



A ESTATÍSTICA DE FICAR EM CASA: UTILIZAÇÃO DE MÉTODOS ESTATÍSTICOS PARA ENTENDER OS CASOS DE COVID-19 NO ESTADO DE SÃO PAULO E A IMPORTÂNCIA DO ISOLAMENTO SOCIAL.

Ana Julia Gonçalves PADUA¹
Ingrid Pikinskeni Moraes SANTOS²
Alisson Fernando Coelho do CARMO³

RESUMO: Este artigo tem como objetivo fazer uma análise estatística, com base nos dados disponíveis no site da Fundação Sistema Estadual de Análise de Dados (SEADE), do desdobramento do número de casos do novo Coronavírus no estado de São Paulo. Esse estado foi escolhido em razão de ser o contexto de maior expressão dos casos da pandemia no cenário nacional e o primeiro lugar no Brasil a apresentar casos da doença e, por isso, a amostra analisada se dá a partir do dia 26 de fevereiro e conclui-se no dia 29 de abril de 2020. Para as análises estatísticas realizadas neste artigo foram utilizadas medidas de posição, dispersão, distribuição normal e correlação linear para entender a dependência entre as variáveis e destacar a relevância do isolamento social.

Palavras-chave: Coronavírus. São Paulo. Distribuição normal. Variação temporal. Isolamento social.

1 INTRODUÇÃO

No dia 30 de janeiro de 2020 a Organização Mundial da Saúde (OMS) declarou o surto causado pelo novo coronavírus (*COVID-19*), mas somente no dia 11 de março de 2020 a *COVID-19* foi decretada como uma pandemia. Este vírus é transmitido pelo ar, causado pelo coronavírus *SARS-CoV-2*. Cerca de 20% dos infectados são assintomáticos (BRASIL DE FATO, 2020). Katelyn

¹ Discente Ana Julia Gonçalves Pádua do 2º ano do curso de administração do Centro Universitário Antônio Eufrásio de Toledo de Presidente Prudente. anajuliadepadua@gmail.com.

² Discente Ingrid Pikinskeni Moraes Santos do 2º ano do curso de administração do Centro Universitário Antônio Eufrásio de Toledo de Presidente Prudente. ingridpmsantos@gmail.com.

³ Docente do curso de Administração do Centro Universitário Antônio Eufrásio de Toledo de Presidente Prudente. Doutor em Ciências Cartográficas pela Universidade Estadual Paulista. alisson.carmo@toledoprudente.edu.br. Orientador do trabalho.

Gostic, pesquisadora da Universidade de Chicago, afirma que, se alguém não percebe que foi infectado e não apresenta sintomas, a detecção se torna impossível, não por erros humanos, mas em razão do comportamento do vírus (MADE FOR MINDS, 2020). Neste cenário, análises afirmam que 44% dos casos foram infectados por pessoas pré-sintomáticas, destacando a importância de se considerar as contaminações dadas por assintomáticos e pré-sintomáticos (NATURE MEDICINE, 2020).

No Brasil, a pandemia teve início em 26 de fevereiro e desde então os números de infectados começaram a aumentar. Em 20 de maio de 2020, foram confirmados no Brasil 293.357 casos de *COVID-19* e 18.894 mortes (G1, 2020). Tendo em vista a dimensão da doença que já naquele momento, matou mais que a dengue em 2019. Segundo o Ministério da Saúde, torna-se necessária a estatística de ficar em casa.

A escolha de analisar o desdobramento da doença no estado de São Paulo foi devido ao primeiro caso de *COVID-19* do Brasil ter sido confirmado na cidade de São Paulo e levando em conta que é um dos estados mais afetados com grande abrangência em toda a área do estado, como pode ser visto na Figura 1. Em maio de 2020, o estado apresentava 69.859 casos e variação diária de óbitos de 4%, enquanto que no Brasil toda essa variação é de 5%.

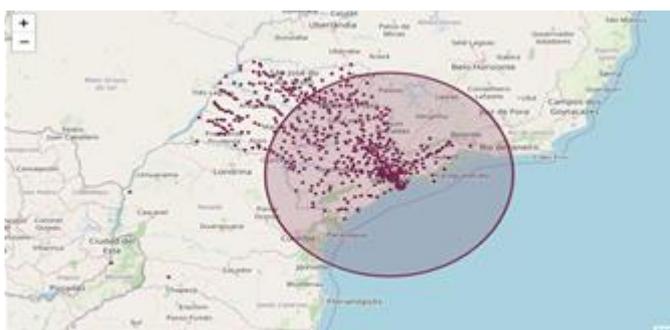


Figura 1: Mapa de casos de Coronavírus no estado de São Paulo, destacando o epicentro na região metropolitana.

Fonte: <https://ciis.fmrp.usp.br/covid19/sp-br/>.

O objetivo deste trabalho é, através da estatística, entender os números de caso do *COVID-19*, sua dimensão e fazer uma projeção do futuro a respeito da pandemia a partir de uma análise estatística baseada na distribuição normal de probabilidade, considerando o mesmo comportamento histórico do passado. Para tanto, foram utilizados os dados reais do início da pandemia no estado de São Paulo,

disponíveis até maio de 2020. Foram utilizados os dados do início da pandemia para verificar a viabilidade da técnica em comparação com os dados reais.

Este trabalho tem o objetivo de entender a variação temporal dos casos no estado, para ressaltar a extrema relevância de entender o quanto a pandemia pode avançar e seus efeitos na sociedade.

2 MATERIAIS UTILIZADOS

Como a área de interesse compreende todo o estado de São Paulo, os dados foram obtidos no site da Fundação Sistema Estadual de Análise de Dados (SEADE), por ser um repositório oficial com dados confiáveis e atualizados. Esta fonte de dados foi importante pois através dela foi encontrado o número de casos confirmados diariamente e o total de óbitos. A amostra analisada neste artigo começou no dia 26 de fevereiro, quando foi confirmado o primeiro caso de COVID-19 de um senhor de 61 anos recém-chegado da Itália, e terminou no dia 29 de abril.

Para a criação de planilhas e gráficos foi utilizado o Planilhas Google, pois é uma ferramenta que permite alto nível de interação e colaboração simultânea, garantindo e favorecendo o trabalho remoto. Ademais, usando o Planilhas Google foram criadas planilhas, gráficos por meio de fórmulas para realização da melhor forma possível a análise estatística do estado de São Paulo. O Google Meet, foi outra plataforma de suma importância, levando em consideração a necessidade das discussões mesmo em época de isolamento social, pois através dele conseguimos interagir com o orientador e realizar reuniões.

2.1 Métodos utilizados

A estatística é uma importante ferramenta analítica capaz de testar e avaliar hipóteses estatísticas, para verificar a força da evidência clínica e, assim, se existem associações entre grupos ou a veracidade de fenômenos de interesse (RODRIGUES; LIMA; BARBOSA, 2020). Sendo assim, a utilização de amostras é de

suma importância para viabilizar o processo analítico e de tomada de decisão, subsidiando a análise realizada nesse artigo, ou seja, através da estatística entender os números de casos da pandemia causada pelo novo Coronavírus.

Primeiramente, foram feitos dois gráficos na Planilha do Google, com todos os dados retirados do site da SEADE, no qual todos os dias são atualizados os números de casos confirmados e óbitos diários. Um dos gráficos contém o número de casos confirmados por dia e no outro gráfico foi representado todas as datas dos óbitos de cada dia. Feito isso, foi realizada a estatística descritiva, que é a etapa inicial de uma análise de dados, ela abrange várias técnicas para descrever um conjunto de dados, ou seja, sintetiza todos os dados de forma direta. Por se tratar de dados quantitativo, ou seja, que podem ser medidos e verificados, a exemplo da quantidade de casos confirmados e número de óbitos, foram utilizadas algumas variáveis quantitativas que podem ser classificadas em dois grupos: o primeiro grupo é chamado de medidas de posição, desse grupo foi utilizado a média, moda e mediana; o segundo grupo é chamado de medidas de dispersão, dele foi utilizado o desvio padrão e coeficiente de variação.

Após esse processo, foram utilizadas as medidas de tendência central, a começar pela média, que consiste em somar todos os valores do conjunto de dados e dividi-los pelo número de elementos deste conjunto para obter um ponto de equilíbrio da distribuição de dados, ou seja, um ponto médio cujo valor demonstra a concentração de dados; a moda, representada pelo valor mais frequente dentro do conjunto de dados. Portanto, a moda foi utilizada para encontrar uma referência rápida e aproximada da posição central dos dados. Foi realizado também os cálculos de mediana, para encontrar o valor que separa o conjunto exatamente pelo ponto médio, ou seja, o valor central do conjunto.

Na sequência, foram realizados os cálculos do grupo chamado de medidas de dispersão. Esses cálculos estatísticos são usados para esclarecer o quanto as amostras variam e também para tornar o estudo das amostras mais confiáveis, tendo em vista que as variáveis citadas no parágrafo acima podem, muitas vezes, não apresentam os resultados de forma concreta, pois são medidas com características centrais. A começar pelo desvio padrão, uma medida que indica a variação dos dados dentro de uma amostra, com relação à média. E por último, o coeficiente de variação (CV), usado para facilitar o estudo dos dados coletados, pois apenas relacionar os resultados do desvio padrão seria inviável em razão da não

padronização das unidades. Dessa forma, o CV serve como um medidor de variabilidade, ou seja, são os dados do desvio padrão, porém expressos por uma porcentagem de forma padronizada.

Com base na distribuição contínua, que é uma forma de descobrir as probabilidades de acontecimentos entre variáveis aleatórias, foi utilizada a modelagem a partir da distribuição normal, pois ela é aplicável para modelar fenômenos naturais, e eventos aleatórios, como pode ser o caso do COVID-19.

Conforme representado na Figura 2, a distribuição normal é facilmente reconhecida pelo seu formato de “sino”, este formato representa que, embora os dados não sigam um padrão linear constante, conforme a quantidade amostral cresce, é esperado uma concentração dos valores próximos a média. Embora não seja a técnica mais indicada para análise de pandemias considerando sua variação ao longo do tempo, nesse trabalho a distribuição normal de probabilidade foi utilizada para analisar o possível comportamento futuro baseada na repetição do comportamento passado.

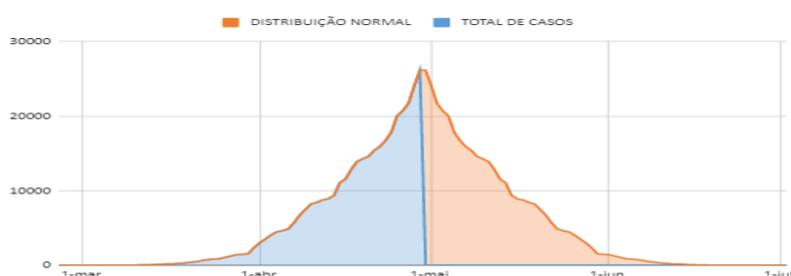


Figura 2: Exemplo de distribuição normal com seu formato peculiar em forma de sino.
Fonte: Autoria própria (2020).

A expectativa de apresentar o comportamento da distribuição normal apresenta coerência uma vez que após o alcance do pico de casos, existe a tendência de desaceleração, comparável à constância de crescimento anterior. No entanto, grande incerteza é inserida na análise em razão da não identificação do pico máximo dos casos. Para simular o melhor cenário, neste trabalho foi considerada a última data registrada da amostra como sendo o pico máximo, permitindo replicar o comportamento passado nos dias subsequentes. Ou seja, as análises deste trabalho assumem a premissa de que o COVID-19 começou a crescer e depois de atingir o seu pico, simulado pelo último dado utilizado na amostra, a tendência é que ele comece a declinar se aproximando da distribuição normal da variação dos dados.

Foi elaborado o gráfico de distribuição de correlação linear entre número de contaminados e quantidade de óbitos, pois ela representa a relação linear entre duas variáveis, dessa forma, elas foram manuseadas para representar os casos confirmados e óbitos por dia. A partir disso, foi analisado o gráfico de diagrama de dispersão, pois, através dele pode-se analisar visualmente se as variáveis possuem correlações positivas ou negativas. Nele observa-se uma alta correlação entre as variáveis de quantidade de óbitos e infectados, sendo elas, positivas, pois conforme uma variável cresce a outra também cresce. Encontrou-se também o coeficiente de determinação, o cálculo que indica o grau de dependência entre as variáveis estudadas.

3 Resultados

O gráfico da Figura 3 representa os dados reais referentes ao total de casos confirmados por dia, registrados no estado de São Paulo, segundo o site da SEADE⁴.

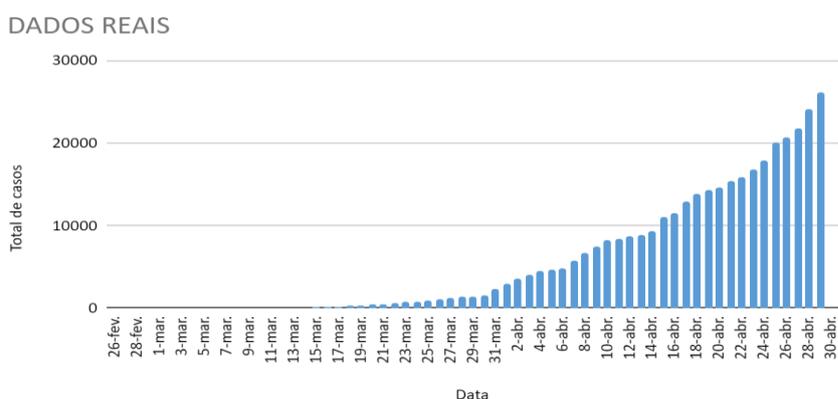


Figura 3: Quantidade de casos confirmados no período utilizado para análise.
Fonte: Autoria própria (2020).

Como podemos observar, o número de casos confirmados nos primeiros dias foi baixo, porém à medida que o tempo foi passando os números cresceram muito rápido, essa velocidade alta se tem, pois, o contágio da doença é muito acelerado, o que leva ao crescimento exponencial. Analisamos que este gráfico está caminhando para o modelo de distribuição normal, pois o aumento de um conjunto de dados que

segue eventos naturais ou aleatórios tende a formar uma distribuição normal em algum momento.

Assumindo as premissas descritas na seção de metodologia, ao replicar o comportamento passado para definir a projeção futura, conforme a Figura 4, é possível perceber que ele segue o padrão das distribuições normais uma vez que, se trata de um conjunto de dados que relaciona o crescimento da curva e seu declínio com o número de dados, ou seja, estamos considerando que o aumento nos casos de COVID-19, causará igualmente o seu declínio após atingir o pico. A grande incerteza é que não sabemos quando será o pico, e por essa razão, estamos assumindo o melhor caso, no qual o pico se apresenta na data do último registro.

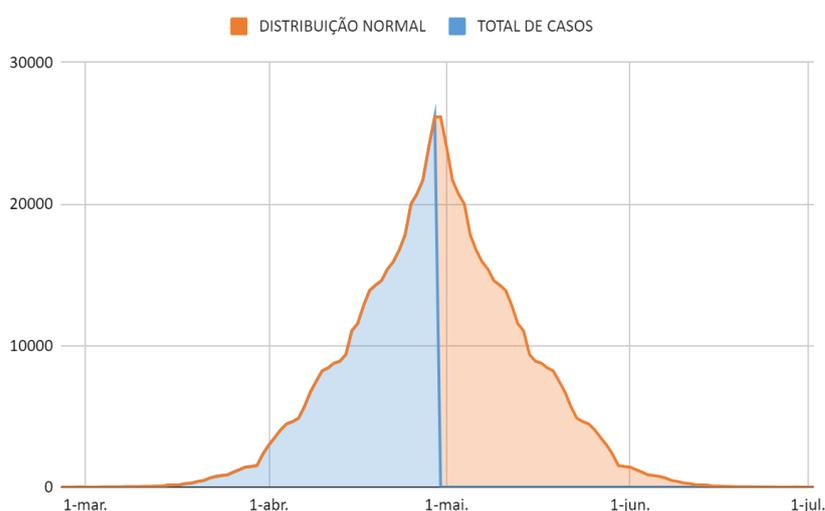


Figura 4: distribuição normal formada a partir da replicação dos dados passados de infectados para projetar o comportamento futuro, estabelecendo o pico como o último registro.

Fonte: Autoria própria (2020).

No gráfico apresentado na Figura 4, sobre os números de casos reais, é de suma importância para compreender a previsão proposta através da distribuição normal deste gráfico, pois naquele gráfico pode-se observar claramente o fluxo contínuo de aumento nos números de pessoas contaminadas pelo novo Coronavírus, ou seja, o aumento no número de dados possibilita, através da distribuição normal, visualizar as probabilidades de acontecimentos com esses dados.

Na Figura 5, o gráfico que corresponde aos óbitos por dia de acordo com as informações retiradas do site da Seade

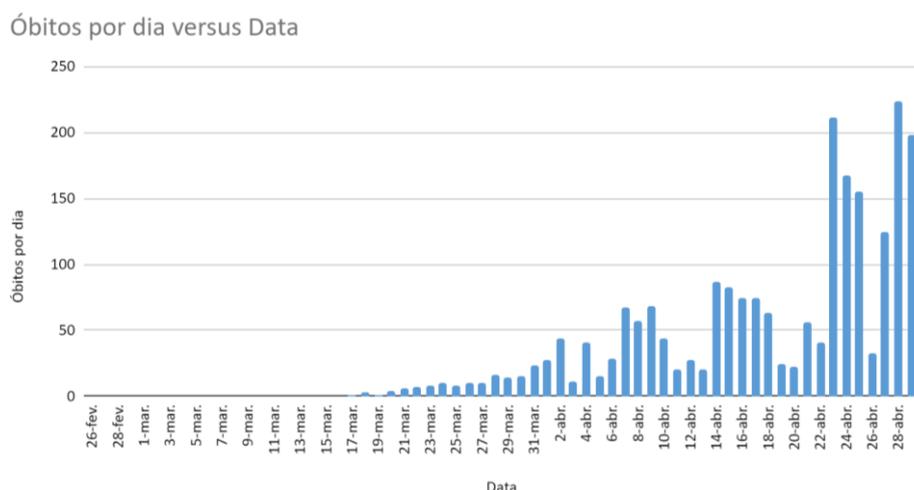


Figura 5: Distribuição dos casos de óbitos em razão do Coronavírus.
 Fonte: (<https://www.seade.gov.br/coronavirus/>).

O prefeito Bruno Covas afirma que, mesmo com todas as tentativas para evitar o aumento da doença, ainda não será o bastante para conter a evolução dos casos (FOLHA DE S.PAULO, 2020). Portanto, relacionando essa informação com os dados representados no gráfico é possível verificar a relação de casos diários confirmados com o avanço dos números de óbitos, ou seja, com isso a probabilidade de uma pessoa infectada vir a óbito aumenta e é provável também que o pico do COVID-19 seja maior do que o suportado pelo Sistema de Saúde.

Utilizando as mesmas premissas e replicação dos dados passados nos dados de óbitos, como exibido na Figura 6, assim como no total de casos confirmados, pode-se observar o padrão de distribuição normal e, a partir disso, explorar o resultado obtido.

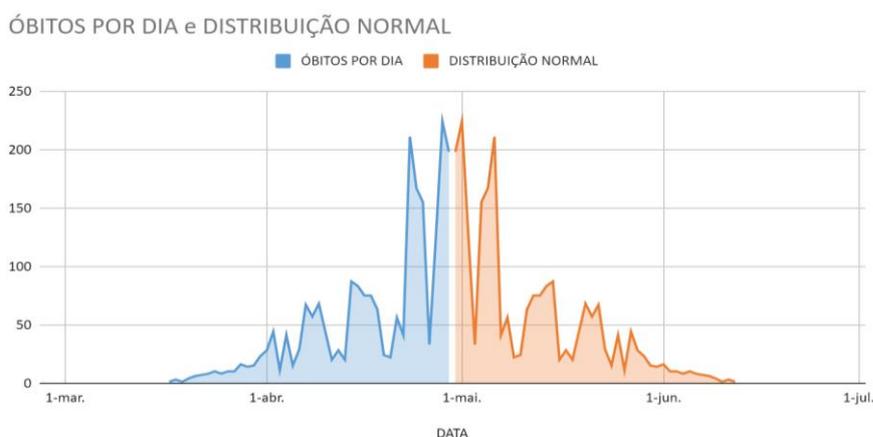


Figura 6: Distribuição normal formada a partir da replicação dos dados de óbitos passados para projetar o comportamento futuro, estabelecendo o pico como o último registro.
 Fonte: Autoria própria (2020).

Ao explorar os dados coletados é possível afirmar que os óbitos diários formam, a longo prazo, uma distribuição normal, mesmo que não de forma tão clara

como os dados de casos confirmados, uma vez que, os casos confirmados só estão avançando e a quantidade de morte por dia oscila bastante, mas ainda sim apresenta uma curva simétrica e um único pico da doença, o qual não se sabe quando acontecerá, então foi realizado uma previsão no período analisado, no caso, o pico simulado foi dia 1 de maio.

E, por último, a representação gráfica da distribuição de correlação linear, na qual foi analisado a possibilidade de relação entre as variáveis casos confirmados e óbitos por dia, conforme apresentado na Figura 7.

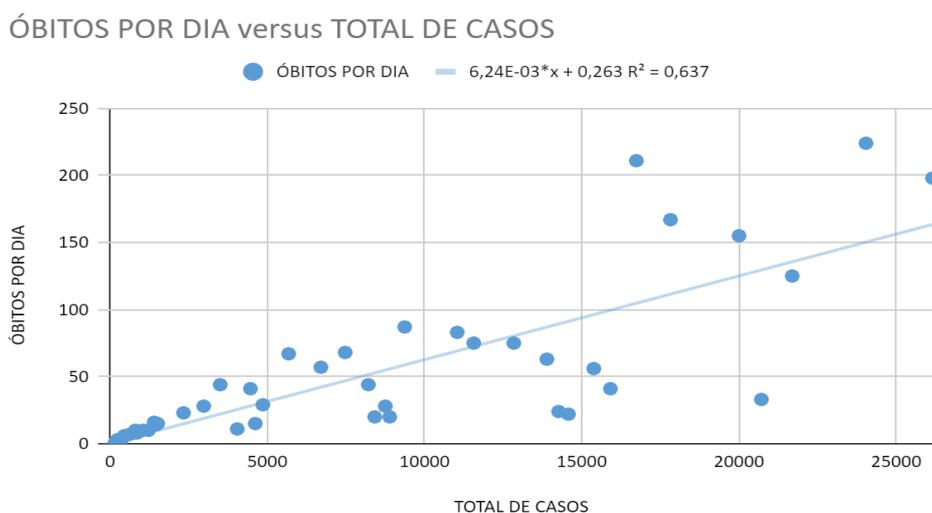


Figura 7: Diagrama de dispersão expressando a correlação entre a quantidade de infectados e óbitos. Fonte: Autoria própria (2020).

Conforme constatado na realização dos cálculos para a análise de correlação, a similaridade entre as duas variáveis é positiva, ou seja, a medida que uma aumenta a outra também aumenta, portanto tende a uma alta dependência. Em uma escala onde +1 indica correlação perfeita positiva e -1 indica correlação perfeita negativa, as variáveis analisadas apresentam um coeficiente de correlação de 0,637, ou seja, uma correlação moderada de 60% (MARKOSSIAN S, ET AL, 2004). Embora a reta não esteja, no período analisado, apresentando uma linha padronizada de grande correlação positiva, é importante ressaltar que os dados dispostos não são despadronizados, mas sim uma relação clara entre a quantidade de pessoas que são infectadas pelo COVID-19 e quantas destas vão a óbito.

Diante dos resultados obtidos é possível perceber a grande relação entre as variáveis e, portanto, o isolamento pode salvar vidas, corroborando com estudos previamente realizados, conforme pode ser visto na Figura 8 (UNICAMP, 2020). Tal

pandemia apresenta uma situação na qual praticamente toda população mundial é suscetível à contaminação (ZINCONE, 2020). Conforme discutido neste trabalho, uma porcentagem dos infectados são assintomáticos, o que dificulta o controle da disseminação da doença, ou seja, o isolamento social faz com que as pessoas tenham contato com menos pessoas, logo contaminam menos.

Total de vidas salvas por dia

	26/05	27/05	28/05	29/05	30/05	31/05	01/06	02/06	03/06	04/06	05/06	06/06	07/06	08/06
Brasil	374	496	638	803	995	1215	1469	1760	2095	2477	2915	3415	3984	4633
SP	133	177	228	288	358	440	534	644	771	918	1088	1286	1514	1776

Figura 8: Vidas salvas pelo isolamento social

Fonte: http://www.ime.unicamp.br/~pjssilva/vidas_salvas.html.

Dessa forma, o isolamento social possibilita o importante achatamento da curva, ou seja, garante que quando a doença atingir seu pico, não irá sobrecarregar o SUS. Embora existam muitas discussões à respeito das iniciativas de isolamento social aplicadas, na Figura 9 podemos perceber a relevância e o impacto positivo relacionado a tais medidas.

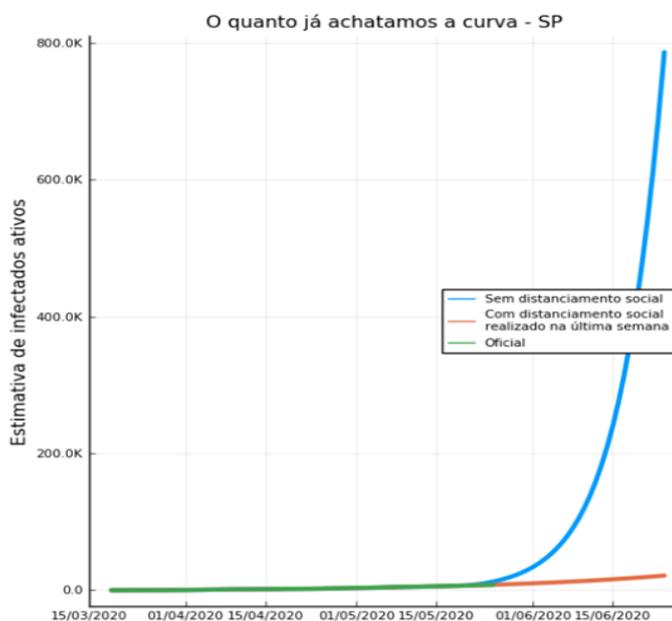


Figura 9: Comparação do comportamento histórico dos dados considerando o isolamento social e sem o mesmo

Fonte: http://www.ime.unicamp.br/~pjssilva/vidas_salvas.html.

3 CONCLUSÃO

Conclui-se que a utilização de medidas de posição, dispersão, distribuição normal e correlação linear são de extrema importância para a veracidade da análise feita, pois a primeira realiza uma análise levando em conta os pontos centrais; através da segunda é possível analisar os dados extremos; com base na terceira podemos estudar as probabilidades de acontecimentos e por fim, a correlação linear permite entender a relação entre uma variável e outra.

O desafio de combater o aumento de casos do novo Coronavírus é complexo pois depende de um Sistema de Saúde com investimentos reduzidos e da cooperação da população que pode ficar em casa, mas corre o risco de desenvolver aspectos negativos para sua saúde mental. Diante disso, o estudo realizado é relevante pois, através dele, é possível entender os números apresentados e afirmar que se a população que pode ficar em casa, continua a sair sem necessidade, a tendência é que os casos confirmados mantenham seu crescimento e, como consequência, o número de óbitos cresça também, pois são variáveis que apresentam uma correlação positiva alta.

No ano de 2015 o Brasil foi protagonista no combate ao Zika Vírus, dessa forma, é importante que o Brasil e mais especificamente o estado de São Paulo que apresenta os maiores índices do Brasil, se atente ao novo coronavírus, pois este acometeu outros países antes, ou seja, pode-se aprender com as informações geradas destes países. Por meio dos dados é possível perceber que a dimensão da doença é muito relevante, pois a medida que mais pessoas são infectadas pelo COVID-19, mais pessoas morrem, uma vez que não é possível atender à severidade das consequências em todas as pessoas infectadas e, muitas vezes, os sintomas demoram a aparecer.

Por fim, o uso da estatística foi importante para fazer análises temporais e perceber que, por ser uma doença facilmente transmitida e, muitas vezes, deixar seus microrganismos no ar por dia, pode piorar muito e gerar um cenário catastrófico até atingir seu auge.

REFERÊNCIAS

BRASIL DE FATO. **Artigo | O que sabemos sobre a covid-19 até agora e o que podemos prever.** Disponível em: <https://www.brasildefato.com.br/2020/05/25/artigo-o-que-sabemos-sobre-a-covid-19-ate-agora-e-o-que-podemos-prever>. Acesso em: Maio de 2020.

FOLHA DE S.PAULO. **Especialistas preveem mais de 100 mil mortes por coronavírus no Brasil.** Disponível em: <https://www1.folha.uol.com.br/equilibrioesaude/2020/05/especialistas-preveem-mais-de-100-mil-mortes-por-coronavirus-no-brasil.shtml>. Acesso em: maio de 2020.

FOLHA INFORMATIVA – COVID-19 (DOENÇA CAUSADA PELO NOVO CORONAVÍRUS). **Organização Pan-Americana da saúde.** Maio, 2020. Disponível em: <https://www.paho.org/pt/covid19>. Acesso em: 22 de maio de 2020.

G1. **Ministro da Saúde diz que isolamento é necessário 'porque você não sabe o que fazer'.** Disponível em: <https://g1.globo.com/politica/noticia/2020/04/29/ministro-da-saude-diz-que-isolamento-e-necessario-porque-voce-nao-sabe-o-que-fazer.ghtml>. Acesso em: 20 de agosto de 2020.

MARKOSSIAN S, SITTAMPALAM GS, GROSSMAN A, ET AL., EDITORS. **Bethesda (MD): Eli Lilly & Company and the National Center for Advancing Translational Sciences;** 2004-. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/>. Acesso em: 16 de setembro de 2020.

NATURE MEDICE. **He, X., Lau, E.H.Y., Wu, P. et al.** Temporal dynamics in viral shedding and transmissibility of COVID-19. 26, 672–675 (2020). Disponível em: <https://doi.org/10.1038/s41591-020-0869-5>. Acesso em: maio de 2020.

SEADE. **Governo do Estado de São Paulo.** Disponível em: <https://www.seade.gov.br/coronavirus/>. Acesso em: 22 de maio de 2020.

UNICAMP. **Vidas salvas pelo isolamento social.** Disponível em: <https://www.unicamp.br/unicamp/coronavirus/vidas-salvas-pelo-isolamento-social>. Acesso em: 26 de maio de 2020.