Statistics for Engineers
TI0111

Homework 1

O objetivo do primeiro *homework* é revisar e se familiarizar com os conceitos de estatística descritiva. Você pode consultar os slides das aulas (de L01 a L04) para os conceitos teóricos e exemplos de código em R [The objective of the first homework is to revise and familiarise with the concepts of descriptive statistics. You may refer to the lecture slides (from L01 to L04) for the theoretical background and R code examples].

Instruções: Você deve entregar os seguintes itens [You must submit the following]:

- Relatório: Resolva cada exercício utilizando cálculos teóricos e descreva, de forma clara, as principais etapas do seu raciocínio. Em seguida, compare os resultados obtidos com aqueles gerados utilizando a linguagem R. Os pontos estão distribuídos igualmente entre as questões [Solve each exercise using theoretical calculations, and clearly describe the main steps of your reasoning. Then, compare your results with those obtained using the R programming environment. Points are equally distributed among the questions].

  - O relatório deve seguir uma estrutura lógica e bem organizada, garantindo clareza e facilitando a avaliação. Para cada exercício, inclua um parágrafo introdutório curto, apresentando brevemente o objetivo geral da tarefa e uma descrição do conjunto de dados utilizado. Descreva as ferramentas, pacotes e métodos estatísticos aplicados. Explique a resolução da questão, com os cálculos relevantes, gráficos (quando aplicável), tabelas e, principalmente, as interpretações [The report should follow a logical and well-organized structure to ensure clarity and ease of evaluation. For each exercise, include a short introductory paragraph briefly presenting the general objective of the task and a description of the dataset used. Describe the tools, packages, and statistical methods applied. Discuss the solution to the question, including relevant calculations, graphs (when applicable), tables, and especially interpretations].
  - Foque em clareza e concisão. Não inclua saídas brutas do R sem explicação, sempre interprete os resultados apresentados. Utilize gráficos e tabelas de forma eficiente, garantindo que cada elemento visual tenha título, legenda e seja pertinente à análise. Evite informações redundantes e mantenha uma formatação limpa e consistente ao longo do documento [Focus on clarity and conciseness. Do not include raw R output without explanation, always interpret the results. Use graphs and tables efficiently, ensuring each visual element has a title, legend, and is relevant to the analysis. Avoid redundancy and keep formatting clean and consistent throughout the document].
  - Utilize uma linguagem técnica, objetiva e impessoal. Evite expressões como "eu acho" ou "nós acreditamos". Prefira construções como "os resultados indicam que..." ou "observou-se que...". Estruture bem os parágrafos, organize o conteúdo por seções

e identifique claramente as respostas para cada item da atividade [Use a technical, objective, and impersonal tone. Avoid expressions like "I think" or "we believe". Prefer formulations such as "the results indicate that..." or "it was observed that...". Structure your paragraphs well, organize the content into sections, and clearly identify the answers to each item of the assignment].

- Listagem de Código: Envie o código utilizado para realizar a análise. O código deve ser funcional e executável, com todos os pacotes e dependências devidamente especificados. Ele pode ser incluído ao final do relatório como apêndice ou enviado separadamente em um arquivo compactado (`.zip`) [Submit the code used to perform the analysis. The code must be functional and executable, with all packages and dependencies clearly specified. It may be included at the end of the report as an appendix or submitted separately in a compressed file (`.zip`)].

- Crie e compartilhe com a professora um repositório Git (por exemplo, no GitHub)[1], de forma que seja possível acompanhar o desenvolvimento do projeto ao longo do tempo. O repositório deve conter: (i) o relatório final em PDF, (ii) todo o código-fonte utilizado na análise, (iii) quaisquer arquivos auxiliares (como scripts de acesso a dados ou arquivos de configuração), e (iv) um arquivo `README` com descrição do projeto, instruções claras de execução (incluindo dependências) e a identificação das contribuições individuais de cada membro do grupo. Submissões sem repositório acessível não serão avaliadas [Create and share with the instructor a Git-based repository (e.g., on GitHub)[2], so that the development can be followed throughout the process. The repository must include: (i) the final report in PDF format, (ii) all source code used in the analysis, (iii) any supporting files (such as data access scripts or configuration files), and (iv) a `README` file containing a project description, clear execution instructions (including dependencies), and a statement of each group member's contributions. Submissions without an accessible repository will not be evaluated].

É permitido consultar recursos externos; no entanto, todas as fontes devem ser devidamente citadas. Caso opte por utilizar ferramentas de IA na realização do trabalho, faça isso de forma responsável, para esclarecer conceitos ou verificar seu raciocínio, e não para gerar soluções completas. Utilizar IA para produzir respostas completas compromete seu aprendizado e infringe as normas de integridade acadêmica. Inclua tanto o prompt quanto a resposta gerada pela IA em um apêndice do relatório ou em um arquivo separado no repositório Git [You are allowed to consult external resources; however, all sources must be properly cited. If you choose to use AI tools while working on your homework, do so responsibly, to clarify concepts or check your reasoning, not to generate complete solutions. Using AI to produce full answers undermines your learning and violates academic integrity policies. Include both the prompt and the AI-generated output in an appendix of your homework report or in a separate file in your Git repository].

[1] Você pode utilizar outra plataforma baseada em Git, desde que o repositório esteja acessível sem restrições à professora.

O trabalho pode ser realizado individualmente ou em dupla. Nesse caso, é obrigatório especificar a contribuição real de cada aluno/aluna. O relatório deve ser enviado até o dia 23 de outubro de 2025, por meio do SIGAA. Submissões com atraso serão penalizadas da seguinte forma: até 24h = 20% de penalidade; até 48h = 40% de penalidade; e assim por diante [The work may be done individually or in pairs. In that case, you must clearly specify each student's actual contribution. The report must be submitted via SIGAA by October 23, 2025. Late submissions will be penalised as follows: up to 24h late = -20%; up to 48h = -40%; and so on].

Questão 1

As emissões diárias de um gás poluente de uma planta industrial foram registradas 80 vezes, em uma determinada unidade de medida. Os dados obtidos estão apresentados na Tabela 1 [The daily emissions of a pollutant gas from an industrial plant were recorded 80 times, using a specific unit of measurement. The data obtained are given in Table 1].

15.8	22.7	26.8	19.1	18.5	14.4	8.3	25.9	26.4	9.8	21.9	10.5
17.3	6.2	18.0	22.9	24.6	19.4	12.3	15.9	20.1	17.0	22.3	27.5
23.9	17.5	11.0	20.4	16.2	20.8	20.9	21.4	18.0	24.3	11.8	17.9
18.7	12.8	15.5	19.2	13.9	28.6	19.4	21.6	13.5	24.6	20.0	24.1
9.0	17.6	25.7	20.1	13.2	23.7	10.7	19.0	14.5	18.1	31.8	28.5
22.7	15.2	23.0	29.6	11.2	14.7	20.5	26.6	13.3	18.1	24.8	26.1
7.7	22.5	19.3	19.4	16.7	16.9	23.5	18.4				

Tabela 1: Emissões diárias de gas poluente (questão 1).

1. Calcule as medidas de tendência central (média, mediana e moda) e as medidas de dispersão (amplitude, variância, desvio padrão e coeficiente de variação) para o conjunto de dados da Tabela 1. Interprete os resultados [Calculate the central tendency measures (mean, median, and mode) and the dispersion measures (range, variance, standard deviation, and coefficient of variation) for the dataset in Table 1. Interpret the results].

2. Construa um histograma e um boxplot para os dados de emissões. Os dados parecem estar simetricamente distribuídos? Existem valores atípicos? [Create a histogram and a boxplot for the emission data. Do the data appear to be symmetrically distributed? Are there any outliers?]

3. Determine os quartis (Q1, Q2, Q3) e o intervalo interquartil (IQR). Utilize esses valores para reforçar sua análise sobre a presença de valores atípicos [Determine the quartiles (Q1, Q2, Q3) and the interquartile range (IQR. Use these values to support your analysis regarding the presence of outliers].

4. Suponha que o limite máximo aceitável diário para as emissões seja de 25 unidades. Qual a proporção de dias em que a planta excedeu esse limite? O comportamento geral das emissões estaria em conformidade com esse padrão regulatório? [Suppose the maximum acceptable daily emission limit is 25 units. What proportion of days exceeded this limit? Would the overall behavior of the emissions comply with this regulatory standard?]

Questão 2

Uma empresa italiana recebeu 20 currículos de cidadãos italianos e estrangeiros na seleção de pessoal qualificado para o cargo de gerente de relações exteriores. A tabela 2 reporta as informações consideradas relevantes na seleção: a idade, a nacionalidade, o nível mínimo de renda desejada (em milhares de euros), os anos de experiência no trabalho [An Italian company received 20 résumés from both Italian and foreign candidates during the hiring process for the position of Foreign Relations Manager. Table 2 reports the information considered relevant in the selection process: age, nationality, minimum desired income (in thousands of euros), and years of work experience].

	Idade	Nacionalidade	Renda	Experiência
1	28	Italiana	2.3	2
2	34	Inglesa	1.6	8
3	46	Belga	1.2	21
4	26	Espanhola	0.9	1
5	37	Italiana	2.1	15
6	29	Espanhola	1.6	3
7	51	Francesa	1.8	28
8	31	Belga	1.4	5
9	39	Italiana	1.2	13
10	43	Italiana	2.8	20
11	58	Italiana	3.4	32
12	44	Inglesa	2.7	23
13	25	Francesa	1.6	1
14	23	Espanhola	1.2	0
15	52	Italiana	1.1	29
16	42	Alemana	2.5	18
17	48	Francesa	2.0	19
18	33	Italiana	1.7	7
19	38	Alemana	2.1	12
20	46	Italiana	3.2	23

Tabela 2: Informações na seleção da empresa italiana (questão 2).

1. Calcule a média, mediana e desvio padrão para as variáveis idade, renda desejada e anos de experiência. O que você pode inferir a partir desses valores sobre o perfil

típico dos candidatos? [Compute the mean, median, and standard deviation for the variables age, desired income, and years of experience. What can you infer from these values about the typical candidate profile?]

2. Agrupe os candidatos por nacionalidade e calcule a renda média desejada e os anos médios de experiência para cada grupo. Qual nacionalidade apresenta a maior renda média desejada? Qual grupo aparenta ser o mais experiente? [Group the candidates by nationality and compute the average desired income and average years of experience for each group. Which nationality has the highest average desired income? Which group appears most experienced?]

3. Existe correlação entre anos de experiência e renda desejada? Utilize ferramentas visuais apropriadas (por exemplo, gráfico de dispersão) e calcule o coeficiente de correlação de Pearson. Interprete o resultado [Is there a correlation between years of experience and desired income? Use appropriate visual tools (e.g., scatter plot) and calculate the Pearson correlation coefficient. Interpret the result].

4. Suponha que a empresa queira priorizar candidatos com pelo menos 10 anos de experiência e renda desejada inferior a 2,0 (mil euros). Quantos candidatos atendem a ambos os critérios? Liste suas nacionalidades e idades [Suppose the company wants to prioritise candidates with at least 10 years of experience and a desired income below 2.0 (thousand euros). How many candidates meet both criteria? List their nationalities, and ages].

5. Construa gráficos que permitam visualizar a distribuição da idade e da renda desejada, separados por nacionalidade. Utilize histogramas, box-plots ou gráficos de barras, e comente as principais diferenças observadas entre os grupos [Create plots to visualize the distribution of age and desired income, separated by nationality. Use histograms, box-plots, or bar charts, and comment on the main differences observed between the groups].

Questão 3

O conjunto de dados em anexo, `HW1_bike_sharing.csv`[3], refere-se ao processo de compartilhamento de bicicletas em uma cidade dos Estados Unidos. O conjunto contém as colunas descritas na Tabela 3. A variável `season` inclui as quatro estações do hemisfério norte: primavera, verão, outono e inverno. A variável `weathersit` representa quatro condições meteorológicas: 'Céu limpo', 'Nublado', 'Chuva fraca', 'Chuva forte'. A variável `temp` é a temperatura normalizada em graus Celsius, ou seja, os valores foram divididos por 41 (valor máximo) [The attached dataset, `HW1_bike_sharing.csv`[3], refers to the bike sharing process in a U.S. city. The dataset comprises the variables listed in Table 3. The variable `season` includes the four seasons in the northern hemisphere: spring, summer, autumn, and winter. The variable `weathersit` represents four weather conditions:

[3] Os dados estão disponíveis no material do homework.

'Clear skies', 'Cloudy', 'Light rain', 'Heavy rain'. The variable `temp` is the normalised temperature in degrees Celsius, i.e., values are divided by 41 (the maximum)].

Tag	Descrição	Description
`instant`	Índice de registro	Record index
`dteday`	Data da observação	Date of observation
`season`	Estação do ano	Season
`weathersit`	Condições meteorológicas	Weather conditions
`temp`	Temperatura em °C (normalizada)	Temperature in °C (normalised)
`casual`	Número de usuários casuais	Number of casual users
`registered`	Número de usuários registrados	Number of registered users

Tabela 3: Variáveis do conjunto `HW1_bike_sharing` (questão 3).

1. Carregue o conjunto de dados `HW1_bike_sharing.csv` no R. Classifique as variáveis quanto ao tipo (categórica ou numérica), identifique o número total de observações e as datas de início e fim da amostra [Load the dataset `HW1_bike_sharing.csv` into R. Classify the variables by type (categorical or numerical), identify the number of observations and the start and end dates of the sample].

2. Calcule medidas de tendência central (média, mediana) e os quartis para cada característica numérica relevante. Apresente os resultados em uma tabela com título apropriado. Comente os principais pontos [Calculate measures of central tendency (mean, median) and quartiles for each relevant numerical feature. Present these statistics in a properly titled table. Comment on the main findings].

3. Atribua os níveis correspondentes às variáveis `season` e `weathersit`. Construa gráficos de barras para ambas. Qual estação do ano apresenta maior número de usuários? O uso de bicicletas depende da estação? Qual é a condição climática mais favorável para o uso do sistema? [Assign the appropriate levels to the variables `season` and `weathersit`. Create bar plots for both variables. Which season has the highest number of users? Is bike sharing usage dependent on the season? What is the most favourable weather condition?]

4. Calcule o número total de usuários por dia, somando `casual` e `registered`. Converta a variável `temp` para temperatura real (multiplicando por 41). Em seguida, construa os gráficos de séries temporais para temperatura e número total de usuários. Essas séries apresentam tendência semelhante? [Compute the total number of users per day by summing `casual` and `registered`. Convert the `temp` variable to actual temperature (multiply by 41). Then, plot time series for temperature and total users. Do these series show similar trends?]