# **Estudo sobre a base Titanic**

1. Introdução

O objetivo deste estudo é aplicar os conhecimentos aprendidos em Estatística Computacional com o professor Luís Machado (MECD) em uma base real, apresentando os passos e comandos utilizados nos softwares R e SPSS e, além disso, procedermos a Análise Descritiva dos Dados.

A base escolhida foi a do Titanic porque se trata de um acidente histórico de um navio luxuoso construído para ser 'inafundável' (segundo o imaginário popular, foi justamente a arrogância humana um fator crucial para esse terrível acidente).

Além disso, referida base é interessante para o propósito do trabalho porque apresenta poucas variáveis, o que facilitará a classificação e análise de todas elas, e ainda assim apresenta os tipos comuns de variáveis (qualitativa - nominal e ordinal - e quantitativa - discreta e contínua).

A quantidade de observações também é interessante (1309 registros), pois não é um número pequeno e nem abrange todas as 3.327 pessoas que embarcaram no Titanic, caracterizando-se, assim, como uma amostra.

892 tripulantes e 2.435 passageiros (de acordo com o site Wikipédia). Por termos, então, apenas uma amostra da população podemos enriquecer o trabalho apresentando uma modelagem estatística para prever se uma pessoa possa ter sobrevivido (ou não) ao acidente.

# , para efeitos deste estudo, vamos supor que trata-se de uma amostra aleatória simples e que a falta de informação das outras pessoas não influencie na chance de sobreviver (ou não) ao acidente.

E por se tratar de uma amostra, além de proceder à Análise Descritiva, podemos construir modelos estatísticos para tentar prever a chance de uma pessoa ter sobrevivido ao acidente, baseando-nos nas características que influenciaram na sobrevivência (ou não) das pessoas à viagem.#

1. Primeiras Análises

2.1 - Dicionário de variáveis

O primeiro passo é conhecermos a base. Inicialmente temos 891 observações e 12 seguintes variáveis:

pclass: Classe do Bilhete. Possíveis valores: 1(primeira), 2(segunda) ou 3 (terceira). Classificação: Variável Qualitativa Ordinal

survived: Sobrevivência (ou não) do passageiro. Possíveis valores: 0 (morreu) ou 1(sobreviveu). Classificação: Variável Resposta (de interesse). Quantidade de sobreviventes na amostra: 500 pessoas(38,20%).

sex: Gênero do passageiro. Possíveis valores: female (feminino) ou male(masculino). Classificação: Variável Qualitativa Nominal

age: Idade em anos. Possíveis valores: Para crianças menores de um ano, a idade pode assumir um valor contínuo (no intervalo ]0,1[). Para mais velhos que isso, usa-se o valores discretos {1,2,3...}. Quando tem o valor da forma xx.5 é porque a idade da pessoa foi estimada. Não obstante essas informações, podemos classificar a Variável como Quantitativa Contínua.

sibsp: Número de irmãos e/ou esposos a bordo. Possíveis valores: {0,1,2,3...}. Classificação: Variável Quantitativa Discreta

parch: Número de pais e/ou filhos a bordo. Possíveis valores: {0,1,2,3...}. Classificação: Variável Quantitativa Discreta

fare: Valor da tarifa paga. Possíveis valores: ]0,+∞[ Classificação: Variável Quantitativa Contínua

embarked: Porto de Embarque. Possíveis valores: C (Cherbourg), Q (Queenstown) ou S (Southampton). Classificação: Variável Qualitativa Nominal.

death: Morte (ou não) do passageiro. Possíveis valores: 0 (não morreu) ou 1(morreu). Esta variável será descartada porque é basicamente oposta à variável survived.

Cabe ressaltar, seria mais interessante se a variável sibsp fosse dividida em duas (uma só para irmãos e outra só para esposas). Com efeito, com essas duas informações separadas, poderíamos facilmente construir a variável sibsp caso desejássemos. Mesmo raciocício se aplica para a variável parch.

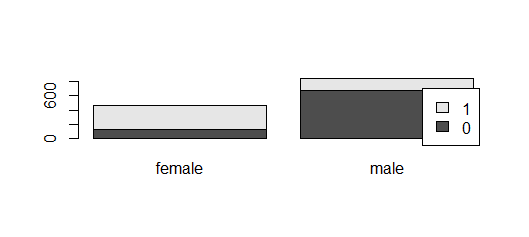
Por fim, podemos pensar em construir outras variáveis. Por exemplo, construir uma marcação(flag) para a idade (real ou estimada); podemos também agrupar valores contínuos (ou até discretos) em classes. Estas (e outras) transformações serão apresentadas no decorrer deste trabalho.

2.2 - Variáveis Qualitativas

2.2.1 - Variáveis Qualitativas Nominais

**Sex (female ou male)**

Das 1309 observações temos 466 mulheres (35,60%) e 843 homens (64,40%), sem valores faltantes, portanto. No gráfico a seguir, veremos para cada uma dessas categorias o comportamento da variável resposta.

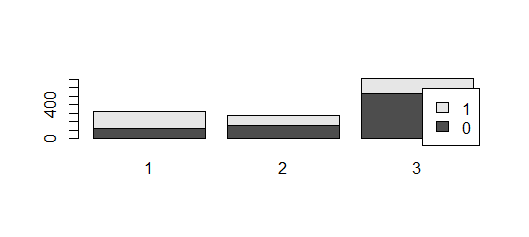


Pela apreciação do gráfico, percebe-se que as mulheres tiveram mais chances de sobrevivência que os homens.

2.2.2 - Variáveis Qualitativas Ordinais

**Pclass (1, 2 ou 3)**

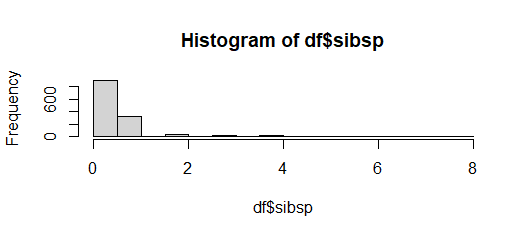
Das 1309 observações temos 323 da primeira (24,68%), 277 da segunda (21,16%) e 709 da terceira (54,16%) , sem valores faltantes, portanto. No gráfico a seguir, veremos para cada uma dessas categorias o comportamento da variável resposta.



Aqui podemos ver que as pessoas da primeira classe tiveram mais chances de sobrevivência, seguindo pela segunda classe, tendo portanto a terceira classe com as maiores taxas de morte.

2.2.3 - Variáveis Quantitativas Discreta

**Sibsp (Número de irmãos e ou esposos)**



Como podemos observar no gráfico, a maioria dos valores assume o valor 0 ou 1. Talvez podemos analisar esta variável agrupando em três categorias (0, 1 e maior que 1) - ou em duas(“Não tem” ou “Tem”).

1. Conclusão

De acordo com o exposto, pessoas ricas e do sexo feminino tem mais chances de sobreviverem ao acidente do Titanic.