

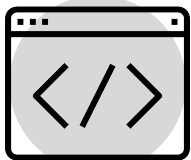
# Profissão: Cientista de Dados



# BOAS PRÁTICAS



# Probabilidade



- **Ilustração básica**
- **Conheça os modelos discretos**
- **Calcule probabilidade**
- **Função probabilidade**
- **Distribuição normal**
- **Distribuição Média**



# Ilustração básica

- Uma variável aleatória é um valor que resulta de um fenômeno aleatório. Compreender como as variáveis aleatórias funcionam pode ajudar a entender melhor os dados e os resultados dos modelos.
- Em um processo estocástico, os dados são gerados continuamente ao longo do tempo. Este é um conceito importante para entender ao trabalhar com dados que são coletados ao longo do tempo.



# Conheça os modelos discretos

- Tente aplicar os conceitos teóricos a situações do mundo real. Por exemplo, o modelo de Bernoulli pode ser usado para determinar se um cliente será adimplente ou inadimplente.
- Use simulações para desenvolver uma intuição sobre os conceitos que estão sendo aprendidos. Por exemplo, você pode simular o lançamento de uma moeda várias vezes para entender a distribuição binomial.



# Calcule probabilidade

- Em muitos casos, a probabilidade de uma variável não é afetada pelo valor da outra. Identificar essas situações pode simplificar a análise.
- Utilizar a distribuição condicional para identificar padrões de dependência: A distribuição condicional é a probabilidade de um evento ocorrer dado que outro evento já ocorreu. Este conceito é útil para identificar padrões de dependência e fazer previsões.



# Função probabilidade

- A função de probabilidade associa um valor a uma probabilidade, enquanto a função densidade de probabilidade associa uma densidade.
- A distribuição acumulada é uma ferramenta importante para variáveis contínuas, pois retorna a probabilidade de uma variável aleatória ser menor ou igual a um valor específico.



# Distribuição Normal

- A padronização é uma técnica útil que transforma uma variável aleatória que segue uma distribuição normal em uma distribuição normal padrão. Isso é feito subtraindo a média e dividindo pelo desvio padrão. A padronização é útil porque as tabelas de distribuição normal geralmente só fornecem valores para a distribuição normal padrão.
- Compreenda a distribuição normal: Como cientista de dados, é crucial entender a distribuição normal, pois muitos fenômenos na natureza seguem essa distribuição. Portanto, familiarize-se com suas propriedades, como a forma de sino e a densidade em torno da média.





# Distribuição Média – Teorema Central do Limite

- O Teorema Central do Limite é fundamental na análise de dados e estatística. Ele afirma que se temos uma amostra de variáveis independentes e identicamente distribuídas (i.i.d), a média dessa amostra terá uma distribuição aproximadamente normal. Esta aproximação melhora à medida que o tamanho da amostra aumenta.
- A distribuição T é usada quando não conhecemos a variância da distribuição. Ela tem um parâmetro a mais, o número de graus de liberdade, que é o número de observações na amostra menos um.
- Esteja ciente de que a distribuição T tem mais massa nas caudas da distribuição: Isso significa que é mais provável observar valores raros. Isso pode ser importante ao analisar os dados e ao fazer previsões.



# Bons estudos!

