***Titanic – Análise Estatística***

João Espada 1, José Almeida 2, Tiago Silva 3

1 PL5, 202100660

2 PL5, 202100661

3 PL5, 202000331

**Resumo:** TODO

**Palavras-chave**: TODO

**1. Introdução**

TODO

**2. Base de Dados**

A base de dados selecionada diz respeito aos passageiros do navio *Titanic*, que naufragou em 1912. O objetivo desta análise é estudar fatores que podem ter influenciado a probabilidade de sobrevivência dos passageiros, com base em características individuais como idade, sexo, classe social e outras variáveis disponíveis.

A base de dados foi obtida a partir da plataforma *Kaggle*, sob o nome "*Titanic: Machine Learning from Disaster*". A versão usada neste trabalho está contida no ficheiro titanic\_data.csv, que contém informações de passageiros para análise.

Esta amostra contém um total de 418 observações (passageiros) e 10 variáveis estatísticas.

**Tabela 1** - Variáveis estatísticas

|  |  |  |
| --- | --- | --- |
| **Nome da Variável** | **Tipo** | **Descrição** |
| **Pclass** | Qualitativa (ordinal) | Classe socioeconómica do passageiro (1ª, 2ª, 3ª classe) |
| **Survived** | Qualitativa (nominal) | Sobrevivência |
| **Sex** | Qualitativa (nominal) | Sexo do passageiro (masculino/feminino) |
| **Age** | Quantitativa (contínua) | Idade do passageiro em anos |
| **SibSp** | Quantitativa (discreta) | Número de irmãos/cônjuges a bordo |
| **Parch** | Quantitativa (discreta) | Número de pais/filhos a bordo |
| **Ticket** | Qualitativa (nominal) | Número do bilhete |
| **Fare** | Quantitativa (contínua) | Valor pago pelo bilhete (em libras) |
| **Cabin** | Qualitativa (nominal) | Número da cabine |
| **Embarked** | Qualitativa (nominal) | Porto de embarque (C = Cherbourg, Q = Queenstown, S = Southampton) |

**2.1. Análise Descritiva**

As legendas das tabelas são apresentadas por cima das mesmas, enquanto as legendas das figuras e dos gráficos são colocadas por baixo das mesmas (ver exemplos, fonte Calibri 10, espaçamento simples). Nestas legendas as palavras **Tabela, Figura, Gráfico** e a respetivanumeração sequencial são escritas a negrito e seguidas de um traço horizontal ( – ).

**Tabela 1** – Blablabla (Calibri 10)

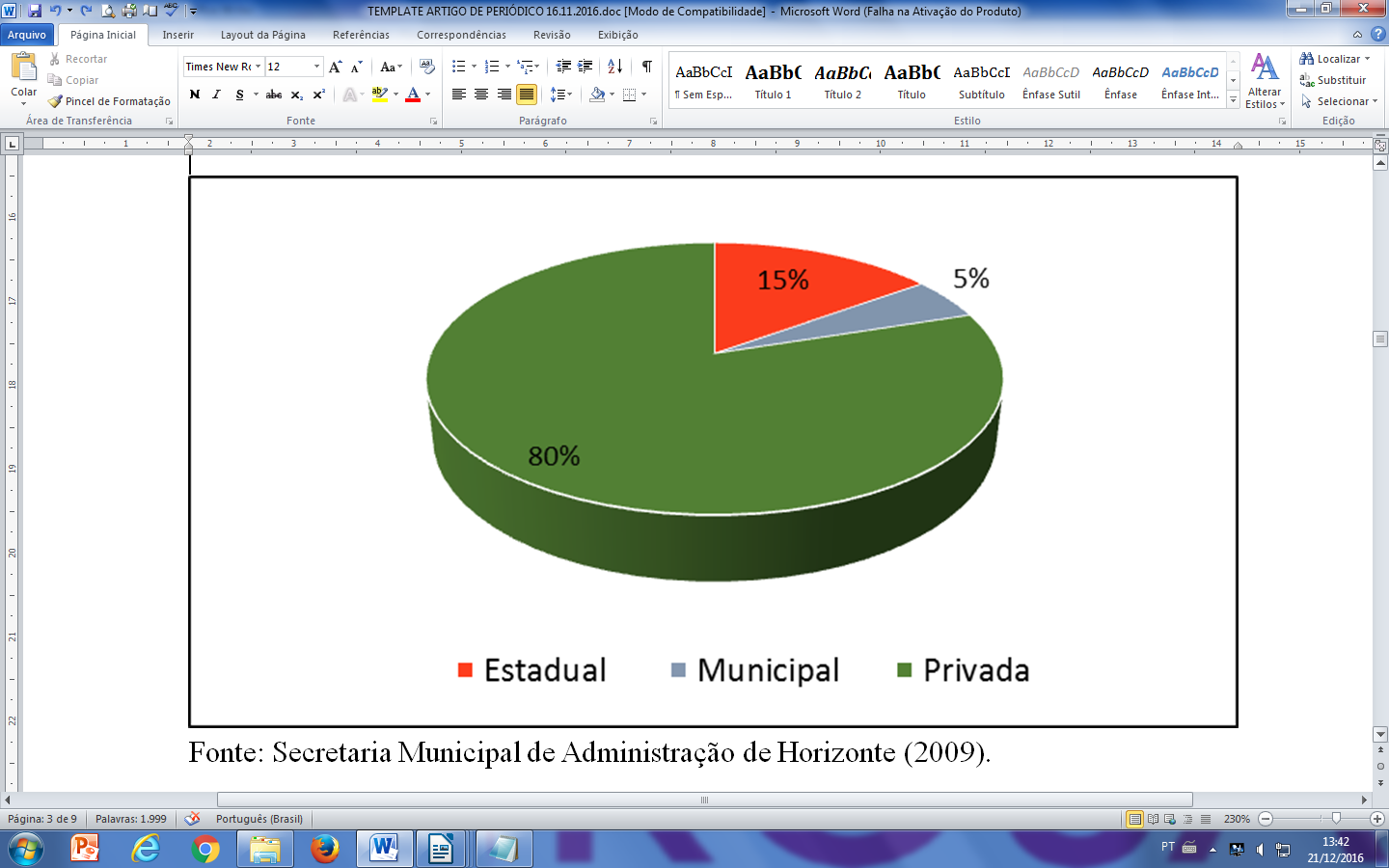
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **E** | **A (cinza 10%)** | **B (cinza 10%)** | **C** | **D** |
| I | 125 | 123 | 234 | 123 |
| II | 23454,83 | 45425,72 | 2456,00 | 2466,13 |
| III | Aiosi | Ioisio | Iosio | Ziipio |

Entre a tabela e o texto que segue deve ser introduzida uma linha em branco. O texto dentro da tabela deve aparecer centrado verticalmente relativamente à célula em que se insere.



**Figura 1** – Blablabla (Calibri 10)

Todas as tabelas, figuras e gráficos devem ser referidos no texto. A referência a tabelas, figuras e gráficos deve ser feita por extenso e iniciar com maiúscula (por exemplo, Tabela 1 e Figura 2 e Gráfico 3).



**Gráfico 1** – Blablabla (Calibri 10)

As equações devem seguir o exemplo abaixo e ser numeradas sequencialmente ao longo do trabalho, sendo a numeração colocada entre parêntesis e alinhada à direita.

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

**2.2. Testes de Ajustamento**

Deve indicar as hipóteses testadas, justificar a escolha e apresentar todos os dados necessários para compreender o que foi feito.

**2.3. Teste de Independência e Regressão Linear Simples**

No Teste de Independência do Qui-Quadrado, deve indicar as hipóteses que está a testar, apresentar a tabela de contingência associada e, se for adequado, quantificar a associação existente. Na Regressão Linear Simples, deve definir a variável dependente e independente, indicar a equação da reta de regressão, interpretar os coeficientes obtidos, avaliar a qualidade do modelo ajustado e fazer uma análise dos resíduos.

**3. IA**

Se recorreu a IA para a realização do trabalho, deve incluir uma secção detalhando os *prompts* específicos que foram utilizados. Além disso, deve também incluir na bibliografia do trabalho o link para a IA utilizada.

**4. Conclusões**

TODO

**Referências Bibliográficas**

* Kaggle (2022). Titanic - Machine Learning from Disaster (Test File). Kaggle Datasets. URL <https://www.kaggle.com/datasets/brendan45774/test-file>.
* [Home - RDocumentation](https://www.rdocumentation.org/)
* [Create Elegant Data Visualisations Using the Grammar of Graphics • ggplot2](https://ggplot2.tidyverse.org/)
* [Bar charts — geom\_bar • ggplot2](https://ggplot2.tidyverse.org/reference/geom_bar.html)
* [A Grammar of Data Manipulation • dplyr](https://dplyr.tidyverse.org/)
* Capítulo 1 - Estatística Descritiva - ME-LEI\_24\_25.pdf - Departamento de Matemática Escola Superior de Tecnologia de Setúbal Instituto Politécnico de Setúbal 2024-2025

**Informação de Suporte:**

**📊 1. Tabelas de Frequência**

Para variáveis:

* Survived (qualitativa nominal)
* Sex (qualitativa nominal)
* Pclass (qualitativa ordinal)
* Embarked (qualitativa nominal)
* SibSp e Parch (quantitativas discretas)

**📉 2. Gráficos**

* Barras ou circulares (variáveis qualitativas)
* Histogramas (quantitativas contínuas: Age, Fare)

**📐 3. Medidas de Localização e Dispersão**

Para Age e Fare: média, mediana, moda, quartis, desvio padrão, amplitude, etc.

**📦 4. Boxplots**

Para comparar distribuição por categorias (ex: idade por Pclass ou Sex)

**📊 Porquê usar tabelas de frequência?**

As **tabelas de frequência**:

* resumem os dados categóricos (qualitativos ou discretos),
* mostram **quantas vezes** cada categoria aparece (**frequência absoluta**),
* mostram **a proporção** de vezes em relação ao total (**frequência relativa**).

**✅ Variáveis que vamos analisar**

**1. Survived (Sobreviveu)**

* Tipo: **Qualitativa nominal**
* Valores: 0 = Não sobreviveu, 1 = Sobreviveu
* **Importância**: é a variável principal que queremos compreender (alvo da análise)

**2. Sex (Sexo)**

* Tipo: **Qualitativa nominal**
* **Importância**: comparar taxas de sobrevivência por género

**3. Pclass (Classe socioeconómica)**

* Tipo: **Qualitativa ordinal**
* Valores: 1, 2, 3 (1ª, 2ª e 3ª classe)
* **Importância**: pode haver relação entre classe social e sobrevivência

**4. Embarked (Porto de embarque)**

* Tipo: **Qualitativa nominal**
* Valores: C (Cherbourg), Q (Queenstown), S (Southampton)
* **Importância**: entender o perfil dos passageiros por porto

As variáveis quantitativas contínuas não foram incluídas na fase de tabelas de frequência pois têm muitos valores únicos e **precisam de ser agrupadas em classes** para serem resumidas numa tabela de frequência