UNIVERSIDADE FEDERAL DE UBERLÂNDIA

FACULDADE DE MATEMÁTICA

PRIMEIRA PROVA DE ESTATÍSTICA - 03/04/2023

Questão 1. Os arquivos `treino_baleias.txt` e `teste_baleias.txt` contém informações sobre as características de algumas espécies de baleias. Os conjuntos de dados possuem, ao todo, 248 observações (198 para treino, 50 para teste). As variáveis incluídas nestes conjuntos de dados são: **especie**: indica a espécie da baleia e é uma variável categórica; **comprimento**: indica o comprimento da baleia em metros e é uma variável numérica contínua; **peso**: indica o peso da baleia em quilos e é uma variável numérica contínua; **profundidade_maxima**: indica a profundidade máxima mergulhada pela baleia em metros e é uma variável numérica contínua; **volume_cranio**: indica o volume do crânio da baleia em centímetros cúbicos e é uma variável numérica contínua.

Os itens de (a) até (g) devem ser respondidos usando apenas os dados de `treino_baleias.txt`. Em (h), os dois conjuntos devem ser utilizados.

(a) Crie um conjunto para cada espécie de baleia; cada data frame criado deverá conter apenas baleias de uma espécie. (1 ponto)

(b) Calcule a média, a variância, o desvio padrão e o coeficiente de variância para a variável peso para cada espécie de baleia. Comente os resultados obtidos. (1 ponto)

(c) Escolha uma espécie de baleia e calcule a mediana da variável volume_cranio para esta espécie escolhida. Interprete o valor obtido. (0.5 ponto)

(d) Apresente o histograma da variável peso para a espécie de baleia azul. Comente os resultados obtidos. (0.5 ponto)

(e) Apresente numa mesma janela os boxplots para cada espécie para a variável comprimento. Comente os resultados obtidos. (1 ponto)

(f) Refaça o item anterior para as outras variáveis numéricas. (3 pontos)

(g) Apresente um gráfico de dispersão de comprimento versus profundidade_maxima. Cada espécie deve ser registrada por uma cor diferente. (2 pontos)

(h) Com base em todas as informações anteriores, construa um modelo de árvore de decisão (não pode utilizar funções em R que retornam a árvore pronta; construa sua própria árvore a partir das técnicas que utilizamos em sala de aula) para prever a espécie de uma baleia com base nas variáveis numéricas do estudo. Justifique as escolhas das variáveis e dos pontos de corte escolhidos. Por fim, utilize o conjunto do arquivo `teste_baleias.txt` para calcular a taxa de acerto. Comente o resultado obtido. (5 pontos)

(i) Este é um item bônus: utilize gráficos de dispersão para registrar por linhas horizontais e verticais os pontos de cortes escolhidos em sua árvore de decisão. As espécies de baleias devem ser registradas por diferentes cores. (1.5 pontos extras)

Questão 2. Considere o seguinte jogo: Steven e Garnit escolherão, cada um, uma sequência de tamanho 3 em que cada entrada da sequência é cara ou coroa; logo em seguida, uma moeda será lançada três vezes; se aparecer a sequência de um dos jogadores, este jogador vence e o jogo acaba; caso não apareça a sequência de nenhum deles, a moeda é lançada pela quarta vez e os três últimos lançamentos são analisados; se nestes três últimos lançamentos aparecer a sequência de um dos jogadores, este jogador vence e o jogo acaba. Se isto não acontecer, a moeda é lançada pela quinta vez e os três últimos resultados são analisados; se aparecer a sequência de um dos jogadores, este jogador vence e o jogo acaba. Este processo é realizado até que apareça a sequência que um dos dois escolheu; se aparecer primeiro a sequência de Steven, ele ganha;

se aparecer primeiro a sequência de Garnit, ela vence. Convencione que cara seja 1 e que coroa seja zero. Supondo que Steven escolheu a sequência (0, 1, 0) e que Garnit escolheu a sequência (0, 0, 1), simule uma partida deste jogo. A simulação deve retornar "steven" caso Steven tenha vencido ou deve retornar "garnit" caso contrário. Após uma simulação, replique o experimento 10 mil vezes e calcule a média de vitórias de Garnit. Comente o resultado obtido. (6 pontos)

Observação: Suponha que os três primeiros lançamentos foram saído (1,0,0). Logo, ninguém ganhou e a moeda é lançada pela quarta vez. Suponha que o quarto lançamento foi 0; logo os três últimos lançamentos foram (0,0,0) e ninguém ganhou. Na quinta vez saiu 1 e, portanto, os três últimos lançamentos foram (0,0,1) e o jogo acaba com vitória de Garnit. As sequências (0, 1, 0), (1, 0, 1, 0) e (1, 1, 0, 1, 0) deixam Steven vitorioso; as sequências (0, 0, 1), (0, 0, 0, 1) e (1, 0, 0, 0, 1) deixam Garnit vitoriosa.

Questão bônus. Benjamin Hartley (Bury St Edmunds, 18 de março de 1945 — Londres, 22 de junho de 2005), conhecido como "Doutor da Morte", foi um médico cardiologista e assassino em série britânico condenado pela morte de mais de 200 pacientes entre as décadas de 1970 e 1990. Dr. Hartley é, talvez, o assassino em série mais prolífico da História Moderna. O arquivo `dados.txt` contém informações sobre o sexo, a idade e a hora da morte (inteira) das vítimas de Benjamin Hartley.

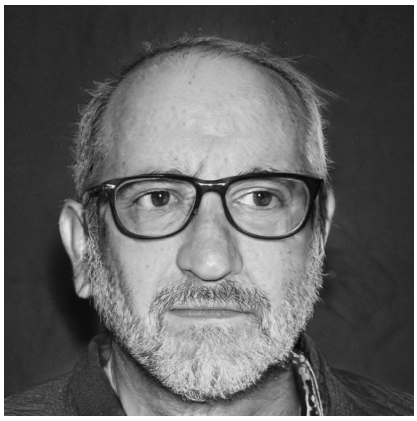

Figura 1: Dr. Benjamin Hartley, 1999.

Antes de responder as questões abaixo, abra o arquivo `dados.txt` e compreenda sua estrutura. Importe o arquivo para o R e utilize-o para responder os seguintes itens.

(a) Escolha um gráfico apropriado para representar as frequências das categorias da variável sexo. Comente os resultados encontrados. (0.5 ponto)

(b) Apresente o histograma da variável idade em 6 intervalos. Comente os resultados obtidos. (1 ponto)

(c) Apresente o boxplot da variável idade. Comente os resultados obtidos. (0.5 ponto)

(d) Apresente o histograma da variável hora. Comente os resultados obtidos. (0.5 ponto)

(e) Com base nas informações obtidas nos itens anteriores, escreva um parágrafo sobre o padrão e o perfil das vítimas de Benjamin Hartley. (1 ponto)