UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Mariana Schulz Bellacosa

Técnicas Básicas e Exploratórias para Análise de Dados de Covid-19

São José do Rio Preto

2022

Mariana Schulz Bellacosa

Técnicas Básicas e Exploratórias para Análise de Dados de Covid-19

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Matemática, junto ao Programa de Pós-Graduação em Matemática em Rede Nacional – PROFMAT, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Campus São José do Rio Preto.
Orientador: Prof. Dr. José Gilberto Spasiani Rinaldi

São José do Rio Preto

2022

Bellacosa, Mariana Schulz
Técnicas Básicas e Exploratórias para Análise de Dados da Covid-19 / Mariana Schulz Bellacosa. - São José do Rio Preto: [s.n.], 2022.
47 f. : il. ; 30 cm.

Orientador: José Gilberto Spasiani Rinaldi
Dissertação (mestrado) - Universidade Estadual Paulista “Júlio de Mesquita Filho”,, Instituto de Biociências, Letras e Ciências Exatas

1. Aplicações de Estatística. 2. Métodos Introdutórios. 3 Médias Móveis. , José Gilberto Spasiani. III. Universidade Estadual Paulista “Júlio de Mesquita Filho”,, Instituto de Biociências, Letras e Ciências Exatas. IV. Título.

CDU -

Mariana Schulz Bellacosa

Técnicas Básicas e Exploratórias para Análise de Dados de Covid-19

Dissertação apresentada como parte dos requisitos para obtenção do título de Mestre em Matemática, junto ao Programa de Pós-Graduação em Matemática em Rede Nacional – PROFMAT, do Instituto de Biociências, Letras e Ciências Exatas da Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus São José do Rio Preto.

Orientador: Prof. Dr. José Gilberto Spasiani Rinaldi

Comissão Examinadora

Prof. Dr. José Gilberto Spasiani Rinaldi
UNESP – Presidente Prudente
Orientador

Prof. Dr. Suetônio de Almeida Meira
UNESP – Presidente Prudente

Prof. Dra. Teresa Cristina Martins Dias
UFSCar – São Carlos

São José do Rio Preto
2022

RESUMO

A Estatística é hoje uma área bastante demandada por vários fatores, seja pela oportunidade de empregos ou pela utilização cada vez mais frequente de medidas estatísticas para analisar ocorrências nos mais diversos setores, de negócios à diverssão. Quando se pensa em Estatística, geralmente, se vislumbram fórmulas e cálculos complexos, que poucos indivíduos poderão realizar.
Neste trabalho são apresentadas técnicas introdutórias de Estatística que, apesar de básicas, podem ser entendidas e aplicadas de forma a se obter bons resultados a respeito de fenômanos atuais.
A Covid-19 apareceu no Brasil no início de 2020 e vários estudiosos tentaram preescrever complexos modelos estatísticos que não apresentaram bons resultados. Neste trabalho, técnicas estatísticas introdutórias são aplicadas à dados de Covid-19 apresentando resultados interessantes em relação aos fenômenos de transmissão e óbitos ocorridos.

Palavras–chave: . Estatística Aplicada. Métodos Introdutórios. Médias Móveis. Probabilidade.

ABSTRACT

Statistics is today an area in high demand due to several factors, whether due to the opportunity for jobs or the increasingly frequent use of statistical measures to analyze occurrences in the most diverse sectors, from business to entertainment. When one thinks about Statistics, one usually envisions complex formulas and calculations, which few individuals will be able to perform.

This work presents introductory statistical techniques that, despite being basic, can be understood and applied in order to obtain good results regarding current phenomena.

Covid-19 appeared in Brazil in early 2020 and several scholars tried to pre-write complex statistical models that did not show good results. In this work, introductory statistical techniques are applied to Covid-19 data presenting interesting results in relation to the phenomena of transmission and deaths that occurred.

Key words: . Applied Statistics. Introductory Methods. Moving Average. Probability.

“Aos meus pais, Mônica e Moacir, aos meus irmãos, Marcella e Murilo, e ao meu namorado, Gustavo, pelo apoio e paciência."

AGRADECIMENTOS

Agradeço, primeiramente, à Deus que me fortaleceu nos momentos difíceis e possibilitou que eu chegasse até aqui me concedendo saúde e força para concluir essa etapa.
Agradeço aos professores do PROFMAT por todo conhecimento que foi transmitido com paciência e carinho durante todas as disciplinas, e em especial ao meu orientador Dr. José Gilberto Spasiani Rinaldi pelos seus conselhos e por sempre estar disposto em ajudar e orientar.
Ao meu namorado, Gustavo, por seu amor, paciência e companheirismo durante os momentos de estudo. E a minha família, pelo apoio e incentivo que foram essenciais e determinantes para essa conquista.
E por fim aos meus colegas de PROFMAT, dentro os quais em especial aos meus amigos Lucas e Izabella pela troca de aprendizagem e lealdade durante os anos de estudo.

SUMÁRIO

Capítulo 1

Introdução

Dados estatísticos são considerados atualamente como sendo elementos indispensáveis a qualquer problema do mundo real que se queira analisar, as empresas tentam arregimentar dados de seus consumidores das formas mais distintas possíveis, ofertam descontos, prêmios, bônus, estre outros.

Informações são preciosas, contudo, é preciso e necessário que sejam analisadas da forma mais adequada, mais relevante e objetiva no sentido de fornecer subsídios à empresa, organização, instituição, enfim, a quem de direito solicitou que as mesmas fossem objeto de análise.

Há casos em que excesso de informações não bem organizadas foram obtidas, geralmente empresas que, na busca desenfreada por dados, coletam várias informações relevantes e não relevantes que vão representar um desafio em “filtrar” o que realmente será de utilidade. Contudo, ainda mais comum, é não realizar uma análise estatística adequada, utilizar ferramentas não apropriadas, não testar hipóteses estatísticas, enfim, não ter a habilidade suficiente para ofertar o que se espera por parte do solicitante da análise.

Atualmente, a área de estatística tem sofrido muito com “especialistas” (geralmente autodenominados assim) que se arriscam neste campo, não raro, realizando previsões que não serão verificadas no cenário real. Isto ocorreu muito em dados e previsões referentes à Covid-19, surgiram vários sujeitos com as mais amplas sugestões, teorias próprias e verificações que, com o simples olhar científico, não se sustentavam.

Devido à complexidade de causas e efeitos da Covid-19, a comunidade científica se debruçou em modelos complexos que, teoricamente, pudessem ser capazes de descrever o fenômeno e suas consequências em termos de contração da doença e dos óbitos subsequentes. Contudo, essa doença, como se sabe hoje, é muito complexa, produz resultados diferentes em cada pessoa que a contrai e depende muito também, do comportamento humano em se expor ao risco em diferentes níveis de intensidade. Todos esses fatores levaram os modelos a não conseguirem descrever satisfatoriamente o fenômeno em termos de previsão do número e casos e consequentes óbitos (Perlman e Yechiali, 2021),

várias previsões não se verificaram e houve então a utilização e valorização de ferramentas básicas de estatística, comumente denomonadas de análises exploratórias de dados.

É comum na área de estatística que se recomende, em qualquer fenômeno novo, que as estatísticas básicas sejam realizadas, são com elas que o pesquisador obterá intimidade com os dados e poderá entender boa parte do funcionamento de eventos e de parte do fenômeno, terá mais facilidade em identificar ferramentas mais adequadas para utilização e, possivelmente, obterá resultados mais precisos e informativos.

Neste trabalho são utilizadas algumas técnicas de estatístiaca básica e probabilidade introdutória a dados reais de Covid-19. Estas ferramentas são acessíveis ao ensino médio, não apresentam cálculos e/ou ferramentas matemáticas complexas, são bastante intuitivas e, algumas delas, como a média móvel, é bastante abordada na mídia.

Os exemplos finais podem ser motivadores para que essas técnicas possam ser estudadas e aplicadas à dados de outras situações práticas, difundindo a área de estatística e suas aplicações.

2. Introdução à Estatística

Neste capítulo introdutório, daremos base suficiente para que se entenda os conceitos (Bussab e Morettin, 2017) que serão abordados durante o decorrer desta pesquisa. Começaremos o capítulo dando uma ideia e mostrando alguns conceitos de probabilidade para o leitor, com isso partiremos para os conceitos de Estatística Descritiva e Inferencial Básica, procederá para Medidas Estatísticas.

2.1 Estatística Descritiva

2.1.1 Conceitos Iniciais

A *estatística descritiva*, pode ser descrita como um dos ramos da estatística, que tem como principal objetivo, fazer a organização e sumarizar determinado conjunto de dados que se esteja trabalhando. Na estatística inferencial básica trabalhamos interessados em utilizar apenas os dados de determinada amostra com o intuito de obter conclusões a respeito de um conjunto maior, conjunto este que a priori não temos acesso.

Descreveremos a seguir algumas fases da aplicação geral de estatística em relação aos dados.

Coleta:

Definição: Ato de recolher, arrecadar ou coletar os dados que são de interesse para o trabalho a ser realizado;

Dentro desta etapa, temos as seguintes divisões de coleta:

- *Contínua:*

  Esta trata de registros contínuos como nascimentos, casamentos, óbitos; ou seja, quando os eventos que acontecem durante determinado estudo, são registrados à medida que ocorrem;

- *Periódica:*

  Esta é realizada em momentos de interesse, um bom exemplo deste tipo de coleta é o censo demográfico, que acontece de ciclo em ciclo;

- *Ocasional:*

  Este tipo, são realizações esporádicas, como uma pesquisa de mercado, ou seja, que não se tem a preocupação de continuidade ou periodicidade.

Análise:

Nesta etapa, que sucede a fase de coleta devemos realizar uma análise em relação aos valores das variáveis de interesse, uma vez que estes valores estão sujeitos ser valores não realísticos, frutos de erros, como digitação ou observação.

Sendo assim, eles devem ser avaliados e retificados, ou mesmo removidos da coleta para que não comprometam os resultados finais.

Dentro desta etapa, é importante ter os valores da média e o desvio padrão, não que o restante não tenha seus valores de importância, mas a média por exemplo se faz útil para determinar se um grupo de dados possui ou não uma tendência geral, com um baixo custo de cálculos, uma vez que para se obter basta somar os valores total de números de uma lista e dividir pela quantidade de números dessa lista; o desvio padrão por sua vez tem seu papel valioso por representar a “distância” de um conjunto de dados em relação a sua média. Se o sigma (letra utilizada para representar o desvio padrão) for alto indica que os dados deste conjunto estão muito dispersos, por outro lado, quanto menor for o valor de sigma, indica que os dados estão mais alinhados com a média do conjunto. Ou seja, com estes dois valores em mãos, conseguimos com rapidez descobrir o quão disperso estão estes dados.

Para que tenhamos uma análise mais completa faremos uso da tecnologia para obtermos cálculos mais rápidos e precisos, por exemplo ao aplicar uma regressão, ação essa que nos faz compreender melhor quais são as relações entre as variáveis, uma vez que essas são traçadas normalmente em um diagrama de dispersão.

Além disso, o uso de softwares que possuem finalidades específicas para realizar uma minuciosa análise de dados, como por exemplo o Minitab, faz com que todas as operações descritas acima possam ser calculadas no mesmo.

Levantamento:

Esta é a fase de processamento dos dados coletados, isso pode ser feito de forma manual ou eletrônica, os dados são agrupados, condensados e tabulados para que facilite o cálculo de medidas e coeficientes, cuja principal finalidade é descrever o fenômeno.

Identificação:

Aqui o intuito é o resumo das informações contidas nestes dados, de forma que evidencie seus aspectos mais importantes e para isso utilizamos ferramentas como tabelas e diferentes tipos de gráficos.

Estudo dos resultados:

Nesta etapa final são aplicadas diversas técnicas estatísticas cujo principal objetivo é inferir sobre os resultados, para isso podemos calcular a média e o desvio padrão, aplicar uma regressão e realizar uma análise estatística.

2.1.2 Variáveis

Quando se trata de variáveis, estamos nos referindo a um conjunto cujos elementos são todos os possíveis resultados. Estas são classificadas em qualitativas que podem ser nominal ou ordinal e quantitativas que podem ser discreta ou contínua. As qualitativas qualificam os dados, como sexo, estado civil e etc, em contrapartida as quantitativas, quantificam os dados, como idade e peso; estas podem ser discretas como número de filhos, número de óbitos e etc. ou contínua como peso, altura, tempo, entre outras. Podemos apresentar o seguinte modelo (Figura 2.1) para esboçar essa separação.

Além disso, existe ainda a possibilidade de discretizar essa variável, que se trata do processo de transformar variáveis contínuas em discretas, por exemplo se considerarmos a classificação por peso (70 kilos, por exemplo). O seguinte diagrama pode exemplificar suas divisões:

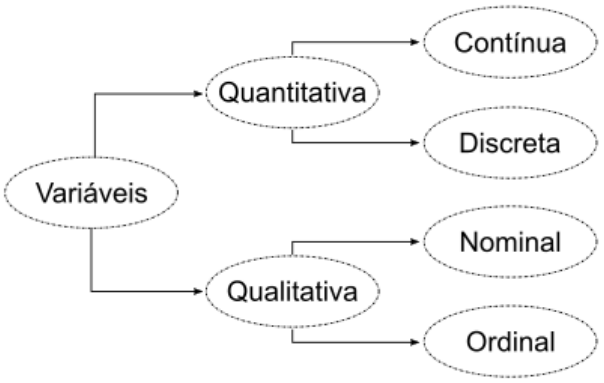


Figura 2.1: Variáveis

2.1.3 População e amostra

Chamamos de população estatística um conjunto de entes (itens, pessoas), finito ou não, portadores de no mínimo uma característica em comum. Por exemplo o censo é uma coleção de dados relativos a todos os elementos de uma população.

Por outro lado a amostra é um subconjunto finito do conjunto população, ou seja uma parte da população. A inferência estatística tem como objetivo estudar a população seguindo evidências decorrentes de uma avaliação minuciosa da amostra.

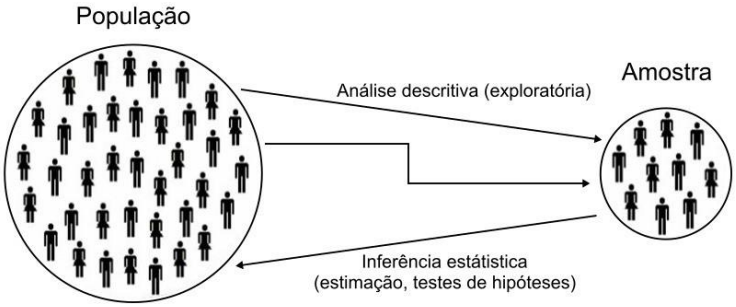


Figura 2.2: População e Amostra

2.1.4 Amostragem

A amostragem é uma técnica de coleta de amostras que visa garantir a mesma chance (ou representatividade) de cada ente da população pertencer à amostra. Dentre todos os tipos de amostragem que existe, podemos citar 3 principais, que são:

- Amostragem aleatória simples: Este podemos dizer que é o processo mais simples de amostragem, e por isso é também frequentemente utilizada, ela corresponde a uma amostra de elementos retirados ao acaso de determinada população, esta consiste em tomar uma população de tamanho n, enumerar os entes de 1 a n e na sequência, ir sorteando os elementos que pertencerão à amostra.
- Amostragem estratificada proporcional: Aqui consideramos que em algumas ocasiões a população se divide em subpopulações, denominadas como estratos. Estes estratos precisam estar representados de forma equitativa (segundo sua ”importância”) logo, deve-se proporcionalizar a amostra segundo o tamanho dos estratos da população; um exemplo simples seria a divisão de uma sala de alunos entre homens e mulheres.
- Amostragem sistemática: A amostragem sistemática é bastante utilizada quando os elementos da população se encontram ordenados. Como por exemplo a numeração de casas de uma rua, os itens em uma linha de produção, lista de alunos, entre outros.

2.2 Apresentação dos Dados

Nesta seção apresentaremos diferentes modos de dispor os dados de uma

determinada pesquisa, esta etapa é de grande importância uma vez que é a partir daí que causaremos uma impressão mais rápida do que está sendo estudado naquele trabalho.

2.2.1 Tabela

Começaremos pela tabela que se trata de uma representação matricial onde os dados ficam dispostos em linhas e colunas, com título e cabeçalho, para exemplificar observe o modelo abaixo: onde ”Temperatura no dia 27/11/2021 ”´e o título e as células com ”Cidade, Temperatura Máxima e Temperatura Mínima” são o que chamamos de cabeçalho da tabela.

Temperatura no dia 27/11/2021

Cidade	Temperatura Máxima	Temperatura Mínima
São Paulo	30°	17°
Presidente Prudente	34°	21°
Mirante do Paranapanema	34°	20°

Figura 2.3: Exemplo de Tabela

2.2.2 Distribuição de Frequência

Para falar de Distribuição de Frequência, vamos pensar na situação onde as variáveis contínuas assumem valores extremamente diferentes, em um curto intervalo, para que esse tipo de problema possa ser contornado, é usual fazer uso das tabelas de distribuição de frequências, estas são construídas de modo que facilite a interpretação, para isso cada classe é constituída por um intervalo de valores da variável. Isso é usual também quando se trata de variáveis discretas que assumem muitos valores diferentes. Segue um exemplo.

Quantidade de sapatos concertados na semana do dia 22/11

Número do Sapato	*Quantidade de Sapatos*
30 ⊢ 33	10
33 ⊢ 36	5
36 ⊢ 39	15
39 ⊢ 42	3
Total:	33

Figura 2.4: Exemplo de Tabela

2.2.3 Gráfico

Nesta forma de apresentação, é buscado uma representação visual de dados e

informações, neles é possível ser colocado tudo que pode ser medido ou quantificado, fazendo assim uma relação entre quantidades e qualidades, dessa forma podemos chegar a uma conclusão de forma mais rápida e simplificada apenas analisando o que está sendo tratado no gráfico em questão. Existem diversos tipos de gráficos, entre eles apresentaremos neste trabalho gráficos em barra, coluna, setores e linhas.

Abaixo segue um exemplo de cada um deles, considerando o problema de se apresentar os dados de consumo de energia mensal no decorrer de um ano.

Gráfico de Barras: Este expressa seus resultados em barras no formato retangular posicionadas na horizontal ou vertical, o comprimento de cada barra varia de acordo com os dados estão sendo representados, segue o exemplo abaixo:

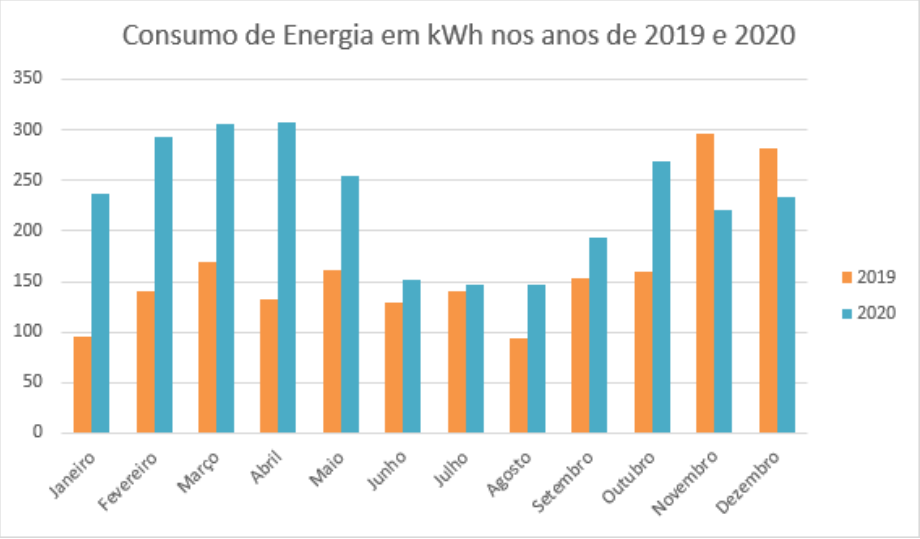


Figura 2.5: Exemplo de Gráfico de Barras

Gráfico de Setores: Este, também conhecido como gráfico de pizza, divide o ângulo de 360° no número de categorias que se deseja apresentar, e aquela que tiver uma quantidade maior é atribuído um setor com um ângulo proporcional a ela, lembrando que neste deve se ter cores e legendas para cada uma delas, segue um exemplo:

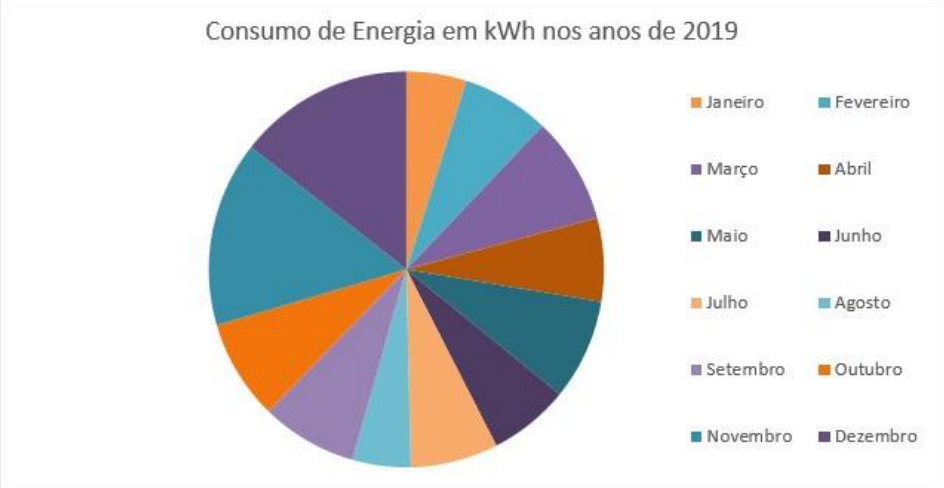


Figura 2.6: Exemplo de Gráfico de Setores

Gráfico de Linhas: Para exibir seus resultados, este tipo de gráfico utiliza pontos distribuídos em uma determinada malha e apresenta uma linha ligando estes, este será frequentemente utilizado neste trabalho uma vez que o mesmo facilita a visualização de crescimento dos casos e mortes causadas pelo COVID-19.

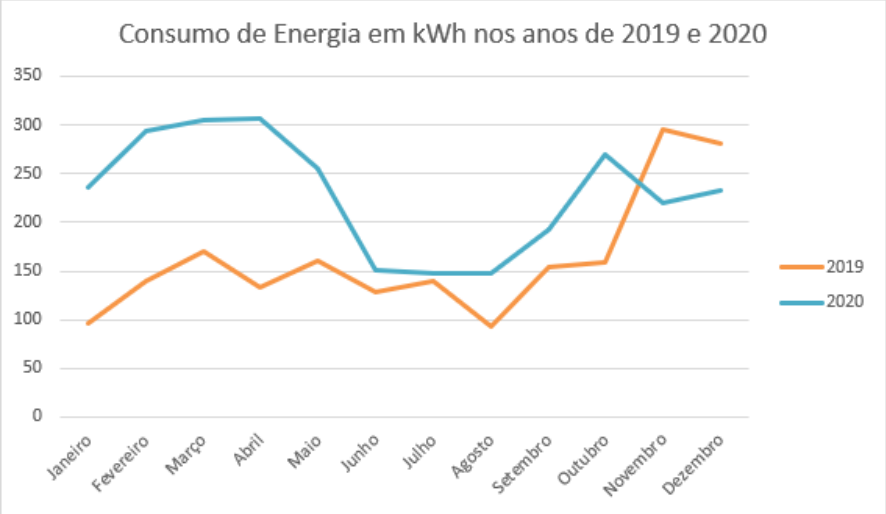


Figura 2.7: Exemplo de Gráfico de Linhas

2.3 Medidas Estatísticas

Nesta seção definiremos algumas medidas estatísticas que podem ser calculadas com o objetivo de obter informações a respeito do comportamento dos dados analisados.

2.3.1 Medidas de posição:

Estas são medidas que fornecem informações a respeito de sua posição em relação aos dados, temos a média, moda e mediana.

Média: Esta é calculada somando-se todos os valores do conjunto de dados e dividindo pelo número de elementos deste conjunto, seu ponto negativo é a sua sensibilidade aos valores da amostra, por isso é mais utilizada em situações em que os dados não tenham grandes discrepâncias na sua distribuição. Esta é dada por

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

onde n é o número de elementos do conjunto. O desvio em relação à média pode ser obtido como

$$d_i = x_i - \bar{x}$$

para $i = 1, 2, \cdots, n$.

Além disso temos as seguinte propriedades:

- $\sum_{i=1}^{n} d_i = 0$;
- $y_i = x_i \pm c \Rightarrow \bar{y} = \bar{x} \pm c, \quad \forall c \in \mathbb{R}$;
- $y_i = x_i \cdot c \Rightarrow \bar{y} = \bar{x} \cdot c, \qquad \forall c \in \mathbb{R}$;

Exemplo: Para sabermos a média final de um aluno na escola basta fazermos a média aritmética entre as médias de cada bimestre. Sabemos que o aluno Lucas ficou com 7 pontos no 1 bimestre, 8 no 2 bimestre, 5 no 3 bimestre e 4 no último bimestre. A média final da escola é 6 entao vamos verificar se Lucas passou de ano, pela definição dada acima sua média será

$$M = \frac{7 + 8 + 5 + 4}{4} = \frac{24}{4} = 6$$

Então concluímos que Lucas passou de ano e agora poderá curtir suas férias.

Moda: Está é denotada por (Mo) e representa o valor que mais aparece no conjunto de dados, ou seja para definir quem é a moda basta observar a frequência em que os valores aparecem e aquele que mais se repetir será a moda. Podem existir conjuntos de dados que não possuem moda, ou seja, nenhum elemento desse conjunto se repete e assim dizemos que ele é amodal. Também pode ocorrer de determinado conjunto possuir dois valores que repetem a mesma quantidade de vezes, então dizemos que ele é bimodal.

Exemplo: Em uma sapataria durante um dia foram vendidos os seguintes números de sapato: $34, 39, 36, 35, 37, 40, 36, 38, 36, 38$ e 41. Podemos observar que a numeração de sapato mais vendida foi o 36. Logo seguindo a definição, 36 é a moda desse conjunto de dados.

Mediana: A mediana denotada por (Md) é o valor que se encontra no centro conjunto de dados. Para encontrar a mediana colocamos os valores em ordem crescente e escolhemos o valor central. Quando o número de elementos é par, somamos os dois valores centrais e dividimos por dois. Além disso existem também outras medidas separatrizes além da mediana, como os quartis, decis, percentis e etc.

Exemplo: Em uma escola, o professor de educação física anotou a altura de um grupo de alunos. Considerando que os valores medidos foram: $1,54m$; $1,67m$; $1,50m$; $1,65m$; $1,75m$; $1,69m$; $1,60m$; $1,55m$ e $1,78m$. Agora para calcularmos a altura do aluno central primeiro vamos colocar essas alturas em ordem crescente, então temos $1,50m$; $1,54m$; $1,55m$; $1,60m$; $1,65m$; $1,67m$; $1,69m$; $1,75m$ e $1,78m$, como temos 9 alunos temos que o aluno que estará no meio é o quinto aluno portanto a mediana será $1,65m$

Média Móvel: A média móvel (Morettin e Toloi, 2018) é um importante indicador da categoria dos Rastreadores de Tendência, pois elas suavizam os dados para formar um indicador de tendência sequencial. Ela calcula o valor médio de preço, volume ou mesmo de outro indicador em determinado período. Em um gráfico ela suaviza os movimentos da série de dados, ou seja, retira os ruídos, representados pelas oscilações mais fortes. Com isso torna-se mais simples entender o comportamento dos dados que estão sendo entregues, se eles estão seguindo uma tendência ou não.

- **Média Móvel Simples:** Está é responsável por minimizar as oscilações mais fortes da sequência do banco de dados, e isso se deve ao fato do seu valor mais atual não depender apenas de um dado mas sim dos n últimos dados, ou seja para calcular a média móvel simples de determinado período basta obter a soma de um conjunto de n valores em sequência e dividir pela quantidade de elementos somados, que é o número n, desta forma, obtemos a seguinte fórmula.

$$MMS_n = \frac{(V_1 + V_2 + \cdots + V_n)}{n}$$

onde $n =$ número de períodos e $V =$ valor

Como o próprio nome indica, uma média móvel é uma média que se move ao longo de um período. Os dados antigos são retirados, a medida que dados mais recentes se tornam disponíveis. Isto faz com que a média se mova ao longo do tempo revelando a tendência que o fenômeno estudado está apresentando.

2.3.2 Medidas de Dispersão ou de Variabilidade:

Nesta seção definiremos a medidas de dispersão que serão utilizadas no decorrer deste trabalho, que são variância, desvio padrão e coeficiente de variação.

Variância: A variância é baseada nos desvios em relação à média e é dada por

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} {d_i}^2$$

é importante observar que nesta formula, a medida da variância está ao quadrado.

Observação: Note que a variância é dada ”elevada ao quadrado”, e isso se se faz necessário pelo fato de estarmos tratando da distância de determinado dado x_i até a sua média $\bar{x}$, caso não houvesse a potência 2 presente na formula, poderíamos ter tanto desvios

negativos quanto positivos e com isso quando somados resultaria em zero. Para voltar ao valor real basta extrai a raiz quadrada de ambos os lados assim voltando à unidade original.

Desvio Padrão: Este por sua vez é obtido raiz quadrada da variância, ou seja

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$S = \sqrt{\frac{1}{n}\sum_{i=1}^{n}{d_i}^2}$$

Podemos citar ainda algumas propriedades, que são:

- $y_i = x_i \pm c \Rightarrow S_y = S_x, \quad \forall c \in \mathbb{R}$;
- $y_i = x_i \cdot c \Rightarrow S_y = S_x \cdot c, \quad \forall c \in \mathbb{R}$;

Coeficiente de Variação: Este é utilizado para avaliar o desvio em relação a média, e é dado por

$$CV = \frac{S}{\bar{x}} \cdot 100\%$$

onde S é o desvio padrão e $\bar{x}$ é a média dos dados.

2.3.3 Medidas de Assimetria e Curtose:

Aqui trataremos do formato das curvas de distribuição de dados, estas nos ajudar à compreender conceitos futuros.

Assimetria: Primeiramente vamos tratar das distribuições que são simétricas, um bom exemplo é a distribuição normal (vide Figura 8). isto ocorre quando a média e a moda coincidem, ou seja $\bar{x} = Mo = Md$.

Alguns tipos de assimetria são baseadas na média e na moda e para tanto, calcula-se $\bar{x} - Mo$ e interpreta-se como (vide Figura 9)

- Se $\bar{x} - Mo = 0$ então é simétrica;
- Se $\bar{x} - Mo < 0$ então é assimétrica à esquerda;
- Se $\bar{x} - Mo > 0$ então é assimétrica à direita;

Chamamos de coeficiente de assimetria o valor utilizado para medir o nível de assimetria, que é dado por

$$AS = \frac{3(\bar{x} - M_d)}{S}$$

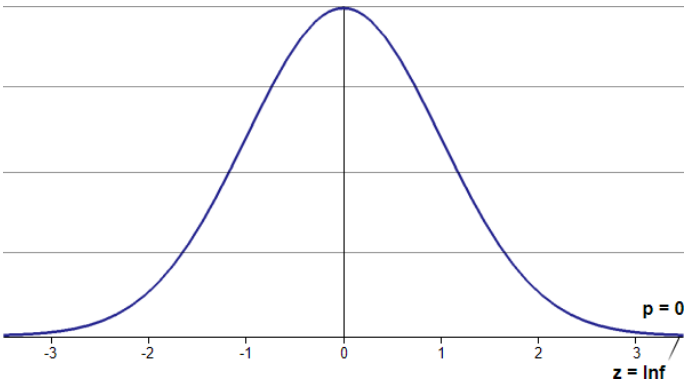
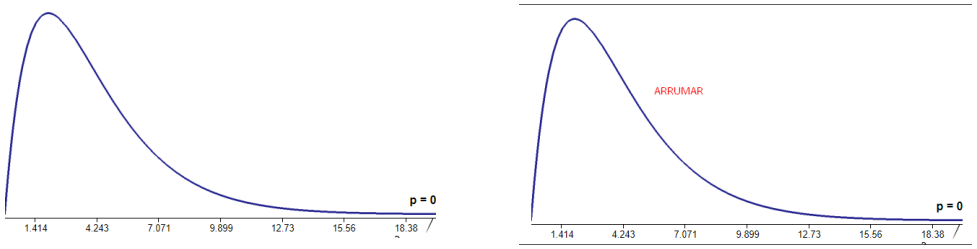


Figura 2.8: Normal



(a) $M_o < M_d < \bar{x}$ (b) $\bar{x} < M_d < M_o$

Figura 2.9: Exemplos de curvas assimétricas

para $0{,}15 < |AS| < 1$ dizemos que a assimetria é considerada moderada, e caso $|AS| > 1$ ela é considerada forte.

A Curtose é uma medida utilizada para avaliar o tipo e o grau de achatamento de uma curva de distribuição. Essa comparação (Figura 10) é dada por uma curva normal e o seus principais tipos são a Mesocúrtica, Platicúrtica e Leptocúrtica.

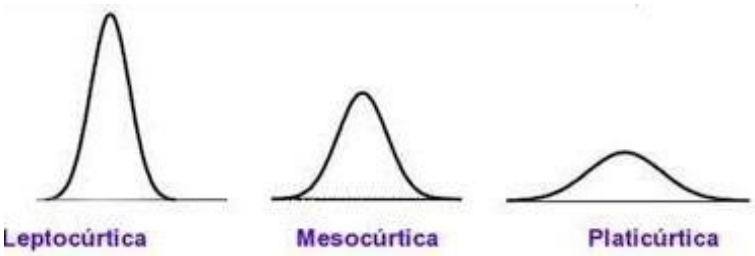


Figura 2.10: Curtose

Assim como na assimetria, a curtose também possui coeficiente, medida essa que quantifica o grau de curtose, e ele é dado por

$$C = \frac{Q_3 - q_1}{2(P_{90} - P_{10})}$$

Caso $C = 0{,}263$ dizemos que a curva é Mesocúrtica, se $C < 0{,}263$ é Leptocúrtica e se $C > 0{,}263$ é chamada de Platicúrtica.

2.3.4 Indicadores (Coeficientes, Taxas e Índices)

O primeiro passo ao se realizar qualquer tipo de estatística, o primeiro passo é a obtenção dos dados e para tal existem duas formas, que são de forma direta (através do contato direto com os dados) ou de forma indireta (coletando através de secundários).

Podemos ainda classifica-los em absolutos ou relativos. No nicho dos absolutos, colocamos aqueles que são obtidos de forma pura, que não possuem qualquer tipo de manipulação seja ela em relação a medidas ou contagens. Por outro lado temos os relativos que são aqueles que são obtidos através dos dados absolutos, realizando-se alguns tipos de transformações, como por exemplo percentagens, gráficos e razoes por exemplo.

Coeficientes:

Quando temos uma relação entre uma parte com o todo, chamamos de coeficiente, ou seja, o mesmo tipo de variável compartilha de variável do mesmo tipo de ente. Como exemplo, seja o coeficiente de mortalidade, esse relaciona o total de óbitos de um determinado local pela população com risco de morrer da seguinte forma:

$$coef.mortalidade = \frac{número\ de\ mortes}{número\ da\ população\ exposta\ ao\ risco}$$

Há outros coeficientes, como de natalidade, de evasão, entre outros.

Taxas:

Podemos dizer de forma simples e direta que taxas são meramente os coeficientes multiplicados por uma potência de 10 (10 ou 100 geralmente). Esta operação faz com que a interpretação da magnitude de algumas taxas seja compreensível de forma mais didática. Como exemplo, imagine a taxa de mortalidade pelo vírus COVID-19 em uma cidade de aproximadamente 300.000 habitantes. Para esta situação, por exemplo, talvez fosse utilizada a taxa de *x* homicídios por 150.000 habitantes.

$$taxa\ de\ mortalidade = \frac{número\ de\ óbitos}{tamanho\ da\ população}$$

Isto fornece um dado comparativo às cidades de tamanhos completamente diferentes à estudada, o risco é inerente à taxa e não ao tamanho da população. Outros exemplos de taxas são a taxa de natalidade, taxa de homicídio e taxa de evasão escolar.

Porcentagens:

Estas por sua vez, são destinadas especialmente para avaliar grandezas quanto ao seu aumento ou redução. Como exemplo, considere uma cidade que inicialmente possuía cerca de 250.000 habitantes em 2015 e 275.000 em 2022. Qual seria a porcentagem de crescimento desta cidade?

$$\frac{valor\ final - valor\ inicial}{valor\ inicial} = \frac{275.000 - 250.000}{250.000} = 0{,}1 \text{ ou } 10\%.$$

Esta porcentagem pode ser realizada por mais de uma cidade podendo, inclusive, compará-las em termos de crescimento.

Índices

Se tratando de índices, esperamos ver razões expressando as quantidades de variáveis distintas, ou seja, não há compartilhamento de variável do mesmo tipo de ente. Por vezes, as variáveis envolvidas podem não possuir nenhuma relação.

Como alguns exemplos, podemos citar:

$$Renda\ per\ capita = \frac{renda\ total\ da\ população}{população\ total}$$

$$Densidade\ populacional = \frac{população\ total}{superfície\ total}$$

$$QI = \frac{Idade\ mental}{Idade\ cronológica}$$

Os intervalos que eles se encontram varia entre 0 e 1. Para isso, são realizadas transformações, tais como considerar o pior e melhor valor do indicador estudado de forma que

$$Índice = \frac{valor\ observado\ do\ indicador - pior\ valor}{melhor\ valor - pior\ valor}.$$

Vale ressaltar que alguns indicadores podem ser mais complexos de se obter pois podem englobar outros indicadores em sua composição.

IDH (Índice de Desenvolvimento Humano)

Este é geralmente utilizado para mensurar o estado de um país considerando a oferta de saúde, renda e conhecimento.

Ele utiliza os seguintes indicadores elementares:

Esperança de vida ao nascer: $I_{esp} = \frac{x-25}{60}$;

Taxa de alfabetização de adultos (15 anos ou mais); $I_{alf} = \frac{x}{100}$;

Taxa bruta de escolaridade (de 7 a 22 anos; matriculados do ensino fundamental ao superior): $I_{esc} = \frac{x}{100}$.

Também, utiliza os seguintes indicadores:

$I_{con} = \frac{2}{3} I_{alf} + \frac{1}{3} I_{esc}$ (índice de conhecimento);

$I_{pib} = \frac{logx - log100}{log40000 - log100}$ (índice do PIB per capita).

Com estas informações, calcula-se o IDH:

$IDH = \frac{1}{3} I_{esp} + \frac{1}{3} I_{con} + \frac{1}{3} I_{pib}$.

Qualquer entidade (ou pessoas) pode propor um índice, é necessário que este seja coerente ao propósito considerado (tem que ser útil), relativamente fácil de ser calculado e interpretado.

2.4 Medida de Correlação e Regressão Linear Simples

A medida de correlação linear é utilizada para avaliar o grau de relação que possuem duas variaveis. Inicialmente, para verificar um possível grau de associação, pode-se realizar um diagrama de dispersão.

Coeficiente de correlação linear: O coeficiente de correlação linear mede a intensidade da relação entre as duas variáveis X e Y, por exemplo. Esta medida varia entre -1 e 1 e é dada por

$$r = corr(X,Y) = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{[n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2][n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2]}}$$

onde n é o número de observações. Esta medida é tal que se $r = 0$ não há correlação, se $0 < |r| < 0{,}3$ a correlação é fraca, se $0{,}3 \leq |r| \leq 0{,}6$ é relativamente fraca e se $0{,}6 < |r| \leq 1$ ela é conclusiva.

Regressão Linear Simples:

Em algumas situações pode-se estudar o comportamento de uma determinada variável em função de outra. Isto se torna especialmente interessante se conseguirmos satisfatoriamente um modelo matemático/estatístico que descreve a relação entre as duas variaveis (aleatórias). A regressão linear simples é do tipo $y = ax + b$, na qual y é a variável dependente, x é a variavel independente a e b são parâmetros dados por

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{(n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2)} \text{ e}$$

$$b = \bar{y} - a\bar{x}$$

Como é possivel perceber, trata-se de uma reta que depende de *a* e *b*, obtidos diretamente dos valores dos dados, dos valores da variável *X* (variável independente) e de *Y* (que, por sua vez, depende da variável *X*).

3. Probabilidades e seus conceitos

3.1 Conceitos básicos

Para introduzir os conceitos iniciais da teoria da probabilidade vamos considerar ocasiões onde seja necessário conhecer como se comporta determinado fenômeno que se tenha o intuito de estudar. Para isso, utiliza-se a repetição do experimento quantas vezes se julgar necessário para que tais resultados sejam verificados. Dizemos que um experimento é aleatório quando ao realizar diversas repetições sob mesmas condições, produzem resultados diferentes. Definiremos alguns conceitos importantes (Bussab e Morettin, 2017), a seguir.

Experimento: Estudo no qual são feitas diversas repetições iguais de forma que possam ser observadas e identificadas os diferentes resultados possíveis.

Espaço Amostral: É o conjunto que contêm todos os possíveis resultados do experimento. Denotamos espaço amostral pela letra S.

Eventos: São todos os subconjuntos do espaço amostral que contenham uma parte das possibilidades de S.

Eventos Mutuamente Exclusivos: Dois eventos são ditos mutuamente exclusivos quando eles não podem ocorrer simultaneamente, ou sejam a ocorrência de um exclui a ocorrência do outro.

Frequência Relativa: A frequência relativa por sua vez indica a porcentagem de vezes que ocorreu um determinado resultado quando comparada ao todo. Para calculá-la precisamos primeiramente encontrar a frequência absoluta, que nada mais é do que o número de vezes que um determinado resultado apareceu, e dividi-la pelo total de resultados obtidos. No geral ela vem para ajudar a compreender a análise dos dados, e para isso é de costume organizar esses dados uma tabela de frequência.

Para exemplificar os conceitos acima vamos pensar no experimento do lançamento de dois dados convencionais honestos. Abaixo temos o espaço amostral, o conjunto S.

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Agora vamos considerar dois eventos distintos. O evento 1 (E_1) será a ocorrência da soma das faces ser maior que 9, enquanto o evento 2 (E_2) será a ocorrência da soma das faces ser menor que 7. Analisando o espaço amostral podemos ver que apenas os elementos em cinza pertencem ao E_1, já os azuis pertencem ao E_2.

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

E_1

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

E_2

Na comparação entre os eventos 1 e 2, dizemos que eles são mutuamente exclusivos, pois podemos observar que não existe nenhum elemento em comum nos dois eventos.

Para exemplificar sobre a frequência relativa vamos considerar o evento do lançamento de um dado não viciado por 100 vezes e foram obtidos os resultados abaixo.

Número no Dado	Quantidade de vezes (frequência absoluta)
1	23
2	16
3	19
4	17
5	15
6	10

Agora para calcularmos a frequência absoluta basta fazermos a divisão da frequência absoluta pela quantidade de vezes que o dado foi lançado, como mostra a figura abaixo

Número no Dado	Quantidade de vezes (frequência absoluta)	Frequência Relativa
1	23	$\frac{23}{100} = 0{,}23$
2	16	$\frac{16}{100} = 0{,}16$
3	19	$\frac{19}{100} = 0{,}19$
4	17	$\frac{17}{100} = 0{,}17$
5	15	$\frac{15}{100} = 0{,}15$
6	10	$\frac{10}{100} = 0{,}10$
Total	100	1

Podemos observar que teve uma discrepancia grande em relação aos resultados da frequência relativa, para que isso diminua é necessário ser realizada mais vezes o lançamento do dado assim fazendo com que as frequências relativas se aproximem.

3.2 Definição de Probabilidade

Temos como exemplo de experimento determinístico quando aquecemos a água em uma temperatura de 100º centígrados , pois sabemos que a água entrará em ebulição e sempre haverá esse resultado assim sendo um fenômeno determinístico. Já o lançamento de uma moeda honesta ou de um dado, determinação da vida útil de um componente eletrônico, entre muitos outros, são exemplos de experimentos aleatórios.

Chamamos de espaço amostral o conjunto de todos os resultados possíveis de um experimento aleatório. Representamos esse conjunto por S. Um subconjunto de S

é chamado de evento. O conjunto $\emptyset$ é dito evento impossível e S evento certo. Por fim o conjunto w onde $w \in S$ é chamado evento elementar ou unitário.

Seja um evento E e S um espaço amostral associado a E. A cada evento A associaremos um número real representado por $P(A)$ e denominado como probabilidade de A, que satisfaz as seguintes propriedades:

(i) $0 \leq P(A) \leq 1$

(ii) $P(S) = 1$

(iii) Se A e B forem eventos mutuamente exclusivos, então, $A \cap B = \emptyset$, sendo também que $A, B \subset S$

$$P(A \cup B) = P(A) + P(B)$$

estes são chamados de Axiomas de Probabilidade.

Vamos considerar que o espaço amostral é enumerável. Este é denominado equiprobabilístico se cada elemento do espaço amostral possui a mesma probabilidade de ocorrência dos eventos unitários. Assim a probabilidade para cada elemento será $\frac{1}{n}$.

Para exemplificar o que foi dito no parágrafo anterior vamos considerar uma urna com 5 bolinhas numeradas de 1 a 5. Assim, pelo item (ii) dos axiomas de probabilidade, temos que

$P(\{1 \cup 2 \cup 3 \cup 4 \cup 5\}) = 1$. Logo como sabemos que este modelo é equiprobabilístico, temos que

$$P(número1) = P(número2) = P(número3) = P(número4) = P(número5) = \frac{1}{5}$$

Para ilustrar o item (iii) vamos considerar dois eventos, onde o que se deve fazer é retirar bolas numeradas de 1 a 5 que estão no interior de uma determinada urna.

- Evento 1: Retirar uma bola com número impar ;
- Evento 2: Retirar uma bola que o número seja múltiplo de 2.

Considerando estes dois eventos, temos que

$$P(número\ ímpar) = P(\{1 \cup 3 \cup 5\}) = P(número\ 1) + P(número\ 3) + P(número\ 5) = \frac{1}{5} + \frac{1}{5} + \frac{1}{5} = \frac{3}{5}$$

$$P(número\ par) = P(\{2 \cup 4\}) = P(número\ 2) + P(número\ 4) = \frac{1}{5} + \frac{1}{5} = \frac{2}{5}$$

Como os eventos são mutuamente exclusivos, ou seja, não existe nenhum número que seja ímpar e par, então a probabilidade da união desses eventos é a soma da probabilidade de cada um, ou seja,

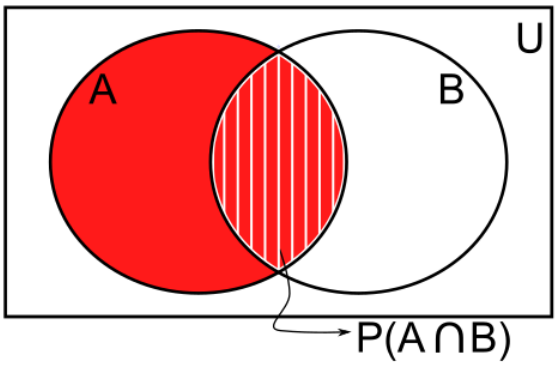
$$P\ (número\ ímpar \cup número\ par) = P\ (número\ ímpar)\ + P\ (número\ par) = \frac{3}{5} + \frac{2}{5} = \frac{5}{5} = 1$$

Como observamos esse tipo de união deu o espaço amostral todo ilustrando também (ii).

3.3 Probabilidade Condicional

Quando estamos interessados em calcular probabilidades, inicialmente surge a preocupação de se comparar eventos com o espaço amostral. No entanto, em muitos casos, existe o interesse em relacionar as probabilidades de dois eventos A e B, vamos supor que o evento A já tenha ocorrido ou que exista a chance de vir a ocorrer. Neste caso, será necessário recalcular a probabilidade do evento B.

Como vamos supor que o evento A já tenha ocorrido, ele passa a ser o novo espaço amostral e as chances de ocorrencia do evento B é a probabilidade de $A \cap B$.



Sejam A e B dois eventos de um espaço amostral S, com $P\ (A)\ >\ 0$. A probabilidade condicional de B, dado A, é definida por 0

$$P(B|A) = \frac{P(B \cap A)}{P(A)}$$

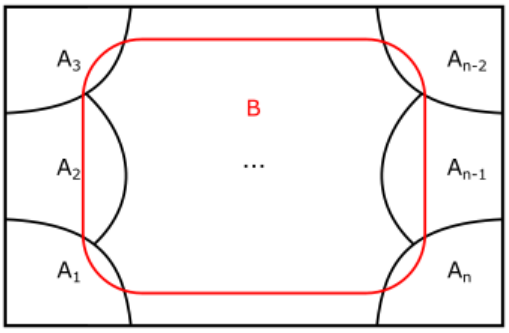
Exemplo: Considere um lançamento de dois dados. Se a soma dos pontos obtidos for 7, qual será a probabilidade do produto ser igual a 12?

Para resolvermos esse tipo de problema temos que utilizar probabilidade condicional, pois queremos saber qual a probabilidade do produto ser igual a 12 dado que a soma dos pontos obtidos foi 7. Então vamos chamar de evento A o lançamento de dois dados com soma das faces igual a 7 e evento B será o produto das faces ser igual a 12. Temos então $A = \{(1,6)\ ;\ (2,5)\ ;\ (3,4)\ ;\ (4,3)\ ;\ (5,2)\ ;\ (6,1)\}$ e $B = \{(2,6)\ ;\ (3,4)\ ;\ (4,3)\ ;\ (2,6)\}$, consequentimente temos $B \cap A = \{(3,4)\ ;\ (4,3)\}$, então calculando $P(B|A)$, temos

$$P(B|A) = \frac{2}{6} = \frac{1}{3}$$

3.4 Teorema da Probabilidade Total

Sejam os eventos $A_1, A_2, \ldots, A_n$ representando uma partição do espaço amostral S e B um evento qualquer de S.



Temos que o evento B pode ser escrito da seguinte forma,

$$B = (A_1 \cap B) \cup (A_2 \cap B) \cup \cdots \cup (A_{n-1} \cap B) \cup (A_n \cap B)$$

Como $A_i \cap B$ são mutuamente exclusivos entre si, para todo i distinto, temos que

$$P(B) = P(A_1 \cap B) + P(A_2 \cap B) + \cdots + P(A_{n-1} \cap B) + P(A_n \cap B)$$

Da probabilidade condicional temos que $P(B|A_i) = \frac{P(B \cap A_i)}{P(A_i)}$, que pode ser reescrita como $P(B \cap A_i) = P(B|A_i) \cdot P(A_i)$ e substituindo na expressão de *P*(*B*) acima, temos que

$$P(B) = P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) + \cdots + P(B|A_{n-1}) \cdot P(A_{n-1}) + P(B|A_n) \cdot P(A_n)$$

Então concluimos que,

$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$ para *i=1,2,...,n*.

Exemplo: Cada uma de 3 caixas possui 2 gavetas. A primeira caixa contém uma moeda de ouro em cada gaveta, a segunda caixa contém 2 moedas de prata em cada gaveta, e a terceira caixa, uma moeda de ouro em uma gaveta, e uma de prata na outra. Uma caixa é escolhida ao acaso; uma de suas gavetas é escolhida ao acaso e aberta, encontrando-se uma moeda de ouro. Qual a probabilidade que a moeda da outra gaveta seja também de ouro?

Para resolvermos vamos nomear os dois eventos, o primeiro é o evento A que é a primeira moeda ser de ouro e o segundo o evento B que é da segunda moeda também ser ouro. Agora teremos que usar probabilidade condicional, pois como foi dito no enunciado que a primeira moeda já foi de ouro, então temos que calcular $P(B|A)$.

Então para isso precisamos encontrar $P(B \cap A)$ que é a probabilidade da primeira e segunda moeda serem de ouro e isso só acontece apenas na caixa I, pois é a unica caixa que tem 2 moedas de ouro. Assim concluimos que

$$P(B \cap A) = P(caixa\ I) = \frac{1}{3}$$

Agora precisamos ver a probabilidade da primeira moeda ser ouro. Para isso vamos usar o teorema da probabilidade total.

$$P(A) = P(A|caixa\ I) \cdot P(caixa\ I) + P(A|caixa\ II) \cdot P(caixa\ II) + P(A|caixa\ III) \cdot P(caixa\ III)$$

Como todas as moedas da caixa I são de ouro temos que a probabilidade da moeda ser ouro dado que foi escolhido a caixa I é 1, ou seja, $P(A|caixa\ I) = 1$. Já na caixa II é zero pois não tem nenhuma moeda de ouro, então $P(A|caixa\ II) = 0$. Na caixa III temos uma ouro e uma prata então a probabilidade de ser a de ouro é $\frac{1}{2}$, com isso temos que $P(A|caixa\ III) = \frac{1}{2}$. E a probabilidade da escolha da caixa é $\frac{1}{3}$, portanto $P(caixa\ I) = P(caixa\ II) = P(caixa\ III) = \frac{1}{3}$. Fazendo as substituições temos,

$$P(A) = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{3} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$$

Então concluimos que a probabilidade da segunda moeda ser ouro dada que a primeira foi ouro é

$$P(B|A) = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

Exemplo (MEYER, 2006):

Uma peça é manufaturada por 3 empresas, 1, 2 e 3. Sabe-se que 1 produz o dobro de 2 que produz o mesmo que 3. Sabe-se também que 2% das peças de 1 e 2 são defeituosas e 4% das peças de 3 também. Todas as peças são colocadas em um depósito e uma delas é retirada ao acaso. Qual é a probabilidade de que ela seja defeituosa?

Solução:

Inicialmente, devemos definir os eventos adequadamente para resolver essas questões. Fazemos

D={a peça é defeituosa}

F_1= {a peça provem da fábrica 1}

F_2= {a peça provem da fábrica 2}

F_3= {a peça provem da fábrica 3}

Devemos calcular a probabilidade da peça selecionada ser defeituosa, que é denotada por *P*(*D*), logo

P(*D*) = *P*(*D*∩ F_1) + *P*(*D*∩ F_2) + *P*(*D*∩ F_3).

Contudo, não são fornecidas as probabilidades das intersecções, então

$P(D) = P(D|F_1)P(F_1) + P(D|F_2)\,P(F_2) + P(D|F_3)\,P(F_3)$.

Trata-se do teorema da probabilidade total. Do problema temos que

$P(F_1) + P(F_2) + P(F_3) = 1$, logo, fazendo a expressão como 2x + x + x =1, então x=1/4 e assim

$P(F_1)$= ½, $P(F_2)$= 1/4 e $P(F_3)$=1/4.

Também,

$P(D|F_1)$=0,02, $P(D|F_2)$=0,02 e $P(D|F_3)$=0,04.

De posse dessas informações, é possível calcular $P(D)$,

$P(D) = P(D|F_1)P(F_1) + P(D|F_2)\,P(F_2) + P(D|F_3)\,P(F_3)$= 0,02.1/2 +0,02.1/4 + 0,04.1/4 = 0,025.

Então, de forma resumida, podemos dizer que 2,5% das peças produzidas pelas fábricas, em média, terão algum defeito.

4. Aplicações

Em qualquer área de conhecimento, as aplicações são, geralmente, os motivadores em estudar conceitos, técnicas e metodologias. Para as áreas mais aplicadas, resolver problemas reais seria como uma meta final de tudo que se empenhou em termos de tempo e esforço. Igualmente, as aplicações podem funcionar como algo que pode ser obtido, contudo, antes disso, tem que ser aprendido, entendido. Na busca de resultados satisfatórios, dominar o conhecimento é fundamental para que isto ocorra e pode servir de incentivo ao aprendizado.

4.1 Aplicação de Indicadores a Dados de Covid-19

A Covid-19 tem sido amplamente combatida pelos setores de saúde do Brasil, moléstia muito contagiosa e que levou muitas vidas antes de se descobrir vacinas que pudessem evitar muitas mortes. As vacinas não impedem que os indivíduos peguem a doença mas diminuem muito seus efeitos causando uma diminuição considerável de mortes.

Antes do início da vacinação, as cidades combatiam a doença com informações sobre os meios de transmissão e de algumas medidas como distanciamento, utilização de máscaras, entre outras. Apesar da doença ser de cunho geral, cidades, estados e países apresentavam, em alguns casos, resultados diferentes, por vários fatores. A estatística pode medir essas diferenças por meio de alguns coeficientes, como pode ser visto no problema a seguir.

Sejam dois países com comportamentos distintos em relação ao número de casos e óbitos de Covid-19.

Dois países A e B têm as seguintes informações sobre a Covid-19:

Características	População Total	Casos confirmados	Número de Óbitos
País A	2.100.000	50.000	2.500
País B	4.800.000	100.000	7.000

a- Em qual desses países você considera mais fácil ser infectado pela doença?

Para comparação devemos calcular o coeficiente de transmissão para cada país.

País A: $coef.transmissão = \frac{número\ de\ confirmados}{número\ da\ população} = \frac{50.000}{2.100.000} \approx 0{,}024$

País B: $coef.transmissão = \frac{número\ de\ confirmados}{número\ da\ população} = \frac{100.000}{4.800.000} \approx 0{,}021$

Logo, no país A o indivíduo pode ser contaminado mais facilmente.

b- Uma vez infectado, em qual país é mais fácil ir a óbito?

Para comparação devemos calcular o coeficiente de óbito para cada país.

País A: $coef.óbito = \frac{número\ de\ óbitos}{número\ de\ confirmados} = \frac{2.500}{50.000} \approx 0{,}05$

País B; $coef.óbito = \frac{número\ de\ óbitos}{número\ de\ confirmados} = \frac{7.000}{100.000} \approx 0{,}07$

Logo, no país B o indivíduo pode ir a óbito mais facilmente.

c- Uma vez não infectado, em qual país é mais fácil ir a óbito?

Nesta situação é necessário realizar um ajuste na população devido ao fato de haver também indivíduos infectados no total. Após o ajuste, devemos calcular o coeficiente de óbito para cada país.

País A: $coef.óbito = \frac{número\ de\ óbitos}{número\ ajustado\ da\ população} = \frac{2.500}{2.100.000-50.000} \approx 0{,}00122$

País B: $coef.óbito = \frac{número\ de\ óbitos}{número\ ajustado\ da\ população} = \frac{7.000}{4.800.000-100.000} \approx 0{,}00149$

Logo, uma vez não infectado, no país B o indivíduo pode ir mais facilmente a óbito.

Neste problema os dados são fictícios devido ao fato de não haver coleta de dados reais confiáveis e completos possíveis para os cálculos acima. Há grande dificuldade em estimar o número de infectados pois não são testados indivíduos suficientemente para esta estimação.

Um problema parecido com esse pode ter uma outra forma de abordagem utilizando técnicas de probabilidade.

4.2 Aplicação com Probabilidade

Três cidades, A, B e C, estão enfrentando casos com a Covid-19, contudo, em suas populações há pessoas contaminadas que não sabem ainda que portam a doença. Sabe-se que a cidade A tem 350.000 habitantes dos quais estima-se que 1% estão nessas condições, a cidade B tem 100.000 habitantes dos quais 2% estão nessas condições e a cidade C tem 50.000 habitantes dos quais 3% também estão nessas condições.

Todo ano uma grande festa é realizada no campo com a presença média de 10% de todos os habitantes de cada uma das três cidades.

Qual é a probabilidade de uma pessoa da festa estar doente?

Se uma pessoa da festa é detectada com a doença, qual é a probabilidade de ter vindo

de cada uma das cidades?

Inicialmente, vamos definir os seguintes eventos:

D: o indivíduo está doente
A: o indivíduo é da cidade A
B: o indivíduo é da cidade B
C: o indivíduo é da cidade C

Da descrição do problema, temos que

$P(D \mid A)$ =0,01
$P(D \mid B)$ =0,02
$P(D \mid C)$ =0,03

Também, se 10% da população vai à festa,
$P(A) = \frac{35000}{50000} = 0,7$
$P(B) = \frac{10000}{50000}$ =0,2
$P(C) = \frac{5000}{50000}$ =0,1

Na primeira pergunta, o que se deseja é a probabilidade de um indivíduo da festa estar doente, denotada por *P*(*D*).

Temos que $D = (D \cap A) \cup (D \cap B) \cup (D \cap C)$, logo
$P(D) = P(D \cap A) + P(D \cap B) + P(D \cap C)$.
Então,
$P(D) = P(A).P(D \mid A) + P(B).P(D \mid B) + P(C).P(D \mid C)$.
$P(D)$ =0,7 . 0,01 + 0,2 . 0,02 + 0,1 . 0,03
$P(D)$ =0,014

Para a segunda pergunta, o que se procura são as probabilidades $P(A \mid D)$, $P(B \mid D)$ e $P(C \mid D)$, logo

$$P(A \mid D) = \frac{P(A \cap D)}{P(D)} = \frac{0{,}007}{0{,}014} = \frac{7}{14}$$

$$P(B \mid D) = \frac{P(B \cap D)}{P(D)} = \frac{0{,}004}{0{,}014} = \frac{4}{14}$$

$$P(C \mid D) = \frac{P(C \cap D)}{P(D)} = \frac{0{,}007}{0{,}014} = \frac{3}{14}$$

Percebe-se que, apesar de 70% da população ser da cidade A, a probabilidade da pessoa detectada como doente na festa é de apenas 0,5.

Estes resultados, apesar de ser intuitivo, é conhecido na literatura como Teorema de Bayes.

4.3 Aplicação utilizando Medidas Estatísticas e Médias Móveis

A Covid-19 representou um grande desafio para o setor de saúde, no início da doença havia muitp desconhecimento das vias de transmissão, sintomas, mortalidade e tratamentos possiveis. Isto também se refletiu na coleta de dados, quais eram as informações realmente importantes diante de tantas dúvidas sobre a doença? Ainda hoje há muitos dados não confiáveis por vários fatores. As informações aqui analisadas se referem a dados de óbitos e casos por dia (foram anotados conforme iam sendo notificados) no Ministério da Saúde (SUS) (Brasil, 2022) até o mês de julho de 2022.

Os dados no Brasil, de forma geral, em boa parte, foram notificados não exatamente nos dias em que ocorreram, isto causou discrepâncias entre dias com poucos registros e outros com muitos registros (era, por exemplo, o caso dos finais de semana). Desta forma, os gráficos de óbitos e casos ficavam com muita variação e apresentavam difícil análise de tendência (vide Figuras 4.1 e 4.2, a seguir).

Pelas Figuras 4.1 e 4.2 pode-se observar que, considerando todo o período, há tendências observáveis, contudo, para uma pequena fração do tempo, os valores ficavam muito discrepantes de forma a não se conseguir vislumbrar se as tendências eram de alta, de baixa ou estáveis (repare que os gráficos parecem ter dois segmentos distintos em alguns setores devido a valores discrepantes).

Modelos complexos foram abordados, sem resultados satisfatórios, as técnicas mais básicas foram consideradas, como é o caso dos chamados modelos de médias móveis (Morettin e Toloi, 2018). Estes apresentam uma suavização dos dados e melhorando as análises de tendências (vide Figuras 4.3 e 4.4, a seguir).

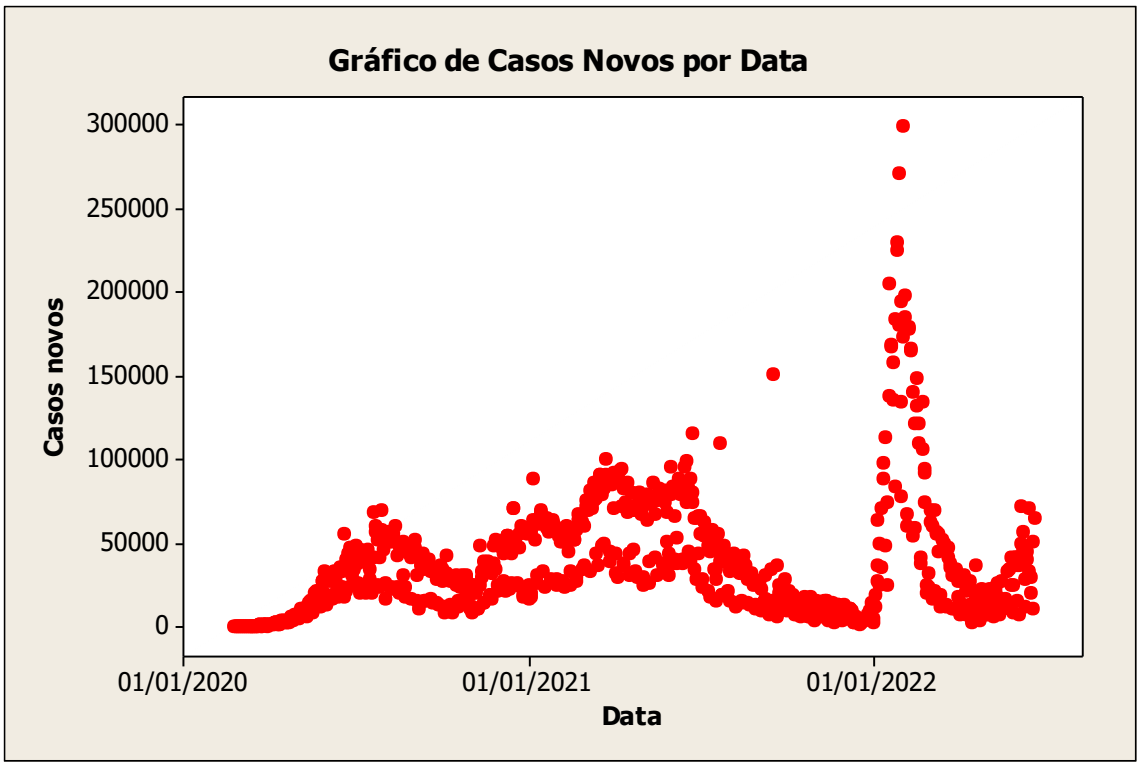


Figura 4.1: Gráfico de casos novos por dia para todo o período até julho de 2022.

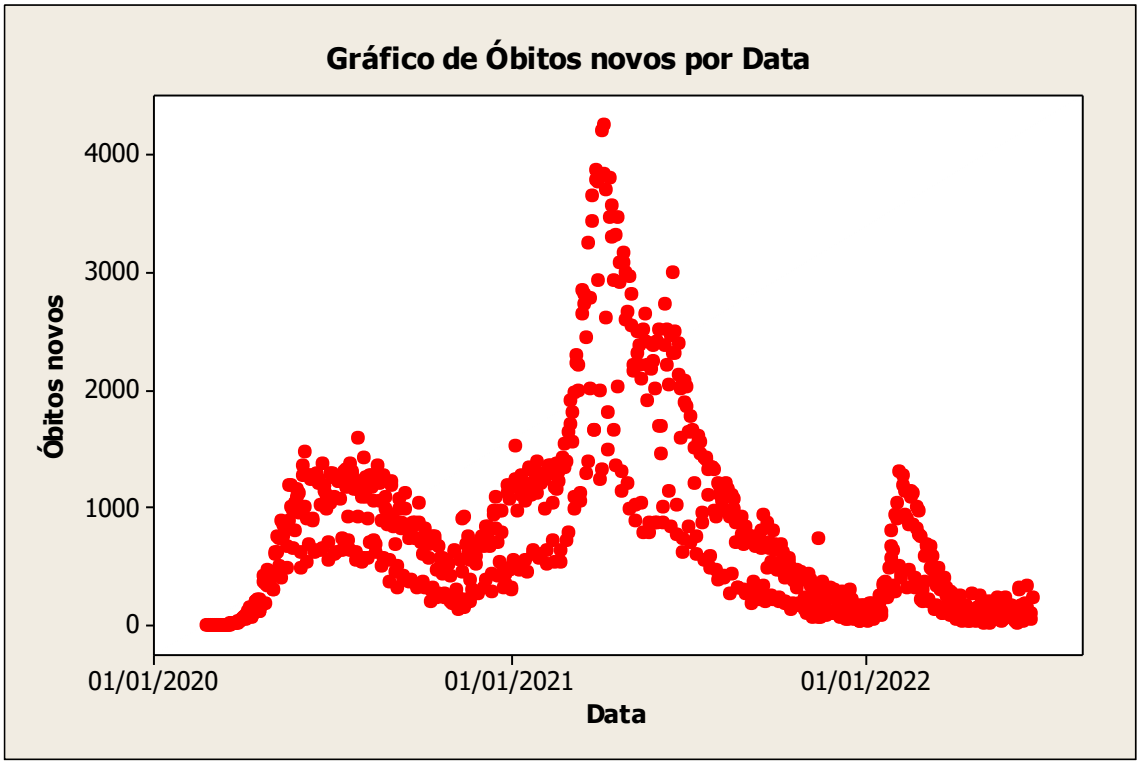


Figura 4.2: Gráfico de óbitos novos por dia para todo o período até julho de 2022.

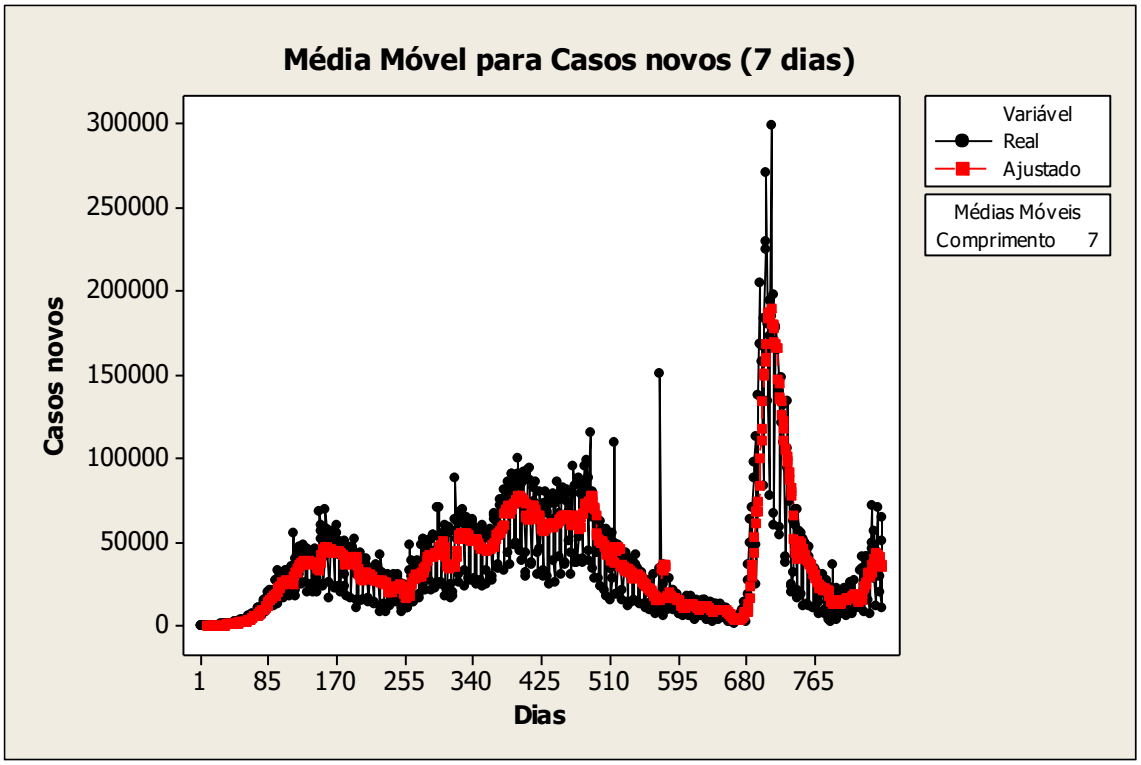


Figura 4.3: Gráfico de médias móveis para casos novos por dia para todo o período até julho de 2022.

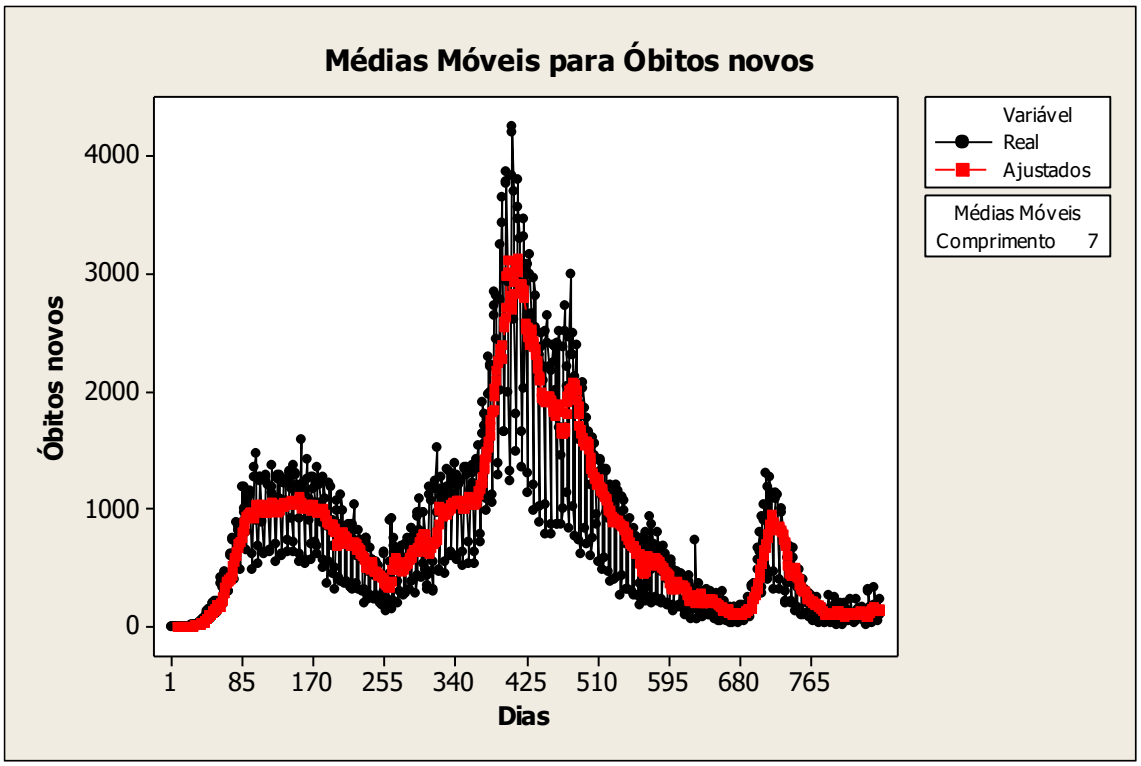


Figura 4.4: Gráfico de médias móveis para óbitos novos por dia para todo o período até julho de 2022.

Nas Figuras 4.3 e 4.4 são observados os pontos suavizados (em vermelho) e os reais (em preto) fornecendo uma visualização mais reveladora de tendências. Nota-se a grande variabilidade que os pontos reais apresentam dificultando bastante realizar uma análise no decorrer do processo, dia após dia com variações muito flutuantes. As médias móveis, apesar de ser uma técnica simples, possibilita que as tendências possam reveladas subsidiando informações e decisões de interesse coletivo.

Para uma análise mais elaborada foi realizada um gráfico que procura relacionar os caos novos e os óbitos novos. Para tanto, foram registrados as médias móveis para ambos e colocados em um mesmo gráfico com dois eixos em escalas difrentes (um ao lado direito e outro ao lado esquerdo). Apesar de estarem em patamares numéricos distintos, é possivel perceber se a tendência de ambos apresentam alguma associação (vide Figura 4.5).

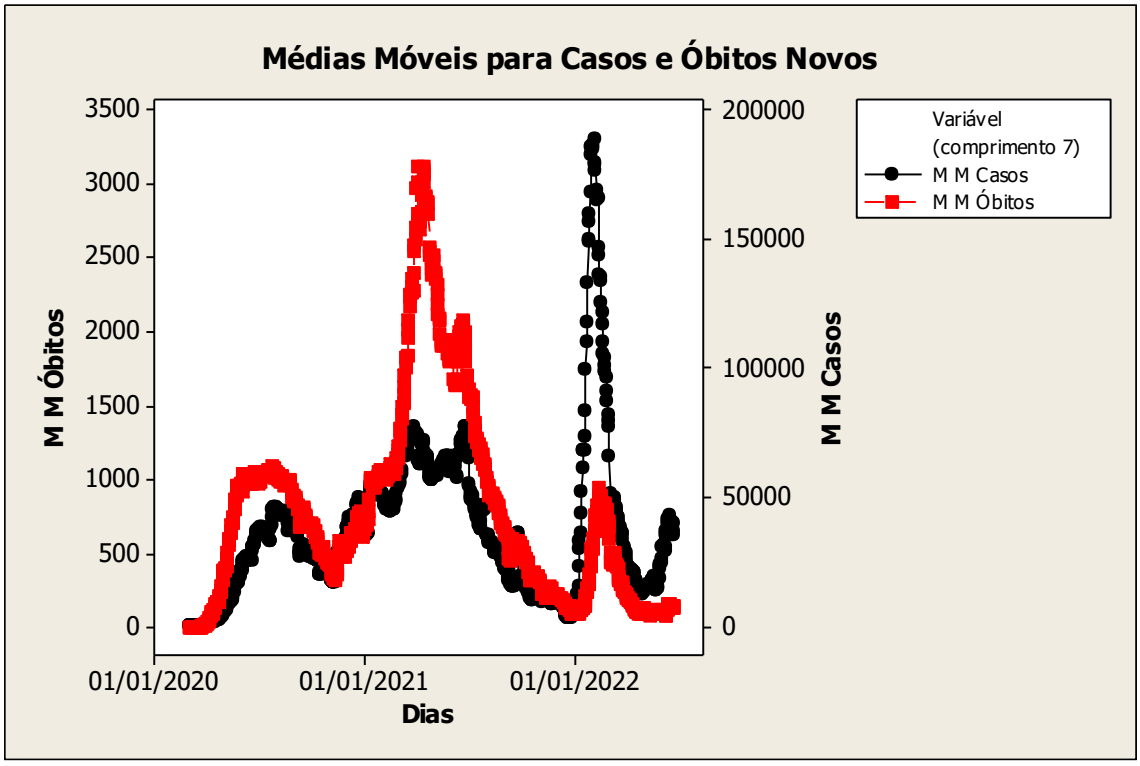


Figura 4.5: Gráfico de médias móveis para casos e óbitos novos por dia com duas escalas.

Percebe-se que há três picos de elevação para o período considerado. No primeiro, juntamente com os casos, houve um aumento de óbitos. Este comportamento fica mais acentuado para o segundo pico no qual o número de óbitos aumenta consideravelmente se desprendendo da curva de casos. O terceiro pico tem um comportamento bem diferente dos demais e será discutido com as técnicas da próxima aplicação.

No início da pandemia se pensava que os casos pudessem se desnvolver conforme uma curva bem conhecida, a curva normal.

No caso da COVID 19, o que tanto se considerou sobre a curtose e a demanda por serviços hospitalares é que, conforme o comportamento de isolamento aumentasse ou diminuísse, teríamos uma curva platicúrtica ou leptocúrtica. Havendo recursos limitados (e, em alguns casos, muito limitados) a curva platicúrtica seria favorável à situação apresentada na maioria dos países, como revela a Figura 4.6.

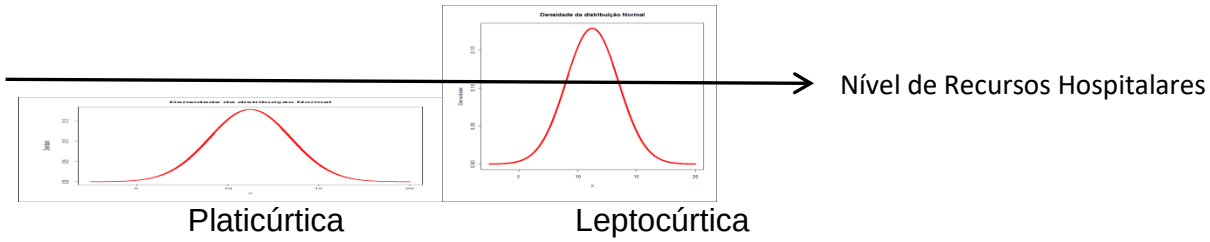


Figura 4.5: Gráfico de curvas normais em relação aos recursos hospitalares.

Evidentemente, na curva leptocúrtica as necessidades extrapolariam o nível de recursos, contudo, como foi visto, a curva de casos e óbitos tiveram mais de um pico, não se apresentaram como esperado.

As medidas estatísticas como média, variância, moda, mediana, entre outras, pouco puderam ajudar nas análises desses dados. Estes apresentam um comportamento temporal e com grande variabilidade.

4.4 Aplicação utilizando Correlação e Regressão Linear

A medida correlação (ou correlação linear) é bastante utilizada para verificar se há alguma associação entre duas variáveis. Existem vários tipos de técnicas para verificar o grau de relação entre duas variáveis, contudo, esta medida é a mais utilizada, dada sua simplicidade de obtenção e de interpretação.

Observando-se a Figura 4.5 pode-se concluir pelo gráfico que não haveria uma associação conclusiva calculando-se a correlação. No entanto, outros tipos de gráficos podem apresentar, além de relações, fenômenos que apontem para alguma mudança de comportamento ou quebra de relação.

Foi realizado um gráfico com os casos novos acumulados e os óbitos novos acumulados (Figura 4.6) e observou-se um comportamento distinto para uma parte da série, a parte final do período.

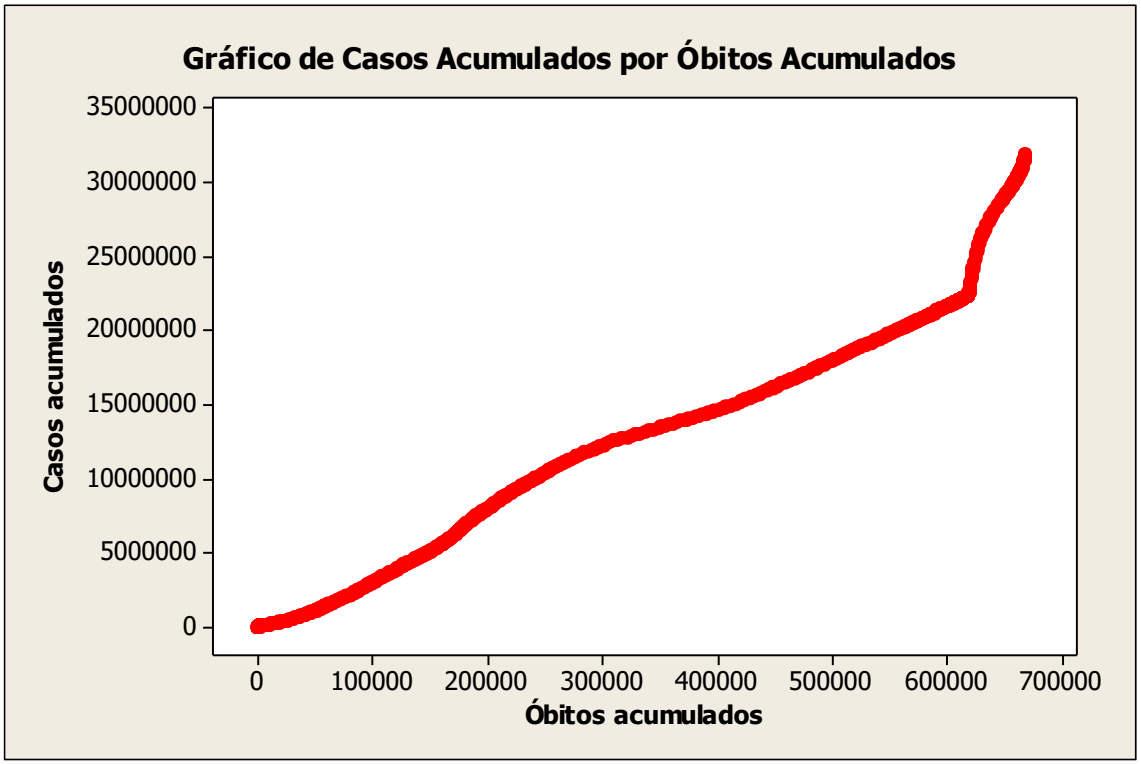


Figura 4.6: Gráfico de casos e óbitos acumulados por dia para todo o período até julho de 2022.

Apesar da mudança de comportamento, a correlação linear para esses dados foi de 0,982, portanto, muito alta. Isto demonstra a importância de associar as técnicas para poder realizar uma análise mais completa, informativa e conclusiva do fenômeno que se está estudando.

Verificando na planilha de dados, observou-se que essa mudança de comportamento teve início no começo do mês de dezembro de 2021. O número de casos continua se acumulando bastante mas não é acompanhado pelo número de óbitos, este apresenta uma desaceleração que faz com que a ponta do gráfico se altere e perca o comportamento anteriormente observado. Sabe-se que, no final de dezembro de 2021, o Brasil alcançou 80% de vacinação de seu público alvo e, a este fato, muito provavelmente se deve a mudança de comportamento em relação aos óbitos, mas não aos casos. A vacina protege dos efeitos da doença, mas não impede o indivíduo de contrair a mesma.

Uma questão interessante é pensar como essa situação se desenvolveria se a vacina não fosse produzida e aplicada na população. Para tanto, foi realizada uma regressão linear simples (vide Figura 4.7, a seguir), contudo, retirando-se o período final, referente à mudança de comportamento observada e atribuída à vacinação.

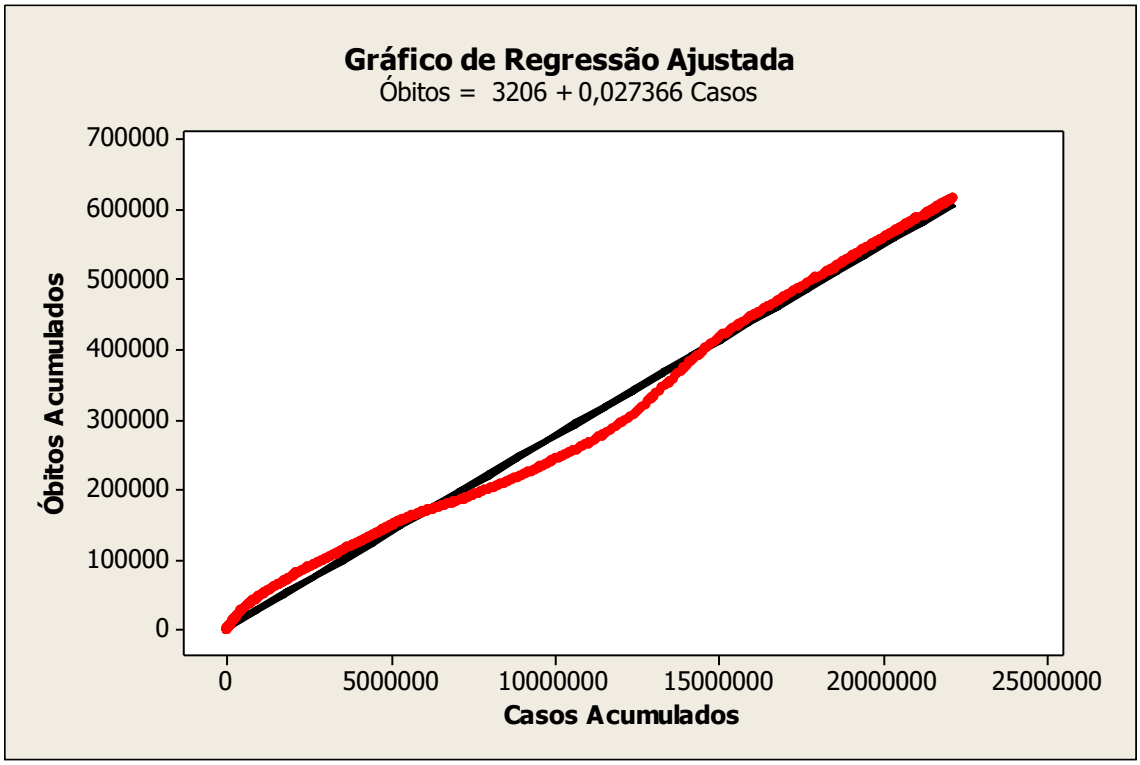


Figura 4.7: Regressão ajustada para casos e óbitos acumulados até início de dezembro de 2021.

A medida de correlação foi novamente calculada e, com a retirada do período final, o resultado foi de 0,997, sendo, como esperado, um pouco mais alto que com os dados anteriores.

A reta de regressão ajustada foi obtida como

$$\text{óbitos} = 3206 + 0{,}027366.\text{casos}$$

ou seja, a variável dependente (óbitos) é determinada pela variável independente (casos), como esperado. Desta forma, sem a vacina, seria possível pensar que esta reta de regressão estivesse satisfatória para um período futuro (entendendo que o comportamento não mudaria, isto deve ser realizado com cautela nas suposições). Para tanto, não haveria vacinas, mudança de comportamento das pessoas, aparecimento de novas variantes, entre outras suposições cabíveis. Assim sendo, aplicando 25.462.033 casos na regressão, teríamos 700.000 óbitos, atingidos entre final de janeiro e começo de fevereiro de 2022.

Deve-se observar o impacto que a vacinação teve na preservação de vidas.

5. Conclusões

Quando se pensa na área de estatística, de forma geral, são vislumbradas técnicas complexas, cálculos matemáticos ou computacionais extensos, díficeis de entender.

Este trabalho procura apresentar técnicas estatísticas que, apesar de introdutórias, podem ser aplicadas à situações reais e práticas do cotidiano, sejam mais complexas ou sejam mais fáceis de compreender.

A pandemia de Covid-19 se revelou desafiadora para ser enfrentada e entendida, houve uma necessidade premente de conhecimento que teve que ser obtido muito rapidamente, técnicas estatísticas mais complexas não se revelaram boas o suficiente para descrever o número de casos para que se evitassem os óbitos.

Para obter informações rapidamente e que fossem compreensíveis, algumas técnicas estatísticas menos sofisticadas foram empregadas a dados reais para a Covid-19. Algumas dessas técnicas foram aqui apresentadas e podem ser ensinadas em nível de ensino médio fornecendo mais visibilidade a uma área muito importante e pouco difundida nesse nível, a área de estatística.

6. Pesquisa futura

Neste trabalho, inicialmente, se pretendia aplicartécnicas de probabilidade e Cadeias Discretas de Markov a dados reais de Covid-19. No entento, essas cadeias precisam de estabilidade para que possam ser analisadas em tempo futuro, o que não ocorreu nos dados levantados. Como considerado anteriormente, tanto o número de casos como o número de óbitos flutuavam muito em suas séries não oferecendo condições para que se considerasse alguma estabilidade razoável.

Com a aplicação de vacinas, ao que parece, os dados começam a apresentar um comportamento mais estável, ainda não suficiente em quantidade para realizar esse tipo de estudo, mas como pesquisa futura.

Também, diante da condição de estabilidade, pode-se realizar uma regressão que, possivelmente, possa descrever o número de óbitos futuros em relação ao número de casos presentes de Covid-19.

Referências Bibliográficas

BRASIL. Ministério da Saúde. Sistema Único de Saúde (SUS). Covid-19 no Brasil. Brasilia: SUS, 2022. Disponível em: https://infoms.saude.gov.br › extensions › covid-19_html Acesso em 30 jul. 2022.

Bussab, W. O.; Morettin, P.A. *Estatística Básica*. São Paulo: Saraiva, 9ª. Ed., 2017.

Meyer, P. L. Probabilidade: Aplicações à estatística. Ed. LTC, São Paulo, 2006

Morettin, P. A.; Toloi M. C. Análise de Séries Temporais: modelos lineares univariados. Ed. Blucher, São Paulo, 2018.

Perlman, Y.; Yechiali, U. The impact of infection risk on customers' joining strategies. *Safety Science*, v. 138, jun. 2021.