



## HOMEWORK III

O objetivo do terceiro *homework* é de revisar e familiarizar-se com os conceitos de inferência estatística. Os conteúdos necessários para a resolução dos exercícios são abordados nos slides e nos códigos em R disponíveis no SIGAA das aulas L19 até L23 [The objective the third homework is to revise and familiarise with the concepts of statistical inference. The contents necessary for solving the exercises are covered in the slides and in the R codes available in SIGAA from the lectures L19 to L23].

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

- o 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].

- o 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)].
- o Crie e compartilhe com a professora um repositório Git (por exemplo, no [GitHub](#))<sup>1</sup>, 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)<sup>2</sup>, 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].

---

<sup>1</sup> Você pode utilizar outra plataforma baseada em Git, desde que o repositório esteja acessível sem restrições à professora.

O trabalho deve ser realizado individualmente ou em dupla. A pasta ZIP contendo a resolução da atividade deve ser enviada pelo SIGAA até o prazo indicado no sistema. Entregas atrasadas não serão aceitas [The work must be completed individually or in pairs. The ZIP folder containing the homework solution must be submitted through SIGAA by the deadline indicated in the system. Late submissions will no be accepted].

## QUESTÃO 1

Assume-se que o tempo de vida  $X$  (medido em anos) de um computador segue uma distribuição exponencial com parâmetro desconhecido  $\lambda > 0$ . Uma amostra aleatória dos tempos de vida dos computadores é apresentada na Tabela 1. Os dados são fictícios e são utilizados apenas para fins ilustrativos [It is assumed that the lifetime  $X$  (measured in years) of a computer follows an exponential distribution with unknown parameter  $\lambda > 0$ . A random sample of computer lifetimes in years is reported in Table 1. The data are fictitious and are used for illustrative purposes only].

0.99	2.31	10.85	6.15	10.81	3.72	5.75	4.15	9.27	7.84
2.31	10.85	6.15	1.81	3.72	5.75	10.40	10.04	4.15	9.27

Tabela 1: Dados usados na questão 1: Tempo de vida (em anos) dos computadores [Data used in question 1: Lifetime (in years) of computers].

1. Escreva a função densidade de probabilidade da distribuição exponencial com parâmetro [Write the probability density function of the exponential distribution with parameter]  $\lambda$ .
2. Dada uma amostra aleatória [Given a random sample]  $X_1, X_2, \dots, X_n$ :
  - (a) Escreva a função de verossimilhança [write the likelihood function]  $L(\lambda)$ .
  - (b) Derive a correspondente função log-verossimilhança [Derive the corresponding log-likelihood function]  $\ell(\lambda)$ .
  - (c) Determine o estimador de máxima verossimilhança (MLE, do inglês)  $\hat{\lambda}$  de  $\lambda$  [Determine the maximum likelihood estimator (MLE)  $\hat{\lambda}$  of  $\lambda$ ].
3. Utilizando os dados fornecidos na Tabela 1, calcule o valor numérico do MLE [Using the data provided in Table 1, compute the numerical value of the MLE]  $\hat{\lambda}$ .
4. Construa o gráfico da função log-verossimilhança  $\ell(\lambda)$  com base nos dados observados, considerando um intervalo adequado de valores para  $\lambda$ . Indique claramente no gráfico o valor do estimador de máxima verossimilhança  $\hat{\lambda}$  [Plot the log-likelihood function  $\ell(\lambda)$  based on the observed data, for a suitable range of  $\lambda$  values. Indicate clearly the value of the maximum likelihood estimate  $\hat{\lambda}$  on the graph].
5. Utilizando o parâmetro estimado [Using the estimated parameter]  $\hat{\lambda}$ :

- (a) Calcule o tempo médio de vida estimado de um computador [Compute the estimated mean lifetime of a computer].
  - (b) Calcule a probabilidade de que um computador funcione por mais de 5 anos [Compute the probability that a computer lasts more than 5 years].
6. A distribuição exponencial possui a *propriedade da falta de memória*, o que significa que a probabilidade de falha no futuro não depende do tempo que o computador já esteve em funcionamento [The exponential distribution has the *memoryless property*, meaning that the probability of failure in the future does not depend on how long the computer has already been operating].
- (a) Explique essa propriedade com suas próprias palavras [Explain this property in simple terms].
  - (b) Discuta brevemente se essa suposição parece razoável para modelar o tempo de vida de computadores [Briefly discuss whether this assumption seems reasonable for modelling computer lifetimes].

## QUESTÃO 2

O conjunto de dados de `penguins`, na biblioteca `palmerpenguins`<sup>3</sup> do R, contém medidas para as três espécies de pinguins (figura 1): ilha no arquipélago Palmer na Antártica, tamanho (comprimento da nadadeira, massa corporal, dimensões do bico) e sexo. Importe o conjunto de dados<sup>4</sup> e familiarize com ele. [The dataset `penguins`, included in the library `palmerpenguins`, includes measurements for the three penguin species (Figure 1): island in Palmer Archipelago in the Antarctic, size (flipper length, body mass, bill dimensions) and sex. Importe the data<sup>4</sup> and familiarise with it].

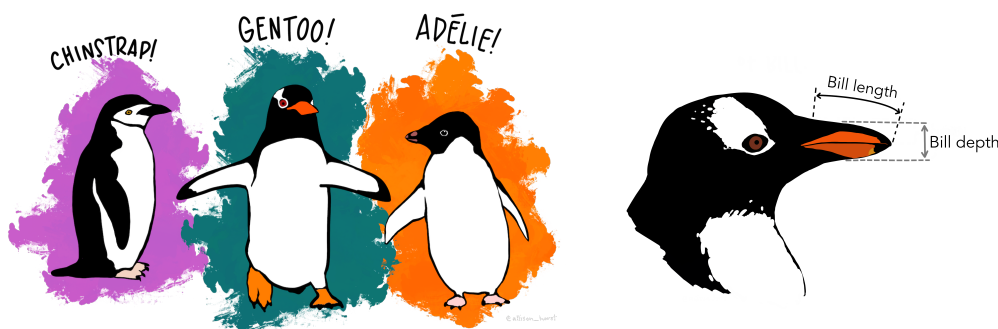


Figura 1: Espécies e características dos pinguins na questão [Penguin species and characteristics in question] 2.

<sup>3</sup> <https://cran.r-project.org/web/packages/palmerpenguins/index.html>

<sup>4</sup> `install.packages("palmerpenguins"); library(palmerpenguins);`  
`penguins_data <- na.omit(penguins) # desconsiderando os dados faltantes`

1. Considere a massa corporal (`body_mass`) em gramas como variável independente,  $x$ , e o comprimento do bico (`bill_length`) em milímetros como variável dependente  $y$ . Construa um gráfico de dispersão entre  $x$  and  $y$ . Com base no gráfico, comente se uma relação linear entre as variáveis parece plausível [Consider the body mass (`body_mass`) as the independent variable,  $x$ , and the bill length (`bill_length`) as the dependent one  $y$ . Build a scatterplot between  $x$  and  $y$ . Based on the plot, comment on whether a linear relationship between the two variables seems plausible].
2. Defina os parâmetros da reta de regressão com o método dos mínimos quadrados e verifique os resultados obtidos com o comando `lm()` no R. Adicione a reta de regressão no gráfico de dispersão [Estimate the parameters of the regression line with the least squares method and check the results obtained with the command `lm()` in R. Add the regression line to the scatterplot].
3. Calcule os resíduos da regressão e apresente uma representação gráfica dos mesmos. Em seguida, calcule a raiz do erro quadrático médio (RMSE, do inglês) e o coeficiente de determinação  $R^2$ . Comente sobre os resultados obtidos [Compute the regression residuals and provide a graphical representation. Next, calculate the root mean squared error (RMSE) and the coefficient of determination  $R^2$ . Comment on the results].
4. O conjunto de dados não apresenta outliers evidentes. Modifique esse conjunto introduzindo artificialmente uma observação extrema, seja por meio de um aumento ou de uma redução substancial no valor da massa corporal ou do comprimento do bico de um dos pinguins. Em seguida, ajuste um modelo de regressão linear utilizando o conjunto de dados modificado. Compare os coeficientes estimados da regressão, as retas ajustadas e os valores do RMSE e do  $R^2$  com aqueles obtidos no item 2. Por fim, discuta a influência da observação artificialmente introduzida sobre os resultados da regressão [The dataset does not present evident outliers. Modify the dataset by artificially introducing an extreme observation, either by substantially increasing or decreasing the body mass or the bill length of one of the penguins. Then, fit a linear regression model using the modified dataset. Compare the estimated regression coefficients, the fitted regression lines, and the RMSE and  $R^2$  values with those obtained in item 2. Finally, comment on the influence of the artificially introduced observation on the regression results].