Aula 11 – Estatística Descritiva para Ciência de Dados

Bem-vindos à Aula 11 do nosso Curso de Matemática Computacional! Hoje, embarcaremos em uma jornada fundamental para quem deseja desvendar os segredos escondidos nos dados: a **Estatística Descritiva**. Em um mundo onde somos bombardeados por informações a todo instante, desde notícias em redes sociais até relatórios financeiros complexos, a capacidade de resumir, organizar e interpretar esses dados se tornou uma habilidade indispensável. Não importa se você está buscando otimizar um algoritmo de inteligência artificial ou simplesmente entender melhor o comportamento de um mercado, a estatística descritiva é a sua bússola.

Imagine-se diante de uma montanha de números, planilhas e gráficos. Sem as ferramentas certas, essa montanha pode parecer intransponível. Nosso objetivo nesta aula é equipá-lo com essas ferramentas, transformando o caos em clareza. Ao final deste encontro, você será capaz de identificar os tipos de dados que encontrará, calcular e interpretar as principais medidas que resumem esses dados e, crucialmente, visualizá-los de forma eficaz para extrair *insights* valiosos.

A relevância deste conteúdo vai muito além da sala de aula. No cenário atual de 2025, com o avanço exponencial da Inteligência Artificial e do Machine Learning, a Estatística Descritiva é a base sólida sobre a qual se constroem modelos preditivos e sistemas inteligentes. Ela é o primeiro passo na **Ciência de Dados**, permitindo que você compreenda o "o quê" antes de tentar prever o "porquê" ou o "como". Prepare-se para transformar dados brutos em conhecimento acionável, uma habilidade que o diferenciará em qualquer campo profissional.

Fundamentos

Desvendando os Dados: Tipos de Variáveis

Antes de começarmos a calcular ou visualizar qualquer coisa, precisamos entender a natureza dos dados com os quais estamos trabalhando. Pense nos dados como ingredientes em uma receita: você não usaria açúcar no lugar do sal, certo? Da mesma mesma forma, diferentes tipos de dados exigem diferentes abordagens estatísticas. Ignorar essa distinção é como tentar assar um bolo sem saber se você tem farinha ou ovos.

A primeira grande divisão que fazemos é entre **variáveis qualitativas** e **variáveis quantitativas**. Essa classificação é a pedra angular para escolher as técnicas de análise corretas. Uma variável é simplesmente uma característica ou atributo que pode ser medido ou observado em um conjunto de dados. Por exemplo, se estamos analisando uma pesquisa de satisfação, a "idade" do respondente é uma variável, assim como sua "opinião sobre o produto".

> 💡 **Por que isso importa?** Compreender essa distinção é vital para a **Análise Exploratória de Dados (EDA)**, um passo crucial em qualquer projeto de Ciência de Dados. Se você está construindo um modelo de Machine Learning para prever a churn de clientes, saber se a "região" é uma variável qualitativa ou se o "tempo de uso do serviço" é quantitativa guiará suas escolhas de pré-processamento e modelagem.

Variáveis Qualitativas: Categorizando o Mundo

As **variáveis qualitativas**, também conhecidas como categóricas, descrevem características ou qualidades que não podem ser medidas numericamente de forma significativa. Elas representam categorias ou grupos. Imagine que você está organizando sua biblioteca: você pode categorizar os livros por gênero (ficção, não ficção, biografia), por autor ou por cor da capa. Essas são classificações, não quantidades.

Dentro das variáveis qualitativas, fazemos uma distinção importante: **nominais** e **ordinais**. Uma variável nominal é aquela em que as categorias não possuem uma ordem natural ou hierarquia. Por exemplo, a cor dos olhos (azul, verde, castanho) ou o estado civil (solteiro, casado, divorciado). Não faz sentido dizer que "azul" é "maior" ou "melhor" que "verde". Já as variáveis ordinais, como o próprio nome sugere, possuem uma ordem ou classificação intrínseca. Pense em um nível de escolaridade (ensino fundamental, médio, superior) ou uma escala de satisfação (muito insatisfeito, insatisfeito, neutro, satisfeito, muito satisfeito). Aqui, a ordem importa, mas a diferença entre as categorias não é necessariamente uniforme ou mensurável.

Variáveis Nominais

- Sem ordem natural
- Exemplos: cor dos olhos, estado civil, tipo sanguíneo
- Técnica: *one-hot encoding*

Variáveis Ordinais

- Com ordem hierárquica
- Exemplos: nível de escolaridade, satisfação, ranking
- Técnica: *label encoding*

No contexto da Ciência de Dados, o tratamento de variáveis qualitativas é um desafio comum. Algoritmos de Machine Learning geralmente trabalham com números, então precisamos de técnicas como *one-hot encoding* ou *label encoding* para transformar essas categorias em um formato que os modelos possam entender. A escolha da técnica depende se a variável é nominal ou ordinal, pois preservar a ordem em variáveis ordinais pode ser crucial para o desempenho do modelo.

Variáveis Quantitativas: Medindo e Contando

Ao contrário das qualitativas, as **variáveis quantitativas** são aquelas que podem ser medidas ou contadas, expressas numericamente. Elas nos dão uma ideia de "quanto" ou "quantos". Se você está medindo a altura de uma pessoa, o número de carros em um estacionamento ou a temperatura de uma sala, você está lidando com variáveis quantitativas. Elas são a base para cálculos matemáticos mais complexos e nos permitem realizar operações como soma, média e desvio padrão.

Variáveis Discretas

Resultam de **contagem** e assumem valores inteiros, geralmente finitos ou contáveis.

- Número de filhos (0, 1, 2, 3...)
- Número de defeitos em produtos
- Quantidade de vendas

💡 *Não faz sentido ter 2,5 filhos*

Variáveis Contínuas

Podem assumir **qualquer valor** dentro de um intervalo, incluindo frações e decimais.

- Altura (1,75m)
- Peso (72,3 kg)
- Tempo (2,5 horas)
- Temperatura (23,8°C)

💡 *Resultam de medição*

A distinção entre discreta e contínua é importante para a escolha de gráficos e modelos estatísticos. Por exemplo, para uma variável discreta, um gráfico de barras pode ser mais adequado, enquanto para uma contínua, um histograma ou um gráfico de densidade faria mais sentido. Em Machine Learning, a forma como lidamos com valores ausentes ou *outliers* pode variar dependendo se a variável é discreta ou contínua, impactando diretamente a qualidade do seu modelo preditivo.

Capítulo 2

Medidas de Tendência Central: O Coração dos Dados

Agora que sabemos classificar nossos dados, o próximo passo é começar a resumi-los. Imagine que você tem uma lista de 1000 salários de funcionários de uma empresa. Ler cada um deles seria exaustivo e pouco informativo. Precisamos de uma forma de representar o "típico" ou o "centro" desses dados. É aqui que entram as **medidas de tendência central**. Elas nos dão um único valor que tenta descrever o ponto central de um conjunto de dados, como um farol que guia os navios em uma noite escura.

As Três Perspectivas do Centro

As três medidas de tendência central mais comuns são a **média**, a **mediana** e a **moda**. Cada uma delas oferece uma perspectiva ligeiramente diferente sobre o centro dos dados e é crucial saber quando usar cada uma. A escolha da medida certa pode mudar completamente a interpretação de um conjunto de informações. Por exemplo, ao analisar o preço de imóveis em uma região, a média pode ser distorcida por algumas poucas mansões caríssimas, enquanto a mediana pode oferecer uma visão mais realista do preço "típico".

01

Média Aritmética

Soma de todos os valores dividida pelo número total. O ponto de equilíbrio dos dados.

02

Mediana

Valor central quando os dados estão ordenados. Resistente a outliers.

03

Moda

Valor que aparece com maior frequência. Útil para dados qualitativos.

Compreender essas medidas é fundamental para qualquer análise de dados, seja você um cientista de dados explorando um novo *dataset* ou um analista de negócios tentando entender o desempenho de vendas. Elas são os primeiros números que você calculará para ter uma ideia geral do que os dados estão tentando lhe dizer, antes mesmo de mergulhar em análises mais complexas ou modelos de IA.

A Média: O Equilíbrio Aritmético

A **média aritmética**, ou simplesmente média, é provavelmente a medida de tendência central mais conhecida e utilizada. Ela é calculada somando-se todos os valores em um conjunto de dados e dividindo-se pelo número total de valores. Pense nela como o ponto de equilíbrio de uma balança: se você colocasse todos os seus dados em uma linha e tentasse encontrar o ponto onde a linha se equilibraria, esse ponto seria a média.

> **Exemplo prático:** Se as notas de um aluno em cinco provas foram 7, 8, 6, 9 e 10, a média seria (7+8+6+9+10) / 5 = 40 / 5 = **8**

A média é intuitiva e fácil de calcular, o que a torna popular. No entanto, ela tem uma vulnerabilidade significativa: é altamente sensível a **valores extremos** (outliers). Um único valor muito alto ou muito baixo pode puxar a média para cima ou para baixo, distorcendo a representação do centro dos dados.

> "No contexto de Ciência de Dados, a média é frequentemente usada para resumir características numéricas, como a idade média dos clientes ou o tempo médio de resposta de um servidor. Contudo, é preciso cautela. Se você está analisando a renda per capita de uma cidade e há alguns bilionários, a média pode sugerir uma renda muito mais alta do que a maioria da população realmente possui."

Mediana e Moda: Outras Perspectivas do Centro

Nem sempre a média é a melhor representação do centro dos dados, especialmente quando temos valores extremos. É aí que a **mediana** e a **moda** entram em cena, oferecendo perspectivas complementares.

Mediana

A **mediana** é o valor central de um conjunto de dados quando eles estão ordenados em ordem crescente ou decrescente. Se você tem um número ímpar de dados, a mediana é o valor do meio. Se for um número par, é a média dos dois valores centrais. Ela é como o "meio da fila": metade dos dados está abaixo dela e metade está acima.

✔ **Vantagem:** Resistente a outliers

Exemplo: Em salários onde a maioria ganha R$ 3.000 e alguns diretores ganham R$ 50.000, a mediana ainda estará próxima de R$ 3.000.

Moda

A **moda** é o valor que aparece com maior frequência em um conjunto de dados. Ela é particularmente útil para dados qualitativos ou discretos, onde a média e a mediana podem não fazer sentido ou serem menos informativas.

✔ **Vantagem:** Única medida aplicável a dados qualitativos

Exemplo: A moda da cor de carros vendidos em uma concessionária pode ser "prata", indicando a preferência dos consumidores.

Em Ciência de Dados, a mediana é frequentemente preferida para resumir distribuições assimétricas, como a renda ou o tempo de vida de um produto. A moda, por sua vez, é essencial para entender as categorias mais populares ou os valores mais comuns em variáveis discretas. Juntas, média, mediana e moda nos dão um panorama muito mais rico do "centro" dos nossos dados, permitindo decisões mais informadas e modelos de Machine Learning mais robustos.

Comparando as Medidas de Tendência Central

Entender as diferenças entre média, mediana e moda é crucial para escolher a ferramenta certa para cada situação. Cada uma delas tem seu momento de brilhar e suas limitações. A média é excelente para dados simétricos e sem outliers, oferecendo uma representação precisa do valor "típico". No entanto, sua sensibilidade a valores extremos pode levar a conclusões enganosas em distribuições assimétricas.

A mediana, por outro lado, é a heroína dos dados assimétricos e com outliers. Por focar no valor central após a ordenação, ela nos dá uma visão mais robusta do que é "comum" no conjunto de dados, sem ser puxada por valores atípicos. Já a moda é a única medida que pode ser aplicada a dados qualitativos, revelando a categoria mais frequente. Ela também é útil para identificar picos em distribuições, mesmo em dados quantitativos.

> 💡 **Dica profissional:** Ao analisar um *dataset* para um projeto de IA, você não deve escolher apenas uma. A prática recomendada é calcular as três. Se média, mediana e moda estiverem próximas, seus dados provavelmente têm uma distribuição simétrica. Se houver grandes diferenças, isso é um sinal de que a distribuição é assimétrica ou que há outliers significativos, e você precisará investigar mais a fundo.

Conceito	Âmbito/Aplicação	Base/Origem	Exemplo
Média	Dados quantitativos, distribuições simétricas	Soma de todos os valores / número de valores	Salário médio de uma equipe sem grandes disparidades
Mediana	Dados quantitativos, distribuições assimétricas ou com outliers	Valor central após ordenação	Preço mediano de imóveis em uma cidade
Moda	Dados qualitativos ou quantitativos discretos	Valor mais frequente	Cor de carro mais vendida

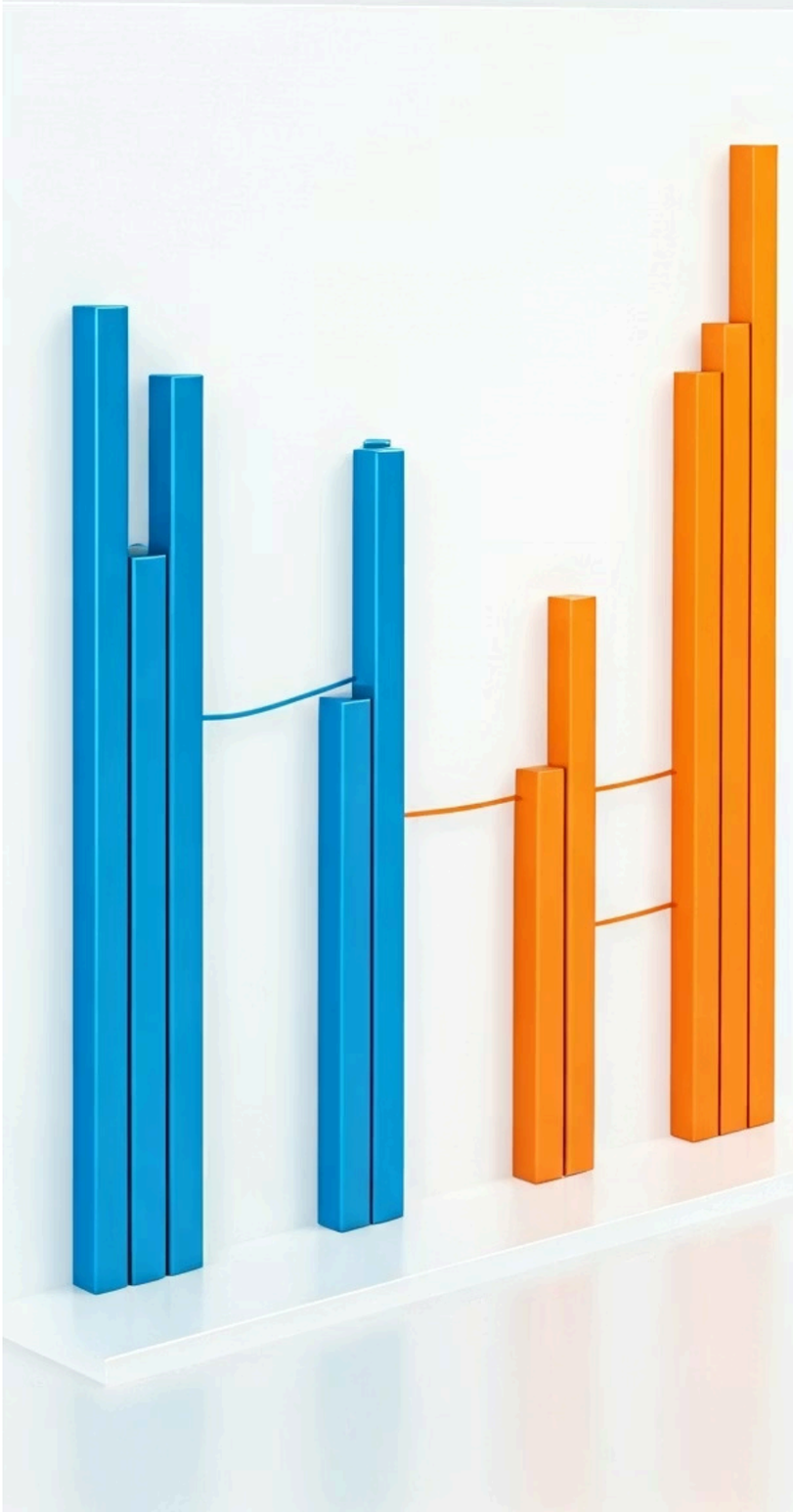
Capítulo 3

Medidas de Dispersão: Entendendo a Variabilidade

Conhecer o centro dos dados é um excelente começo, mas não é o suficiente. Imagine que duas turmas de alunos têm a mesma nota média de 7. Isso significa que as turmas são idênticas em desempenho? Não necessariamente! Uma turma pode ter todos os alunos com notas entre 6 e 8, enquanto a outra pode ter alunos com notas de 0 a 10. Ambas têm média 7, mas a segunda turma mostra uma variação muito maior. É aqui que entram as **medidas de dispersão**.

As medidas de dispersão nos dizem o quão espalhados ou concentrados os dados estão em torno do seu centro. Elas quantificam a variabilidade, a amplitude das diferenças entre os valores. Sem essas medidas, teríamos uma visão incompleta e potencialmente enganosa dos nossos dados. É como saber a temperatura média de uma cidade, mas não saber se ela varia entre 0°C e 40°C ou se permanece estável em torno de 20°C. A variabilidade é um fator crítico em muitas áreas, desde o controle de qualidade na indústria até a análise de risco em investimentos.

Para um cientista de dados, entender a dispersão é tão importante quanto entender a tendência central. Um modelo de Machine Learning que prevê um valor com alta variabilidade pode ser menos confiável do que um que prevê com baixa variabilidade. A dispersão nos ajuda a avaliar a consistência dos dados e a identificar potenciais problemas ou oportunidades.



Amplitude: O Alcance dos Dados

A **amplitude** é a medida de dispersão mais simples de calcular e entender. Ela nos dá uma ideia rápida do "alcance" total dos nossos dados. Para calculá-la, basta subtrair o menor valor do maior valor em um conjunto de dados.

47

Anos de amplitude

Se as idades variam de 18 a 65 anos

Cálculo: 65 - 18 = 47

Apesar de sua simplicidade, a amplitude é uma medida útil para ter uma primeira impressão da variabilidade. Ela nos diz o quão "longe" os dados se estendem. No entanto, sua principal desvantagem é que ela é extremamente sensível a outliers. Um único valor atípico, seja ele muito alto ou muito baixo, pode inflar drasticamente a amplitude, dando uma falsa impressão de grande variabilidade quando a maioria dos dados está, na verdade, bem concentrada.

⚠️ Limitações da Amplitude

- Extremamente sensível a outliers
- Ignora a distribuição dos dados intermediários
- Raramente usada como única medida

Uso recomendado: Como ponto de partida para verificar rapidamente o intervalo de uma variável antes de aplicar transformações.

Por essa razão, a amplitude raramente é usada como a única medida de dispersão em análises mais aprofundadas. Ela serve como um ponto de partida, um "primeiro olhar" sobre a extensão dos dados. Em Ciência de Dados, você pode usá-la para verificar rapidamente o intervalo de uma variável antes de aplicar transformações ou identificar potenciais erros de entrada de dados que resultaram em valores absurdamente altos ou baixos.

Variância e Desvio Padrão: A Essência da Variabilidade

Se a amplitude nos dá o alcance total, a **variância** e o **desvio padrão** nos oferecem uma visão muito mais sofisticada e robusta de como os dados se espalham em torno da média. A **variância** mede a média dos quadrados das diferenças de cada valor em relação à média do conjunto de dados. Parece um pouco complexo, mas a ideia é simples: quanto maior a variância, mais os dados estão dispersos da média.

Variância

Média dos quadrados das diferenças em relação à média

Problema: unidades quadradas

Raiz Quadrada

Aplicamos a raiz quadrada da variância

Solução: retorna às unidades originais

Desvio Padrão

Medida intuitiva de dispersão nas unidades originais

Resultado: fácil interpretação

> **Analogia prática:** Imagine que você está medindo a precisão de um atirador. A média dos tiros pode ser o centro do alvo, mas o desvio padrão dirá o quão "espalhados" os tiros estão. Um desvio padrão pequeno indica que os tiros estão agrupados perto do centro (alta precisão), enquanto um desvio padrão grande significa que os tiros estão espalhados (baixa precisão).

O problema da variância é que ela está em unidades quadradas (por exemplo, se os dados são em metros, a variância é em metros quadrados), o que dificulta a interpretação direta. Para resolver isso, calculamos o **desvio padrão**, que é simplesmente a raiz quadrada da variância. O desvio padrão retorna a medida de dispersão para as unidades originais dos dados, tornando-o muito mais intuitivo. Ele nos diz, em média, o quão longe cada ponto de dado está da média.

Em Machine Learning, o desvio padrão é crucial para a normalização de dados, avaliação de modelos e compreensão da incerteza nas previsões.

Capítulo 4

Visualização de Dados: Pintando a História dos Números

Até agora, falamos sobre números e cálculos. Mas, como diz o ditado, "uma imagem vale mais que mil palavras". No mundo da Ciência de Dados, uma boa visualização pode valer mais que mil tabelas. A **Visualização de Dados** é a arte e a ciência de representar informações numericamente de forma gráfica, tornando padrões, tendências e *insights* complexos imediatamente compreensíveis. É a ponte entre os números brutos e a intuição humana.

Ferramentas Visuais Essenciais

Quando você está diante de um novo *dataset*, a primeira coisa que um cientista de dados faz, após a limpeza inicial, é visualizá-lo. Gráficos nos ajudam a identificar outliers, a entender a distribuição dos dados, a detectar correlações e a comunicar nossas descobertas de forma eficaz para públicos não técnicos. Sem visualização, estaríamos navegando às cegas, perdendo a oportunidade de descobrir histórias fascinantes que os dados têm para contar.

Histograma

Ferramenta visual poderosa para entender a **distribuição** de uma variável quantitativa. As barras representam intervalos de valores (bins), e a altura indica a frequência de dados naquele intervalo.

Uso: Revelar a forma da distribuição (simétrica, assimétrica, bimodal, uniforme)

Box Plot

Excelente para visualizar a distribuição de forma concisa e **comparar grupos**. Exibe mediana, quartis, valores mínimo/máximo e identifica outliers potenciais.

Uso: Comparar desempenho entre regiões, departamentos ou períodos

Gráfico de Dispersão

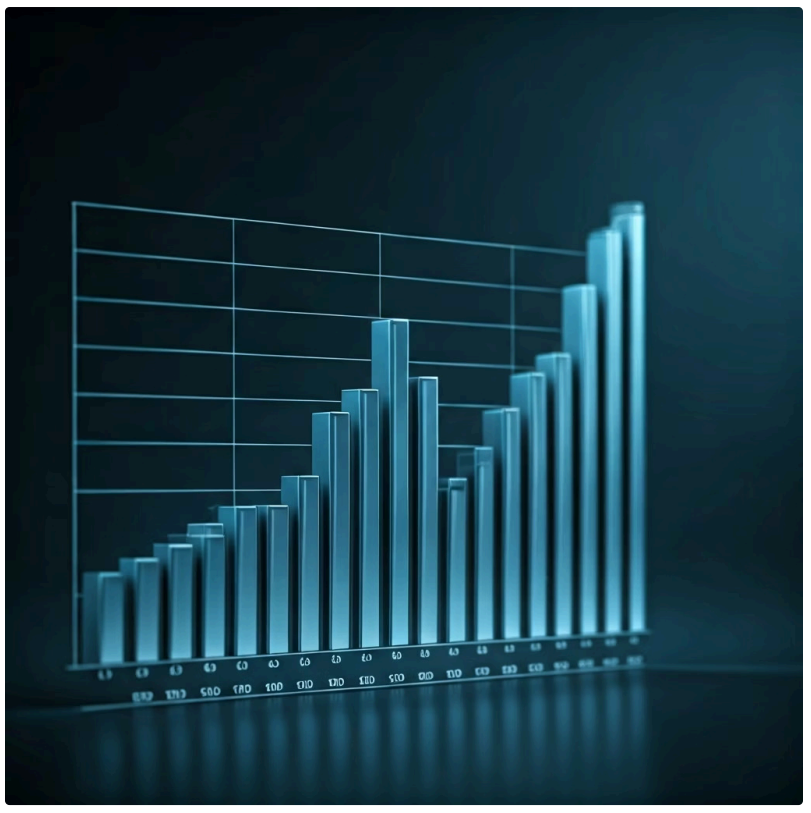
Ferramenta ideal para explorar a **relação entre duas variáveis** quantitativas. Cada ponto representa um par de valores (X, Y), revelando correlações.

Uso: Identificar correlações positivas, negativas ou ausência de correlação

💡 **Dica de ouro:** Nesta seção, exploramos alguns dos gráficos mais poderosos e comuns para a Estatística Descritiva. Cada um deles serve a um propósito específico e, quando usados corretamente, podem revelar camadas profundas de informação que seriam invisíveis em uma tabela de números.

Histograma: A Distribuição em Barras

O **histograma** é uma ferramenta visual poderosa para entender a distribuição de uma variável quantitativa. Ele se parece com um gráfico de barras, mas com uma diferença crucial: as barras representam intervalos de valores (chamados "bins" ou classes), e a altura de cada barra indica a frequência (ou contagem) de dados que caem naquele intervalo. Pense em um histograma como uma forma de "agrupar" seus dados e ver onde eles se concentram.



Por exemplo, se você está analisando as alturas de 1000 pessoas, um histograma pode mostrar que a maioria das pessoas tem entre 1,60m e 1,70m, com menos pessoas nas extremidades (muito baixas ou muito altas). Ele revela a forma da distribuição dos dados: se é simétrica, assimétrica (enviesada para a esquerda ou direita), bimodal (com dois picos) ou uniforme. Essas informações são vitais para escolher os modelos estatísticos e de Machine Learning apropriados.

- **Identifica distribuição normal**

  Verifica se os dados seguem uma curva em forma de sino

- **Detecta outliers visíveis**

  Valores extremos aparecem como barras isoladas

- **Sugere transformações**

  Indica se a variável precisa de transformação logarítmica

Em Ciência de Dados, o histograma é um dos primeiros gráficos a serem gerados para qualquer variável numérica. Ele ajuda a identificar rapidamente se os dados seguem uma distribuição normal (curva em forma de sino), se há *outliers* visíveis, ou se a variável precisa de alguma transformação (como logarítmica) antes de ser usada em um algoritmo. É uma janela para a "personalidade" dos seus dados.

Box Plots e Gráficos de Dispersão: Detalhes e Relações

Além dos histogramas, o **box plot** (ou diagrama de caixa) e o **gráfico de dispersão** são ferramentas indispensáveis na caixa de ferramentas de qualquer analista de dados.

Box Plot

Excelente para visualizar a distribuição de uma variável quantitativa de forma concisa, especialmente útil para **comparar distribuições entre diferentes grupos**. Ele exibe a mediana, os quartis (25% e 75%), e os valores mínimo e máximo, além de identificar *outliers* potenciais.

- Resume 5 números-chave em uma imagem
- Facilita comparação entre grupos
- Identifica outliers visualmente

Gráfico de Dispersão

A ferramenta ideal para explorar a **relação entre duas variáveis quantitativas**. Cada ponto no gráfico representa um par de valores (X, Y). Ao observar o padrão dos pontos, podemos identificar correlações.

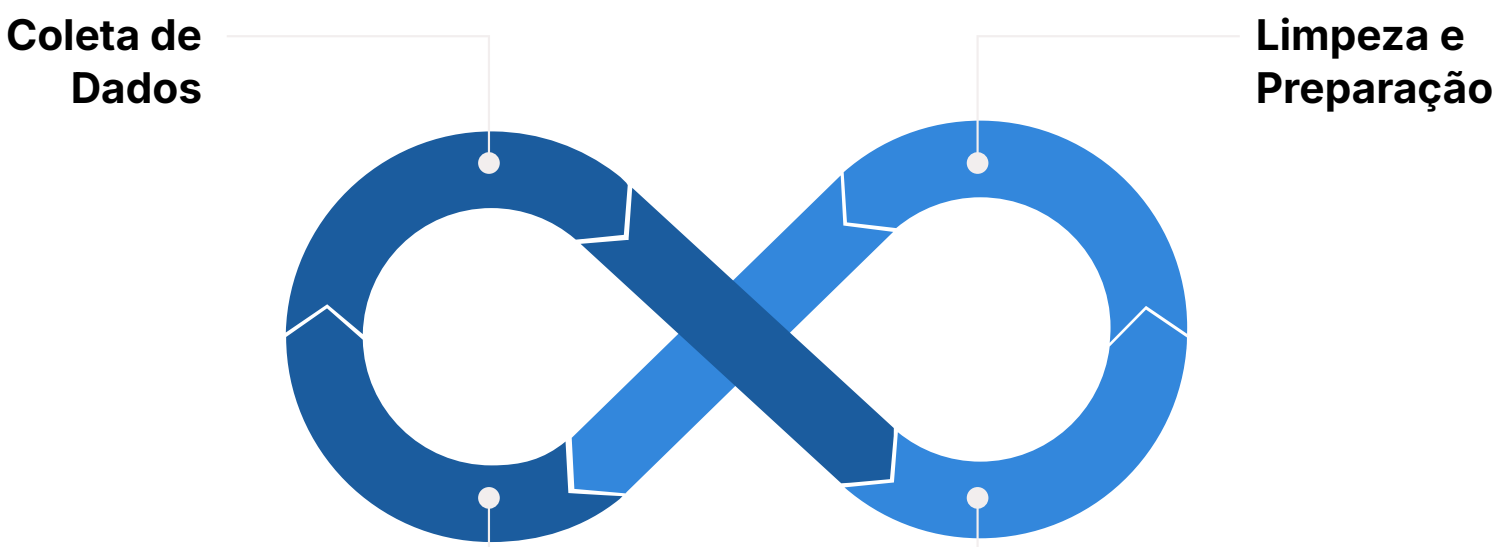
- Correlação positiva (X↑ → Y↑)
- Correlação negativa (X↑ → Y↓)
- Nenhuma correlação aparente

> **Exemplo prático:** Imagine que você está comparando o desempenho de vendas de três regiões diferentes. Um box plot pode rapidamente mostrar qual região tem a mediana de vendas mais alta, qual tem a maior variabilidade e se há vendas excepcionalmente altas ou baixas em alguma delas.

Um gráfico de dispersão pode mostrar a relação entre horas de estudo e notas em uma prova. Em Machine Learning, gráficos de dispersão são cruciais para identificar *features* correlacionadas e entender a natureza das relações que seu modelo tentará aprender.

A Importância da Análise Exploratória de Dados (EDA)

Chegamos a um ponto crucial que amarra todos os conceitos que vimos até agora: a **Análise Exploratória de Dados (EDA)**. A EDA não é apenas uma etapa em um projeto de Ciência de Dados; é uma mentalidade, uma filosofia. É o processo de investigar *datasets* para resumir suas principais características, muitas vezes usando métodos visuais. Pense na EDA como um detetive que examina a cena do crime antes de tentar resolver o caso. Ele procura pistas, padrões, anomalias, e tenta entender o contexto antes de formular uma teoria.



Identificar outliers e anomalias

Detectar valores atípicos que podem distorcer análises

Entender distribuições

Compreender como as variáveis se comportam

Detectar relações entre variáveis

Identificar correlações e dependências

Descobrir insights valiosos

Extrair conhecimento acionável dos dados

Sem a EDA, você estaria alimentando dados "cegamente" em algoritmos de Machine Learning, sem entender suas peculiaridades, seus problemas ou suas oportunidades. Isso pode levar a modelos ruins, previsões imprecisas e decisões de negócios equivocadas. A EDA é o que nos permite fazer perguntas inteligentes aos dados e obter respostas significativas.

No cenário atual de 2025, onde a complexidade dos dados só aumenta, a EDA se torna ainda mais vital. Ela é a base para a engenharia de *features*, a seleção de modelos e a interpretação dos resultados. Um bom cientista de dados passa uma parte significativa do seu tempo fazendo EDA, pois é nesse estágio que a verdadeira compreensão dos dados acontece, pavimentando o caminho para a construção de sistemas de IA e Machine Learning eficazes e confiáveis.

Em Prática: De Dados a Decisões

Nesta aula, desvendamos os fundamentos da Estatística Descritiva, uma área essencial para qualquer pessoa que lida com dados. Começamos classificando os dados em **variáveis qualitativas e quantitativas**, entendendo que a natureza do dado define como ele deve ser tratado. Em seguida, exploramos as **medidas de tendência central** – média, mediana e moda – que nos ajudam a encontrar o "coração" dos nossos dados, cada uma com suas particularidades e aplicações.

Aprofundamos nosso conhecimento com as **medidas de dispersão** – amplitude, variância e desvio padrão – que nos revelam o quão espalhados os dados estão, complementando a visão do centro. Finalmente, mergulhamos na **visualização de dados**, utilizando histogramas, box plots e gráficos de dispersão para transformar números em histórias visuais, facilitando a identificação de padrões e *insights*. Tudo isso culmina na compreensão da **Análise Exploratória de Dados (EDA)**, a abordagem sistemática para extrair conhecimento dos dados antes de qualquer modelagem complexa.

01

Classifique suas variáveis

Sempre comece um projeto de dados entendendo o tipo de cada variável

02

Calcule as três medidas centrais

Média, mediana e moda para visão completa do centro

03

Avalie a dispersão

Use o desvio padrão para entender variabilidade e consistência

04

Visualize seus dados

Histogramas, box plots e gráficos de dispersão revelam padrões ocultos

05

Pratique EDA constantemente

A EDA é sua bússola para construir modelos de IA mais robustos

Autoavaliação

1. Qual das seguintes afirmações sobre variáveis é **correta**?
   a) Variáveis qualitativas sempre podem ser ordenadas.
   b) A idade de uma pessoa é um exemplo de variável qualitativa nominal.
   c) O número de carros em um estacionamento é uma variável quantitativa discreta.
   d) Variáveis contínuas só podem assumir valores inteiros.
2. Um conjunto de dados possui os seguintes valores: 10, 12, 15, 15, 18, 20, 22. Qual é a mediana deste conjunto?
   a) 15
   b) 15.5
   c) 16
   d) 17
3. Em uma análise de salários onde a maioria dos funcionários ganha um valor baixo, mas alguns diretores ganham salários muito altos, qual medida de tendência central seria a mais adequada para representar o salário "típico" da maioria dos funcionários?
   a) Média
   b) Mediana
   c) Moda
   d) Desvio Padrão
4. Qual gráfico é mais adequado para visualizar a relação entre duas variáveis quantitativas e identificar possíveis correlações?
   a) Histograma
   b) Box Plot
   c) Gráfico de Barras
   d) Gráfico de Dispersão

Gabarito: 1. c) | 2. a) | 3. b) | 4. d)

Questão Discursiva: Explique a importância da Análise Exploratória de Dados (EDA) no contexto da construção de modelos de Machine Learning, citando como as medidas de tendência central, dispersão e visualização contribuem para esse processo.

Próxima Aula

Na Aula 12, daremos um passo adiante e exploraremos as **Distribuições de Probabilidade**. Entender como os dados se distribuem é fundamental para fazer inferências e previsões mais precisas, conectando diretamente com os conceitos de variabilidade que vimos hoje.

Recursos Adicionais

- **Livro:** "Estatística Essencial para Ciência de Dados" (para aprofundamento teórico e prático).
- **Plataforma:** Kaggle (para praticar EDA em *datasets* reais).
- **Artigo:** "The Importance of Exploratory Data Analysis" (para insights sobre a prática profissional).

NOTA IMPORTANTE: As informações regulatórias/legais/técnicas desta aula estão atualizadas até 2025. Consulte sempre fontes oficiais para verificar alterações.