Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2023/2024

Ficha Teórico-Prática 1

Estatística Descritiva

Objetivos:

- Familiarização com a linguagem de programação Phyton no suporte à Análise Exploratória de Dados;
- Breve revisão de Estatística Descritiva;
- Análise e discussão dos resultados.

1. A PORDATA, Base de Dados de Portugal Contemporâneo (https://www.pordata.pt/) , constitui uma base de dados com "informação sobre múltiplas áreas da sociedade, para Portugal, municípios e países europeus". Em **`Pordata>Portugal>Ambiente, Energia e Território>Poluição Atmosférica e Clima`**, encontra o número de dias sem chuva em estacoes meteorológicas portuguesas, nas últimas décadas. Escolha o botão Exportar para Excel.

a) Utilizando a função `read_excel`, importe da folha de calculo os dados numéricos com os respetivos rótulos de linhas e de colunas. Ajuste o parâmetro `range` de forma a incluir apenas estes dados;

b) Por visualização direta dos dados, verifique os rótulos das colunas. Se necessário, utilize a função `colnames` para renomear a coluna 1;

c) Construa um gráfico com os diagramas de extremos e quartis (*box plot*) que permita comparar os números de dias sem chuva nas estacoes meteorológicas da base de dados. Qual a estacão que registou mais e a que registou menos dias sem chuva nas últimas décadas? Analise o gráfico, referindo a concentração e a dispersão dos dados;

d) Comente a existência de zeros, nos dados e nos gráficos. Como podemos interpretar estes valores? Substitua todos os zeros por NA e repita o exercício anterior. Quais as diferenças mais notórias?

e) Refaça os gráficos restringindo às estacoes meteorológicas do continente e removendo os *outliers*. Comente os resultados;

f) Descreva em termos de quartis a distribuição do número de dias sem chuva em Castelo Branco. Repita com os dados relativos ao Porto;

g) Construa uma tabela de frequências para o número de dias sem chuva no Porto. Repita o exercício com os dados em classes definidas empiricamente e compare com a utilização da regra de Sturges;

h) Construa um gráfico que permita visualizar a forma da distribuição do número de dias sem chuva no Porto. O que pode observar no gráfico? Analise a tabela com as frequências observadas no gráfico;

i) Analise graficamente a evolução temporal do número de dias sem chuva no Porto. O que conclui? Repita a análise para o número de dias sem chuva em Faro.

2. ***Our world in data*** (https://ourworldindata.org/) , constitui simultaneamente uma base de dados e a organização que a suporta, baseada na Universidade de Oxford. Segundo a sua própria descrição, é composta de Investigação e dados para a evolução contra os maiores problemas mundiais. Em ***Technology Adoption*** encontra dados relativos à percentagem da população com acesso á Internet. Importe para CSV este conjunto de dados.

a) Crie uma estrutura de dados em R utilizando a função `read.csv`;

b) Altere de forma conveniente o nome da coluna 4;

c) Compare a distribuição de valores percentuais da população com acesso `a internet no ano 2000, com as mesmas distribuições nos anos 2010 e 2019;

d) Por inspeção direta dos dados, verifique se todas as linhas constituem dados relativos a países. Qual a importância deste facto para o resultado na alínea anterior? Filtre os dados de forma a considerar apenas os dados relativos a países e refaça os gráficos da alínea anterior;

e) Calcule:

i. O número de países ou organizações constantes dos dados;

ii. De entre os anteriores, quantos são países e quantos constituem regiões ou grupos de países?

iii. Os valores m´máximo, mínimo, médio e mediano da percentagem de população com acesso à Internet em 2019;

iv. A variância e a amplitude interquartil dos mesmos dados.

f) Identifique os 10 países com maior percentagem da população com acesso à Internet;

g) Represente graficamente a evolução temporal da percentagem da população com acesso à Internet em Portugal, Espanha, Dinamarca, União Europeia e na população global;

Exercícios de consolidação

1. Na PORDATA, em ***Pordata>Europa>Educação>Escolaridade da População***, encontram-se os dados da população europeia com ensino superior em percentagem da população, entre 25 e os 64 anos, por grupo etário. Exporte para Excel os dados disponíveis.

a) Utilizando a função `read_excel`, importe da folha de cálculo os dados numéricos com os respetivos rótulos de linhas e de colunas;

b) Altere de forma conveniente os rótulos das colunas;

c) Construa um gráfico com os diagramas de extremos e quartis (box plot) que permita comparar as percentagens da população com ensino superior, em 2020, nos distintos grupos etários. Qual o grupo etário com maior percentagem de população com ensino superior? Analise o gráfico, referindo a concentração e a dispersão dos dados;

d) Visualizando os dados, verifique a existência de zeros. Como podemos interpretar estes valores? Substitua todos os zeros por NA e repita o exercício anterior. Quais as diferenças mais notórias?

e) Compare em termos de média, as percentagens globais da população com ensino superior em 1992 e em 2020. Considerando o número de dados omissos, estes valores podem considerar-se sem reservas?

f) Descreva em termos de quartis a distribuição da percentagem da população com ensino superior em 2020. Compare com os valores observados em 1992;

g) Utilizando a regra de Sturges para classificar os dados, construa uma tabela de frequências para a percentagem da população europeia do grupo etário 25 − 34 anos com ensino superior em 2020. Repita com o grupo etário 55 − 64 e compare os resultados.

h) Represente graficamente as distribuições da alínea anterior. Customizando os parâmetros da função `hist`, represente adequadamente as duas distribuições de frequências no mesmo gráfico e descreva o resultado em termos das modas, da simetria e da dispersão dos dados.

2. Numa aula prática laboratorial de Algoritmia e Programação, o docente decidiu realizar um estudo do desempenho dos alunos, no sentido de avaliar qual o tipo de erro mais realizado. Para tal, sugeriu aos alunos a codificação de um dado algoritmo em C++. De seguida, pediu-lhes que compilassem o programa e analisassem o n.º de erros léxicos, sintáticos e semânticos cometidos.

Aluno	Erros Léxicos	Erros Sintáticos	Erros Semânticos
1	2	5	1
2	3	2	0
3	0	1	0
4	0	0	0
5	3	2	1
6	2	4	1
7	1	5	0
8	2	6	0
9	1	3	1
10	2	6	0
11	2	4	1
12	3	7	1
13	4	12	1

a) Construa um gráfico com os ***diagramas de extremos e quartis (box plot)*** que nos permita analisar o comportamento dos alunos pelo n.º de erros cometidos de cada tipo. Qual o tipo de erro mais cometido? Analise o gráfico, referindo a concentração e a dispersão dos dados.

b) Construa as ***tabelas de frequências*** para cada tipo de erro. Da análise da tabela indique o valor mediano de cada tipo de erro. Qual é o número de erros mais comum em cada tipo de erro?

c) Determine a média, o desvio padrão, o mínimo e o máximo para cada tipo de erro. Com base nestas medidas o que pode afirmar sobre os dados?

d) Construa um gráfico que permita visualizar a forma da distribuição de frequências da amostra. O que pode observar no gráfico?