RELATÓRIO - ANÁLISE EXPLORATÓRIA DE DADOS - ESP1A5

Nome: Felipe Gustavo de Lima Santos **Prontuário:** SP3093875

1. INTRODUÇÃO

Para esse relatório, foram utilizados os dados de testes realizados (covid-testing-all-observations.csv) e os dados de casos registrados (total_cases.csv).

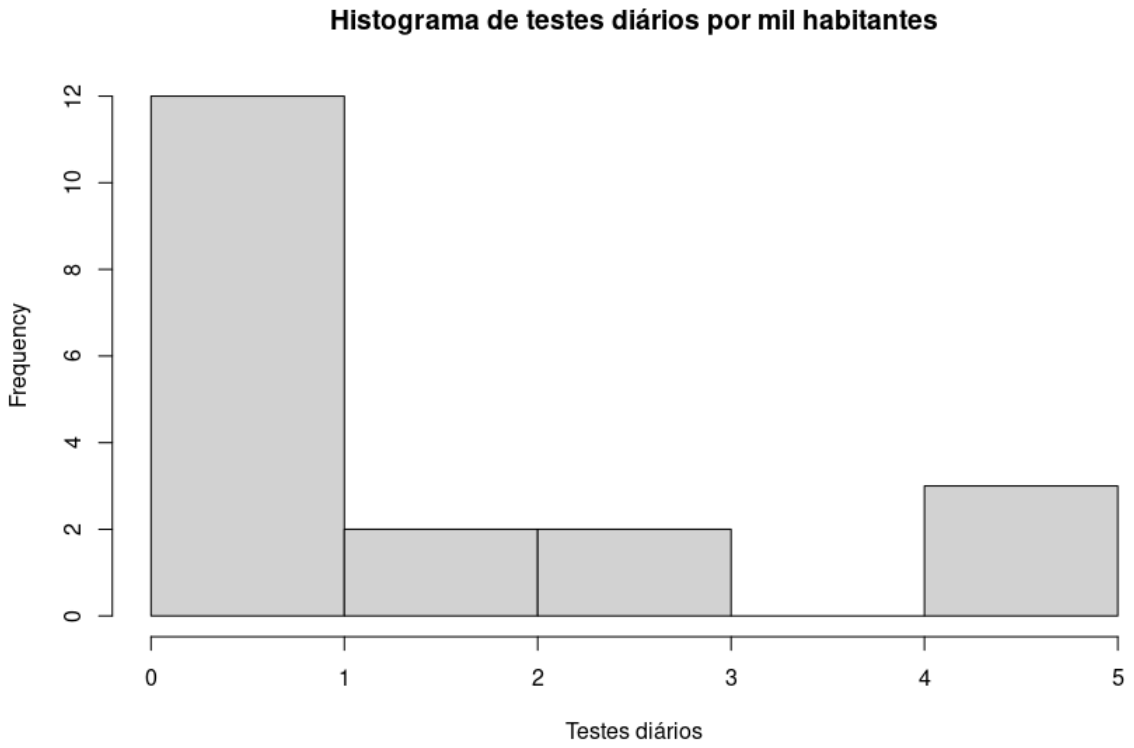
Na análise de estatística descritiva, foram utilizados dados de quatro dos cinco países mais populosos do mundo(Índia, Estados Unidos, Indonésia e Paquistão), o único dos cinco que não foi considerado foi a China pois não existem registros de testes. Já na análise de probabilidades e inferências, foram utilizados apenas os dados do Brasil. Essa escolha se deu pelo conteúdo das tabelas. Na tabela de testes, por exemplo, não há uma seção contendo os testes acumulados de todos os países presentes na tabela, além de não haver uma constância nas datas registradas entre cada país, dessa forma tornando-se complicado obter os dados mundiais dessa tabela. Por isso, filtrar apenas os dados referentes a apenas um país (nesse caso, o Brasil) se provou uma melhor abordagem.

2. ESTATÍSTICA DESCRITIVA

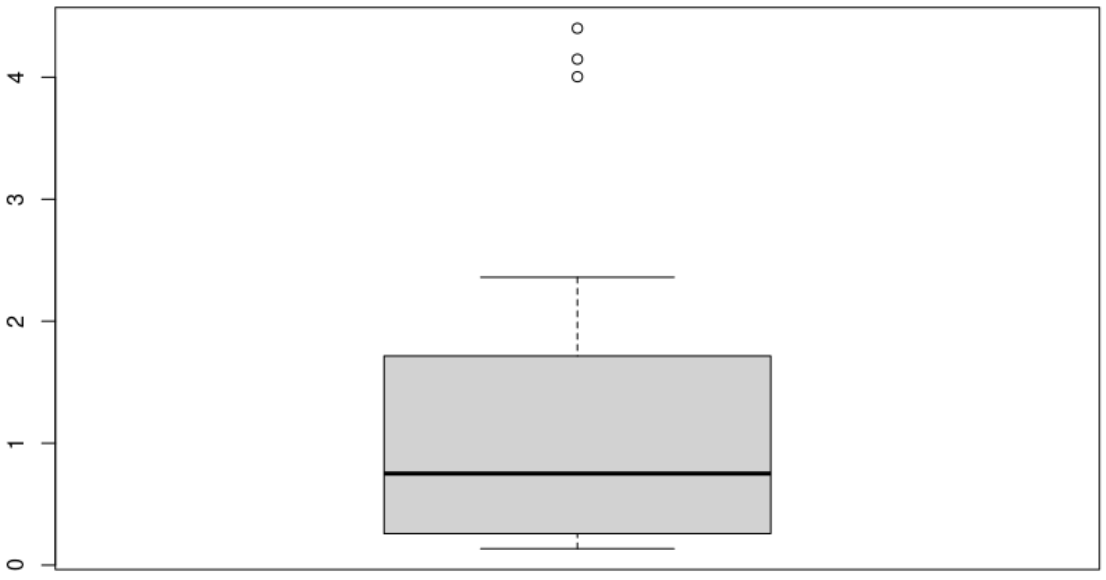
Para filtrar a tabela, foram considerados os dados de 3 em 3 meses, começando no dia 19 de março de 2021 e as análises (linha 31 do arquivo R) foram divididas entre a análise de testes diários entre os quatro países e as análises da quantidade de casos entre os quatro países.

2.1. TESTES DIÁRIOS

Utilizando algumas das funções mais básicas, foi encontrado os valores de média, mediana, 1,3046 e 0,7510, respectivamente. Após isso, foram gerados o histograma e o boxplot dessa variável.



Nesse histograma, observa-se que a realização de testes por dia se manteve, majoritariamente, entre 0 e 1 testes por mil habitantes.



O boxplot confirma a observação do histograma, mostrando também os três outliers, nos eventos nos quais houveram entre 4 e 5 testes realizados por mil habitantes.

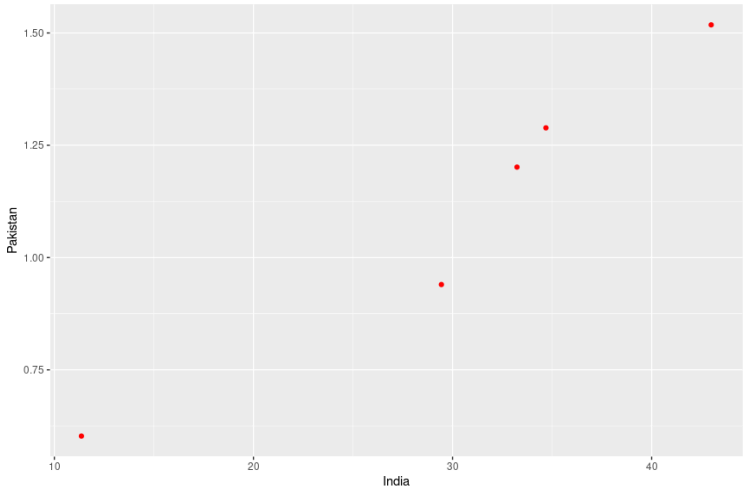
2.2. CASOS

Para a análise do número de casos (linha 38 do arquivo R), foi feito o cálculo da covariância e do coeficiente de correlação, além da geração do gráfico de covariância. Como esses cálculos são utilizados na análise bivariada, os quatro países foram agrupados em duas duplas, cada uma com seus cálculos próprios.

2.2.1. ÍNDIA X PAQUISTÃO

Para esse agrupamento, o cálculo da covariância resultou em 3,997563e+12 no método Pearson e 2,5 no método Spearman. O coeficiente de correlação no método Pearson foi de, aproximadamente, 0,972 e, no método Spearman, 1.

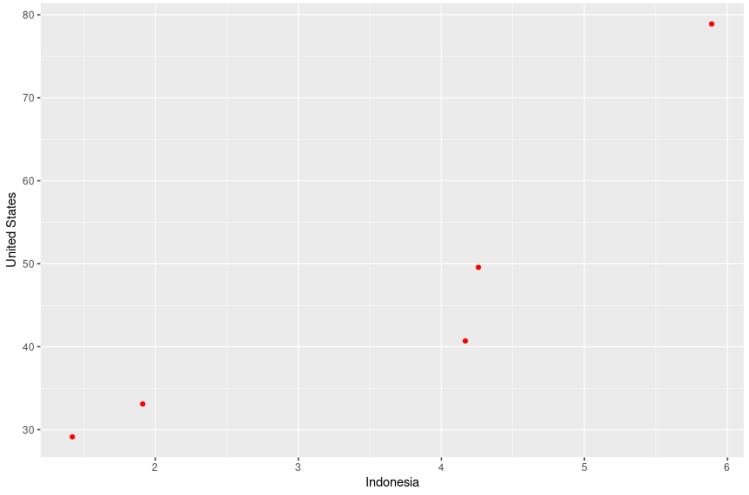
A seguir, o gráfico gerado, em milhões de casos:



2.2.2. INDONÉSIA X EUA

Para esse agrupamento, o cálculo da covariância resultou em 3,333704e+13 no método Pearson e 2,5 no método Spearman. O coeficiente de correlação no método Pearson foi de, aproximadamente, 0,912 e, no método Spearman, 1.

A seguir, o gráfico gerado, em milhões de casos:



3. PROBABILIDADE

Para esses cálculos, foi utilizado a data mais recente presente em ambas as listas, 11 de março de 2022, que indica 70.923.215 testes realizados e 28.973.799 casos registrados.

3.1. PROBABILIDADE CONDICIONAL

Para calcular a probabilidade condicional, foi primeiro obtido os valores das probabilidades de que um brasileiro tenha tido um caso de Covid-19 (levando-se em consideração que nenhum brasileiro tenha tido Covid-19 mais de uma vez) através do cálculo “número de casos / população brasileira”, e de que um brasileiro tenha sido testado para Covid-19 (levando-se em consideração que nenhum brasileiro tenha sido testado mais de uma vez) através do cálculo “número de testes / população brasileira”.

Considerando-se que a população brasileira segundo o Censo 2022 é de, aproximadamente, 215,3 milhões, a probabilidade de que um brasileiro tenha sido testado (P(A)) é igual a 0,329 e a probabilidade de que um brasileiro tenha tido um caso de Covid-19 (P(B)) é igual a 0,134. Além disso, segundo dados da Associação Brasileira de Redes de Farmácias e Drogarias (Abrafarma) reportados pelo jornal O Globo, a porcentagem de testes positivos em março de 2022 era de 7,28%, o que significa que a probabilidade de que um brasileiro que tenha sido testado teve um teste positivo (P(B|A)) é igual a 0,0728.

Aplicando o teorema de Bayes (linha 70 do arquivo R), foi encontrado o valor da probabilidade condicional reversa, isto é, a probabilidade de que um brasileiro que teve um caso de Covid-19 tenha sido testado. Esta probabilidade (P(A|B)) é igual a 0,178.

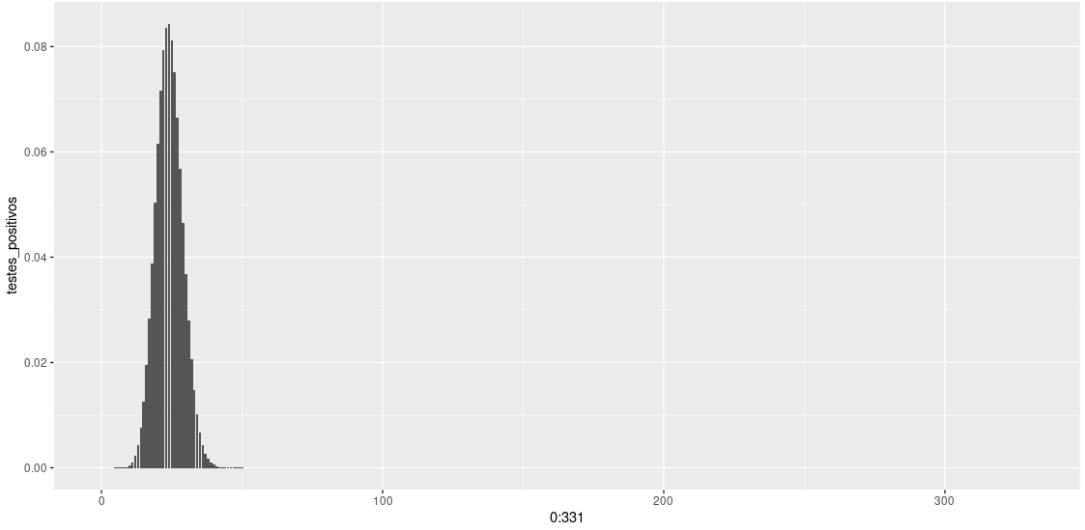
Esses resultados indicam que, como apenas cerca de 33% da população realizou testes de Covid-19, há uma alta probabilidade de que brasileiros que tiveram casos de Covid-19 não foram testados, o que se comprova dado que, considerando a taxa de testes positivos e a quantidade de testes realizados, a quantidade de testes positivos é igual a, aproximadamente, 5.163.210, o que representa cerca de 17,8% dos casos de Covid-19 registrados.

3.2. DISTRIBUIÇÃO BINOMIAL

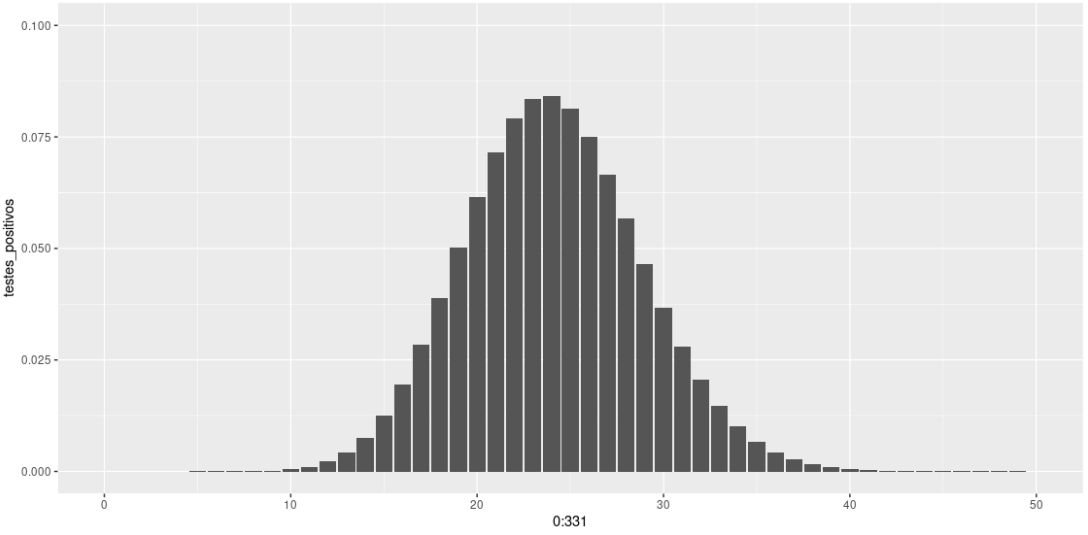
Para a distribuição binomial (linha 83 do arquivo R), foram calculadas todas as probabilidades de taxa de testes positivos. Para esse cálculo, foram utilizados os valores por mil habitantes arredondados, para que o cálculo fosse possível.

Primeiro foi obtida a probabilidade de que todos os casos de Covid-19 tenham sido testados (considerando a probabilidade de sucesso como a taxa de testes positivos). Nesse cálculo, o resultado obtido foi de 6,9567e-66, uma probabilidade extremamente pequena.

A seguir, foi gerado o gráfico de todas as probabilidades:



Nesse gráfico, é possível notar que a probabilidade se torna extremamente pequena após cerca de 50 sucessos e que a maior probabilidade (0,084245) é de que ocorram 24 de sucessos para cada mil brasileiros. Abaixo há um gráfico que mostra apenas os sucessos de 0 a 50 para melhor visualização.



3.3. DISTRIBUIÇÃO POISSON

Para os cálculos da distribuição Poisson (linha 97 do arquivo R), foi necessário primeiro calcular as médias de novos testes e casos por mês. Para isso, foi considerado o período completo em ambas as tabelas, de forma que, na tabela dos casos, foram considerados os dados após 11 de março de 2022. Os resultados obtidos foram de 3,526 novos casos por mil brasileiros por mês e, aproximadamente, 15,369 testes por mil brasileiros realizados por mês.

A seguir foram realizados os cálculos de diversas probabilidades:

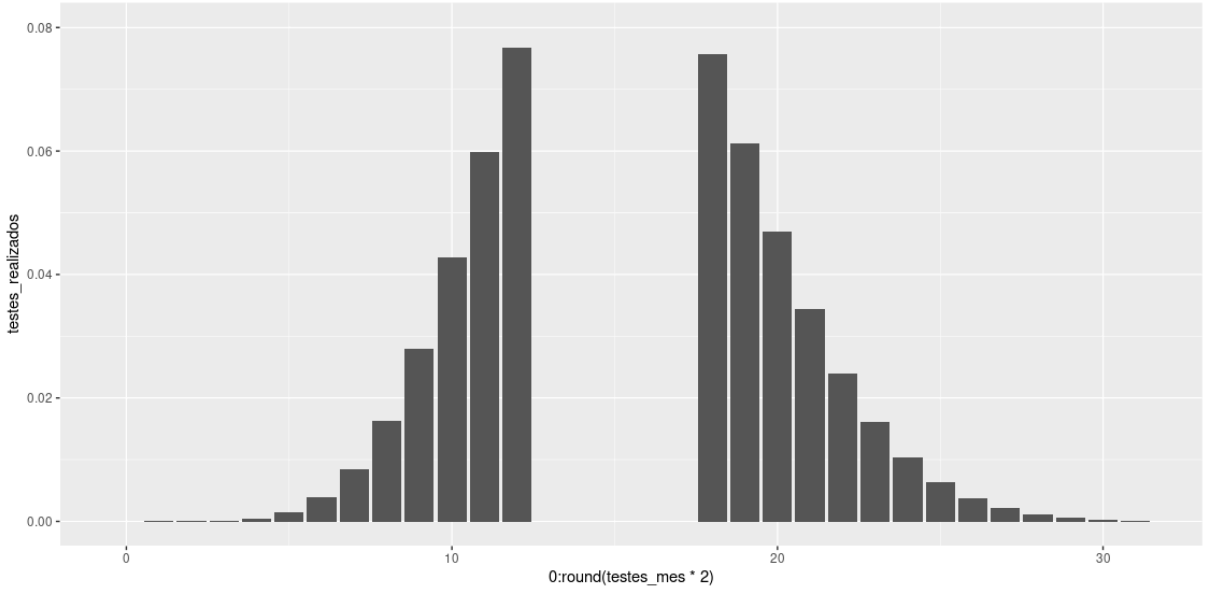
Dobro de testes	0,0001572
Metade dos testes	0,016327
Nenhum teste	2,1145e-07
Testes de jan/2022	0,0015111
Testes de 2021	0,019804

Dobro de casos	0,039588
Metade dos casos	0,18284
Nenhum caso	0,029399
Casos de jan/2022	6,1657e-05
Casos de 2021	4,1629e-05

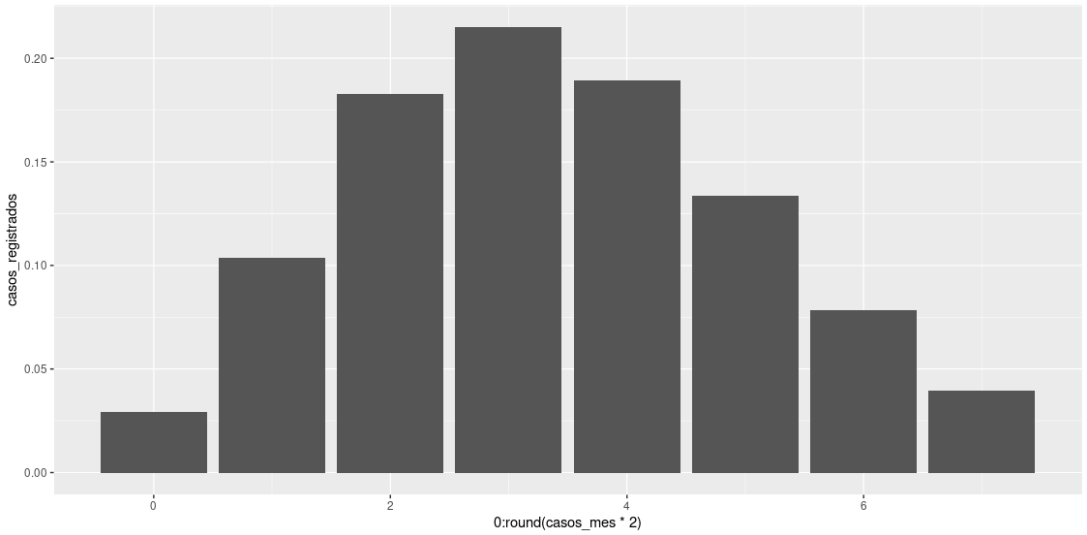
Obs: a quantidade de testes e casos por mil brasileiros em janeiro de 2022 foi, respectivamente, 5,15 e 13,18 e a quantidade de testes e casos por mil brasileiros em 2021 foi de, respectivamente, 171,94 e 69,47.

Também foram gerados gráficos com a probabilidade na distribuição Poisson.

Testes Realizados



Casos Registrados



3.4. DISTRIBUIÇÃO NORMAL

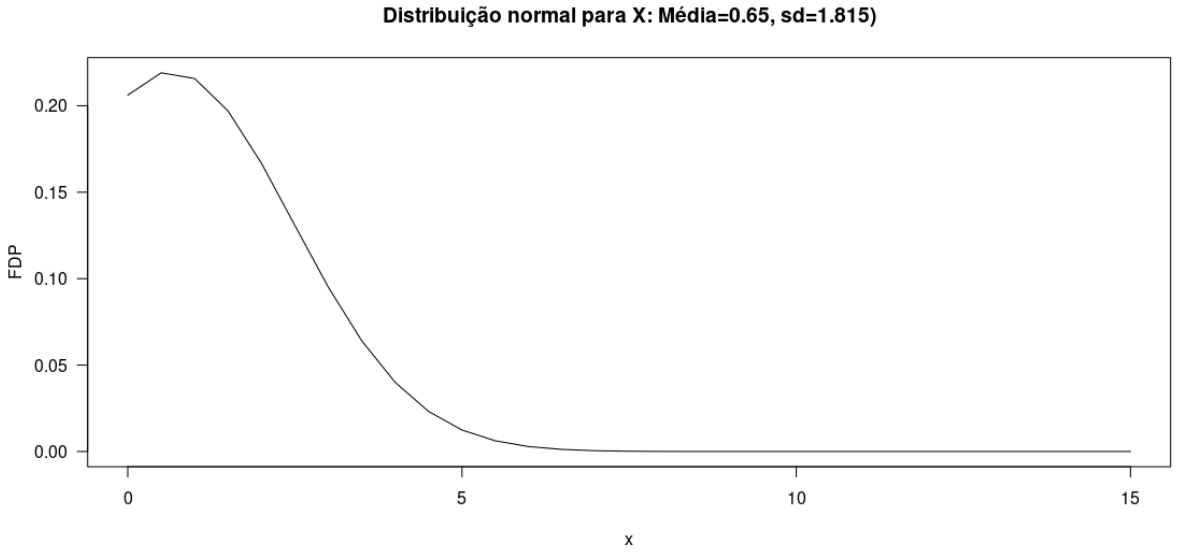
Para a distribuição normal (linha 137 do arquivo R), foram utilizados apenas os dados de novos testes por dia por mil habitantes, pois, diferentemente das variáveis de casos e testes totais, essa variável não utiliza valores acumulados.

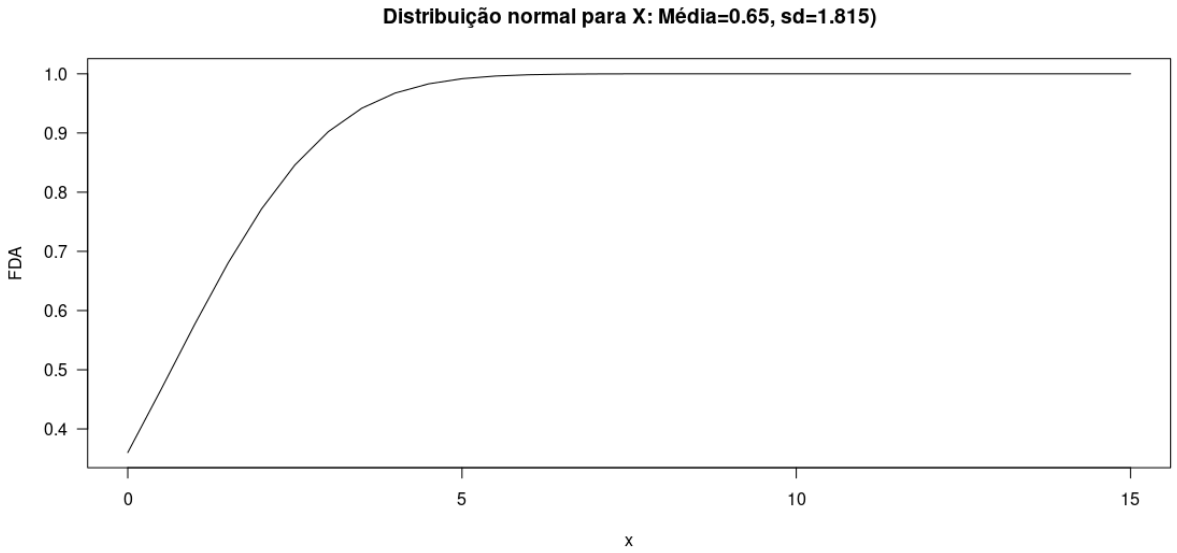
Utilizando os dados encontrados na análise de estatística descritiva, como desvio padrão, mediana e quartis, foi primeiro calculado diversas probabilidades relacionadas a posição de um evento no boxplot:

Abaixo da mediana	0,41708
Acima da mediana	0,58292
Entre os quartis	0,06677
Outlier	0,44196
Outlier extremo	0,34377

Obs: como foi observado anteriormente, não é possível que existam outliers para valores abaixo dos quartis. Dessa forma, os cálculos de outliers e outliers extremos consideraram apenas a probabilidade de que um outlier esteja acima do terceiro quartil.

Abaixo estão os gráficos da FDP e FDA dessa variável:





4. INFERÊNCIA

4.1. INTERVALOS DE CONFIANÇA

Foram calculados dois intervalos de confiança (linha 170 do arquivo R), para testes diários por mil habitantes do Brasil e dos quatro países utilizados na análise de estatística descritiva.

Para o intervalo do Brasil, foram utilizados os dados encontrados anteriormente, como a média (0,65) e o desvio padrão (1,815), também levando em consideração o número de amostras (calculado pelo número total de fileiras subtraído pelo número de valores nulos), que é igual a 265. Os resultados alcançados indicam o intervalo de confiança entre 0,431 e 0,868.

Já no intervalo dos quatro países, não foi utilizada a tabela filtrada de 3 em 3 meses, mas sim todos os registros desses países. Com isso, foram encontrados os valores 1,375 para a média e 1,729 para o desvio padrão, em uma amostra de 2.878 registros. O intervalo de confiança nesse caso foi entre 1,312 e 1,438.

4.2. TESTES DE HIPÓTESE

Para o teste de hipótese (linha 185 do arquivo R), foi testada a relação entre a quantidade de testes diários por mil habitantes do Brasil e dos quatro países.

Realizando o teste de Shapiro-Wilk, foi constatado que ambas as distribuições são normais, visto que seu p-value é menor que 0,05. Ao realizar o teste de variância, no entanto, o p-value observado é maior que 0,05, tornando o H0 falso para a variância.

Por fim, ao comparar as médias, o p-value foi novamente menor que 0,05, indicando que não há diferença estatisticamente significante entre as médias de testes diários do Brasil e dos quatro países.

5. REFERÊNCIAS BIBLIOGRÁFICAS

YONESHIGUE, Bernardo. Covid-19: taxa de testes positivos aumenta 28% no Brasil. **O Globo**, 2022. Disponível em : https://oglobo.globo.com/saude/medicina/noticia/2022/04/covid-19-taxa-de-positivos-em-testes-de-farmacia-crescem-28percent-no-brasil.ghtml. Acesso em 08 jun 2024