UNIVERSIDADE ABERTA

Metodologias estatísticas para análise de dados agrupados em estudos epidemiológicos – Aplicação real aos dados da COVID-19

Tausene Mune Tausene

Mestrado em Bioestatística e Biometria

2024

UNIVERSIDADE ABERTA

Metodologias estatísticas para análise de dados agrupados em estudos epidemiológicos – Aplicação real aos dados da COVID-19

Tausene Mune Tausene

Mestrado em Bioestatística e Biometria

Dissertação orientada pela
Professora Doutora Teresa Paula Costa Azinheira Oliveira,
Departamento de Ciências e Tecnologia, Universidade Aberta

2024

RESUMO

A análise estatística de dados agrupados é frequentemente utilizada na área de epidemiologia para permitir uma conclusão fidedigna em estudos de grupos de indivíduos. Vários são os problemas enfrentados pelos investigadores relacionados com a aplicação de metodologias estatísticas em pesquisa científica. O da presente pesquisa é explorar a aplicação eficaz das metodologias estatísticas na análise de dados agrupados em epidemiológicos, especialmente em relação à COVID-19. A metodologia da pesquisa adotou uma abordagem exploratória, combinando revisão bibliográfica detalhada com análise prática dos dados utilizando o software R. Os dados da COVID-19 foram obtidos da base Our World in Data, permitindo análises exploratórias, correlacionais e de modelagem, incluindo análises espaciais para entender a propagação da doença em Moçambique. São destacadas técnicas como regressão logística, análise de variância e modelos lineares generalizados, visando identificar padrões de propagação, fatores de risco e avaliar o impacto das intervenções de saúde pública. Os resultados da pesquisa demonstraram a eficácia dessas metodologias na análise dos dados da COVID-19 em Moçambique. Foram identificados padrões de propagação da doença, fatores de risco significativos e a eficácia de algumas intervenções de saúde pública. A pesquisa contribui significativamente para o avanço do conhecimento sobre análise de dados agrupados e epidemiologia da COVID-19 em Moçambique, fornecendo subsídios valiosos para a compreensão e enfrentamento da pandemia.

Palavra-Chave: Metodologias estatísticas, análise de dados agrupados, estudos epidemiológicos, COVID19.

Abstract

Statistical analysis of grouped data is frequently used in the field of epidemiology to allow reliable conclusions in studies of groups of individuals. Many are the challenges faced by researchers regarding the application of statistical methodologies in scientific research. The aim of this present study is to explore the effective application of statistical methodologies in the analysis of grouped data in epidemiological studies, particularly concerning COVID-19. The research methodology adopted an exploratory approach, combining detailed literature review with practical data analysis using the R software. COVID-19 data were obtained from the Our World in Data database, allowing for exploratory, correlational, and modeling analyses, including spatial analyses to understand the disease spread in Mozambique. Techniques such as logistic regression, analysis of variance, and generalized linear models are highlighted, aiming to identify propagation patterns, risk fators, and evaluate the impact of public health interventions. The research results demonstrated the effectiveness of these methodologies in analyzing COVID-19 data in Mozambique. Propagation patterns of the disease, significant risk fators, and the effectiveness of some public health interventions were identified. The research significantly contributes to advancing knowledge on grouped data analysis and COVID-19 epidemiology in Mozambique, providing valuable insights for understanding and addressing the pandemic.

Keyword: Statistical methodologies, grouped data analysis, epidemiological studies, COVID19.

DEDICATÓRIA

Dedico este trabalho a minha família, em especial ao meu filho Christian Mune Tausene.

AGRADECIMENTOS

Este trabalho é correspondente a uma meta académica que me propus a cumprir, no contexto do Mestrado em Bioestatística e Biometria, e a sua conclusão é o resultado de apoio de várias pessoas que me acompanharam durante este processo de estudo.

O meu agradecimento primordial vai para a minha orientadora, Professora Teresa Oliveira, por todo apoio e compreensão ao longo deste percurso, prazer de trabalhar durante a parte curricular e dissertativa do mestrado.

Agradeço igualmente a Universidade Aberta, todos Colaboradores e Professores de quem tive o prazer de ser um aluno durante a minha caminhada de aprendizagem, especialmente os do Mestrado em Bioestatística e Biometria.

A minha família pelo apoio incondicional, compreensão e motivação, sobretudo em momentos difíceis. Um especial agradecimento aos meus avós, Albano Mune Tausene e Chapanja Mozolande, minha companheira Teresa Mogas que sempre tiveram uma palavra de apoio e motivação para todos os dias e etapas da minha vida, estudos e trabalho, mesmo nos mais cansativos ou menos produtivos, em que eu duvidei da minha capacidade e recebi do seu lado apoio.

Agradeço aos meus colegas do Mestrado que em muito contribuíram para a minha chegada até aqui. O agradecimento também é extensivo aos amigos que, direta ou indiretamente, contribuíram para a concretização deste objetivo.

ÍNDICES

Índice

Índice de Figuras

LISTA DE ABREVIATURAS, SIGLAS OU ACRÓNIMOS

ANOVA	Análise de Variância
EE	Estado de Emergência
IEA	Associação Internacional de Epidemiologia
ICTV	Comité Internacional de Taxonomia de Vírus
MLG	Modelos Lineares Generalizados
OMS	Organização Mundial da Saúde

CAPÍTULO I

1. ENQUADRAMENTO

1.1. Introdução

O presente capítulo contém a caracterização dos principais temas do trabalho, abrangendo sobre metodologias estatísticas de análise de dados, estudos epidemiológicos e COVID-19. Sobre a COVID-19, é abordado a sua génese, problemática, a gestão de dados, análises e entre outros aspectos.

Todos dias nos deparamos com informação estatística sobre áreas tão diversas como a educação, economia, a medicina, o desporto ou a política. A nossa vida é em larga medida governada por dados que, conscientemente ou não, utilizamos na tomada de decisões.

Somos permanentemente confrontados com ocorrência de vários problemas de saúde pública como as epidemias, que a curto, médio e longo prazo exigem respostas científicas em relação a sua ocorrência, transmissão e propagação, que usualmente transcende os limites geográficos e populacionais próprios de um surto.

Os dados para a medição da saúde provêm de diversas fontes, motivo pelo qual devem ser considerados os aspetos relacionados como a invalidez, qualidade, integridade e cobertura dos próprios dados e suas fontes (Organização Pan-Americana da Saúde, 2010).

Os dados, quantitativos ou qualitativos, que se obtêm e se registram dos serviços de saúde e das estatísticas vitais representam a "matéria prima" para o trabalho epidemiológico. Quando os dados são incompletos ou inconsistentes, serão obtidas medidas enviesadas ou inexatas e as intervenções derivadas do seu uso não serão efetivas.

A quantificação dos problemas de saúde na população requer procedimentos e técnicas estatísticas diversas, algumas delas de relativa complexidade. Dadas as características de múltiplos fatores dos problemas de saúde, as técnicas qualitativas são também valiosas para aproximar-se do conhecimento dos determinantes da saúde. É por isso que existe a necessidade de incorporar, de forma dialética, métodos e técnicas quantitativas e qualitativas que permitam estudar os diversos componentes dos objetos de estudo.

Assim, esta pesquisa tem como objectivo explorar como as metodologias estatísticas para análise de dados agrupados podem ser aplicadas de forma eficaz na análise de dados em situações epidemiológicas, concretamente a de COVID-19. Especificamente, busca compreender como técnicas como a regressão logística, análise da variância e os GLMs podem contribuir para identificação de padrões de propagação, a identificação de factores de risco e a avaliação do impacto das intervenções de saúde pública.

Em relação ao tema em estudo, de antemão sabe-se que existem várias abordagens para a análise de dados (agrupados ou não agrupados), dependendo dos objetivos e da natureza dos dados em questão. A análise descritiva é muitas vezes o primeiro passo, envolvendo a criação de resumos estatísticos para entender a distribuição dos dados.

O foco deste trabalho de pesquisa são os dados agrupados. Entretanto, a análise de dados agrupados é uma ferramenta essencial na investigação epidemiológica, permitindo a compreensão dos padrões de propagação de doenças e a identificação de factores de risco em populações afetadas. Neste contexto, várias metodologias estatísticas são empregadas para explorar a relação entre as variáveis explicativas e de interesse, de modo a fornecer informações e intervenções direcionadas.

São metodologias estatísticas de análise de dados agrupados análise de Regressão Logística, Análise de Variância (ANOVA), Séries Temporais, Modelos Lineares Generalizados, modelos de Sobrevivência, modelos de Mistura Multinomial, análise de Sobrevivência Multinível, modelos Hierárquicos Bayesianos, modelos de Séries Temporais Espaciais, modelos de rede e modelos de equações estruturais.

Segundo Alvarenga (2015), dentre as metodologias estatísticas anteriormente descritas as mais utilizadas na análise de dados agrupados e que serão objecto de estudo nesta dissertação de mestrado destacam-se Análise de Variância (ANOVA), análise de regressão logística, Séries temporais, os Modelos Lineares Generalizados (GLMs) e os modelos de Sobrevivência.

O estudo sobre metodologias estatísticas para análise de dados agrupados, sobretudo em estudos epidemiológicos é crucial devido aos desafios apresentados na recolha e análise de dados. Compreender a propagação do vírus, por exemplo, os fatores de risco e impactos na saúde pública requer uma abordagem estatística sólida e adequada aos dados agrupados.

Na pesquisa epidemiológica, a análise de dados agrupados desempenha um papel importante, fornecendo informações valiosas sobre a disseminação de doenças e a eficácia das medidas de controle. Estudos epidemiológicos utilizam com frequência modelos materiais para prever cenários e implementar medidas de contenção de doenças (Heidecher de Oliveira, 2022).

A análise estatística de dados (sobretudo os agrupados) em tais estudos apresenta desafios metodológicos que devem ser abordados para garantir resultados confiáveis e conclusões sólidas. Um dos grandes e principais desafios em relação a análise de dados agrupados em estudos epidemiológicos têm a ver com a influência do agrupamento na variabilidade dos dados e nos resultados obtidos.

Quando os dados são agrupados ocorre perda de informação individual, pois as observações dentro de cada grupo são resumidas em estatísticas agregadas, como médias, taxas ou proporções. Contudo, esta proporção pode reduzir a precisão da estimativa e afetar a validade das inferências estatísticas realizadas.

Outra questão de suma importância relacionada ao problema é a presença de correlações intra-classe, isto é, a tendência de observações dentro de um mesmo grupo serem mais semelhantes entre si do que em grupos diferentes. Essas correlações podem surgir de fatores compartilhados pelos indivíduos dentro do mesmo grupo, como características genéticas, comportamentais ou ambientais.

Erros na análise e a subestimação da variabilidade real dos dados podem ser resultantes da não consideração dessas correlações, levando a intervalos de confiança muito estreitos e conclusões equivocadas. Além disso, pode introduzir-se enviesamentos nos resultados dos estudos epidemiológicos por meio da estrutura de agrupamento, especialmente se os grupos não representam adequadamente a população estudada ou se houver desigualdades no tamanho dos grupos.

De acordo com Ahlbom (2005), a agregação de dados pode levar a ignorar a heterogeneidade intra-regional e, consequentemente, a conclusões errôneas. Portanto, o autor enfatiza a importância de métodos estatísticos que levem em conta a estrutura hierárquica dos dados para obter resultados mais precisos e confiáveis.

White (2009) refere a necessidade de abordar adequadamente a questão dos dados ausentes ou omissos em estudos epidemiológicos. O autor argumenta que por haver dados omissos isso pode comprometer a validade e a precisão das análises, e métodos apropriados de imputação são essenciais para lidar com essa questão. Ele destaca a imputação múltipla com equações de predição como uma abordagem robusta para lidar com dados omissos em estudos agrupados.

Das epidemias que assolam o mundo nos dias atuais, a pandemia da COVID-19 teve um impacto negativo muito grande e este tem sido atualmente um dos temas de estudos epidemiológicos em todo mundo. De acordo com Oliveira (2021), a COVID-19 é uma doença com alta potência de transmissão/propagação. Contudo, a pandemia de COVID-19 ilustra a importância da aplicação dessas metodologias estatística em epidemiologia. No contexto específico de Moçambique, compreender os padrões de propagação e os fatores determinantes da disseminação da COVID-19 é importante para o desenvolvimento de estratégias de respostas eficazes. Com a pandemia, diversos estudos foram desenvolvidos para encontrar solução para o problema. Na perspetiva estatística foi possível obter dados e produzir diversas análises capazes de ajudar na interpretação do comportamento da doença no espaço de ocorrência por meio de análise exploratória. Ver Teodoro et al. (2023) e Teodoro et al. (2024).

A natureza destes dados, como o número de casos em uma região geográfica específica ou em um determinado período de tempo, pode afetar a precisão e interpretação dos resultados. Porém, é fundamental aplicar as metodologias e ferramentas estatísticas adequadas para lidar com essa complexidade e obter resultados significativos.

Como parte de desenvolvimento do trabalho, será utilizado a técnica de análise estatística de dados agrupados utilizando o software R. Os dados sobre COVID-19 estão disponíveis na base de dados Our World in Data. A escolha para uso do software R foi motivada por sua reconhecida versatilidade e aplicabilidade em diversas áreas de pesquisa, uma vez que, sua flexibilidade e recursos robustos permitem explorar, visualizar e analisar dados, ajustar modelos matemáticos e obter resultados precisos.

O R studio é uma plataforma de computação estatística amplamente utilizada e reconhecida pela flexibilidade e poder analítico. Optou-se por usar R studio para realizar as análises estatísticas devido à sua capacidade de manipular dados complexos, executar uma ampla gama de técnicas estatísticas e gerar visualizações gráficas de alta qualidade. Além disso, sua natureza de código aberto permite uma maior transparência e replicabilidade das análises realizadas.

Ao desenvolver o tema sobre metodologias estatísticas para análise de dados e explorar base de dados abrangentes, espera-se contribuir significativamente para o avanço do conhecimento sobre análise de dados agrupados e epidemiologia da COVID-19 em Moçambique.

Objetivo Geral

- Estudar as metodologias estatísticas na análise de dados agrupados em estudos epidemiológicos.

Objetivo específico

- Comparar diferentes abordagens de análise de dados epidemiológicos, com ênfase aos dados agrupados;
- Identificar e descrever as principais metodologias estatísticas utilizadas na análise de dados agrupados em estudos epidemiológicos, especialmente em relação à COVID-19;
- Aplicar metodologias estatísticas de análise de dados agrupados em estudos epidemiológicos reais relacionados a COVID-19, utilizando o software R.

1.2. Estrutura

A pesquisa encontra-se dividida em quatro (4) capítulos, a saber: Introdução, revisão de literatura, metodologia, resultados e discussão. Na introdução será apresentado a delimitação do tema em estudo e os objetivos.

Na revisão de literatura, será apresentado o quadro conceptual teórico sobre o tema, olhando sobre diferentes abordagens relacionadas a metodologias estatísticas de análise de dados em estudos epidemiológicos.

No terceiro capítulo, será apresentado a metodologia do estudo em causa, explicando o procedimento metodológico através do qual se orientou a pesquisa e apresentação dos diferentes métodos de análise estatística dos dados em diferentes contextos.

No último capítulo serão analisados e discutidos os resultados, com base na metodologia de análise escolhida durante o desenvolvimento da presente dissertação.

CAPÍTULO II: DESENVOLVIMENTO DE CONTEÚDOS TEÓRICOS ENVOLVIDOS. DETALHE DAS METODOLOGIAS ESTATÍSTICAS A USAR E INDICAÇÃO DE ALGUNS ESTUDOS CONSULTADOS

2. EPIDEMIOLOGIA E ANÁLISE DE DADOS

Nesta secção são apresentadas informações sobre o tema consultados em estudos relevantes e que abordam sobre as metodologias estatísticas na análise de dados agrupados em estudos epidemiológicos, COVID-19 e a aplicação prática da ferramenta R.

2.1.Conceitos básicos de epidemiologia e sua importância na saúde pública

Epidemiologia estuda os padrões de saúde e doença, além dos fatores associados em nível populacional. A epidemiologia pode ser definida como o estudo da distribuição e dos determinantes de situações ou eventos relacionados com a saúde em populações humanas, e a aplicação destes estudos para o controlo das doenças e outros problemas de saúde. A palavra epidemiologia deriva do étimo grego (*epi)* = entre, (demos) = pessoas e (*logos*) = doutrina (Organização Mundial da Saúde, 2014).

De acordo com Rouquayrol *et. al* (2013) citado por Gomes (2015) a epidemiologia pode ser definida como a ciência que estuda o processo saúde doença em coletividades humanas, analisando a distribuição e os fatores determinantes das enfermidades, danos à saúde e eventos associados à saúde coletiva, propondo medidas específicas de prevenção, controle ou erradicação de doenças e fornecendo indicadores que sirvam de suporte ao planeamento, administração e avaliação das ações de saúde.

De acordo com a Associação Internacional de Epidemiologia (IEA), o termo epidemiologia refere-se ao "estudo dos fatores que determinam a frequência e a distribuição das doenças nas coletividades humanas", ou seja, "enquanto a clínica se dedica ao estudo da doença no indivíduo, analisando caso a caso, a epidemiologia debruça-se sobre os problemas de saúde em grupos de pessoas, às vezes grupos pequenos, na maioria das vezes envolvendo populações numerosas" (Camargo e Villar, 2021).

Segundo Gomes (2015), a epidemiologia tem como princípio básico o entendimento de que os eventos relacionados à saúde, como doenças, seus determinantes e o uso de serviços de saúde, não se distribuem ao acaso entre as pessoas. Há grupos populacionais que apresentam mais casos de certo agravo e há outros que morrem mais por determinada doença.

Tais diferenças ocorrem porque os fatores que influenciam o estado de saúde das pessoas não se distribuem igualmente na população, portanto, acometem mais alguns grupos do que outros (Pereira, 2013).

A epidemiologia tornou-se ao longo dos anos uma ciência ampla que abriga inúmeras áreas do conhecimento e muitas subdivisões, tais como (Pereira, 2013 citado por Gomes, 2015):

- epidemiologia clínica;
- epidemiologia investigativa;
- epidemiologia nutricional;
- epidemiologia de campo;
- epidemiologia descritiva;
- etc.

Segundo a autora acima citada, em linhas gerais a epidemiologia apresenta 3 principais áreas de atuação a saber:

- Descrição das condições de saúde da população por meio da construção de indicadores de saúde.
- Investigação dos fatores determinantes da situação de saúde.
- Avaliação do impacto das ações para alterar a situação de saúde.

De acordo com Pereira (2013) citado por Gomes (2015), são principais áreas de atuação da epidemiologia as seguintes:

- ***Diagnóstico da situação de saúde***

O diagnóstico da situação de saúde consiste na recolha sistemática de dados sobre a saúde da população, informações demográficas, econômicas, sociais, culturais e ambientais, que servirão para compor os indicadores de saúde. Apesar de parecer uma tarefa simples, o diagnóstico da situação de saúde apresenta minúcias importantes para a sua realização.

O diagnóstico de situação de saúde tem como principais objetivos a construção de um plano de ação em saúde que venha a minimizar os problemas identificados e a formulação de hipóteses

sobre os fatores envolvidos na construção e manutenção de um cenário epidemiológico. Tais hipóteses poderão e deverão ser testadas (Gomes, 2015).

- ***Investigação etiológica***

A investigação dos agentes etiológicos das doenças sempre foi, desde os seus primórdios, um objetivo prioritário da epidemiologia. No final do século XIX até meados do século XX, foi dado um grande enfoque às doenças infetocontagiosas, tendo em vista a evolução da microbiologia e a grande prevalência de doenças infeciosas no mundo (Pereira, 2013 citado por Gomes, 2015).

Inicialmente, foi adotada uma abordagem uni-causal para o processo de adoecimento, ou seja, toda doença apresentava um agente etiológico que, uma vez identificado, poderia ser combatido. Tal abordagem solucionou vários problemas de saúde pública:

Controle de doenças por meio da vacinação	Controle de doenças por meio do tratamento
Exemplo: poliomielite, varíola, febre tifoide.	Exemplo: tuberculose, hanseníase

Tabela 1. problemas de saúde pública. *Fonte: (Gomes, 2015).*

Tal abordagem também serviu para doenças não infeciosas, como é o caso do "bócio endêmico", que foi praticamente eliminado pela iodação do sal de cozinha. Com a evolução do conhecimento científico, a abordagem uni-causal não foi capaz de explicar as causas de várias doenças, surgindo assim a abordagem multicausal para a investigação dos agentes etiológicos.

- ***Determinação de risco***

O conceito de risco na epidemiologia está diretamente associado à ocorrência de doenças na população, fugindo um pouco das conceções de causalidade individual. Em epidemiologia, o risco pode ser definido como "o grau de probabilidade da ocorrência de um determinado evento" (Pereira, 2013 citado por Gomes, 2015); ou como "a probabilidade de ocorrência de um resultado desfavorável, um dano ou um fenômeno indesejado.

Deste modo, estima-se o risco ou probabilidade de que uma doença exista por meio dos coeficientes de incidência e prevalência" (CLAP-OPAS/OMS, 1988). Existem várias medidas do para o cálculo do risco na epidemiologia, mas as duas mais importantes são as seguintes medidas

de associação: Risco Relativo ou Razão de Risco (RR) e Razão de Chances ou Odds Ratio (OR) (Gomes, 2015). Até ao século XX, os estudos epidemiológicos tinham como objeto central de estudo as doenças infeciosas, mas atualmente as doenças não transmissíveis são também alvo de estudo da epidemiologia (Martcheva, 2015 citado por Theodoro, 2022).

Uma característica que diferencia doenças infeciosas das demais, como a influenza, é o facto de indivíduos infetados causarem diretamente novas infeções por meio de microrganismos. Uma doença infeciosa é resultante de um agente microbiano patogênico, este agente pode ser bateriano, viral, fúngico, parasitário ou podem ser proteínas tóxicas (Theodoro, 2022). As doenças infeciosas causam graves crises sanitárias e económicas, e a modelação matemática permite orientar respostas de políticas públicas e individuais para o controle dessas doenças. Muitos modelos matemáticos que estudam a dinâmica populacional da COVID-19 vêm sendo trabalhados e discutidos desde o início da pandemia, tal como se pode constatar em Leal et al (2023). No entanto, na sua maioria, são modelos clássicos SIR (suscetível-infetado-recuperado/removido) e modelos SEIR (suscetível-exposto-infecioso-recuperado/removido).

As doenças infeciosas podem ser classificadas em agudas ou crônicas. As infeções agudas são aquelas em que a resposta imune é rápida e remove os patógenos após um curto período de tempo (dias ou semanas), como a influenza. Já doenças infeciosas crônicas perduram por longos períodos (meses ou anos) como a herpes. As doenças agudas, como a COVID-19, podem ser descritas pelo modelo SIR, estudado inicialmente por Kermack e McKendrick (ROHANI; Keeling, 2008; Kermack; Mckendrick, 1927 citado por Theodoro, 2022).

Quando uma doença se espalha em uma população, a mesma pode ser dividida em compartimentos, no modelo SIR existem 3 compartimentos:

- Os indivíduos saudáveis, mas que podem contrair a doença. Estes são chamados suscetíveis (S);
- A classe dos que contraíram a doença são chamados de infetados (I), que nesse caso vamos considerar também infeciosos, no qual o indivíduo é capaz de transmitir a doença;
- Os indivíduos que se recuperam e não podem contrair a doença novamente são chamados removidos ou recuperados (R).

O número de indivíduos em cada compartimento muda ao longo do tempo da epidemia, ou seja, $S(t), I(t) e\ R(t)$ estão em função do tempo, e o tamanho da população $N = S(t) + I(t) + R(t)$.

Diferença entre dados agrupados e dados individuais

De acordo com Camargo e Villar (2021), os dados podem ser encontrados em diferentes formatos, sendo comumente classificados em dois principais grandes grupos:

- *os dados estruturados*, normalmente de estrutura predeterminada; e os
- *dados não estruturados*, que não possuem estruturas preestabelecidas, alinhadas ou bem definidas, como é o caso de imagens.

Os dados estruturados, são conjuntos de informações organizadas em colunas (atributos, variáveis, features etc.) e linhas (registos, itens, observações etc.). São dados mais comummente encontrados diretamente em bancos de dados, por exemplo ou arquivos com algum tipo de separação entre as colunas, Excel, arquivos com campos de tamanho fixo (Guerra, Oliveira e McDonells, 2018).

Os dados não estruturados, não têm uma estrutura previsível, ou seja, cada conjunto de informações possui uma forma única. Geralmente são arquivos com forte teor textual. Não podemos dizer que são dados "desorganizados", e sim que são organizações particulares para cada conjunto de informações. Analisar este tipo de dados é muito mais complexo e exige conhecimento avançado em mineração de dados. Apesar disso, é o tipo de dados mais abundante na realidade.

Dados semiestruturados, são dados que também possuem uma organização fixa, porém não seguem o padrão de estrutura linha/coluna, ou seja, seguem uma estrutura mais complexa e flexível, geralmente hierárquica, estruturada em *tags* ou marcadores de campos. São exemplos de arquivos semiestruturados: JSON, XML, HTML e YAML (Guerra, Oliveira e McDonells, 2018).

Dados semiestruturados, algumas vezes, são facilmente transformados em dados estruturados. Destes, a preferência por dados estruturados na presente pesquisa está relacionada à sua maior facilidade de análise.

Quando se deseja modelar o comportamento dos dados para efeitos de tomada de decisão, podem ser empregues dois métodos: os que estão pautados predominantemente em modelos pré-concebidos, como é o caso dos métodos que se utilizam maioritariamente do histórico de dados e da aplicação de técnicas estatísticas; e os que consistem na aplicação de *Machine Learning*, ou

seja, em métodos apoiados em programação dinâmica e inteligência artificial que permitem a exploração de grandes quantidades de dados na procura por padrões de consistência (Camargo e Villar, 2021).

2.2. Tipos comuns de estudos epidemiológicos e suas características

De acordo com Barreto e Lima-Costa (2003), os estudos epidemiológicos podem ser classificados em observacionais e experimentais. De uma maneira geral, os estudos epidemiológicos observacionais podem ser classificados em descritivos e analíticos.

Estudos Descritivos: sem grupo de comparação

Os estudos descritivos têm por objetivo determinar a distribuição de doenças ou condições relacionadas à saúde, segundo o tempo, o lugar e/ou as características dos indivíduos (Barreto & Lima-Costa, 2003). Ou seja, responder à pergunta: quando, onde e quem adoece?

A epidemiologia descritiva pode fazer uso de dados secundários (dados pré-existentes de mortalidade e hospitalizações, por exemplo) e primários (dados coletados para o desenvolvimento do estudo).

A epidemiologia descritiva examina como a incidência (casos novos) ou a prevalência (casos existentes) de uma doença ou condição relacionada à saúde varia de acordo com determinadas características, como sexo, idade, escolaridade e renda, entre outras (Barreto & Lima-Costa, 2003).

- *Ecológicos* - Nos Estudos Ecológicos as medidas usadas representam características de grupos populacionais. Portanto a unidade de análise é a população e não o indivíduo. A limitação principal do estudo ecológico é que a relação entre o fator de exposição e o evento pode não estar ocorrendo ao nível do indivíduo.;
- *Transversais* - Nos Estudos Transversais, cada indivíduo é avaliado para o fator de exposição e a doença em determinado momento. Muitas vezes o estudo transversal é realizado apenas com objetivo descritivo sem nenhuma hipótese para ser avaliada.

Estudos Analíticos: com grupo(s) de comparação

- *Observacionais* -Nos Estudos Observacionais não existe nenhuma manipulação do fator de estudo. Podem ser subdivididos em descritivos e analíticos. De acordo com Kjellström,

Bonita, & Beaglehole (2010), os estudos observacionais permitem que a natureza determine o seu curso: o investigador mede, mas não intervém. Esses estudos podem ser descritivos e analíticos: um estudo descritivo limita-se a descrever a ocorrência de uma doença em uma população, sendo, frequentemente, o primeiro passo de uma investigação epidemiológica; um estudo analítico aborda, com mais profundidade, as relações entre o estado de saúde e as outras variáveis.

- *Coorte (prospetivo ou histórico)* - Uma coorte é um grupo de indivíduos definido a partir de suas características pessoais (idade, sexo, etc.), nos quais se observa, mediante exames repetidos, a aparição de uma enfermidade (ou outro desfecho) determinada. No Estudos de Coorte Longitudinais, cada indivíduo é avaliado inicialmente para os fatores de exposição ou característica.;
- *Caso-controle (retrospetivo)* - Nos estudos de caso-controle avalia-se inicialmente quem tem (caso) ou não (controle) o evento de interesse. Os estudos de caso - controle são sempre, obviamente, estudos comparados.

Estudos experimentais: Ensaio Clínico ou do tipo "*crossover*".

Estudos experimentais ou de intervenção envolvem a tentativa de mudar os determinantes de uma doença, tais como uma exposição ou comportamento, ou cessar o progresso de uma doença através de tratamento. São similares a experiências realizadas em outras ciências. Ver (Kjellström, Bonita, & Beaglehole, 2010).

2.3. Sequência lógica dos estudos epidemiológicos

Na investigação epidemiológica, o atual estágio de conhecimento determina, muitas vezes, a conceção mais lógica do desenho do estudo. Normalmente, há uma progressão dos estudos de desenhos de estudo para geração de hipóteses até desenhos de estudo para a testagem da hipótese (OMS, 2014).

Por exemplo, as hipóteses são muitas vezes geradas por métodos como a vigilância, notificação de casos, séries de casos ou estudos ecológicos. Essas hipóteses são depois testadas usando dados provenientes de experiências, de estudos transversais anteriores, de estudos caso controlo ou de estudos de coorte retrospetivos, que podem ser feitos relativamente depressa e são mais acessíveis.

Se estes estudos apoiarem a hipótese, poderá então realizar-se um estudo de coorte prospetivo. Finalmente, em algumas situações, pode ser adequada a realização de um ensaio clínico aleatório.

O fluxograma abaixo ilustra a aplicação dos vários tipos de estudos primários. Em todos os tipos de estudos, a formulação da hipótese deve preceder a análise.

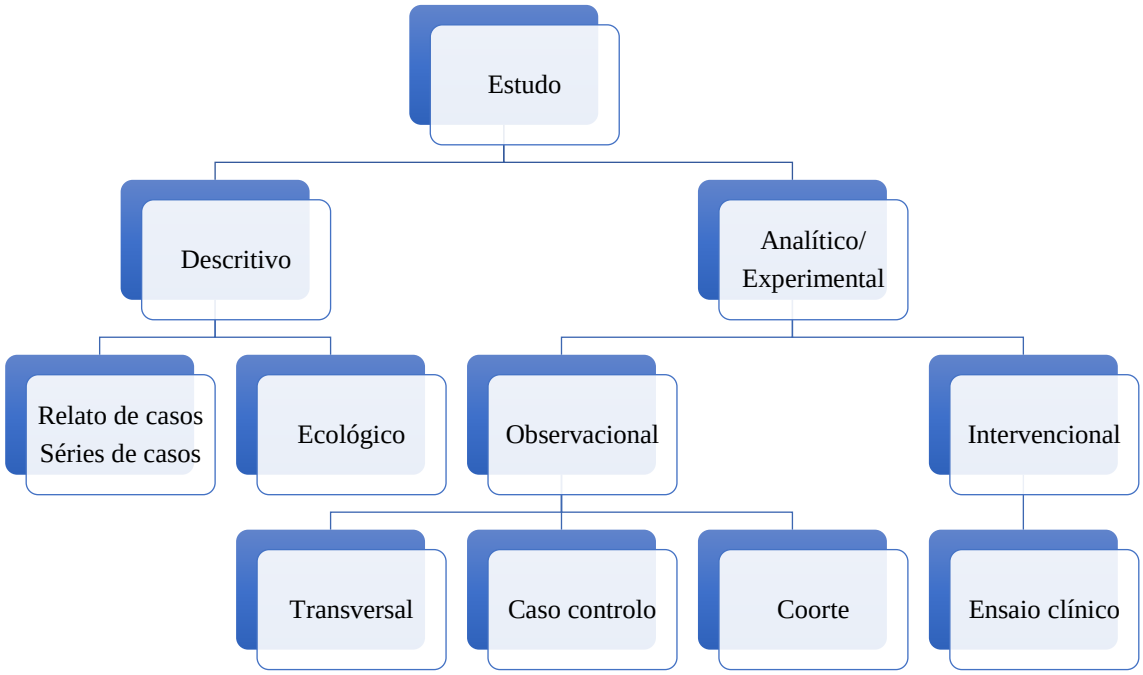


Ilustração 1. O fluxograma ilustrativo da aplicação dos vários tipos de estudos primários. Fonte: OMS (2014).

Pesquisa Clínica

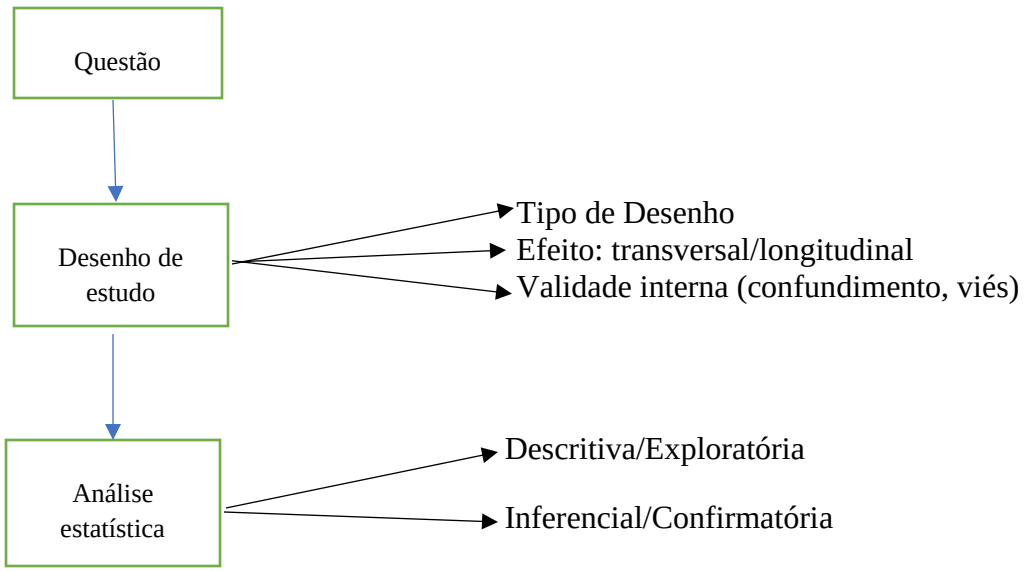


Ilustração 2. Ilustração de pesquisa clínica. (Autor, 2023).

2.4. Métodos essenciais para a epidemiologia aplicada às doenças infeciosas

De acordo com o Centro Europeu de Prevenção e Controlo das Doenças (2022), são Métodos essenciais para a epidemiologia aplicada às doenças infeciosas:

Epidemiologia descritiva

- Descrição dos perfis demográficos das populações, incluindo as pirâmides etárias, e os fatores com impacto na estrutura populacional
- Identificação dos métodos utilizados a nível nacional e internacional para assegurar uma notificação abrangente das doenças infeciosas às agências regulamentadas pertinentes.
- Identificação das fontes disponíveis de dados individuais e agregados sobre doenças infeciosas,
- Cálculo e interpretação das medidas relativas à frequência das doenças (incidência, prevalência, taxas por idade, taxas de letalidade) e tendências nas taxas de doença ao longo do tempo.
- Realização da análise e a comparação das taxas de doença entre regiões, entre populações e ao longo do tempo, utilizando procedimentos de normalização diretos e indiretos, conforme o que for relevante.
- Interpretação das tendências da doença a partir de análises de séries temporais.

- Obtenção e interpretação dados sobre as taxas de doença representados em gráficos e explique-os aos públicos relevantes.

Métodos de investigação epidemiológica

- Realizar uma avaliação crítica da literatura científica utilizando instrumentos estabelecidos, como listas de verificação para revisões sistemáticas, avaliações rápidas e de longo prazo dos riscos, ensaios aleatórios controlados, estudos de coorte, estudos de casos-controlo, avaliações económicas, estudos de diagnóstico e estudos qualitativos, conforme o que for relevante.
- Redigir um protocolo de estudo com informações detalhadas sobre o problema de saúde pública a investigar e as técnicas de investigação adequadas compatíveis com o problema e o contexto.
- Conceber estudos epidemiológicos (p. ex., estudos com base na população, estudos transversais, estudos ecológicos) para investigar a carga da doença numa população, utilizando estratégias de amostragem adequadas.
- Conceber estudos epidemiológicos para investigar os fatores determinantes da doença, para determinar as associações e/ou o nexo de causalidade da doença (p. ex., estudos de coorte, estudos de casos-controlo, estudos transversais, reconhecendo a natureza multifatorial da doença).
- Conceber estudos qualitativos com base nas ciências comportamentais para explorar aspetos qualitativos do impacto das doenças infeciosas nas pessoas, na comunidade e nos serviços de saúde.
- Conceber, testar e avaliar métodos de recolha de dados, incluindo formulários de notificação de casos e questionários.
- Avaliar os instrumentos de estudo e as respetivas propriedades de medição, em particular a sua validade, fiabilidade e aplicabilidade transcultural.
- Explicar e aplicar os conceitos de correlação e associação nos estudos observacionais e aplique critérios relevantes para inferir o nexo de causalidade a partir dos estudos observacionais.

- Reconhecer fontes de enviesamento, de confundimento, de interação e modificação dos efeitos, e a forma de as reconhecer e de fazer ajustamentos em função delas na conceção do estudo e nas técnicas analíticas.
- Calcular e interprete as medidas de efeito a partir dos estudos de coorte, dos estudos de casos-controlo e dos ensaios aleatórios controlados.

3. Dados estatísticos

Dados brutos podem ser definidos como o conjunto completo de dados recolhidos num estudo, antes de qualquer arredondamento, edição ou organização estatística. Eles usam-se, em primeiro lugar, para ajudar os planificadores e os administradores da área de saúde a determinar as necessidades em cuidados de saúde (OMS, 2014).

A recolha de dados é uma atividade primária e crucial para a vigilância epidemiológica. Entre os principais dados recolhidos, podem ser citados os dados demográficos, ambientais, socioeconómicos e outros. Após a etapa de recolha de dados, os mesmos precisam ser registados em banco de dados, que, na maioria das vezes, geram a necessidade de criação dos denominados Sistemas de Informação em Saúde, que têm por função a consolidação de informações em saúde, possibilitando a análise de situações de risco.

O número de casos que caracteriza a presença de uma epidemia, varia segundo o agente infecioso, o tamanho e o tipo da população exposta, a sua experiência prévia com a doença e o tempo e o lugar da ocorrência (Pereira, 2013 citado por Gomes, 2015). A base da ciência de dados é, obviamente, o dado. Portanto, é fundamental ter boas fontes de dados, preferencialmente dados estruturados para iniciar sua análise. Porém, eventualmente recorre-se a fontes de dados não estruturados ou semiestruturados.

Após a recolha de dados, tal como na construção de qualquer outro indicador de saúde, é necessário proceder à sua transformação em informações úteis em saúde. Os métodos utilizados para organizar dados compreendem o arranjo desses dados em subconjuntos que apresentem características similares, como por exemplo a mesma idade (ou "faixa etária"), a mesma finalidade, a mesma escola ou o mesmo bairro de residência. Os dados agrupados podem ser resumidos em tabelas ou gráficos e, a partir desses, podemos obter as estatísticas descritivas associadas, a média, a mediana, o desvio padrão, etc..

Dados organizados em grupos ou categorias/classes podem ser contabilizados numa tabela de "distribuição de frequência", como a que se segue:

Regra de Sturges (Logaritmo)	
Quantidade de dados (n)	Quantidade de Classes (k)
1	1
2	2
3 a 5	3
6 a 11	4
12 a 23	5
24 a 46	6
47 a 93	7
94 a 187	8
188 a 376	9
377 a 756	10

Regra da Potência de 2	
Quantidade de dados (n)	Quantidade de Classes (k)
1 e 2	1
3 e 4	2
5 a 8	3
9 a 16	4
17 a 32	5
33 a 64	6
65 a 128	7
129 a 256	8
257 a 512	9
513 a 1024	10

Bom Senso		
Quantidade de dados (n)	Quantidade MÍNIMA de Classes (k)	Quantidade MÁXIMA de Classes (k)
até 50	5	10
51 a 100	8	16
101 a 200	10	20
201 a 300	12	24
301 a 500	15	30
mais de 500	20	40

Assim, a distribuição de frequência para dados agrupados é a série estatística que condensa um conjunto de dados conforme as frequências ou repetições de seus valores. Os dados encontram-se dispostos em classes ou categorias com as frequências correspondentes.

A distribuição de frequência é composta por seguintes elementos:

- Amplitude total
- Frequência simples absoluta
- Classe
- Intervalo de classe ou amplitude de intervalo de classe
- Limites de classe (limites inferior e limites superior).

3.1. Fontes de dados

As fontes de dados podem ser primárias e secundárias.

Fontes primárias: dizemos que a fonte é primária quando o próprio pesquisador gera a informação. As fontes primárias mais utilizadas são a observação direta do fenômeno, que é um método clássico na pesquisa científica, e o questionário, que é um instrumento de pesquisa usado para levantamento de dados.

Fontes secundárias: São bancos de dados ou arquivos previamente existentes, onde estão armazenadas as informações que serão utilizadas no levantamento, ou seja, os dados já existem e

o pesquisador irá lançar mão deles para desenvolver seu estudo. As fichas de cadastro de estudantes ou de clientes de uma loja de departamentos são exemplos de fontes de dados secundária.

3.2. Gestão de dados e bioestatística

Para melhor gestão dos dados tem que ser ter em conta os seguintes aspetos:

- Distinguir entre as variáveis e as observações e descrever os atributos das variáveis, incluindo tipos de variáveis e nível de medição.
- Descrever os princípios da gestão de dados, incluindo a normalização na recolha de dados pessoais, a compilação eletrónica dos dados e a garantia da validade dos dados numa base de dados.
- Seguir as regras relativas à privacidade dos dados pessoais e os quadros jurídicos em matéria de proteção de dados, demonstrando a proteção e a segurança dos dados em todas as vertentes do trabalho.
- Realizar a gestão de dados e a análise estatística como utilizador independente de, pelo menos, um tipo estatístico de software (p. ex., SPSS, R, STATA, SAS).
- Descrever o conceito básico de probabilidade e aplique procedimentos estatísticos básicos, como estatísticas descritivas e estatísticas básicas de inferência.
- Determinar e interpretar as estimativas pontuais, os intervalos de confiança, as estimativas do risco e os níveis de significância, incluindo os valores de p.
- Descrever os princípios da análise multivariável e da análise de sobrevivência, leve a cabo a análise e interprete os resultados.
- Participar no desenvolvimento e na interpretação dos protocolos estatísticos.

Modelação de doenças

- Comunicar com os especialistas em modelação de doenças para determinar os pressupostos e os processos de modelação preditiva de doenças infeciosas.
- Descrever as aplicações e as limitações da modelação preditiva das doenças infeciosas no planeamento da preparação, na previsão e na orientação para os decisores políticos.
- Comparar e interprete os resultados de diferentes modelos e cenários de doenças infeciosas, tendo em conta os seus pressupostos.

Vários avanços metodológicos e instrumentais vêm sendo propostos e aperfeiçoados, notadamente no que se refere aos Sistemas de Informação em Saúde (Tasca, 1993; Morais, 1994) aos processos de análise da situação (Castellanos,1991, 1991b, 1994) ao planeamento de ações da chamada vigilância da saúde (Teixeira, 1994; Sá & Artmann, 1994), às metodologias de capacitação gerencial (Teixeira, 1992) e de avaliação de sistemas locais de saúde (Barata, 1990 citado por Teixeira, 1999).

A vigilância epidemiológica representa um conjunto de ações estratégicas que proporcionam o conhecimento, a deteção ou prevenção de qualquer mudança nos fatores determinantes e condicionantes de saúde individual ou coletiva, com a finalidade de recomendar e adotar as medidas de prevenção e controle das doenças e agravos (Dias, *et al*., 2016).

O uso das ferramentas e ações de vigilância epidemiológica torna possível romper com as cadeias de transmissão das enfermidades, bem como sustenta e valoriza o serviço de modo que a informação em saúde possa ser uma ponte para a integração entre o olhar clínico e o epidemiológico (Gomes, 2015).

3.3. Modelos para representar os aspetos etiológicos do processo saúde-doença

Existem vários modelos para se representar o processo saúde-doença, principalmente quando este está associado aos aspetos etiológicos das doenças. Portanto, apresenta-se alguns modelos, realçando que não existe um modelo ideal de representação deste processo, mas sim um que melhor se ajuste ao cenário individual ou coletivo para a ocorrência da doença. Segundo Pereira (2013) citado por Gomes (2015), como principais modelos podem ser citados:

- cadeia de eventos;
- modelos ecológicos;
- rede de causas;
- múltiplas causas - múltiplos efeitos;
- abordagem sistêmica da saúde;
- etiologia social da doença.

3.4. Medidas de frequência de doenças

Como o termo sugere, medidas de frequências de doenças são indicadores construídos com o objetivo de medir a ocorrência de doenças na população. Em termos gerais, as principais medidas em saúde são:

- índices;
- coeficientes;
- taxas;
- indicadores.

Segundo Lima, Pordeus e Rouquayrol (2013), estas medidas podem ser definidas de acordo com os conceitos abaixo:

- **Índice**: termo genérico apropriado para referir-se a todos os descritores da vida e da saúde; inclui todos os termos numéricos existentes e incidentes que trazem a noção de grandeza.
- **Coeficientes**: são medidas secundárias que, ao serem geradas pelos quocientes entre medidas primárias de variáveis independentes, deixam de sofrer influência dessas variáveis para expressar somente a intensidade dos riscos de ocorrência. Em outras palavras, trata-se da frequência com que um evento ocorre na população.
- **Taxas:** são medidas de risco aplicadas para cálculos de estimativas e projeções de incidências e prevalências em populações de interesse.
- **Indicadores:** são os índices críticos capazes de orientar a tomada de decisão em prol das evidências ou providências.

3.5. Potenciais erros nos estudos epidemiológicos

Uma medição tem imperfeições que dão origem a um erro no resultado da medição. O erro costuma ser classificado em dois componentes: erro aleatório e erro sistemático (Tabacnicks, 2009).

Erro aleatório

De acordo com Tabacnicks (2009), o erro aleatório tem origem em variações imprevisíveis também chamadas efeitos aleatórios. Esses efeitos são a causa de variações em observações repetidas do mensurando. O erro aleatório é a divergência, devida apenas ao acaso, de uma

observação numa amostra, do verdadeiro valor de uma população, que leva à falta de precisão na medição de uma associação. Há três principais fontes de erro aleatório: variação individual/biológica, erro de amostragem e erro de medição. O erro aleatório pode ser minimizado, mas nunca poderá ser completamente eliminado, uma vez que apenas se pode estudar uma amostra da população; ocorre sempre uma variação individual e nenhuma medição é perfeitamente rigorosa. O erro aleatório pode ser reduzido através da medição cuidadosa da exposição e do resultado, da adequada seleção dos participantes do estudo e da inclusão de uma amostra de tamanho suficiente. Ainda de acordo com Tabacnicks (2009), o desvio padrão da média não é o erro aleatório da média, mas representa uma medida da incerteza da média devido aos efeitos aleatórios.

Erro sistemático (Viés ou enviesamento)

O viés ou erro sistemático ocorre quando há tendência para produzir resultados que divergem sistematicamente dos valores verdadeiros ou reais. Diz-se que um estudo é de precisão elevada, quando tem apenas um pequeno viés ou erro sistemático. O erro sistemático, em geral, não pode ser eliminado, mas pode eventualmente ser reduzido ou, caso seja identificado, deve ser corrigido (Tabacnicks, 2009).

O viés (ou erro sistemático) pode conduzir a uma sobre- estimativa ou uma sub- estimativa da força de uma associação. As fontes de viés ou enviesamento em epidemiologia são muitas, tendo sido identificados mais de 30 tipos específicos de viés.

De acordo com a OMS (2014), os principais tipos de viés são:

- Viés de seleção
- Viés de informação
- Viés de confusão ou de confundimento.

Víes de seleção

O Viés de seleção ocorre quando há uma diferença sistemática entre as características dos participantes incluídos no estudo e as características da população fonte.

Viés de informação (também chamado Viés de medição)

O viés de informação ocorre quando surgem problemas de qualidade (rigor) na recolha, registo, codificação ou análise de dados entre grupos de comparação.

Confusão/confundimento

Num estudo sobre a associação entre exposição e uma causa (ou fator de risco ou de proteção) e a ocorrência de uma doença, confundimento poderá ocorrer se existir outro fator na população de estudo e que esteja associado com ambos, a doença e o fator inicial em estudo.

Se este segundo fator estranho estiver desigualmente distribuído entre os subgrupos de exposição, surge um problema. O confundimento ocorre quando os efeitos de dois fatores de risco ou protetores não tiverem sido separados, concluindo-se, portanto, de forma incorreta que o efeito se deve mais a uma variável do que a outra (OMS, 2014).

4. Pandemia da COVID-19 - Aspetos gerais

É o nome dado à ocorrência epidêmica caracterizada por larga distribuição espacial, atingindo várias nações. Em outras palavras, a pandemia pode ser tratada como a ocorrência de uma série de epidemias localizadas em diferentes regiões e que ocorrem em vários países ao mesmo tempo (Rouquayrol; Barbosa; Machado, 2013 citado por Gomes, 2015).

O novo coronavírus em humanos desencadeou uma pandemia no mundo, devido a sua fácil transmissão e o intenso fluxo de pessoas e transportes nos dias atuais. Em março de 2020, à época com 118.326 casos confirmados e 4.292 óbitos, a Organização Mundial da Saúde (OMS) declarou a disseminação da COVID-19 como pandemia (OMS, 2020 citado por Fagundes, 2021).

O vírus da doença corona (COVID-19) teve seu gene em Wuhan, China em dezembro e rapidamente espalhou-se pela China e demais países e regiões no mundo. Denominado como SARSCoV2, resultou em uma pandemia, pois a doença causada pelo vírus; a COVID-19, se espalhou rapidamente por todas as nações do mundo e resultou em milhões de mortes (Wangping, *et al*., 2020; Proença & Schmidt, 2021).

4.1. Transmissão e manifestação clínica

Até ao final do mês de março de 2021, em todo o mundo tinha sido reportado um cumulativo de 128 milhões de casos de COVID-19. Nessa altura, a Europa era o continente com maior número

de casos cumulativos, com 39 milhões, seguido da América do Norte, com 35 milhões de casos (ONS, 2021). O continente Africano era na altura o segundo continente com menor número cumulativo de casos, com 4 milhões de casos (figura).

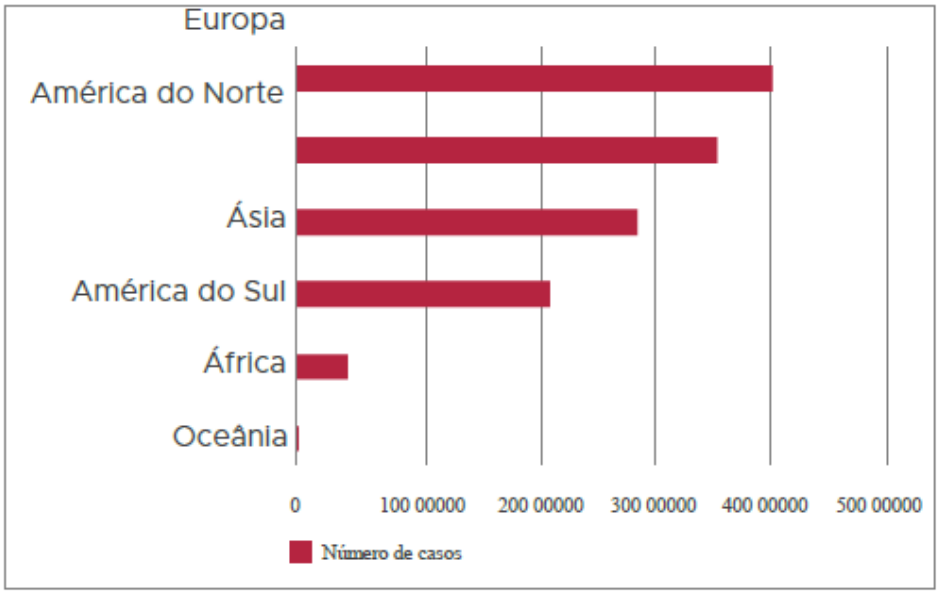


Figura 1.Número cumulativo de casos de COVID-19 por continente, até 31 de março de 2021. Fonte: ONS, 2021.

Por se tratar de uma infeção respiratória aguda, o SARS-CoV-2 dissemina-se principalmente por gotículas, secreções respiratórias e contato direto com o paciente infetado. Diante dessa perspetiva, destaca-se a capacidade de o vírus ser transmitido de humano para humano (transmissão direta), principalmente entre membros familiares, entre os quais existe maior contato próximo e por tempo prolongado.

Um estudo realizado por Van Doremalen *et al.* demonstrou que o SARS-CoV-2 pode permanecer viável e infecioso em aerossóis por até 3 h após ser eliminado no ambiente. No entanto, este tempo de sobrevida pode variar a depender do local, da quantidade, da espessura da secreção liberada pelo paciente e da superfície em que ela irá se depositar (Brito *et al*, 2020).

4.2. Diagnóstico

O diagnóstico confirmatório da COVID-19 é feito por teste molecular das secreções respiratórias. Os sintomas inespecíficos da patologia e a ausência de achados tomográficos patognomónicos tornam imperativo o uso de outros exames complementares para auxiliar no diagnóstico

diferencial. Em tempos de pandemia, a primeira hipótese torna-se quase sempre automática, porém, outros vírus como influenza, vírus sincicial respiratório e metapneumovírus não devem ser excluídos. Portanto, o diagnóstico etiológico deve ser sempre priorizado (Brito *et al*, 2020).

A pandemia do novo coronavírus (COVID-19) gerou inúmeros problemas para a população, pois além da doença e dos males causados, a população foi obrigada a se adaptar a determinadas condições, em que a conscientização de não aglomeração é um dos maiores problemas enfrentados, tendo em vista que tal necessidade limita completamente a liberdade de ação e atitudes de qualquer indivíduo (Dias, Carla, Pamplona, & Barbosa, 2020).

A pandemia da COVID-19 demonstrou a fragilidade dos sistemas de saúde ao redor do mundo após a situação emergencial na China. As autoridades sanitárias locais implementaram ações, buscando soluções para frear o avanço da COVID-19, minimizar o impacto sobre a população e fornecer precioso tempo para melhor adequação do sistema de saúde. Neste cenário, o entendimento sobre a circulação de pessoas e o sistema de transportes no estado é muito importante para direcionar a tomada de decisões (Silva, Matto, & Silva, 2020).

Com base no comportamento dos casos de COVID-19, os governos decidiram reabrir vários segmentos com algumas restrições, tais como o uso obrigatório de máscaras e número reduzido de pessoas atendidas ou encontradas em um determinado local, como supermercados, escolas, cinemas, shoppings, feiras livres e comércio em geral.

Dentro do contexto de doenças infeciosas, a modelação matemática desempenha um papel de grande destaque seja no sentido de compreender a dinâmica celular ou espalhamento de uma doença na população. Mais especificamente no caso da COVID-19 vários matemáticos, epidemiologistas e pesquisadores do mundo todo se mobilizaram para utilizar dados reais, realizar ajustamentos das curvas, bem como realizar predições sobre o andamento da pandemia nas mais diversas localidades.

4.3. Caracterização de Moçambique e Histórico sobre COVID-19

Moçambique (10°27'S e 26°52'S;30°12'E e 40°51'E) é um país na costa sudeste de África (Mapa 1), limitado pelo Oceano Índico (leste), Tanzânia e Malawi (Norte), Zâmbia e Zimbabué (Oeste), África do Sul e Essuatíni (sudoeste)13.

A população é de aproximadamente 27.909.798 habitantes, tendo como densidade 28,7 hab/km214. Moçambique tem dez províncias mais a capital e maior área urbana, Cidade de Maputo (considerada uma província), localizada na zona sul.

Em 2015, aproximadamente 46% da população vivia abaixo do limiar da pobreza (paridade de poder de compra de 1,9 dólares americanos) e 40% tinham acesso a saneamento melhorado. O acesso limitado ao saneamento está entre as causas da cólera e de outras doenças relacionadas com a higiene18, aumentando certamente o risco de COVID-19 entre a população moçambicana. De acordo com o Grupo Banco Mundial, apenas 14% da população abaixo da linha da pobreza e 42% acima dela tinham instalações sanitárias em 2015.

A 11 de Agosto de 2020 um total de 1.128.245 casos e 25.884 de mortes por COVID-19 (CFR: 2%) havia sido relatado em 55 países africanos. Isto representa cerca de 5% de todos casos relatados globalmente. Desde o último relatório, foram notificados 87.944 novos casos da COVID-19, o que representa o aumento de 7% nos novos casos relatados em relação aos que foram relatados a 11 de agosto de 2020 (União Africana, 2020).

Eis a proporção casos da COVID-19 relatados por região: região austral 56% (31.922), região norte 17% (16.323), região oriental 9% (13.045), região ocidental 13% (7.457), região central 5% (1.423).

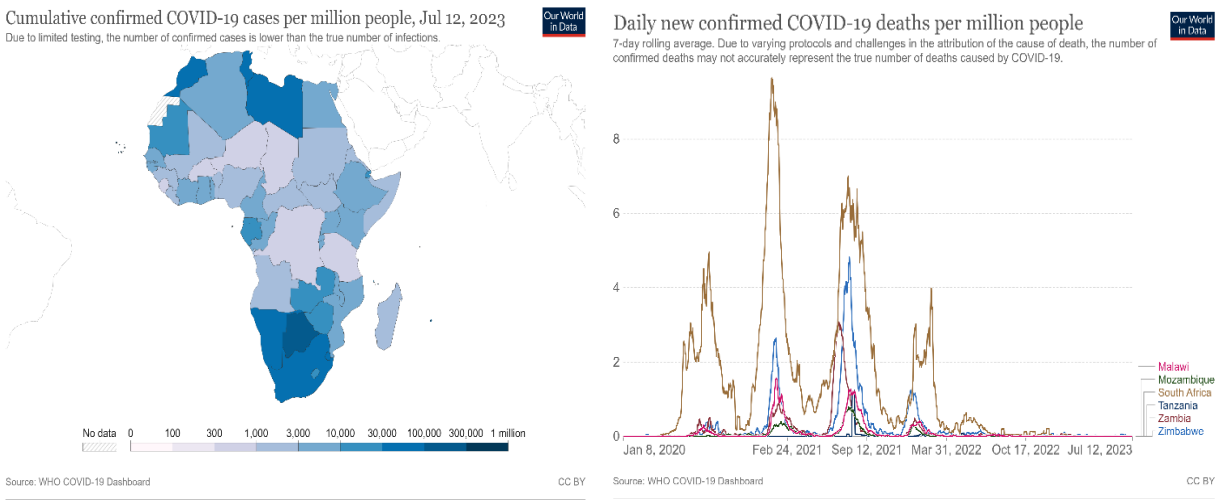


Figura 2. Distribuição dos casos confirmados na Africa.

Segundo Pereira & Forquilha (2020) citado por Zavala (2021), o primeiro caso do coronavírus oficialmente registado nas estatísticas das autoridades sanitárias de Moçambique foi diagnosticado a 22 de março de 2020, apenas alguns dias após a OMS ter declarado a pandemia da COVID-19.

Considerado importado pelo facto de o paciente ter contraído o vírus fora do País, o caso esteve envolto em certa polemica, por se tratar de um político de renome e não ter sido oficialmente comunicado.

A 30 de Março de 2020, no contexto dos esforços do Governo de Moçambique com vista a evitar a rápida propagação da doença, o Presidente da República, num discurso à Nação, declarou, pela primeira vez na história da jovem democracia moçambicana, o Estado de Emergência (EE) por razões de calamidade pública (Decreto Presidencial 11/2020, de 30 de março).

Com efeitos a partir de 1 de abril de 2020 e com uma duração de 30 dias, o documento que decretava o EE continha um conjunto de medidas que impunham limites não só à entrada e saída de pessoas como também à livre circulação de pessoas e bens dentro do território nacional.

Nessa ocasião, foram anunciadas algumas medidas de reforço para conter a disseminação da pandemia, tais como: a submissão à quarentena obrigatória de todas as pessoas que haviam viajado para estrageiro ou que mantiveram contacto com casos confirmados da COVID-19, considerando o período de transmissão ou contaminação do ou pelo vírus; a proibição de quaisquer eventos públicos ou privados; a limitação do movimento ao nível nacional e a partir de todas fronteiras de entrada no País e o encerramento das atividades comerciais ou semelhantes e, em alguns casos limitando a sua atividade (Nyusi, 2020 citado por Zavala, 2021).

A declaração do estado de emergência e imposição das medidas restritivas e de prevenção da COVID-19 em Moçambique visa responder a esse desiderato. Entretanto, o risco de aumento do número de infeção por COVID-19 parece iminente e, a inobservância das medidas emanadas parece estar entre as causas.

Devido a disposição quase anárquica das moradias, a maior densidade populacional e vias de acesso deficientes, a presença de residências que albergam um número elevado de membros, os mercados pouco estruturados, deficientes canais de circulação, foram identificados como fatores que concorrem para o aumento de risco de infeção e propagação da COVID-19, visto que não favorecem a prática do distanciamento físico (Zavala, 2021).

A estas condições aliam-se as questões normativas e comportamentais (Frederico & Matsinhe, 2021 citado por Zavala, 2021).

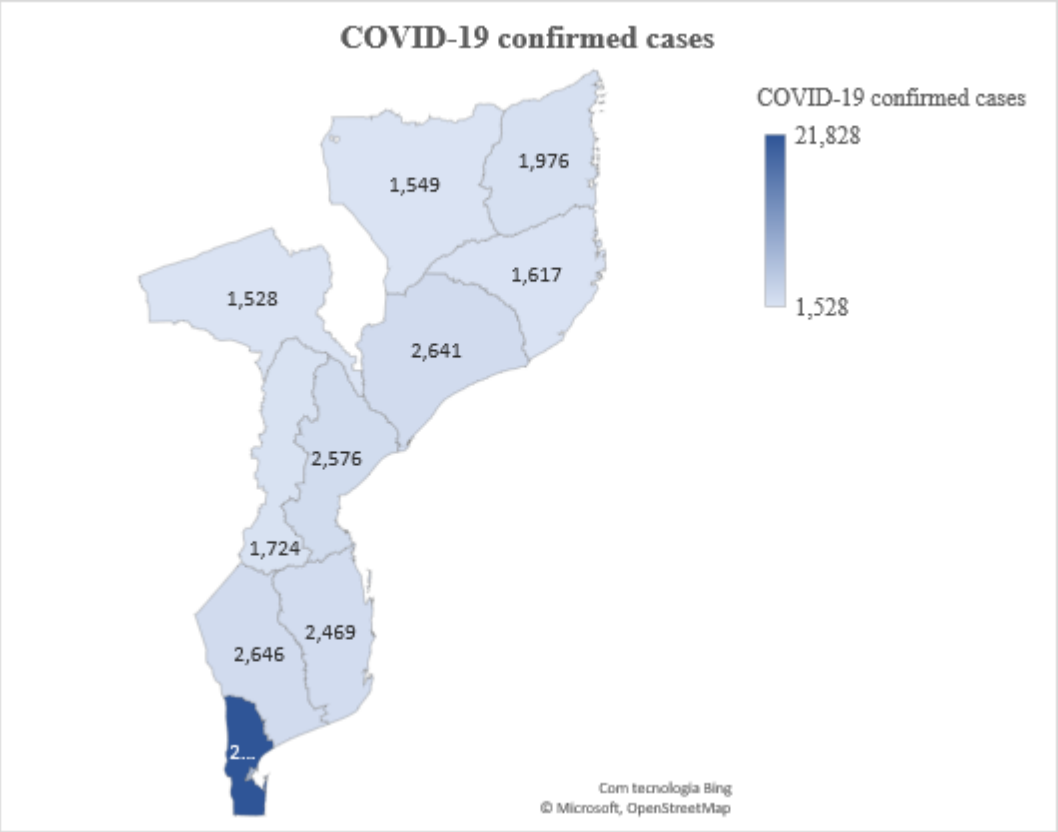


Figura 3. Distribuição dos casos confirmados em Moçambique.

5. Vigilância e modelação epidemiológica

É necessária uma monitorização contínua das informações de saúde na perspetiva de identificar o mais brevemente possível as situações de risco para a saúde do indivíduo e das populações. Portanto, faz-se necessária a implantação de sistemas de vigilância epidemiológica, que pode ser definida como o processo sistemático e contínuo de coleta, análise, interpretação e disseminação de informação com a finalidade de recomendar e adotar medidas de prevenção e controle de problemas de saúde (Braga; Werneck, 2009).

5.1. Modelação epidemiológica

Os modelos epidemiológicos são ferramentas importantes para a análise e compreensão de epidemias na população humana, que ajudam a levantar e testar estratégias de controle que auxiliam no combate a essas doenças. Os modelos matemáticos são bastante utilizados para esse

propósito e possuem um alto grau de precisão, porém são de difícil implementação, por utilizarem equações diferencias e terem alta dependência de dados (Biencourt, 2010).

Originaram-se vários modelos que, atualmente, ajudam a descrever a dinâmica de doenças, incluindo a COVID-19. Através deles é possível analisar fenômenos como crescimento, picos de prevalência, entre outras descrições do fenômeno (Theodoro, 2022). A seguir alguns modelos conhecidos serão apresentados.

5.2. Medidas estatísticas associadas a variáveis quantitativas

O resumo dos dados por meio de tabelas de frequências e gráficos de dispersão fornecem muito mais informação sobre o comportamento dos dados de uma variável do que a própria tabela original de dados.

5.2.1. Medidas de posição ou de tendência central

Mostram o valor representativo em torno do qual os dados se distribuem. São utilizadas para sintetizar, em um único número, o conjunto de dados observados. Talvez a medida mais conhecida desse tipo seja o que normalmente é conhecido como "média" ou, mais precisamente média aritmética de um conjunto de dados.

A média é considera a medida de posição mais importante. Podemos ter 4 tipos de médias:

1. Média Aritmética
2. Média Ponderada
3. Média Geométrica
4. Média Harmônica

Modelos para Resposta Quantitativa.

- Regressão linear múltipla;
- Modelo de Cox: a resposta é o tempo até a ocorrência de um evento de interesse (presença de censuras).

Modelos para Resposta Qualitativa ou Contagem.

- Modelo de Regressão Logística (binaria ou poliatómica);
- Modelo de Regressão de Poisson: contagem.
- Outros: modelos beta, gama, etc.
- Métodos estatísticos para comparar as proporções.

5.3. Modelos estatísticos - Métodos Paramétricos e não paramétricos

Descrição	**Métodos Paramétricos**	**Métodos Não Paramétricos**
Estatísticas descritivas	Média, desvio padrão	Mediana, intervalo interquartil
Amostra com população (ou valor hipotético)	Teste *t* de uma amostra (n <30) e teste Z de uma amostra (n ≥30)	Uma amostra do teste de classificação assinada de Wilcoxon
Dois grupos não pareados	Teste *t* de amostras independentes (teste t de amostras não pareadas)	Teste U de Mann Whitney/teste de soma de classificação de Wilcoxon
Dois grupos emparelhados	Teste t de amostras pareadas	Amostras relacionadas - Teste de postos sinalizados de Wilcoxon

Tabela 2. Métodos Paramétricos e não paramétricos.

6. Processamento de dados usando R

Modelos matemáticos são importantes para previsão da possibilidade e severidade de ocorrência e fornece informações para determinar o tipo e intensidade de intervenção da doença. (Wangping, *et al*., 2020).

A atividade de inteligência está voltada para obtenção e análise sistemática de dados, informações e de produção e difusão de conhecimentos, relativos a fatos e situações de imediata ou potencial influência sobre o processo decisório, a ação governamental.

Na análise quantitativa o uso de programas informatizados facilita o tratamento e a análise de dados, mas não se deve superestimar seu alcance e aplicações. A respetiva utilidade é maior quando são estabelecidas redes de colaboração e sistemas de informação em saúde, que permitem o tratamento eficiente de grandes bases de dados e geram informação oportuna e útil para a tomada de decisões.

Um programa informatizado reduz notavelmente o tempo de cálculo, processamento e análise dos dados, mas é o trabalho humano o que aporta resultados racionais e válidos para o desenvolvimento dos objetivos de saúde pública (Organizacao Pan Americana, 2010).

A utilidade da modelação é permitir a construção diária de uma previsão confiável para a evolução das doenças, bem como, acompanhar o desenvolvimento dos fenômenos; armazenando o estudo histórico das ocorrências, em relação às ações de contenção utilizadas naquele dado momento (Proença & Schmidt, 2021).

Ao debruçar sobre a programação R, inicialmente importa salientar particularidades da linguagem e afirmar que existem muitos conceitos de linguagem apresentados em diferentes fontes. Buscou-se aqui fazer um resumo e trazer conceitos sobre a linguagem no geral e a computação em particular. Da leitura feita em diversas obras pode-se entender que a linguagem computacional desempenha um papel de suma importância na entrada de dados, sua visualização, manipulação e o processamento. Portanto, para a implementação de um algoritmo em um computador, é necessário descrevê-lo sob maneira que a sua linguagem seja reconhecida pelo computador caso queira executá-lo. A linguagem de programação faz descrição do algoritmo/códigos introduzidos para que seja entendido pelo computador.

Ao longo da história de computação foram desenvolvidas diversas linguagens de programação, sendo que cada linguagem foi desenvolvida em seu teu tempo, com objetivo de introdução de facilidades e tarefas tornaram ao longo da história a tarefa do programador mais fácil e pouco suscetível a erros. Antigamente programar era considerada uma arte restrita a apenas alguns grupos de indivíduos para tornar uma ferramenta computacional de uso comum, mas, hoje em dia, com as linguagens visuais, o programador deixou de fazer parte de um grupo restrito de pessoas pela facilidade no manuseio e a clareza na manipulação dos algoritmos e comando.

A linguagem de programação é classificada quanto ao nível, a geração e quanto ao paradigma. Quanto ao nível a linguagem de programação pode ser:

✓ baixo nível - linguagem de baixo nível são as linguagens escritas usando microprocessadores, geralmente estes tipos de linguagem de programação é designada de linguagens Assembly. Estes programas apresentam maior velocidade na execução e ocupam poucos espaços de memoria. Por outro lado, o código gerado para um tipo de processador não tem utilidade noutros tipos de processado, e, estas características constituem desvantagens deste tipo de linguagem.

✓ medio nível - esta linguagem constitui a juncão entre o nível baixo e nível alto de programação, estes estão mais voltados aos seres humanos e as máquinas. Os seus comandos são explicito e implícito em parte na execução.

✓ alto nível

Quando à geração a linguagem pode ser de:

- 1ª geração,
- 2ª geração,
- 3ª geração,
- 4ª geração e
- 5ª geração.

Indo ao conceito do R, Sfair (2015) conceitua R como uma linguagem e ambiente de desenvolvimento integrado para cálculos estatísticos e produção de gráficos. De acordo com Lauretto (2015), R é uma linguagem e um ambiente de desenvolvimento voltado principalmente para computação estatística inferência, simulações, data mining, etc) e gráficos.

O desenvolvimento da ferramenta R foi inspirado em duas linguagens

- S (John Chambers e colegas Bell Labs): sintaxe
- Scheme (Hal Abelson and Gerald Sussman): implementação e semântica

Foi desenvolvido originalmente por Ross Ihaka e Robert Gentleman (Departamento Estatística da Universidade de Auckland, Nova Zelândia, e atualmente desenvolvido pelo R Development Core

Team. O R está disponível como um software livre, nos termos da GNU GPL (General Public License).

- Windows, Linux, OS X (Mac).

***RStudio*: Ambiente de desenvolvimento**

RStudio é um ambiente de desenvolvimento integrado (IDE Integrated Development Environment) para R (Lauretto, 2015).

- Envolve a Integração de:
- Editor de programas
- Facilidade de execução parcial ou total de scripts R
- Visualização de dados (Tabelas, Gráficos
- Documentação (Ferramentas de depuração de programas

Disponível para Windows, Linux, OS X (Mac)

De acordo com Landeiro (2013), o R é um programa leve (ocupa pouco espaço e memória) e geralmente roda rápido, até em computadores não muito bons. Isso porque ao instalarmos o R apenas as configurações mínimas para seu funcionamento básico são instaladas (pacotes que vem na instalação "base"). Para realizar tarefas mais complicadas pode ser necessário instalar pacotes adicionais (packages).

Website oficial

–https:// www.rstudio.com

Ambiente do RStudio

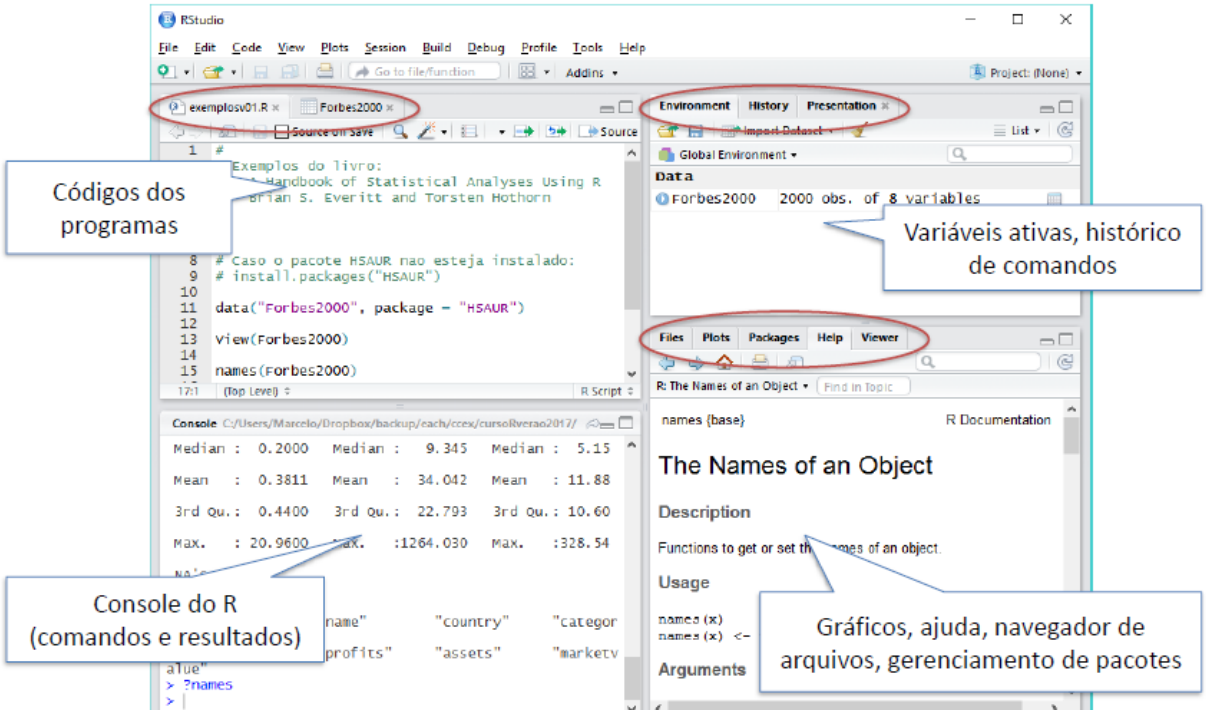


Figura 4: Ambiente do RStudio. **Fonte:** *(Lauretto, 2015)*.

6.1. Tipos de dados no R

Tipos de objetos mais usuais em R:

- vector: o mais elementar e um dos mais importantes
- matrix e array: generalizações multidimensionais de vetores
- data frame: conjunto de dados retangular no qual:
  - linhas representam os casos (sujeitos do estudo)
  - colunas representam as variáveis descritivas dos casos.

O *dataframe* é o formato ideal para o armazenamento dos dados do R. De acordo com Ribeiro Jr. (2005), para entrar com os dados no R usa-se o editor que vem no programa. E, para digitar rapidamente os dados a serem introduzidos no programa o mais fácil é usar códigos para as variáveis categóricas.

Segundo Torgo (2006), o R tem vários objetos onde podemos armazenar diversos tipos de dados, mas, os *dataframe* revelam-se como os mais adequados para armazenar tabelas de dados de

problema qualquer. O R tem disponível diversas tabelas com dados que podem ser usados no treinamento.

Cada objeto em R (e cada coluna em um data frame) possui um dos seguintes tipos básicos:

- numeric: para variáveis numéricas (reais, complexas)
- fator: representação de variáveis categóricas nominais ou ordinais
- logical: TRUE/FALSE
- character: texto string

De acordo com Silva, Barttoluzi e Dinis (2009), cada objeto possui atributos que podem ser, tamanho (lenght) ou tipo (mode). As informações contidas nestes atributos são bastante importantes na manipulação dos dados.

6.2. Tipos de gráficos

As componentes de estrema importância e versátil do ambiente R são as capacidades gráficas. Estas capacidades permitem que o R consiga plotar desde gráficos bidimensionais simples até gráficos tridimensionais mais complexos por meio de comandos simples. Os gráficos estatísticos tais como histogramas, curvas de distribuições, gráficos de dispersão, assim como os gráficos de barra são dados mais enfase no R (Bortolluzi, Silva e Dinis, 2009).

No programa R existem três tipos de funções gráfica:

- funções de alto nível (high-level), que produzem gráficos completos;
- funções de baixo nível (low-level), que adicionam informações a um gráfico existente e,
- funções para trabalhar interactivamente com gráficos.

São vários comandos a serem executados na produção de gráficos, dentre eles o Plot, Boxplot, Barplot e Pie. Para manipular os comandos, na programação R existem pacotes que geram os gráficos, estes pacotes são designados de *packpages*. Para além destes pacotes, existem outros diversos que possuem funções similares em diferentes especificações.

A linguagem de programação R se apresenta como uma alternativa robusta e eficiente para a ciência de dados em geral. Por ser uma linguagem desenvolvida por estatísticos, muitas funções e fórmulas que nas linguagens mais difundidas precisam ser desenvolvidas do zero, no R já vêm

prontas para aplicação com breves comandos. Esta característica, além de permitir códigos mais simples, também contribui para minimização de erros (Fagundes, 2021).

O R é gratuito e de código aberto, o que gera potencial vantagem competitiva em relação a ferramentas como SAS e SPSS. Dentre as principais funções da linguagem está extensa relação de modelos estatísticos, que vão desde a modelação linear e não-linear, a análise de séries temporais, os testes estatísticos clássicos, análise de agrupamento e classificação, etc, além da apresentação gráfica dos resultados contando com variadas técnicas, passando também pela criação e manipulação de mapas.

R é um software estatístico gratuito de código aberto, excelente para gráficos, modelação estatística clássica e vários métodos não paramétricos, bem como para muitos modelos multiníveis.

Para além dos modelos específicos que podem ser enquadrados pelos pacotes R existentes, este programa é totalmente programável e pode assim adaptar-se a qualquer modelo, desde que seja feita programação suficiente. R contém vários pacotes que estimam modelos multiníveis: o nlme e o lme4; lme4 é particularmente valioso para lidar com resultados não normalmente distribuídos e estruturas de dados parcialmente cruzadas (Hritcu, 2015).

Desta forma, a análise estatística da linguagem ‘R’ se apresenta como uma boa alternativa para análise de dados de forma simples e dinâmica, uma vez que dispõe de uma enorme quantidade de pacotes desenvolvidos com as principais manipulações mais comumente utilizadas na análise de dados, o que facilita tanto o aprendizado quanto a utilização da linguagem.

Contudo, todos dados usados no R encontram-se armazenado na memoria do computador no formato de uma objeto, estes objetos por sua vez apresentam um nome pelo qual estão associados e podem armazenar diferentes tipos, podendo ser elas os números, textos, vectores, matrizes, expressões, chamadas, funções e gráficos. O que faculta o armazenamento dos objetos na memoria são usados o operador de atribuição.

Os comandos utilizados no R dependem da finalidade de execução dos dados pelo que, o R é aplicado em diversos tipos de estudos, e cada caso é um caso, mas de uma forma organizada, irei aqui apresentar alguns comandos que acho ser de suma importância na manipulação de dados.

Contudo para análise dos dados e processamento da informação no R, são necessárias instalações de pacotes específicos.

III. Metodologia

Nesta secção são apresentados os principais procedimentos metodológicos para o desenvolvimento do presente trabalho, que trata das metodologias estatísticas de análise de dados agrupados em estudos epidemiológicos. Aborda-se as etapas de análise de dados, incluindo os conceitos básicos necessários para a sua compreensão.

O tipo de pesquisa adotado, segundo Gil (1991), é bibliográfico documental e de levantamento, sendo um estudo exploratório, conforme os objetivos estabelecidos.

Foi realizada uma revisão bibliográfica detalhada com o objetivo de investigar diversas abordagens estatísticas para análises de dados, abrangendo tanto a literatura teórica quanto os métodos empiricamente comprovados. Consultamos estudos relevantes sobre metodologias estatísticas de análise de dados agrupados em estudos epidemiológicos, utilizando dados da COVID-19 de Moçambique para aplicação prática das respetivas metodologias na ferramenta R.

As metodologias estatísticas amplamente utilizadas para analisar dados agrupados incluem Análise de Variância (ANOVA), análise de regressão logística, Séries Temporais, Modelos Lineares Generalizados (GLMs) e modelos de Sobrevivência. Previamente realizamos uma análise exploratória para examinar os dados provenientes da aplicação de técnicas estatísticas, facilitando a compreensão básica dos dados e das relações entre as variáveis.

O estudo seguiu uma abordagem analítica observacional transversal, baseada no registro do número de casos de COVID-19 em diferentes países, com foco em Moçambique. Após a revisão teórica, adotamos a abordagem de implementação prática dessas metodologias, utilizando o software R para análise exploratória e descritiva dos dados coletados.

Os dados foram obtidos da base Our Wold in Data (OWID) da OMS (link) e outras fontes de pesquisa, incluindo o boletim informativo diário do Ministério da Saúde de Moçambique. Realizamos adaptações dos modelos existentes para o contexto específico de Moçambique, conduzindo análises e testes de correlação, análises de séries temporais e regressão linear.

Our World in Data (OWID), é uma plataforma online que fornece acesso gratuito a dados e pesquisas sobre os principais problemas globais, incluindo a pandemia da COVID-19. Esta base de dados oferece uma fonte rica de informações sobre os casos confirmados, óbitos, medidas de

intervenção e outras variáveis relevantes para o estudo da pandemia agrupados por dias e diferentes regiões, como Moçambique, o que permite uma análise temporal e aeroespacial da propagação da doença no País. Usou-se essa base de dados para realizar análise estatísticas exploratórias, correlacionais, de regressão e modelagem de dados agrupados, explorando a relação entre as variáveis disponíveis e os padrões de propagação e o impacto da COVID-19 em Moçambique. A análise se baseou nos dados disponíveis no período de 1 de março de 2020 até 24 de maio de 2023, totalizando 1238 dias de registros. Os dados explorados estão estruturados em diversas variáveis, agrupadas de acordo com diferentes aspetos relacionados à pandemia. Essas variáveis incluem informações sobre casos confirmados, novos casos, óbitos entre outros.

Exploramos modelos estatísticos adequados para análise de dados agrupados, como modelos de regressão e modelos de séries temporais, e aplicamos técnicas de análise exploratória para identificar padrões e tendências nos dados agrupados, incluindo a visualização de dados por meio de gráficos e mapas.

A escolha do software R para o processamento dos dados foi motivada por sua versatilidade, aplicabilidade em diversas áreas de pesquisa e recursos robustos para análise estatística e visualização de dados. Sua ampla utilização na comunidade acadêmica estatística e a disponibilidade de pacotes e funções para análise, visualização e modelagem estatística foram fatores determinantes.

IV. RESULTADOS E DISCUSSÃO

4. Metodologias estatísticas para análise de dados agrupados

Existem várias metodologias estatísticas para análise de dados agrupados, dependendo do tipo de dados e das perguntas da pesquisa, e cada uma com suas próprias aplicações e vantagens, e, um passo muito importante na análise dos dados é a seleção do método estatístico apropriado.

Para cada situação específica estão disponíveis métodos estatísticos para análise e interpretação dos dados. Para selecionar os métodos estatísticos apropriados, é necessário conhecer os pressupostos e as condições dos métodos estatísticos, para que o método estatístico adequado seja selecionado para análise de dados.

A seleção do método estatístico apropriado depende de três fatores:

- Finalidade e objetivo do estudo
- Tipo e distribuição dos dados
- Natureza das observações (emparelhada ou não emparelhadas).

Assim, um dos primeiros passos na análise de dados é a compreensão dos dados em si. Os dados agrupados geralmente envolvem categorização ou agregação de informações em grupos distintos. Isto, incorre por vários motivos, como a conveniência na apresentação de dados, a simplificação de uma grande quantidade de informação ou a necessidade de comparar diferentes categorias.

Uma seleção errada do método estatístico para além de criar problemas sérios durante a interpretação dos resultados, afeta a conclusão do estudo. A aplicação de metodologias estatísticas não apropriadas podem ser vistos em muitas condições, como o uso de testes t não pareados em dados pareados ou uso de testes paramétricos que não seguem a distribuição normal, entre outros erros.

Os diversos métodos estatísticos usados na análise de dados agrupados em estudos epidemiológicos são dependentes do estudo que se pretende desenvolver. São várias as metodologias estatísticas para análise e estes podem ser encontrados descritos na tabela a seguir:

Metodologia	Autor e ano	Tipos de dados	Aplicação
Análise de regressão Logística	David W. Hosmer Jr. e Stanley Lemeshow - Ano: 1989	Variável dependente binaria	Investigação de fatores de risco, previsão de eventos binários;
Series temporais	George E.P. Box e Gwilym M. Jenkins - Ano: 1976	Observações sequencias ao longo do tempo	Modelagem de tendências temporais, previsão de séries temporais;
Modelos Lineares Generalizados	Peter J. Diggle, Patrick Heagerty, Kung-Yee Liang e Scott L. Zeger - Ano: 2002	Diversos, incluindo variáveis continuas	Modelagem de diferentes tipos de variáveis de resposta, adaptação a distribuições específicas
Modelos de Sobrevivência	David R. Cox - Ano: 1972	Tempo até um evento de interesse	Avaliação da sobrevida, análise de tempo até o evento, modelagem de riscos
Modelos de Mistura Multinominal	Michael D. Escobar e Mike West - Ano: 1994	Variáveis categóricas com múltiplas categorias	Segmentação de populações em grupos distintos, identificação de padrões latentes
Analise de sobrevivência Multinível	Ronald H. Heck, Scott L. Thomas e Lynn N. Tabata Ano: 2010	Dados de sobrevivência organizados em níveis hierárquicos	Modelagem de variação entre grupos, consideração de efeitos de cluster
Modelo Hierárquico Bayesiano	Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari e Donald B. Rubin - Ano: 2013	Dados organizados em múltiplos níveis hierárquicos	Incorporação de incertezas, modelagem de estruturas complexas, análise de efeitos aleatórios
Modelos de Series Espaciais Temporais Espaciais	Noel Cressie e Christopher K. Wikle - Ano: 2011	Observações sequenciais ao longo do tempo e espaço	Modelagem de padrões temporais e espaciais, previsão de séries espaciais
Modelos de rede	Mark S. Handcock, Adrian E. Raftery e Jeremy M. Tantrum Ano: 2007	Dados de integrações entre unidades	Modelagem de estruturas de rede, análise de influência e propagação de eventos
Modelos de equações estruturais	Karl G. Jöreskog e Dag Sörbom - Ano: 1977	Dados com estrutura de relação entre as variáveis	Modelagem de relações complexas entre variáveis, análise de causalidade

Tabela 3. metodologias estatísticas para análise agrupados. Fonte: Autor, 2024.

Dentre as metodologias estatísticas acima descritas como mais utilizadas em estudos epidemiológicos, destacam-se a ANOVA, a análise de regressão logística, as Séries temporais, os Modelos Lineares Generalizados (GLMs) e os modelos de Sobrevivência. Para fundamentar a prevalência das metodologias estatísticas anteriormente destacadas, foi conduzida uma revisão abrangente da literatura em epidemiologia.

Durante essa análise, identifiquei tendências consistentes no uso de certas abordagens estatísticas em uma variedade de estudos. São apresentadas a seguir alguns estudos em que estas metodologias foram aplicadas:

Metodologia	**Autores**	**Âmbito da Aplicação**
Regressão Logística	Hosmer Jr, D. W., Lemeshow, S., Altman, D. G., Hosmer Jr, D. W., Lemeshow, S., Zou, G.	Análise de dados binários, como estudos de prevalência e fatores de risco.
ANOVA	Altman, D. G., Kleinbaum, D. G., Kupper, L. L., Morgenstern, H.	Estudos experimentais e observacionais, análise de variáveis contínuas e categóricas.
Regressão Linear	Altman, D. G., Clayton, D., & Hills, M., Durrleman, S.	Modelagem de dados utilizando variáveis contínuas.
Modelos Lineares Generalizados	Diggle, P. J., Heagerty, P., Liang, K. Y., & Zeger, S. L., McCullagh, P., & Nelder, J. A., Clayton, D., & Hills, M., Hosmer Jr, D. W., Lemeshow, S.	Análise de dados longitudinais, como estudos de coorte e painel, e modelagem de diferentes tipos de variáveis de resposta.
Modelos de Sobrevivência	Cox, D. R., Collett, D., Prentice, R. L., & Gloeckler, L. A., Hosmer Jr, D. W., Lemeshow, S., Rosner, B.	Análise de dados de tempo até o evento, como estudos de sobrevivência e estudos longitudinais.

Tabela 4. Autor, 2024.

Ao examinar esses estudos, ficou evidente que as metodologias mencionadas são utilizadas de forma consistente para abordar uma ampla gama de questões de pesquisa em epidemiologia. Importa ressaltar que estas observações são respaldadas por uma análise de uma amostra representativa de estudos de referência na área.

Assim, a análise de regressão logística é especialmente relevante quando a variável de interesse é binária e as observações estão agrupadas em unidades como regiões geográficas ou grupos demográficos. A regressão logística permite entender como as variáveis independentes afetam as chances de um resultado específico ocorrer. Enquanto as Séries temporais se destacam na análise de dados recolhidos ao longo do tempo, permitindo a investigação de tendências, sazonalidades e padrões de longo prazo na propagação de doenças.

Os estatísticos usam os modelos de regressão em problemas onde o objetivo é estudar a relação entre as variáveis (Alvarenga, 2015). O modelo linear é o mais utilizado para modelar a relação entre as variáveis estudadas, sendo que, este modelo assume, entre outras, que o valor esperado da variável resposta é uma combinação linear das variáveis explicativas e que a variável resposta segue a distribuição normal.

Como uma das metodologias estatísticas usadas na análise de dados agrupados, os Modelos lineares generalizados são uma classe poderosa e flexível de modelos estatísticos que ampliam a aplicabilidade dos modelos estatísticos lineares tradicionais, ou seja, conjunto de técnicas de análises estatística que se baseiam na suposição de que existe uma relação linear entre as variáveis independentes (preditoras) e uma variável dependente (resposta). Os Modelos lineares generalizados foram desenhados para lidar com uma variedade mais ampla de tipos de dados e distribuições, tornando-os uma ferramenta essencial em análises estatísticas avançadas.

Ao abordar em específico sobre as principais metodologias estatísticas de análise dados agrupados, deve-se realçar que estes são divididos e podem ser aplicados a dados simples e dados complexos.

Os comummente utilizados têm sido aplicados a dados que se apresentam de forma simples de análise, como a aplicação de testes de correlação, e, à medida que as complexidades dos dados agrupados aumentam, métodos mais avançados são empregues para análise de dados, como a regressão logística entre outras metodologias.

Os GLMs oferecem flexibilidade na modelagem de uma variedade de variáveis de respostas, enquanto que os Modelos de Sobrevivências são aplicáveis quando o tempo até um evento de interesse é o foco de análise.

De acordo com Alvarenga (2015), o modelo linear generalizado foi introduzido em 1972 por Nelder e Wedderburn como resposta a resolução do problema do modelo linear, pois o modelo linear é limitado, e não pode ser utilizado se a distribuição da variável resposta é diferente da normal. Este modelo foi um grande impulso na área de modelação estatística.

Os Modelos lineares generalizados apresentam três componentes principais:

- ***Função de ligação*** – Descreve a relação entre o valor esperado da variável resposta e uma combinação linear das variáveis independentes. Conecta parte do modelo linear com a distribuição da variável resposta. São exemplos desta função: a função *logit*, a função de identidade e a função *log*.
- ***Modelo linear*** – as variáveis independentes são combinadas linearmente com os coeficientes.
- ***Distribuição de probabilidade*** – Os Modelos lineares generalizados permitem que a variável resposta siga diferentes distribuições de probabilidade, como a distribuição normal, binomial, Poisson, gama, entre outras distribuições. Assim, a escolha de distribuição depende da natureza dos dados e os objetivos de análise.

Componente aleatória – é representada por um conjunto de variáveis aleatórias independentes $y_1 \ldots . y_n$, obtidas de uma mesma distribuição que faz parte da família de distribuição exponencial com médias $\mu_1 \ldots . . \mu_n . E(Y_i) = \mu_i, i = 1, \ldots . n$

Componente sistemático – A variáveis explanatórias entram na forma linear de seus efeitos

$$n_i = \sum_{r=1}^{p} x_{ri} . \beta_r = x_i^T . \beta \; ou \; \eta = X . \beta$$

Onde, X é a matriz do modelo, β é o vetor dos parâmetros desconhecidos e η é o preditor linear.

Função ligação – é uma função que relaciona o componente aleatório ao componente sistemático, vinculando a média ao preditor linear.

$$\eta_i = g(\mu_i)$$

Onde g é uma função monótona e diferenciável, podendo ser:

Identidade: $\eta = \mu$

Potencia: $\eta = \mu^{\lambda}$, em que λ é um número real qualquer.

Logit: $\eta = \log(\frac{\mu}{1-\mu})$

Probit: $\eta = \Phi^{-1}(\mu)$

Complemento Log-Log: $\eta = \log[-\log(1-\mu)]$

Logaritmo: $\eta = \log(\mu)$

Segundo Roquim (2014), dentre as funções acima apresentadas, a função ligação mais simples é a função identidade, utilizada no modelo linear.

São denominadas de funções de ligações canónicas quando as funções de ligação fornecem estatísticas suficientes. As funções canónicas dos modelos Normal, de Poisson, Binomial, Gamma e funções inversas são respetivamente:

$$\eta = \mu\,, \qquad \eta = \log(\mu)\,, \qquad \eta = \log(\frac{\mu}{1-\mu})\,, \quad \eta = \mu^{-1}\,, \qquad \eta = \mu^{-2}$$

Segundo Monfardini (2016), a vantagem do uso destes modelos é que a variável resposta pode assumir qualquer distribuição nesta família e a relação entre a variável resposta e as covariáveis, que podem contribuir ou não para explicar sua variabilidade, dá-se através de uma função ligação.

Outros autores afirmam que o modelo é generalizado pelo facto da variável resposta apresentar um leque maior de distribuição e não só a normal como nos modelos clássicos ou modelo linear.

Segundo Nascimento de Jesus (2015), uma das condições necessárias para que seja possível a utilização de Modelos Lineares Generalizados é que a variável resposta dependente pertença à família exponencial de distribuições. Sendo assim, antes da abordagem aprofunda dos Modelos Lineares Generalizados faz-se necessário falar desta família.

De acordo com Turkman *et al*. (2000), os Modelos Lineares Generalizados (MLG) são limitados por manterem a estrutura de linearidade, isto porq as distribuições se restringirem à família exponencial e por exigirem a independência das respostas. Posto isto, são casos particulares dos Modelos Lineares Generalizados, entre outros:

- Modelos de Análise de Variância e Covariância;
- Modelo de Regressão de Poisson;
- Modelo de Regressão Logística;
- Modelo de Regressão Linear Clássico;

Família exponencial de Distribuições

Uma distribuição é dita ser da família exponencial se a sua densidade pode ser escrita na seguinte forma:

$f(y \setminus \theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$ Equação 1. Família exponencial de Distribuições

A família exponencial configura-se como de suma importância nos modelos lineares generalizados, pelo facto de neste ser possível a utilização de dados discretos ou contínuos, dados que possuem assimetrias e dados que são restritos a um intervalo do conjunto dos reais.

Existem várias distribuições que pertencem a esta família exponencial, dentre elas:

Distribuição Normal (geralmente utilizadas para dados contínuos simétricos)

$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{\frac{-(y-\mu)^2}{2\sigma^2}\right\} = exp\left\{\frac{y*\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{1}{2}ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right\}$ Equação 2

Em que,

$$\theta = \mu\,;\;\; b(\theta) = \frac{\mu^2}{2} \;\;;\;\; a(\phi) = \sigma^2 \;\; ; c(y, \phi) = \frac{1}{2}\ln(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}$$

Neste caso, a média e a variância são:

$$E(y) = b'(\theta) = \mu$$

$$Var(y) = b''(\theta)a(\phi) = \sigma^2$$

Distribuição Binomial (Geralmente utilizadas para dados de proporções)

$$f(y;\mu) = \binom{n}{y} p^y (1-p)^{n-y} = exp\left\{y \ln\left(\frac{p}{1-p}\right) + n + \ln(1-p) + ln\binom{n}{y}\right\}$$

Sendo:

$$\theta = \ln\left(\frac{p}{1-p}\right) = \left[p = \frac{\mu}{n}\right] = \ln\left(\frac{\mu}{n-p}\right); b(\theta) = -n\ln(1-p) = n\ln\left(1+e^{\theta}\right) : a(\phi)$$
$$= 1\,; c(y,\phi) = ln\binom{n}{y}$$

A média e a variância podem ser obtidas

$$E(y) = b'(\theta) = \frac{ne^{\theta}}{1+e^{\theta}} = \left[e^{\theta} = (\frac{p}{1-p})\right] = np$$

$$Var(y) = b''(\theta)a(\phi) = \frac{ne^{\theta}}{(1+e^{\theta})^2} = \left[e^{\theta} = \left(\frac{p}{1-p}\right)\right] = np(1-p)$$

Distribuição Poisson

É geralmente utilizada para dados de contagem

$$f(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = \ exp\{y\ln(\lambda) - \lambda - \ln(y!)\}$$

Sendo:

$$\theta = \ln(\lambda) \rightarrow \ \lambda = e^{\theta}; \quad b(\theta) = \lambda = e^{\theta} \ ; a(\phi) = 1; c(y,\phi) = -\ln(y!)$$

Podendo ser a média e a variância.

$$E(y) = \ b'(\theta) = e^{\theta} = \lambda$$

$$Var(y) = b''(\theta)a(\phi) = e^{\theta}1 = \lambda$$

Distribuição Gama

Geralmente utilizadas para dados contínuos assimétricos.

$$f(y;\mu,\alpha) = \frac{1}{\Gamma(\alpha)}\left(\frac{\alpha}{\mu}\right) y^{\alpha-1} e^{-\alpha(\frac{y}{\mu})}$$

$$= exp\left[\frac{y\frac{-1}{\mu} + \ln(\frac{1}{\mu})}{\frac{1}{\alpha}} + \alpha\ln(\alpha y) - \ln(y) - ln[\Gamma(\alpha)]\right]$$

Em que,

$$\theta = -\frac{1}{\mu}; b(\theta) = -l\left(\frac{1}{\mu}\right) = -\ln(-\theta)\,; a(\phi) = \frac{1}{\alpha}$$

$$c(y,\phi) = \alpha\ln(\alpha y) - \ln(y) - ln[\Gamma(\alpha)]$$

Cuja média e variância são:

$$E(y) = b'(\theta) = -\frac{1}{\theta}$$

$$Var(y) = b''(\theta)a(\phi) = \mu^2\frac{1}{\alpha} = \frac{\mu^2}{\alpha}$$

Distribuição Normal inversa

Esta distribuição é geralmente utilizados para dados continuos assimetricos

$$f(y;\mu,\sigma^2) = -\frac{1}{\sqrt{2\pi\sigma^2 y^3}} exp\left\{\frac{-(y-\mu)^2}{2\sigma^2\mu^2 y}\right\} = exp\left\{\frac{y\frac{-1}{2\mu^2} + \frac{1}{\mu}}{\sigma^2} - \frac{1}{2}\left[\ln(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 + y}\right]\right\}$$

Sendo:

$$\theta = -\frac{1}{2\mu^2};\;\; b(\theta) = \frac{1}{\mu} = -(-2\theta)^{\frac{1}{2}};\quad a(\phi) = \sigma^2;\quad c(y,\phi) = -\frac{1}{2} + \left[\ln(2\pi\sigma^2 y^3) + \frac{1}{\sigma^2 + y}\right]$$

Sendo a média e a variância

$$E(y) = b'(\theta) = (-2\theta)^2 \qquad\qquad Var(y) = b''(\theta)a(\phi) = \mu^3\sigma^2$$

Estimação de parâmetros

De acordo com Cordeiro e Dantas (2000), a estimação do vetor parâmetro desconhecido de um Modelo Linear Generalizado, $\beta = (\beta_1, \dots\dots, \beta_k)'$, geralmente é feita utilizando-se o método de máxima verosimilhança, que consiste em resolver o sistema $\mathrm{U}(\hat{\beta}) = 0$, onde $\mathrm{U}(\hat{\beta})$ é conhecido como função escore ou função suporte, dada por:

$$U(\hat{\beta}) = \frac{\partial L(\beta)}{\partial \beta} = \left(\frac{\partial L(\beta)}{\partial \beta_1}, \dots, \frac{\partial L(\beta)}{\partial \beta_k}\right)'$$

Sendo: $\frac{\partial L(\beta)}{\partial \beta_r} = \sum_{l=1}^{n} \phi_l [Y_l - \mu_l] W_l \frac{d\eta_l}{d\mu_l} X_{lr}$ $\quad r=1,\dots,k$

Onde $W = \frac{(\frac{d\mu_l}{d\eta_l})^2}{V}$ é denominado função peso.

- **A função de máxima verosimilhança** – A solução das equações de máxima verosimilhança é equivalente a uma iteração de métodos de mínimos quadrados ponderados com uma função de peso (Nascimento de Jesus, 2015). Esta é uma abordagem estatística usada para estimar parâmetros de um modelo probabilísticos.

Quando aplicada a um modelo linear, a solução da equação de máxima verosimilhança pode ser vista como uma interação de métodos de mínimos quadrados ponderados. Ora vejamos as componentes da equação:

$$W = \frac{(\frac{d\mu}{dY})^2}{V}$$

Esta é a função de peso usada nos mínimos quadrados, ponderados. Ela pondera a contribuição de cada observação para a estimação dos parâmetros do modelo. Aqui, $\frac{d\mu}{dY}$, representa a derivada da média do modelo em relação a variável dependente Y, e V é uma medida de variância.

A variável dependente modificada: $y = Y + (Z - \mu)/(\frac{d\mu}{dY})$.

Esta é uma forma modificada da variável dependente Y. Ela incorpora uma correção baseada na diferença entre a média do modelo (μ) e uma variável Z, dividida pela derivada da média do modelo em relação a variável dependente Y.

Isso sugere que, ao modificar a variável dependente Y, estamos ajustando os dados para que sejam mais adequados ao modelo, levando em consideração as propriedades do modelo e das observações. Tem bases nas estimativas atuais, ou seja, a expressão (Z-*μ)/(dμ/dy)* traduz uma medida de quão longe o valor de *Z* está da media μ, em termos de mudanças na variável dependente *Y*, com base nas estimativas atuais.

Assim, ao adicionar essa correção a variável dependente *Y*, estamos ajustando-a com base nas relações e nas condições atuais, garantindo que a modificação da variável dependente Y leve em consideração as estimativas atuais das variáveis envolvidas.

- **Estatística Suficiente** – Ocorre quando θ, o parâmetro de distribuição do componente aleatório, e Y o valor previsto do modelo linear coincidirem. Sendo:

$$L = ZY - g(Y) + h(z) \;\; e \;\; \frac{\partial L}{\partial B_i} = \alpha(\phi)(Z - \mu)x_i \;,$$

tendo desta forma as equações de máxima verosimilhança $\sum_k (z - \hat{\mu})x_{ik} = 0$.

Estando o somatório sob observação teremos:

$$\sum_k z_k x_{ik} = \sum_k \hat{\mu}_k x_{ik}$$

ANOVA

A ANOVA foi desenvolvida por Fisher, no início do século XX, na Estação Experimental Agrícola de Rothmstead, em Inglaterra. Tal como referido anteriormente, no capítulo da introdução, o principal intuito de proceder ao estudo de variância é fazer a comparação de mais de dois grupos em relação à sua localização.

Este estudo permite igualmente resumir o modelo de regressão linear recorrendo à decomposição de soma dos quadrados: soma de quadrados total *(SQT)* que traduz a variação da variável resposta; soma dos quadrados explicada *(SQR)* que traduz a variação da variável resposta que é explicada

pelo modelo e a soma dos quadrados dos resíduos *(SQE)* que traduz a variação da variável resposta que não é explicada pelo modelo. Matematicamente tem-se: *SQT = SQR + SQE*.

O cálculo do teste F, permite verificar qual das variáveis independentes será a responsável pela explicação. Ao determinar este teste, individualmente para cada uma das variáveis é possível verificar quais são as variáveis significativas.

Análise de Variância e Modelo Linear

Apesar da análise de variância ter surgido de forma independente, quer a Regressão linear quer a ANOVA, são consideradas, segundo Cadima (2014) casos particulares do Modelo Linear. Como tal existem conceitos abordados no tópico anterior que se podem considerar também neste tópico:

1. Variável Resposta (Y): variável quantitativa (numérica) que se pretende estudar;
2. Fator: variável preditora qualitativa categórica;
3. Níveis do fator: as várias categorias do fator, isto é, as várias possibilidades experimentais onde se realiza observações de Y.

ANOVA a um fator – Modelo de Efeitos Fixos

Seguidamente assume-se que o investigador escolheu, especificamente, os níveis do fator utilizado no planeamento de experiências, e como consequência, as inferências realizadas são com base na análise nesses mesmos níveis. O modelo matemático pode ser escrito da seguinte forma:

$$Y_{ij} = \mu + \alpha_j + e_{ij}, \quad j = 1, \dots, k \;\; e \;\; i = 1, \dots, n_j$$

Onde:

μ é o valor médio global, α_j é o efeito do j-ésimo tratamento e e_{ij} são os erros aleatórios. Os erros aleatórios são variáveis independentes, identicamente distribuídas e seguem uma distribuição de parâmetros $(0, \sigma^2)$. É com base na partição da soma dos quadrados que é construída a tabela ANOVA, como se pode ver abaixo.

F.V	GL	S.Q	Q.M	E.T
Entre os grupos	$k-1$	$SRQ = \sum_{i=1}^{k} n_i Y_i^2 - nY^2$	$QMR = SQR/(K-1)$	$F_0 = QMR/QME$
Erro ou resíduos	$n-k$	$SQR = SQT - SQR$	$QME = SQE/(n-k)$	
Total	$n-1$	$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}^2 - nY^2$		

Tabela 5. tabela ANOVA para a comparação de níveis de um fator.

Onde k é o número de grupos e n a dimensão da amostra. A tabela anteriormente apresentada é para caso geral, em caso equilibrado, ou seja, quando as amostras possuem mesma dimensão, deve-se ser considerado:

$$SRQ = r\sum_{i=1}^{k} Y_i^2 - nY^2$$

$$SQT = \sum_{i=1}^{k} (Y_{ij} - Y)^2$$

Onde r é o número de observações por tratamento, sendo esse fixo.

ANOVA a um fator – Modelo de efeitos aleatórios

Contrariamente ao modelo de efeitos fixos, por impossibilidade de considerar todos os tratamentos, o investigador seleciona aleatoriamente alguns dos tratamentos a testar. Tendo a escolha dos tratamentos sido aleatória, será possível inferir acerca de toda a população. É de considerar que nesse modelo as observações não são independentes.

No tocante aos cálculos e tabela ANOVA, não há diferença entre o modelo de efeitos fixos e modelos aleatórios, entretanto exige-se ter uma interpretação mais rigorosa (Oliveira, 2004).

ANOVA fatorial

Quando numa experiência, por se suspeitar que possa existir outros fatores que influenciam a variável estatística, e pretende analisar para além dos tratamentos, deve ser aplicada a ANOVA fatorial (Sousa, 2017).

Tendo em consideração a ANOVA de 2 fatores, o modelo matemático pode ser escrito da seguinte forma:

$$Y_{ij} = \mu + \alpha_i + \beta_i + \gamma_{ij} + e_{ij}$$

Onde

α_i a influência do nível *i* do fator 1.

β_j a influência do nível *j* no fator 2;

γ_{ij} influência combinada de ambos os níveis de seus respetivos fatores.

Com esse modelo, tem-se não apenas um teste de hipótese como na ANOVA a um fator, mas três: teste ao fator 1, teste ao fator 2 e teste à interação do fator 1 e 2. Logo, deve ser adicionado às fontes de variação na tabela ANOVA, sendo que SQR será igual ao QMR e o SQE igual ao QME, pois os graus de liberdade serão iguais a 1.

Modelos de suavização exponencial

Modelos ARIMA

Modelos ARIMA são classe de modelos populares em *forecasting*. A diferença entre modelos de suavização exponencial e os ARIMA é que os primeiros são baseados em descrever a tendência e a sazonalidade na série, enquanto os segundos se baseiam nas autocorrelações presentes nos dados.

Estacionariedade, diferenciação e autocorrelação

Uma série temporal é dita estacionária se suas propriedades (média, variância, etc.) não dependem do tempo da observação. Portanto, séries que apresentam tendência ou sazonalidade não são estacionárias. Por outro lado, uma série composta por valores gerados aleatoriamente (ex. pela função rnorm) são estacionárias, visto que a "aparência" da série é basicamente a mesma para qualquer período *t*.

Modelos autoregressivos

A diferença entre modelos regressivos e autoregressivos é que os primeiros preveem o valor de uma variável de interesse usando uma combinação linear (equação) das variáveis explanatórias. Já os segundos usam uma combinação linear de valores passados da própria variável. Matematicamente, um modelo autoregressivo é descrito como:

$$Yt = C + \emptyset_1 Y_{t-1} + \emptyset_1 Y_{t-2} + \cdots + \emptyset_p Y_{t-p} + e_t$$

Onde C é uma constante e e_t é um erro aleatório (ruído branco). Esse tipo de modelo é chamado de modelo AR(p) e são normalmente restritos a séries estacionárias.

Modelos de média móvel

Modelos de média móvel utilizam valores passados de erro de previsão de maneira semelhante a um modelo de regressão:

$$yt = c + e_t + \theta_1 Y_{t-1} + \theta_1 Y_{t-2} + \cdots + \theta_p e_{t-p}$$

Um modelo acima é chamado de modelo MA(q) e pode ser interpretado como um modelo onde *yt* é uma média ponderada dos erros de previsão passados.

A combinação entre os métodos de diferenciação e os modelos de autoregressão e média móvel resultam em um modelo ARIMA (AutoRegressive Integrated Moving Average model) não-sazonal, que pode ser descrito matematicamente como:

$$\acute{y} = c + \emptyset_1 \acute{y}_{t-1} + \ldots + \emptyset_p \acute{y}_{t-p} + \theta_1 + \theta_1 e_{t-1} + \ldots \theta_p e_{t-p} + e_t$$

Onde:

ý: é a série diferenciada. A equação acima é o que descreve o modelo ARIMA (p, d, q), onde:

- *p* é a ordem do modelo autoregressivo;
- *d* é o grau de diferenciação;
- *q* é a ordem do modelo de média móvel.

É possível saber o comportamento da previsão de um modelo ARIMA apenas baseado nos valores dos coeficientes c e d:

- Se c=0 e d=0, previsões em longo prazo serão iguais a zero;
- Se c=0 e d=1, serão iguais a uma constante maior que zero;
- Se c≠0 e d=2, seguirão uma linha reta;
- Se c≠0 e d=0, convergirão para a média da série;
- Se c≠0 e d=1, seguirão uma linha reta;
- Se c≠0 e d=2, seguirão uma tendência quadrática;

Benvenuto *et al*. (2020), construíram modelos ARIMA, que são modelos autorregressivos integrados a médias móveis, para a previsão da propagação da doença, com base em dados diários de prevalência e incidência da COVID19, fornecidos pela Universidade Johns Hopkins.

De acordo com Camargo e Villar (2021), um desafio da seleção dos modelos de séries temporais é a estimação dos parâmetros iniciais. Como cada modelo consiste em um conjunto de suposições, sejam elas explícitas ou implícitas, e um ou mais parâmetros que devem ser ajustados com o auxílio dos dados históricos disponíveis, desenvolver a melhor maneira de se chegar à boas primeiras estimativas é um dos entraves no ajuste do modelo aos dados.

Uma maneira eficiente de se estimar a relação entre diferentes conjuntos de dados é a análise de correlação, onde são estudadas potenciais relações de causa e efeito dentre os dados disponíveis (Fagundes, 2021). Para tal estudo, podem ser utilizadas diferentes linguagens de programação como ferramentas para manipular e analisar dados tais como Python, Matlab, Java e R.

SÉRIES TEMPORAIS

A distribuição da doença no tempo é um conceito amplamente difundido na área da saúde e no conhecimento geral da população sobre a ocorrência de doenças, portanto, não é incomum escutar comentários sobre a expectativa de se registrar elevação na incidência de certa doença em determinada época do ano (Gomes, 2015).

Exemplo: asma nos períodos de inverno, leptospirose nos períodos de chuva, etc.

Segundo Medronho, Werneck e Perez (2009), o estudo sobre a distribuição da doença no tempo fornece valiosas informações para a compreensão, previsão, busca etiológica, prevenção de doenças e avaliação do impacto de intervenções em saúde. Dentro desta perspetiva, faz-se necessário o registo e acompanhamento da evolução temporal das doenças para que seja possível se reconhecer padrões e tendências para a ocorrência de doenças ao longo do tempo (dias, semanas,

meses e anos) e se determinar os limites para as variações periódicas de um evento, fazendo com que seja possível se identificar elevação da incidência ou prevalência de uma doença para além do que se espera num dado período.

Os modelos de séries temporais são apropriados para analisar a evolução temporal dos dados da COVID-19 ao longo do tempo. Esses modelos permitem identificar tendências, sazonalidades e padrões cíclicos nos dados, auxiliando na compreensão da dinâmica temporal da doença e na projeção de cenários futuros. Segundo Ehlers (2007) citado por Fagundes (2021), uma série temporal consiste em um conjunto de observações feitas sequencialmente ao longo do tempo. Para Camargo e Villar (2021), uma série temporal pode ser definida como uma sequência de dados obtida em intervalos regulares de tempo durante um período específico. Por conseguinte, a previsão de séries temporais é a área de previsão que se concentra na análise das observações anteriores de uma variável para desenvolver um modelo que melhor capture os seus relacionamentos e padrões, podendo o modelo construído com base nessas observações ser usado para prever os valores futuros dessa variável aleatória.

De acordo com Brockwell e Davis (2002), o estudo por meio de séries temporais é utilizado nas mais diversas áreas de conhecimento, para diferentes finalidades, com o objetivo de fornecer informação ou explicação relativas ao problema estudado (Souza, 2022). Segundo Morettin e Toloi (1986) citado por Pinheiro *et al* (2010), quando se analisa as observações de um mesmo fenômeno, durante um determinado intervalo de tempo, os objetivos básicos são:

- modelar o fenômeno em estudo;
- obter conclusões estatísticas e
- adaptar o modelo para realizar previsões.

Bruni (2007) citado por Pinheiro *et al* (2010) aborda que o estudo de uma série de tempo costuma requerer a decomposição de seus componentes, e a análise individual de cada um deles. Estes componentes ou tipos de movimentos que atuam sobre o comportamento do fenômeno são: tendência *(Tt),* ciclo *(Ct*), variação sazonal (*St*) e termo aleatório (*at*).

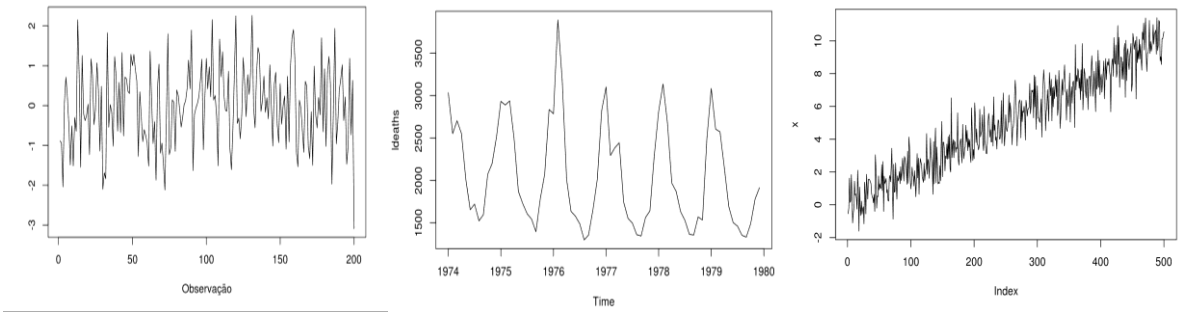
Matematicamente é representada da seguinte maneira:

$$Z_t = T_t + C_t + S_t + a_{t.}$$

A caraterística mais importante deste tipo de dados está no fato de que as observações vizinhas são dependentes, tornando necessário que essas dependências sejam levadas em consideração durante as análises e modelações.

Auffarth (2021) citado por Oliveira (2022) aponta que a análise de séries temporais envolve diversas técnicas de exploração de dados, como a visualização de distribuições, a análise de tendências e de padrões cíclicos e sazonais e a análise de relacionamentos entre variáveis de interesse. Conforme entendimento consolidado nos referenciais teóricos, uma série temporal é composta pela agregação dos seguintes elementos:

a) *Tendência:* representa o grau de variação da média dos dados da série com o decorrer do tempo. A tendência demonstra o comportamento de longo prazo da série, podendo apresentar padrões de crescimento ou de decréscimo;
b) *Sazonalidade:* caracteriza-se pela ocorrência de padrões de flutuação em intervalos regulares de tempo, muito comum, por exemplo, em dados sobre economia e clima;
c) *Variações Cíclicas:* caracterizam-se por flutuações periódicas observadas em séries temporais, alterando o comportamento da série, mas que ocorrem em frequência menor se comparadas com a flutuação sazonal – normalmente, com intervalos superiores a um ano;
d) *Ruído:* são variações irregulares na série de dados, que não apresentam padrão identificável. Enquanto os demais componentes da série temporal podem ser identificados e utilizados para modelação e previsão, o ruído é o componente não modelável, compondo grande parte dos erros existentes em modelos de previsão. Ver figura abaixo com exemplos.



De acordo Freitas *et al* (2019), uma série temporal possui tendência quando apresenta um comportamento monótono na série ao longo do tempo t, retratando a evolução global no sentido

do crescimento ou decrescimento do nível da série. A aplicação de análises estatísticas para a previsão futura de séries temporais, podem ser utilizadas na previsão de mortalidade, mesmo esses modelos tendo sido destinados a previsões de curto prazo eles têm-se mostrado úteis na realização dessas antevisões mesmo a longo prazo (Booth e Tickle 2008 citado por Proença & Schmidt, 2021).

Enquanto em modelos de regressão, por exemplo, a ordem das observações é irrelevante, para a análise em séries temporais a ordem dos dados é crucial. Vale ressaltar também que o tempo pode ser substituído por outras variáveis como espaço, profundidade, etc. Assim, de acordo com Oliveira (2022), a partir da realização da análise exploratória dos dados, é possível avaliar a aplicabilidade de diferentes modelos para previsão de séries temporais. Modelos estatísticos e de aprendizagem de máquina (inclusive métodos de aprendizagem profunda, ou *deep learning*) são comumente citados pela literatura especializada como opções para realização de previsões em séries temporais.

Dados de séries temporais surgem em vários campos do conhecimento, tais como: economia (preços diários de ações, taxa mensal de desemprego, produção industrial); medicina (eletrocardiograma, eletroencefalograma); epidemiologia (número mensal de novos casos da doença)e meteorologia (precipitação pluviométrica, temperatura diária, velocidade do vento), (Fagundes, 2021).

Correlação e estatística de associação

Correlação de Kendall

A correlação de kendall é uma medida de associação para variáveis ordinais, ela dará uma medida de grau de associação para duas variáveis.

Formula:

Sendo: X: Quantidades de Pares concordantes e Y: Quantidade de pares discordantes.

$$\tau = \frac{(X)-(Y)}{n(n-1)/2}$$

Análise do τ

τ =1: Indica uma correlação positiva muito forte

τ=0: Indica que não há correlação

τ= -1: Indica uma correlação negativa muito fraca.

Correlação de Spearman

A correlação de Spearman mensura a intensidade da relação, com o uso de uma função monótona, entre duas variáveis. Diferente da correlação de Pearson essas variáveis podem ser lineares ou não e também podem ser contínuas ou ordinais.

Sendo:

$d_1 =$ distância do alcanse do elemento n

n: número de observações.

$$\rho R = 1 - \frac{\sum_i d_i^2}{n(n^2-1)}$$

Análise para o ρR

ρR=1: Indica uma correlação positiva muito forte

ρR=0: Indica que não há correlação

ρR= -1: Indica uma correlação negativa muito fraca.

Correlação de Person

A correlação de Pearson mensura a associação linear entre variáveis contínuas, sendo este o valor que expressa o quanto a relação entre as variáveis pode ser descrita em uma reta.

Formula da correlação de Pearson:

Sendo: $x_i \; e \; x_y$: Os valores medidos em ambas variáveis

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\rho = \frac{\sum_{n}^{i=1}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{n}^{i=1}(x_i - \bar{x})^2][\sum_{n}^{i=1}(y_i - \bar{y})^2]}}$$

Análise do ρ

$\rho = 1$: Indica uma correlação positiva muito forte

$\rho = 0$: Indica que não há correlação

$\rho = 1$: Indica uma correlação negativa muito fraca

Modelos de Regressão

A técnica chamada de regressão é usada para predizer o valor de uma variável Y (chamada de variável resposta ou dependente) baseado em uma ou mais variáveis X (variável explanatória ou independente). Se a regressão utiliza apenas uma variável explanatória, é chamada de regressão simples. O objetivo da regressão é representar a relação entre as variáveis respostas e explanatória por meio de uma equação matemática linear do tipo:

$$Y = \beta_1 + \beta_2 X + \epsilon \qquad \textit{Equação 3}$$

Onde: β_1 a interceptação da reta com o eixo vertical e β_2 o coeficiente de inclinação associado à variável explanatória. Tais elementos são chamados coeficientes da regressão. O termo ϵ representa o termo do erro, que é a parte de *Y* que a regressão é incapaz de explicar (por existir outras variáveis que explicariam *Y* mas que não foram incorporadas ao modelo)

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{i2} + \cdots \beta_k x_{ki} + \epsilon_i, \quad i = 1, \ldots N$$

onde y é a variável dependente, xk é a k-ésima variável explicativa, k é o parâmetro estimado para k-ésima variável e ϵ é o termo de erro. A função lm() estima esse modelo pelo método denominado de mínimos quadrados ordinários (MQO).

O conceito de regressão remonta ao ano de 1805, quando pelo conhecimento de Legrende foi publicado o método dos Mínimos Quadrados, e anos mais tarde por Gauss em 1809. Este método foi aplicado a um problema de determinar as órbitas dos corpos em tono do Sol, partindo de

observações astronómicas. Em 1921, Gauss viria a publicar um desenvolvimento da teoria dos mínimos quadrados que incluía uma versão do teorema de Gauss-Markov.

Ainda no século XIX, o termo regressão foi utilizado por Francis Galton para descrever um fenómeno biológico. Este fenómeno defendia que as alturas dos descendentes de antepassados altos tendem a regredir para uma média de alturas normal. Para Galton, a regressão apresentava apenas um significado biológico. Ao longo dos anos, este conceito foi sendo aprimorado com o conhecimento de vários estatísticos, como Fisher, Yule ou Pearson.

Nos dias de hoje, os métodos de regressão continuam a ser alvos de pesquisa. Nas últimas décadas tem vindo a ser desenvolvidos novos métodos que tornam a regressão robusta, envolvendo dados correlacionados, como é o caso das séries temporais e das curvas de crescimento.

Em regressão linear são estudadas as relações entre uma variável resposta normalmente distribuída e as variáveis dependentes. Esse estudo é muito usado, particularmente em situações relacionadas com humanos, por exemplo, idade, raça ou sexo.

Segundo Forthofer, *et al*. (2007), a regressão linear é considerada como sendo uma extensão do coeficiente de correlação. A regressão linear permite fazer uma abordagem de várias variáveis independentes e dependentes.

Modelo de Regressão Linear Simples

O modelo de regressão linear simples é definido como sendo a relação linear entre uma variável independente (X) e uma variável resposta (Y), ou seja, é possível verificar se Y é influenciada por X. É importante ressaltar que este modelo é utilizado para fazer previsões acerca do comportamento de uma variável numérica.

Sejam os pares $(x_i,\ y_i)$, $i = 1,2,3,\dots,m$ uma serie de dados, $x_i,\ y_i \in \mathbb{R}$, com $\{(x_i, y_i)\}i$ linearmente independente.

Seja a matriz $X \in \mathbb{R}^{m\times n}$ tal que $X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_m \end{pmatrix}_{m\times n}$. Chamaremos de x_i as colunas do x.

Queremos com o modelo de regressão linear:

$$\min_{\beta} \|X\beta - Y\|_2^2$$

Para tal, basta que o resíduo da minimização seja ortogonal ao subespaço gerado pelas colunas de X, ou seja, $X\beta - Y \perp X_i$, o que implica que $(X\beta - Y)^T X_i = 0$. Graficamente pode ser representado.

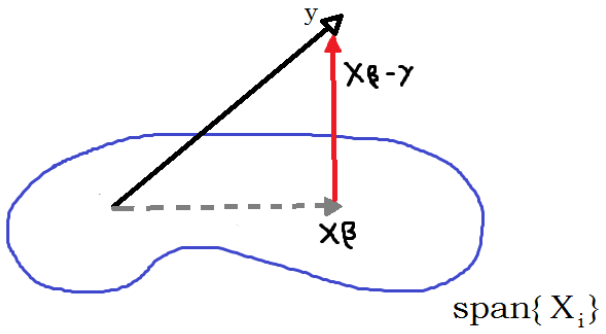


Figura 1. Ortogonalidade entre resíduo e x.

span{X_i}- espaço gerado pelas colunas de \(X \), descreve-se como conjunto de todas as combinações lineares possíveis das variáveis explicativas. O vetor \(y \) pode ser decomposto como a soma de uma parte no espaço gerado por \(X \) e uma parte ortogonal a ele.

Dai, obtém-se

$$X^T(X\beta - Y) = 0$$

$$X^T X\beta - X^T Y = 0$$

$$X^T X\beta = X^T Y$$

$$\beta = (X^T X)^{-1} X^T Y$$

A inversa de $(X^T X)^{-1}$ sempre existe, pois, supondo que as colunas de X são linearmente independentes.

O modelo teórico associado a esta regressão é dado pela seguinte equação:

$$Y_i = \beta_0 + \beta_1 \mathcal{X}_1 + \mathcal{E}_i, \quad i = 1, \dots, n$$

Onde:

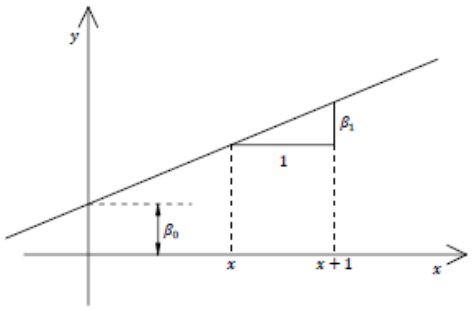
Y_i representa o valor associado à variável resposta, na observação i, $i = 1, \dots, n$

x_i representa o valor associado à variável independente, X, na observação $i = 1, \dots, n$

$\mathcal{E}_i$ representa o erro aleatório, isto é, a variável que permite explicar a variabilidade existente em *Y* que não é explicada por *X*, $i = 1, \dots, n$

β_0 *e* β_i são os parâmetros do modelo. O parâmetro β_0 representa o ponto de interseção entre a reta de regressão e o eixo das coordenadas, ou seja, quando *x* toma o valor zero. Este parâmetro pode ser também chamado de intercepto ou coeficiente linear. O parâmetro 1 representa o declive da reta de regressão, sendo que o seu valor indica a mudança na média da distribuição de probabilidade de *Y* quando ocorre o incremento de uma unidade na variável *X*.

É importante frisar que se faz necessário testar hipóteses acerca de β_1 de modo a verificar a existência de regressão linear, indicando se o modelo se ajusta ao conjunto de dados. Para uma interpretação geométrica dos conceitos representados pelos parâmetros anteriormente esclarecidos, pode-se observar o seguinte gráfico.



É necessário que as condições subjacentes ao modelo sejam atendidas, essas podem ser visualizadas mais claramente a seguir:

. $E(\varepsilon_i) = 0$;

. $Var(\varepsilon_i) = \sigma^2$

. $\{\varepsilon_i\}_{i=1}^{2}$, variáveis aleatórias independentes e identicamente distribuídas;

.$\varepsilon_i \frown N(0, \sigma^2)$

Consequentemente tem-se que:

$E(Y_i) = E(Y_i|x_i) = \beta_0 + \beta_1 x_i = \mu$;

$Var(Y_i) = \sigma^2$;

$\{Y_i\}_{i=1}^{n}$, variáveis aleatórias independentes e identicamente distribuídas;

$Y_i \frown N(\beta_0 + \beta_1 x_i, \sigma^2)$.

Após o entendimento inicial da técnica de regressão linear, para colocá-la em prática é preciso preceder a algumas análises. Assim, recomenda-se a caracterização das variáveis para se ter uma visão geral dos dados, bem como a aplicação do diagrama de dispersão e a verificação de correlação entre essas variáveis através do cálculo do coeficiente de correlação de Pearson ou de Spearman.

O coeficiente de correlação ao quadrado é o coeficiente de determinação do modelo, o qual explica a proporção que a variável resposta é explicada pelo modelo ajustado. Além de calcular esse coeficiente, faz-se necessário ainda testar a existência de correlação significativa por meio de teste de hipótese adequado.

A regressão linear simples pode ser aplicada a diversas áreas, Santos (2015), por exemplo, aplica-se esta análise para verificar a interação de um fármaco anestésico com o sinal cerebral BIS (índice bi-espectral), já no âmbito da análise risco, Silva (2016) aplica a regressão linear simples para modelar o número de pedidos de indenização e os montantes pagos pelas seguradoras.

Modelo de Regressão Linear Múltipla

O modelo de regressão linear múltipla, ao contrário do modelo abordado no tópico anterior, apresenta *k* variáveis explicativas. O modelo teórico associado é dado pela seguinte expressão:

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i, i = 1, \ldots, n$$

Onde:

Y_i representa o valor associado à variável resposta, na observação *i, i =1,..., n*;

x_i representa o valor associado à i-ésima observação das *k* variáveis independentes, *X, i = 1, ..., n;*

ε_i representa os erros aleatórios; β_0, β_1, β_2, …, β_k são os parâmetros de regressão do modelo. Estes parâmetros representam a média esperada na variável resposta, quando ocorre o incremento de uma unidade em x_i, *i = 1, ..., n*, sendo que as restantes variáveis x_k , $k \neq j$ se mantem constantes. O parâmetro β_0, tal como no modelo anterior, representa a interseção do plano k-dimensional com o eixo das coordenadas.

No que diz respeito às condições subjacentes ao modelo, essas são idênticas as do modelo de regressão linear simples e podem ser vistas no ponto anterior.

Tendo em consideração as formas de representação do modelo de regressão linear múltipla, esse pode ser também escrito matricialmente. Terá tantas linhas quanto o número de observações, genericamente consideram-se n observações, e tantas colunas quanto o número de variáveis explicativas, genericamente consideram-se *k* variáveis explicativas.

A representação matricial deste tipo de regressão tem uma dimensão dada por $n \times (p+1), n \geq p$
As condições subjacentes ao modelo em notação matricial são:

$E(\varepsilon) = 0$

$Var(\varepsilon) = \sigma^2 i_n$

X matriz característica $(p+1)$

Relativamente à aplicação da regressão linear múltipla, faz-se necessário seguir os passos de caracterização das variáveis e todo o processo descrito na regressão linear simples. Em Santos (2015), por exemplo, utilizou-se a regressão linear múltipla visando saber se dois fármacos propofol Ce e remifentanil Ce influenciam o sinal BIS, tendo em vista que nesse caso há duas variáveis independentes.

Modelos logístico

O uso da regressão logística tem estado presente nas duas últimas décadas para estimar a probabilidade de eventos dicotômicos, com aplicações em economia, medicina, análise de risco e tomadas de decisão (Frota, 2011). É uma técnica recomendada para situações em que a variável dependente é de natureza dicotômica ou binária. Quanto às independentes, tanto podem ser categóricas ou não.

A regressão logística é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias. Busca estimar a probabilidade de a variável dependente assumir um determinado valor em função dos conhecidos de outras variáveis;

• Os resultados da análise ficam contidos no intervalo de zero a um.

Na regressão logística, a probabilidade de ocorrência de um evento pode ser estimada diretamente. No caso da variável dependente Y assumir apenas dois possíveis estados (1 ou 0) e haver um conjunto de p variáveis independentes X_1, X_2, ... , X_p, o modelo de regressão logística pode ser escrito da seguinte forma:

$$P(Y = 1) = \frac{1}{1+e^{-g(x)}}$$

Onde:

$$g(x) = B_0 + B_1X_1 + B_2X_2 + \cdots + B_PX_P$$

Os coeficientes B_0, B_1, ..., Bp são estimados a partir do conjunto dados, pelo método da máxima verossimilhança, em que encontra uma combinação de coeficientes que maximiza a probabilidade da amostra ter sido observada.

Considerando uma certa combinação de coeficientes B_0, B_1, ..., B_p e variando os valores de X. Observa-se que a curva logística tem um comportamento probabilístico no formato da letra S, o que é uma característica da regressão logística. (Hosmer e Lemeshow ,1989).

a) Quando $g(x) \to +\infty$, então P(Y=1)→ 1
b) Quando $g(x) \to -\infty$, então P(Y=1)→ 0

Queremos encontrar o parâmetro que maximizam a probabilidade de obter os dados observados $(x_i, y_i), i = 1,2,3, \ldots$

Curva de regressão logística

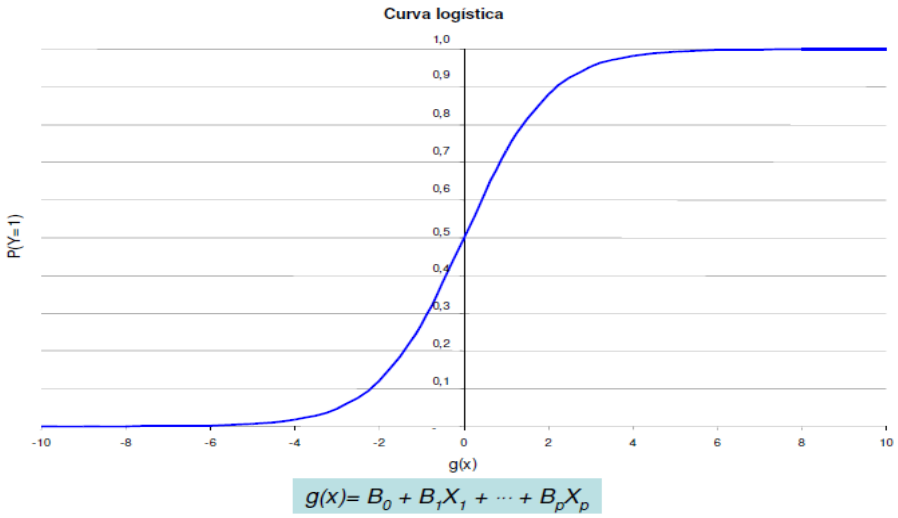


Figura 6. Curva de regressão logística.

O modelo que tem-se mostrado mais adequado para a modelação que se pretende é o Modelo Logístico. O nome vem da transformação que se faz da probabilidade de adoecer θ_i.

$$logito(\theta_i) = log(\frac{\theta_i}{1-\theta_i}) \qquad \textit{Equação 4}$$

sendo, ao longo deste trabalho, log(.) o logaritmo natural.

O modelo logístico é linear, sendo:

$$\lambda_i = log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \sum_{S=1}^{p} X_{is}\beta_s \qquad \textit{Equação 5}$$

para $i = 1,2,3,\dots,n$.

Aplicação do modelo aos estudos epidemiológicos

O modelo logístico é muito versátil, permitindo que praticamente todas as situações de interesse do pesquisador possam ser modeladas. Neste trabalho nos restringiremos a estudar o modelo com uma variável resposta binária. Na literatura, alguns autores apresentam a extensão do modelo logístico para respostas politómicas (e.g. Prentice e Pyke, 1979 citado por Barros e Lima Filho, 1994).

A seguir é mostrado como pode ser construído este modelo para uma variedade de situações, desde a mais simples, com apenas uma variável de exposição binária, até casos em que se incluem no modelo variáveis de estratificação, interação e confundimento.

Inicialmente assumindo que temos um estudo prospetivo para analisar, onde os parâmetros de interesse, incluindo os riscos relativos, são diretamente estimáveis. Em seguida estenderemos estes resultados também aos estudos de caso-controle (Barros e Lima Filho, 1994).

	Exposição	
Doença	Sim	não
Sim	d_1	d_0
não	$n_1 - d_1$	$n_0 - d_0$
Total	n_1	n_0

Tabela 6. Dados referentes a uma resposta e uma exposição binárias na forma de uma tabela 2 ´ 2.

Regressão logística

À função logística, que descreve a forma matemática na qual o modelo logístico é baseado. Tal função, chamada *f(z)*, é denotada por $f(Z) = \frac{1}{1+e^{-z}}$. Na figura a seguir, foram plotados os valores dessa função com z variando de $-\infty$ a $+\infty$.

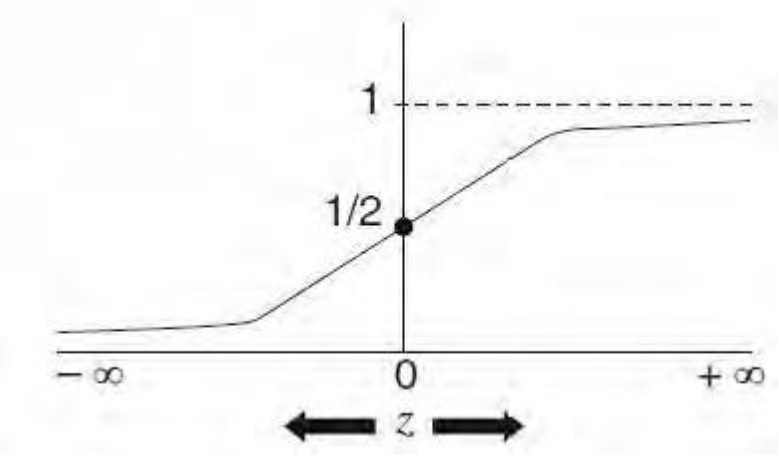


Figura 2: gráfico da função logística com z variando de -∞ a +∞. Fonte: KLEINBAUM, D. G.; KLEIN, M. Logistic Regression, a Self-Learning Text, 3ª edição, Londres-ENG, Springer, 2002, p. 5.

Modelo Multinível

O termo multinível refere-se aos níveis distintos ou unidades de análise, que geralmente, mas nem sempre, consistem em indivíduos (no nível inferior) que estão aninhados em unidades contextuais/agregadas (no nível superior) (Fisher e Getis, 2010).

Métodos multiníveis consistem em procedimentos estatísticos que são pertinentes quando (i) as observações que estão sendo analisadas são correlacionadas ou agrupadas, ou (ii) os processos causais são pensados para operar simultaneamente em mais de um nível, e/ou (iii) existe um interesse intrínseco em descrever a variabilidade e heterogeneidade do fenômeno, para além do foco na média (Diez Roux 2002; Subramanian *et al*. 2003; Subramanian 2004a, 2004b citado por Fisher, 2010).

Modelos estatísticos multiníveis são frequentemente usados em áreas como processamento de imagem e sensoriamento remoto (Kolaczyk et al. 2005). Os métodos multinível são especificamente voltados para a análise estatística de dados que possuem uma estrutura aninhada.

O modelo de regressão multinível assume que os dados são hierárquicos, com a variável de resposta medida no nível mais baixo e variáveis explicativas medidas em todos os níveis existentes (Hritcu, 2015).

A análise multinível também é conhecida como: Modelo Hierárquico Linear, Modelo de Efeitos Mistos, Modelo de Efeitos Aleatórios e Regressão Hierárquica. Ela é uma extensão do modelo de regressão tradicional quando variáveis são analisadas dispostas em vários níveis de agregação (Laros e Marciano, 2008).

De acordo com Laros e Marciano (2008), análise multinível aplica-se a uma população com estrutura hierárquica. A obtenção de uma amostra de tal população se dá pela escolha aleatória dentre unidades do nível macro (por exemplo, hospitais).

Uma vez selecionadas essas unidades, o segundo passo seria escolher, também de modo aleatório, as unidades do nível micro (por exemplo, pacientes dentro dos hospitais). Este procedimento descreve sumariamente a técnica adequada para se conseguir uma amostra na qual o pressuposto de independência entre os sujeitos não é violado; porém, por motivos práticos, de natureza financeira ou logística, o que é feito frequentemente é proceder a uma amostragem de todos os indivíduos disponíveis, depois de escolhidas as unidades do nível macro.

Nieto (2018) destaca que os dados agrupados apresentam correlações intrínsecas devido à estrutura hierárquica dos estudos, como unidades agrupadas dentro de regiões geográficas. Portanto, Nieto argumenta que é fundamental utilizar técnicas estatísticas apropriadas, como modelos de regressão multivariada hierárquica, para capturar a dependência entre as observações e obter resultados corretos.

Software estatístico para análise multinível

A capacidade de modelar relacionamentos mais complexos tem um custo computacional. Os modelos mais complexos podem atolar facilmente, não convergir para uma solução ou produzir resultados questionáveis.

Modelos multiníveis são mais exigentes em termos de dados, pois tamanhos de amostra adequados em vários níveis podem ser necessários para garantir poder suficiente para detectar efeitos; como resultado, os modelos podem se tornar bastante complicados, difíceis de estimar e ainda mais difíceis de interpretar. Esses tipos de modelos "exploratórios" geralmente são ainda mais difíceis de estimar com resultados categóricos do que com resultados contínuos (Hritcu, 2015).

Cada pacote multinível possui uma interface diferente e recursos diferentes; portanto, a escolha de qual usar é importante. Nossos dados são modelados com SPSS e R, dois dos pacotes estatísticos gerais capazes de realizar análises multiníveis.

A estimativa dos componentes de variância é uma questão importante. No caso de y contínuo, a maioria dos softwares estatísticos procede da maneira clássica usando métodos de máxima verossimilhança completa (FML) ou, mais comumente, métodos de máxima verossimilhança restrita (REML) para o modelo normal.

O método de estimativa REML geralmente é o método padrão na maioria dos pacotes. Ambos FML e REML produzem estimativas de efeitos fixos idênticas, mas REML produz estimativas de componentes de variância que são menos enviesadas.

Em pequenas amostras com dados balanceados, REML é geralmente preferível a FML porque é imparcial. Em grandes amostras, no entanto, as diferenças entre as estimativas são insignificantes (Snijders e Bosker, 1999 citado por Hritcu, 2015).

Estimar uma equação multinível pode produzir estimativas de efeito fixo e efeito aleatório para as variáveis de nível individual. O efeito fixo de uma variável é o efeito médio em toda a população dos países, expresso pelo coeficiente de regressão; o efeito aleatório fornece informações sobre se esse efeito difere ou não entre os países (Hritcu, 2015).

A escolha de definir um efeito como fixo ou aleatório nem sempre é fácil de fazer. Se faz sentido supor ou prever em bases teóricas ou metodológicas que a relação entre uma variável de nível 1 e o resultado difere entre as unidades de nível 2, isso sugere definir o efeito como aleatório. Se os testes mostrarem que os dados são inconsistentes com essa suposição, o modelo pode ser reestimado definindo o efeito a ser corrigido. A este estudo, na aplicação da análise multinível, Hritcu (2015) abordou um estudo por ele desenvolvido em que procurava investigar diferenças entre países no efeito de satisfação com a saúde sobre a satisfação com a vida. Segundo o autor para investigar diferenças entre países no efeito da satisfação com a saúde sobre a satisfação com a vida, será necessário especificar também um efeito aleatório dessa variável, o que significa que se assume que o efeito varia aleatoriamente dentro da população dos países, e o pesquisador está interessado para testar e estimar a variância desses efeitos aleatórios nessa população.

Assim, quando não há orientações teóricas ou outras prévias sobre quais variáveis devem ter efeito aleatório, o pesquisador pode ser conduzido pelo foco da investigação. As variáveis explicativas que são especialmente importantes ou têm efeitos especialmente fortes, conforme evidenciado por sua significância e tamanho, podem ser modeladas com efeitos aleatórios.

A regressão multinível é uma abordagem estatística adequada para analisar dados agrupados, como os dados da COVID-19 coletados em diferentes regiões geográficas. Nesse tipo de modelo, são incorporados níveis hierárquicos, permitindo que as observações dentro de cada região sejam consideradas dependentes umas das outras.

No contexto da COVID-19, os Modelos de Equações de Estimação Generalizada (GEE) têm sido aplicados em estudos longitudinais da COVID-19, nos quais os mesmos indivíduos são acompanhados ao longo do tempo. Segundo Horton *et al*. (2020), os modelos GEE permitem analisar a correlação entre as observações dentro dos grupos, considerando a dependência dos dados ao longo do tempo.

Modelos de equações de estimação generalizadas (GEE) são uteis quando os dados agrupados apresentam correlação intraclasse, ou seja, quando há dependência entre as observações dentro do mesmo grupo ou cluster. Essa abordagem é amplamente utilizada em estudos epidemiológicos.

Os GEEs permitem o ajuste dos parâmetros de interesse, levando em consideração a correlação entre as observações, ao mesmo tempo que fornecem estimativas consistentes pra efeitos de interesse (Zeger, Liang e Abert, 1988).

Vários foram os trabalhos científicos desenvolvidos ao longo do tempo que fornecem fundamentos importantes para embasar sobre o GEE, e dentre vários trabalhos desenvolvidos, destacam-se no presente estudo o trabalho de Wang *et al*. (2020), que contem uma revisão abrangente sobre aspetos relevantes da epidemiologia da COVID-19 e discute metodologias estatísticas aplicada ao estudo da doença.

Firth (2020) apresenta em seu estudo uma abordagem estatística para reduzir os enviesamentos em estimativas de máxima verossimilhança. Essa metodologia pode ser útil na análise de dados agrupados em estudos epidemiológicos, onde o agrupamento pode introduzir viés nos resultados. A compreensão e aplicação da abordagem de Firth (2020) ajuda a melhorar a precisão das estimativas e inferências em estudos similares a este.

Além disso, para a identificação de fatores de risco e entender a variação da gravidade da doença entre diferentes grupos, é destacado o estudo de Verity *et al.* (2020) que estima a gravidade da COVID-19 por meio de modelos estatístico, considerando dados agrupados de diferentes regiões. Também, nessa abordagem, o artigo de Raigor *et al*. (2020) destaca a importância da metodologia estatística adequada na obtenção de estimativas confiáveis, como a taxa de letalidade da COVID-19.

O trabalho de Raigor *et al.* (2020), ressalta a necessidade de considerar as limitações e incertezas ao analisar dados agrupados e apresenta uma reflexão critica sobre diferentes estimativas.

Outra metodologia usada na análise de dados agrupados em estudos epidemiológicos são a regressão logística multinível. A Regressão Logística multinível é adequado quando os dados possuem uma estrutura hierárquica, com indivíduos agrupados em diferentes níveis.

Um exemplo de aplicação desta abordagem é um estudo epidemiológico sobre doenças infeciosas em diferentes regiões geográficas, onde os indivíduos seriam agrupados por regiões. A regressão logística multinível permite modelar a variabilidade entre os grupos, considerando os diferentes níveis de agrupamentos (Sniiders e Bosker, 2012).

Segundo Merlo *et al*. (2017), essa abordagem é adequada para lidar com a estrutura hierárquica dos dados, permitindo a modelação dos efeitos individuais e dos efeitos do grupo, o que é essencial para capturar a variação tanto dentro dos grupos como entre os grupos.

De acordo com Muller *et al*. (2021), no contexto da COVID-19, muito estudo epidemiológico tem sido conduzido para compreender a propagação da doença, os fatores de risco e a eficácia das intervenções. Dentre os diversos estudos, um estudo desenvolvido por Raigor *et al*. (2020) investigou a prevalência da COVID-19 em uma determinada região, utilizando amostras de indivíduos em diferentes comunidades. A análise foi realizada utilizando regressão logística multinível, levando em consideração tanto os efeitos individuais como os efeitos do grupo.

Além disso, estudos de coorte tem sido amplamente utilizado para acompanhar indivíduos ao longo do tempo e avaliar a taxa de infecção da COVID-19, bem como identificar os fatores de riscos associados. O estudo realizado por Punge *et al*. (2020) acompanhou uma coorte de trabalhadores de saúde durante a pandemia. A análise foi conduzida utilizando modelos de sobrevivência fracionais, considerando a estrutura de agrupamento dos dados e dependência entre os indivíduos dentro dos grupos.

Os Modelos de Sobrevivência Fracionais têm sido amplamente utilizados em estudos de sobrevida, que avaliam o tempo até a ocorrência de um evento, como a mortalidade pela COVID-19. Segundo Crowther *et al*. (2019), esses modelos são adequados para lidar com a dependência e a correlação entre os indivíduos dentro dos grupos.

Um dos estudos de destaque desenvolvido sobre modelos de sobrevivências é o estudo de Goldstein *et al*. (2021), que investigaram a taxa de mortalidade pela COVID-19 em diferentes países, considerando a estrutura de dados por meio de modelos multinível. Os autores puderam

identificar fatores individuais (como idades e comorbidades) e fatores do grupo (como medidas de controle da pandemia) que influenciaram a taxa de mortalidade.

Análise de variância (ANOVA) com efeitos fixos ou mistos, a ANOVA é aplicada quando há interesse em comparar medias entre grupos, levando em consideração o agrupamento dos dados. A ANOVA de efeitos fixos é apropriada quando os grupos são selecionados intensionalmente, enquanto a ANOVA de efeitos mistos é utilizada quando os grupos são considerados uma amostra aleatória da população (Maxwell e Delaney, 2004).

Aplicação dos dados reais sobre COVID-19 no Software R

Análise descritiva dos dados

A análise descritiva foi realizada para obter resumo das características dos dados recolhidos. Essa etapa é essencial para a compreensão inicial dos dados agrupados em estudos epidemiológicos. Utilizou-se os seguintes comandos no R para realizar a análise descritiva:

```
R
#Carregar os dados ("dados_covid.csv")
Dados<-read.csv("dados_covid.csv")
#Estatísticas descritivas
summary(dados_covid$Total_cases)
#Histograma
hist(dados_covid$Total_cases, main = "Distribuição de casos totais de COVID-19", xlab = "Total de casos"
```

Tabela 7. Análise descritiva dos dados.

A análise descritiva permite ter uma visão geral dos dados recolhidos, identificando medidas de tendência central, como média, mediana e quartis, bem como a distribuição dos dados por meio de

gráficos, como histogramas. Essa etapa inicial é importante para a compreensão básica dos dados (Montgomery, Peck & Vining, 2012).

```
> head(dados_covid)
 [1] total_cases                          new_cases                          new_cases_smoothed
 [4] total_deaths                         new_deaths                         new_deaths_smoothed
 [7] total_cases_per_million              new_cases_per_million              new_cases_smoothed_per_million
[10] total_deaths_per_million             new_deaths_per_million             new_deaths_smoothed_per_million
[13] reproduction_rate                    icu_patients                       icu_patients_per_million
[16] hosp_patients                        hosp_patients_per_million          weekly_icu_admissions
[19] weekly_icu_admissions_per_million    weekly_hosp_admissions             weekly_hosp_admissions_per_million
[22] total_tests                          new_tests                          total_tests_per_thousand
[25] new_tests_per_thousand               new_tests_smoothed                 new_tests_smoothed_per_thousand
[28] positive_rate                        tests_per_case                     tests_units
[31] total_vaccinations                   people_vaccinated                  people_fully_vaccinated
[34] total_boosters                       new_vaccinations                   new_vaccinations_smoothed
[37] total_vaccinations_per_hundred       people_vaccinated_per_hundred      people_fully_vaccinated_per_hundred
[40] total_boosters_per_hundred           new_vaccinations_smoothed_per_million  new_people_vaccinated_smoothed
[43] new_people_vaccinated_smoothed_per_hundred  stringency_index            population_density
[46] median_age                           aged_65_older                      aged_70_older
[49] gdp_per_capita                       extreme_poverty                    cardiovasc_death_rate
[52] diabetes_prevalence                  female_smokers                     male_smokers
[55] handwashing_facilities               hospital_beds_per_thousand         life_expectancy
[58] human_development_index              population                         excess_mortality_cumulative_absolute
[61] excess_mortality_cumulative          excess_mortality                   excess_mortality_cumulative_per_million
<0 rows> (or 0-length row.names)
>
```

Figura 7. Visualização dos dados (variáveis) carregados do ficheiro Owid_data_COVID do site da OMS no ambiente de trabalho do software R.

Com base no comando "*summary"* apresentado na tabela 1, a seguir é apresentado o resumo dos dados contidos na base de dados a ser manipulados.

```
> summary(dados_covid)
  total_cases       new_cases      new_cases_smoothed  total_deaths      new_deaths     new_deaths_smoothed total_cases_per_million new_cases_per_million
 Min.   :     1   Min.   :   0.0   Min.   :   0.000   Min.   :   1.0   Min.   : 0.000   Min.   : 0.000      Min.   :   0.03        Min.   :  0.000
 1st Qu.: 19573   1st Qu.:   0.0   1st Qu.:   6.143   1st Qu.: 600.8   1st Qu.: 0.000   1st Qu.: 0.000      1st Qu.: 593.68         1st Qu.:  0.000
 Median :151185   Median :  22.0   Median :  26.429   Median :1938.5   Median : 0.000   Median : 0.286      Median :4585.58         Median :  0.667
 Mean   :133339   Mean   : 188.5   Mean   : 189.308   Mean   :1443.6   Mean   : 1.812   Mean   : 1.819      Mean   :4044.32         Mean   :  5.719
 3rd Qu.:229696   3rd Qu.: 108.8   3rd Qu.: 106.000   3rd Qu.:2219.0   3rd Qu.: 1.000   3rd Qu.: 1.286      3rd Qu.:6966.92         3rd Qu.:  3.299
 Max.   :233417   Max.   :6640.0   Max.   :3837.143   Max.   :2243.0   Max.   :34.000   Max.   :26.143      Max.   :7079.78         Max.   :201.398
 NA's   :80                        NA's   :5          NA's   :144                       NA's   :5           NA's   :80
 new_cases_smoothed_per_million total_deaths_per_million new_deaths_per_million new_deaths_smoothed_per_million reproduction_rate icu_patients
 Min.   :  0.000                Min.   : 0.03            Min.   :0.00000         Min.   :0.00000                Min.   :0.0000    Mode:logical
 1st Qu.:  0.186                1st Qu.:18.22            1st Qu.:0.00000         1st Qu.:0.00000                1st Qu.:0.6925    NA's:1238
 Median :  0.802                Median :58.80            Median :0.00000         Median :0.00900                Median :0.9200
 Mean   :  5.742                Mean   :43.79            Mean   :0.05493         Mean   :0.05517                Mean   :0.9401
 3rd Qu.:  3.215                3rd Qu.:67.31            3rd Qu.:0.03000         3rd Qu.:0.03900                3rd Qu.:1.1800
 Max.   :116.385                Max.   :68.03            Max.   :1.03100         Max.   :0.79300                Max.   :2.6200
 NA's   :5                      NA's   :144                                      NA's   :5                      NA's   :272
 icu_patients_per_million hosp_patients  hosp_patients_per_million weekly_icu_admissions weekly_icu_admissions_per_million weekly_hosp_admissions
 Mode:logical             Mode:logical   Mode:logical              Mode:logical          Mode:logical                     Mode:logical
 NA's:1238                NA's:1238      NA's:1238                 NA's:1238             NA's:1238                        NA's:1238
```

Figura 8. Sumário dos dados de Covid-19 em Moçambique, com todas variáveis apresentadas.

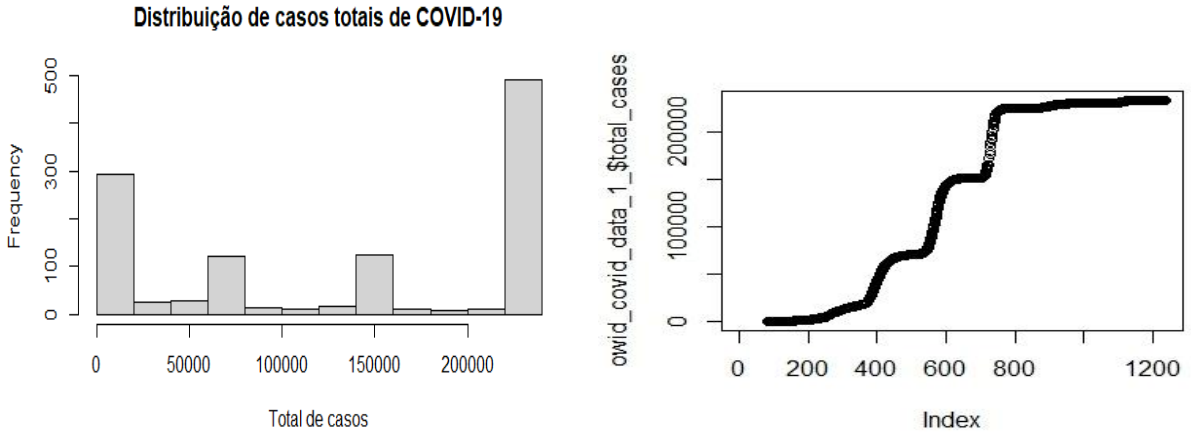


Figura 9. Total de casos da COVID-19 em Moçambique.

Os casos de COVID-19 em Moçambique apresentaram um crescimento exponencial, o que é consistente com a natureza altamente transmissível do vírus, como evidenciado em outros estudos desenvolvidos.

Gráfico de dispersão

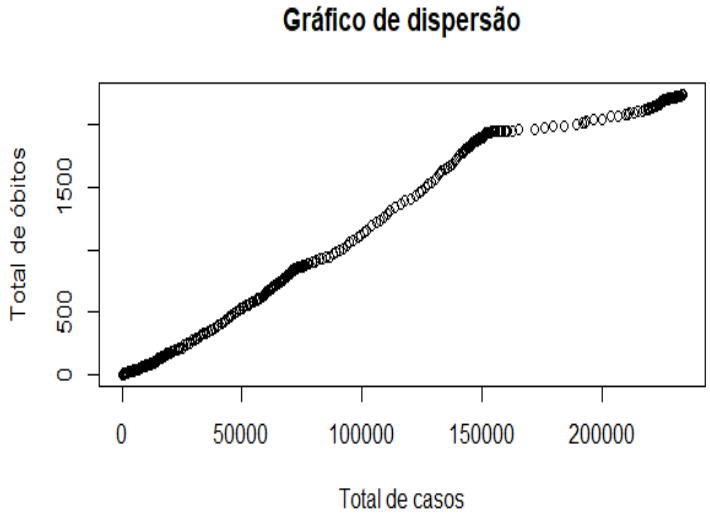


Figura 10. Gráfico de dispersão totais de casos de COVID e Óbitos

Dado o carácter temporal dos dados da COVID-19, foram realizadas análise de séries temporais para identificar padrões sazonais, tendências e flutuações nos casos ao longo do tempo. Foram aplicadas técnicas como modelos de suavização exponencial.

Segundo Ehlers (2007) citado por Fagundes's e Oliveira (2021), uma série temporal consiste em um conjunto de observações feitas sequencialmente ao longo do tempo. A característica mais importante deste tipo de dados está no facto de que as observações vizinhas são dependentes, tornando necessário que essas dependências sejam levadas em consideração durante as análise e modelações.

Para a análise de tendências, o R possui funções e pacotes especializados para análise de séries temporais, como o *"forecast*" e o "*tsibble".* Esses pacotes permitem realizar análises de decomposição de séries temporais, identificação de padrões sazonais e construção de modelos de previsão. No presente trabalho foi usado o *pacote forecast*.

R

#Gráfico de linhas para casos diários

Plot(dados_covid$New_cases, type = "1", main= "Evolução temporal de casos de COVID-19", xlab= "Dias", ylab="Novos casos")

#Suavização dos novos casos usando média móvel

Dados_covid$New_case_smoothed<-ma(dados_covid$New_cases, order=7, centre=TRUE)

plot(dados_covid$New_case_smoothed, type="1", main="Suavização de novos casos de COVI-19", xlab="Dias", ylab="Novos casos suavizados")

Tabela 8. Gráfico de linhas para casos diários. Autor, 2023.

De acordo com Chatfiel (2004), a análise auxilia na compreensão da dinâmica da propagação da doença e pode fornecer insights para medidas de controle e intervenção.

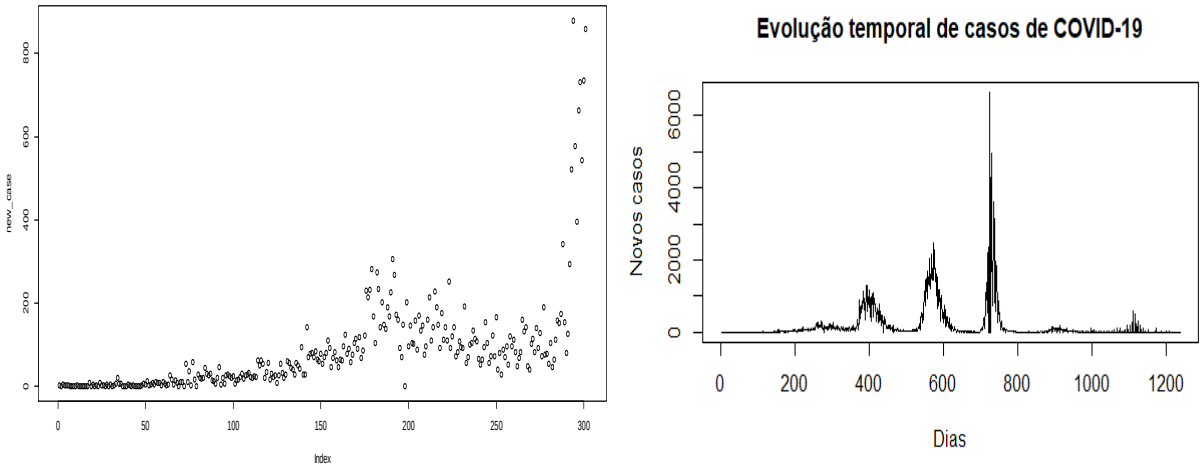


Figura 11. Evolução temporal de casos de COVID-19 em Moçambique

A análise temporal permitiu visualizar a evolução dos casos ao longo do tempo, possíveis picos, sazonalidades e tendências, onde, conforme o gráfico é evidente a tendência crescente durante a fase inicial, e um certo abrandamento de número de casos depois de algum tempo.

Suavização dos novos casos usando média móvel

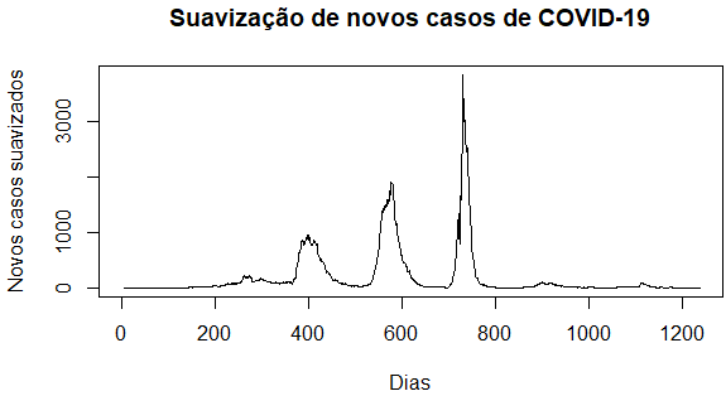
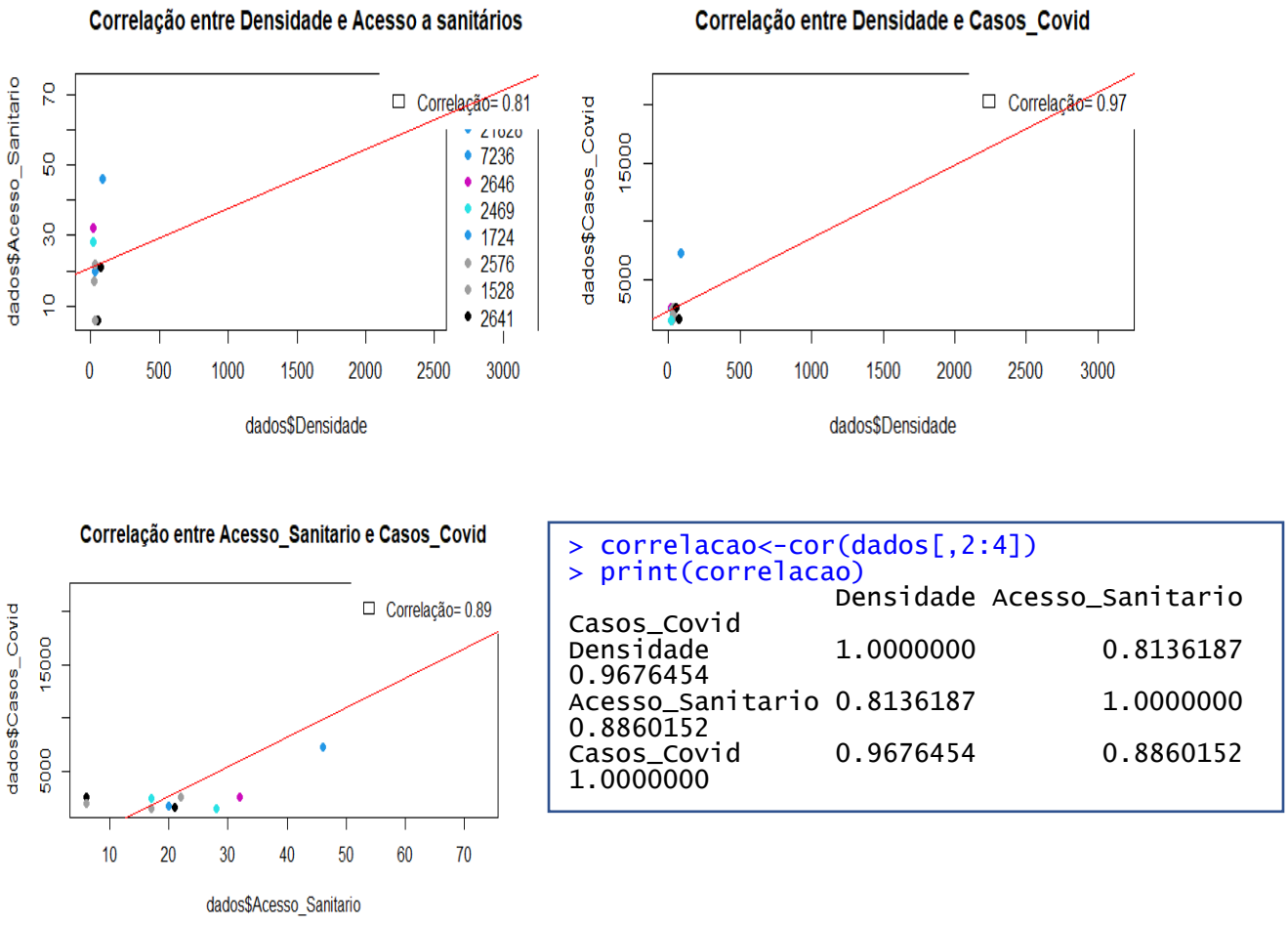


Figura 12. Novos casos suavizados.

Para compreender a relação entre as variáveis, foram realizadas análises de correlação e regressão entre estes. A análise de correlação avalia a relação entre duas variáveis, auxiliando na identificação de possíveis associações entre os dados agrupados em estudos epidemiológicos (Hair, Black Barbin e Anderson, 2014). De acordo com Cavalcante e Vasconcelos (2018), para descrever as relações entre uma variável respostas e algumas covariáveis em dados agrupados de

acordo com um ou mais fatores são usados os modelos de efeitos mistos. Segundo os autores, a associação dos efeitos comuns a observação que compartilham o mesmo nível de classificação, os modelos mistos representam uma forma flexível de estrutura de correlação induzida pelo agrupamento dos dados.

Esta análise tem como objetivo investigar a relação entre variáveis, fornecendo informações sobre a intensidade e direção dessa relação. O R possui funções para calcular coeficientes de correlação, como "*cor()"* e "*cor.test()*", bem como pacotes como o *"corrplot"* para visualização de matrizes de correlação. Procurando compreender a existência da correlação entre as variáveis densidades populacionais, pessoas com acesso a instalações sanitárias e casos confirmados da COVID-19, foram obtidos da análise estatística no R, os seguintes resultados:




```
> correlacao<-cor(dados[,2:4])
> print(correlacao)
                 Densidade Acesso_Sanitario
Casos_Covid
Densidade        1.0000000         0.8136187
0.9676454
Acesso_Sanitario 0.8136187         1.0000000
0.8860152
Casos_Covid      0.9676454         0.8860152
1.0000000
```

Figura 13. correlação entre as variáveis densidades populacionais, pessoas com acesso a instalações sanitárias e casos confirmados da COVID-19

Da análise feita da correlação entre a densidade, acesso sanitário e os casos de COVID-19, foi obtido uma correlação entre 0.81, 0.89 e 0.97 o que indica que existe uma correlação positiva forte entre as duas variáveis. Isso significa que a medida que a densidade populacional aumenta, a taxa de acesso a instalações sanitárias também tende a aumentar de forma significativa.

Considerando ainda a correlação, foram realizados o cálculo da matriz de correlação entre o total de casos por milhão de habitantes e o índice de restrições aplicadas. Os resultados incluem o valor do *teste t*, os graus de liberdade, o valor de *p*, a hipótese alternativa, o intervalo de confiança de 95% e o coeficiente de correlação.

Na tabela a seguir, apresenta-se o resultado do cálculo da matriz de correlação entre as variáveis "Total_cases_per_milion"(total de casos por milhão de habitantes) e "Stringency_index" (índice de restrições aplicadas).

Antes disso, importa realçar que a matriz de correlação é uma representação tabular que apresenta os coeficientes de correlação entre as variáveis. Valores aproximados de 1 indicam uma correlação positiva forte, valores aproximados a -1 indicam uma correlação negativa forte, e valores próximos de 0 indicam uma correlação fraca ou inexistente.

Em dados não agrupados

```
cor_total_cases_stringency_index<-cor.test(covid_data$total_cases,covid_data$stringency_index)
> print(cor_test_tests_cases2)

        Pearson's product-moment correlation

data:  covid_data$total_cases and covid_data$stringency_index
t = -56.687, df = 1012, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8860585 -0.8564841
sample estimates:
      cor
-0.8720652


>
```

Tabela 1. Correlação entre total de casos e índice de rigor.

Em dados agrupados:

```
> agrupado<-aggregate(covid_data[,c("total_cases","stringency_index")],by=list(covid_data$total_cases,covid_data$string
ency_index),FUN=mean)
 cor_total_cases_stringency_index2<-cor.test(agrupado$total_cases,agrupado$stringency_index)
 >
 > print(cor_total_cases_stringency_index2)

                Pearson's product-moment correlation

 data:  agrupado$total_cases and agrupado$stringency_index
 t = -50.539, df = 872, p-value < 2.2e-16
 alternative hypothesis: true correlation is not equal to 0
 95 percent confidence interval:
  -0.8793797 -0.8455134
 sample estimates:
       cor
 -0.8634161



 >
```

Tabela 10. Correlação entre total de casos e índice de rigor.

Contudo, os resultados da análise revelam um valor extremamente baixo para o *teste t* (-56.68), com 1012 graus de liberdade. O valor do *p* encontrado foi menor que 2.2e-16, indicando que a probabilidade de obter uma correlação forte entre o total de casos e as medidas de controle e restrição é praticamente nula. Portanto, rejeitamos a hipótese nula de que não há correlação entre essas variáveis.

O coeficiente de correlação encontrado foi de -0.8720652, o que indica uma correlação negativa muito forte entre o total de casos de COVID-19 e as medidas de controle e restrição. Este valor está dentro do intervalo de confiança de 95%, que varia de -0.8860585 -0.8564841, o que reforça a robustez da relação negativa observada.

Contudo, da análise da correlação de dados brutos e dados agrupados foram obtidos valores com diferenças muito pequenas, o que não é substancial e, em termos práticos, ambas as correlações indicam uma relação negativa entre as variáveis. A pequena variação entre estas pode advir de algumas razões: Arredondamentos de dados originais, grupo de dados, variabilidade de dados.

A relação destas variáveis também é explicada por meio de gráfico de dispersão.

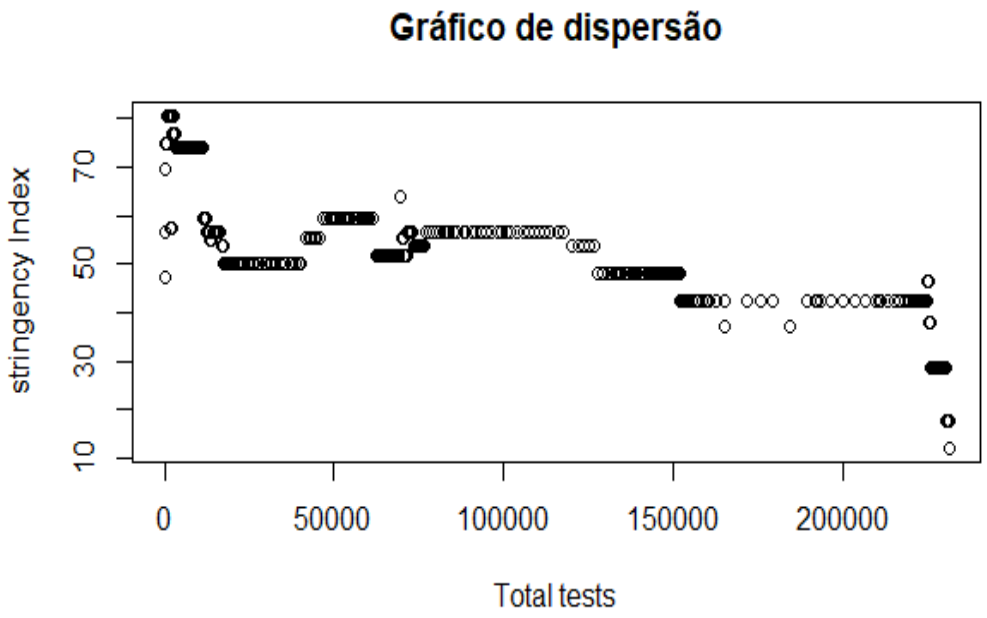


Figura 14. gráfico de dispersão entre total de testes e índice de restrição.

A correlação negativa encontrada neste estudo sugere que a medida que as medidas de controle e restrição são intensificadas, o total de casos da COVID-19 tende a diminuir. Isso é consistente com as teorias e evidencias existentes de que a implementação de medidas rigorosas pode reduzir a disseminação do vírus.

Correlação de Pearson entre total de testes realizados e total de casos confirmados.

```
cor_test_tests_cases<-cor.test(covid_data$total_tests,covid_data$total_cases)
> print(cor_test_tests_cases)

        Pearson's product-moment correlation

data:  covid_data$total_tests and covid_data$total_cases
t = 171.55, df = 614, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9879778 0.9912265
sample estimates:
      cor
0.9897291
```

Tabela 11. Correlação de Person entre total de testes realizados e total de casos confirmados.

A análise da tabela 5 revelou um valor impressionante para o *teste t* (171.55), com 614 graus de liberdade. O valor de *p* encontrado foi menor do que 2.2e-16, indicando a probabilidade de obter uma correlação forte entre o total de testes e o total de casos de COVID-19 ao acaso é praticamente nula. Portanto rejeitamos a hipótese nula de que não há correlação entre essas variáveis.

O coeficiente de correlação encontrado foi de 0.9897291, o que indica uma correlação positiva muito forte entre o total de testes e o total de casos de COVID-19. Este valor está dentro do intervalo de confiança de 95%, que varia de 0.9879778 a 0.9912265, destacando mais a robustez da relação positiva observada entre as variáveis.

Assim sendo, a correlação observada nesta análise sugere que, à medida que o número total de testes para a COVID-19 aumenta, o número total de casos confirmados também tendem a aumentar, o que se justifica pelo facto de que, quanto mais testes são realizados, maior a probabilidade de identificar casos positivos.

Análise de regressão

A análise de regressão permite investigar a relação entre uma variável dependente e uma ou mais variáveis independentes, auxiliando na compreensão dos fatores que podem influenciar os dados agrupados em estudos epidemiológicos (Kutner, Nachtsheim, Neter e Li, 2004).

R

#Regressão linear

```
>
> modelo<-lm(casos_covid~dados$Densidade,data = dados)
> summary(modelo)
```

Gráfico de dispersão com linhas de regressão

```
>
> plot(dados_covid$population_density,dados_covid$total_cases,
main="Regressao linear", xlab="Densidade populacional", ylab=
"Total de casos")
>
```

```
summary(modelo)

Call:
lm(formula = casos_covid ~ dados$Densidade, data = dados)

Residuals:
   Min     1Q  Median     3Q    Max
-1189.0  -854.2   -68.3    2.3  4342.6

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2350.2713   512.7512   4.584  0.00132 **
dados$Densidade    6.2427     0.5426  11.505  1.1e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1600 on 9 degrees of freedom
Multiple R-squared:  0.9363,        Adjusted R-squared:  0.9293
F-statistic: 132.4 on 1 and 9 DF,  p-value: 1.101e-06


>
```

Tabela 1. Analise de regressão linear entre os variáveis casos de COVID-19 e Densidade populacional.

O R-Square é de 0.9363, o que significa que aproximadamente 93.63% da variabilidade casos da COVID-19 é explicada pela densidade populacional. Este valor sugere uma forte relação entre as variáveis. Estes dados indicam que a densidade populacional é uma variável importante na explicação dos casos de COVID-19.

Foi obtido um valor de p-value 1.101e-06 associado ao modelo de regressão. Este valor sugere que a relação entre a densidade populacional e casos de COVID-19 é significativa. Estes resultados sugerem que a densidade populacional tem um impacto significativo na propagação dos casos de COVID-19. Assim, áreas com maiores densidades populacional tendem a ter mais casos de COVID-19.

Os resultados acima justificam-se pois em áreas densamente povoadas, pessoas tendem a ter contactos mais próximos umas das outras devido a proximidade física, seja transportes públicos, moradias próximas ou espaços comerciais movimentados e ou lotados, o que facilita a transmissão do vírus pessoa para pessoa. Por outro lado, em áreas densamente urbanas, a mobilidade de pessoa é maior.

Um estudo desenvolvido na China no ano de 2020, analisou os dados iniciais da pandemia em Wuham e constatou que as áreas mais densamente povoadas da cidade tinham uma taxa de infeção significativamente alta.

Gráfico de dispersão com linha de regressão

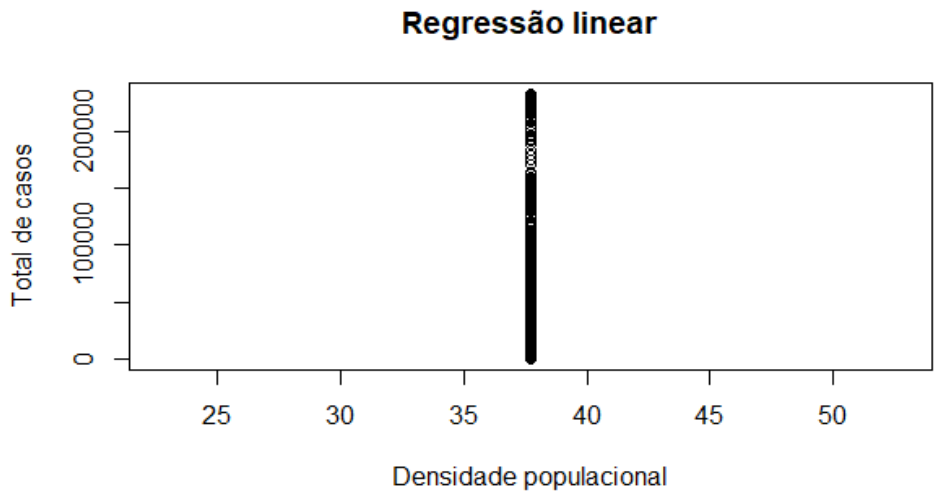


Figura 15. Gráfico de dispersão com linha de regressão. (Autor, 2023)

Os resultados demonstram qua a metodologia estatística de análises de dados agrupados é altame nte eficaz na investigação de epidemiologias, especificamente COVID-19. Ao explorar as tendên cias temporais das taxas de casos de COVID-19 em Moçambique, por meio de análises, os result ados do presente estudo reforçam as descobertas de outros pesquisadores que investigaram sobre COVID-19 em Moçambique e no mundo.

As análises de séries temporais revelam tendências bem como padrões de crescimento sendo que Os modelos de regressão dão informações sobre a influência de variáveis económicas e medidas de intervenção. Buscando e comparando estudos que investigaram a propagação da COVID-19 em contextos semelhantes desenvolvidos na área, observamos alinhamento dos resultados e concordâncias nas tendências e na influência de medidas de intervenção.

Silva *et al*. (2022) conduziram uma pesquisa abrangente em Moçambique, analisando o efeito das medidas de intervenção na disseminação da COVID-19. Assim como no presente estudo, eles identificaram uma correlação significativa entre implementação de medidas rigorosa e a redução nas taxas de casos. Assim, ao comparar o resultado presente estudo com os resultados obtidos por

Silva *et al*. (2022), pode-se inferir que as medidas de intervenção desempenharam um papel crucial no controle de disseminação do vírus em Moçambique.

Com o passar dos tempos é possível observar a redução de queda das taxas de casos confirmados, que foi o resultado da implementação de medidas rigorosas de prevenção, que reforça a eficácia dessas intervenções, ressaltando a importância de políticas publicas proativas para conter a pandemia.

De forma análoga, Muller *et al*. (2021) documentaram em seu estudo que medidas rigorosas de *lockdown* na Europa ocasionaram reduções significativas nas taxas de casos e óbitos, corroborando a observação do presente estudo.

Ao investigar as relações entre variáveis independentes e dependentes usando modelos de regressão, os resultados do presente estudo contemplam estudos anteriores em Moçambique. Goncalves *et al*. (2021) de forma semelhante identificaram associações entre as variáveis económicas e a disseminação da COVID-19 no país. As coerências entre esses estudos destacam a relevância continua dos modelos de regressão na exploração dos fatores adjacentes à epidemiologia da doença em Moçambique.

De forma semelhante, Mabunda *et al*. (2023) desenvolveram analise espacial da incidência da COVID-19 em Moçambique. Eles destacam a presença de clusters geográficos de casos, concordando com a nossa pesquisa sobre os padrões de propagação.

Esta convergência entre os resultados e os de estudos anteriores reforça a consistência das conclusões e a robustez em cada metodologia empregada para analise no presente estudo. O padrão espacial sugere a influencia de fatores locais de propagação da COVID-19, como densidade populacional e conectividade entre as regiões.

Abordando propagação do vírus em aglomerados, Khan *et al*. (2020) investigaram a propagação do vírus em um aglomerado urbano na índia e identificam clusters geográficos de casos, reforçando a importância dos fatores locais na disseminação. A correspondência transcultural sublinha a influencia crucial do contexto geográfico na dinâmica da pandemia.

Em consonância com Goncalves *et al*. (2021), observa-se associações entre variáveis económicas e a disseminação da covid-19. Smith *et al*. (2020) examinaram em seu estudo disparidades

socioeconómicas nos Estados Unidos, descobrindo que comunidades desfavorecidas eram mais suscetíveis a doença. Esses padrões foram também relatados em diferentes continentes, reforçando a complexidade interconectada entre fatores económicos e saúde.

A abordagem metodológica deste estudo, integrada análise estatísticas sobre a variabilidade e distribuição dos casos da COVID-19, é escoada em uma investigação africana de destaque. El-Sadr *et al*. (2022) discutem abordagens globais abrangentes para a COVID-19 em países africanos, realçando a importância de uma perspetiva multifacetada.

Constata-se consistência nos resultados do presente estudo com os dos estudos locais, o que reforça a validade e a aplicabilidade deste trabalho em contextos epidemiológicos em Moçambique.

CONCLUSÃO E RECOMENDAÇÕES

A análise de dados é parte essencial de muitas áreas de pesquisa científica, constituindo uma técnica valiosa da análise estatística. Ao organizar dados em intervalos, torna-se mais fácil compreender tendências, padrões e resultados nos conjuntos extensos de dados. Os dados agrupados geralmente envolvem categorização ou agregação de informações em grupos distintos. Isto, incorre por vários motivos, como a conveniência na apresentação de dados, a simplificação de uma grande quantidade de informação ou a necessidade de comparar diferentes categorias.

Existem várias metodologias estatísticas para análise de dados, dependendo do tipo de dados e das perguntas da pesquisa, e cada uma com suas próprias aplicações e vantagens, e um passo muito importante na análise dos dados é a seleção do método estatístico apropriado. Nos estudos com análises dos dados as metodologias mais usadas são a estatística descritiva e a estatística inferencial.

Para análise de dados agrupados, sobretudo no contexto epidemiológico, diferente abordagem estatística tem sido amplamente utilizada na literatura científica. Algumas abordagens amplamente utilizadas incluem os Modelos Lineares Generalizados, a regressão Multinível, os modelos de sobrevivência, Análise de variância (ANOVA) com efeitos fixos ou mistos, series temporais entre outras metodologias dependendo do objetivo da pesquisa.

Para cada situação especifica estão disponíveis métodos estatísticos para análise e interpretação dos dados. Para selecionar os métodos estatísticos apropriados é necessário conhecer os pressupostos e as condições dos métodos estatísticos, para que o método estatístico adequado seja selecionado para análise de dados. Os fatores cruciais para a seleção do método estatístico apropriado, contemplam a finalidade e objetivo do estudo, o tipo e distribuição dos dados e a natureza das observações (emparelhada ou não emparelhadas), sendo nesta etapa fundamental a compreensão dos dados em si. Uma seleção errada do método estatístico, para além de criar problemas sérios durante a interpretação dos resultados, afeta a conclusão do estudo. A aplicação de metodologias estatísticas incorretas infelizmente pode ser constatada em diversos trabalhos de aplicações a dados reais, sobretudo por falta de entendimento entre os especialista e pela não inclusão de estatísticos em estudos de caso de certas especialidades.

Este estudo demonstrou que, a combinação de metodologias estatísticas com o *software* R, pode ser aplicada de forma eficaz na análise de dados agrupados em estudos epidemiológicos.

Para compreender a relação entre as variáveis, foram realizadas por meio de aplicação prática no R análises de correlação e regressão entre variáveis sobre COVID-19 em Moçambique. Da análise da correlação feita entre a densidade, o acesso sanitário e os casos de COVID-19, foram obtidos uma correlação positiva forte entre as variáveis densidades populacionais e acesso a instalações sanitárias.

Ainda concernente à correlação, foram realizados o cálculo da matriz de correlação entre o total de casos por milhão de habitantes e o índice de restrições aplicadas. Os resultados incluem o valor do teste t, os graus de liberdade, o valor de p, a hipótese alternativa, o intervalo de confiança de 95% e o coeficiente de correlação, revelando um valor extremamente baixo para o teste t e valor de p menor, indicando que a probabilidade de obter uma correlação forte entre o total de casos e as medidas de controlo e restrição é praticamente nula, tendo sido obtida uma correlação negativa muito forte entre o total de casos de COVID-19 e as medidas de controlo e restrição.

Contudo, da análise da correlação de dados brutos e dados agrupados foram obtidos valores com diferenças muito pequenas, o que não é substancial e, em termos práticos, ambas correlações indicam uma relação entre as variáveis. A pequena variação entre estas pode advir de algumas razões tais como arredondamentos de dados originais, falta de sensibilidade no agrupamento de dados e da própria variabilidade destes.

Por outro lado, os resultados sugerem que à medida que o número total de testes para a COVID-19 aumenta, os números totais de casos confirmados também tendem a aumentar, o que se justifica pelo facto de que, quanto mais testes são realizados, maior a probabilidade de identificar casos positivos.

Outras análises, como a de regressão evidenciaram que aproximadamente 93.63% da variabilidade casos da COVID-19 é explicada pela densidade populacional, o que indica que a densidade populacional é uma variável importante na explicação dos casos de COVID-19 e tem um impacto significativo na propagação dos casos. Assim, áreas com maiores densidades populacional tendem a ter mais casos de COVID-19.

O resultado da aplicação prática deste estudo é consistente com as teorias científicas desenvolvidas em temas similares, no que concerne a estudo sobre COVID-19, sendo que os resultados obtidos evidenciam que a implementação de medidas rigorosas pode reduzir a disseminação do vírus.

Comparando estudos que investigaram sobre o tema em contextos semelhantes desenvolvidos na área, observamos alinhamento dos resultados e concordâncias nas tendências e na influência de medidas de intervenção. Conclui-se haver consistência entre os resultados do presente trabalho e os estudos locais, o que reforça a validade e a aplicabilidade do estudo do tema em contextos epidemiológicos em Moçambique.

Recomenda-se a replicação do estudo para outras doenças, comparação com outras regiões e realização de estudos longitudinais.

REFERÊNCIAS BIBLIOGRÁFICAS

Ahlbom, A. (2005). *Biases Introduced by Merging Different Geographical Scales in Regionalized Health Data." International Journal of Health Geographics*, vol. 4, no. 4, pp. 1-9.

Araújo, M. V. (2021). *Métodos de Clustering em Aprendizado de Máquinas Não Supervisionado*. Niteroi - RJ, Brasil.

Barreto, E. F. (2021). *Ciência de dados aplicados a pandemia do Coronavírus no Brasil, uma análise socioeconómica*. Universidade Estadual Paulista (UNESP).

Barros, A.J.D; Lima Filho, E.C. (1994). *O modelo logístico aplicado a estudos epidemiológicos*.

Beltekian D, Gavrilov D, Giattino C, *et al*. (2020). *Data on COVID-19 (coronavirus). GitHub, Inc. Our World in Data* Web site. https://github.com/owid/covid-19-data/tree/ master/public/data.

Brito, S. B. P. *et al*. (2020). *Pandemia da COVID-19: O maior desafio do seculo XXI*. São Paulo.

Camargo, T. A; Villar, J. (2021). *Aplicativo Web para Análise e Previsão de Séries Temporais Epidemiológicas*.

Chatfield, C. (2004). *The Analysis of Time Series: An Introduction (6th ed.)*. Chapman and Hall.

Cordeiro, G. M e Dantas, R. A. (2000). *Avaliação de mercado de apartamentos de recife utilizando Modelos Lineares Generalizados*. Porlamar.

Cordeiro, G. M e Demétrio, C.G.B. (2008). *Modelos Lineares Generalizados e extensões*. Piracicaba.

Dias, G. N.; Carla, J., Pamplona, V., e Barbosa, E. D. (2020). *Análise matemática e estatística da doença COVID-19 e implicações em projeções futuras*. Amazônia.

El-Sadr, W. M. *et al.* (2022). *Comprehensive Approaches to COVID-19 in Africa Countries*. The Lancet, 399(10310), 2021-1023. Disponível em: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(22)00391-7/fulltext

Fagundes, J. A. (2021). *A linguagem R na análise de dados: Um estudo de caso dos transportes públicos do Rio de Janeiro durante a pandemia da COVID-19*. Rio de Janeiro.

Fagundes, J; Fagunde, V. e Oliveira, M. H. S. (2021). *A linguagem R na análise de dados: Um estudo de casos de transportes públicos do Rio de Janeiro durante a pandemia da COVID-19*. Rio de Janeiro.

Faustino, E. *et al*. (2020). *Análise da Incidência de COVID-19 em diferentes faixas etárias em Moçambique*. Saúde Pública, 25 (1), 45-52.

Freitas, J. R., Pereira, M.M.A; Santana, L.I.T; Silva, A.S.A & Filho, M.C. (2019). *Estudo comparativo de séries temporais para previsão dos números de casos semanais de dengue em alguns municípios de Pernambuco*. Pernambuco.

Fischer, M.M. e Getis, A. (2010). *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, DOI 10.1007/978-3-642-03647-7_24, © Springer-Verlag Berlin Heidelberg.

Gomes, E. C. (2015). *Conceitos e ferramentas da epidemiologia*. Recife: Ed. Universitária da UFPE.

Goncalves, R. *et al*. (2021). *Impactos Socioeconómicos da Pandemia de COVID-19 em Moçambique*. Economia e Sociedade, 30(2), e00213520. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-06182021000200515&tlng=pt

Guerra, S.; Oliveira, F; McDonell. (2018). *Ciência de dado em R- Introdução*.

Hair, J.F., Black, W.C., Babin, B.J., e Anderson, R.E. (2014). *Multivariate Data Analysis.* (7th ed.). Pearson.

Heidecher de Oliveira, M. (2022). *Análise comparativa entre três modelos matemáticos aplicados a COVID-19 no Município de São Paulo*. São Paulo.

Jonhson, R. A., Wichern D. W. (2002). *Applied Multivariate Statistical Methods*, Prentice Hall

Kac, G., Sichieri, R., e Gigante, D. P. (2007). *Epidemiologia nutricional*. SciELO-Editora FIOCRUZ.

Khan, K. *et al*. (2020). *Spread of a Novel Coronavirus in an Urban Slum, New Delhi, India, 2020*. Public Health, 183, 1-3. Disponível em: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7336557/

Konzen, E; They, N. H; e Ritter. (2019). *Introdução ao software estatístico R*. Rio Grande do Sul.

Kutner, M.H., Nachtsheim, C.J., Neter, J., e Li, W. (2004). *Applied Linear Statistical Models*. (5th ed.). McGraw-Hill.

Lakatos, E. M., e Marconi, M. A. (2003). *Fundamentos de Metodologia Científica*. São Paulo: Atlas, 5ª edição.

Leal, C., Morgado, L., Oliveira, T.A. (2023). *Mathematical and Statistical Modelling for Assessing COVID-19 Superspreader Contagion: Analysis of Geographical Heterogeneous Impacts from Public Event*s. Mathematics 11 (5), 1156 (19 pages).

Mabunda, J. *et al*. (2023). *Análise Espacial da Incidência da COVID-19 em Moçambique*. Cadernos de Saúde Pública, 39(3), e00213520. Disponível em: https://www.scielosp.org/article/csp/2023.v39n3/e00213520/

Maxwell, S. E., e Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparasion perspective (2nd ed.)*. Lawrence Erlbaum Associates.

Medri, W. (2011). *Análise exploratória de dados*. Londrina/Pr.

Ministério da Saúde de Moçambique. (2021). *Relatórios de Situação COVID-19*. Disponível em: https://www.misau.gov.mz/index.php/covid-19/situacao-epidemiologica

Monfardini, F. (2016). *Modelos Lineares Generalizados Bayesianos para dados Longitudinais*. São Paulo.

Montgomery, D.C. (2017). *Design and Analysis of Experiments*, Wiley, 9th Edition.

Muller, *et al*. (2021). Impact of Lockdown Measures on COVID-19 Cases and Death's in Europe. The Lancet, 396(10265), 1524-1534. Disponível em: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31917-2/fulltext

Nascimento de Jesus, M.F. (2015). *Estudo comparativo entre as funções de ligação LOGIT e PROBIT: Estimando parâmetros.* São Cristóvão.

Nelder, J. A., Wedderburn, W. M. (1972) *Generalized Linear Models. Wiley Journal of the Royal Statistical Society*. Series A (General), Vol. 135 pp.370-38

Nienov, O. H., *et al*. (2021). *Epidemiologia Aplicada Básica. Universidade Federal do Rio Grande do Sul*; Porto Alegre.

Nieto, J.R. (2018). *Epidemiology: Beyond the Basics. 4th edition. Jones & Bartlett Learning*.

Oliveira, E. F. (2022). *Análise de séries temporais para previsão de demanda no INSS*. Brasília – DF.

Oliveira, G.T. (2021). *Análise espacial dos casos da COVID-19 no estado do Rio de Janeiro*. Niterói, Rio de Janeiro.

Oliveira, T.A. (2004). *Estatística Aplicada*. Edições Universidade Aberta, n. 287

Organização Pan-Americana da Saúde. (2010). *Módulos de Princípios de Epidemiologia para o Controle de Enfermidades. Módulo 5: pesquisa epidemiológica de campo – aplicação ao estudo de surtos / Organização Pan-Americana da Saúde*; Ministério da Saúde. Brasília. 98 p.: il. 7 volumes. ISBN 978-85-7967-023-7.

Observatório Nacional de Saúde. (2021). *COVID-19 em Moçambique. Relatório do 1º ano*. Maputo.

Proença, C. S., e Schmidt, C. A. (2021). *Previsão estatística e comparação de dados da mortalidade por COVID-19*. Medianeira PR.

Ribeiro Júnior, P. J. (2005). *Curso sobre o programa computacional R*. Rio de Janeiro.

Roquim, F.V. (2014). *Modelos Lineares Generalizados: Estudo do método de Newton -Ramphon para estimação de parâmetros através de um modelo logístico*. Lavras.

Silva, A. *et al*. (2022). *Efeitos das Medidas de Intervenção na Propagação da COVID-19 em Moçambique*. Revista de Epidemiologia, 20(2), 241-248. Disponível em: https://www.scielo.br/scielo.php?pid=S1415-790X2022000200241

Silva, R. J., Matto, J., e Silva, K. B. (2020). *Análise espacial sobre a dispersão da covid-19 no Estado da Bahia / Spatial analysis on the dispersion of covid-19 in the state of Bahia*. Bahia.

Silva, B. F; Bortoluzzi, M. A., e Dinis, J. (2009). *Curso de estatística básica: Introdução ao Software R*. Santa Maria.

Silva, Y. L. (2020). *Clusterização automática em seleção de materiais e processos de fabricação utilizando pso-particle swarm optimization*. João Pessoa.

Smith, J., *et al*. (2020). *Socioeconomic and Geographic Disparities in Coronavirus Disease 2019 Reported Incidence, Knowledge, and Behaviour in United States*. Clinical Infectious Disease, 72(6), e113-e121. Disponível em: https://academic.oup.com/cid/article/72/6/e113/5901132

Snijders, T. A., e Bosker, R. J. (2012). *Multinivel analysis: An introduction to basic and advanced multinivel modelling (2nd ed.)*. Sage Publications.

Souza, B. L. (2022). *Modelos de Series Temporais para a Predição de Incidência de Tuberculose no Brasil*. Niteroi - RJ, Brasil.

União Africana. (2020). *Boletim de Informação #31: Pandemia da Doença do Coronavírus 2019 (COVID-19)*.

Teodoro, M.F., Oliveira, T.A., Arune, F. (2024). *COVID-19 Infection: A Mozambican Case Study. Biometrics & Biostatistics International Journal*, Volume 13, Issue 1, Pages 7-14, MedCrave.

Teodoro, M.F., Oliveira, T.A., Taero, E. (2023). *COVID-19 infection and risk analysis: a short introduction*. Biometrics & Biostatistics International Journal, Volume 12, Issue 4, Pages 121-125, MedCrave.

Theodoro, M. M. (2022). *Modelação Fracionária da Dinâmica da COVID-19*. Botucatu.

Turkaman, M. A. A., e Silva, G. L. (2000) *Modelos Lineares Generalizados – da teoria à prática*. Lisboa, SPE.

Wangping, J., Ke, H., Yang, S., Wenzhe, C., Shengshu, W., Shanshan, Y., . . . Liu Miao 1, H. Y. (2020). *Extended SIR prediction of the epidemics trend of COVID-19 in Italy and compared with Hunan, China*. Beijing.

White, Ian R., et al. (2009). *Multiple Imputation for Missing Data in Epidemiological and Clinical Research: Potential and Pitfalls*. British Medical Journal, vol. 338, pp. b2393.

World Health Organization. (2019). *Epidemiological data analysis for the early warning alert and response network (EWARN) in humanitarian emergencies: a quick reference handbook / World Health Organization*. Regional Office for the Eastern Mediterranean.

Zavala, V. G. M. (2021). *Perceção e gestão de risco da contaminação da COVID-19: Caso de consumidores de bebidas alcoólicas na cidade de Maputo*. Maputo.