CE009 - Introcução à Estatística

Prova 1

1. (2 pontos) Uma pesquisa foi realizada na universidade para analisar a relação entre a experiência prévia dos alunos em programação e o seu desempenho na disciplina de Algoritmos e estruturas de dados. Foram coletados dados de 200 estudantes, categorizados da seguinte forma:

   - **Experiência prévia em programação**: "Sim" ou "Não".
   - **Desempenho na disciplina**: "Aprovado" ou "Reprovado".

   A tabela abaixo apresenta a distribuição dos estudantes de acordo com essas variáveis:

   Table 1: Distribuição dos estudantes por experiência prévia e desempenho

Experiência Prévia	**Aprovado**	**Reprovado**	**Total**
Sim	80	20	100
Não	50	50	100
Total	130	70	200

   Com base nos dados fornecidos, responda às seguintes questões:

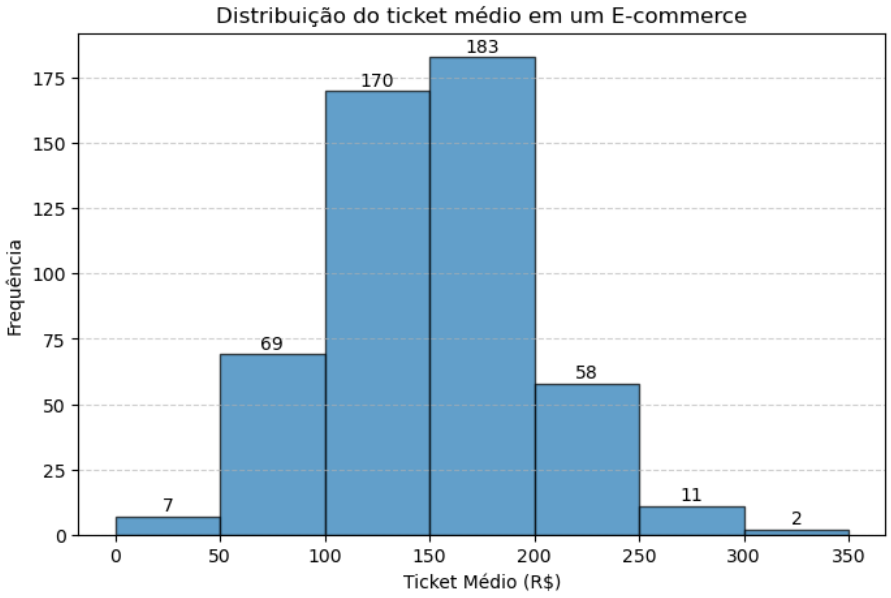
   (a) (1 ponto) Qual a proporção de alunos aprovados, entre aqueles que tinham experiência prévia em programação?

   A proporção é $\frac{80}{100} = 0{,}80$ ou 80%.

   (b) (1 ponto) Dos alunos reprovados, qual a proporção daqueles que não tinham experiência prévia em programação?

   A proporção é $\frac{50}{70} \approx 0{,}714$ ou 71,4%.

2. (3 pontos) Uma empresa de e-commerce deseja analisar o ticket médio das compras realizadas em sua plataforma. Para evitar distorções causadas por valores atípicos, a empresa decidiu eliminar todas as compras cujo valor esteja acima do quantil de 90% da distribuição dos tickets médios. O histograma abaixo apresenta a distribuição dessa variável nos últimos meses.



Com base no histograma apresentado, responda às seguintes questões:

(a) (1 ponto) Qual é o valor do quantil de 90% para o ticket médio das compras?

$$\frac{q(0.9) - 200}{21} = \frac{250 - 200}{58} \quad \Rightarrow \quad q(0.9) = 218,10$$

(b) (1 ponto) Qual é a média do ticket médio das compras antes da remoção dos valores acima do quantil de 90%?

$$\bar{x} = \frac{\sum(\text{frequência} \times \text{ponto médio})}{\sum \text{frequência}} \quad \Rightarrow \quad \bar{x} = 150,7$$

(c) (1 ponto) A equipe de marketing da empresa deseja criar campanhas promocionais focadas nos clientes que realizam compras de menor valor, visando aumentar sua

recorrência na plataforma. Para isso, é necessário entender a proporção de compras cujo ticket médio é inferior a 100. Com base no histograma apresentado, calcule essa porcentagem e justifique sua resposta.

Aproximadamente 15,2% das compras possuem ticket médio inferior a R$ 100, considerando as frequências dos dois primeiros intervalos do histograma (0-50 e 50-100), que somam 76 ocorrências de um total de 500.

3. (3 pontos) Uma empresa de segurança cibernética está desenvolvendo um novo algoritmo de detecção de anomalias em redes, utilizando aprendizado de máquina. Para validar sua eficácia, um estudo baseado em amostragem sistemática é conduzido, coletando um pacote a cada 130 segundos. A análise das observações ocorre até que um dos seguintes critérios seja atingido:

   - Se 40% ou mais dos pacotes analisados forem classificados incorretamente (falsos positivos ou falsos negativos), o modelo é considerado inadequado e a coleta de dados é interrompida;
   - Se após 150 pacotes analisados a taxa de erro estiver abaixo de 8%, o modelo é considerado eficiente e a coleta é encerrada;
   - Caso contrário, a análise continua até atingir 300 pacotes para uma avaliação mais robusta.

   (a) (1 ponto) Após a análise de 120 pacotes, foram registrados 42 erros (falsos positivos ou falsos negativos). O estudo deve continuar? Justifique sua resposta.

   Sim, o estudo deve continuar. A taxa de erro é de $\frac{42}{120} = 35\%$, abaixo do limite de 40%.

   (b) (1 ponto) Durante a fase de validação do modelo, um pesquisador sugere o uso de amostragem intencional, onde apenas pacotes que possuam características previamente identificadas como indicativas de ataques são analisados. Como essa estratégia pode comprometer a generalização dos resultados? Em quais cenários essa abordagem pode ser útil?

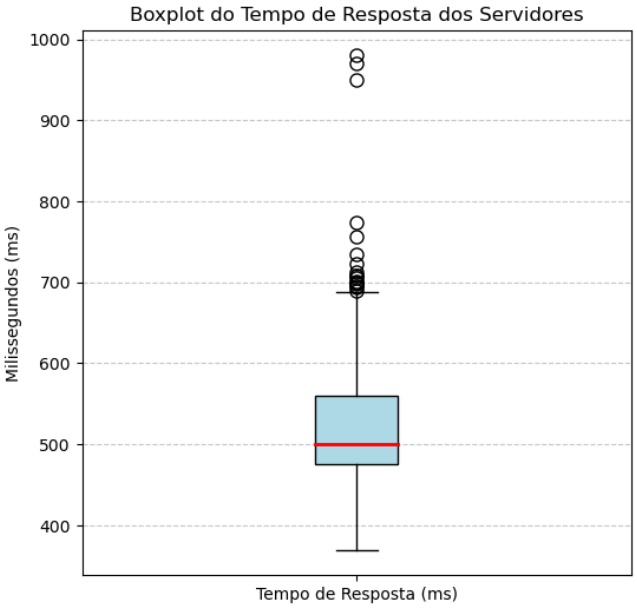
   A amostragem intencional compromete a generalização dos resultados porque os pacotes analisados não representam a diversidade total do tráfego real da rede. Ao

focar apenas em pacotes com características suspeitas, o modelo pode se tornar enviesado, tendo desempenho elevado apenas em casos semelhantes aos da amostra, mas com desempenho ruim em situações normais ou diferentes. Por outro lado, essa abordagem pode ser útil em fases específicas de testes, como avaliação da sensibilidade do modelo em detectar ataques conhecidos, criação de conjuntos desbalanceados propositalmente para verificar o comportamento do algoritmo em cenários de risco elevado, simulação de ataques dirigidos ou testes de estresse, quando o objetivo é testar a robustez do modelo diante de ameaças conhecidas.

(c) (1 ponto) Qual outro método de amostragem seria adequado para testar a eficácia do algoritmo? Justifique sua escolha considerando cenários como ataques em horários de pico ou padrões distintos de tráfego na rede.

Um método mais adequado seria a amostragem estratificada. Neste método, o tráfego da rede seria dividido em estratos, como horários de pico vs. horários de baixa atividade, tipos de protocolos (HTTP, FTP, DNS etc.), localizações geográficas ou sub-redes diferentes. A amostragem estratificada garante que o modelo seja exposto a diversas condições operacionais, o que aumenta a representatividade e robustez da validação, permitindo uma análise mais precisa do desempenho em cenários variados - inclusive em momentos críticos como ataques durante horários de pico.

4. (2 pontos) Uma empresa de hospedagem na nuvem está analisando o tempo de resposta (em milissegundos) de suas máquinas virtuais, sob diferentes cargas de trabalho. Para isso, um conjunto de medições foi observado e está representado no boxplot abaixo:



Com base no boxplot apresentado, responda:

(a) (1 ponto) Como o intervalo entre os quartis pode ser interpretado no contexto da análise de desempenho dos servidores?

O intervalo entre o primeiro e o segundo quartil é menor do que o intervalo entre o segundo e o terceiro quartil. Isso indica que os tempos de resposta estão mais concentrados no intervalo inferior da mediana, ou seja, há menor variabilidade entre os tempos de resposta mais rápidos. Por outro lado, o aumento do intervalo entre $q(0,5)$ e $q(0,75)$ sugere uma maior dispersão nos tempos de resposta mais altos dentro dos 50% centrais da amostra. Em termos de desempenho, isso revela que os servidores tendem a apresentar respostas mais consistentes quando operam em condições menos exigentes, mas há uma maior flutuação no desempenho sob cargas mais altas.

(b) (1 ponto) Suponha que um novo tempo de resposta de 650 ms foi registrado. Esse valor pode ser considerado um outlier, segundo o critério considerado no boxplot? Justifique.

O limite superior do boxplot está próximo de 700 ms. Como o novo tempo de resposta registrado, de 650 ms, está abaixo desse limite, ele não é considerado um outlier com base no critério do boxplot.