ANADI - Análise de Dados em Informática - Trabalho prático 1

André Novo
3DE
ISEP
1181628@isep.ipp.pt

Diogo Ribeiro
3DE
ISEP
1180782@isep.ipp.pt

Tiago Ribeiro
3DE
ISEP
1181444@isep.ipp.pt

***Abstract*—Este artigo científico serve para documentar e desenvolver processos de Análise Exploratória de Dados, Inferência Estatística, Correlação e Regressão de um determinado conjunto de dados facultados. O artigo é constituído por uma introdução, onde são descritos os objetivos do trabalho, propósito do mesmo, assim como uma curta contextualização do tema abordado e um enquadramento teórico das várias áreas em estudo. De seguida, são expostos os problemas que nos foram fornecidos para cada uma das secções e propostas de resolução aos mesmos, apresentando algumas conclusões que podem ser retiradas dessas mesmas resoluções. Por último, é apresentada uma breve conclusão do trabalho desenvolvido.**

***Index Terms*—análise, dados, correlação, regressão, inferência estatística**

I. Introdução

O covid-19 é uma doença infeciosa que abalou o mundo e teve um impacto significativo nas nossas vidas, tendo originado um grande número de infetados, de mortos, entre outros. Para este artigo científico, iremos utilizar um conjunto de dadosTodos esses dados provenientes do impacto do vírus nos vários países e continentes do mundo.

Este artigo científico tem como finalidade a realização de uma análise de um conjunto de dados que nos foi fornecido através de um ficheiro contendo dados reais sobre o Covid-19, relativos ao período compreendido entre os dias 01/01/2020 e 27/02/2021. Esse ficheiro foi retirado da base de dados internacionais "Our World In Data" [1]. Para além da análise, também tem como finalidade a execução de outros processos relativos a inferências estatísticas, correlações e regressões desses mesmos dados, assim como as respetivas conclusões que podem ser retiradas.

O desenvolvimento dos processos acima mencionados é feito através da resolução de um conjunto de exercícios fornecidos pelos docentes desta unidade curricular, sendo, por isso, o propósito deste artigo puramente académico.

Por forma a executar todas as atividades pedidas, é feito um uso à linguagem de alto nível R, que possui um ambiente vocacionado para a análise de dados e para a componente gráfica [2].

Todos os processos pedidos incidem num determinado conjunto de conteúdos teóricos descritos abaixo. Neste artigo, não são mencionados todos os conteúdos, apenas os mais importantes para o entendimento do mesmo.

A. Análise Exploratória de Dados

A análise exploratória de dados permite representar a informação contida num determinado conjunto de dados de uma forma organizada, através de gráficos, tabelas de frequência, medidas de tendência central (como média, mediana e moda), medidas de tendência não central (como quartis e percentis), medidas de dispersão (como desvio padrão, variância e amplitude interquartil), entre outros [3].

A média consiste no quociente entre a soma de todos os valores observados e o número de observações.

Um boxplot, ou caixa de bigodes, é uma ferramenta gráfica que permite visualizar a distribuição de um determinado conjunto de dados. É constituído por 3 quartis, sendo que um quartil corresponde a um quantil de ordem 0.25, 0.50 ou 0.75. Um quantil corresponde ao valor que separa os 100q% valores menores de uma determinada amostra dos 100(1-q)% valores maiores [4].

Dito de uma outra forma, os quartis definem valores aos quais x% da amostra é igual ou inferior a esse valor. Para o 1º quartil, x corresponde a 25, para o 2º 50 (equivalendo, assim, à mediana, devido ao facto de 50% da amostra estar acima desse valor e 50% estar abaixo dele), e finalmente para o 3º, 75. Ainda é constituído por limites inferior e superior. O limite inferior e superior são calculados da seguinte forma [5]:

$$LimiteInferior = 1Quartil – 1,5 * (3Quartil – 1Quartil) \quad (1)$$

$$LimiteSuperior = 3Quartil + 1,5 * (3Quartil – 1Quartil) \quad (2)$$

A diferença entre o 3º quartil e o 1º quartil também é chamada de amplitude interquartil [4].

B. Inferência Estatística

A inferência estatística preocupa-se com o raciocínio necessário para que, a partir dos dados, se consiga obter conclusões gerais. Isto é, o seu objetivo é obter uma afirmação acerca de uma população com base numa amostra, que consiste numa parte de dados retirada dessa mesma população. Um dos tipos de inferência diz respeito a testes de hipóteses [6].

Os testes de hipóteses permitem responder se uma determinada afirmação acerca de uma população é verdadeira ou falsa, recorrendo à informação contida em amostras aleatórias.

Dado um parâmetro desconhecido θ de uma população e um valor fixo $\theta 0$ iremos considerar os seguintes três testes de hipótese:

Caso	H_0	H_1	
(a)	$\theta = \theta_0$	$\theta \neq \theta_0$	teste bilateral
(b)	$\theta \geq \theta_0$	$\theta < \theta_0$	teste unilateral (à esquerda)
(c)	$\theta \leq \theta_0$	$\theta > \theta_0$	teste unilateral (à direita)

Tabela retirada de *Aulas Teóricas - Testes de Hipóteses Paramétricos*

Onde, H0 é a hipótese nula e H1 á hipótese alternativa.

Podemos decidir se rejeitamos H0, ou se não o rejeitamos usando o valor de prova (p-value). Pode-se interpretar o p-value como sendo o menor nível de significância para o qual a H0 é rejeitada para o valor observado da estatística teste. Caso p-value $< \alpha$ (nível de significância), rejeita-se H0. Caso contrário, não se rejeita [7].

Os testes de hipóteses podem ser de dois tipos: paramétricos e não paramétricos.

Os testes paramétricos assumem uma distribuição normal dos valores de uma amostra, sendo feitos para dados idealizados. Já os não paramétricos são usados para situações em que os paramétricos não são apropriados, como quando a distribuição não é normal, ou é desconhecida ou até mesmo quando o tamanho da amostra é pequeno (isto é, quando o tamanho é menor de 30) para que se possa assumir uma distribuição normal. Ao contrário dos anteriores, estes testes são desenhados para dados reais, ou seja, dados com algumas irregularidades, valores muito discrepantes e com algumas lacunas [8].

O teste de Shapiro-Wilk é um teste não paramétrico que serve para verificar se uma variável aleatória X, retirada de uma amostra aleatória, segue uma distribuição normal, onde temos as seguintes hipóteses:

H0 : X segue uma distribuição normal vs H1 : X não segue uma distribuição normal

O teste de Lilliefors é um teste não paramétrico que serve como uma correção a um outro teste não paramétrico, que tem por nome Kolmogorov-Smirnov (K-S), de modo a testar, independentemente dos valores da média e do desvio padrão (considerados para o teste de K-S), se a amostra é proveniente ou não de uma distribuição normal.

Ambos os testes de Shapiro-Wilk e Lilliefors são idênticos, no entanto geralmente efetua-se o de Lilliefors para amostras consideradas grandes (isto é, com tamanho igual ou superior a 30), enquanto que o de Shapiro-Wilk é usado para amostras mais pequenas [9].

O teste One-Way-ANOVA (Análise de variância com um factor) é usado para comparar médias de duas ou mais amostras independentes e testar se existem diferenças significativas entre essas amostras. O teste possui 2 variáveis: a variável independente, que é categórica e define os grupos (ou fatores) que serão comparados, e a variável dependente, que é uma variável numérica, cujas médias serão comparadas.

Este teste possui vários pressupostos, nomeadamentes os seguintes: a variável dependente deve ser contínua e normalmente distribuída para cada grupo (podendo ser usado o teste de Shapiro-Wilk referido anteriormente), a independente deve ter 2 ou mais grupos independentes, as observações devem ser independentes, isto é, não devem existir relações entre as observações de grupos diferentes, e não devem conter outliers significativos, e não deve existir homogeneidade de variâncias.

Este último ponto pode ser verificado recorrendo ao teste de Levene. Este teste é usado para verificar se k populações têm variâncias iguais [7].

O teste de Kruskal-Wallis é um teste que serve como alternativa não paramétrica do teste One-Way ANOVA. Deve ser utilizado quando a hipótese de normalidade for rejeitada ou se o tamanho das amostras a serem utilizadas for pequeno (tamanho inferior a 30) [9].

Quando são executados testes para verificar se existem ou não diferenças entre grupos, por vezes pode ser necessário obter mais informações relativamente a essas diferenças, uma vez que esses testes só informam se existem ou não diferenças. Nestes casos, utilizam-se testes post-hoc.

Os testes para comparar médias entre duas amostras independentes permitem comparar médias. Neste tipo de testes, as variâncias podem ser conhecidas, desconhecidas, mas assumidas como iguais ou desconhecidas, mas assumidas como diferentes (teste de Welsh). Por consequência do Teorema do limite central, poderá se supor, ao usar o teste, que ou as amostras têm distribuição normal ousão grandes, isto é, o seu tamanho é igual ou maior que 30 [7].

C. Correlação

Podemos definir correlação como sendo uma relação entre dois termos, seja através do crescimento dos mesmos como o contrário. Dito isto, os coeficientes de correlação são métodos estatísticos para se medir as relações entre variáveis e o que elas representam. Sendo os principais: o de Pearson, o de Spearman e o de Kendall.

O coeficiente de correlação de Pearson (r), também chamado de correlação linear, é um grau de relação entre duas variáveis contínuas e exprime o coeficiente de correlação (r) através de valores situados entre -1 e 1. Quando o coeficiente de correlação é próximo de 1, podemos afirmar que se observa um aumento no valor de uma variável quando a outra também aumenta, isto é, há uma relação linear positiva. Quando o coeficiente é próximo de -1, também é possível dizer que as variáveis são correlacionadas, mas neste caso, quando o valor de uma das variáveis aumenta, o valor da outra diminui. Isso é o que é chamado de correlação negativa ou inversa. Um coeficiente de correlação próximo de zero indica que não existe uma relação entre as duas variáveis, e quanto mais o r se aproximam de 1 ou -1, mais forte é a relação entre as variáveis.

O coeficiente de correlação de Spearman (ρ), é usado quando as variáveis que se pretendem estudar são ambas ordinais ou quando uma das mesmas é contínua e a outra é ordinal. A avaliação do nível de correlação efetua-se da mesma forma que no de Pearson ($-1 \leq \rho \leq 1$).

O coeficiente de correlação de Kendall (τ), é uma alternativa ao coeficiente de correlação de Spearman, sobretudo, quando as amostras são pequenas e/ou existem muitos empates. Este fator baseia-se no número de pares de observações concordantes (consistentes) e no número de pares de observações discordantes (inconsistentes). Dados dois pares de observações, podemos dizer que são: concordantes, discordantes ou empatados [10].

D. Regressão

Podemos afirmar que a análise de uma regressão linear é usada para explicar a relação entre uma variável aleatória e um conjunto de variáveis não aleatórias. Dizemos que a variável aleatória é dependente e aquelas que pertencem ao conjunto das não aleatórias, independentes. Ora, quando temos apenas uma variável independente, a regressão linear em causa assume a nomenclatura de simples, em contrapartida caso haja mais do que uma variável preditora (não aleatória), estamos na presença de uma regressão múltipla. Ora, acrescentando a isto, a variável aleatória deve ser sempre contínua enquanto que as variáveis não aleatórias podem ser contínuas, discretas ou categóricas. Relativamente ao resíduos de uma regressão linear, existem alguns testes importantes, tais como: o teste de Shapiro, o teste de Durbin-Watson e o teste de Breusch-Pagan.

O teste de Shapiro ajuda-nos a avaliar a normalidade dos resíduos.

O teste de Durbin-Watson serve para verificar a independência dos resíduos, ou seja, calcula a existência de autocorrelações residuais, bem como o p-value inerente.

Usa-se o teste de Breusch-Pagan para averiguar a existência de homocedasticidade num conjunto de dados [11].

E. Nota

O ficheiro a ser utilizado contém diversos campos relativos a determinados dados. Um dos campos, designado por location indica o país, no entanto, também indica continentes, pelo que, por vezes, esse campo é utilizado quando é necessário extrair informações de continentes.

II. Análise Exploratória de Dados

A. Gráfico que mostra o número total de infetados ao longo do período de tempo, por continente

Partindo da exportação dos dados do ficheiro a ser utilizado, procedeu-se à extração das colunas necessárias para a realização do exercício, nomeadamente as colunas referentes aos campos LOCATION, TOTAL_CASES, e DATE, sendo cada um deles, respetivamente, localização geográfica, podendo este campo se referir a um país ou continente, número total de infetados diários e data. Esta extração originou a criação de uma nova estrutura de dados, contendo apenas as colunas referidas.

De seguida, procedeu-se à criação de várias estruturas, partindo da anteriormente criada, para cada continente, contendo os dados referentes ao continente em si, utilizando-se o campo LOCATION.

Com as estruturas criadas, foi construído um gráfico com as estruturas anteriores, sendo a variável Y referente ao número total de infetados, e a variável X ao intervalo de tempo analisado, desde a apariação do primeiro caso mundialmente registado até ao último registo presente no dataset fornecido. Cada linha é referente a um continente, com uma cor única e uma legenda a identificar. O gráfico obtido foi o seguinte, com recurso ao uso das funções ggplot() e $\text{geom}_line(), ambasexportadasdabibliotecaggplot2$:

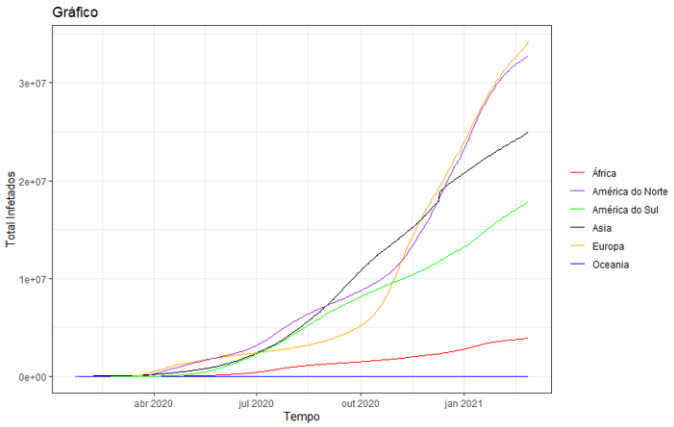

Fig. 1. Gráfico com número total de infetados ao longo do tempo, por continente.

Observando o gráfico, podemos concluir que a Europa registou, no início de 2021, o maior número total de infetados entre os outros continentes, enquanto que a Oceania, em termos gerais, foi o continente com menor número de casos totais desde o início da pandemia.

B. Gráfico do total de infetados por milhão de habitantes, ao longo do período de tempo, por continente

Para a realização deste exercício, partimos do uso do mesmo código usado no ponto acima referido, fazendo apenas uma alteração na exportação das colunas do ficheiro de dados. Em vez de ter sido feita a exportação do campo TOTAL_CASES (número total de casos), foi feita a exportação do campo TOTAL_CASES_PER_MILLION (número total de casos por milhão de habitantes), como requisitado pelo enunciado.

Com recurso às mesmas funções mencionadas na alínea A, obtivemos o seguinte gráfico:

Analisando o mesmo, podemos constatar que a América do Norte apresenta, desde finais de 2020, um maior número de casos por milhão de habitantes do que os restantes continentes. Assim como no gráfico anterior, a Oceania registou os valores mais baixos, de forma geral.

*C. Um boxplot do número de mortos diários por milhão de habitantes para cada um dos seguintes países: Portugal, Espanha, Itália e Reino Unido. Remova os outliers usando o critério: x é outlier sse x não pertencer a [Q1-1.5*IQR , Q3+1.5*IQR]*

Por forma a conseguirmos resolver o problema proposto, foram criadas numa fase inicial, uma estrutura com os campos LOCATION e NEW_DEATHS_PER_MILLION, respetivamente, localização e novos mortos por milhão de habitantes.

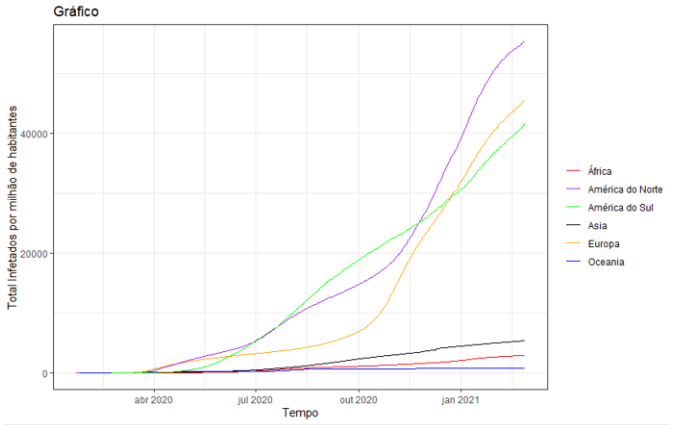

Fig. 2. Gráfico com número total de infetados por milhão de habitantes ao longo do tempo, por continente.

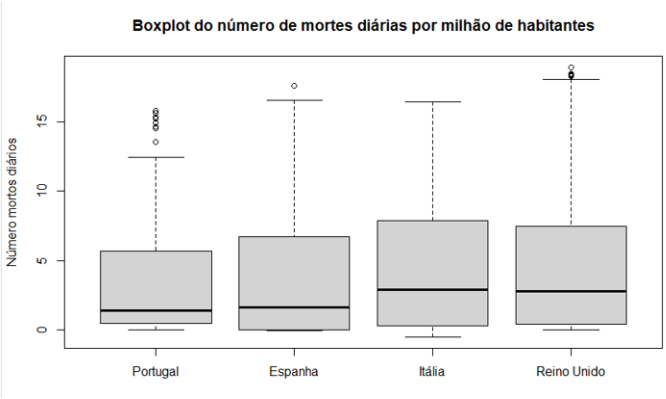

Fig. 3. Boxplot do número de mortos diários por milhão de habitantes para cada um dos países.

Em seguida, para cada país da lista pedida, foi criada uma estrutura de dados apenas desse país e um vetor apenas com os dados relativos ao número de mortos diários por milhão de habitantes.

Partindo dos dados relativos a esse país, foram determinados o 1º e 3º quartis, usando as percentagens de 25 e 75%, respetivamente. Também é calculada a amplitude interquartil dos dados. Com os 3 valores, procedeu-se ao cálculo dos limites superior e inferior. A tabela seguinte mostra os resultados obtidos destes 5 valores para cada um dos países:

TABLE I
TABELA COM DADOS RELATIVOS AO BOXPLOT

Paises	1ºQ	3ºQ	AIQ	LS	LI
Portugal	0.563	6.792	6.228	16.1	-8.78
Espanha	0	7.13	7.133	17.8	-10.7
Itália	0.281	7.856	7.575	19.2	-11.1
Reino Unido	0.442	7.851	7.409	19	-10.7

Tendo os limites determinados, eliminámos os outliers de acordo com o critério exigido.

Finalmente, com os outliers removidos de cada estrutura, obtivemos o seguinte boxplot:

D. Um gráfico de barras com o número total de mortos, por milhão de habitantes, e o nº de testes diários por milhar de habitantes, para os países: Albânia, Dinamarca, Alemanha e Rússia

Em relação ao enunciado supracitado, o enunciado possui um erro, sendo que, em vez de ser nº de testes diários por milhar de habitantes, o correto é nº total de testes por milhar de habitantes.

Num primeiro momento, duas estruturas foram criadas, sendo que ambas possuem o campo LOCATION (localização), no entanto, uma delas contém o campo TOTAL_DEATHS_PER_MILLION, já a outra o campo TOTAL_TESTS_PER_THOUSAND, correspondendo a número total de mortos por milhão e número total de testes por milhar, respetivamente.

Para cada uma das estruturas criadas, foi criada uma outra contendo os dados relativos apenas aos países exigidos pelo enunciado. Em seguida, partindo da anterior, foi criada uma nova com o valor máximo dos valores totais por cada país. Por fim, esses valores máximos foram transferidos para um vetor numérico.

Com ambos os vetores, procedeu-se, então, á construção de um novo data frame. Para efeitos de visualização dos resultados no gráfico de barras, a estrutura foi convertida numa matrix e, em seguida, essa mesma matrix foi transposta, ou seja, uma matrix obtida através da troca entre as linhas e colunas da original [9].

Por último, recorrendo ao uso dessa matrix, obtivemos o seguinte gráfico de barras, utilizando a função barplot():

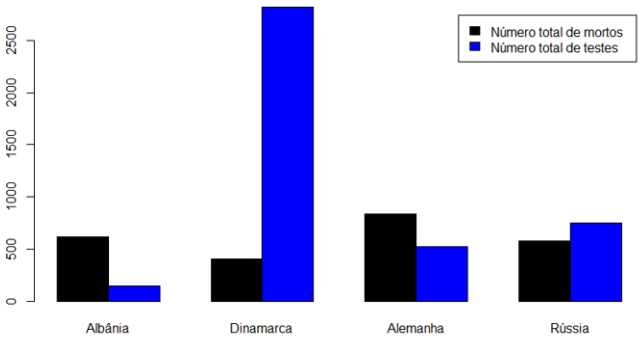

Fig. 4. Gráfico de barras com o número total de mortos, por milhão de habitantes, e o nº total de testes por milhar de habitantes.

Observando o gráfico, conclui-se que, dentro da lista de países visualizados, a Dinamarca apresenta o maior valor total de testes realizados, em contrapartida com a Albânia, que apresenta o menor. Em relação ao número total de mortos, a Alemanha possui o maior valor, enquanto que a Dinamarca regista o mais baixo.

E. Indique qual o país europeu que teve o maior número de infetados, por milhão de habitantes, num só dia

Para efeitos de realização do exercício, foi criada uma estrutura com dados relativos ao continente Europa e, de seguida, foi criado um data frame com os campos LOCATION (localização) e NEW_CASES_PER_MILLION (novos casos por milhão) dessa mesma estrutura criada.

Após este processo, determinou-se o índice da estrutura onde se encontra o maior valor da coluna relativa aos novos casos por milhão de habitantes. Com o índice, extraiu-se o valor de localização correspondente.

O país obtido foi o Vaticano.

F. Indique em que dia, e em que pais, se registou a maior taxa de transmissibilidade do vírus

Este exercício é bastante semelhante ao anteriormente apresentado, em termos de código e raciocínio. No entanto, neste, numa primeira fase, apenas é criada uma estrutura com os campos LOCATION, DATE e REPRODUCTION_RATE, respetivamente, localização, data e taxa de transmissibilidade.

Seguindo a sequência da alínea E, determinou-se o índice da estrutura onde se encontra o maior valor da coluna referente à taxa de transmissibilidade e, com esse mesmo índice, extraiu-se os valores de localização e data correspondentes.

O país obtido foi a Coreia do Sul, no dia 22/02/2020.

*G. Efetue um boxplot do nº de mortos diários por milhão de habitantes, em cada continente. Remova os outliers usando o critério: x é outlier sse não pertencer a [Q1-1.5*IQR , Q3+1.5*IQR]*

A resolução deste exercício é bastante idêntica ao exercício da alínea C. A única diferença reside na localização que, neste caso, é referente a cada continente.

Aplicando os mesmos princípios enunciados na alínea C, obtivemos o seguinte boxplot:

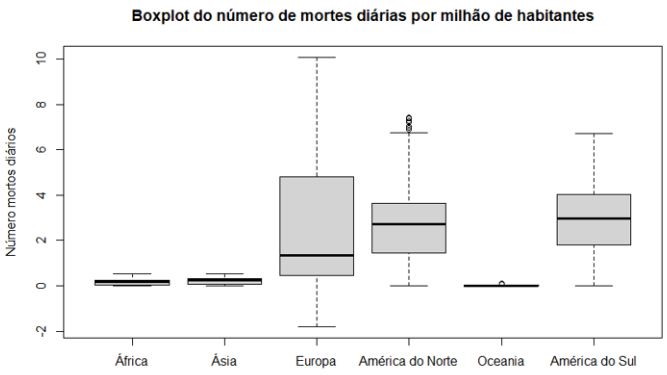

Fig. 5. Boxplot do nº de mortos diários por milhão de habitantes, em cada continente.

III. Inferência Estatística

Os exercícios a serem executados nesta secção partem da geração de amostras pseudoaleatórias dos dados contidos no ficheiro de dados, no período compreendido entre os dias 01/04/2020 e 27/02/2021. Para essa geração, é pedido o uso da função set.seed().

Esta função define o número inicial usado para gerar uma sequência de números aleatórios. Ou seja, assegura que, ao serem gerados números aleatórios, usando o mesmo número passado na função, os valores gerados serão sempre os mesmos, no caso do processo se repetir [12].

A. Considerando apenas os dados relativos, a 30 dias, da amostra pseudoaleatória (usando o valor 118 no parâmetro da função set.seed()) que obteve, verifique se a média da taxa transmissibilidade no Reino Unido é superior à média da taxa de transmissibilidade em Portugal

Numa primeira fase, construímos uma estrutura de dados com os campos LOCATION, DATE e REPRODUCTION_RATE, que correspondem, respetivamente, a localização, data e taxa de transmissibilidade do vírus.

De seguida, foram criadas estruturas para cada país (Portugal e Reino Unido), com a restrição do período temporal. Depois, gerámos amostras pseudoaleatórias de 30 dias para cada um dos países do enunciado, recorrendo ao uso da função set.seed() com o valor 118.

Como é necessário comparar as médias de 2 amostras, realizámos testes de médias, no caso, entre amostras independentes. Considerou-se que as variâncias são desconhecidas e diferentes, ou seja, estamos perante um teste de Welsh. Como as amostras têm tamanho igual a 30, são consideradas grandes e, portanto, o teste poderá ser aplicado.

Formulou-se 2 hipóteses a serem testadas. As hipóteses são as seguintes:

H0: média da taxa de transmissibilidade do Reino Unido $\leq$ *média da taxa de transmissibilidade de Portugal vs H1: média da taxa de transmissibilidade do Reino Unido* $>$ *média da taxa de transmissibilidade de Portugal*

Este teste trata-se, portanto, de um teste unilateral à direita.

Após se realizar o teste, verificou-se que o p-value tem o valor 0.8096. Como este valor é maior que 0.05, então não se rejeita H0. Logo, não existem evidências de que a média da taxa de transmissibilidade no Reino Unido seja superior à media da taxa de transmissibilidade em Portugal.

B. Considerando apenas os dados relativos, a 15 dias, da amostra pseudoaleatória (usando o valor 115 no parâmetro da função set.seed()) que obteve, verifique se há diferenças significativas entre o nº de mortes diárias, por milhão de habitantes, em Espanha, França, Portugal e Itália. No caso de haver, efetue uma análise post-hoc

Numa primeira fase, a metodologia é semelhante ao da alínea A até à execução do teste de Shapiro, para verificar se as amostras têm distribuição normal, no entanto, na estrutura inicial, o campo REPRODUCTION_RATE (taxa de transmissibilidade) é substituído pelo campo NEW_DEATHS_PER_MILLION e, no caso deste exercício, são usados outros países, no caso, Espanha, França, Portugal e Itália. Na função set.seed(), o valor é 115.

TABLE II
TABELA COM DADOS RELATIVOS AO P-VALUE

Paises	p-value
Espanha	0.001544
França	0.0001338
Portugal	2.819e-06
Itália	0.0008006

Os valores de p-value obtidos com o teste de Shapiro para cada país encontram-se descritos na seguinte tabela:

Como podemos constatar, todos os p-value são inferiores a 0.05, isto é, rejeita-se H0. Ou seja, nenhuma das amostras tem distribuição normal, porque H0 é a hipótese que diz que existe distribuição normal e, em vez de usarmos ANOVA, iremos recorrer ao teste de Kruskall-Wallis, que é uma alternativa não paramétrica para o ANOVA.

Consideramos as seguintes hipóteses:

H0: nº de mortes diárias, por milhão de habitantes, em Espanha = nº de mortes diárias, por milhão de habitantes, em França = nº de mortes diárias, por milhão de habitantes, em Portugal = nº de mortes diárias, por milhão de habitantes, na Itália vs H1: nº de mortes diárias, por milhão de habitantes, em Espanha ≠ nº de mortes diárias, por milhão de habitantes, em França ≠ nº de mortes diárias, por milhão de habitantes, em Portugal ≠ nº de mortes diárias, por milhão de habitantes, na Itália

Em seguida, foi criado um data frame agrupando todos os dados relativos ao número de mortos de cada amostra que foi criada. Em seguida, foram criados grupos recorrendo à função factor. Com ambos os dados, fez-se o teste.

Obtivemos p-value = 0.5104. Como p-value é superior a 0.05, leva-nos a não rejeitar H0. Logo, podemos concluir que não existem diferenças significativas entre as amostras e, portanto, não será necessário realizar uma análise Post-Hoc.

C. Para cada Continente gere uma amostra pseudoaleatória, de 30 dias. Para a África use a seed 100, para a Ásia use a seed 101, para a Europa use a seed 102, para a América do Norte use a seed 103, e para a América do Sul use a seed 104. Verifique se existe diferença significativa entre os números médios diários de mortes, por milhão de habitantes, entre os continentes. No caso de haver, efetue uma análise post-hoc

A resolução deste exercício é semelhante ao da alínea B. As únicas diferenças residem no valor da função set.seed(), que é diferente para cada continente, no uso do campo CONTINENT em vez do campo LOCATION e no tamanho das amostras pseudoaleatórias, que passa a ser 30.

Para que se verifique se cada amostra segue uma distribuição normal, neste caso, recorre-se ao teste de Lilliefors, devido ao facto de estarmos perante uma amostra grande, ou seja, cujo tamanho é igual ou superior a 30 (o das amostras é 30).

Os valores de p-value obtidos com o teste de Lilliefors para cada país encontram-se descritos na seguinte tabela:

Como podemos observar, todos os p-value são inferiores a 0.05, isto é, rejeita-se H0. Ou seja, nenhuma das amostras tem distribuição normal e, em vez de usarmos ANOVA, iremos recorrer ao teste de Kruskall-Wallis.

TABLE III
TABELA COM DADOS RELATIVOS AO P-VALUE

Continentes	p-value
África	1.043e-14
Ásia	2.722e-05
Europa	1.003e-05
América do Norte	2.298e-12
América do Sul	0.0002225

Consideramos as seguintes hipóteses:

H0: nº médio diário mortes, por milhão de habitantes, na África = nº médio de mortes diárias, por milhão de habitantes, na Ásia = nº médio de mortes diárias, por milhão de habitantes, na Europa = nº médio de mortes diárias, por milhão de habitantes, na América do Norte = nº médio de mortes diárias, por milhão de habitantes, na América do Sul vs H1: nº médio diário mortes, por milhão de habitantes, na África ≠ nº médio de mortes diárias, por milhão de habitantes, na Ásia ≠ nº médio de mortes diárias, por milhão de habitantes, na Europa ≠ nº médio de mortes diárias, por milhão de habitantes, na América do Norte ≠ nº médio de mortes diárias, por milhão de habitantes, na América do Sul

Em seguida, foi criado um data frame agrupando todos os dados relativos ao número de mortos de cada amostra que foi criada. Em seguida, foram criados grupos recorrendo à função factor. Com ambos os dados, fez-se o teste.

Obtivemos p-value = 0.001283. Como p-value é inferior a 0.05, leva-nos a rejeitar H0. Logo, podemos concluir que existem diferenças significativas entre as amostras e, portanto, será necessário realizar uma análise Post-Hoc.

Para a realização desta análise, fizemos uso da função kruskalmc, proveniente da bilioteca pgirmess.

Obtivemos a seguinte tabela ao correr a função supracitada:

```
Multiple comparison test after Kruskal-Wallis
p.value: 0.05
Comparisons
                                    obs.dif critical.dif difference
África-América do Norte          16.316667     31.48803      FALSE
África-América do Sul            35.250000     31.48803       TRUE
África-Asia                      13.816667     31.48803      FALSE
África-Europa                    39.366667     31.48803       TRUE
América do Norte-América do Sul  18.933333     31.48803      FALSE
América do Norte-Asia             2.500000     31.48803      FALSE
América do Norte-Europa          23.050000     31.48803      FALSE
América do Sul-Asia              21.433333     31.48803      FALSE
América do Sul-Europa             4.116667     31.48803      FALSE
Asia-Europa                      25.550000     31.48803      FALSE
```

Fig. 6. Valores obtidos da análise post hoc.

IV. CORRELAÇÃO

Neste exercício foi-nos pedido para verificar se existia correlação entre 2 pares de valores, em 2021, para países europeus com mais de 10 milhões de habitantes. Para isso, começámos por criar uma estrutura de dados auxiliar com os países, que pertencessem ao continente europeu e que ao mesmo tempo tivessem mais de 10 milhões de habitantes. A estas condições juntámos também a restrição de data, ficando apenas com os registos de 2021. Durante a criação da mesma,

removemos os campos ”NA”. Esta estrutura servirá de base para as alíneas adiante.

A. o valor máximo da taxa diária de transmissibilidade e a densidade populacional de todos os países da Europa com mais de 10 milhões de habitantes

Nesta alínea foi-nos pedido para verificar se existia correlação nas situações enumeradas em cima entre: o valor máximo da taxa diária de transmissibilidade e a densidade populacional. Dito isto, usando a estrutura anterior como base, nós procedemos ao agrupamento do valor máximo da taxa diária de transmissibilidade de um país com a sua densidade populacional. Ou seja, para cada país existente antes, ficaríamos com uma entrada que a esse país corresponderiam a sua densidade populacional e o seu valor máximo da taxa de transmissibilidade. Como pressupostos, considerámos que as variáveis eram contínuas e que não existia nenhum outlier significativo. Feito isto, limitámos-nos a chamar a função R responsável por calcular o coeficiente de Pearson, r.

```
        Pearson's product-moment correlation

data:  max_rep_rate and pop
t = -0.64421, df = 13, p-value = 0.5306
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.6312714  0.3696854
sample estimates:
      cor
-0.175886
```

Fig. 7. Valores obtidos com o cor.test().

Com este valor de r, sendo mais próximo de 0 do que de -1, podemos concluir que existe uma correlação negativa moderada entre as variáveis em estudo.

B. o total de mortos por milhão de habitantes e a percentagem da população com 65 anos ou mais em todos os países da Europa com mais de 10 milhões de habitantes

Nesta alínea foi-nos pedido para verificar se existia correlação nas situações enumeradas em cima entre: o total de mortos por milhão de habitantes e a percentagem da população com 65 anos ou mais. Dito isto, usando a estrutura anterior como base, nós procedemos ao agrupamento do valor máximo do total de mortos por milhão de habitantes de um país com a sua percentagem da população com 65 anos ou mais. Ou seja, para cada país existente antes, ficaríamos com uma entrada que a esse país corresponderiam a sua percentagem da população com 65 anos ou mais e o seu valor máximo registado em 2021 do campo de total de mortos por milhão de habitantes. Como pressupostos, considerámos que as variáveis eram contínuas e que não existia nenhum outlier significativo. Feito isto, limitámos-nos a chamar a função R responsável por calcular o coeficiente de Pearson, r.

Com este valor de r, sendo mais próximo de 0 do que de 1, podemos concluir que existe uma correlação positiva moderada entre as variáveis em estudo.

```
        Pearson's product-moment correlation

data:  age65 and deathsM
t = 1.4745, df = 13, p-value = 0.1642
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1659203  0.7461077
sample estimates:
      cor
0.3785143
```

Fig. 8. Valores obtidos com o cor.test().

V. Regressão

A. Construa o modelo de regressão linear múltipla

Nesta alínea foi-nos pedido para construir o modelo de regressão linear múltipla, tendo por base a variável dependente, o ”Índice de rigor” e como variáveis independentes, a média mensal por mortes diárias por milhão de habitantes (D), a média mensal de casos diários por milhão de habitantes (C) e a média mensal da taxa de transmissibilidade (R). Foi-nos também pedido, para apenas considerarmos as entradas de 2020-04-01 até 2021-02-27, de Portugal. Dito isto, começámos por criar uma estrutura de dados auxiliar capaz de suportar os dados na forma que nos é pedido pelo enunciado, isto é, a cada um dos 11 (onze) meses, ficaram as médias requisitadas, como podemos ver na figura em baixo.

	mediaMensalCasos.datesPortugalMeses	mediaMensalIndice.stringency_index	mediaMensalCasos.new_cases_per_million	mediaMensalMortes.new_deaths_per_million	mediaMensalTaxa.reproduction_rate
1	abril	83.33500	57.54140	2.7009333	1.0000000
2	agosto	71.76129	21.95532	0.2750323	1.0625806
3	dezembro	76.25742	365.76326	7.5957419	0.9741935
4	julho	72.00032	28.25387	0.5029032	0.9151613
5	junho	70.60500	31.51673	0.5425000	1.0630000
6	maio	74.58323	23.58455	1.3318710	0.9367742
7	novembro	73.04133	512.52503	6.5315000	1.1150000
8	outubro	68.02742	207.96413	1.6956129	1.3441935
9	setembro	66.08167	57.30810	0.4869333	1.2273333
10	fevereiro	77.20946	302.66856	13.7806667	0.9600000
11	janeiro	76.64494	970.70556	17.6400968	1.2154839

Fig. 9. Tabela usada para criar o modelo.

Posto isto, criámos o modelo pretendido chamando a função R diretamente responsável por essa função (lm). Os gráficos resultados desta operação são os seguintes:

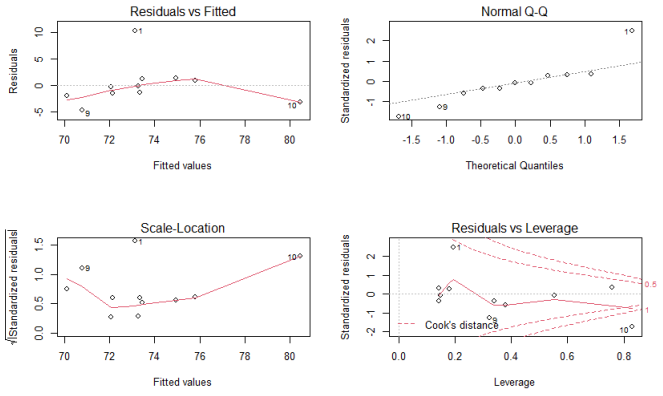

Fig. 10. Modelo de Regressão Linear Múltipla (Normal Q-Q).

B. Verifique se as condições de: Homocedasticidade, Autocorrelação nula e de Multicolinearidade são satisfeitas

Nesta alínea foi-nos pedido para verificar se as condições de: Homocedasticidade, Autocorrelação nula e de Multicolinearidade se encontram satisfeitas.

Usando o modelo da alínea A como base, para a parte da autocorrelação nula, realizámos o teste de Shapiro, bem como o teste de Durbin-Watson de modo a avaliarmos a normalidade e a independência dos resíduos, correspondentemente. O teste de Shapiro indicou que não havia normalidade dos resíduos e o teste de Durbin-Watson apontou para a existência de correlação entre os mesmos. Já na parte da homocedasticidade, realizámos o teste de Breusch-Pagan e o mesmo indicou a existência de homocedasticidade. Por fim, na parte da multicolinearidade, utilizámos o cálculo do VIF e o mesmo também apontou para a existência de multicolinearidade.

C. Estime o valor de Ir para os valores Dm = 10, Cm = 460 e Rm = 1.1

Nesta alínea foi-nos pedido para efetuarmos uma estimativa do valor do Ir (Índice de rigor) para os seguintes valores: Dm = 10, Cm = 460 e Rm = 1.1. Para isso, criámos uma estrutura auxiliar com os valores enunciados na pergunta e posteriormente, aplicámos a mesma ao modelo da alínea A usando a função R predict como mecanismo.

VI. Conclusões

Foi realizado, com recurso a um ficheiro de dados fornecido para o efeito, um estudo que consistiu na apresentação de propostas de resolução para um conjunto de exercícios, exercícios esses que foram realizados com recurso à linguagem R, e que consistiram na análise dos dados provenientes desses ficheiros, sob a forma de vários tipos de gráficos, como gráficos de barras, de evolução ao longo de um determinado período de tempo e diagramas de caixa de bigodes. A partir destes, podemos verificar o comportamento de vários países ou continentes em relação a determinadas variáveis, e avaliar as diferenças entre os mesmos.

Também foram feitas inferências estatísticas, por forma a tentar verificar, através do uso de testes de hipóteses, se certas afirmações feitas sobre uma determinada população eram verdadeiras ou falsas. Para além disto, averiguou-se se existia algum tipo de correlação entre algumas variáveis pedidas pelos exercícios e, através de regressões lineares, foi possível verificar a existência de relações entre uma variável aleatória e um conjunto de variáveis não aleatórias.

O artigo contém, para além das propostas de resolução dos vários exercícios pedidos, um enquadramento teórico, por forma a assimilar todos os conceitos teóricos necessários para a compreensão e realização desses mesmos exercícios.

References

[1] Ritchie, H. (n.d.). Coronavirus source data. Retrieved April 19, 2021, from https://ourworldindata.org/coronavirus-source-data

[2] Informática, E., Isep, D. E. I. (2021). 1 . 1 Instalar o R e o RStudio 1 . 2 Iniciar o R no RStudio. 1–29

[3] Análise Exploratória. (2016, April 13). Retrieved April 23, 2021, from https://analise-estatistica.pt/analise-exploratoria

[4] Moura, A. (2019). Aulas TP - Estatística Descritiva. 1–24.

[5] Oliveira, B. (2021, January 25). Boxplot: Você Sabe como INTERPRETAR Esse tipo de gráfico? Retrieved May 01, 2021, from https://operdata.com.br/blog/como-interpretar-um-boxplot/

[6] Ferreira, P. L. (2005). Estatística Descritiva e Inferencial. Faculdade de Economia - Universidade de Coimbra, 120

[7] Madureira, A. (2021). Aulas Teóricas - Testes de Hipóteses Paramétricos. 1–36.

[8] Parametric test. (n.d.). Retrieved May 01, 2021, from https://www.sciencedirect.com/topics/medicine-and-dentistry/parametric-test

[9] Madureira, A. (2020). Aulas T - Testes de Hipóteses Não Paramétricos. 1–24.

[10] Madureira, A. (2020a). Aulas T - Testes de Correlação. 1–9.

[11] Inform, E. (2020). Regressão Linear Análise de Dados em Informática. 1–33.

[12] Dichotomize, L. (2018, January 22). Live free or dichotomize - a set.seed() + ggplot2 adventure. Retrieved May 01, 2021, from https://livefreeordichotomize.com/2018/01/22/a-set-seed-ggplot2-adventure/