Revisão e Notas sobre Estatística I

Pedro Rupf Pereira Viana

10 de dezembro de 2025

1 Introdução e Objetivos

Trata-se de uma revisão teórica e expositiva sobre Conceitos gerais, Análise Exploratória de Dados (EDA), conceitos de Amostragem, Probabilidade e os diferentes tipos de Distribuições, bem como suas aplicações na área de Análise e Ciência de Dados.

2 Conceitos Gerais

A palavra **Estatística** têm sua origem na palavra latina *STATUS* (Estado). Há indícios de que há 3000 anos A.C. já eram realizados censos na Babilônia, China e Egito. Também podemos ver no livro de Números, do Velho Testamento, uma instrução do Altíssimo dada à Moisés, para que fosse realizado um levantamento dos homens de Israel que estivessem aptos para guerrear (Nm. 1:1-3). Usualmente, estas informações eram utilizadas para a taxação de impostos ou para o alistamento militar.

A estatística envolve técnicas para coletar, organizar, descrever, analisar e interpretar dados, quer sejam provenientes de experimentos, ou vindos de estudos observacionais. A análise estatística de dados geralmente tem por objetivo a tomada de decisões, resoluções de problemas ou produção de conhecimento. Mas novos conhecimentos normalmente geram problemas de pesquisas, resultando em um processo iterativo.

Normalmente o termo estatística esta associado a números, tabelas, gráficos, mas a importância da estatística fica melhor representada por dois pontos comuns em nosso dia a dia:

- Dados
- Variabilidade

Podemos dividir a estatística em três áreas:

- Estatística Descritiva
- Probabilística
- Estatística Inferencial

A estatística descritiva é a etapa inicial da análise utilizada para descrever e resumir os dados. A disponibilidade de uma grande quantidade de dados e de métodos computacionais muito eficientes revigorou está área da estatística.

Inferência Estatística é, em suma, o estudo de técnicas fundamentada na teoria das probabilidades, que possibilitam a extrapolação a um grande conjunto de dados, das informações e conclusões obtidas a partir da amostra.

2.1 Estatística Descritiva

A estatística descritiva é um ramo fundamental da estatística que se concentra em resumir, organizar e apresentar dados de forma clara e concisa, sem fazer inferências sobre populações maiores. Ela permite que pesquisadores, analistas e tomadores de decisão compreendam as características principais de um conjunto de dados, facilitando a identificação de padrões, tendências e anomalias. De acordo com definições padrão, a estatística descritiva utiliza coeficientes breves para sintetizar um conjunto de dados, representando

uma população inteira ou uma amostra dela. Essa abordagem é essencial em diversas áreas, como economia, saúde, educação e ciências sociais, onde o volume de dados pode ser esmagador sem ferramentas de síntese.

Historicamente, a estatística descritiva evoluiu a partir de métodos simples de contagem e agrupamento, ganhando sofisticação com o avanço da computação. Hoje, ela é distinguida da estatística inferencial, que usa amostras para fazer previsões sobre populações. Em resumo, enquanto a inferencial projeta para o desconhecido, a descritiva descreve o conhecido. Para ilustrar, imagine um conjunto de dados sobre as alturas de alunos em uma turma: a descritiva calcularia médias e variações, mas não generalizaria para todos os alunos do país.

2.1.1 Tipos de dados e variáveis em Estatística Descritiva

Antes de mergulhar nas medidas, é crucial entender os tipos de dados manipulados na estatística descritiva. Os dados podem ser classificados como qualitativos (categóricos) ou quantitativos (numéricos). Os qualitativos incluem categorias nominais (sem ordem, como cores) e ordinais (com ordem, como níveis de satisfação: baixo, médio, alto). Já os quantitativos dividem-se em discretos (valores inteiros, como número de filhos) e contínuos (valores reais, como peso).

Essa classificação influencia as técnicas descritivas aplicadas. Por exemplo, para dados qualitativos, usamos frequências e percentuais, enquanto para quantitativos, aplicamos medidas numéricas. Além disso, as variáveis podem ser independentes (causadoras) ou dependentes (afetadas), embora na descritiva o foco seja mais na descrição do que na causalidade. Uma análise descritiva eficaz começa com a identificação correta desses tipos, evitando erros na interpretação.

2.1.2 Medidas de Tendência Central

As medidas de tendência central são o cerne da estatística descritiva, resumindo um conjunto de dados em um único valor representativo. As principais são a **média**, a **mediana** e a **moda**.

- **Média Aritmética:** Calculada como a soma de todos os valores, divididos pelo número de observações:

$$\bar{x} = \frac{\sum x_i}{n} \tag{2.1}$$

  Ela é sensível a valores extremos (chamados de *outliers*), o que a torna útil para dados simétricos, mas menos robusta em distribuições assimétricas. Por exemplo, em um conjunto de salários 1000, 1200, 1300, 50000, a média é alta devido ao outlier.

- **Mediana:** O valor central quando os dados são ordenados. Para um número ímpar de observações, é o meio; para par, a média dos dois centrais. Ela é resistente a outliers, ideal para dados enviesados, como rendas familiares.

- **Moda:** O valor mais frequente. Pode ser unimodal (uma moda), bimodal (duas) ou multimodal. Útil para dados categóricos, como a cor mais comum em uma pesquisa.

Essas medidas fornecem insights sobre o "centro"dos dados, mas devem ser usadas em conjunto para uma visão completa. Em aplicações reais, como análise de desempenho escolar, a mediana pode revelar uma tendência mais realista do que a média influenciada por notas extremas.

2.1.3 Medidas de Dispersão

Enquanto as medidas de tendência central indicam onde os dados se concentram, as de dispersão mostram o quão espalhados eles estão. Elas quantificam a variabilidade, essencial para entender a consistência dos dados.

- **Amplitude:** A diferença entre o valor máximo e mínimo (Amplitude = Máx - Mín). Simples, mas sensível a outliers.
- **Variância:** A média dos quadrados das diferenças em relação à média:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{2.2}$$

  Mede a dispersão quadrática, útil em análises financeiras para risco.
- **Desvio Padrão:** A raiz quadrada da variância ($s = \sqrt{s^2}$). Na mesma unidade dos dados, facilita interpretações, como em controle de qualidade industrial.
- **Intervalo Interquartil (IQR):** A diferença entre o terceiro quartil (Q3) e o primeiro (Q1), capturando a dispersão do meio 50% dos dados. Robusto contra outliers, usado em boxplots.

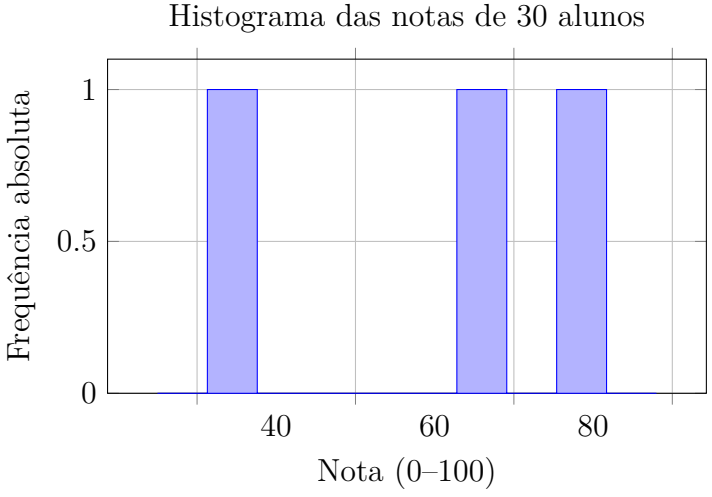
Essas medidas ajudam a comparar conjuntos de dados. Por exemplo, duas turmas com a mesma média de notas podem ter desvios padrões diferentes, indicando variabilidades distintas.

2.1.4 Distribuições de Frequência e Representações Gráficas

A estatística descritiva não se limita a números; ela emprega gráficos para visualização. As distribuições de frequência agrupam dados em classes, mostrando contagens ou percentuais. Por exemplo, uma tabela de frequência para idades pode revelar picos em faixas etárias.

Gráficos comuns incluem:

- **Histograma:** Barras representando frequências em intervalos, ideal para dados contínuos, revelando formas como normal (sino) ou assimétrica.



Histograma com 10 classes

- **Gráfico de Barras:** Para dados categóricos, comparando frequências.

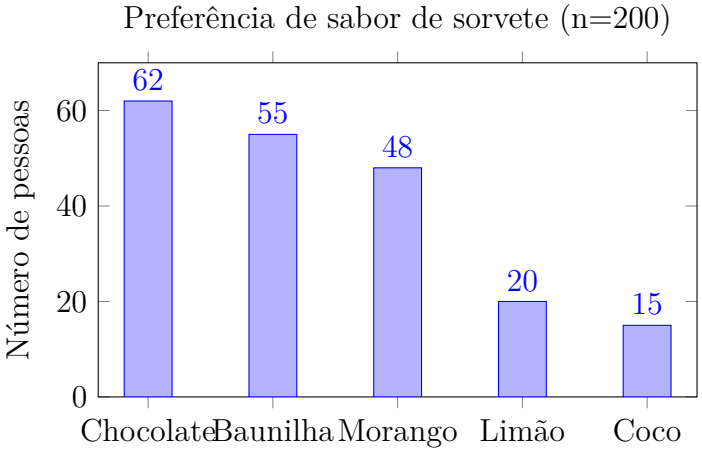
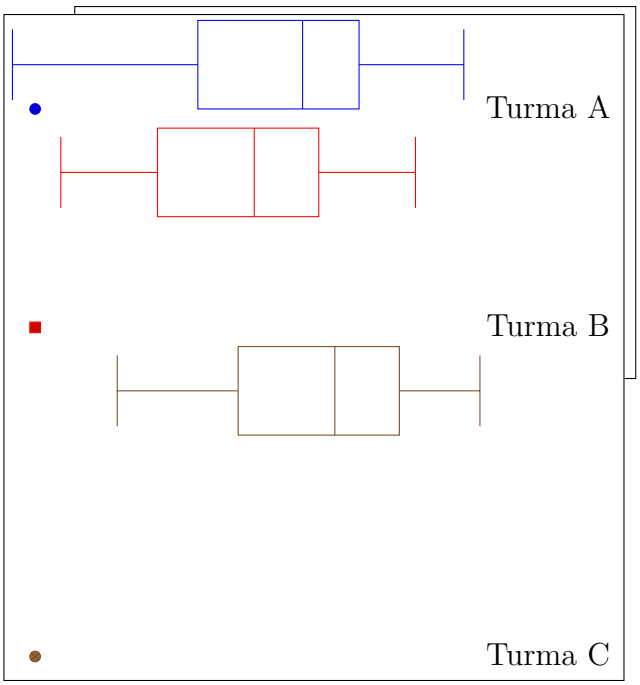


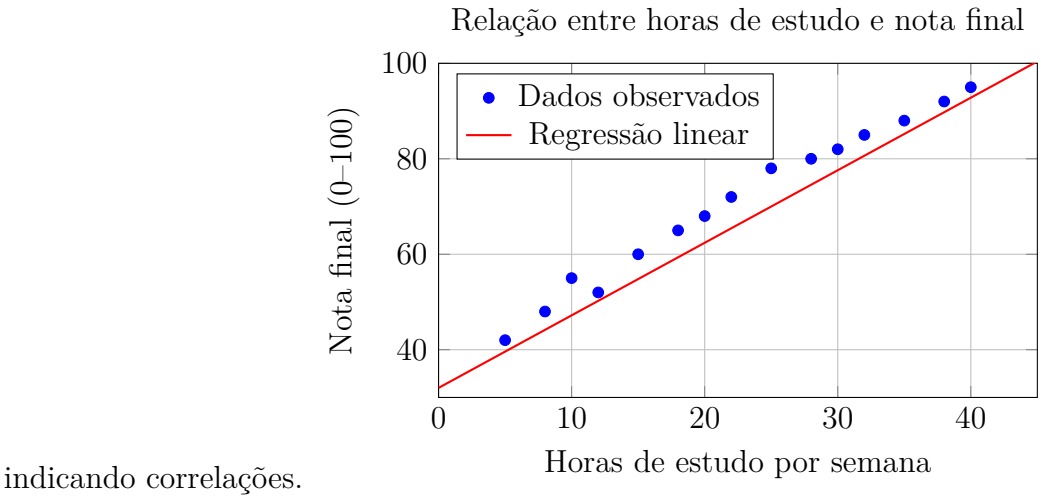
Gráfico de barras categóricas

- **Boxplot:** Mostra mediana, quartis e outliers, facilitando detecção de anomalias.



Boxplot comparativo de três turmas

- **Gráfico de Dispersão:** Para relações entre duas variáveis quantitativas,



indicando correlações.

Gráfico de dispersão com linha de tendência

Essas ferramentas visuais resumem dados complexos, tornando-os acessíveis. Em pesquisas sociais, um histograma pode destacar desigualdades em distribuições de renda.

2.1.5 Assimetria e Curtose

Para uma descrição completa, avaliamos a forma da distribuição:

- **Assimetria (Skewness):** Mede o desvio da simetria. Positiva (cauda direita longa), negativa (cauda esquerda) ou zero (simétrica).

$$Skew = \frac{3(\bar{x} - Mediana)}{s} \tag{2.3}$$

- **Curtose:** Indica o achatamento das caudas. Alta curtose significa caudas pesadas (mais outliers); baixa, caudas leves.

Essas medidas ajudam a escolher testes estatísticos apropriados e detectar desvios de normalidade.

3 Análise Exploratória de Dados

3.1 Conceito básico

A Análise Exploratória de Dados (Exploratory Data Analysis, ou EDA, na sigla em inglês) representa um pilar fundamental na ciência de dados e na estatística aplicada, servindo como etapa inicial e indispensável para compreender conjuntos de dados complexos antes de proceder a modelagens mais avançadas. A EDA enfatiza a importância de investigar os dados de forma aberta e iterativa, sem pressuposições rígidas, para revelar padrões, anomalias e relacionamentos subjacentes. Diferentemente da análise confirmatória, que testa hipóteses pré-definidas, a EDA adota uma perspectiva indutiva, guiada pela curiosidade empírica e pela robustez estatística.

Deve se entender que a EDA não é apenas como uma coleção de técnicas descritivas, mas como um *framework* metodológico ancorado em princípios probabilísticos e inferenciais. Em um contexto onde os dados crescem exponencialmente (impulsionados por

avanços em *Big Data* e *Machine Learning*), a EDA assume um papel crucial na mitigação de vieses, na detecção de erros e na formulação de hipóteses informadas.

3.2 Fundamentos Teóricos da Análise Exploratória de Dados

A EDA baseia-se em premissas estatísticas que priorizam a resiliência dos dados a suposições paramétricas. Em via de regra, não necessariamente os dados irão aderir a distribuições ideais, como a normal, e que métodos resistentes (isto é, aqueles menos sensíveis a outliers ou violações de normalidade) devem ser privilegiados. Por exemplo, em vez de depender exclusivamente da média aritmética, que é vulnerável a valores extremos, a EDA promove o uso da mediana como medida de tendência central mais robusta.

Do ponto de vista probabilístico, a EDA incorpora conceitos de distribuições empíricas e funções de distribuição cumulativa (CDF). A CDF empírica, definida como:

$$\widehat{F}_{(x)} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leqslant x) \tag{3.1}$$

onde I é a função indicadora e n o tamanho da amostra, permite uma representação não paramétrica da distribuição dos dados, facilitando comparações com distribuições teóricas via testes como o de Kolmogorov-Smirnov. Essa abordagem estatística permite identificar desvios sistemáticos, como assimetria (skewness) ou curtose (kurtosis), quantificados respectivamente por:

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \text{ e } \gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \tag{3.2}$$

onde μ_k são os momentos centrais e σ o desvio padrão, respectivamente.

Além disso, a EDA integra princípios de inferência bayesiana informal, onde priors subjetivos são atualizados com evidências dos dados. Embora não formalize modelos bayesianos completos, a iteração exploratória refina crenças iniciais sobre a estrutura dos dados, alinhando-se à filosofia de Tukey de "detecção de surpresas". Em resumo, os fundamentos teóricos da EDA com viés estatístico enfatizam a robustez, a não parametricidade e a iteração, preparando o terreno para análises mais rigorosas.

3.3 Técnicas Estatísticas Centrais na EDA

As técnicas estatísticas formam o cerne da EDA, divididas em univariadas, bivariadas e multivariadas. Na análise univariada, o foco recai sobre a distribuição individual de variáveis. Medidas de tendência central — média ($\bar{x} = \frac{1}{n}\sum x_i$), mediana e moda — são complementadas por medidas de dispersão, como variância ($s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$), desvio padrão e intervalo interquartil (IQR = Q3 - Q1). O IQR, em particular, é estatisticamente robusto para detectar outliers via regra de Tukey: valores abaixo de Q1 - 1.5 $\times$ IQR ou acima de Q3 + 1.5 $\times$ IQR são considerados anômalos.

A análise de distribuições envolve testes de normalidade, como o Shapiro-Wilk, que testa a hipótese nula de que os dados seguem uma distribuição normal:

$$W = \frac{(\sum a_i x_{(i)})^2}{\sum(x_i - \bar{x})^2} \tag{3.3}$$

onde a_i são coeficientes tabulados. Se rejeitada, distribuições alternativas (e.g., log-normal, Poisson) são exploradas. Para dados categóricos, frequências relativas e testes

qui-quadrados revelam associações inesperadas:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3.4}$$

Na análise bivariada, coeficientes de correlação são pivôs. O coeficiente de Pearson:

$$cr = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{3.5}$$

assume linearidade e normalidade, enquanto o de Spearman, baseado em ranks, é não paramétrico e robusto a outliers. Para variáveis categóricas, o coeficiente de contingência ou o teste de independência qui-quadrado é aplicado. Análises multivariadas estendem isso via matrizes de correlação e análise de componentes principais (PCA), onde autovalores e autovetores decompõem a variância: $\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$, reduzindo dimensionalidade e identificando clusters latentes.

3.4 Ferramentas Visuais e Computacionais com Ênfase Estatística

A visualização é o braço direito da EDA, transformando abstrações estatísticas em insights intuitivos. Histogramas e gráficos de densidade kernel (KDE) ilustram distribuições, com o KDE calculado pela expressão abaixo:

$$\hat{f}(x) = \frac{1}{nh} \sum K\left(\frac{x - x_i}{h}\right) \tag{3.6}$$

onde K é o kernel e h o bandwidth. Boxplots, introduzidos por Tukey, resumem quartis e outliers, facilitando comparações entre grupos via testes não paramétricos.

Scatterplots e heatmaps de correlação visualizam relacionamentos bivariados, com linhas de regressão linear:

$$\hat{y} = b_0 + b_1 x, \text{ onde } b_1 = r\frac{s_y}{s_x} \tag{3.7}$$

destacando tendências. Para dados multivariados, pairplots e gráficos de coordenadas paralelas revelam interações complexas. Em contextos estatísticos avançados, Q-Q plots comparam quantis empíricos com teóricos, diagnosticando desvios de normalidade.

Computacionalmente, linguagens como **R** (com pacotes como *ggplot2* e *exploratory*) e Python (com *pandas*, *seaborn* e *statsmodels*) operacionalizam essas técnicas. Por exemplo, em Python, a função **pandas.describe()** gera sumários estatísticos, enquanto **scipy.stats** executa testes. A integração com machine learning estende a EDA para descoberta de padrões não supervisionados, mantendo o viés estatístico via validação cruzada e métricas como o coeficiente de silhueta.

4 Amostragem

4.1 Conceito básico

A amostragem constitui um dos pilares fundamentais da estatística inferencial, permitindo que conclusões sobre uma população inteira sejam extraídas a partir do estudo de uma subparte dela, denominada amostra. Em contextos onde o censo — a análise exaustiva de todos os elementos da população — é impraticável devido a limitações de tempo, custo ou recursos, a amostragem emerge como uma ferramenta indispensável. Conceitualmente, a população refere-se ao conjunto completo de indivíduos ou unidades de interesse, caracterizada por parâmetros como média μ e variância σ^2, enquanto a amostra é um subconjunto selecionado, cujas estatísticas (e.g., média amostral $\bar{x}$ e variância s^2) servem como estimadores desses parâmetros.

A teoria da amostragem enfatiza a importância de métodos que minimizem vieses e erros amostrais, garantindo a representatividade e a validade inferencial. Distinguem-se dois grandes paradigmas: a amostragem probabilística, onde cada elemento da população tem probabilidade conhecida e não nula de ser selecionado, permitindo o cálculo de erros padrão e intervalos de confiança; e a amostragem não probabilística, baseada em critérios subjetivos, mais suscetível a vieses, mas útil em pesquisas exploratórias ou com populações de difícil acesso.

4.2 Fundamentação Teórica da Amostragem

A amostragem probabilística baseia-se no princípio de que a seleção aleatória reduz o viés de seleção, permitindo que a distribuição amostral de uma estatística siga leis probabilísticas conhecidas, como o Teorema Central do Limite (TCL). Pelo TCL, para amostras suficientemente grandes ($n > 30$), a média amostral $\bar{x}$ aproxima-se de uma distribuição normal com média μ e variância σ^2/n, independentemente da distribuição populacional.

O erro amostral, quantificado pelo erro padrão, é dado por $SE = \sigma/\sqrt{n}$ para a média, com correção para populações finitas:

$$SE = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \tag{4.1}$$

onde N é o tamanho populacional. Já o viés ocorre quando o método de seleção sistematicamente super ou subestima o parâmetro. Em amostragens não probabilísticas, o viés é incontrolável, impedindo generalizações rigorosas. Outro conceito chave é o quadro amostral (sampling frame), lista completa e atualizada da população, essencial para evitar erros de cobertura.

4.3 Métodos de Amostragem Probabilística

Os métodos probabilísticos garantem inferência estatística válida. São eles:

- **Amostragem Aleatória Simples (AAS):** Cada elemento tem probabilidade igual de seleção ($p = n/N$). Pode ser com ou sem reposição. Fórmula de variância da média:

$$Var(\bar{x}) = \frac{\sigma^2}{n}\left(1 - \frac{n}{N}\right) \tag{4.2}$$

  Simples e sem viés, mas ineficiente para populações heterogêneas.

- **Amostragem Sistemática:** Seleciona-se um elemento inicial aleatório k e, em seguida, cada i-ésimo ($i = N/n$). Eficiente operacionalmente, mas suscetível a periodicidades na lista.

- **Amostragem Estratificada:** A população é dividida em estratos homogêneos (e.g., por idade ou renda), com seleção aleatória proporcional ou ótima dentro de cada. Reduz a expressão da variância por:

$$Var(\bar{x}_{est}) = \sum W_h^2 \frac{\sigma_h^2}{n_h} \tag{4.3}$$

  onde W_h é o peso do estrato. Ideal para heterogeneidade conhecida.

- **Amostragem por Conglomerados (Cluster):** Divide-se em clusters heterogêneos (e.g., escolas), selecionando clusters aleatoriamente e amostrando todos ou subamostra dentro. Útil para populações dispersas, mas aumenta variância se clusters forem semelhantes internamente.

4.4 Métodos de Amostragem Não Probabilística

Embora menos rigorosos, são comuns em pesquisas qualitativas:

- **Amostragem por Conveniência:** Seleção de elementos acessíveis. Rápida, mas propensa a viés.

- **Amostragem por Cotas:** Similar à estratificada, mas sem aleatoriedade; fixa cotas por características.

- **Amostragem Intencional ou por Julgamento:** É onde o especialista seleciona elementos típicos.

- **AAmostragem "Bola de Neve":** Indicada para populações raras; participantes indicam outros.

Esses métodos não permitem cálculo de erros padrão, limitando a generalização.