Projeto de Trabalho Final

ME323 | Turma D | 1°S de 2021

Análise de dados de COVID-19 em R

Distribuição de Mortalidade por Município

Nome / RA: Daniel Credico de Coimbra (155077)

Nosso objetivo é utilizar a linguagem R para aplicar a maior parte dos conceitos e das ferramentas estatísticas introduzidas na disciplina ME323. Para tal, começaremos por uma análise *descritiva* de um conjunto de dados e, então, exercitaremos em cima desses dados as ferramentas de análise *inferencial* fornecidas no curso.

O objeto de estudo escolhido é a demografia da epidemia de SARS-CoV-19 no Brasil. Nossos dados sobre esse tema serão obtidos a partir do website https://covid.saude.gov.br/, que fornece logo na página principal a opção de *download* de **três arquivos .csv** com dados relevantes.

Interface pública para download de nossos dados

Esses arquivos .csv contêm informações diárias de infecção, mortalidade, e recuperação em cada um dos 5.570 municípios brasileiros. Esses mesmos arquivos também agregam os dados desse municípios em 27 unidades federativas (os “estados”), 5 macrorregiões, e o país como um todo. O primeiro arquivo .csv trata do 1°S/2020, o segundo do 2°S/2020, e o terceiro do 1°S/2021.

Nossa análise será apresentada em um **Jupyter Notebook** utilizando a linguagem R, a ser realizada nas seguintes etapas.

1. Importaremos os três arquivos .csv no ambiente Jupyter em R.
2. Determinaremos o **tipo** de nossos dados: quantativos, qualitativos, etc.
3. Criaremos uma **pivot table** unificando as três bases de dados, reformatando os dados para que as linhas correspondam a cada um dos municípios, as colunas correspondam às datas, e os valores correspondam à taxa de mortalidade acumulada (mortes / população) para aquele município naquela data.
4. Fixada uma data, tal taxa é uma **variável aleatória discreta**. Estudaremos sua distribuição ao longo do país, analisando suas medidas de posição e de dispersão, além de visulizá-la em gráficos como **histogramas** e **boxplots**. Pode-se então avaliar municípios *outliers*.
5. Ao visualizar a distribuição dessa variável aleatória, algumas discussões podem ocorrer.
    i. Pode-se entender cada caso de COVID como um **ensaio de Bernoulli**, com uma probabilidade *p* de óbito. Pode-se **estimar o valor de *p*** com um intervalo de confiança.
    ii. Com *n* casos da doença e *m* mortes, pode-se definir uma **variável binomial** com parâmetros ⟨n, m, p⟩. Pode-se então **aproximar essa distribuição binomial a uma distribuição normal**, e averiguar se a distribuição aproximada resultante se assemelha à distribuição original. Ou seja, pode-se avaliar a normalidade da distribuição original.
6. Para fins de exercício, poderíamos realizar **amostragens aleatórias** dos os 5.570 municípios para realizar várias **estimavas** a média e a variância da taxa de mortalidade, usando o **Teorema Central do Limite** para construir um **intervalo de confiança**. Então comparamos nossas estimativas com os valores reais.

7. Por fim, como tal variável aleatória é indexada a uma data, e como nossa análise descritiva será automatizada, será possível estudar a **evolução temporal** das propriedades estatísticas dessa variável aleatória ao longo do tempo.

Além do trabalho estatístico que se pode fazer com os dados elencados acima, existem dois outros trabalhos que poderiam ser realizados mediante a obtenção de novos dados. Estes são:

8. Possibilidade: Se for possível obter um banco de dados com as coordenadas em Latitude e Longitude de cada município brasileiro, então também será possível examinar a **dispersão geográfica** da pandemia ao longo do tempo.
9. Possibilidade: Se for possível obter um banco de dados com índices municipais, tais como saneamento e renda média, então também será possível levantar hipóteses sobre a **correlação** da taxa de mortalidade de cada município com outros índices. Com isso, poderíamos desenvolver **probabilidades condicionais** de mortalidade. Ademais, um **teste de hipótese** poderia ser realizado, tomando como hipótese nula que a correlação foi gerada aleatoriamente (isto é, trata-se de uma correlação espúria).