INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Departamento de Engenharia Química

Impacto do túnel do Marquês de Pombal na qualidade do ar da Avenida da Liberdade

FILIPA ANDREIA ALMIRO AFONSO SILVA
(Licenciada em Sociologia)

Dissertação para obtenção do grau de Mestre
em Engenharia da Qualidade e Ambiente

Orientadores: Doutora Célia Maria da Silva Fernandes (ISEL)
Doutor Paulo José Raimundo Ramos (ISEL)

Júri:
Presidente: Doutora Isabel Maria da Silva João (ISEL)
Vogais: Doutora Sandra Maria da Silva Figueiredo Aleixo (ISEL)
Doutora Célia Maria da Silva Fernandes (ISEL)

Setembro de 2025

INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Departamento de Engenharia Química

Impacto do túnel do Marquês de Pombal na qualidade do ar da Avenida da Liberdade

FILIPA ANDREIA ALMIRO AFONSO SILVA
(Licenciada em Sociologia)

Dissertação para obtenção do grau de Mestre
em Engenharia da Qualidade e Ambiente

Orientadores: Doutora Célia Maria da Silva Fernandes (ISEL)
Doutor Paulo José Raimundo Ramos (ISEL)

Júri:
Presidente: Doutora Isabel Maria da Silva João (ISEL)
Vogais: Doutora Sandra Maria da Silva Figueiredo Aleixo (ISEL)
Doutora Célia Maria da Silva Fernandes (ISEL)

Setembro de 2025

Declaração de integridade

Declaro que esta dissertação é o resultado da minha investigação pessoal e independente. O seu conteúdo é original e todas as fontes listadas nas referências bibliográficas foram consultadas e estão devidamente mencionadas no texto. Mais declaro que todas as referências científicas e técnicas relevantes para o desenvolvimento do trabalho estão devidamente citadas e constam das referências bibliográficas.

O autor

Assinado por: **Filipa Andreia Almiro Afonso Silva**
Num. de Identificação: 12027869
Data: 2025.12.12 12:27:28+00'00'

Lisboa, 30 de setembro de 2025

Agradecimentos

A realização desta dissertação representa não apenas o culminar de uma etapa académica exigente, mas também o reflexo de um percurso feito com o apoio e contributo de várias pessoas, às quais quero demonstrar a minha mais profunda gratidão.

Em primeiro lugar, agradeço à minha orientadora, Professora Doutora Célia Maria da Silva Fernandes, e ao meu orientador, Professor Doutor Paulo José Raimundo Ramos, pela orientação, disponibilidade e acompanhamento atento ao longo de todo este processo. O seu conhecimento, dedicação e incentivo foram fundamentais para a concretização deste trabalho.

À Professora Doutora Isabel Maria da Silva João, coordenadora do Mestrado, agradeço todo o apoio, empenho e dedicação com que sempre acompanhou os estudantes, contribuindo de forma significativa para o bom funcionamento do curso e para o ambiente académico vivido.

À minha filha, Beatriz Silva, agradeço do fundo do coração por ser a minha maior fonte de inspiração e motivação. A sua presença, ainda que silenciosa, acompanhou cada momento deste percurso. Ao meu marido, Nuno Silva, agradeço a paciência, o apoio incondicional e a força com que me ajudou a superar os momentos mais difíceis.

Aos meus colegas de turma, com quem partilhei desafios, aprendizagens e conquistas, deixo um sincero agradecimento. Em especial, à Rita Silva e ao Marco Freitas, pelo companheirismo, amizade e entreajuda constantes ao longo destes anos.

Agradeço ainda ao Jornal local "LPP/Lisboa para pessoas" pela permissão de utilização da imagem da Estação de monitorização da qualidade do ar da Avenida da Liberdade, publicada em https://lisboaparapessoas.pt/2024/01/17/avenida-da-liberdade-poluicao-2023-dados/.

A todos, o meu muito obrigada.

Resumo

A poluição ambiental é um dos maiores problemas com que as grandes cidades se confrontam atualmente. Este é também um problema da cidade de Lisboa e ao longo dos anos foram sendo implementadas medidas para reduzir as emissões de poluentes atmosféricos. Neste trabalho pretende-se apresentar um estudo estatístico sobre os principais poluentes atmosféricos existentes na zona da Avenida da Liberdade, nomeadamente, o monóxido de carbono, o dióxido de azoto e as partículas finas de matéria em suspensão com diâmetro aerodinâmico inferior a 10 micrómetros. Foram usados dados recolhidos na estação de monitorização da qualidade do ar situada nesse local. Tentou-se, ainda, avaliar o impacto ambiental associado à construção do túnel do Marquês de Pombal, quer na poluição provocada pelas obras de construção/ampliação, quer nas mudanças no tráfego automóvel provocadas pela sua implementação, focando-nos na qualidade do ar. A análise comparou o nível de poluição dos três poluentes em seis períodos diferentes, considerando-se um período anterior à construção do túnel, períodos associados às várias obras realizadas, um período após a finalização das obras e, por fim, o período associado ao confinamento devido à pandemia provocada pelo vírus SARS-CoV-2.

Palavras-chave: Poluição ambiental; Monóxido de carbono; Dióxido de azoto; Partículas finas de matéria em suspensão; Estação de monitorização da qualidade do ar; ANOVA; Teste de Kruskall-Wallis; Teste de Nemenyi; Análise discriminante linear.

Abstract

The Environmental pollution is one of the biggest problems that large cities are facing today. This is also a problem in the Lisbon city, and over the years, measures have been implemented to reduce air pollutant emissions. In this paper, we present a statistical study of the main air pollutants in the Avenida da Liberdade area, like carbon monoxide, nitrogen monoxide, and fine matter particles suspended with an aerodynamic diameter of less than 10 micrometers. The data collected of the air quality was monitorized by the station located in that place. We also attempted to assess the environmental impact associated with the construction of the Marquês Tunnel, both in terms of pollution caused by the construction/expansion works and in the changes in vehicle traffic, resulting from its implementation, with a focus on air quality. The analysis compared the pollution levels of the three pollutants in six different periods, considering a period prior to the construction of the tunnel, periods associated with the various works carried out, a period after the completion of the works and, finally, the period associated with the confinement due to the pandemic caused by the SARS-CoV-2 virus.

Keywords: Environmental pollution; Carbon monoxide; Nitrogen monoxide; Fine matter particles suspended; Air quality monitoring station; ANOVA; Kruskal-Wallis test; Nemenyi test; Linear discriminant analysis.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

APA	Agência portuguesa do ambiente
AVC	Acidente vascular cerebral
$CCDR$	Comissões de Coordenação e Desenvolvimento Regional
COV_s	Compostos Orgânicos Voláteis
$DPOC$	Doença pulmonar obstrutiva crónica
$EMQAr$	Estações de monitorização da qualidade do ar
LRN	Laboratório de Referencia Nacional
NP	Norma Portuguesa
OMS	Organização Mundial de Saúde
PM	Material Particulado
PM_{10}	Material Particulado com diâmetro aerodinâmico inferior a 10 micrómetros
$PM_{2.5}$	Material Particulado com diâmetro aerodinâmico inferior a 2.5 micrómetros
VL	Valores limite

Lista de Elementos Químicos

CO	Monóxido de carbono
CO_2	Dióxido de carbono
NH_3	Amoníaco
NO	Monóxido de Azoto
NO_2	Dióxido de Azoto
NO_x	Óxidos de Azoto
N_2O	Óxido Nitroso
O_3	Ozono troposférico
SO_2	Dióxido de Enxofre

Capítulo 1

Introdução

1.1 Estações de monitorização da qualidade do ar

O crescimento acelerado dos centros urbanos tem intensificado a concentração de poluentes na atmosfera, despertando uma preocupação cada vez maior da sociedade em relação a essa questão.

A poluição do ar é amplamente reconhecida como a principal ameaça ambiental à saúde e ao bem-estar humano. De acordo com a Organização Mundial da Saúde (OMS), cerca de 99% da população global está exposta a condições insalubres, respirando altos níveis de partículas finas e dióxido de azoto (NO_2).

A poluição atmosférica de origem humana é a principal causa da degradação da qualidade do ar. Ela resulta, principalmente, de grandes fontes emissoras, como os meios de transporte, sistemas de aquecimento residencial, atividades agrícolas, domésticas e industriais (Gomes, 2010). O Decreto-Lei n.º 102/2010, de 23 de setembro, na sua redação atual, regula o regime de avaliação e gestão da qualidade do ar ambiente, definindo os critérios mínimos para essa avaliação, sustentada nos princípios estipulados no diploma (APA, 2021a).

Segundo o Decreto-Lei n.º 102/2010, de 23 de setembro existe uma diferenciação entre "zona" e "aglomeração":

- Zona – "são áreas geográficas de características homogéneas, em termos de qualidade do ar, ocupação do solo e densidade populacional" (APA, 2021b).
- Aglomeração - "são zonas caracterizadas por um número de habitantes superior a 250 000, ou que se situe entre 50 000 e 250 000 e tenha uma densidade populacional superior a 500 habitantes/km2" (APA, 2021b).

As zonas são delimitadas com base na análise da qualidade do ar ambiente, enquanto as aglomerações correspondem a áreas onde os critérios de definição se limitam a parâmetros estatísticos relacionados com a população residente.

A delimitação das zonas esta disponível para consulta na base de dados do sistema de informação do Qualar (`https://qualar.apambiente.pt/intro`). O QualAr é o sistema oficial usado para informar o público sobre a qualidade do ar, sendo gerido pela Agência Portuguesa do Ambiente (APA). Através deste sistema — disponível online e também em aplicação móvel — é possível consultar, em tempo real, o estado da qualidade do ar em diferentes regiões do país. O Qualar recolhe dados das estações de monitorização espalhadas pelo território e calcula o Índice de Qualidade do Ar.

Portugal dispõe de estações e redes fixas de medição para monitorizar a qualidade do ar ambiente, sendo a maioria gerida e operada pelas Comissões de Coordenação e Desenvolvimento

Regional (CCDR).

As estações de medição da qualidade do ar possuem tipologias distintas, ajustadas às emissões predominantes na área onde estão localizadas. Estas estações refletem diferentes formas de exposição da população à poluição atmosférica e, por essa razão, estão equipadas com analisadores específicos para medir diferentes poluentes.

Para cada poluente, existe um conjunto específico de locais de medição (estações), cuja localização é definida com base em requisitos que asseguram a representatividade das medições em cada zona:

- Estação de tráfego - o objetivo é monitorizar as concentrações máximas de poluentes provenientes do tráfego rodoviário, às quais a população pode estar exposta. Estas concentrações, frequentemente altas durante curtos períodos de tempo, são medidas em locais próximos a vias de tráfego intenso (APA, 2021c).
- Estações de fundo - o objetivo é avaliar a exposição média da população a concentrações de fundo, em locais afastados da influência direta de vias de tráfego ou de qualquer fonte próxima de poluição (APA, 2021c).
- Estações industriais - têm como finalidade avaliar as concentrações máximas de determinados poluentes de origem industrial, estando localizadas nas proximidades de áreas industriais (APA, 2021c).

Depois de definida a localização de uma estação, a escolha do ponto de entrada de amostra segue critérios de micro-localização, garantindo que as medições refletem adequadamente o tipo de ambiente pretendido.

Uma forma de "medir" a poluição do ar é através de estações de monitorização da qualidade do ar (EMQAr), que fazem a monitorização da concentração dos poluentes atmosféricos. As estações estão equipadas com um dispositivo de amostragem que recolhe o ar do exterior da estação e o distribui por um conjunto de analisadores que medem os vários poluentes atmosféricos, em contínuo e de forma automática, determinando as suas concentrações no ar ambiente "em tempo real" (APA, 2021c).

Para garantir a comparabilidade das medições de um poluente feitas em diferentes locais e períodos, os métodos utilizados para medir o mesmo poluente devem ser equivalentes e cumprir critérios rigorosos de garantia e controlo de qualidade.

Estão definidos métodos de medição de referência para avaliar a concentração de cada poluente, bem como a metodologia necessária para demonstrar a equivalência de outros métodos em relação ao método de referência. Estes métodos alternativos devem ser aprovados pela Agência Portuguesa do Ambiente (APA), no âmbito das suas responsabilidades como Laboratório de Referência Nacional (LRN). As normas que estabelecem os métodos de medição também especificam os requisitos de controlo e garantia de qualidade a serem seguidos, incluindo mecanismos para avaliar o cumprimento desses critérios, com o LRN da APA desempenhando um papel fundamental.

O sistema de informação Qualar, além de apresentar os resultados da avaliação da qualidade do ar, inclui informações sobre os métodos de medição utilizados em cada estação das redes de monitorização em funcionamento em Portugal (APA, 2021d).

Os dados obtidos através do Qualar são analisados tendo por base o enquadramento legal aplicável à qualidade do ar ambiente, nomeadamente a legislação europeia e nacional que estabelece os valores-limite e os critérios de avaliação dos diferentes poluentes atmosféricos. Em Portugal, a principal referência normativa é o Decreto-Lei n.º 102/2010, que transpõe para o ordenamento jurídico nacional a Diretiva 2008/50/CE, definindo os valores-limite, os valores-alvo e os níveis de alerta aplicáveis a poluentes como as partículas inaláveis (PM10 e

PM2.5), o dióxido de azoto, o dióxido de enxofre, o ozono, o monóxido de carbono e o benzeno. O Decreto-Lei prevê ainda a adoção de ações corretivas e planos de melhoria sempre que os valores-limite sejam ultrapassados, garantindo a proteção da saúde pública e do ambiente. Por fim, obriga à divulgação periódica dos resultados da qualidade do ar e à harmonização das avaliações com a legislação comunitária, promovendo uma gestão rigorosa e cientificamente fundamentada da poluição atmosférica em Portugal.

1.2 Importância dos estudos da qualidade do ar

A qualidade do ar é um dos pilares fundamentais para a saúde pública e o bem-estar das populações, dado que a poluição atmosférica tem efeitos diretos e indiretos sobre o organismo humano. Nos últimos anos, o aumento das emissões de poluentes provenientes de fontes como o tráfego rodoviário, a atividade industrial, a queima de combustíveis fósseis e a agricultura tem intensificado as preocupações sobre os impactos na saúde. Os estudos sobre a qualidade do ar são essenciais para compreender a composição do ar que respiramos e as concentrações de poluentes como partículas finas de matéria em suspensão com diâmetro aerodinâmico inferior a 2.5 ou 10 micrómetros ($PM_{2.5}$ e PM_{10}), dióxido de azoto (NO_2), dióxido de enxofre (SO_2), ozono (O_3) e monóxido de carbono (CO), entre outros. Estes poluentes têm a capacidade de penetrar nas vias respiratórias e atingir os pulmões, podendo afetar o sistema cardiovascular, respiratório e até mesmo provocar cancro.

A exposição contínua a esses poluentes está associada ao aumento de doenças respiratórias crónicas, como asma, bronquite e doença pulmonar obstrutiva crónica (DPOC), além de problemas cardiovasculares como enfarte do miocárdio, acidente vascular cerebral (AVC) e hipertensão. Estudos epidemiológicos têm demonstrado que a poluição do ar também está ligada ao aumento da mortalidade precoce, afetando, principalmente, populações vulneráveis, como crianças, idosos e indivíduos com doenças pré-existentes. O Relatório de Saúde e Ambiente, produzido pelo Observatório Português da Saúde e Ambiente estima que cerca de 8% das mortes registadas em Portugal estão associadas a fatores ambientais.

Além dos efeitos diretos na saúde, a qualidade do ar tem implicações significativas na qualidade de vida e na produtividade das populações, já que a poluição pode levar a uma maior incidência de doenças, absentismo laboral e aumento dos custos com o sistema de saúde. Por conseguinte, os estudos da qualidade do ar desempenham um papel essencial na identificação de áreas de risco, na monitorização dos níveis de poluentes e na implementação de políticas públicas de saúde e ambientais, que visam a redução das emissões e a promoção de ambientes mais saudáveis. Esses estudos também permitem que sejam avaliados os impactos de medidas de mitigação, como a adoção de tecnologias mais limpas, o incentivo ao transporte público, a regulamentação das emissões industriais e a implementação de zonas verdes nas cidades. Dessa forma, a investigação sobre a qualidade do ar não só contribui para a compreensão dos riscos para a saúde, mas também orienta ações concretas para melhorar a qualidade do ar, proteger a saúde das populações e promover um desenvolvimento sustentável.

A poluição do ar continua a ter impactos significativos na saúde da população europeia, particularmente em áreas urbanas. Os poluentes mais graves na Europa, em termos de danos para a saúde humana, são o material particulado, NO_2 e O_3 (EEA, 2022). Nesta dissertação serão apenas referidas PM_{10} em $\mu g/m^3$, NO_2 em $\mu g/m^3$ e CO em mg/m^3.

1.2.1 Partículas PM_{10}

As partículas finas de matéria em suspensão com diâmetro aerodinâmico inferior a 10 micrómetros englobam substâncias minerais e/ou orgânicas que se podem encontrar na atmosfera sob a forma líquida ou sólida. Devido ao seu pequeno tamanho, são inaláveis e podem penetrar nas vias respiratórias superiores, chegando até os brônquios.

As PM_{10} são uma das principais componentes da poluição do ar e são monitorizadas devido ao seu impacto na saúde humana e no meio ambiente.

As partículas PM_{10} podem ser classificadas como primárias ou secundárias dependendo da sua origem. São classificadas como primárias aquelas que são emitidas diretamente para a atmosfera a partir de fontes naturais (como poeira do solo e erupções vulcânicas) ou atividades humanas (como emissões industriais e tráfego rodoviário). As que são classificadas como secundárias são as que se formam na atmosfera através de reações químicas entre outros poluentes gasosos como SO_2, Óxidos de Azoto (NO_x), Amoníaco (NH_3) e Compostos Orgânicos Voláteis (COV_s), que podem ter origem tanto em processos naturais como em atividades humanas (APA, 2021e). Após serem libertadas na atmosfera são transportadas pelo ar, podendo mesmo criar concentrações elevadas em locais distantes da sua fonte.

Quanto menor for o tamanho das partículas, maior será a sua capacidade de penetrar profundamente no sistema respiratório, aumentando os impactos negativos na saúde. As PM_{10} são as mais prejudiciais, pois conseguem alcançar as vias respiratórias e, em alguns casos, atingir os pulmões, causando diversos problemas de saúde.

1.2.2 Dióxido de azoto (NO_2)

Os NO_x referem-se a uma série de compostos formados por Azoto e Oxigénio, incluindo o Monóxido de Azoto (NO), o NO_2 e o Óxido Nitroso (N_2O), entre outros. De entre estes, o NO e o NO_2 são os mais significativos como poluentes do ar, enquanto o N_2O é principalmente reconhecido como um gás de efeito estufa. O NO_2 é resultante da combustão de combustíveis fósseis, principalmente em veículos automóveis, indústrias e centrais termoelétricas (APA, 2021f).

O NO_2, o único dos NO_x que é regulamentado, é um gás acastanhado, com um odor característico, muito corrosivo e fortemente oxidante (APA, 2021f).

Nos ambientes urbanos, os transportes representam a principal fonte de NO_x. As emissões provenientes dos escapes dos veículos são, em grande parte, compostas por NO, uma molécula instável que reage rapidamente com o oxigénio atmosférico, originando NO_2. Em áreas com tráfego intenso, as concentrações de NO_x variam em função do fluxo de veículos (APA, 2021f).

O NO_2 provoca uma série de doenças graves respiratórias como enfisema pulmonar, bronquites, traqueítes e em casos mais graves cancro (Castro et al., 2013).

1.2.3 Monóxido de carbono (CO)

O CO é um gás resultante da combustão incompleta de combustíveis fósseis ou de outras substâncias orgânicas ricas em carbono. A sua origem pode ser antropogénica, proveniente de fontes como a geração de eletricidade, processos industriais, aquecimento residencial e comercial, além das emissões dos veículos movidos a motores de combustão. No entanto, também pode ser liberado por fenómenos naturais, como erupções vulcânicas e incêndios florestais (APA, 2021g).

Nas zonas urbanas a maior produção de CO dá-se devido aos transportes rodoviários, e este poluente é principalmente emitido quando os motores estão em altas rotações como, por

exemplo, no "pára-arranca".

A exposição ao CO pode provocar sintomas como dor de cabeça, tonturas, mal-estar, náuseas e vómitos. Além disso, pode comprometer a capacidade de aprendizagem, reduzir o desempenho no trabalho e afetar a coordenação motora. Os seus efeitos nocivos na saúde resultam da sua forte afinidade com a hemoglobina do sangue, formando uma ligação estável que impede o transporte de oxigénio dos pulmões para os tecidos e de dióxido de carbono no sentido inverso. Em casos de exposição prolongada, a falta de oxigenação pode levar a consequências graves, incluindo a morte (APA, 2021g).

Em Portugal, a concentração de CO é legislada através do Decreto-Lei n.º 102/2010, de 23 de setembro.

Capítulo 2

Métodos estatísticos

As informações contidas neste capítulo foram retiradas de livros (Alkarkhi & Alqaraghuli, 2020; Härdle & Hlávk, 2015; Johnson & Wicher, 2014; Marôco, 2021; Pestana & Gageiro, 2014; Zelterman, 2015), dissertações de doutoramento (Nemenyi, 1963) e apontamentos das disciplinas de Complementos de Estatística para a Engenharia do Mestrado de Engenharia da Qualidade e Ambiente, de Estatística Multivariada da Licenciatura em Matemática Aplicada à Tecnologia e à Empresa e de Técnicas de Estatística Multivariada das Licenciaturas e apontamentos das disciplinas de Complementos de Estatística para a Engenharia do Mestrado de Engenharia da Qualidade e Ambiente, de Estatística Multivariada da Licenciatura em Matemática Aplicada à Tecnologia e à Empresa e de Técnicas de Estatística Multivariada das Licenciaturas em Engenharia Biomédica, em Engenharia Informática e de Computadores e em Engenharia Química e Biológica (Fernandes & Ramos, P, 2025a,b,c,d) do Instituto Superior de Engenharia de Lisboa.

2.1 Análise estatística descritiva dos dados

2.1.1 Medidas de localização de tendência central

Vamos considerar as seguintes medidas de localização de tendência central:

- Média;
- Mediana;
- Moda.

A média é uma medida de tendência central muito usada, é o valor obtido somando todos os valores da amostra e dividindo pelo número total de observações. A média é fortemente influenciada quando a amostra tem outliers. A fórmula da média é a seguinte:

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

onde n representa o número total de observações e x_i representa o i-ésimo valor observado da amostra relativa à variável em estudo.

A mediana é outra medida de tendência central muito utilizada. Esta divide a distribuição de valores em duas partes iguais. Para se poder calcular a mediana de uma amostra os valores observados têm de estar ordenados.

Para o calculo da mediana é necessário considerar duas hipóteses:

- Se o número de observações é par:

$$Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2};$$

- Se o número de observações é ímpar:

$$Me = x_{\left(\frac{n+1}{2}\right)};$$

onde $x_{(i)}$ é o i-ésimo valor observado na amostra ordenada.

Também se pode obter o valor da mediana com a fórmula de cálculo dos quartis, pois a mediana corresponde ao segundo quartil, que iremos ver mais à frente.

A moda é o valor que ocorre com mais frequência num conjunto de dados.

- Se o conjunto de valores observados tiver uma moda, ele diz-se unimodal;
- Se o conjunto de valores observados tiver duas modas, ele diz-se bimodal;
- Se o conjunto de valores observados tiver mais do que duas modas, ele diz-se multimodal;
- Se o conjunto de valores observados não tiver moda, ele diz-se amodal.

2.1.2 Medidas de localização de tendência não central

As medidas de localização de tendência não central designam-se por quantis. Os quantis são valores que dividem uma distribuição de dados ordenados em partes iguais. Existem vários tipos de quantis:

- Quartis: Dividem os dados em quatro partes iguais, pelo que, existem três quantis;
- Decis: Dividem os dados em dez partes iguais, pelo que, existem nove decis;
- Percentis: Dividem os dados em cem partes iguais, pelo que, existem noventa e nove percentis.

O quantil de ordem p, Q_p, obtém-se da seguinte forma:

$$Q_p = (1-k)\, x_{(i)} + k x_{(i+1)},$$

onde

$$i = \lfloor np + 1 - p \rfloor,$$

$$k = np + 1 - p - i,$$

$0 < p < 1$, $x_{(i)}$ é o i-ésimo valor observado na amostra ordenada e $\lfloor y \rfloor$ é a parte inteira de y. Quando $p = \frac{1}{2}$, o quantil $Q_{\frac{1}{2}}$ corresponde à mediana pois divide o conjunto de dados em duas partes iguais. Para obter os quartis, considera-se p igual a $\frac{1}{4}$, $\frac{1}{2}$ e $\frac{3}{4}$. Para obter os decis, p assume os valores $\frac{1}{10}$, $\frac{2}{10}$, $\frac{3}{10}$, $\ldots$, $\frac{9}{10}$, e para obter os percentis, o valor de p será $\frac{1}{100}$, $\frac{2}{100}$, $\frac{3}{100}$, $\ldots$, $\frac{99}{100}$.

O *outlier* é uma observação que se destaca das demais por estar muito distante do padrão geral dos dados ou por ser inconsistente em relação a eles. Estes valores extremos podem influenciar fortemente medidas estatísticas como a média e o desvio padrão, podendo distorcer a interpretação dos dados. A presença de *outliers* pode ocorrer por motivos reais, como

variações naturais no fenómeno estudado ou por erros humanos e falhas no processo de recolha ou análise dos dados.

Apesar do intervalo interquartis ser uma medida de dispersão irá ser definida neste ponto pois necessitamos dela para pesquisar a existência de *outliers*. Para obter este intervalo usamos a seguinte fórmula:

$$IQ = Q_{\frac{3}{4}} - Q_{\frac{1}{4}}.$$

Qualquer observação que verifique uma das seguintes condições:

$$Q_{\frac{1}{4}} - 3 \times IQ \leq x_i < Q_{\frac{1}{4}} - 1,5 \times IQ$$

ou

$$Q_{\frac{3}{4}} + 1,5 \times IQ < x_i \leq Q_{\frac{3}{4}} + 3 \times IQ$$

designa-se por *outlier* moderado. Se a observação verificar uma das seguintes condições:

$$x_i < Q_{\frac{1}{4}} - 3 \times IQ$$

ou

$$x_i > Q_{\frac{3}{4}} + 3 \times IQ$$

designa-se por *outlier* extremo ou severo.

2.1.3 Medidas de dispersão

As medidas de dispersão dão uma ideia da variabilidade dos dados, ou seja, se estes são muitos distintos dos outros. Existem várias medidas de dispersão, como por exemplo:

- Variância;
- Desvio padrão;
- Coeficiente de variação.

A variância é uma medida de variabilidade ou dispersão dos dados relativamente à média. A fórmula da variância é:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}{n-1}.$$

O desvio padrão é uma medida que expressa o quanto os valores de uma variável se afastam da média, e mostra, de forma absoluta o grau de dispersão dos dados, em relação ao valor da média. O desvio padrão é a raiz quadrada da variância:

$$s = \sqrt{s^2}.$$

O coeficiente de variação é uma medida de dispersão de natureza relativa, ou seja, independente das unidades de medida dos dados. O valor do coeficiente de variação é dado pelo quociente entre o desvio padrão e o valor absoluto da média em valor percentual:

$$cv = \frac{s}{\mid \overline{x} \mid} \times 100\%.$$

Este coeficiente é particularmente útil quando pretendemos comparar a dispersão de duas distribuições de valores em que:

- as variáveis não estão expressas na mesma unidade de medida;
- as médias amostrais são muito diferentes.

O coeficiente de variação é também designado por desvio padrão relativo.

O coeficiente de variação resistente, cvr, é uma medida estatística de dispersão relativa robusta, ou seja, deve ser usada quando a amostra apresenta *outliers*, pois o valor do IQ e da Me não são influenciados pela existência de *outliers*:

$$cvr = \frac{IQ}{Me} \times 100\%.$$

Diz-se que:

- se $cv \leq 15\%$ os dados apresentam uma variabilidade fraca;
- se $15\% < cv < 30\%$ os dados apresentam uma variabilidade média;
- se $cv \geq 30\%$ os dados apresentam uma variabilidade elevada.

A interpretação é análoga para o coeficiente de variação resistente.

2.2 Testes de aderência ou de qualidade de ajuste

De entre os vários testes existentes, destacam-se os seguintes:

- Teste de Kolmogorov-Smirnov com correção de Lilliefors;
- Teste de Shapiro-Wilk.

Em ambos os testes, as hipóteses em teste são as seguintes:

H_0: A amostra é proveniente de uma população com distribuição normal;

H_1: A amostra é proveniente de uma população com distribuição diferente da distribuição normal.

Nas secções seguintes iremos abordar, de uma forma mais pormenorizada, cada um destes testes.

2.2.1 Teste de Kolmogorov-Smirnov com correção de Lilliefors

O teste de Kolmogorov-Smirnov com correção de Lilliefors é uma adaptação do teste Kolmogorov-Smirnov feita para lidar com o facto de, em muitos casos, os parâmetros da distribuição normal (como o valor médio e o desvio padrão) serem desconhecidos e precisarem de ser estimados a partir dos dados. Quando estes parâmetros são estimados, o teste de Kolmogorov-Smirnov tradicional tende a ser mais conservador e pode subestimar a probabilidade de rejeitar a hipotese nula de normalidade. Portanto a correção de Lilliefors ajusta o teste de Kolmogorov-Smirnov, para tornar os resultados mais precisos quando se trabalha com dados amostrais e parâmetros desconhecidos (Lilliefors, 1967).

Estatística de teste:

$$D = \max\left\{D^+, D^-\right\},$$

onde

$$D^{+} = \sup\left\{\left|F\left[x_{(i)}\right] - S_n\left[x_{(i)}\right]\right|\right\}$$

e

$$D^{-} = \sup\left\{\left|F\left[x_{(i)}\right] - S_n\left[x_{(i-1)}\right]\right|\right\},$$

sendo $F\left[x_{(i)}\right] = \Phi\left[x_{(i)}\right]$ a função de distribuição de probabilidade da distribuição normal reduzida e $S_n\left(x\right)$ a função de distribuição empírica:

$$S_n\left(x\right) = \begin{cases} 0 & , \text{ se } x < x_{(1)} \\ \frac{i}{n} & , \text{ se } x_{(i)} \leq x < x_{(i+1)} \\ 1 & , \text{ se } x \geq x_{(n)} \end{cases},$$

onde $x_{(i)}$ é o i-ésimo valor observado na amostra ordenada;

Regra de decisão: Para decidir se rejeitamos ou não H_0 temos de determinar o valor crítico (D_n). O D_n depende da dimensão da amostra, n, e do nível de significância usado. Este valor é obtido através de tabelas especificas para o teste Kolmogorov-Smirnov com correção de Lilliefors que consideram a dimensão da amostra e o nível de significância desejado, normalmente um valor entre 1% e 10%. Se o valor observado da estatística de teste ($D_{observado}$) for maior ou igual ao valor crítico, ou seja se $D_{observado} \geq D_n$, ou se o nível de significância for maior ou igual que o $p - value$, ou seja se $\alpha \geq p - value$, devemos rejeitar H_0.

2.2.2 Teste de Shapiro-Wilk

O teste de Shapiro-Wilk é um teste muito utilizado para avaliar a normalidade de uma distribuição de dados. Criado por Samuel Shapiro e Martin Wilk em 1965, este teste é reconhecido pela sua alta sensibilidade na deteção dos desvios em relação à normalidade. Este teste premite obter bons resultados mesmo quando o número de observações é pequeno (Shapiro & Wilk, 1965).

Estatística de teste:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_{(i)} - \overline{x}\right)^2},$$

onde $x_{(i)}$ é o i-ésimo valor observado na amostra ordenada e a_i são valores constantes obtidos a partir da média, das variâncias e das covariâncias de uma amostra de dimensão n de uma distribuição normal.

Regra de decisão: O valor crítico (W_n) da distribuição da estatística W encontra-se em Shapiro & Wilk (1965). O seu valor depende da dimensão da amostra e do nível de significância usado. Se o valor observado da estatística de teste for menor ou igual ao valor crítico, ou seja se $W_{observado} \leq W_n$, ou se $\alpha \geq p - value$ devemos rejeitar H_0.

2.3 Testes de homocedasticidade ou de homogeneidade das variâncias

De entre os vários testes existentes, destacamos os seguintes:

- Teste de Levene;
- Teste de Bartlett.

Em ambos os testes, as hipóteses em teste são as seguintes:

$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

$H_1 : \exists i, j : \sigma_i^2 \neq \sigma_j^2$, com $i \neq j$; $i, j = 1, \ldots, k$

Nas secções seguintes iremos abordar, de uma forma mais pormenorizada, cada um destes testes.

2.3.1 Teste de Levene

O teste de Levene é um dos testes mais utilizados para verificar a homogeneidade das variâncias. Este teste tem como finalidade verificar se as variâncias populacionais são idênticas, ou se existem pelo menos duas que são diferentes.

Estatística de teste:

$$F = \frac{n-k}{k-1} \times \frac{\sum_{i=1}^{k} n_i \left(\overline{Z}_i - \overline{Z}\right)^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Z_{ij} - \overline{Z}_i\right)^2} \sim F_{k-1;n-k},$$

em que n_i é a dimensão de cada uma das k amostras, $i = 1, \ldots, k$, e n é a dimensão da amostra global:

$$n = n_1 + n_2 + \cdots + n_k.$$

A variável Z pode definir-se como $Z_{ij} = \left|Y_{ij} - \overline{Y}_i\right|$, com $i = 1, \ldots, k$ e $j = 1, \ldots, n_i$, em que Y_{ij} é a observação j da amostra i e $\overline{Y}_i$ é a média da amostra i, quando a variável $Y \sim N(\mu, \sigma)$. Se existirem fortes suspeitas de que a variável não tem distribuição normal então Z deve calcular-se por $Z_{ij} = \left|Y_{ij} - \widetilde{Y}_i\right|$, em que $\widetilde{Y}_i$ é a mediana da amostra i. Para além disso, $\overline{Z}_i$ é a média de Z_{ij} na amostra i e $\overline{Z}$ é a média de Z_{ij} na amostra global:

$$\overline{Z}_i = \sum_{j=1}^{n_i} \frac{Z_{ij}}{n_i} \qquad \text{e} \qquad \overline{Z} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{Z_{ij}}{n}.$$

Regra de decisão: Rejeitar H_0 se $F_{\text{observado}} \geq F_{k-1;n-k;1-\alpha}$, sendo k o número de grupos e $n = n_1 + n_2 + \cdots + n_k$ ou se $\alpha \geq p - value$.

2.3.2 Teste de Bartlett

O objetivo deste teste é o mesmo do teste de Levene. O teste de Bartlett é bastante sensível a desvios da normalidade, pelo que para o aplicar de forma adequada os dados devem ser normalmente distribuídos.

Estatística de teste:

$$Q = \frac{(n-k)\ln s_p^2 - \sum_{i=1}^{k} (n_i - 1)\ln s_i^2}{1 + \frac{1}{3(k-1)} \times \left[\left(\sum_{i=1}^{k} \frac{1}{n_i - 1}\right) - \frac{1}{n-k}\right]} \sim \chi^2_{k-1},$$

em que n_i é a dimensão de cada uma das k amostras, $i = 1, \ldots, k$, e n é a dimensão da amostra global:

$$n = n_1 + n_2 + \cdots + n_k,$$

s_i^2 é a variância de cada uma das k amostras, $i = 1, \ldots, k$, e s_p^2 é a variância combinada, que é uma média ponderada das variâncias do grupo e é dada por:

$$s_p^2 = \sum_{i=1}^{k} \frac{(n_i - 1)\, s_i^2}{n-k}.$$

Regra de decisão: Rejeitar H_0 se $Q_{\text{observado}} \geq \chi^2_{k-1;1-\alpha}$, sendo k o número de variáveis ou se $\alpha \geq p - value$.

2.4 ANOVA

A designação de ANOVA vem da expressão inglesa **AN**alysis **O**f **VA**riance. Os pressupostos da ANOVA são os seguintes:

- Independência: verifica-se quando as observações em cada grupo são independentes umas das outras. Isso significa que os dados de um grupo não devem influenciar ou afetar os dados de outro grupo. Cada participante ou unidade de observação de um grupo deve ser único e não ter qualquer relação com os participantes de outros grupos.
- Normalidade, descrita na Secção 2.2: refere-se à suposição de que os dados dentro de cada grupo seguem uma distribuição normal. Esta suposição é muito importante porque a ANOVA baseia-se na análise das médias dos grupos, e a normalidade das distribuições ajuda a garantir que os resultados do teste são confiáveis e válidos.
- Homocedasticidade, descrita na Secção 2.3: refere-se à suposição de que os diferentes grupos em estudo possuem caraterísticas semelhantes ou comparáveis, especialmente em termos das variâncias. A dispersão dos dados em torno da média é igual em todos os grupos, o que vai permitir a comparação direta das médias entre grupos. Caso a homogeneidade seja violada, isto é, caso as variâncias dos grupos sejam significativamente diferentes, os resultados dos testes podem ser distorcidos, levando a conclusões erradas.

As hipóteses são as seguintes:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$

$H_1 : \exists i, j : \mu_i \neq \mu_j$, com $i \neq j$; $i, j = 1, \ldots, k$

Estatística de teste:

$$F = \frac{\frac{SQF}{k-1}}{\frac{SQE}{n-k}} = \frac{QMF}{QME},$$

sendo a soma de quadrados do factor dada por

$$SQF = \sum_{i=1}^{k} n_i \left(\overline{Y}_i - \overline{Y}\right)^2$$

e a soma dos quadrados dos erros dada por

$$SQE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \overline{Y}_i\right)^2 = \sum_{i=1}^{k} \left(n_i - 1\right) S_i^2,$$

onde S_i^2 é a estimativa da variância da amostra i, Y_{ij} representa a observação ij, $\overline{Y}_i$ é a média da amostra i, n_i é a dimensão da amostra i e $\overline{Y}$ é a média global da amostra.

Regra de decisão: Rejeitar H_0 se $F_{\text{observado}} \geq F\left(k-1; n-k; 1-\alpha\right)$ ou se $\alpha \geq p - value$.

No caso de não se verificar algum ou alguns dos pressupostos necessários para a utilização da ANOVA devemos utilizar um teste não paramétrico. Neste tipo de situação podemos recorrer ao teste de Kruskal-Wallis. Alguns autores referem no entanto que a ANOVA é um teste bastante resistente á violação de pressupostos, dando resultados muito semelhantes aos do teste de Kruskall-Wallis.

2.5 Testes de comparação múltipla

Quando realizamos o teste ANOVA pretendemos testar se as médias dos grupos são todas iguais, ou se existem pelo menos duas diferentes. Quando se deteta que existe pelo menos um par de médias diferentes não é possível identificar quais são. Para se conseguir essa informação é necessário utilizar testes de comparação múltipla.

Existem diversos testes de comparação múltipla, mas nesta dissertação vão abordar-se apenas dois deles:

- Teste HSD de Tukey;
- Teste de Scheffé.

Em ambos os testes, as hipóteses em teste são as seguintes:

$H_0 : \mu_i = \mu_j$

$H_1 : \mu_i \neq \mu_j$, com $i \neq j$; $i, j = 1, \ldots, k$

2.5.1 Teste HSD de Tukey

Este teste é mais preciso quando as amostras têm igual dimensão. O teste de Tukey consiste em comparar todos os pares de médias possíveis, baseando-se na diferença mínima significativa.

Estatística de teste:

$$W = \frac{\left|\overline{Y}_i - \overline{Y}_j\right|}{\sqrt{\frac{QME}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim q\left(k, n-k\right),$$

sendo $q\left(k, n-k\right)$ a distribuição Studentized Range com k e $n-k$ graus de liberdade.

Regra de decisão: Rejeitar H_0 se $\left|\overline{Y}_i - \overline{Y}_j\right| \geq q\left(k, n-k, 1-\alpha\right)\sqrt{\frac{QME}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$, sendo $q\left(k, n-k, 1-\alpha\right)$ o quantil de probabilidade $1-\alpha$ da distribuição Studentized Range, ou se $\alpha \geq p - value$.

2.5.2 Teste de Sheffé

O teste de Sheffé permite a utilização de amostras com dimensões diferentes. É um teste mais robusto no que diz respeito aos pressupostos de normalidade e igualdade das variâncias.

Estatística de teste:

$$F = \frac{\left|\overline{Y}_i - \overline{Y}_j\right|}{\sqrt{\frac{QME}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim \sqrt{\left(k-1\right)F\left(k-1, n-k\right)}.$$

Regra de decisão: Rejeitar H_0 se $\left|\overline{Y}_i - \overline{Y}_j\right| \geq \sqrt{\left(k-1\right)F\left(k-1, n-k, 1-\alpha\right)\frac{QME}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$, sendo $F\left(k-1, n-k, 1-\alpha\right)$ o quantil de probabilidade $1-\alpha$ da distribuição F-Snedecor, ou se $\alpha \geq p - value$.

2.6 Teste de Kruskall-Wallis

O teste de Kruskal-Wallis é usado para testar se duas ou mais amostras provêm de populações com a mesma distribuição e para isso é necessário comparar dois ou mais grupos independentes, quando são violados os pressupostos de normalidade ou homogeneidade das variância, ou seja, os pressupostos do teste ANOVA.

As hipóteses são:

H_0 : A distribuição dos valores da variável é idêntica nas k populações ou, de outra forma, $F(X_1) = F(X_2) = \cdots = F(X_k)$

H_1 : Existe pelo menos uma população onde a distribuição da variável é diferente de uma das outras populações em estudo ou, de outra forma, $\exists i, j : F(X_i) \neq F(X_j)$, com $i \neq j$; $i, j = 1, \ldots, k$.

Sendo, no entanto, este teste particularmente sensível a diferenças de localização, as hipóteses deste teste são, frequentemente, escritas como:

$H_0 : \widetilde{\mu}_1 = \widetilde{\mu}_2 = \cdots = \widetilde{\mu}_k$

$H_1 : \exists i, j : \widetilde{\mu}_i \neq \widetilde{\mu}_j$, com $i \neq j$; $i, j = 1, \ldots, k$,

sendo $\widetilde{\mu}_i$ a mediana da i-ésima população. No entanto deve ter-se alguma cautela quando se usa esta formulação porque se as distribuições das populações forem iguais então as medianas são iguais, mas o contrário nem sempre se verifica, ou seja, se as medianas forem iguais, as distribuições das populações podem ser iguais ou não. Para comparar as medianas devemos primeiro certificar-nos que as funções de distribuição são pelo menos idênticas.

Para calcular a estatística de teste devemos começar por ordenar, por ordem crescente, todas as n observações das diferentes amostras, obtendo uma amostra global, atribuindo a cada observação a sua ordem na amostra global e mantendo a origem da observação, ou seja, de que amostra provém essa observação.

No caso de existirem observações com o mesmo valor, a ordem dessas observações é dada pela média aritmética das ordens que essas observações teriam se não fossem iguais.

A estatística de teste, caso não hajam observações iguais, é dada por

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} \frac{R_j^2}{n_j} - 3(n+1) \sim \chi^2_{k-1}.$$

Caso hajam observações iguais a estatística de teste é dada por

$$H_E = \frac{H}{1 - \frac{\sum_{i=1}^{g} \left(t_i^3 - t_i\right)}{n^3 - n}} \sim \chi^2_{k-1},$$

onde R_j representa a soma das ordens em cada uma das j amostras e $n = n_1 + n_2 + \cdots + n_k$. A expressão

$$1 - \frac{\sum_{i=1}^{g} \left(t_i^3 - t_i\right)}{n^3 - n}$$

é uma correção necessária caso existam mais de dois grupos com observações iguais. Nesta expressão, g representa o número de grupos com observações iguais e t o número de observações em cada grupo de observações iguais.

Regra de decisão: Rejeitar H_0 se $H_{\text{observado}} \geq \chi^2_{k-1;1-\alpha}$, sendo $\chi^2_{k-1;1-\alpha}$ o quantil de probabilidade $1 - \alpha$ da distribuição Qui-quadrado, ou se $\alpha \geq p - value$.

2.7 Teste de Nemenyi

O teste de Nemenyi é um teste estatístico não paramétrico utilizado para realizar comparações múltiplas entre pares de amostras, especialmente após a rejeição da hipótese nula na realização do teste de Kruskall-Wallis. Este teste é uma alternativa para comparação múltipla quando não se pode assumir a normalidade dos dados. Este teste é usado após se ter verificado no teste Kruskall-Wallis uma diferença significativa entre pelo menos a distribuição de duas populações, pois o teste de Kruskall-Wallis só indica que existe a diferença entre as várias distribuições das k populações, mas não identifica entre que pares isso ocorre. O teste de Nemenyi vai indicar quais os pares onde se verificam as diferenças significativas.

Para identificar em qual ou quais dos grupos as distribuições são significativamente diferentes é necessário proceder à comparação múltipla das médias das ordens, também designada, em alguns casos de forma algo abusiva, de comparação múltipla das medianas.

As hipóteses em teste são:

$H_0 : F(X_i) = F(X_j)$

$H_1 : F(X_i) \neq F(X_j)$, com $i \neq j$; $i, j = 1, \ldots, k$.

Estatística de teste:

$$q = \frac{\mid \overline{R}_i - \overline{R}_j \mid}{\sqrt{\frac{k(n+1)}{12}}} \sim q(\infty, k),$$

se as amostras têm todas a mesma dimensão ou

$$q = \frac{\mid \overline{R}_i - \overline{R}_j \mid}{\sqrt{\frac{n(n+1)}{24}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \sim q(\infty, k),$$

se pelo menos uma das amostras tem dimensão diferente, sendo $\overline{R}_i = \frac{R_i}{n_i}$ a ordenação média das observações da amostra i. A estatística de teste q tem distribuição Studentized Range, com graus de liberdade iguais a ∞. Esta distribuição encontra-se tabelada.

Regra de decisão: Rejeitar H_0 se $\left|\overline{R}_i - \overline{R}_j\right| \geq q(\infty, k, 1-\alpha)\sqrt{\frac{n(n+1)}{24}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$, sendo $q(\infty, k, 1-\alpha)$ o quantil de probabilidade $1-\alpha$ da distribuição Studentized Range, ou se $\alpha \geq p - value$.

2.8 Análise discriminante de dados

Um dos pressupostos da análise discriminante é a normal multivariada, a qual pode ser verificada, por exemplo, através do teste de Shapiro-Wilk multivariado. No entanto, a análise discriminante é uma técnica da estatística multivariada utilizada quando se pretende estudar uma variável dependente de natureza qualitativa (ou categórica), com base em variáveis independentes quantitativas.

O principal objetivo deste método é identificar quais variáveis explicativas melhor distinguem os diferentes grupos existentes. Assim, ao conhecer as características de um novo caso, torna-se possível prever a que grupo ele pertence.

Para isso, são construídas funções discriminantes - combinações lineares das variáveis explicativas - que procuram maximizar a separação entre os grupos (aumentando as diferenças entre as suas médias) e, ao mesmo tempo, reduzir a probabilidade de erros na classificação dos casos.

2.8.1 Pressupostos

A análise discriminante mantém-se robusta mesmo quando a suposição de normalidade multivariada não é totalmente cumprida, desde que os grupos analisados tenham dimensões semelhantes, com pelo menos vinte elementos cada.

Nos casos em que as dimensões dos grupos são desiguais, recomenda-se que o grupo com menor número de observações tenha, no mínimo, vinte casos, especialmente quando o modelo inclui até cinco variáveis independentes.

No âmbito da análise discriminante, assume-se como pressuposto fundamental a homogeneidade das matrizes de variâncias-covariâncias entre os grupos, ou seja, pressupõe-se que os grupos apresentam uma variabilidade semelhante. Esta condição pode ser testada estatisticamente através do Teste M de Box.

Contudo, é frequente que, em amostras de grande dimensão, este teste conduza à rejeição da hipótese de homogeneidade das referidas matrizes, mesmo quando as discrepâncias entre elas são marginalmente significativas do ponto de vista prático.

Apesar desta limitação, a análise discriminante revela-se suficientemente robusta à violação deste pressuposto, desde que se verifiquem duas condições essenciais:

(i) O grupo com menor dimensão amostral tem um número de observações superior ao número de variáveis independentes consideradas no modelo;

(ii) Não existe proporcionalidade entre as médias dos grupos e as respetivas variâncias.

A estes dois pressupostos metodológicos poderão ser ainda acrescentados os seguintes pressupostos:

- Existir um critério pré-definido que nos permite dividir os indivíduos em dois ou mais grupos;
- O número de indivíduos em cada grupo é pelo menos dois;
- O número de variáveis discriminantes poderá ser qualquer, desde que verifique a condição de ser menos que o número total de indivíduos menos dois;
- Nenhuma das variáveis discriminantes poderá ser combinação linear das restantes.

2.8.2 Função discriminante

A análise discriminante é realizada através de uma ou mais combinações lineares das variáveis independentes utilizadas (X_i). Cada combinação linear (Y_i) constitui uma função discriminante:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p,$$

em que:

- a_{ij} são os coeficientes de ponderação;
- X_j são as variáveis discriminantes não normalizadas.

Idealmente, uma função discriminante deverá assumir valores semelhantes para todos os indivíduos pertencentes ao mesmo grupo, refletindo a homogeneidade interna de cada categoria.

As funções discriminantes são construídas de forma a maximizar a separação entre os grupos, isto é, procuram acentuar as diferenças entre as categorias com base nas combinações lineares das variáveis independentes.

Uma vez estimadas, estas funções permitem alcançar os dois objetivos centrais da análise discriminante: a análise e a classificação.

2.8.3 Estimação das funções discriminantes

Dadas p variáveis e k grupos é possível estabelecer:

$$m = \min\{k-1; p\}$$

funções discriminantes que são combinação linear das p variáveis:

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p.$$

Os coeficientes a_{ij}, associados à função discriminante, permitem identificar as variáveis independentes que mais contribuem para a diferenciação entre os grupos, no âmbito de uma dada função Y_i. Quanto maior for o valor absoluto de a_{ij} maior será a influência da variável correspondente, na capacidade discriminativa da função.

A função resultante desta combinação linear de variáveis é designada por função discriminante linear de Fisher, sendo construída com o propósito de maximizar a separação entre os grupos previamente definidos, com base nas variáveis explicativas disponíveis.

Em termos matriciais, a função discriminante pode escrever-se como:

$$\mathbf{Y} = \boldsymbol{X}'\boldsymbol{a},$$

onde $\boldsymbol{X}'$ é a matriz transposta da matriz $\boldsymbol{X}$, com as p variáveis e $\boldsymbol{a}$ é o vector dos pesos.

A matriz:

$$\mathbf{T} = \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}\right)\left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}\right)'$$

é a matriz da soma de quadrados e produtos cruzados totais da matriz $\boldsymbol{X}$, com p variáveis e $\overline{\boldsymbol{X}}$ é o vector das médias totais de cada variável.

Para cada grupo j, a matriz da soma dos quadrados e produtos cruzados dentro dos grupos é:

$$\mathbf{W}_j = \sum_{i=1}^{n_j}\left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_j\right)\left(\boldsymbol{X}_{ij} - \overline{\boldsymbol{X}}_j\right)',$$

com

$$\mathbf{W} = \sum_{j=1}^{k}\mathbf{W}_j$$

ou

$$\mathbf{W} = (n_1 + \cdots + n_k - k)\,\mathbf{S}_T,$$

sendo

$$\mathbf{S}_T = \frac{1}{(n_1 + \cdots + n_k - k)}\left[(n_1 - 1)\,\mathbf{S}_1 + \cdots + (n_k - 1)\,\mathbf{S}_k\right]$$

a matriz total de variâncias-covariâncias.

A matriz da soma de quadrados e produtos cruzados entre os grupos é dada por:

$$\mathbf{B} = \mathbf{T} - \mathbf{W} = \mathbf{T} - \sum_{j=1}^{k}\mathbf{W}_j$$

ou

$$\mathbf{B} = \sum_{j=1}^{k}\left(\overline{\boldsymbol{X}}_j - \overline{\boldsymbol{X}}\right)\left(\overline{\boldsymbol{X}}_j - \overline{\boldsymbol{X}}\right)',$$

sendo $\overline{\boldsymbol{X}}_j$ o vector das médias em cada grupo de cada variável.

Sendo $\mathbf{T} = \mathbf{B} + \mathbf{W}$ a soma dos quadrados totais para a função discriminante, pode agora escrever-se $\mathbf{Y}'\mathbf{Y}$ como:

$$\mathbf{Y}'\mathbf{Y} = \boldsymbol{a}'\mathbf{T}\boldsymbol{a} = \boldsymbol{a}'\left(\mathbf{B} + \mathbf{W}\right)\boldsymbol{a} = \boldsymbol{a}'\mathbf{B}\boldsymbol{a} + \boldsymbol{a}'\mathbf{W}\boldsymbol{a}.$$

Uma vez que $\boldsymbol{a}'\mathbf{B}\boldsymbol{a}$ e $\boldsymbol{a}'\mathbf{W}\boldsymbol{a}$ são, respectivamente, a soma de quadrados entre os grupos e a soma de quadrados dentro dos grupos para a função discriminante Y, a obtenção da função discriminante resume-se a encontrar o vector $\boldsymbol{a}$ tal que:

$$\lambda = \frac{\boldsymbol{a}'\mathbf{B}\boldsymbol{a}}{\boldsymbol{a}'\mathbf{W}\boldsymbol{a}}$$

seja máximo, pois o objectivo é maximizar as diferenças entre grupos e minimizar a variação dentro dos grupos.

A solução deste problema de maximização ocorre para:

$$\left(\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}\right)\boldsymbol{a} = 0,$$

com a restrição

$$\mid \mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I} \mid = 0.$$

Este problema tem:

$$m = \min\left\{k - 1; p\right\}$$

soluções que correspondem aos valores próprios da matriz $\mathbf{W}^{-1}\mathbf{B}$, onde k representa o número de grupos e p o número de variáveis discriminantes. O maior valor próprio λ_1 tem um vetor próprio correspondente à primeira função discriminante, o segundo maior valor próprio λ_2 está ligado ao vetor próprio da segunda função discriminante, e assim sucessivamente. Cada função discriminante obtida é ortogonal às anteriores, ou seja, os seus scores não são correlacionados. Os coeficientes a_{ij} das funções discriminantes correspondem às componentes dos vetores próprios associados aos respetivos valores próprios da matriz $\mathbf{W}^{-1}\mathbf{B}$.

2.8.4 Importância de cada função discriminante

Uma forma de avaliar a importância de cada função discriminante é através da percentagem do valor próprio a ela associado, dado que a soma de todos os valores próprios representa a variância total explicada pelas variáveis originais. Assim, a importância relativa de cada função discriminante pode ser determinada pelo quociente entre o seu valor próprio e o somatório dos valores próprios:

$$\frac{\lambda_j}{\sum_{i=1}^{m}\lambda_i},$$

com $j = 1, \ldots, m$. Este quociente indica a proporção da variabilidade entre os grupos que é explicada pela respetiva função discriminante linear. Os valores próprios resultam do rácio entre a variância entre os grupos e a variância dentro dos grupos. Quanto mais elevado for o valor próprio, maior será a quantidade de variação explicada por essa função discriminante.

2.8.5 Testes de hipóteses para as funções discriminantes

A avaliação da significância estatística das funções discriminantes é realizada através do teste lambda de Wilks, cuja estatística segue, sob certas condições, uma distribuição aproximada do

tipo qui-quadrado. Este teste permite determinar quais funções discriminantes contribuem de forma significativa para a separação entre os grupos.

Desta forma, torna-se possível identificar as funções estatisticamente relevantes, permitindo selecionar apenas aquelas que devem ser consideradas na interpretação e utilização do modelo discriminante. Wilks definiu um teste para a igualdade de médias dos k grupos, a partir dos valores próprios λ_j da matriz $\mathbf{W}^{-1}\mathbf{B}$, com estatística de teste dada por:

$$\Lambda = \prod_{j=1}^{m} \left(\frac{1}{1+\lambda_j} \right).$$

Para este teste têm-se as hipóteses:

$$H_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_m = 0$$
$$H_1 : \exists \lambda_j \neq 0, \text{ com } j = 1, \ldots, m$$

A distribuição de Λ não é conhecida, mas a transformação proposta por Bartlett segue uma distribuição Qui-quadrado com $p\,(k-1)$ graus de liberdade:

$$V = -\left[(n-1) - \frac{1}{2}\,(p+k) \right] \ln \Lambda = \left[(n-1) - \frac{1}{2}\,(p+k) \right] \sum_{j=1}^{m} \ln\,(1+\lambda_j)\,.$$

Regra de decisão estatística: Rejeitar H_0 se $V_{\text{observado}} \geq \chi^2_{p(k-1);1-\alpha}$, sendo $\chi^2_{p(k-1);1-\alpha}$ o quantil de probabilidade $1-\alpha$ da distribuição Qui-quadrado, ou se $\alpha \geq p - value$, com $p - value = P\,[V \geq V.E.T.]$, sendo $V.E.T.$ o valor da estatística de teste e, neste caso, a região crítica à direita. Quando se rejeita H_0 pode concluir-se que a solução discriminante é estatisticamente significativa, o que significa que as p variáveis discriminantes têm elevado poder para discriminar os k grupos de indivíduos antes de remover qualquer função discriminante.

A rejeição da hipótese nula, H_0, no teste de Wilks permite concluir que a solução discriminante é estatisticamente significativa. Isto indica que o conjunto das p variáveis discriminantes possui um elevado poder discriminatório, sendo eficaz na separação dos k grupos de indivíduos considerados, antes de ser removida qualquer função discriminante do modelo.

Também é usual realizar-se um teste de significância para cada uma das funções discriminantes. Assim, as hipóteses para um teste de significância para a j-ésima função discriminante são as seguintes:

$$H_0 : \lambda_j = 0$$
$$H_1 : \lambda_j \neq 0$$

A estatística de teste é dada por:

$$V_j = \left[(n-1) - \frac{1}{2}\,(p+k) \right] \ln\,(1+\lambda_j)$$

e segue uma distribuição Qui-quadrado com $p+k-2j$ graus de liberdade.

Regra de decisão estatística: Rejeitar H_0 se $V_{\text{observado}} \geq \chi^2_{p+k-2j;1-\alpha}$, sendo $\chi^2_{p+k-2j;1-\alpha}$ o quantil de probabilidade $1-\alpha$ da distribuição Qui-quadrado, ou se $\alpha \geq p - value$, com $p - value = P\,[V \geq V.E.T.]$, sendo, neste caso, a região crítica à direita. Quando se rejeita H_0 pode concluir-se que a j-ésima função discriminante é significativa.

2.8.6 Classificação dos indivíduos em k grupos

A análise discriminante constitui uma técnica estatística de classificação particularmente relevante, permitindo identificar o grupo mais provável, entre k existentes, ao qual um determinado indivíduo pertence, com base nos seus valores observados para as variáveis discriminantes.

Embora as funções discriminantes possam ser utilizadas diretamente para a classificação dos indivíduos, esta abordagem torna-se menos clara e mais complexa quando o número de grupos é superior a dois. Nesses casos, a utilização de funções de classificação revela-se mais adequada, proporcionando regras de decisão mais explícitas e eficazes na atribuição dos indivíduos aos respetivos grupos.

Quando se consideram k grupos, é necessário definir k funções de classificação, também designadas por funções classificatórias.

Assumindo que as matrizes de variâncias-covariâncias são iguais para todos os grupos e que as populações associadas a cada grupo seguem uma distribuição normal multivariada, a função de classificação para o grupo j é dada por:

$$C_j = C_{j0} + C_{j1}X_1 + C_{j2}X_2 + \cdots + C_{jp}X_p = C_{j0} + \boldsymbol{C}'_j\boldsymbol{X}.$$

Os coeficientes podem ser obtidos a partir da matriz total de variâncias-covariâncias amostral, $\mathbf{S}_T$, e do centróide do grupo j, $\overline{\boldsymbol{X}}_j$, o vector das p médias das variáveis discriminantes para o grupo j, do seguinte modo:

$$\boldsymbol{C}'_j = \overline{\boldsymbol{X}}'_j\mathbf{S}_T^{-1}$$

e

$$C_{j0} = -\frac{1}{2}\boldsymbol{C}'_j\overline{\boldsymbol{X}}_j = -\frac{1}{2}\overline{\boldsymbol{X}}'_j\mathbf{S}_T^{-1}\overline{\boldsymbol{X}}_j = -\frac{1}{2}\boldsymbol{C}'_j\overline{\boldsymbol{X}}_j,$$

onde

$$\mathbf{S}_T = \frac{1}{n_1 + n_2 + \cdots + n_k - k}\left[(n_1 - 1)\,\mathbf{S}_1 + (n_2 - 1)\,\mathbf{S}_2 + \cdots + (n_k - 1)\,\mathbf{S}_k\right].$$

Para cada indivíduo a ser classificado, calcula-se um score classificatório para cada grupo, obtido ao aplicar a função de classificação correspondente aos valores das variáveis discriminantes desse indivíduo.

Quando os grupos não apresentam dimensões idênticas ou probabilidades a priori p_j iguais, é necessário introduzir nas funções de classificação um ajuste que considere a dimensão relativa de cada grupo, de forma a refletir adequadamente a sua representatividade:

$$C_j = C_{j0} + \boldsymbol{C}'_j\boldsymbol{X} + \ln\left(\frac{n_j}{n}\right)$$

ou à probabilidade a priori

$$C_j = C_{j0} + \boldsymbol{C}'_j\boldsymbol{X} + \ln p_j.$$

As funções classificatórias, ou de classificação, podem ser utilizadas para:

- Avaliar a eficácia do modelo de análise discriminante na correta atribuição dos indivíduos aos grupos;
- Classificar novos indivíduos nos grupos previamente definidos pelo modelo.

Para avaliar a eficiência do processo de classificação podem calcular-se probabilidades de classificação incorrecta.

Existem vários métodos para estimar estas probabilidades, nomeadamente:

- Método de resubstituição;
- Método de validação cruzada;
- Método de Jacknife.

No método de resubstituição utilizam-se as funções de classificação para atribuir cada indivíduo a um grupo e registar quando essa atribuição coincide com o grupo original a que o indivíduo pertence. Este procedimento consiste na construção de uma matriz de classificação (ou matriz de confusão), que compara os grupos iniciais, previamente definidos, com os grupos atribuídos a posteriori, resultantes da aplicação da análise discriminante. As linhas desta matriz correspondem aos grupos originais, enquanto as colunas representam os grupos previstos pela análise discriminante. O elemento genérico n_{ij} da matriz indica o número de indivíduos que pertenciam originalmente ao grupo i e que foram classificados no grupo j pela análise discriminante. Além disso, é possível incluir nesta matriz os resultados da classificação de indivíduos cujo grupo inicial é desconhecido, permitindo assim a atribuição destes indivíduos aos grupos mais prováveis.

O método de validação cruzada é usado quando a dimensão da amostra total é suficientemente grande para permitir a sua divisão em duas subamostras - uma de estimação e outra de validação. Pode-se desenvolver o processo classificativo com base na primeira subamostra e avaliar a sua eficácia através da segunda. Os indivíduos da subamostra de validação são classificados utilizando as funções discriminantes derivadas da subamostra de estimação, possibilitando a construção da matriz de classificações que permite comparar a classificação prevista com os grupos originais.

O método de Jacknife é um método alternativo para situações em que a amostra é demasiado pequena para ser dividida em duas subamostras. Neste procedimento, exclui-se uma observação de um determinado grupo e as funções discriminantes são construídas com base nas observações remanescentes. Em seguida, classifica-se a observação excluída utilizando as funções discriminantes obtidas. Este processo é repetido sucessivamente para cada observação da amostra original, permitindo uma avaliação rigorosa da capacidade classificativa do modelo.

2.8.7 Violação dos pressupostos

Quando os pressupostos da análise discriminante - nomeadamente, a normalidade das variáveis e a igualdade das matrizes de variâncias-covariâncias entre os grupos - são satisfeitos, as funções discriminantes lineares produzem resultados ótimos, minimizando as taxas de erro de classificação.

Se a percentagem de classificações corretas for elevada, a violação desses pressupostos não compromete significativamente a validade do modelo.

Por outro lado, quando a taxa de classificações corretas é baixa, torna-se difícil determinar se essa situação resulta da violação dos pressupostos ou do fraco poder discriminatório das variáveis utilizadas.

Especificamente, a violação da igualdade das matrizes de variâncias-covariâncias pode afetar negativamente o desempenho da análise discriminante. Nestes casos, se a normalidade das variáveis for mantida, mas a homogeneidade das matrizes não se verificar, é recomendada a utilização de funções discriminantes quadráticas em substituição das lineares.

Capítulo 3

Resultados

Todo o estudo estatístico apresentado neste capítulo foi realizado recorrendo ao software estatístico R. As informações contidas neste capítulo foram retiradas de livros (Alkarkhi & Alqaraghuli, 2020; Dalgaard, 2008; Venables & Smith, 2025; Zelterman, 2015) e apontamentos das disciplinas de Complementos de Estatística para a Engenharia do Mestrado de Engenharia da Qualidade e Ambiente, de Estatística Multivariada da Licenciatura em Matemática Aplicada à Tecnologia e à Empresa e de Técnicas de Estatística Multivariada das Licenciaturas e apontamentos das disciplinas de Complementos de Estatística para a Engenharia do Mestrado de Engenharia da Qualidade e Ambiente, de Estatística Multivariada da Licenciatura em Matemática Aplicada à Tecnologia e à Empresa e de Técnicas de Estatística Multivariada das Licenciaturas em Engenharia Biomédica, em Engenharia Informática e de Computadores e em Engenharia Química e Biológica (Fernandes & Ramos, P, 2025e) do Instituto Superior de Engenharia de Lisboa.

O estudo foi feito utilizando os poluentes CO, NO_2 e PM_{10}. Para cada poluente foram analisados seis períodos temporais diferentes que são os seguintes:

- $nxst$ - período compreendido entre 1 janeiro de 2001 e 30 de junho de 2003 e corresponde ao período em que ainda não existia nenhum túnel construído.
- $obin$ - corresponde ao período em que foi realizada a obra inicial e está compreendido entre 1 de julho 2003 e 30 de abril de 2007.
- $fpar$ - período em que o túnel se encontrava em funcionamento parcial e está compreendido entre 1 de maio de 2007 e 31 de março de 2010.
- $obam$ - período em que foram realizadas as obras de ampliação do túnel e está compreendido entre 1 de abril de 2010 e 31 de março de 2012.
- $fpln$ - período de funcionamento pleno do túnel e está compreendido entre 1 de abril de 2012 e 20 de março de 2020.
- $covd$ - período em que ocorreu o confinamento devido à pandemia provocada pelo vírus SARS-CoV-2 e está compreendido entre 21 de março de 2020 e 31 de dezembro de 2022.

Em todos os testes de hipóteses realizados ao longo deste capítulo foi considerado um nível de significância de $\alpha = 0,05$ e foi considerada a regra de decisão baseada no valor do $p-value$ do respetivo teste:

$$\text{Rejeitar } H_0 \text{ se e só se } \alpha \geq p - value.$$

3.1 Estatística descritiva para CO

	nxst	obin	fpar	obam	fpln	covd
nobs	63138.0000	63138.0000	63138.0000	63138.0000	63138.0000	63138.0000
NAs	43201.0000	34481.0000	38515.0000	46587.0000	0.0000	31211.0000
Minimum	0.0000	0.0170	0.0410	0.0360	0.0120	0.0630
Maximum	6.9800	5.3780	5.0680	2.4840	3.7860	2.2260
1. Quartile	0.3310	0.3090	0.2650	0.2380	0.2000	0.1920
3. Quartile	0.8540	0.7160	0.5330	0.4555	0.3660	0.3400
Mean	0.6805	0.5867	0.4443	0.3808	0.3161	0.2916
Median	0.5450	0.4690	0.3760	0.3250	0.2650	0.2530
Sum	13567.3100	16813.3890	10940.7250	6302.4370	19956.0840	9310.2340
SE Mean	0.0039	0.0026	0.0018	0.0018	0.0008	0.0009
LCL Mean	0.6728	0.5816	0.4407	0.3774	0.3145	0.2898
UCL Mean	0.6882	0.5919	0.4480	0.3842	0.3176	0.2934
Variance	0.3048	0.1971	0.0842	0.0507	0.0388	0.0264
Stdev	0.5521	0.4439	0.2901	0.2252	0.1970	0.1626
Skewness	3.0379	2.7320	3.1541	2.4048	3.5927	2.9805
Kurtosis	16.0226	11.9994	18.7211	9.1766	24.6484	14.9897

Figura 3.1: Medidas de estatística descritiva das concentrações de CO

Na Figura 3.1 verificamos uma diminuição na concentração de CO, ao nível dos valores máximos observados, no 1^{o} e 3^{o} quartis, na média, na mediana, na variância, no desvio padrão e nos limites do intervalo de confiança para a média, ao longo do tempo, ou seja, do período mais longínquo para o período mais atual. Todos os períodos têm assimetria positiva, ou seja, existe uma maior concentração de valores na faixa dos valores mais baixos da amostra.

Relativamente à curtose ou achatamento, temos uma curva leptocúrtica, ou seja, menos achatada quando comparada com curva da distribuição normal. Existe uma maior concentração de valores em torno da média do que na distribuição normal e há uma maior concentração de valores extremos ou discrepantes nas caudas (possivelmente *outliers*).

Na Figura 3.2 observa-se que a média da concentração de CO tem vindo a diminuir ao longo do tempo.

Na Figura 3.1, observa-se, ainda, uma diminuição dos valores do desvio padrão ao longo dos vários períodos, o que aparenta indicar uma menor variabilidade dos dados ao longo do tempo. No entanto, na Figura 3.3 observam-se vários outliers nos valores superiores, em todos os períodos.

Período	**Coeficientes de variação**	**Coeficientes de variação resistente**
$nxst$	81.1241	95.9633
$obin$	75.6621	86.7804
$fpar$	65.2994	71.2766
$obam$	59.1372	66.9231
$fpln$	62.3226	62.6415
$covd$	55.775	58.498

Tabela 3.1: Coeficientes de variação e de variação resistente

Na Tabela 3.1 apresentamos os valores dos coeficientes de variação e de variação resistente. Apesar do desvio padrão ter vindo a diminuir ao longo do tempo, os valores dos dois coeficientes

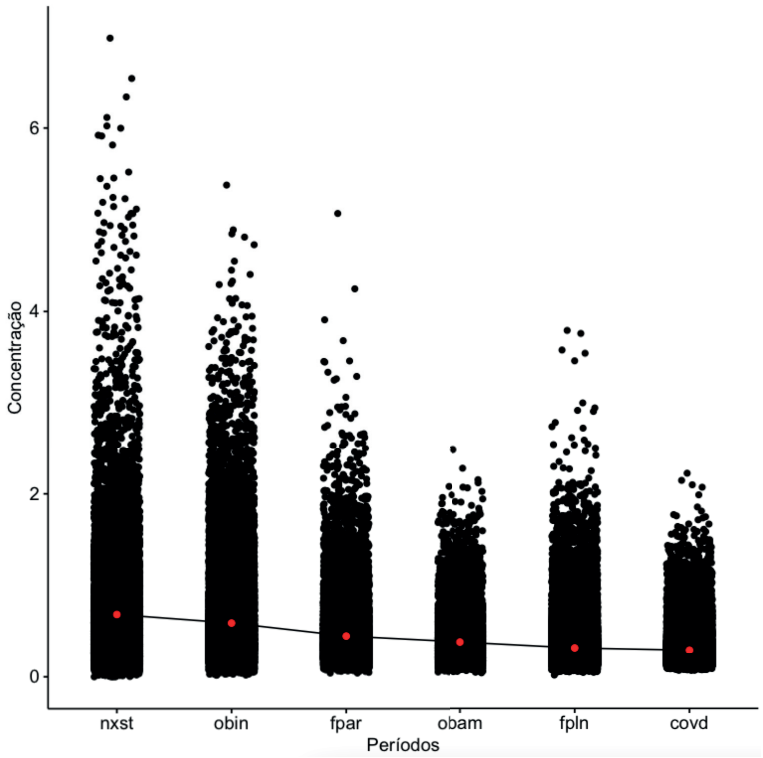

Figura 3.2: Distribuição gráfica dos valores por período, com representação da média

de variação são sempre muito elevados, apresentando também ligeiras diminuições. Os valores obtidos indicam uma grande variabilidade das concentrações de CO, sendo justificada no caso do coeficiente de variação, em grande parte, pela existência de outliers.

3.2 ANOVA

Vamos verificar se os pressupostos da ANOVA são verificados. A independência dos grupos (períodos) está assegurada por se tratarem de períodos mutuamente exclusivos. Nas secções que se seguem iremos analisar a normalidade e a homogeneidade dos grupos (períodos).

3.2.1 Normalidade dos grupos (períodos)

O gráfico de comparação de quantis ou *Q-Q Plot*, é um gráfico que nos permite avaliar se um conjunto de dados provém de uma distribuição teórica como, por exemplo, a distribuição normal. Permite uma avaliação meramente visual se os dados em estudo seguem uma determinada distribuição, não servindo de prova e sendo sempre necessária a realização de um teste de hipóteses para uma verificação analítica.

No gráfico de comparação de quantis, obtido para os dados do período correspondente à fase plena, que se apresenta na Figura 3.4, representa-se os pares ordenados dos quantis de duas distribuições, a distribuição normal e a distribuição empírica. Se os dois conjuntos de quantis fossem provenientes da mesma distribuição, o conjunto de pontos obtido formaria uma linha aproximadamente reta. Observando o gráfico conclui-se que os dados da concentração de CO, neste período, não parecem ser provenientes de uma população com distribuição normal.

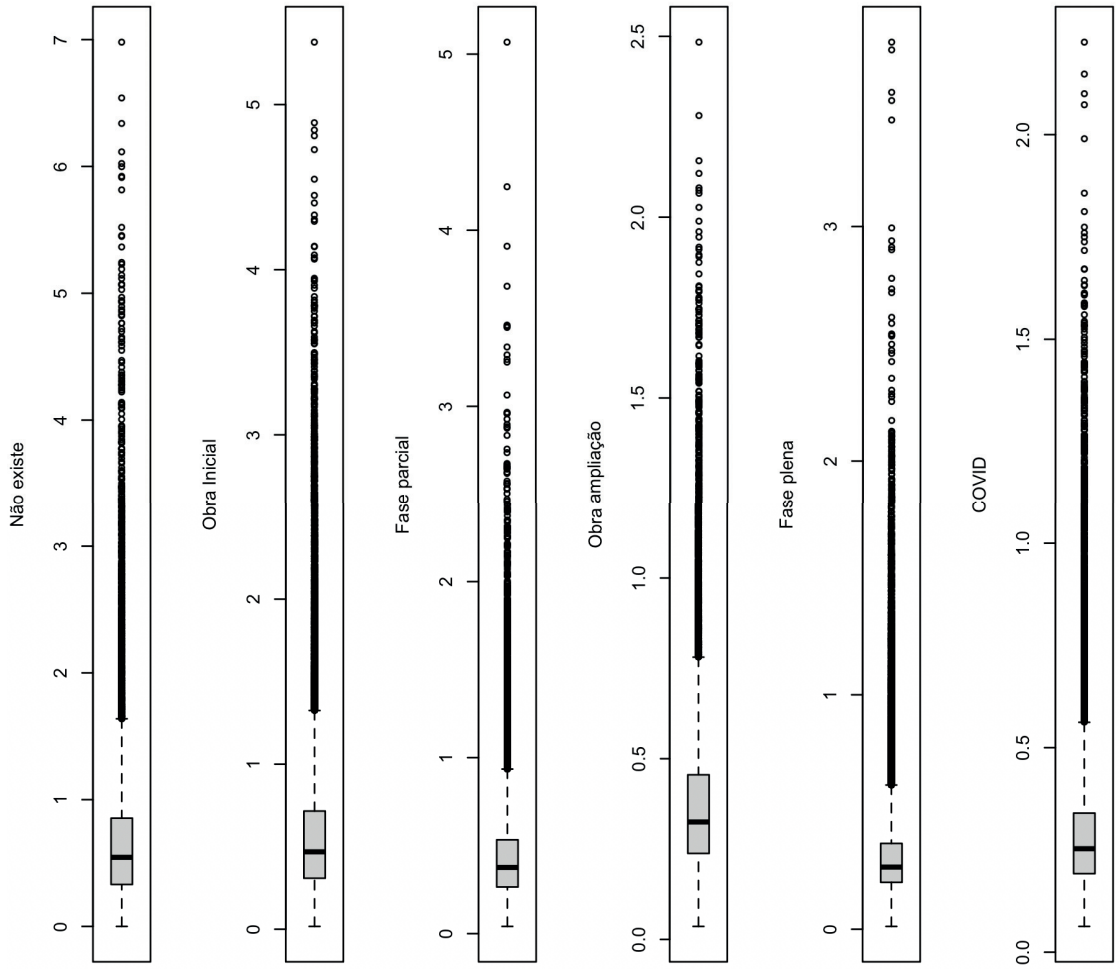

Figura 3.3: Diagrama de extremos e quartis

Para confirmar a perceção de não normalidade dos grupos vamos usar o teste de Kolmogorov-Smirnov com correção de Lilliefors. As hipóteses em teste são as seguintes:

H_0: A amostra do i-ésimo grupo é proveniente de uma população com distribuição normal

H_1: A amostra do i-ésimo grupo é proveniente de uma população com distribuição diferente da distribuição normal,

com $i = nxst, obin, fran, obam, fpln, covd$.

Na Tabela 3.2 são apresentados os valores da estatística de teste e do $p-value$, por período, obtidos recorrendo ao *software* estatístico R.

Como se pode verificar, na Tabela 3.2, o valor dos $p-value$ é aproximadamente zero em todos os períodos, pelo que, devemos rejeitar H_0 para todos esses períodos. Assim, concluí-se que as observações de cada um dos grupos não são provenientes de populações com distribuição normal, como se suspeitava.

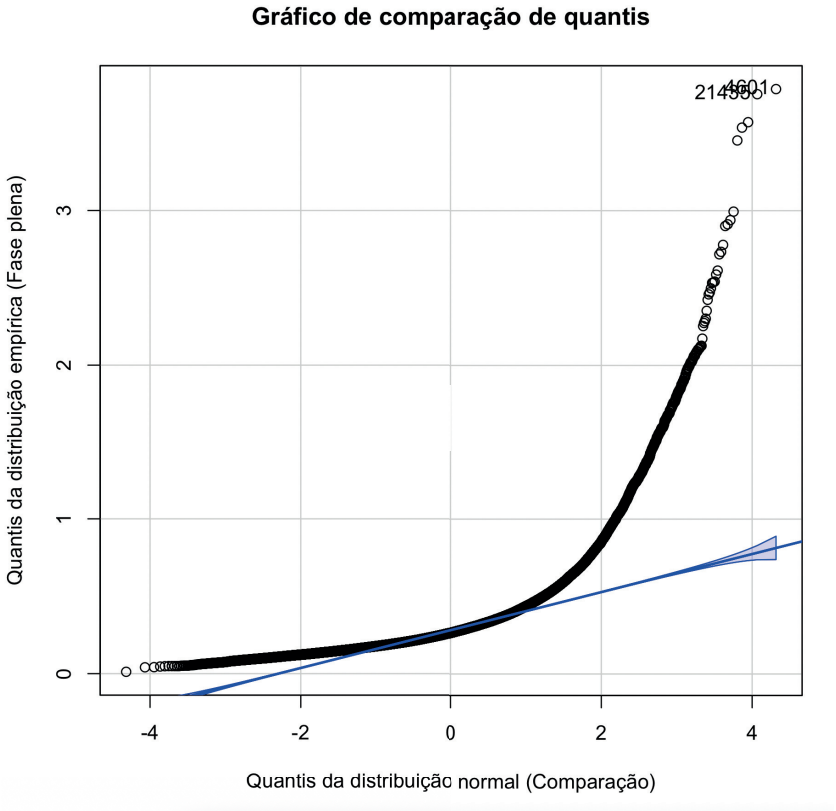

Figura 3.4: Gráfico de comparação de quantis da fase plena

Período	**Valor da estatística de teste**	**Valor do** $p - value$
$nxst$	0.13218	$\simeq 0$
$obin$	0.14545	$\simeq 0$
$fpar$	0.13329	$\simeq 0$
$obam$	0.13123	$\simeq 0$
$fpln$	0.15477	$\simeq 0$
$covd$	0.14003	$\simeq 0$

Tabela 3.2: Valor da estatística de teste e do $p - value$ para o teste de Kolmogorov-Smirnov com correção de Lilliefors

3.2.2 Homogeneidade das variâncias

Nesta secção irão realizar-se os testes de Bartlett e de Levene e pretendemos verificar se todos os períodos têm variância idêntica. Em ambos os testes as hipóteses em teste são as seguintes:

$$H_0 : \sigma^2_{nxst} = \sigma^2_{obin} = \sigma^2_{fpar} = \sigma^2_{obam} = \sigma^2_{fpln} = \sigma^2_{covd}$$
$$H_1 : \exists i, j : \sigma^2_i \neq \sigma^2_j, \text{ com } i \neq j;\ i, j = nxst, obin, fran, obam, fpln, covd$$

Através das Figuras 3.5 e 3.6 observamos que os valores dos $p-value$ são aproximadamente zero, pelo que se deve rejeitar H_0, concluindo-se que pelo menos duas das variâncias são

```
        Bartlett test of homogeneity of variances

   data:  list(nxst, obin, fpar, obam, fpln, covd)
   Bartlett's K-squared = 70445, df = 5, p-value < 2.2e-16
```

Figura 3.5: Teste de Bartlett

```
Levene's Test for Homogeneity of Variance (center = mean)
          Df F value    Pr(>F)
group      5  5495.1 < 2.2e-16 ***
      184827
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 3.6: Teste de Levene

diferentes.

Apesar dos pressupostos não serem verificados apresenta-se o teste de ANOVA e de seguida o teste alternativo à ANOVA quando os pressupostos não são verificados - o teste de Kruskal-Wallis.

3.2.3 Teste de ANOVA

As hipóteses em teste são as seguintes:

$H_0 : \mu_{nxst} = \mu_{obin} = \mu_{fpar} = \mu_{obam} = \mu_{fpln} = \mu_{covd}$

$H_1 : \exists i, j : \mu_i \neq \mu_j$, com $i \neq j$; $i, j = nxst, obin, fran, obam, fpln, covd$

```
                Df Sum Sq Mean Sq F value Pr(>F)
periodos         5   3395   679.0    6999 <2e-16 ***
Residuals   184827  17929     0.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
193995 observations deleted due to missingness
```

Figura 3.7: Teste de ANOVA

Analisando a Figura 3.7 observamos que o valor do $p - value$ é aproximadamente zero, pelo que devemos rejeitar H_0, concluindo-se que a concentração média de CO é diferente em pelo menos dois dos períodos.

Para analisar quais os períodos que diferem significativamente entre si foi aplicado um teste de comparação múltipla. Optou-se pelo teste de Scheffé pois é o mais robusto à violação dos pressupostos da ANOVA e porque as amostras têm dimensões distintas.

3.2.4 Teste de comparação múltipla de Scheffé

As hipóteses em teste são as seguintes:

$H_0 : \mu_i = \mu_j$

$H_1 : \mu_i \neq \mu_j$,

para $i \neq j$, com $i, j = nxst, obin, fran, obam, fpln, covd$.

```
  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$periodos
                 diff      lwr.ci      upr.ci   pval
fpar-covd  0.15271944  0.14393029  0.16150859 <2e-16 ***
fpln-covd  0.02446084  0.01734435  0.03157734 <2e-16 ***
nxst-covd  0.38889906  0.37954491  0.39825320 <2e-16 ***
obam-covd  0.08917885  0.07925314  0.09910455 <2e-16 ***
obin-covd  0.29510137  0.28666870  0.30353403 <2e-16 ***
fpln-fpar -0.12825860 -0.13604461 -0.12047258 <2e-16 ***
nxst-fpar  0.23617961  0.22630654  0.24605269 <2e-16 ***
obam-fpar -0.06354059 -0.07395679 -0.05312440 <2e-16 ***
obin-fpar  0.14238193  0.13337708  0.15138677 <2e-16 ***
nxst-fpln  0.36443821  0.35601961  0.37285681 <2e-16 ***
obam-fpln  0.06471800  0.05566856  0.07376744 <2e-16 ***
obin-fpln  0.27064052  0.26325929  0.27802175 <2e-16 ***
obam-nxst -0.29972021 -0.31061736 -0.28882305 <2e-16 ***
obin-nxst -0.09379769 -0.10335479 -0.08424059 <2e-16 ***
obin-obam  0.20592252  0.19580532  0.21603972 <2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 3.8: Teste de comparação múltipla de Scheffé

Da observação da Figura 3.8 temos que o valor do $p - value$ é aproximadamente zero em todas as comparações de cada conjunto de dois períodos. Assim, deve-se rejeitar H_0, em todos os testes de hipóteses da igualdade dos valores médios, concluindo-se que as concentrações médias de CO são significativamente diferentes em todos os períodos.

De seguida apresentamos o teste que é uma alternativa não paramétrica à ANOVA - o teste de Kruskal-Wallis.

3.2.5 Teste de Kruskal-Wallis

As hipóteses em teste são as seguintes:

H_0 : A distribuição dos valores da concentração de CO é idêntica nos seis períodos

H_1 : Existe pelo menos um período onde a distribuição dos valores da concentração de CO é diferente

ou de outra forma,

$H_0 : F(X_{nxst}) = F(X_{obin}) = F(X_{fpar}) = F(X_{obam}) = F(X_{fpln}) = F(X_{covd})$

$H_1 : \exists i, j : F(X_i) \neq F(X_j)$, com $i \neq j$ e $i, j = nxst, obin, fran, obam, fpln, covd$

Na Figura 3.9 observamos que o valor do $p - value$ é aproximadamente zero, pelo que devemos rejeitar H_0, concluindo-se que a distribuição das concentrações de CO é diferente em pelo menos dois períodos, o que pode ser visualizado na Figura 3.10.

```
    Kruskal-Wallis rank sum test

data:  concentracao by periodos
Kruskal-Wallis chi-squared = 33188, df = 5, p-value < 2.2e-16
```

Figura 3.9: Teste de Kruskal-Wallis

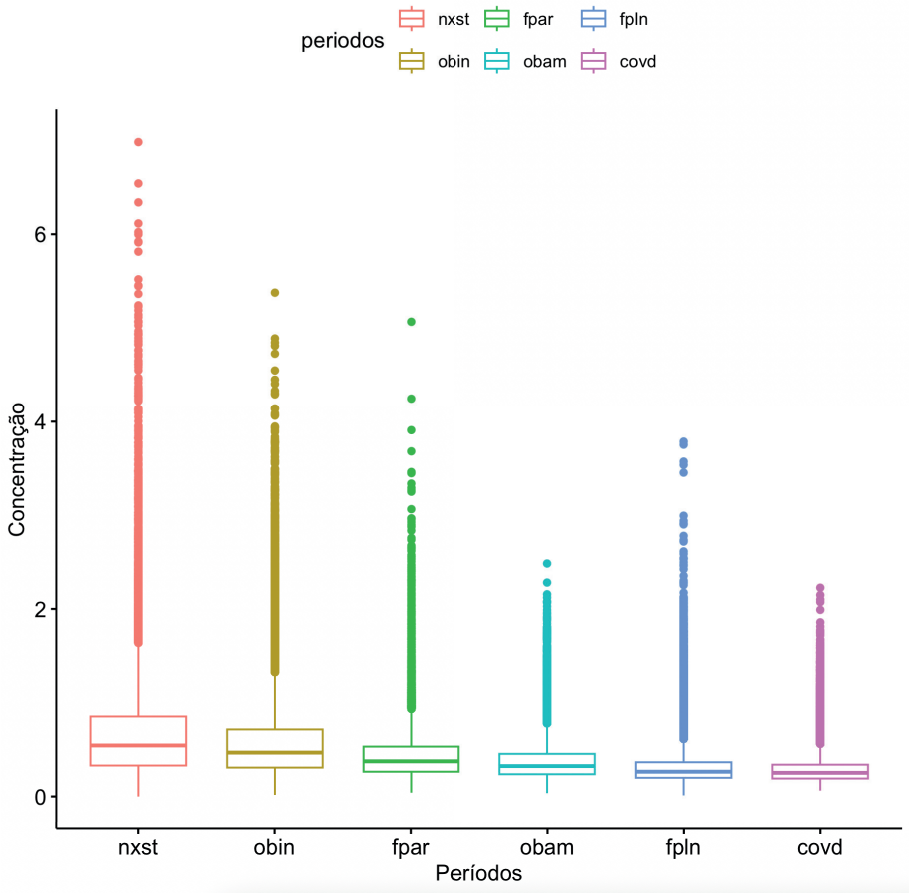

Figura 3.10: Gráficos de extremos e quartis para o teste de Kruskal-Wallis

Para analisar quais os pares que diferem significativamente entre si realizamos o teste de comparação múltipla de Nemenyi.

3.2.6 Teste de comparação múltipla de Nemenyi

As hipóteses em teste são:

$$H_0 : F(X_i) = F(X_j)$$
$$H_1 : F(X_i) \neq F(X_j),$$

com $i \neq j$ e $i, j = nxst, obin, fran, obam, fpln, covd$.

```
Nemenyi's test of multiple comparisons for independent samples (tukey)

          mean.rank.diff   pval
fpar-covd      37490.673 <2e-16 ***
fpln-covd       6083.137 <2e-16 ***
nxst-covd      60290.608 <2e-16 ***
obam-covd      24928.521 <2e-16 ***
obin-covd      53580.117 <2e-16 ***
fpln-fpar     -31407.536 <2e-16 ***
nxst-fpar      22799.935 <2e-16 ***
obam-fpar     -12562.153 <2e-16 ***
obin-fpar      16089.444 <2e-16 ***
nxst-fpln      54207.471 <2e-16 ***
obam-fpln      18845.384 <2e-16 ***
obin-fpln      47496.980 <2e-16 ***
obam-nxst     -35362.088 <2e-16 ***
obin-nxst      -6710.491 <2e-16 ***
obin-obam      28651.596 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 3.11: Teste de Nemenyi

Na Figura 3.11 observa-se que o valor do $p - value$ é aproximadamente zero em todas as comparações de cada dois períodos. Assim deve-se rejeitar H_0, em todos os testes de hipóteses, concluindo-se que as distribuições das concentração de CO são significativamente diferentes em todos os períodos.

3.3 Estatística descritiva para NO_2

Na Figura 3.12 verificamos uma diminuição na concentração de NO_2, ao nível dos valores máximos observados, no 1^{o} e 3^{o} quartis, na média, na mediana, na variância, no desvio padrão e nos limites do intervalo de confiança para a média, apenas nos três últimos períodos. Todos os períodos têm uma assimetria positiva, ou seja, existe uma maior concentração de valores na faixa de valores mais baixos da amostra.

Relativamente à curtose ou achatamento, temos uma curva leptocúrtica, ou seja, menos achatada quando comparada com curva da distribuição normal. Existe uma maior concentração de valores em torno da média do que na distribuição normal e há uma maior concentração de valores extremos ou discrepantes nas caudas (possivelmente *outliers*)

Na Figura 3.13 observa-se que a média da concentração de NO_2 aumentou nos três primeiros períodos e depois diminuiu nos últimos três.

Na Figura 3.12 observa-se, ainda, uma diminuição dos valores do desvio padrão, o que aparenta indicar uma menor variabilidade dos dados. No entanto na Figura 3.14 observam-se vários outliers nos valores superiores, em todos os períodos.

Na Tabela 3.3 apresentam-se os valores dos coeficientes de variação e de variação resistente. Apesar do desvio padrão ter vindo a diminuir nos três últimos períodos, os valores dos dois coeficientes de variação são sempre muito elevados, e apresentam um aumento no seu valor. Os valores obtidos indicam uma grande variabilidade das concentrações de NO_2, justificada

	nxst	obin	fpar	obam	fpln	covd
nobs	63138.0000	63138.0000	63138.0000	63138.0000	63138.0000	63138.0000
NAs	43201.0000	34481.0000	38515.0000	46587.0000	0.0000	31211.0000
Minimum	2.4000	1.9000	2.9000	0.7000	1.7000	1.0000
Maximum	309.4000	472.9000	390.2000	366.0000	361.4000	243.2000
1. Quartile	35.8000	40.2000	38.7000	37.5000	30.5000	20.2000
3. Quartile	80.7000	87.0000	91.8000	85.1000	76.2000	58.1000
Mean	60.8841	66.9021	68.8862	64.4415	57.0457	42.0611
Median	54.0000	63.1000	64.1000	59.6000	52.1000	37.6000
Sum	1213846.5000	1917212.7000	1696184.9000	1066571.8000	3601752.3000	1342885.8000
SE Mean	0.2366	0.2163	0.2508	0.2800	0.1382	0.1531
LCL Mean	60.4205	66.4780	68.3946	63.8928	56.7749	41.7611
UCL Mean	61.3478	67.3261	69.3778	64.9903	57.3165	42.3611
Variance	1115.5939	1341.2168	1549.1096	1297.3187	1205.1171	748.0193
Stdev	33.4005	36.6226	39.3587	36.0183	34.7148	27.3499
Skewness	0.9510	1.0665	0.9804	1.0571	1.1521	0.9864
Kurtosis	1.0176	2.7590	1.7127	2.4801	2.4947	1.1382

Figura 3.12: Medidas de estatística descritiva das concentrações de NO_2

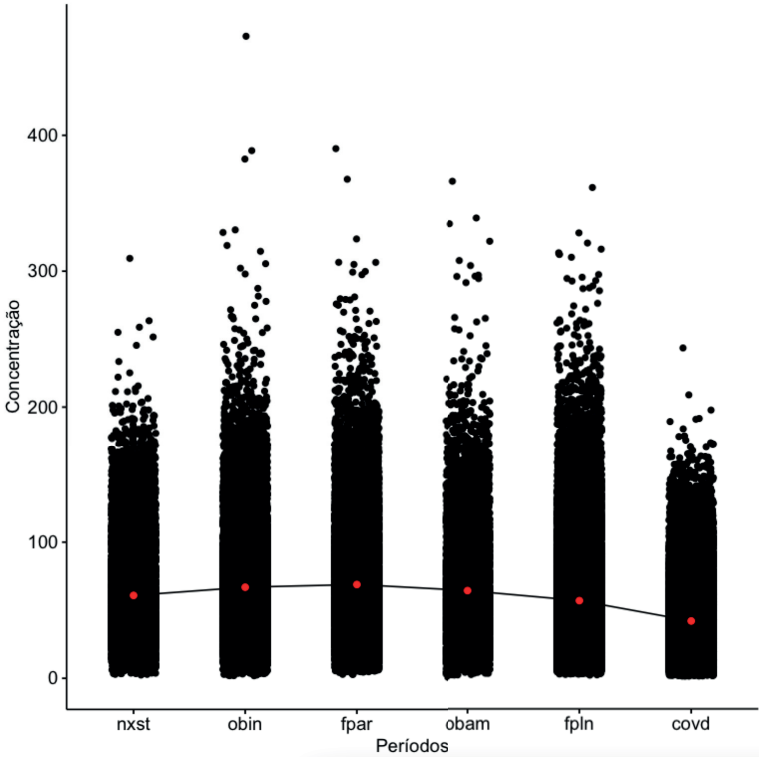

Figura 3.13: Distribuição gráfica dos valores por período, com representação da média

no caso do coeficiente de variação, em grande parte, pela existência de *outliers*.

3.4 ANOVA

Em primeiro lugar vão verificar-se os pressupostos da ANOVA. A independência dos grupos (períodos) está assegurada por se tratarem de períodos mutuamente exclusivos. Nas secções que se seguem irão analisar-se a normalidade e a homogeneidade dos grupos (períodos).

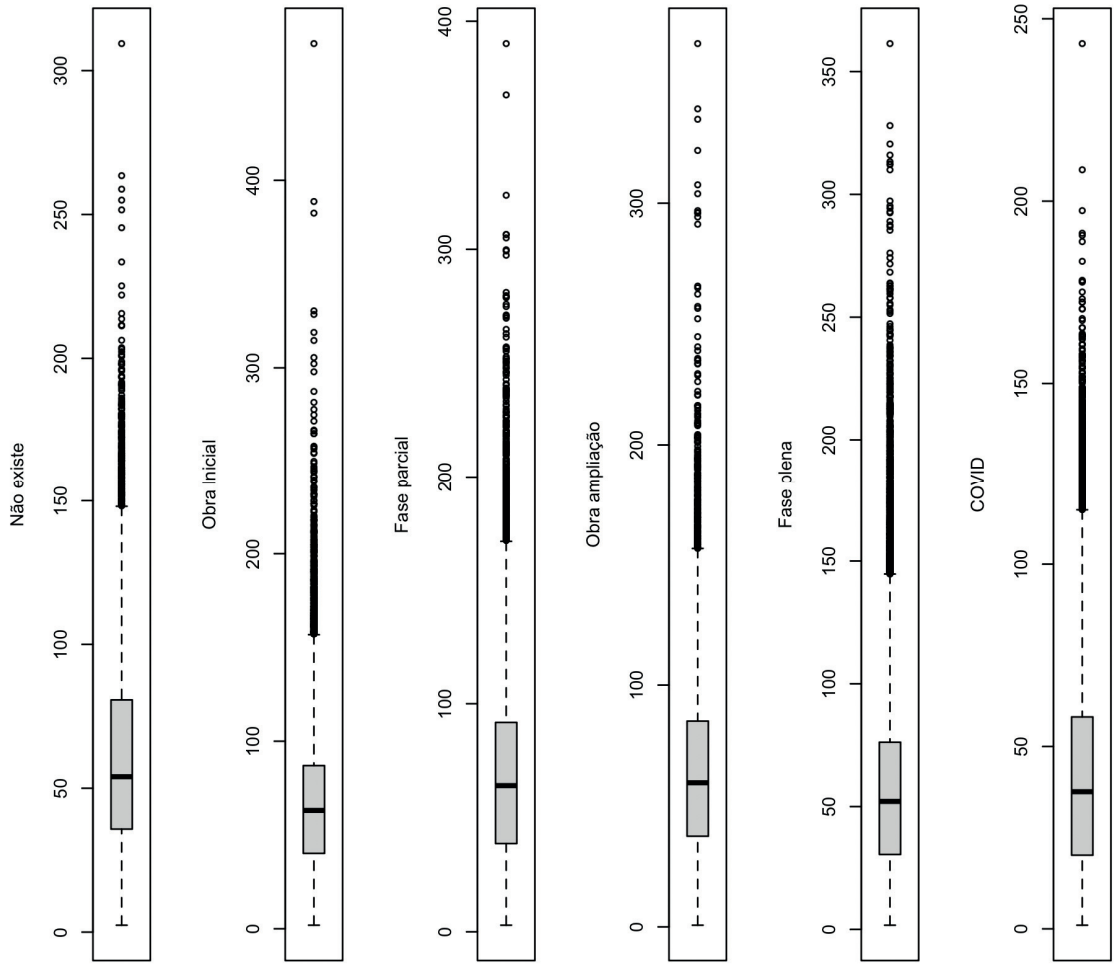

Figura 3.14: Diagrama de extremos e quartis NO2

Período	Coeficientes de variação	Coeficientes de variação resistente
nxst	54.8592	83.1481
obin	54.7406	74.1680
fpar	57.1359	82.8393
obam	55.8930	79.8658
fpln	60.8543	87.7159
covd	65.0243	100.7979

Tabela 3.3: Coeficientes de variação e de variação resistente

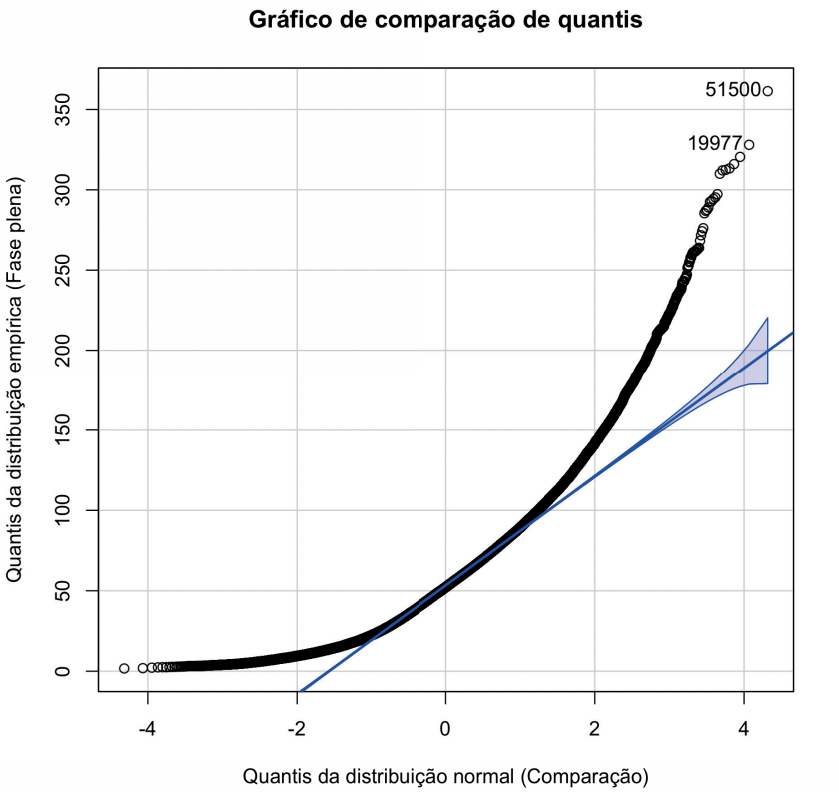

Figura 3.15: Gráfico de comparação de quantis da fase plena

3.4.1 Normalidade dos grupos (períodos)

No gráfico de comparação de quantis, obtido para os dados do período correspondente à fase plena, que se apresentam na Figura 3.15, representam-se os pares ordenados dos quantis de duas distribuições, a distribuição normal e a distribuição empírica. Observando o gráfico conclui-se que os dados da concentração de NO_2, neste período, não parecem ser provenientes de uma população com distribuição normal.

Para analisar a normalidade dos grupos vai usar-se o teste de Kolmogorov-Smirnov com correção de Lilliefors.

As hipóteses em teste são as seguintes:

H_0: A amostra do i-ésimo grupo é proveniente de uma população com distribuição normal

H_1: A amostra do i-ésimo grupo é proveniente de uma população com distribuição diferente da distribuição normal,

com $i = nxst, obin, fran, obam, fpln, covd$.

Período	**Valor da estatística de teste**	**Valor do** $p - value$
$nxst$	0.085383	$\simeq 0$
$obin$	0.053423	$\simeq 0$
$fpar$	0.05728	$\simeq 0$
$obam$	0.054555	$\simeq 0$
$fpln$	0.06538	$\simeq 0$
$covd$	0.074479	$\simeq 0$

Tabela 3.4: Valor da estatística de teste e do $p - value$

Como se pode verificar, na Tabela 3.4, o valor do $p - value$ é aproximadamente zero em todos os períodos, pelo que se deve rejeitar H_0 para todos os períodos. Assim, conclui-se que as observações de cada um dos grupos não são provenientes de populações com distribuição normal.

3.4.2 Homogeneidade das variâncias

Nesta secção vão realizar-se testes de Bartlett e de Levene e pretende-se verificar se todos os períodos têm variância idêntica. Em ambos os testes as hipóteses em teste são as seguintes:

$H_0 : \sigma^2_{nxst} = \sigma^2_{obin} = \sigma^2_{fpar} = \sigma^2_{obam} = \sigma^2_{fpln} = \sigma^2_{covd}$

$H_1 : \exists i, j : \sigma^2_i \neq \sigma^2_j$, com $i \neq j$; $i, j = nxst, obin, fran, obam, fpln, covd$

```
        Bartlett test of homogeneity of variances

data:  list(nxst, obin, fpar, obam, fpln, covd)
Bartlett's K-squared = 4224.1, df = 5, p-value < 2.2e-16
```

Figura 3.16: Teste de Bartlett

Através das Figuras 3.16 e 3.17 observa-se que os valores dos $p-value$ são aproximadamente zero, pelo que se deve rejeitar H_0 concluindo-se que pelo menos duas variâncias são diferentes. Apesar dos pressupostos não serem verificados apresenta-se o teste ANOVA e de seguida o teste alternativo à ANOVA, quando os pressupostos não são verificados - o teste de Kruskal-Wallis.

```
Levene's Test for Homogeneity of Variance (center = mean)
          Df F value    Pr(>F)
group      5  586.17 < 2.2e-16 ***
      184827
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 3.17: Teste de Levene

3.4.3 Teste de ANOVA

As hipóteses em teste são as seguintes:

$$H_0 : \mu_{nxst} = \mu_{obin} = \mu_{fpar} = \mu_{obam} = \mu_{fpln} = \mu_{covd}$$

$$H_1 : \exists i, j : \mu_i \neq \mu_j, \text{ com } i \neq j;\ i, j = nxst, obin, fran, obam, fpln, covd$$

```
                Df    Sum Sq Mean Sq F value Pr(>F)
periodos         5  14134587 2826917    2372 <2e-16 ***
Residuals   184827 220255937    1192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
193995 observations deleted due to missingness
```

Figura 3.18: Teste de ANOVA

Analisando a Figura 3.18 observa-se que o valor do $p - value$ é aproximadamente zero, pelo que devemos rejeitar H_0, concluindo-se que a concentração média de NO_2 é diferente em pelo menos dois dos períodos.

Para analisar quais as médias que diferem significativamente entre si foi aplicado um teste de comparação múltipla. Optou-se pelo teste de Scheffé pois é o mais robusto à violação dos pressupostos da ANOVA.

3.4.4 Teste de comparação múltipla de Scheffé

As hipóteses em teste são as seguintes:

$$H_0 : \mu_i = \mu_j$$

$$H_1 : \mu_i \neq \mu_j,$$

para $i \neq j$, com $i, j = nxst, obin, fran, obam, fpln, covd$.

Da observação da Figura 3.19 conclui-se que o valor do $p - value$ é aproximadamente zero em todas as comparações de cada conjunto de dois períodos. Assim deve-se rejeitar H_0, em todos os testes de hipóteses da igualdade dos valores médios, concluindo-se que as concentrações médias de NO_2 são significativamente diferentes em todos os períodos.

De seguida apresenta-se o teste que é uma alternativa não paramétrica à ANOVA - o teste de Kruskal-Wallis.

```
  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$periodos
               diff     lwr.ci      upr.ci    pval
fpar-covd  26.825067  25.850893  27.7992399 < 2e-16 ***
fpln-covd  14.984581  14.195802  15.7733598 < 2e-16 ***
nxst-covd  18.822977  17.786180  19.8597731 < 2e-16 ***
obam-covd  22.380399  21.280252  23.4805457 < 2e-16 ***
obin-covd  24.840940  23.906279  25.7756003 < 2e-16 ***
fpln-fpar -11.840486 -12.703473 -10.9774986 < 2e-16 ***
nxst-fpar  -8.002090  -9.096403  -6.9077768 < 2e-16 ***
obam-fpar  -4.444668  -5.599179  -3.2901559 < 2e-16 ***
obin-fpar  -1.984127  -2.982208  -0.9860463 2.6e-08 ***
nxst-fpln   3.838396   2.905293   4.7714981 < 2e-16 ***
obam-fpln   7.395818   6.392795   8.3988415 < 2e-16 ***
obin-fpln   9.856359   9.038236  10.6744808 < 2e-16 ***
obam-nxst   3.557422   2.349602   4.7652429 < 2e-16 ***
obin-nxst   6.017963   4.958671   7.0772543 < 2e-16 ***
obin-obam   2.460541   1.339169   3.5819122 2.9e-10 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 3.19: Teste de Scheffé

3.4.5 Teste de Kruskal-Wallis

As hipóteses em teste são as seguintes:

H_0 : A distribuição dos valores da concentração de NO_2 é idêntica nos seis períodos

H_1 : Existe pelo menos um período onde a distribuição dos valores da concentração de NO_2 é diferente,

ou de outra forma,

$$H_0 : F(X_{nxst}) = F(X_{obin}) = F(X_{fpar}) = F(X_{obam}) = F(X_{fpln}) = F(X_{covd})$$
$$H_1 : \exists i, j : F(X_i) \neq F(X_j), \text{ com } i \neq j \text{ e } i, j = nxst, obin, fran, obam, fpln, covd$$

```
    Kruskal-Wallis rank sum test

data:  concentracao by periodos
Kruskal-Wallis chi-squared = 12065, df = 5, p-value < 2.2e-16
```

Figura 3.20: Teste de Kruskal-Wallis

Na Figura 3.20, observa-se que o valor do $p - value$ é aproximadamente zero, pelo que devemos rejeitar H_0, concluindo-se que a distribuição das concentrações de NO_2 é diferente em pelo menos dois períodos, o que pode ser visualizado na Figura 3.21.

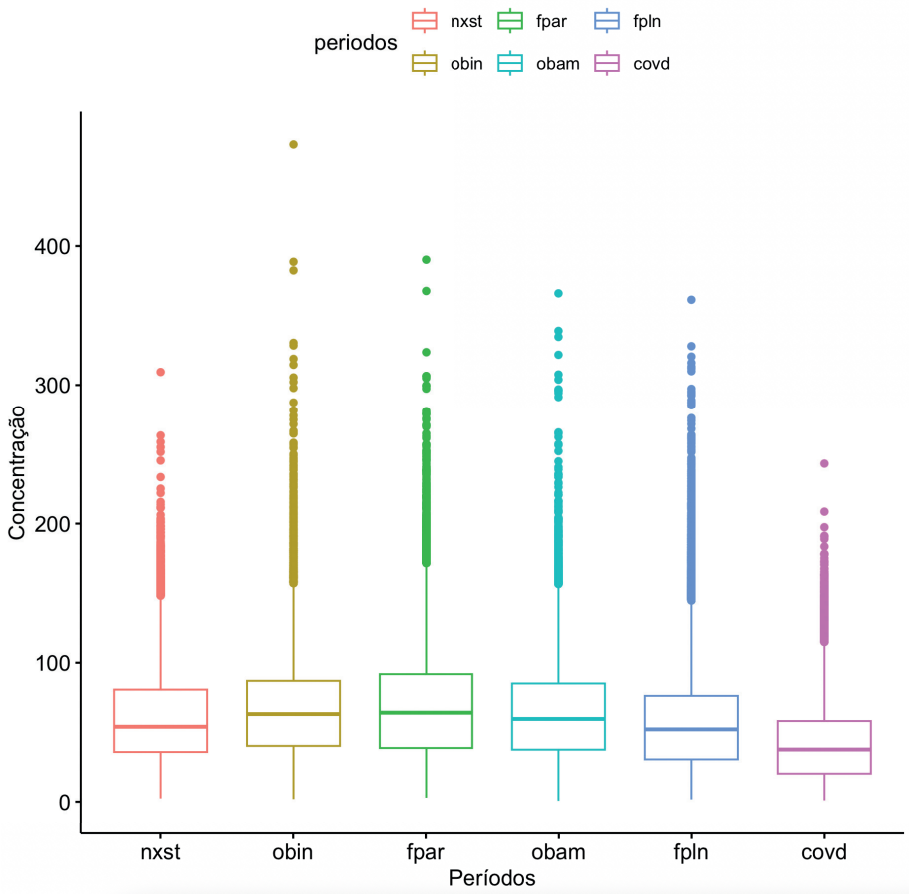

Figura 3.21: Gráficos de extremos e quartis para o teste de Kruskal-Wallis

Para analisar quais os pares que diferem significativamente entre si realizamos o teste de comparação múltipla de Nemenyi.

3.4.6 Teste de comparação múltipla de Nemenyi

As hipóteses em teste são as seguintes:

$$H_0 : F(X_i) = F(X_j)$$

$$H_1 : F(X_i) \neq F(X_j),$$

com $i \neq j$ e $i, j = nxst, obin, fran, obam, fpln, covd$.

Observa-se na Figura 3.22 que o valor do $p - value$ é aproximadamente zero em todas as comparações de cada dois períodos, com exceção dos períodos $obin - fpar$. Assim deve-se rejeitar H_0, em todos os testes de hipóteses, com exceção do teste de comparação dos períodos $obin - fpar$ porque o valor de $p - value$ é 0.1600. Conclui-se que as distribuições das concentração de NO_2 são significativamente diferentes em todos os períodos, com exceção dos períodos $obin - fpar$.

```
 Nemenyi's test of multiple comparisons for independent samples (tukey)

          mean.rank.diff    pval
fpar-covd      40635.742 < 2e-16 ***
fpln-covd      24043.381 < 2e-16 ***
nxst-covd      30739.196 < 2e-16 ***
obam-covd      35744.870 < 2e-16 ***
obin-covd      39528.087 < 2e-16 ***
fpln-fpar     -16592.361 < 2e-16 ***
nxst-fpar      -9896.546 < 2e-16 ***
obam-fpar      -4890.872 7.5e-14 ***
obin-fpar      -1107.654  0.1600
nxst-fpln       6695.815 < 2e-16 ***
obam-fpln      11701.489 < 2e-16 ***
obin-fpln      15484.707 < 2e-16 ***
obam-nxst       5005.674 4.2e-14 ***
obin-nxst       8788.892 < 2e-16 ***
obin-obam       3783.217 5.7e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 3.22: lteste de Nemenyi

3.5 Estatística descritiva para PM_{10}

Na Figura 3.23 verificam-se algumas oscilações nos valores máximos ao longo do tempo, começando por aumentar, depois diminuindo, voltando novamente a aumentar. No 1º e 3º quartis, na média, na mediana, na variância, no desvio padrão e nos limites do intervalo de confiança para a média verifica-se uma diminuição ao longo do tempo, até ao período $fpar$, depois os valores aumentam no período *obam* e voltam a diminuir nos dois últimos períodos.

	nxst	obin	fpar	obam	fpln	covd
nobs	63138.0000	63138.0000	63138.0000	63138.0000	63138.0000	63138.0000
NAs	43201.0000	34481.0000	38515.0000	46587.0000	0.0000	31211.0000
Minimum	0.0000	0.0000	2.4000	1.8000	1.1000	0.8000
Maximum	447.2000	539.3000	289.1000	180.5000	219.5000	477.4000
1. Quartile	31.7000	29.2000	26.0000	28.3000	19.6000	14.0000
3. Quartile	78.3000	68.4000	53.1000	55.5000	38.9000	26.3000
Mean	58.8224	52.7607	41.6166	44.0682	31.2077	21.5671
Median	52.0000	46.3000	38.7000	41.3000	28.0000	19.4000
Sum	1172741.9000	1511962.0000	1024726.1000	729373.6000	1970392.8000	688572.8000
SE Mean	0.2712	0.1974	0.1366	0.1682	0.0683	0.0781
LCL Mean	58.2908	52.3737	41.3488	43.7386	31.0738	21.4140
UCL Mean	59.3540	53.1476	41.8844	44.3979	31.3417	21.7202
Variance	1466.3359	1116.9087	459.6990	468.0429	294.8950	194.8253
Stdev	38.2928	33.4202	21.4406	21.6343	17.1725	13.9580
Skewness	1.5010	1.9666	1.3441	1.0196	1.6669	9.2830
Kurtosis	4.9231	10.9391	4.2009	1.8385	5.5330	208.1654

Figura 3.23: Medidas de estatística descritiva das concentrações de PM_{10}

Todos os períodos têm uma assimetria positiva, ou seja, existe uma maior concentração de valores na faixa dos valores mais baixos da amostra.

Relativamente à curtose ou achatamento, temos uma curva leptocúrtica, ou seja, menos achatada quando comparada com curva da distribuição normal. Existe uma maior concentração de valores em torno da média do que na distribuição normal e há uma maior concentração de

valores extremos ou discrepantes nas caudas (possivelmente *outliers*).

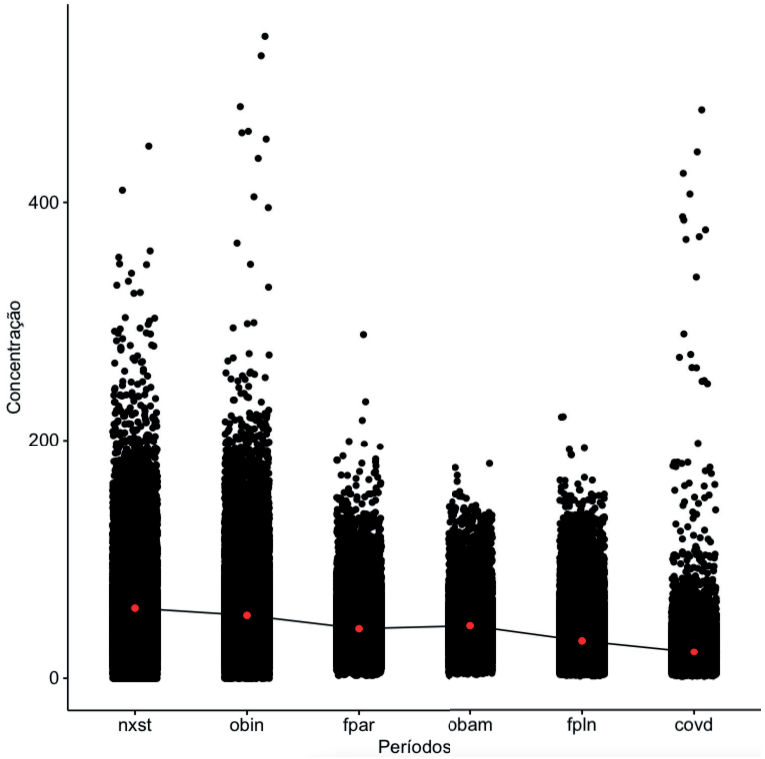

Figura 3.24: Distribuição gráfica dos valores por período, com representação da média

Na Figura 3.24 observa-se que a média da concentração de PM_{10} diminuiu até ao período $fpar$, aumentando depois no período *obam* e diminuindo nos dois últimos períodos.

Período	**Coeficientes de variação**	**Coeficientes de variação resistente**
nxst	65.0990	89.6154
obin	63.3430	84.6652
fpar	51.5193	70.0258
obam	49.0928	65.8596
fpln	55.0265	68.9286
covd	64.7189	63.4021

Tabela 3.5: Coeficientes de variação e de variação resistente

Na Figura 3.25 observam-se vários *outliers* nos valores superiores, em todos os períodos. Na Tabela 3.5 apresentamos os valores dos coeficientes de variação e de variação resistente. Apesar do desvio padrão ter tendência a diminuir ao longo do tempo, os valores dos dois coeficientes de variação são sempre muito elevados, apresentando também ligeiras diminuições. Os valores obtidos indicam uma grande variabilidade das concentrações de PM_{10}, sendo justificada no caso do coeficiente de variação, em grande parte, pela existência de *outliers*.

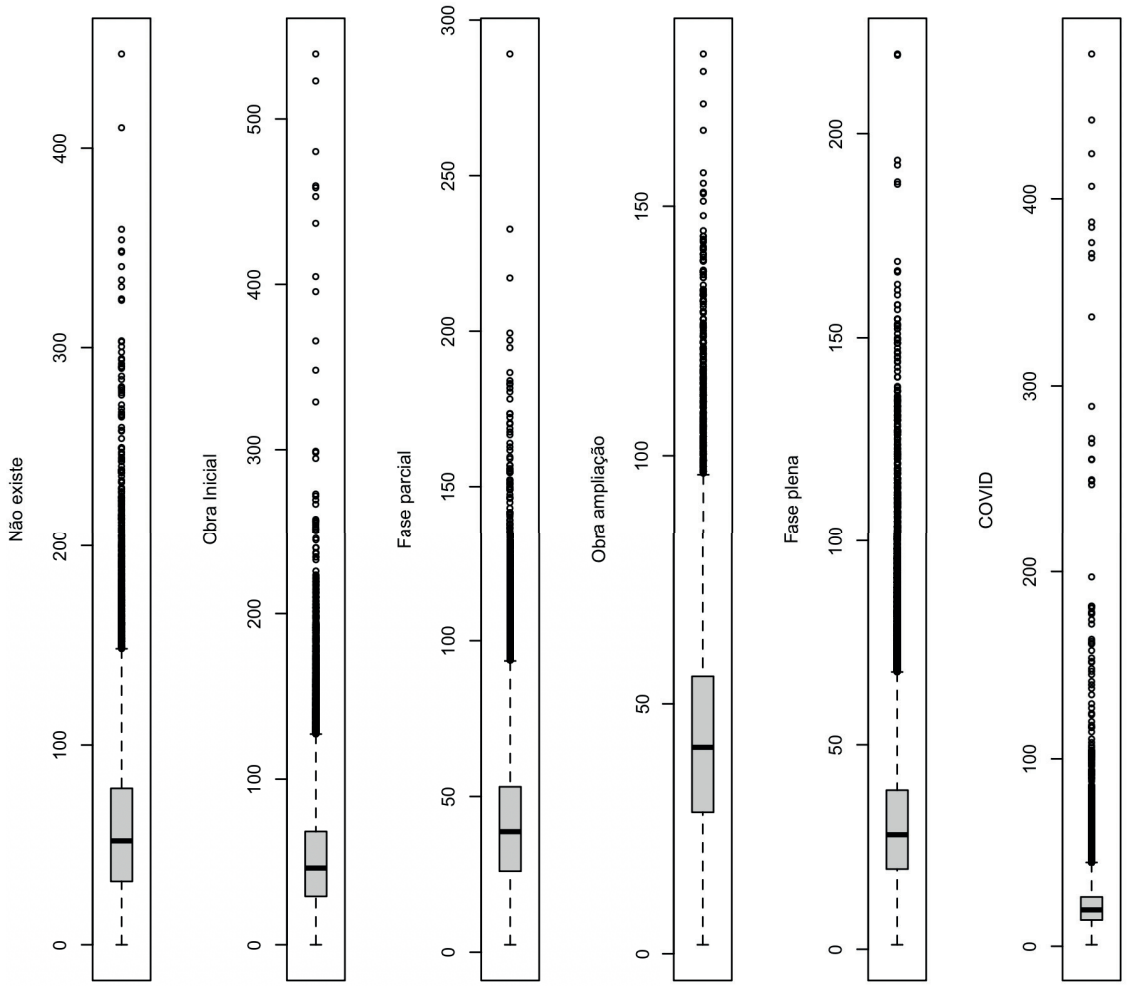

Figura 3.25: Diagrama de extremos e quartis

3.6 ANOVA

Vamos verificar se os pressupostos da ANOVA são verificados. A independência dos grupos (períodos) está assegurada por se tratarem de períodos mutuamente exclusivos. Nas secções que se seguem iremos analisar a normalidade e a homogeneidade dos grupos (períodos).

3.6.1 Normalidade dos grupos (períodos)

No gráfico de comparação de quantis, obtido para os dados do período correspondente à fase plena, apresentado na Figura 3.26, representamos os pares ordenados dos quantis da distribuição normal e da distribuição empírica. Observando o gráfico conclui-se que os dados de concentração de PM_{10}, neste período, não parecem ser provenientes de uma população com distribuição normal.

Para analisar a normalidade dos grupos vamos usar o teste de Kolmogorov-Smirnov com correção de Lilliefors. As hipóteses em teste são as seguintes:

H_0: A amostra do i-ésimo grupo é proveniente de uma população com distribuição

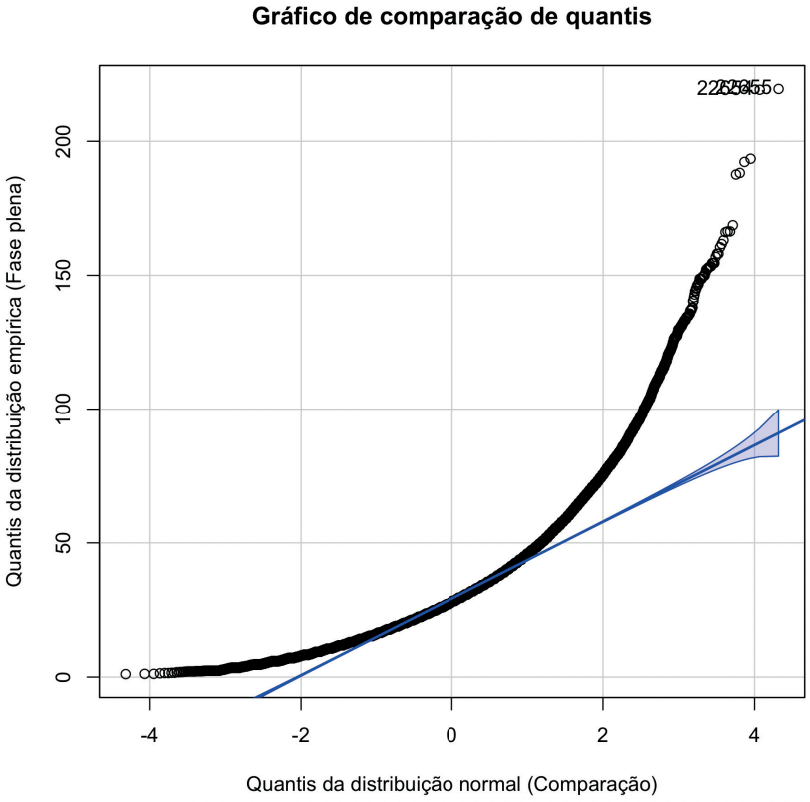

Figura 3.26: Gráfico de comparação de quantis da fase plena

normal

H_1: A amostra do i-ésimo grupo é proveniente de uma população com distribuição diferente da distribuição normal,

com $i = nxst, obin, fran, obam, fpln, covd$.

Período	**Valor da estatística de teste**	**Valor do** $p - value$
$nxst$	0.074786	$\simeq 0$
$obin$	0.086925	$\simeq 0$
$fpar$	0.07336	$\simeq 0$
$obam$	0.06309	$\simeq 0$
$fpln$	0.09696	$\simeq 0$
$covd$	0.12035	$\simeq 0$

Tabela 3.6: Valor da estatística de teste e do $p - value$

Como se pode verificar, na Tabela 3.6 o valor do $p - value$ é aproximadamente zero em todos os períodos, pelo que devemos rejeitar H_0 para todos os períodos. Assim, concluímos que os dados de cada um dos grupos são provenientes de uma população com distribuição diferente da normal.

3.6.2 Homogeneidade das variâncias

Nesta secção iremos realizar testes de Bartlett e de Levene e pretendemos verificar se todos os períodos têm variância idêntica.

Em ambos os testes as hipóteses em teste são as seguintes:

$H_0 : \sigma^2_{nxst} = \sigma^2_{obin} = \sigma^2_{fpar} = \sigma^2_{obam} = \sigma^2_{fpln} = \sigma^2_{covd}$

$H_1 : \exists i, j : \sigma^2_i \neq \sigma^2_j$, com $i \neq j$; $i, j = nxst, obin, fran, obam, fpln, covd$

```
        Bartlett test of homogeneity of variances

data:  list(nxst, obin, fpar, obam, fpln, covd)
Bartlett's K-squared = 45958, df = 5, p-value < 2.2e-16
```

Figura 3.27: Teste de Bartlett

```
Levene's Test for Homogeneity of Variance (center = mean)
          Df F value    Pr(>F)
group      5  6337.9 < 2.2e-16 ***
      184827
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figura 3.28: Teste de Levene

Através das Figuras 3.27 e 3.28 observamos que o valor de $p - value$ é aproximadamente zero, pelo que se deve rejeitar H_0, concluindo-se que pelo menos duas variâncias são diferentes. Apesar dos pressupostos não serem verificados apresenta-se o teste ANOVA e de seguida o teste alternativo à ANOVA, quando os pressupostos não são verificados - o teste de Kruskal-Wallis.

3.6.3 Teste de ANOVA

As hipóteses em teste são as seguintes:

$H_0 : \mu_{nxst} = \mu_{obin} = \mu_{fpar} = \mu_{obam} = \mu_{fpln} = \mu_{covd}$

$H_1 : \exists i, j : \mu_i \neq \mu_j$, com $i \neq j$; $i, j = nxst, obin, fran, obam, fpln, covd$

```
                Df    Sum Sq Mean Sq F value Pr(>F)
periodos         5  27324086 5464817    9606 <2e-16 ***
Residuals   184827 105142609     569
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
193995 observations deleted due to missingness
```

Figura 3.29: Teste de ANOVA

Analisando a Figura 3.29 temos que o valor do $p - value$ é aproximadamente zero, pelo que devemos rejeitar H_0, concluindo-se que a concentração média de PM_{10} é diferente em pelo menos dois períodos.

Para analisar quais as médias que diferem significativamente entre si foi aplicado um teste de comparação múltipla. Optou-se pelo teste de Scheffé pois é o mais robusto à violação dos pressupostos da ANOVA.

3.6.4 Teste de comparação múltipla de Scheffé

As hipóteses em teste são as seguintes:

$$H_0 : \mu_i = \mu_j$$
$$H_1 : \mu_i \neq \mu_j,$$

para $i \neq j$, com $i, j = nxst, obin, fran, obam, fpln, covd$.

```
  Posthoc multiple comparisons of means: Scheffe Test
    95% family-wise confidence level

$periodos
                diff     lwr.ci     upr.ci   pval
fpar-covd  20.049523  19.376450  20.722595 <2e-16 ***
fpln-covd   9.640616   9.095636  10.185597 <2e-16 ***
nxst-covd  37.255286  36.538946  37.971625 <2e-16 ***
obam-covd  22.501150  21.741040  23.261259 <2e-16 ***
obin-covd  31.193552  30.547779  31.839324 <2e-16 ***
fpln-fpar -10.408906 -11.005158  -9.812654 <2e-16 ***
nxst-fpar  17.205763  16.449684  17.961842 <2e-16 ***
obam-fpar   2.451627   1.653956   3.249298 <2e-16 ***
obin-fpar  11.144029  10.454439  11.833620 <2e-16 ***
nxst-fpln  27.614669  26.969973  28.259365 <2e-16 ***
obam-fpln  12.860533  12.167528  13.553539 <2e-16 ***
obin-fpln  21.552935  20.987681  22.118190 <2e-16 ***
obam-nxst -14.754136 -15.588639 -13.919633 <2e-16 ***
obin-nxst  -6.061734  -6.793616  -5.329852 <2e-16 ***
obin-obam   8.692402   7.917628   9.467176 <2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 3.30: Teste de Scheffé

Analisando a Figura 3.30 observamos que o valor do $p - value$ é aproximadamente zero em todas as comparações de cada conjunto de dois períodos. Assim deve-se rejeitar H_0 em todos os testes de hipóteses da igualdade das médias, concluindo-se que as concentrações médias de PM_{10} são significativamente diferentes em todos os períodos.

De seguida apresentamos o teste que é uma alternativa não paramétrica à ANOVA - o teste de Kruskal-Wallis.

3.6.5 Teste de Kruskal-Wallis

As hipóteses em teste são as seguintes:

H_0 : A distribuição dos valores da concentração de PM_{10} é idêntica nos seis períodos

H_1 : Existe pelo menos um período onde a distribuição dos valores da concentração de PM_{10} é diferente,

ou de outra forma,

$$H_0 : F(X_{nxst}) = F(X_{obin}) = F(X_{fpar}) = F(X_{obam}) = F(X_{fpln}) = F(X_{covd})$$
$$H_1 : \exists i, j : F(X_i) \neq F(X_j), \text{ com } i \neq j \text{ e } i, j = nxst, obin, fran, obam, fpln, covd$$

```
        Kruskal-Wallis rank sum test

data:  concentracao by periodos
Kruskal-Wallis chi-squared = 43596, df = 5, p-value < 2.2e-16
```

Figura 3.31: Teste de Kruskal-Wallis

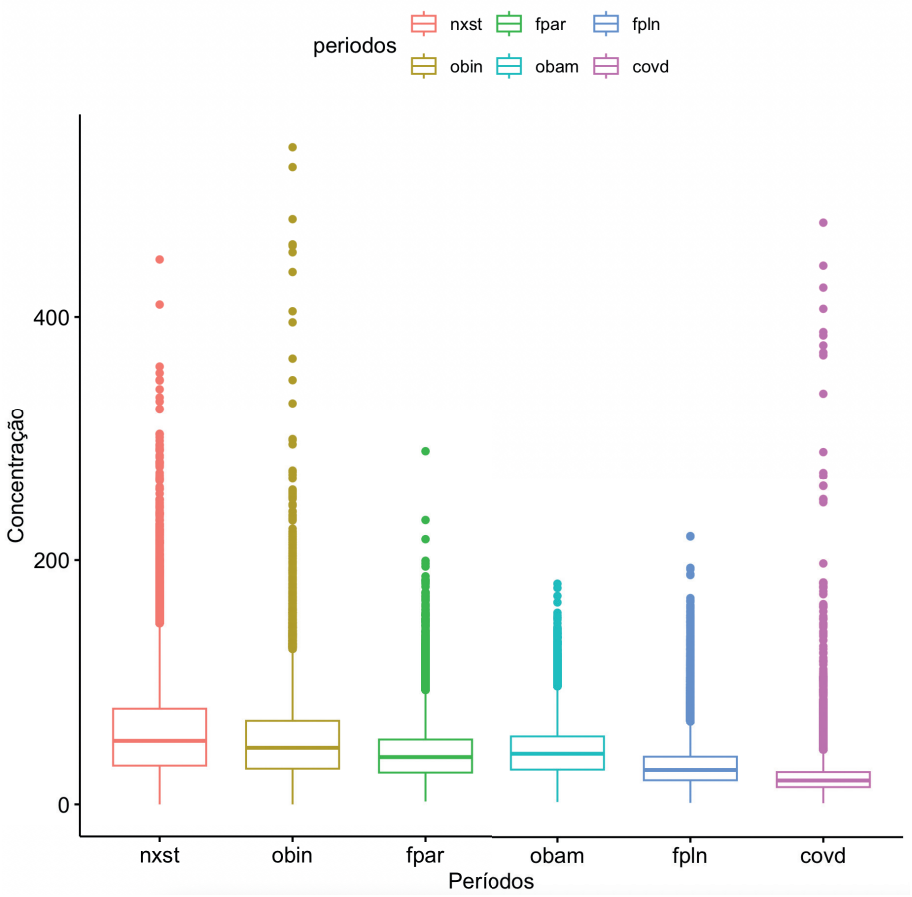

Figura 3.32: Gráficos de extremos e quartis para o teste de Kruskal-Wallis

Na Figura 3.31 observamos que o valor do$p - value$ é aproximadamente zero, pelo que devemos rejeitar H_0, concluindo-se que a distribuição das concentrações de PM_{10} é diferente em pelo menos dois períodos, o que pode ser visualizado na Figura 3.32.

Para analisar quais os pares que diferem significativamente entre si realizamos o teste de comparação múltipla de Nemenyi.

3.6.6 Teste de comparação múltipla de Nemenyi

As hipóteses em teste são:

$H_0 : F(X_i) = F(X_j)$

$H_1 : F(X_i) \neq F(X_j)$,

com $i \neq j$ e $i, j = nxst, obin, fran, obam, fpln, covd$.

```
Nemenyi's test of multiple comparisons for independent samples (tukey)

          mean.rank.diff    pval
fpar-covd      57142.185 < 2e-16 ***
fpln-covd      30691.753 < 2e-16 ***
nxst-covd      76448.440 < 2e-16 ***
obam-covd      63052.847 < 2e-16 ***
obin-covd      70473.598 < 2e-16 ***
fpln-fpar     -26450.432 < 2e-16 ***
nxst-fpar      19306.256 < 2e-16 ***
obam-fpar       5910.662 3.9e-14 ***
obin-fpar      13331.413 < 2e-16 ***
nxst-fpln      45756.688 < 2e-16 ***
obam-fpln      32361.094 < 2e-16 ***
obin-fpln      39781.845 < 2e-16 ***
obam-nxst     -13395.594 < 2e-16 ***
obin-nxst      -5974.843 < 2e-16 ***
obin-obam       7420.751 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 3.33: Teste de Nemenyi

Observa-se na Figura 3.33 que o valor do $p - value$ é aproximadamente zero em todas as comparações de cada dois períodos. Assim deve-se rejeitar H_0 em todos os testes de hipóteses, concluindo-se que também as distribuições das concentração de PM_{10} são significativamente diferentes em todos os períodos.

3.7 Análise discriminante linear

A aplicação de uma análise discriminante linear aos dados tem como objetivo verificar se os poluentes usados (variáveis) permitem distinguir os grupos usados (períodos temporais). Após a aplicação desta análise obtêm-se $m = \min\{k - 1; p\}$ funções discriminantes, sendo k o número de grupos e p o número de variáveis discriminantes. A análise tem dois pressupostos metodológicos:

- Cada grupo é uma amostra de uma população normal multivariada. Devido à quantidade de observações não foi possível realizar o teste de Shapiro-Wilk para a normal multivariada. No entanto como cada grupo tem mais de 20 observações, a técnica de análise discriminante linear é relativamente robusta à sua violação.
- As matrizes de variâncias-covariâncias são iguais para todos os grupos. Para verificar este pressuposto é usado o teste M de Box.

3.7.1 Teste M de Box

As hipóteses em teste são as seguintes:

- H_0 : As matrizes de variâncias-covariâncias são iguais para todos os grupos.
- H_1 : As matrizes de variâncias-covariâncias não são iguais para todos os grupos.

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  dados
Chi-Sq (approx.) = 158984, df = 30, p-value < 2.2e-16
```

Figura 3.34: Teste M de Box

Observando a Figura 3.34 concluí-se que o $p - value$ é aproximadamente zero, logo deve-se rejeitar H_0, concluindo-se que pelo menos duas matrizes de variâncias-covariâncias não são idênticas.

Apesar da violação deste pressuposto, foi realizada a análise discriminante linear, pois a técnica é bastante robusta à violação deste pressuposto, desde que as dimensões das amostras de cada grupo sejam suficientemente grandes, o que realmente se verifica.

```
Call:
lda(est ~ ., data = dados, na.action = "na.omit")

Prior probabilities of groups:
    Covid  FParcial    FPlena   NExiste    ObrAmp    ObrIni
0.1727343 0.1332176 0.3415948 0.1078649 0.0895457 0.1550427

Group means:
             PM10      NO2        CO
Covid    21.56710 42.06113 0.2916100
FParcial 41.61662 68.88620 0.4443295
FPlena   31.20772 57.04571 0.3160709
NExiste  58.82239 60.88411 0.6805091
ObrAmp   44.06825 64.44153 0.3807889
ObrIni   52.76065 66.90207 0.5867114

Coefficients of linear discriminants:
             LD1         LD2         LD3
PM10  0.03068500  0.02467848 -0.03613439
NO2  -0.01460094  0.02987904  0.01842046
CO    2.07396472 -3.28997288  2.43106376

Proportion of trace:
   LD1    LD2    LD3
0.7760 0.2083 0.0157
```

Figura 3.35: Análise discriminante linear

Foi aplicada a técnica de análise discriminante linear aos dados e obteve-se o output da Figura 3.35. Como existem $k = 6$ grupos e $p = 3$ variáveis obtiveram-se $m = \min\{6 - 1; 3\} = 3$ funções discriminantes, sendo estas dadas por:

$$Y_1 = 0.030685X_1 - 0.01460094X_2 + 2.07396472X_3,$$

$$Y_2 = 0.02467848X_1 + 0.02987904X_2 - 3.28997288X_3$$

e

$$Y_3 = -0.03613439X_1 + 0.01842046X_2 + 2.43106376X_3,$$

sendo $X_1 = PM_{10}$, $X_2 = NO_2$ e $X_3 = CO$. A 1ª função explica 77.6% da variabilidade entre grupos, a 2ª função explica 20.83% e a 3ª função explica apenas 1.57%.

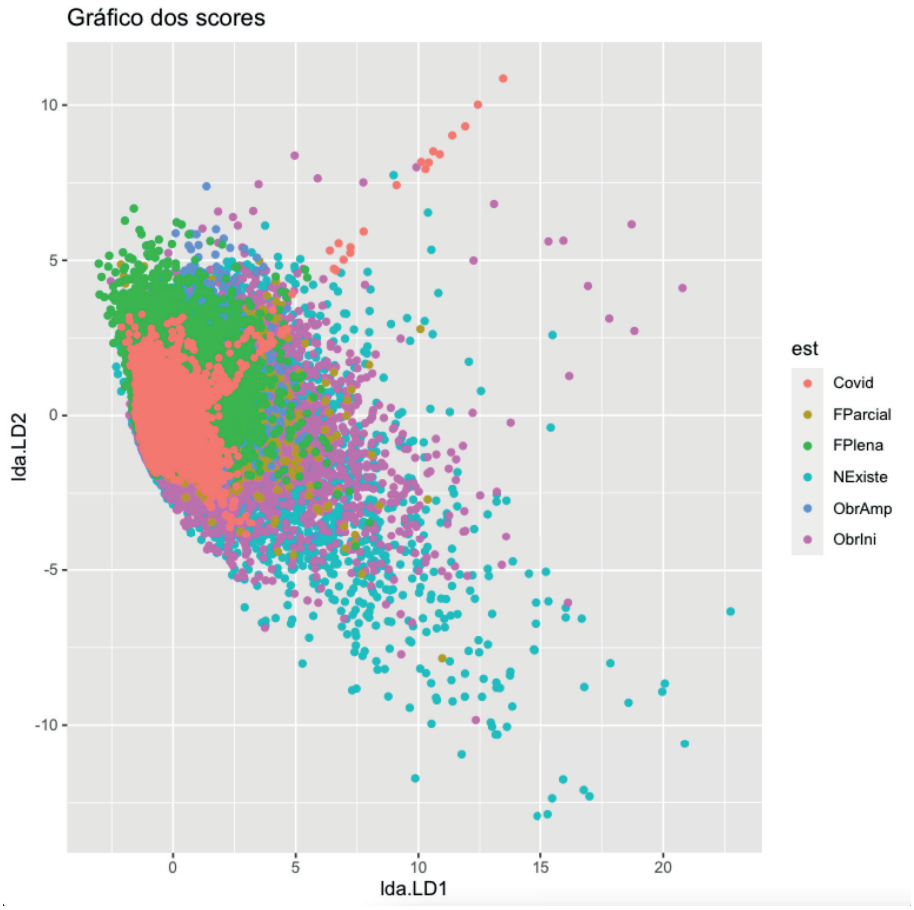

Figura 3.36: *Scores* das funções discriminantes 1 e 2

Em seguida apresentam-se os Gráficos 3.36, 3.37 e 3.38 dos *scores* da análise discriminante linear, onde a cor azul claro corresponde ao período *nxst*, a cor roxo corresponde ao período *obin*, a cor verde seco ao período $fpar$, a cor azul escuro ao período *obam*, a cor verde ao período $fpln$ e a cor vermelha ao período *covd*. Através destes gráfico observa-se uma grande dispersão dos *scores* relativos ao período *nxst*, relativamente às três funções. Os *scores* do período *obin* também apresentam alguma dispersão relativamente às três funções. Os *scores* dos períodos *covd*, $fpar$ e *obam* têm *scores* com valores mais semelhantes, aparecendo, nos gráficos, muito sobrepostos. No período $fpln$ os *scores* têm uma menor variabilidade quando comparados com os períodos *nxst* e *obin*, mas com valores relativamente diferentes, quando comparados com os períodos *covd*, $fpar$ e *obam*.

No Gráfico 3.39 podemos visualizar os *scores* relativos às três funções em simultâneo.

O teste lambda de Wilks cuja estatística de teste segue uma distribuição qui-quadrado permite testar se a solução discriminante global é estatisticamente significativa.

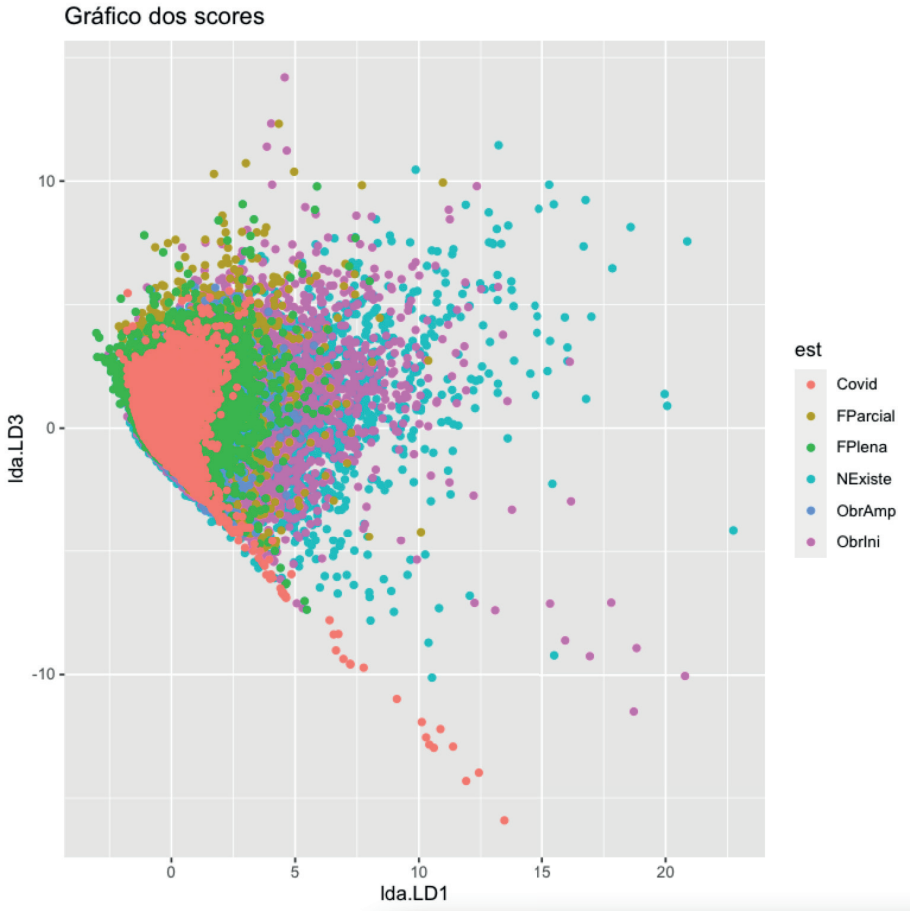

Figura 3.37: *Scores* das funções discriminantes 1 e 3

As hipóteses em teste são as seguintes:

$$H_0 : \lambda_1 = \lambda_2 = \lambda_3 = 0$$
$$H_1 : \exists \lambda_j \neq 0, \text{ com } j = 1, 2, 3,$$

sendo λ_i o valor próprio associado à função discriminante i.

Na Figura 3.40 apresenta-se o output do teste lambda de Wilks. Analisando o output verificamos que o valor do $p - value$ é aproximadamente zero, pelo que se deve rejeitar H_0, ao nível de significância de 5%, concluindo-se que a solução discriminante global é estatisticamente significativa. Assim, conclui-se que os três poluentes permitem fazer a diferenciação dos seis períodos temporais.

Para avaliar a eficácia classificativa da análise discriminante pode estimar-se as probabilidades de classificação correta em cada grupo. Para isso pode usar-se três métodos:

- Método de resubstituição.
- Método de Jacknife.
- Método de validação cruzada.

Ao aplicar cada um destes métodos é obtida uma matriz de classificações, que compara as classificações iniciais em cada grupo com as classificações resultantes da aplicação da análise

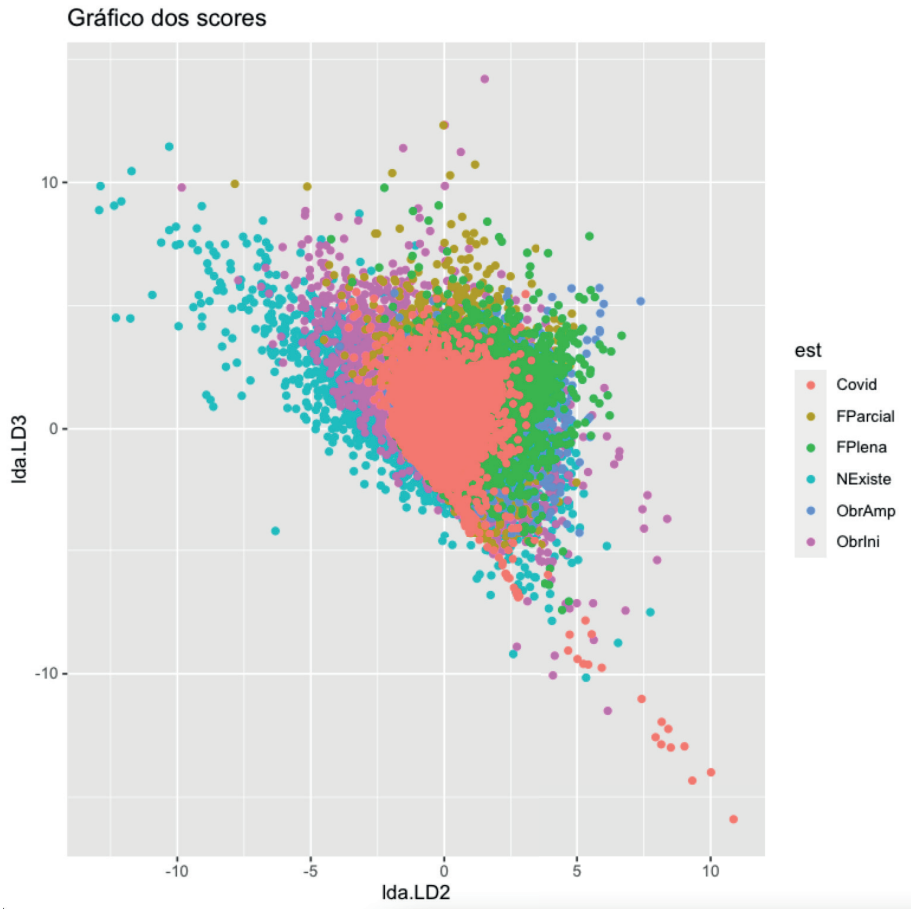

Figura 3.38: *Scores* das funções discriminantes 2 e 3

discriminante linear. O termo geral desta matriz n_{ij} representa o número de observações classificadas inicialmente no grupo i e cujo grupo previsto é j. Quando $i = j$, n_{ii}, concluí-se que a observação está bem classificada e corresponde aos elementos da diagonal da matriz.

3.7.2 Método de resubstituição

Na Figura 3.41 apresentamos a matriz de classificações do método de resubstituição. A estimativa da probabilidade de uma observação estar corretamente classificada é dada por:

$$p = \frac{2257 + 183 + 59523 + 5543 + 137 + 4967}{184833} = 0.3928,$$

ou seja, aproximadamente 40% das observações estão classificadas no grupo onde estavam originalmente. A estimativa da probabilidade de uma observação estar corretamente classificada em cada um dos grupos é:

- *covd*: $p = \frac{2257}{31927} = 0.0707$.
- *fpar*: $p = \frac{183}{24623} = 0.0074$.
- *fpln*: $p = \frac{59523}{63138} = 0.9427$.
- *nxst*: $p = \frac{5543}{19937} = 0.2780$.

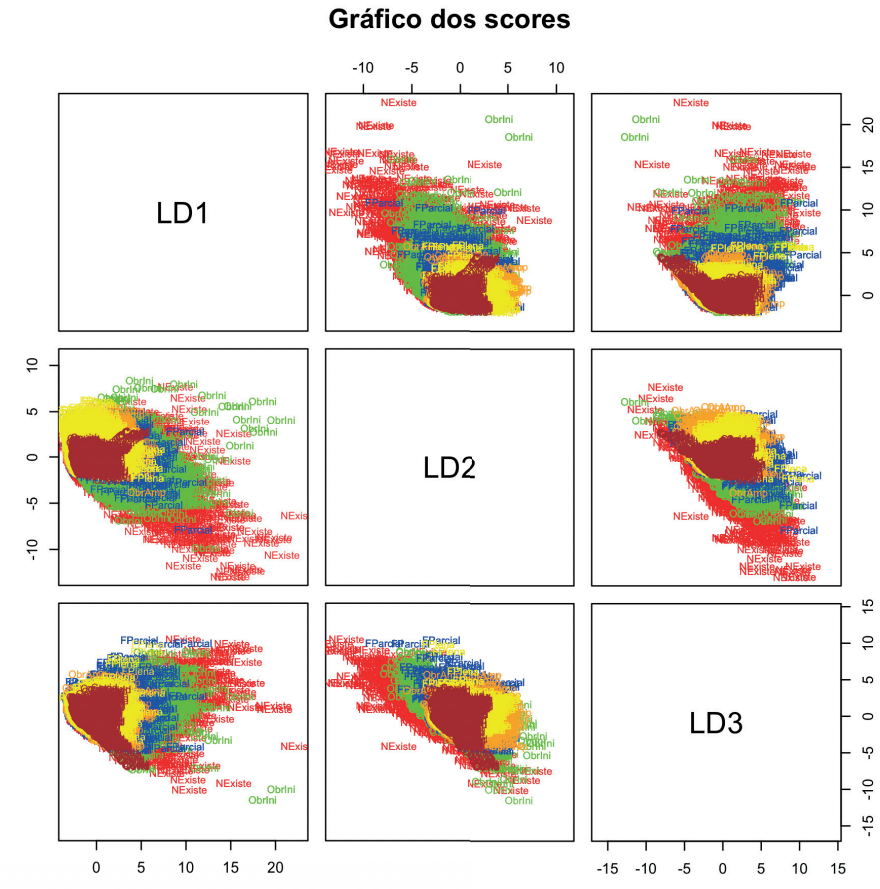

Figura 3.39: *Scores* das três funções discriminantes

```
    One-way MANOVA (Bartlett Chi2)

data:  x
Wilks' Lambda = 0.69587, Chi2-Value = 67018, DF = 15, p-value < 2.2e-16
sample estimates:
             PM10      NO2        CO
Covid    21.56710 42.06113 0.2916100
FParcial 41.61662 68.88620 0.4443295
FPlena   31.20772 57.04571 0.3160709
NExiste  58.82239 60.88411 0.6805091
ObrAmp   44.06825 64.44153 0.3807889
ObrIni   52.76065 66.90207 0.5867114
```

Figura 3.40: Teste lambda de Wilks

- *obam*: $p = \frac{137}{16551} = 0.0083$.
- *obin*: $p = \frac{4967}{28657} = 0.1733$.

Os três grupos onde se observam os valores dos *scores* mais diferenciados é onde a probabilidade da classificação correta é maior, ou seja, $fpln$, $nxst$ e $obin$.

	predicted					
original	Covid	FParcial	FPlena	NExiste	ObrAmp	ObrIni
Covid	2257	3	29161	234	0	272
FParcial	287	183	21793	533	39	1788
FPlena	1223	227	59523	428	47	1690
NExiste	1141	41	9664	5543	154	3394
ObrAmp	178	110	14808	203	137	1115
ObrIni	882	143	18671	3849	145	4967

Figura 3.41: Matriz de classificações do método de resubstituição

3.7.3 Método de Jacknife

est	Covid	FParcial	FPlena	NExiste	ObrAmp	ObrIni
Covid	2256	3	29162	234	0	272
FParcial	287	182	21793	533	39	1789
FPlena	1225	227	59520	428	48	1690
NExiste	1141	41	9664	5543	154	3394
ObrAmp	178	110	14808	203	137	1115
ObrIni	882	143	18671	3852	145	4964

Figura 3.42: Matriz de classificações do método de Jacknife

Na Figura 3.42 apresentamos a matriz de classificações do método de Jacknife. Ao observar a matriz obtida concluímos que as probabilidades vão ser praticamente iguais às obtidas com o método anterior.

3.7.4 Método de validação cruzada

	predicted					
original	Covid	FParcial	FPlena	NExiste	ObrAmp	ObrIni
Covid	942	12	14811	124	0	74
FParcial	234	342	11059	151	1	524
FPlena	734	361	29933	124	10	407
NExiste	435	266	6318	1341	36	1572
ObrAmp	79	288	7442	74	6	386
ObrIni	220	504	10932	871	15	1786

Figura 3.43: Matriz de classificações do método de validação cruzada

Na Figura 3.43 apresentamos a matriz de classificações do método de validação cruzada. A estimativa da probabilidade de uma observação estar corretamente classificada é dada por:

$$p = \frac{942 + 342 + 29933 + 1341 + 6 + 1786}{92414} = 0.3717,$$

ou seja, aproximadamente 37% das observações estão classificadas no grupo onde estavam originalmente, sendo este um valor semelhante ao obtido pelos outros dois métodos. A estimativa da probabilidade de uma observação estar corretamente classificada em cada um dos grupos é:

- $covd$: $p = \frac{942}{15963} = 0.0590$.
- $fpar$: $p = \frac{342}{12311} = 0.0278$.
- $fpln$: $p = \frac{29933}{31569} = 0.948$.
- $nxst$: $p = \frac{1431}{99687} = 0.2$.
- $obam$: $p = \frac{6}{8275} = 0.1345$.
- $obin$: $p = \frac{1786}{14328} = 0.0007$.

Embora estas probabilidades sejam ligeiramente diferentes em relação aos outros dois métodos, as mais elevadas continuam a ser as que estão associadas aos períodos $fpln$, $nxst$ e $obin$.

De destacar, ainda, que nos três métodos, o período $fpln$ tem uma probabilidade de classificação correta de aproximadamente 94%. As restantes probabilidades de classificação correta são bastante inferiores. Este facto poderá ser justificado por três motivos:

- Os seis períodos terem muitas medições semelhantes para os três poluentes.
- As três variáveis discriminantes não serem suficientes para discriminar os seis períodos.
- Os pressupostos desta técnica não terem sido verificados.

Capítulo 4

Conclusões e trabalhos futuros

Neste trabalho foram analisados dados referentes às concentrações de CO, NO_2 e PM_{10}, observados na EMQAr da Avenida da Liberdade, entre 1 de janeiro de 2001 e 31 de dezembro de 2022.

Os dados foram divididos em seis períodos disjuntos, tendo em conta o período antes do início da construção do túnel do Marquês de Pombal, o período de início de construção, o funcionamento parcial e obras de ampliação. Por fim, o período após a finalização da construção do túnel e o período de confinamento devido à pandemia provocada pelo vírus SARS-CoV-2.

Conseguiu concluir-se que ao longo do tempo houve uma diminuição dos valores da concentração de CO e NO_2 na maioria das medidas estatísticas analisadas, principalmente, a partir do período das obras de ampliação do túnel. Em relação às concentrações das PM_{10} observou-se um comportamento ligeiramente diferente. Houve uma diminuição dos valores das concentrações de PM_{10} até ao período de funcionamento parcial do túnel, tendo-se verificado um aumento no período das obras de ampliação e por fim observamos uma diminuição nos valores das medidas estatísticas analisadas. Em todos os períodos e para todos os três poluentes foram observados muitos picos de concentração (*outliers*).

A diminuição das concentrações nos dois últimos períodos poderão ser justificadas por vários fatores. No caso do penúltimo período, esta diminuição poderá ser justificada pela entrada em funcionamento pleno do túnel e pelo fim das obras. No caso do último período, a justificação para esta diminuição poderá estar associada à diminuição do tráfego devido às restrições de deslocação provocadas pela pandemia. A diminuição verificada na maioria dos valores dos três poluentes ao longo do tempo também poderá ser justificada pela proibição de circulação de veículos com matrícula anterior a 1992, a partir do 2º semestre de 2012, e de veículos com matricula anterior a 2000, a partir de 1 de janeiro de 2015. Outra possível justificação poderá, ainda, estar ligada às várias alterações na circulação em redor da Avenida da Liberdade. Foi possível concluir que os períodos considerados apresentam diferentes valores nas concentrações médias dos poluentes e também na distribuição dessas concentrações.

A análise discriminante linear permitiu concluir que as concentrações dos três poluentes permitem diferenciar os seis períodos considerados, embora com pouca eficácia, pois na avaliação da eficácia da análise obtivemos uma percentagem de classificação correta total entre 37% e 40%, dependendo do método usado.

De destacar, ainda, o período de funcionamento pleno do túnel que obteve uma probabilidade de classificação correta de aproximadamente 0.94. Aparentemente neste período as concentrações dos poluentes foram mais diferenciadas das dos outros períodos, pois há poucas concentrações a serem consideradas erradas, ou seja, a serem classificadas como concentrações

de outros períodos por terem maior semelhança com estas.

Considera-se que em trabalhos futuros seria interessante considerar um conjunto mais diversificado de poluentes, apesar de isso ser de difícil concretização se as observações continuarem a ser recolhidas exclusivamente na EMQAr da Avenida da Liberdade. Esta limitação é imposta pela reduzida informação disponibilizada pela generalidade das EMQAr da região de Lisboa.

Também parece que seria importante ter em conta possíveis ventos locais, que poderão influenciar a forma como os poluentes ambientais se espalham pela zona da Avenida da Liberdade, nomeadamente os que foram/são originários do túnel do Marquês de Pombal.

Bibliografia

Alkarkhi, A. F., & Alqaraghuli, W. A. (2020). *Applied statistics for environmental science with R*. Elsevier.

Agência Portuguesa do Ambiente (2021a). Redes de medição. Ministério do Ambiente e da Ação Climática.https://apambiente.pt/ar-e-ruido/avaliacao-da-qualidade-do-ar

Agência Portuguesa do Ambiente (2021b). Redes de medição. Ministério do Ambiente e da Ação Climática.https://apambiente.pt/ar-e-ruido/delimitacao-zonas-e-aglomeracoes

Agência Portuguesa do Ambiente (2021c). Redes de medição. Ministério do Ambiente e da Ação Climática.https://apambiente.pt/ar-e-ruido/redes-de-medicao

Agência Portuguesa do Ambiente (2021d). Redes de medição. Ministério do Ambiente e da Ação Climática.https://apambiente.pt/ar-e-ruido/metodos-de-medicao

Agência Portuguesa do Ambiente (2021e). Redes de medição. Ministério do Ambiente e da Ação Climática.https://apambiente.pt/ar-e-ruido/particulas-em-suspensao-pm

Agência Portuguesa do Ambiente (2021f). Redes de medição. Ministério do Ambiente e da Ação Climática.https://apambiente.pt/ar-e-ruido/oxidos-de-azono-nox

Agência Portuguesa do Ambiente (2021g). Redes de medição. Ministério do Ambiente e da Ação Climática.https://apambiente.pt/ar-e-ruido/monoxido-de-carbono-co

Castro, A., Araújo, R., & Silva, G. (2013). Qualidade do ar - Parâmetros de controle e efeitos na saúde humana: uma breve revisão. *Holos*, *5*, 107-121. https://redalyc.org/articulo.oa?id=481548607010

Dalgaard, P. (2008). *Introductory statistics with R* (2nd ed.). Springer.

European Environment Agency (2022). Air quality in Europe 2022. https://www.eea.europa.eu/en/analysis/publications/air-quality-in-europe-2022

Fernandes, R., & Ramos, P. (2025a). *Estatística descritiva e análise exploratória de dados*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2025b). *Amostragem e distribuições amostrais*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2025c). *Estimação*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2025d). *Testes de hipóteses*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2025e). *Análise discriminante linear*. Instituto Superior de Engenharia de Lisboa.

Fernandes, R., & Ramos, P. (2025f). *O software estatístico R*. Instituto Superior de Engenharia de Lisboa.

Gomes, J. (2010). *Poluição atmosférica* (2ª ed.). Publindústria.

Härdle, W., & Hlávka, Z. (2015). *Multivariate statistics* (2nd ed.). Springer.

Johnson, R., & Wichern, D. (2014). *Applied multivariate statistical analysis* (6th ed.). Pearson.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *318*(62), 399-402. https://doi.org/10.2307/2283970

Marôco, J. (2021). *Análise estatística com SPSS Statistics* (8ª ed.). ReportNumber.

Nemenyi, P. B. (1963). Distribution-free multiple comparisons (Publication No. 6406278) [Doctoral dissertation, Princeton University]. ProQuest Dissertations & Theses Global.

Pestana, M. H., & Gageiro, J. N. (2014). *Análise de dados para ciências sociais - A complementaridade do SPSS* (6ª ed.). Edições Sílabo.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*, 591-611. https://doi.org/10.2307/2333709

Venables, W. N., & Smith, D. M., & the R Core Team (2020). *An introduction to R*. https://cran.r-project.org/doc/manuals/R-intro.pdf

Zelterman, D. (2015). *Applied multivariate statistics with R*. Springer.