Análise de Dados em Informática

José Maia  
Departamento de Engenharia Informática  
Instituto Superior de Engenharia do PortoPorto, Portugal  
1191419@isep.ipp.pt

Lourenço Melo  
Departamento de Engenharia Informática  
Instituto Superior de Engenharia do PortoPorto, Portugal  
1190811@isep.ipp.pt

Artur Muiria  
Departamento de Engenharia Informática  
Instituto Superior de Engenharia do PortoPorto, Portugal  
1161274@isep.ipp.pt

# Introdução

Este artigo foi criado no âmbito do primeiro trabalho prático realizado na unidade curricular de Análise de Dados em Informática (ANADI), integrada na Licenciatura em Engenharia Informática (LEI) do Instituto Superior de Engenharia do Porto.

O enunciado é dividido em três ficheiros de dados. Foi feita a Análise Exploratória de Dados, Análise Inferencial, Correlação e Regressão para cada um dos dados fornecidos. Vai ser explicado como foram filtrados os dados, apresentadas as resoluções e demonstradas as conclusões retiradas.

# Metodologia de trabalho

Primeiramente, analisamos o enunciado em grupo e estudamos os requerimentos de todas as alíneas. Achamos que seria mais eficiente visto que são 3 exercícios dividir um exercício por aluno. Durante a realização do trabalho, o grupo trabalhou e tirou dúvidas durante a aula e trabalhávamos em casa comunicando via "Discord". Mesmo tendo havido divisão de tarefas, houve entreajuda e pariticipação de todos os elementos do grupo na totalidade do trabalho. Para a resolução dos exercícios foi utilizada a linguagem R e a plataforma RStudio, para além desta ferramenta usamos toda a informação disponibilizada no moodle da unidade curricular. Finalmente, depois da resolução de todos os exercícios passamos à elaboração do artigo científico em grupo onde revimos todos os problemas e discutimos os resultados.

Identify applicable funding agency here. If none, delete this text box.

# Análise e Discussão de Resultados

## Análise de rendimento da produção de 3 ESP - bomba elétrica para poços profundos

Primeiramente, leram-se os dados do ficheiro (“DADOS1.csv”) que contém dados medidos, a cada 5 minutos no espaço temporal de 1 de junho de 2013 às 00:00 até 12 de junho de 2014 às 14:50 das três bombas. Na leitura destes dados foram ignoradas as duas primeiras linhas do ficheiro visto que estas são linhas com propósitos informativos e influenciariam negativamente a facilidade como os dados seriam analisados.

De seguida, foram também renomeadas as colunas que iriam ser utilizadas, de maneira a facilitar o acesso.

***a***) Na resolução desta alínea utilizamos a função “as.POSIXct” para converter a coluna do tempo em objetos POSIXct. Assim, foi criada a coluna Tempo\_POSIX com o seguinte aspeto:

Nota-se que com a criação desta coluna o acesso e a filtração de dados por data tornaram-se significativamente mais simples.

***b)*** Para criar o gráfico primeiro teríamos de que filtrar os dados. A temperatura do motor das 3 bombas no dia 4 de agosto de 2013 foi filtrada utilizando a função subset() referindo a data e as colunas pretendidas. Com o dados filtrados usamos o ggplot() para desenhar o gráfico onde temos a temperatura do motor no eixo do y e a hora do dia no x. O gráfico obtido:

Analisando o gráfico podemos verificar que a bomba com temperaturas mais altas durante o dia é a bomba 3. A bomba 2 é a bomba com menores temperaturas durante o dia, mas é a que tem maior oscilação de temperaturas.

***c)*** Com os dados obtidos na alínea b) foi feito um boxplot com as temperaturas das bombas:

Com este boxplot podemos concluir que a bomba 2 tem a maior oscilação de temperaturas e as mais baixas das três bombas. A bomba 3 é a que tem as temperaturas mais altas, mas menos oscilação durante o dia. A bomba 1 tem maiores temperaturas que a 2 e menos que a 3 e as temperaturas variam mais do que a 3 e menos do que a 2.

***d)*** Nesta alínea somos informados que a quantidade de barris produzida num dia é igual ao cálculo da média das medições do “oil rate” no dia.

***d)i)*** Para obter os barris produzidos diariamente no mês de março de 2014 pelas bombas 1 e 2 temos novamente que começar por filtrar os dados usando um subset() onde referimos a data e as colunas necessárias. De seguida com os dados filtrados utilizamos a função aggregate() para calcular a média por dia do “oil rate” de cada bomba.

Finalmente usando o gplot() criamos o seguinte gráfico de barras :

Analisando o gráfico concluímos que a bomba 1 (A) produz, em todos os dias de março de 2014, mais barris de petróleo que a bomba 2.

***d)ii)*** Para conseguirmos descobrir em que mês a bomba 1 extraiu mais petróleo tivemos que filtrar os dados do “oil rate” entre os dias 1-6-2013 e 31-5-2014. De seguida, calculamos a media de “oil rate” diária de cada dia que nos dá a quantidade de barris total do dia. Para termos os barris totais do mês sumamos os barris produzidos diariamente desse mês usando um aggregate(). Finalmente ficamos com uma tabela “total-por-mês” onde temos o total de barris produzidos por mês:

Analisando a tabela obtida verificamos que o mês de agosto é o que tem maior produção de barris de petróleo.

***d)iii)*** Para obter os dias da amostra aleatória fizemos um subset() dos dias onde o índice correspondia a um valor da amostra aleatória. Ficamos assim, com os seguintes dias:

Finalmente, calculamos a produção diária de cada bomba nos dias da amostra e fizemos o boxplot:

Conseguimos verificar que em todos os dias da amostra a bomba 1 produz mais barris de petróleo que a bomba 2.

***d)iv)*** Visto que as amostras das duas bombas partilham os mesmos dias consideramos que as amostras são emparelhadas, logo, efetuamos um t.test() onde colocamos a variável “paired = true” e “alternative = "greater".

H0: A média da produção diária de petróleo da bomba 1 é igual à da bomba 2.

H1: A média da produção diária de petróleo da bomba 1 é maior do que da bomba 2.

Utilizamos os valores da alínea anterior e obtivemos o seguinte resultado:

O resultado indica que existe uma diferença significativa entre a bomba 1 e 2. O valor de t = 8.0043 e o valor de p = 2.175e-05 é extremamente baixo. Como p < 0.05 rejeitamos H0. Há evidências estatísticas nesta amostra para concluir que a média de produção diária de petróleo da bomba 1 é maior do que da bomba 2, para um nível de significância de 0.05.

***d)v)*** Analisando o boxplot da alínea iii) podemos concluir que o resultado obtido na alínea anterior corresponde à realidade. Conseguimos verificar que os valores de produção diária da bomba 1 são todos maiores que os da bomba 2.

## Análise da aceleração em viaturas

a) O ficheiro continha dados de 4 variáveis, neste caso, aceleração, número de cilindros, peso e potência de 99 viaturas. O problema consiste em descobrir se existiam diferenças significativas na aceleração entre os grupos de viaturas com 4, 6 e 8 cilindros. Para este problema foi definido o “α” para todas as situações como “0.05” como a convenção demanda, pois não foi referido um valor diferente para o “α” no enunciado do problema.

Começamos assim por importar o ficheiro “DADOS3.csv”. Apos isso, com comando “str” observamos com que tipo de dados estávamos a lidar e alguns exemplos dos primeiros valores. De seguida, criamos subconjuntos de dados para cada número de cilindros diferentes para no futuro termos a capacidade de comparar os subconjuntos.

Depois verificamos as medidas de tendência central de cada subconjunto e verificamos que a mediana das viaturas com 4 cilindros é bastante próxima comparativamente com as viaturas com 6 cilindros.

Para obtermos uma ideia visual de como os dados se posicionavam críamos um gráfico “boxplot” em que remove-mos os “outliers”, no intuito, de obter um resultado mais próximo da realidade. No gráfico pudemos observer uma disparidade do grupo das viaturas com 8 cilindros comparativamente com os grupos de viaturas com 4 e 6 cilindros, como podemos observar na [Figura 10].

Antes de realizar um teste de comparação, por boa prática, foram realizados teste à normalidade em todos os grupos na sua aceleração. Mas como a quantidade de observações de dados para cada grupo era diferente acabou-se por realizar-se testes diferentes à normalidade para cada grupo. Assim, para grupo com 30 ou mais observações realizamos o teste de “Lilliefors”, já para grupos com menos de 30 observações realizamos o teste de “Shapiro”.

Então para o grupo de 4 cilindros que possui 51 observações foi realizado o teste de “Lilliefors”.

H0: A aceleração do grupo segue uma distribuição normal

H1: A aceleração do grupo não segue uma distribuição normal

Como o “p-value” é igual a 0.2894 podemos concluir que não existem evidências estatísticas suficientes para rejeitar a hipótese nula pois p > 0.05.

Como o grupo de 6 cilindros possui 17 observações foi realizado o teste de "Shapiro”.

H0: A aceleração do grupo segue uma distribuição normal

H1: A aceleração do grupo não segue uma distribuição normal

Ao observar que o “p-value” é igual a 0.03628 indica que existem evidências suficientes para se rejeitar H0 e aceitar-se a hipótese alternativa. Assim a aceleração do grupo de viaturas com 6 cilindros não seguem uma distribuição normal.

Como o grupo de 8 cilindros possui 31 observações foi realizado o teste de “Lilliefors”.

H0: A aceleração do grupo segue uma distribuição normal

H1: A aceleração do grupo não segue uma distribuição normal

Ao observar que o “p-value” é igual a 0.7804 não evidências estatísticas suficientes para se rejeitar a hipótese nula.

Desta maneira, não foi possível realizar um teste “ANOVA”, porque a aceleração do grupo de viaturas com 6 cilindros não segue uma distribuição normal e por esse motivo, realizamos o teste “Kruskal-Wallis”.

H0: Não há diferença estatisticamente significativa entre as medianas dos grupos.

H1: Há evidencias estatísticas que existem diferenças significativas nas medianas entres os grupos.

Como o “p-value” é igual a 2.795e-11, a condição seguinte “p < 0.05” confirma-se. Então, por esta condição ser verdadeira rejeitamos a hipótese nula e aceitamos a hipótese alternativa, ou seja, existem evidencias estatísticas significativas para se rejeitar a hipótese nula e afirmar que há pelo menos uma diferença significativa na aceleração entre pelo menos dois grupos de viaturas com cilindros diferentes.

Com o intuito de descobrir quais grupos tinham essas diferenças significativas nas medianas entre os grupos foi realizado o teste de “Dunn” com um ajuste de “Bonferroni”.

Este teste irá determinar quais pares de grupos apresentam diferenças estatisticamente significativas entre si.

H0: Não existem diferenças significativas na aceleração entre os grupos.

H1: Existem diferenças significativas na aceleração entre os grupos.

Com estes resultamos podemos concluir que existem evidências estatísticas que confirma que há diferenças significativas na aceleração entre as viaturas de 4 e 8 cilindros e também entre as viaturas de 6 e 8 cilindros. Isto deve-se por se rejeitar a hipótese nula pois os “p-values” representados na [Figura 15] são menores que 0.00001 assim são menores do que o nível de significância que é 0.05. Desta forma, rejeita-se H0 e aceita-se H1 para os grupos referidos anteriormente.

Já para os valores de comparação entre os grupos de 4 e 6 cilindros não existem evidências suficientes para se rejeitar H0 pois o “p-value” é maior que o nível de significância.

Ao comparar o “boxplot” e os valores obtidos do teste de “Kruskal-Wallis” e “Dunn” dá nos uma forte evidência estatística do que querias resolver neste problema.

b)

i)

Nesta alínea o objetivo era encontrar o modelo de regressão linear, na qual era necessário considerar a variável número de cilindros uma variável “Dummy”.

Com base nos resultados da regressão linear podemos concluir o seguinte:

Os cilindros, o peso e a potência tem uma influência significativa na aceleração dos veículos de acordo com os valores dos “t-values” e “p-values”. O coeficiente de regressão do peso é positivo, ou seja, quanto maior o peso maior a aceleração, contrariamente às outras variáveis. Pelo valor estimado, o peso não tem grande influência na aceleração das viaturas comparativamente com os cilindros e o “Horsepower”. É possível apurar que os cilindros têm uma grande influência na aceleração das viaturas pelos valores “estimados”.

Como Pr(>|t|) nas variaveis de cilindros, horsepower e peso foram menores que o nível de significância, podemos afirmar que mais significativo é o seu coeficiente, ou seja, estes são menos prováveis de ter acontecido por acaso.

Assim por cada unidade que é incrementada no peso existe um aumento médio de 0.0033 na aceleração da viatura. Já por cada unidade que é incrementada na potência existe uma diminuição média de 0.0657 na aceleração da viatura. Já para cada incremento em cilindros nas viaturas é afetada negativamente a aceleração em média de 1.088 m/s^2.

ii)

Para este problema apenas era necessário estimar a aceleração de uma viatura com um peso de 2950 kg, 100hp de potência e 4 cilindros. Ao utilizar o modelo de regressão linear obtido anteriormente obtemos o resultado seguinte:

Ou seja, o resultado da aceleração previsto médio para uma viatura com aquelas características é de 17.30784 m/s^2.

## Análise de algoritmos de “Machine Learning”

Começou-se por buscar os dados contidos no documento “DADOS2.csv”, continha o cálculo de precisão dos 6 algoritmos de “Machine Learning” em 10 tipos de dados diferentes.

Instalou-se o pacote “readr” no “RStudio” para permitir a leitura de ficheiros em formatos “.csv”. Após a leitura colocou-se os dados numa variável “dados” para facilitar o acesso da informação.

#### A picture containing text Description automatically generated Utilizou-se a biblioteca “Hmisc” para possibilitar o uso comando ”rcorr” que cria uma matriz de correlações. Selecionou-se a Correlação Linear de Pearson, porque este caso trata-se de dados contínuos com amostras independentes.

Com base na matriz de correlações conseguiu-se notar que existem correlações positivas fortes entre os pares de algoritmos “GB” e “SVM” (0.86), “RF” e “DT” (0.88) e entre “ML” e “KN” (0.85).

Uma correlação positiva forte é indicada quando o coeficiente de correlação de “Pearson” é próximo de 1, o que se aplica neste caso. Pode-se também observar que existe pelo menos uma forte relação linear entra duas variáveis em que ambas aumentam ou diminuem juntas.

#### Aqui decidiu-se usar o Teste ANOVA para efetuar o teste de hipótese com o objectivo de verificar se existem diferenças significativas entre a precisão dos diferentes algoritmos. Optou-se pela ANOVA porque serve para comparar médias de duas ou mais amostras independentes e testar se existem diferenças estatísticamente significativas. Foi essencial formatar os dados e portanto dividiu-se apenas em duas colunas, uma chamada “algoritmo” e a outra de “precisao”.

Iniciou-se o processo de verificar os pressupostos do teste ANOVA ao efetuar o teste “Shapiro”, considerou-se um nível de significância de 5% (alfa = 0.05). Considerou-se que H0: p-value > 0.05 vs H1: p-value <= 0.05

A screenshot of a computer

Description automatically generated with medium confidenceConclui-se que como “p-value” é 0.0007857 < 0.05, rejeita-se H0 e assumiu-se que os dados da amostra não seguem uma distribuição normal. Como a normalidade foi rejeita decidiu-se recorrer a métodos alternativos como o teste de “Kruskal-Wallis”. Este teste é a alternativa não paramétrica do teste ANOVA.

Utilizou-se o mesmo nível de significância, onde H0: p-value > 0.05 vs H1: p-value <= 0.05.

H0: Não há evidencias de diferenças significativas entre a precisão dos diferentes algoritmos.

H1: Há evidencias de diferenças significativas entre a precisão dos diferentes algoritmos.

Text

Description automatically generated

Conclui-se que “p-value” é 0.3335 > 0.05 portanto rejeita-se H1 e confirmou-se não existência de evidencias de diferenças significativas entre a precisão dos diferentes algoritmos.

#### Realizou-se um teste de hipoótese com o apoio da função “kruskal.test()” para comparar a precisão dos diferentes algoritmos de “Machine Learning”. A decisão de utilização deste teste foi baseada no facto de que o teste de normalidade de “Shapiro” indicar que a distribuição dos dados em cada grupo não ser nomal.

O teste de “Kruskal-Wallis” indicou que não há diferenças significativas (p-value > 0.05). Como não houve diferenças significativas não essencial fazer o estudo “post-hoc” do teste.

##### Conclusão

O uso do R na análise de dados apresenta diversas vantagens, tais como a flexibilidade, disponibilidade de uma vasta variedade de ferramentas, pacotes e bibliotecas para diferentes tipos de análise. Facilita a visualização dos resultados através de gráficos personalizáveis e relatórios.

Durante a realização do trabalho em grupo, conseguirmos melhorar as nossas habilidades em R e técnicas aprendidas ao longo das aulas de Análise de Dados. Além disso, tivemos a oportunidade de expandir o nosso conhecimento na área, consolidando assim a aprendizagem.

##### References

# ISEP. (16 de Fevereiro de 2023). *Análise Exploratória de Dados - Estatística DescritivaFicheiro*. Obtido de Moodle ISEP: https://moodle.isep.ipp.pt/pluginfile.php/277049/mod\_resource/content/9/ANADI2122\_AED\_T.pdf

ISEP. (16 de Fevereiro de 2023). *Regressão Linear*. Obtido de Moodle ISEP: https://moodle.isep.ipp.pt/pluginfile.php/277082/mod\_resource/content/6/ANADI-LEI-1718\_RegressaoLinear.pdf

ISEP. (16 de Fevereiro de 2023). *Testes Paramétricos*. Obtido de Moodle ISEP: https://moodle.isep.ipp.pt/pluginfile.php/277061/mod\_resource/content/4/TestesParametricosPT.pdf

**from your pap**