

REVISÃO PARA PRIMEIRA PROVA DE ESTATÍSTICA COMPUTACIONAL

Questão 1. Vovó Juju e seu neto Jorel estão apostando num jogo de cara ou coroa. Uma moeda é lançada: se sair cara, Vovó Juju recebe 1 real de Jorel; se sair coroa, Vovó Juju paga 1 real para seu neto. No início do jogo, Vovó Juju tem 18 reais e Jorel tem 7 reais. O jogo termina quando algum dos jogadores perde todo o seu dinheiro. Simule uma partida deste jogo. A simulação deve retornar o nome do vencedor do jogo. Agora, simule 10 mil vezes este experimento e calcule a média de vezes em que Jorel ganhou e a média de vezes que Vovó Juju ganhou. Interprete o resultado.

Questão 2. Uma urna contém bilhetes numerados de 1 até 30, todos do mesmo tamanho. Considere o seguinte experimento: retirar um bilhete da urna, registrar o resultado e devolver o bilhete para a urna. Você continuará a sortear bilhetes até que todos os números sejam retirados. Seja N o número de sorteios que você precisou realizar até que todos os números fossem registrados. Simule este experimento; a simulação deve retornar o valor N . Agora, simule 10 mil vezes este experimento. Calcule a média dos resultados obtidos e interprete o valor encontrado.

Questão 3. O arquivo `chicago.csv` contém dados de mortalidade diária (todas as causas exceto acidentes (`death`), doenças cardiovasculares (`cvd`), respiratórias(`resp`)) e clima (temperatura média em °F (`temp`), umidade relativa média (`rhum`)) para a cidade de Chicago no período 1997-2000. Os dados são do National Morbidity, Mortality and Air Pollution Study. Importe a partir da função `read.csv` o arquivo `chicago.csv` para o R. Além das colunas descritas acima, o conjunto também contém as seguintes variáveis: `time`: sequência da observação; `year`: ano da observação; `month`: mês da observação; `day`: dia da observação; `season`: estação do ano.

- (a) Quantas mortes provocadas por doenças cardiovasculares ocorreram no período do estudo? Em qual estação ocorreram mais mortes provocadas por doenças cardiovasculares? E em qual estação ocorreram menos mortes provocadas por doenças cardiovasculares? (4 pontos)
- (b) Em qual ano ocorreram mais mortes relacionadas a doenças respiratórias? (2 pontos)
- (c) Qual a média de temperatura em cada uma das estações? E qual a média da umidade relativa em cada uma das estações? Apresente, numa mesma janela e na mesma escala, um boxplot para a variável `temp` para cada estação do ano. (4 pontos)
- (d) Apresente um gráfico de pontos para representar as variáveis `time` (eixo x) e `temp` (eixo y). Cada ponto do gráfico deve ser colorido de acordo com sua estação. (4 pontos)

Questão 4. Papagaio-do-mar é o nome comum dado às aves charadriiformes da família dos alcídeos, pertencentes ao gênero *Fratercula*. Existem três espécies de papagaios-do-mar conhecidas: *arctica*, *corniculata* e *cirrhata*. O conjunto `papagaio.txt` apresenta informações sobre o peso (em gramas), o tamanho (em centímetros), a envergadura da asa (em centímetros) e a espécie de 500 papagaios-do-mar.

- (a) Leia o arquivo `papagaio.txt`. Em seguida, comece a analisar os dados a partir das funções `head`, `tail`, `str`, `summary`.
- (b) Faça um gráfico que apresente a quantidade de aves de cada espécie.

- (c) Faça análises gráficas para compreender as características de cada espécie (exemplo: tem alguma espécie que se diferencia em relação à envergadura?). Tente usar diferentes tipos de gráficos. Escreva um pequeno texto comentando todos os resultados obtidos.
- (d) Construa um modelo de árvore de decisão usando estruturas condicionais para prever a espécie do papagaio a partir das outras variáveis. Em seguida, calcule a sua taxa de acerto para o modelo construído.
- (e) Construa agora um modelo knn para prever a espécie. Você pode usar a função pronto para o knn. Compare a taxa de acerto deste modelo com a árvore de decisão.