Estatística Descritiva

Stefano Mozart

12/02/2025

Sumário

- Frequência
- Variáveis e gráficos associados
- Prática

Frequência

Análise univariada

É o processo estatístico de examinar e descrever uma única dimensão, ou variável, dos elementos da amostra. Seu objetivo principal é descrever e resumir as características essenciais da população em relação a essa variável.

- Distribuição de Frequência;
- Medidas de Tendência Central;
- Medidas Separatrizes;
- Medidas de Dispersão;
- Medidas de forma (assimetria e curtose);
- Identificação de padrões e *outliers* (observações anômalas);

Análise paramétrica

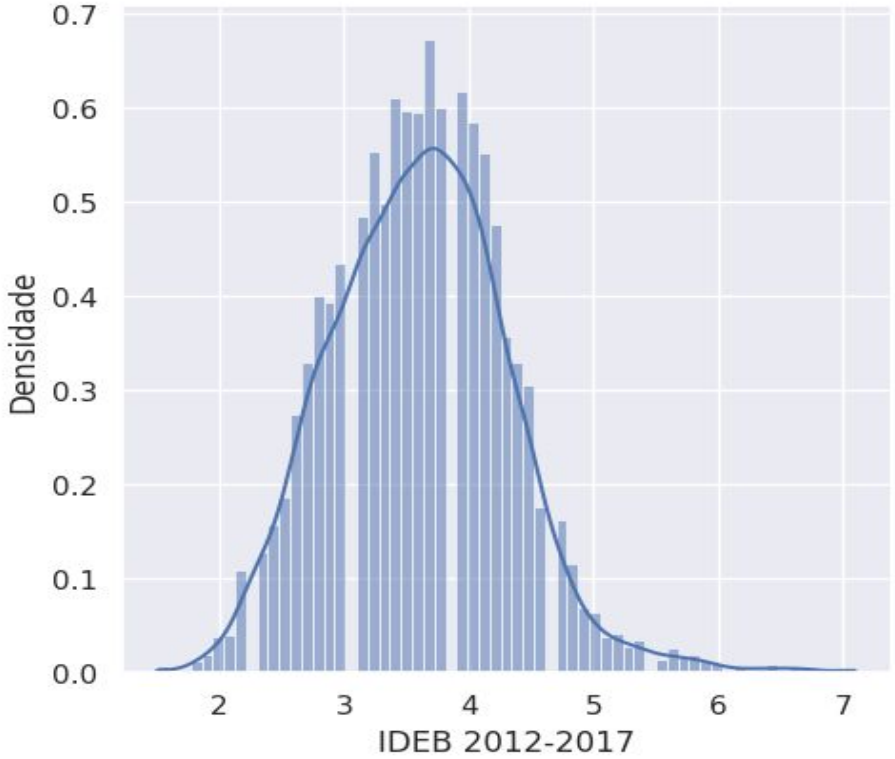
Assume que os dados seguem uma distribuição específica, geralmente a distribuição normal (gaussiana). Dessa forma, a análise envolve a estimação de parâmetros dessa distribuição, como média, variância e desvio padrão.

- Requisitos:
  - Normalidade dos dados;
  - Homogeneidade de variâncias (em alguns testes);
  - Escala de medida intervalar ou racional.

- Testes comuns:
  - Teste t de Student: comparar as médias de dois grupos.
  - ANOVA: comparar as médias de três ou mais grupos.
  - Regressão Linear: modela a relação entre uma variável dependente e uma ou mais variáveis independentes.
  - Correlação de Pearson: Mede a força e a direção da relação linear entre duas variáveis contínuas.

Análise paramétrica

Exemplo: score/nota do IDEB 2012-2017, no dataset do Ideb



Análise não-paramétrica

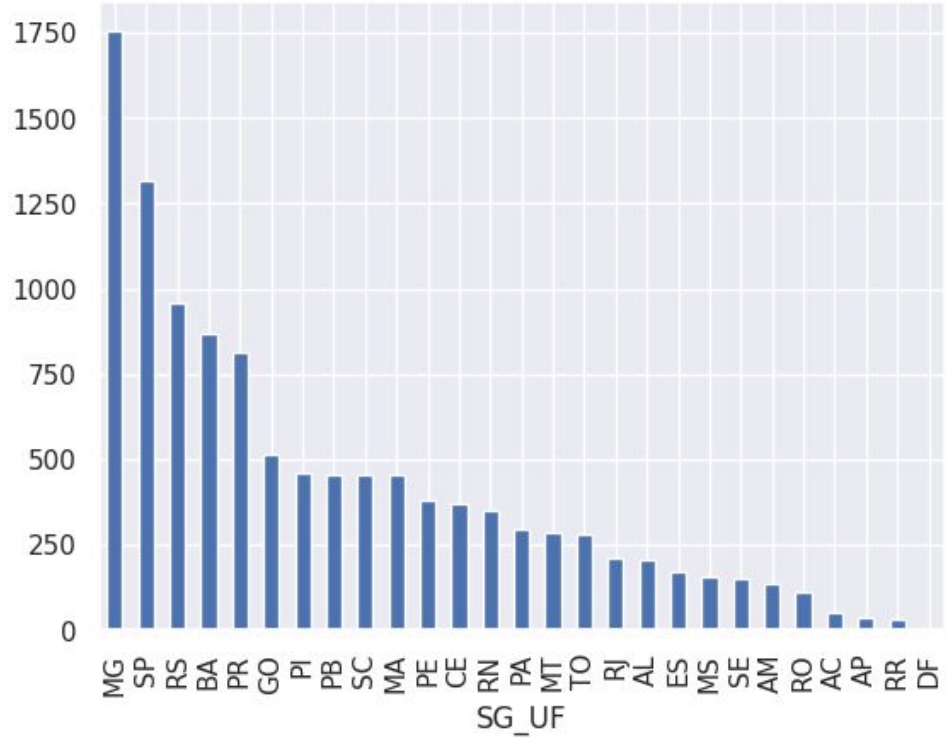
Não assume uma forma específica para a distribuição dos dados. É mais flexível e pode ser aplicada a dados que não seguem uma distribuição normal ou que são medidos em escalas ordinais.

- Vantagens:
  - Não requer normalidade nem homogeneidade;
  - Resistentes a amostras pequenas ou com amplitude significativa;

- Exemplos de Testes Não Paramétricos:
  - Teste de Mann-Whitney U: comparar médias quando a normalidade não pode ser assumida.
  - Teste de Wilcoxon para Amostras Pareadas: comparar duas condições ou medições relacionadas.
  - Teste de Kruskal-Wallis: comparar três ou mais grupos.
  - Correlação de Spearman: medir a relação entre variáveis quando a relação linear não pode ser assumida ou quando os dados são ordinais.

Análise não-paramétrica

Exemplo: UF, no dataset do Ideb



Frequência

Quantidade de observações de um valor específico para uma variável. Obtida a partir da simples contagem das observações na amostra, permite a identificação de padrões, bem como da existência de eventos anômalos, a partir da esporadicidade ou raridade de certas observações.

- **Absoluta:**
  Número exato de vezes que um valor específico aparece em um conjunto de dados.

- **Relativa:**
  Proporção ou porcentagem de ocorrências de um valor específico frente ao total de observações. É calculada dividindo a frequência absoluta pelo total de elementos.

- **Frequência Acumulada:**
  Soma das frequências de todas as classes anteriores até a classe atual. Ajuda a entender quantas observações recaem em intervalos ou categorias.

Distribuição de Frequência

Uma sumarização ou tabulação dos dados brutos que tem por objetivo demonstrar como as observações são distribuídas ao longo de diferentes categorias ou intervalos numéricos.

- Para variáveis quantitativas, os valores são agrupados em intervalos (também conhecidos como *bins*).

- Para variáveis qualitativas, as categorias são geralmente definidas pelas próprias características dos dados (os distintos valores admitidos ou efetivamente observados para cada variável).

- Pode se apresentar em diferentes escalas:
  - Nominal;
  - Ordinal;
  - Intervalar;
  - Razão;

Distribuição de Frequência

Escala Nominal:

Dados classificados em categorias sem uma ordem natural. As categorias servem apenas para identificação.

Exemplo: Frequência da rede ‘Pública’ (etc.) no dataset IDEB:

Rede	Frequência
Pública	5420
Estadual	5417
Federal	334
Municipal	91

Distribuição de Frequência

Escala Ordinal:

As categorias possuem uma ordem ou hierarquia, mas os intervalos entre elas não são necessariamente iguais ou mensuráveis..

Exemplo: faixas etárias no dataset ‘consumidor’

Faixa Etária	Frequência
até 20 anos	270
21 a 30 anos	10906
31 a 40 anos	21679
41 a 50 anos	9656
51 a 60 anos	4652
61 a 70 anos	2128
mais de 70 anos	575

Distribuição de Frequência

Escala Intervalar:

Mede dados numéricos onde os intervalos entre os valores são iguais. Contudo, não há um zero absoluto que represente a ausência da característica medida.

Exemplo: faixas de nota no IDEB

Faixa	Frequência
1-2	42
2-3	1997
3-4	5497
4-5	3035
5-6	287
6-7	48
7-8	8

Distribuição de Frequência

Escala Intervalar:

Mede dados numéricos onde os intervalos entre os valores são iguais. Contudo, não há um zero absoluto que represente a ausência da característica medida.

Exemplo: faixas de nota no IDEB

Faixa	Frequência
1-2	42
2-3	1997
3-4	5497
4-5	3035
5-6	287
6-7	48
7-8	8

Distribuição de Frequência

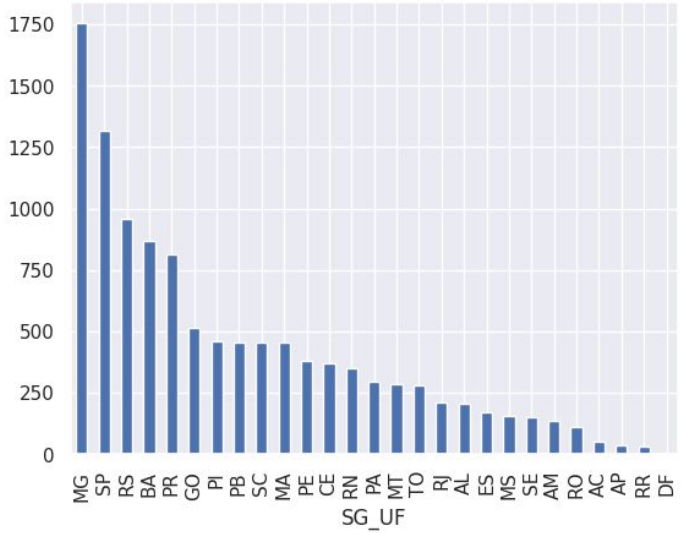
Escala Razão:

Na escala de razão, os dados numéricos possuem intervalos iguais e um zero absoluto, o que permite operações aritméticas (como multiplicação e divisão). Exemplos comuns incluem altura, peso e renda.

Faixa	Frequência
0 a 2000	10
2000 a 4000	25
4000 a 6000	15
6000 a 8000	5

Desbalanceamento

Ocorre quando as frequências associadas às classes ou categorias de uma amostra são muito discrepantes. Em outras palavras, uma ou mais classes possuem significativamente mais observações do que outras. Esse problema é bastante comum em tarefas de classificação, onde uma classe majoritária pode dominar o conjunto e mascarar a presença e a importância da(s) classe(s) minoritária(s).



Desbalanceamento

Medidas Separatrizes:

Medidas separatrizes (como a mediana, quartis e percentis) têm a função de "dividir" a distribuição dos dados em partes iguais. Quando a amostra é desbalanceada:

- Influência do Grupo Majoritário:
  Se um grupo domina a amostra, as medidas separatrizes tenderão a refletir a distribuição desse grupo. Por exemplo, a mediana do conjunto global pode se posicionar de forma a representar predominantemente os dados da classe majoritária, mascarando as características (como centralidade) dos grupos minoritários.

- Perda de Informações Sobre Subgrupos:
  Quando diferentes subgrupos possuem distribuições distintas, o cálculo de quartis ou percentis sobre o conjunto agregado pode não capturar essas variações. Assim, a interpretação dos dados pode ser equivocada, pois não se sabe se os cortes (por exemplo, o primeiro ou o terceiro quartil) se aplicam igualmente a todas as subpopulações.

Desbalanceamento

Medidas de Dispersão

Medidas de dispersão (como desvio padrão, variância, amplitude e intervalo interquartílico) quantificam a variabilidade dos dados. O desbalanceamento pode levar a:

- Distorção do Valor Global:
  Se a classe majoritária possui uma variabilidade relativamente baixa e a minoritária uma variabilidade alta (ou vice-versa), o cálculo global pode subestimar ou superestimar a verdadeira dispersão. O grupo com maior número de observações "puxa" o valor final, ocultando a variabilidade real presente na(s) classe(s) menos representada(s).

- Sensibilidade a Outliers:
  Em amostras desbalanceadas, se a classe minoritária contiver outliers ou valores extremos, sua influência pode ser diluída na análise global. Dessa forma, medidas como a amplitude ou mesmo o desvio padrão podem não refletir adequadamente a dispersão dos dados em cada subgrupo.

Desbalanceamento

Viés do Modelo:

- Tendência à Classe Majoritária: Algoritmos de aprendizado de máquina podem se tornar tendenciosos, favorecendo a classe que possui maior número de instâncias. Isso pode levar a um desempenho aparentemente bom (alta acurácia) mas, na realidade, o modelo pode estar ignorando a classe minoritária.
- Previsões Inadequadas: O modelo pode ter dificuldades em reconhecer ou prever corretamente as instâncias da classe minoritária, resultando em um alto número de falsos negativos ou falsos positivos.

Avaliação Enganosa:

- Métricas Infladas: métricas medidas na amostra podem ser enganosas em contextos desbalanceados.
- Overfitting para a Classe Majoritária: O modelo pode se ajustar muito bem aos padrões da classe majoritária, perdendo a capacidade de generalizar para novas amostras que contenham as classes minoritárias.

Variáveis e gráficos associados

Variável

Uma característica, ou atributo, associada a cada elemento de uma população ou amostra. Também pode ser entendida como qualquer dimensão do conjunto de dados em análise que permite medir e expressar características de interesse de forma quantificável.

- Essas características podem se apresentar de diversas formas, manifestando uma natureza qualitativa ou quantitativa;
- Identificar e classificar corretamente as variáveis é crucial para o sucesso de qualquer análise estatística, pois isso influencia desde o design da coleta de dados até as técnicas de análise e a interpretação dos resultados;
  - Na coleta, a definição das variáveis garante a precisão e relevância dos dados, tendo em vista as perguntas de pesquisa;
  - Na análise, a correta interpretação e comunicação das variáveis ajuda tanto a moldar os quanto a comunicar resultados;

Variável qualitativa

Informação de natureza categórica, não quantificável. Tem forte relação com a análise de frequência e de distribuição de frequência.

- **Nominal:** Representam categorias sem qualquer ordem natural entre elas. A definição da variável é complicada pelo fato de que elementos podem ser associados a um ou mais valores. Exemplos incluem cor dos cabelos, gênero, ocupação, nacionalidade.
- **Ordinal:** Semelhantes às variáveis nominais, mas com uma ordem ou hierarquia definida. Elementos podem estar associados a apenas um valor. Exemplos incluem níveis de educação (fundamental, médio, superior); e classificações de percepção de qualidade (uma, duas, três estrelas) ou intensidade (discordo totalmente, discordo, etc);

Variável qualitativa

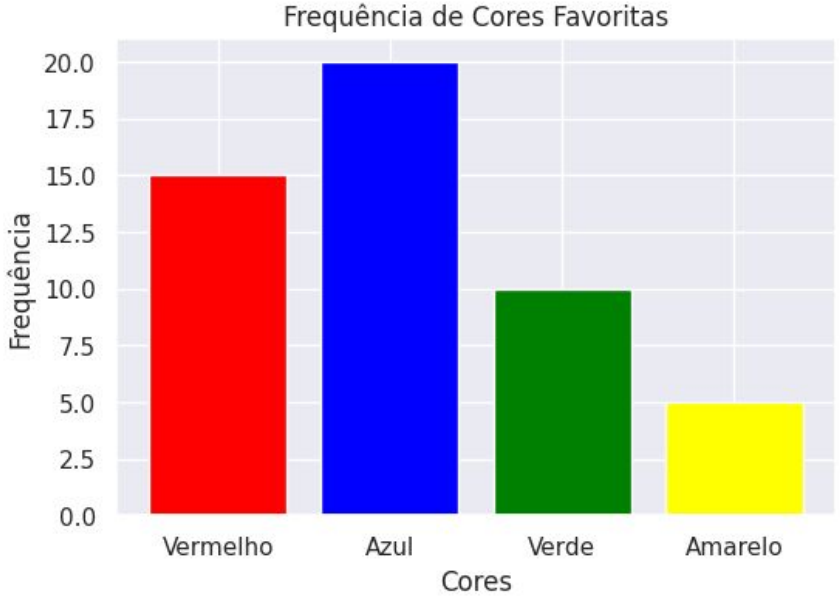
Os gráficos mais utilizados para analisar essas variáveis são:

- **Barras;**
- **Gráfico de área (Pizza);**
- **Gráfico de Pareto;**
- **Gráfico de Mosaico;**

Variável qualitativa

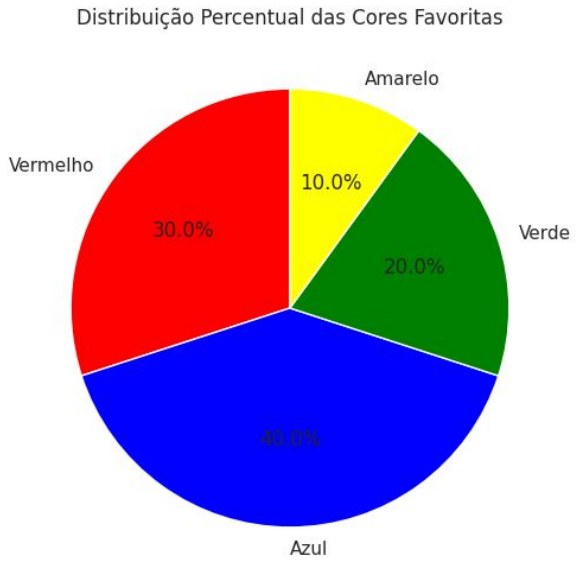
Barras:

Exibe a frequência ou a proporção de cada categoria por meio de barras, que podem ser dispostas horizontal ou verticalmente. É útil para comparar diferentes categorias.



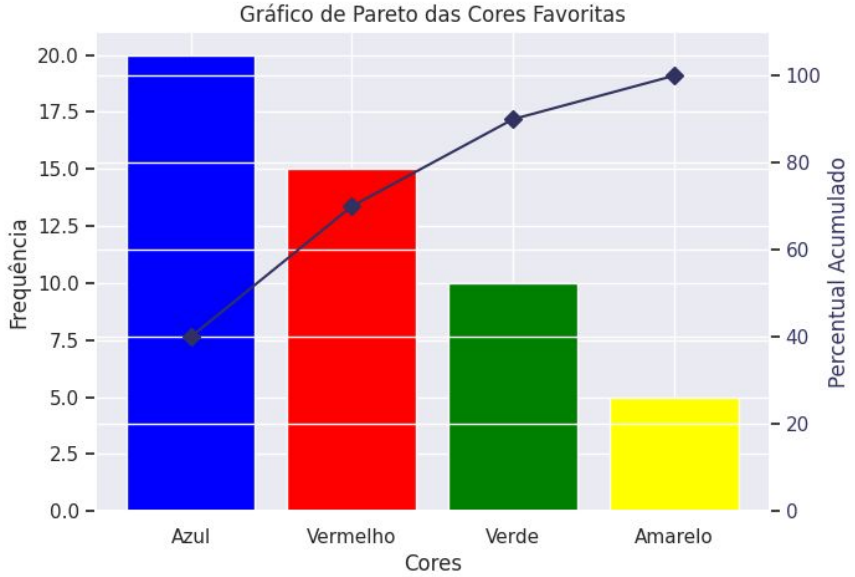
Variável qualitativa

Gráfico de setor (Pizza): Representa as proporções de cada categoria dentro da amostra, de forma que "fatia" do círculo indica a porcentagem correspondente à categoria. É indicado quando o número de categorias é pequeno e se deseja evidenciar a composição percentual.



Variável qualitativa

Gráfico de Pareto: Combina um gráfico de barras (ordenadas de forma decrescente pela frequência) com uma linha que mostra o acumulado das frequências. Esse gráfico ajuda a identificar quais categorias são as mais relevantes em termos de contribuição.

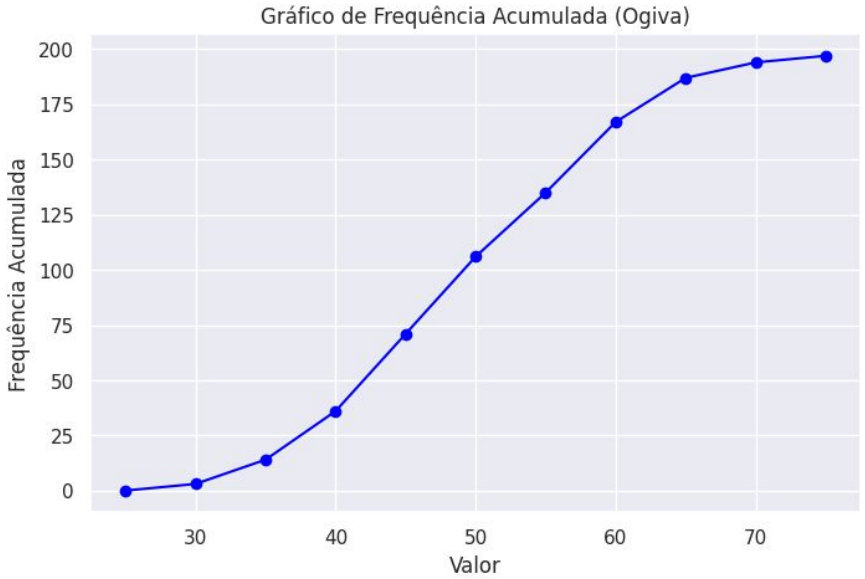


Variável qualitativa

Gráfico de Frequência Acumulada:
O gráfico de frequência acumulada, também conhecido como ogiva, é uma representação gráfica que mostra a soma acumulada das frequências à medida que se percorre os valores ou intervalos de uma variável. Em outras palavras, para cada valor ou classe, o gráfico exibe o total de observações que estão abaixo daquele ponto.

Essa visualização é útil para:

- Identificar quantos elementos estão abaixo de um determinado valor.
- Determinar percentis, mediana e outros quantis.
- Comparar distribuições acumuladas de diferentes conjuntos de dados.



Variável qualitativa

Gráfico de Mosaico: Utilizado para visualizar a relação entre duas variáveis qualitativas, exibindo a proporção das categorias em áreas proporcionais.

Variável quantitativa

Informação de natureza numérica que quantifica uma característica do elemento. São fundamentais numa análise estatística descritiva, uma vez que permitem a aferição direta de medidas estatísticas.

- **Discreta:** Assumem valores inteiros ou contáveis. Geralmente resultam de processos de contagem e têm um número finito ou enumerável de valores possíveis. Exemplos incluem o número de filhos, número de carros em um estacionamento.
- **Contínua:** Podem assumir qualquer valor dentro de um intervalo contínuo. Geralmente resultam de medições e podem incluir valores fracionários. Exemplos incluem altura, peso, temperatura;

Variável quantitativa

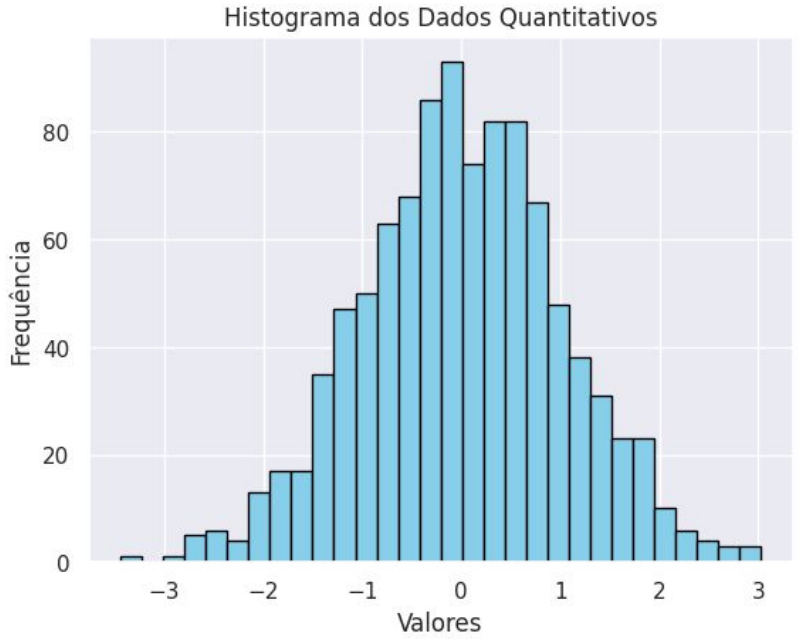
Os gráficos comumente utilizados na análise são:

- **Histograma;**
- **KDE (Kernel Density Estimation);**
- **Boxplot;**
- **Gráfico de dispersão (scatter plot);**
- **Gráfico de linha;**

Variável quantitativa

Histograma:

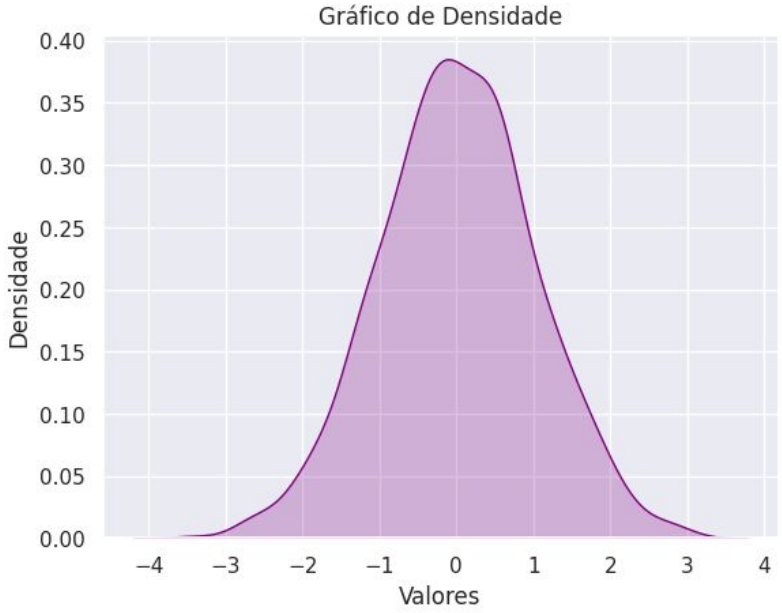
Exibe a distribuição dos dados em intervalos (bins), facilitando a visualização de sua distribuição de frequência e forma (distribuição normal, assimetria, etc.).



Variável quantitativa

KDE (Kernel Density Estimation):

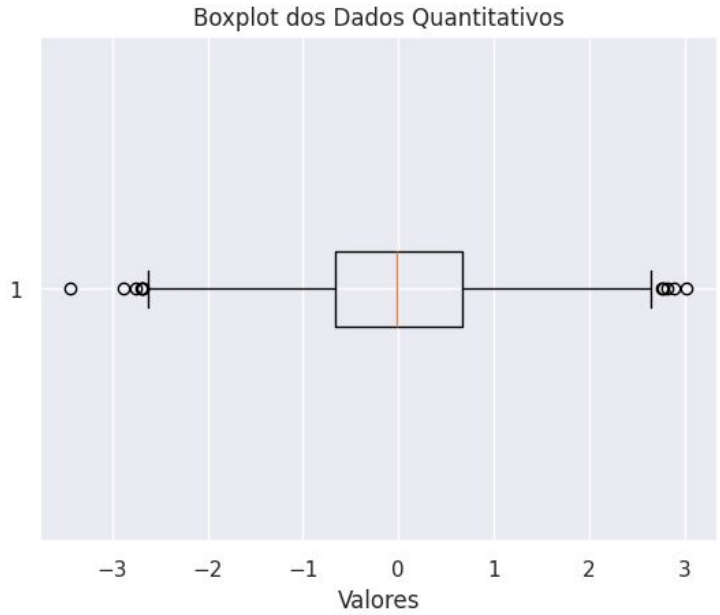
Fornece uma estimativa suavizada da função densidade de probabilidade dos dados, sendo útil para identificar a forma da distribuição sem a rigidez dos bins do histograma.



Variável quantitativa

Boxplot:

Resume a distribuição dos dados destacando a mediana, os quartis e os possíveis outliers, facilitando a comparação entre diferentes conjuntos de dados.



Variável quantitativa

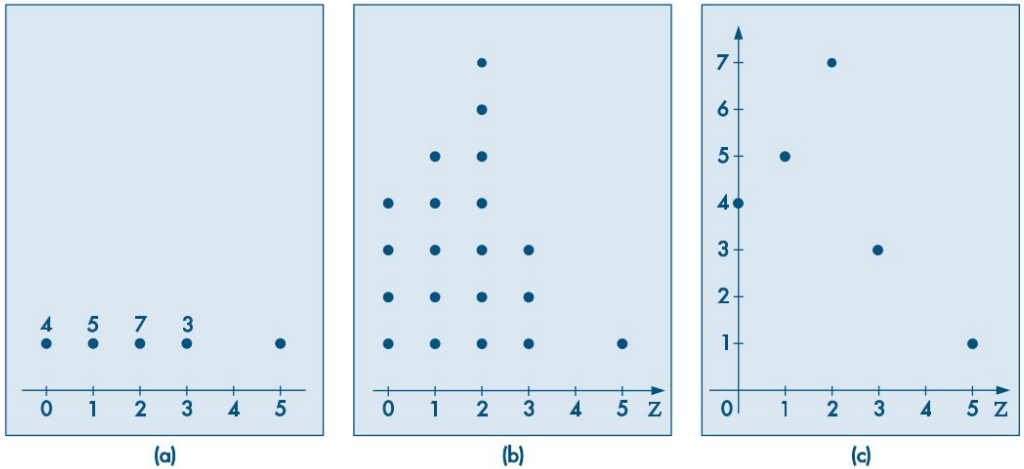
Gráfico de dispersão (scatter plot) unidimensional:

Os valores são representados por pontos ao longo da reta (provida de uma escala).

(a) Valores repetidos são acompanhados pelo número de repetições;

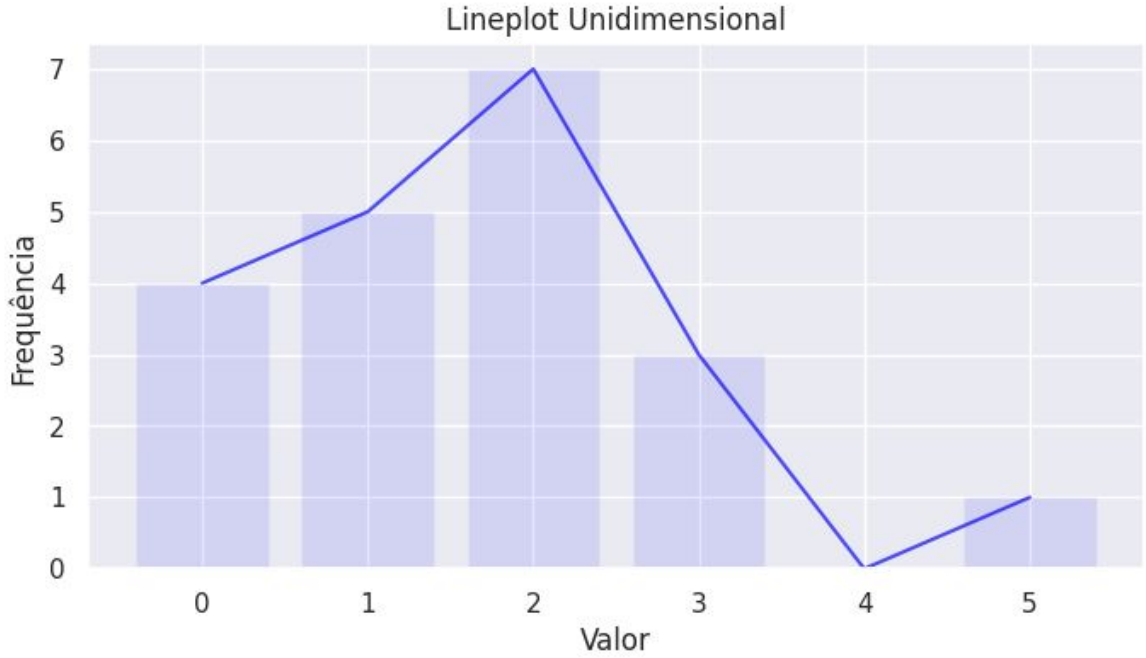
(b) Valores repetidos são “empilhados”;

(c) Apresenta apenas o ponto mais alto da pilha;



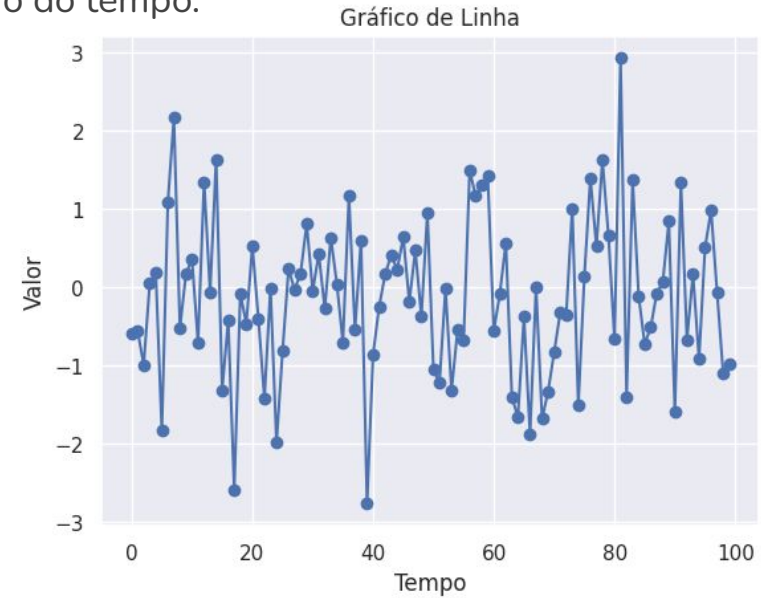
Variável quantitativa

Gráfico de linha unidimensional:
Exibe a frequência de observações ligando o topo dos bins por meio de uma linha.



Variável quantitativa

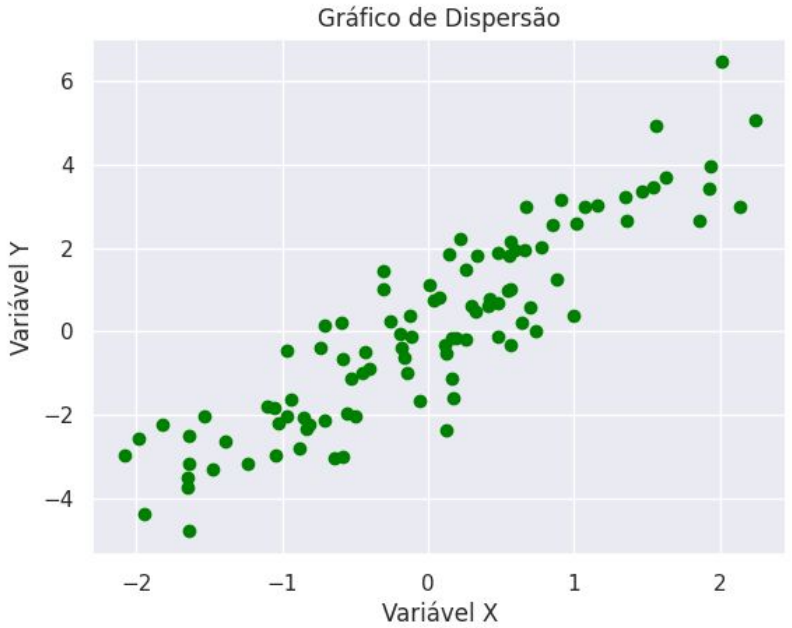
Gráfico de linha bidimensional: Conecta os pontos de dados em sequência, exibindo tendências e variações de uma variável dependente em relação a outra variável, independente e igualmente quantitativa. Muito utilizado para análise de séries temporais, evidenciando a evolução da variável de interesse ao longo do tempo.



Variável quantitativa

Gráfico de dispersão (scatter plot):

Utilizado para investigar a relação entre duas variáveis quantitativas, mostrando cada par de valores como um ponto no gráfico.



Prática

Exercícios:

1. Baixe um dataset de sua escolha no Colab/Jupyter Notebook;
2. Exiba as variáveis disponíveis na amostra;
3. Identifique o tipo de cada variável (qualitativa nominal/ordinal, quantitativa discreta/contínua);
4. Exiba a distribuição de frequência de cada variável;
5. Indique, caso existam, os possíveis desbalanceamentos de sua amostra;
6. Escolha uma variável qualitativa nominal e exiba um gráfico de setores (pizza);
7. Escolha uma variável quantitativa e exiba um histograma;

Obrigado

Stefano Mozart
linkedin.com/in/stefano-mozart/
github.com/stefanomozart