

Revisão para G1

Reforçar os conteúdos abordados até o momento para a prova.

Márcio Nicolau

2025-09-24

Table of contents

Introdução	1
Conteúdo Coberto	1
SEÇÃO I - QUESTÕES DE MÚLTIPLA ESCOLHA	2
Questão 1	2
Questão 2	2
SEÇÃO II - QUESTÕES DISSERTATIVAS	3
Questão 3	3
Questão 4	3
Questão 5	4
SEÇÃO III - QUESTÕES DE COMPLETAR CÓDIGO	5
Questão 6 - Python	5
Questão 7 - R	6
Referências Bibliográficas	7

List of Figures

Introdução

Esta aula de revisão tem como objetivo preparar você para a G1, cobrindo os conteúdos abordados nas aulas de 1 a 9. A revisão será composta por questões que avaliam sua compreensão dos conceitos fundamentais e sua capacidade de aplicá-los em contextos da Ciência da Computação.

Conteúdo Coberto

As aulas abrangeram os seguintes tópicos principais:

1. **Apresentação da Disciplina e Pensamento Estatístico** (Aula 1)
2. **Análise Combinatória: Permutação e Arranjo** (Aula 2)
3. **Introdução à Probabilidade e Eventos** (Aula 3)
4. **Reforço: Análise Combinatória e Probabilidade** (Aula 4 - incluindo Combinação)
5. **Probabilidade Condicional e Teorema de Bayes** (Aula 5)
6. **Introdução à Estatística Descritiva** (Aula 6)
7. **Análise Exploratória de Dados (AED)** (Aula 7)
8. **Amostragem e Tipos de Amostras** (Aula 9)

SEÇÃO I - QUESTÕES DE MÚLTIPLA ESCOLHA

Questão 1

Uma empresa de software está analisando os logs de acesso ao seu sistema. Em um dia típico, ocorrem 1000 tentativas de login, das quais 950 são bem-sucedidas e 50 falharam. Se 30 das tentativas que falharam são de ataques maliciosos, qual é a probabilidade de uma tentativa de login falhada ser um ataque malicioso?

Dica

Esta questão envolve **probabilidade condicional**. Identifique:

- Evento condicionante: “login falhado”
- Evento de interesse: “ataque malicioso”
- Use a fórmula: $P(A|B) = \frac{\text{número de casos favoráveis}}{\text{número total de casos do evento condicionante}}$

- a) 0.03
- b) 0.05
- c) 0.30
- d) 0.60
- e) 0.95

Questão 2

Um cientista de dados está realizando uma Análise Exploratória de Dados (AED) em um dataset de vendas de uma loja online. Ele observa que o histograma da variável “valor_pedido” apresenta uma cauda longa à direita. Com base nas propriedades das medidas de tendência central, qual das afirmações a seguir é CORRETA sobre esta distribuição?

Dica

Assimetria à direita (cauda longa à direita) significa:

- A maioria dos valores estão concentrados à esquerda
- Valores extremos “puxam” a média para a direita
- Lembre-se da relação: em distribuições assimétricas à direita $\rightarrow \text{Moda} < \text{Mediana} < \text{Média}$

- a) Média = Mediana = Moda, caracterizando uma distribuição simétrica

- b) Média < Mediana < Moda, caracterizando assimetria à esquerda
- c) Moda < Mediana < Média, caracterizando assimetria à direita
- d) A mediana é a medida menos adequada para representar o centro desta distribuição
- e) O desvio padrão será igual à amplitude neste tipo de distribuição

SEÇÃO II - QUESTÕES DISSERTATIVAS

Questão 3

Uma empresa de desenvolvimento de jogos online quer analisar o comportamento de seus usuários para melhorar a experiência de jogo. Eles têm acesso a dados de 10 milhões de jogadores ativos, distribuídos globalmente em diferentes fusos horários, plataformas (PC, mobile, console) e níveis de experiência (iniciante, intermediário, avançado).

Dica

Parte a: Pense nas limitações práticas de um censo:

- Aspectos financeiros e de recursos
- Questões de tempo e logística
- Problemas técnicos e de privacidade
- Compare vantagens da amostragem vs censo

Parte b: A empresa quer representatividade de subgrupos específicos. Considere:

- Qual método garante que todos os subgrupos sejam representados?
- Como dividir a população em grupos homogêneos?
- Que método evita sub-representação de grupos minoritários?

a) Explique por que seria impraticável realizar um censo de todos os jogadores e justifique a necessidade de usar amostragem.

b) Considerando que a empresa quer garantir que cada combinação de plataforma e nível de experiência esteja adequadamente representada na amostra, qual método de amostragem probabilística você recomendaria? Justifique sua escolha e explique como implementaria este método.

Questão 4

Um sistema de recomendação de uma plataforma de e-learning utiliza o Teorema de Bayes para sugerir cursos aos estudantes. O sistema observou os seguintes dados históricos:

- 60% dos estudantes são da área de Tecnologia, 25% de Negócios e 15% de Design
- Entre os estudantes de Tecnologia, 80% completam os cursos recomendados
- Entre os estudantes de Negócios, 70% completam os cursos recomendados
- Entre os estudantes de Design, 65% completam os cursos recomendados

 Dica

Parte a: Aplique o Teorema de Bayes: $P(A|B) = P(B|A) \times P(A) / P(B)$

- Identifique: A = “ser da área de Tecnologia”, B = “completou curso”
- $P(A)$ = probabilidade a priori (% estudantes de Tecnologia)
- $P(B|A)$ = verossimilhança (% que completam dado que são de Tecnologia)
- $P(B)$ = probabilidade total (use Lei da Probabilidade Total)

Parte b: Pense em como Bayes permite:

- Atualização contínua de probabilidades com novos dados
- Personalização baseada no histórico do usuário
- Adaptação de algoritmos de classificação

a) Se um estudante completou um curso recomendado, qual é a probabilidade de ele ser da área de Tecnologia? Mostre todos os cálculos.

b) Explique como este tipo de análise probabilística pode ser utilizada para melhorar algoritmos de machine learning em sistemas de recomendação.

Questão 5

Uma startup de análise de dados está desenvolvendo um algoritmo para detectar anomalias em transações financeiras. Durante a fase de desenvolvimento, eles coletaram dados de 50.000 transações e observaram as seguintes estatísticas descritivas para o valor das transações (em R\$):

- Média: R\$ 2.450,00
- Mediana: R\$ 180,00
- Desvio Padrão: R\$ 8.320,00
- Amplitude: R\$ 245.000,00

 Dica

Parte a: Compare as medidas descritivas:

- O que a diferença entre média e mediana indica sobre assimetria?
- O que um desvio padrão muito alto (maior que a média) sugere?
- Uma amplitude muito grande indica que tipos de problemas?

Parte b: Para detectar anomalias, considere técnicas que mostram:

- Distribuição geral dos dados (histograma)
- Valores extremos claramente (box plot)
- Padrões temporais ou relacionais (scatter plots)
- Cada técnica deve revelar aspectos específicos dos outliers

a) Com base nas medidas apresentadas, caracterize a distribuição dos valores das transações e identifique possíveis problemas nos dados. Justifique sua resposta.

b) Que técnicas de Análise Exploratória de Dados você recomendaria para visualizar e investigar melhor estes dados antes de desenvolver o algoritmo de detecção de anomalias? Cite pelo menos 3 técnicas específicas e explique o que cada uma revelaria.

SEÇÃO III - QUESTÕES DE COMPLETAR CÓDIGO

Questão 6 - Python

Complete o código Python abaixo para implementar um sistema de amostragem estratificada que selecione 100 usuários de um dataset de uma rede social, garantindo que a proporção de usuários por faixa etária na amostra seja a mesma da população:

Dica

Amostragem Estratificada em Python:

- Use `value_counts(normalize=True)` para calcular proporções
- Multiplique proporções pelo tamanho total da amostra e use `.round().astype(int)`
- Para filtrar DataFrame: `df[df['coluna'] == valor]`
- Use `.sample(n=tamanho, random_state=42)` para amostragem aleatória
- Para exibir proporções da amostra: `df['coluna'].value_counts(normalize=True)`

```
import pandas as pd
import numpy as np

# Dataset simulado de usuários de rede social
np.random.seed(42)
usuarios = pd.DataFrame({
    'user_id': range(1, 5001),
    'idade': np.random.randint(13, 70, 5000),
    'posts_mes': np.random.poisson(15, 5000),
    'faixa_etaria': pd.cut(np.random.randint(13, 70, 5000),
                           bins=[12, 18, 25, 35, 50, 70],
                           labels=['13-18', '19-25', '26-35', '36-50', '51-70'])
})

# COMPLETE O CÓDIGO ABAIXO:
tamanho_amostra_total = 100

# 1. Calcular as proporções de cada faixa etária na população
proporcoes_faixa = _____

# 2. Calcular o tamanho da amostra para cada faixa etária
tamanhos_por_faixa = _____

# 3. Realizar a amostragem estratificada
amostra_estratificada = pd.DataFrame()
for faixa, tamanho in tamanhos_por_faixa.items():
    if tamanho > 0: # Evitar erro quando tamanho = 0
        sub_amostra = _____
        amostra_estratificada = pd.concat([amostra_estratificada, sub_amostra])
```

```
# 4. Exibir os resultados
print("Distribuição original por faixa etária:")
print(proporcoes_faixa)
print("\nDistribuição na amostra estratificada:")
print(_____)
```

Questão 7 - R

Complete o código R abaixo para implementar uma análise de probabilidade condicional que calcule a probabilidade de diferentes tipos de bugs serem críticos em um sistema de software:

Dica

Probabilidade Condicional em R:

- Use `table(df$col1, df$col2)` para tabela de contingência
- Para filtrar: `nrow(df[df$coluna == "valor", ])` conta linhas que atendem condição
- Para filtrar com múltiplas condições: `df$col1 == "valor1" & df$col2 == "valor2"`
- $P(A|B)$ = número de casos (A e B) / número de casos (B)
- Use `sprintf("%.3f", valor)` para formatar números com 3 casas decimais

```
library(dplyr)

# Dados históricos de bugs reportados
bugs_data <- data.frame(
  bug_id = 1:1000,
  tipo = sample(c("UI", "Performance", "Security", "Logic"), 1000,
    replace = TRUE, prob = c(0.4, 0.25, 0.15, 0.2)),
  criticidade = sample(c("Baixa", "Media", "Alta", "Critica"), 1000,
    replace = TRUE, prob = c(0.3, 0.35, 0.25, 0.1))
)

# COMPLETE O CÓDIGO ABAIXO:

# 1. Criar tabela de contingência entre tipo e criticidade
tabela_contingencia <- _____

# 2. Calcular P(Crítico | Security) - probabilidade de um bug de Security ser Crítico
bugs_security <- _____
bugs_security_criticos <- _____
prob_critico_dado_security <- _____

# 3. Calcular P(Performance | Crítico) - probabilidade de um bug crítico ser de Performance
bugs_criticos <- _____
bugs_criticos_performance <- _____
```

```

prob_performance_dado_critico <- _____

# 4. Exibir resultados
cat("Tabela de Contingência:\n")
print(_____)

cat(sprintf("\nP(Crítico | Security) = %.3f\n", _____))
cat(sprintf("P(Performance | Crítico) = %.3f\n", _____))

# 5. Interpretar os resultados
cat("\nInterpretação:\n")
if(prob_critico_dado_security > 0.15) {
  cat("Bugs de Security têm alta probabilidade de serem críticos.\n")
} else {
  cat("Bugs de Security têm baixa probabilidade de serem críticos.\n")
}

```

Referências Bibliográficas

BUSSAB, Luiz O. de M.; MORETTIN, Pedro A. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.

MAGALHÃES, Marcos N.; LIMA, Antonio C. P. de. **Noções de Probabilidade e Estatística**. 7. ed. São Paulo: Edusp, 2013.

TRIOLA, Mario F. **Introdução à Estatística**. 12. ed. Rio de Janeiro: LTC, 2017.