Aula 04 - Análise univariada: dispersão e forma

MBA
Ciência de Dados e Inteligência Artificial Aplicadas

Estatística Descritiva

Stefano Mozart

12/02/2025

Sumário

- Medidas de dispersão
- Assimetria e curtose

Dispersão

Medidas de dispersão

Dispersão é o conceito estatístico que descreve como os elementos se espalham em torno de uma medida de tendência central (como a média ou a mediana).

- Em última instância, a dispersão nos informa sobre a heterogeneidade ou homogeneidade da amostra, indicando se os elementos estão concentrados em torno do centro amostral ou mais amplamente distribuídos.

- A maioria das medidas de dispersão tem a mesma unidade de medida que os elementos da amostra. Em outras palavras, se as medições estão em metros ou segundos, a medida de dispersão também está. Exemplos de medidas de dispersão incluem:
  - Amplitude
  - Intervalo Interquartílico
  - Coeficiente de Dispersão Quartílica
  - Desvio Absoluto Médio
  - Desvio Absoluto Mediano
  - Variância
  - Desvio Padrão
  - Coeficiente de Variação

Amplitude amostral

Também conhecida como amplitude total, é a medida de escala da dispersão indicada pela distância máxima entre os elementos em uma amostra. Para uma amostra ordenada de tamanho n, é calculada pela diferença:

$$A = x_n - x_1$$

- Útil na comparação de amostras/sub amostras;
- Não revela informação sobre a distribuição interna, mas, quando aliada a outras medidas de dispersão pode identificar a presença de *outliers*.

Intervalo interquartílico (IIQ)

É a medida de escala obtida pela diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1.

$$\text{IIQ} = Q_3 - Q_1$$

- Robusto contra *outliers*, por abranger os 50% elementos centrais, fornece uma boa medida da variabilidade central da amostra;

- Complementa a análise da Amplitude, informando a existência (ou não) de consistência entre a variabilidade observada no centro e nas caudas da distribuição.

Coeficiente de Dispersão Quartílica (CDQ)

É uma medida de dispersão adimensional, útil para comparação da variabilidades entre diferentes subgrupos ou diferentes amostras. Dada pela expressão:

$$CDQ = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

- Em muitos casos, será necessário expressar a variabilidade em termos relativos porque, por exemplo, um desvio-padrão igual a 1 pode ser muito pequeno se a magnitude dos dados é da ordem de $\times 10^3$, mas pode ser considerado muito elevado se esta magnitude da ordem de $\times 10^1$.

Desvio Absoluto Médio

É a média das distâncias entre cada elemento e a média amostral. Ele nos dá, portanto, uma noção da variabilidade média na amostra.

$$\mathrm{DAM} = \frac{\sum |x_i - \bar{x}|}{n}$$

- Assim como qualquer média aritmética, é extremamente sensível a *outliers*
- Dá, por outro lado, uma noção balanceada do espalhamento dos elementos nos dois lados da medida central;

Desvio Mediano Absoluto

É a mediana da série de distâncias absolutas em relação à mediana. Dessa forma, tem-se uma medida de posição central dos desvios (indica que 50% dos desvios são menores/maiores que esse valor).

$$\text{MAD} = median(|x_i - Md|_{i=1}^{n})$$

- Tem interpretação bastante simples e intuitiva
- Tem a mesma unidade de medida da variável estudada
- Mais resistente a *outliers* que o Desvio Absoluto Médio

Momentos amostrais

São ferramentas essenciais porque fornecem um resumo numérico que descreve as principais características da distribuição dos dados, permitindo uma compreensão inicial sobre sua forma e comportamento.

- Em termos gerais, o k-ésimo momento (ou momento de ordem k) de uma amostra é calculado como a média das k-ésimas potências dos valores (ou das diferenças dos valores em relação a algum ponto de referência, geralmente a média).
  - **Momento bruto**: o primeiro momento bruto é sempre a média aritmética. A interpretação dos demais depende de contexto.

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

  - **Momento central:** o primeiro momento central é sempre zero. O segundo é a variância.

$$m_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^k$$

Momentos amostrais

- **Identificação de Padrões:**
  - A média (primeiro momento) indica a tendência central.
  - A variância (segundo momento central) indica a dispersão, ajudando a identificar a homogeneidade ou heterogeneidade dos dados.
  - A assimetria (terceiro momento central) revela se os dados estão inclinados para a direita ou para a esquerda.
  - A curtose (quarto momento central) indica se os dados possuem caudas pesadas ou leves comparadas a uma distribuição normal.

- **Comparação de Distribuições:** Permitem comparar diferentes conjuntos de dados ou verificar se os dados seguem uma distribuição conhecida (como a normal), o que pode ser crucial para a escolha de métodos estatísticos e modelagem.

- **Identificação de Outliers:** Uma alta variância ou uma assimetria acentuada podem sugerir a presença de *outliers*, que precisam ser investigados ou tratados adequadamente.

Variância

Como segundo momento central, é a média dos quadrados das distâncias entre cada elemento e a média (populacional ou amostral). Indica, portanto, de forma robusta, a variabilidade dos elementos da amostra.

- O cômputo da variância pressupõe o conhecimento da média populacional;

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- A variância amostral, ou empírica, requer apenas o conhecimento da média amostral

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Desvio padrão

É a raiz quadrada da variância e informa, em média, o quanto os valores individuais se afastam do valor central (média) de uma amostra.

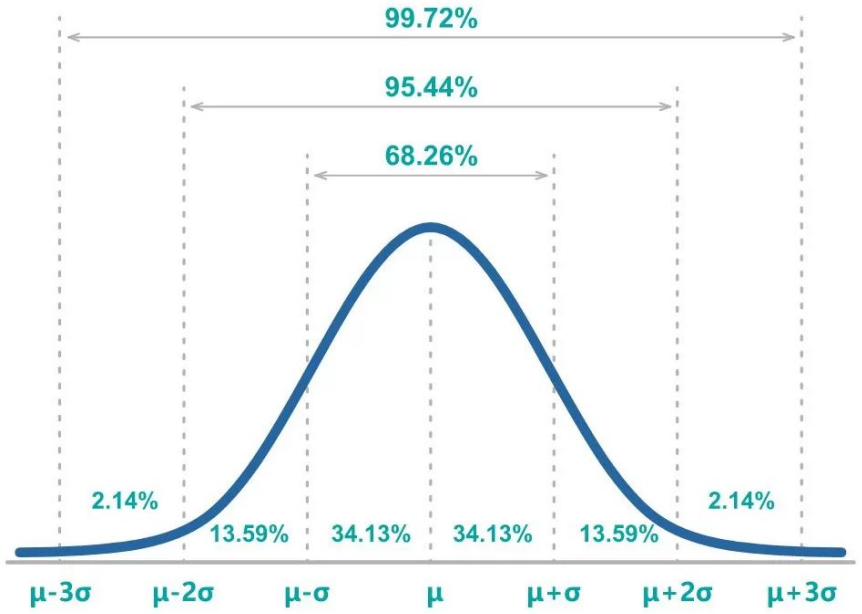
$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Indica a dispersão dos elementos, tendo como referência a média;
- É uma grandeza expressa na mesma unidade de medida da variável analisada;
- Altamente sensível a *outliers*;
- Também é sensível ao ajuste (ou não) da amostra à distribuição normal;
- Não pode ser confundido com o erro amostral ou erro padrão, que são associados à diferença estimada entre qualquer estatística computada sobre a amostra e sua correspondente populacional;

Desvio padrão

Numa distribuição Normal (Gaussiana) ou aproximadamente normal, caracterizada por μ e σ:

- 68,2% das observações distam até 1σ da média;
- 95,4% distam até 2σ;
- 99,7% distam até 3σ;



Coeficiente de Variação

É uma medida de variabilidade relativa, expressa na forma da seguinte grandeza adimensional:

$$CV = \frac{\sigma}{\mu} \qquad \text{ou} \qquad \bar{CV} = \frac{S}{\bar{x}}$$

- Útil para comparar a variabilidade de diferentes variáveis no mesma amostra, com distintas unidades de medida;
- Também utilizada para comparar a variabilidade de populaçẽos completamente heterogêneas;

Prática

Exercícios

1. Selecione uma variável contínua de interesse e calcule todas as medidas de dispersão apresentadas (variância e desvios amostrais, bem como amplitude, intervalo interquartílico, coeficiente de dispersão quartílica e coeficiente de variação). Anote em seu caderno a interpretação das medidas obtidas:
    1.1. Sua amostra é concentrada ou dispersa?
    1.2. A amplitude interna (IIQ) é proporcional à amplitude total da amostra?
    1.3. Quais a implicações dessas informações em relação ao contexto/significado concreto dessa variável?

2. Exiba um gráfico de dispersão (KDE) da variável estudada
    2.1. Sinalize os componentes do IIQ (*i.e.* Q1 e Q3) com linhas verticais na cor cinza;
    2.2. Sinalize, com linhas verticais azuis, a média amostral e o primeiro desvio padrão à esquerda e à direita

3. Exiba um gráfico KDE, marcando, quando existir, os *outliers*;

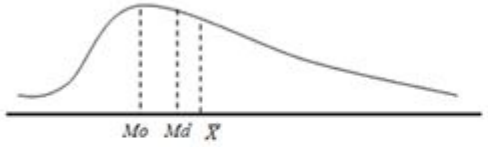
Forma

Assimetria

A assimetria, ou skewness, quantifica o grau de simetria de uma distribuição em torno da sua média.

Assimetria Positiva:

- Cauda à direita longa ou pesada
- Média maior que mediana, que é maior que a moda

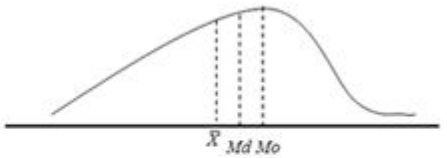


a) Assimetria à direita ou positiva

$\bar{x} > Md > Mo$

Assimetria Negativa:

- Cauda longa à esquerda
- Moda maior que mediana, que é maior que a média



b) Assimetria à esquerda ou negativa

$Mo > Md > \bar{x}$

Distribuição Simétrica:

- Média, mediana e moda são coincidentes
- Forma de sino com dados distribuídos igualmente à direita e à esquerda

Assimetria

A assimetria pode ser medida através de diferentes coeficientes que avaliam como os dados se distribuem em relação às medidas de tendência central.

Terceiro Momento Central:

- Medida robusta de assimetria;
- Dá a intensidade e direção da assimetria

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$$

Coeficiente de Fisher-Pearson:

- Medida ajustada a partir do terceiro momento central.
- Tem a mesma interpretação, mas tende a corrigir o viés de amostras menores.
- Mais comumente utilizado entre os coeficientes de assimetria.

$$\mathrm{CFP} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S}\right)^3$$

Assimetria

Primeiro Coeficientes de Pearson:

- Compara a média com a moda. Em uma distribuição simétrica, média e moda são iguais.
- Se a média amostral for maior que a moda, haverá assimetria positiva; se for menor, negativa.
- A dificuldade está em definir a moda de forma consistente, especialmente em distribuições contínuas.

$$\mathrm{CAP}_1 = \frac{\bar{x} - Mo}{S}$$

É importante notar que é quase impossível observar simetria pura em uma amostra, por isso os coeficientes de assimetria assumem valores quase sempre diferentes de zero

Assimetria

Segundo Coeficiente de Pearson:

- Multiplica por 3 a diferença entre a média e a mediana, padronizada pelo desvio padrão.
- Assume que a mediana é uma boa aproximação para a moda em distribuições levemente assimétricas e é mais prática de calcular.
- É amplamente utilizada como um indicador rápido da direção da assimetria.

$$\mathrm{CAP}_2 = \frac{3(\bar{x} - Md)}{S}$$

Coeficiente Quartílico de Bowle

- Medida bastante simples e robusta.
- Menos sensível a outliers.
- É muito útil para obter uma visão rápida sobre a simetria dos dados no centro da amostra.

$$\mathrm{CAQ} = \frac{(Q3 - Q2) - (Q2 - Q1)}{(Q3 - Q1)}$$

Curtose

A curtose mede o "peso" das caudas de uma distribuição e a concentração dos dados em torno da média, ou seja, o quão "pontiaguda" ou "achatada" a distribuição é em comparação com uma distribuição normal.

- Uma distribuição com curtose igual à da normal (geralmente 3 quando usamos a definição tradicional) é considerada mesocúrtica.
- Se a curtose for maior que 3, a distribuição é leptocúrtica (mais pontiaguda e com caudas mais pesadas).
- Se a curtose for menor que 3, a distribuição é platicúrtica (mais achatada e com caudas mais leves).

Curtose

Quarto momento amostral: Calculado de forma semelhante à assimetria

$$k = m_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$$

Excesso de curtose: Como a distribuição normal possui uma curtose igual a 3, é comum calcular o excesso de curtose, que é a diferença entre a curtose da amostra e 3:

$$k - 3$$

- Excesso positivo: indica caudas mais pesadas do que as da normal (distribuição leptocúrtica).
- Excesso negativo: indica caudas mais leves (distribuição platicúrtica).

Curtose

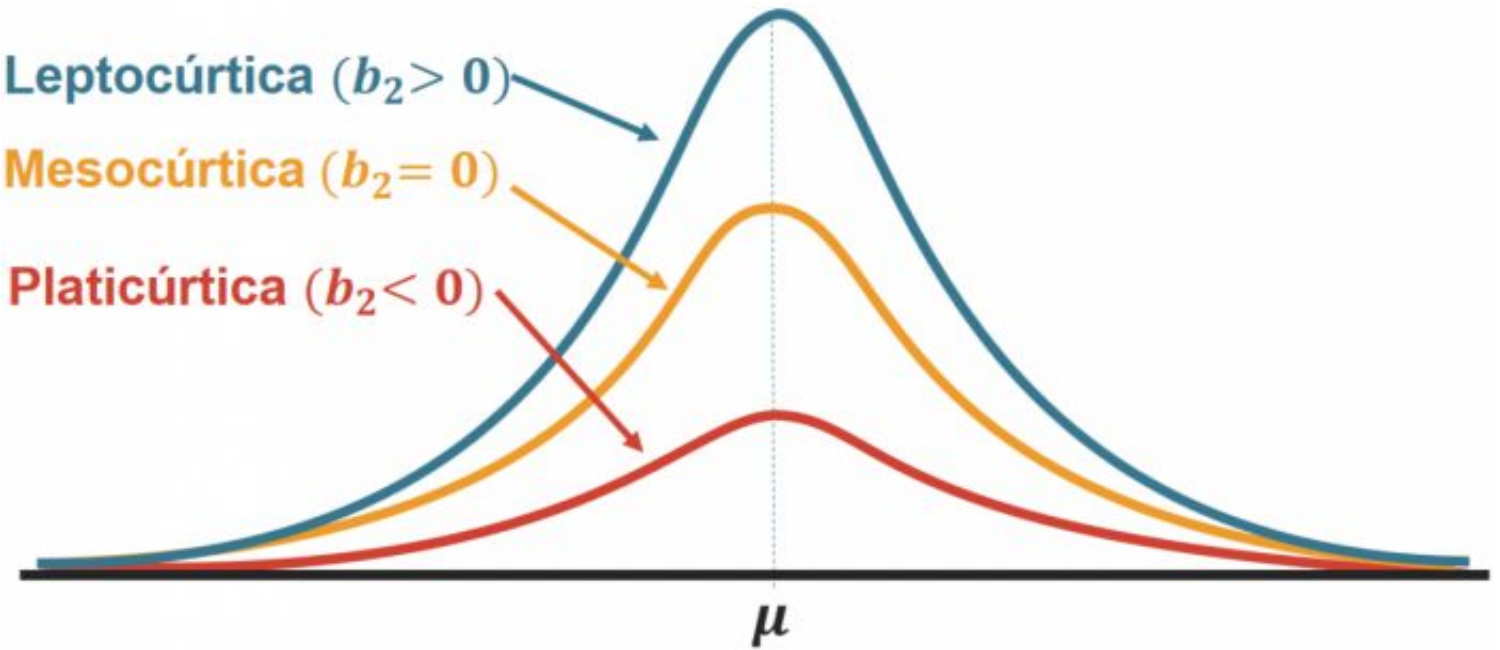
Detecção de Valores Extremos:

- Auxilia na identificação de outliers e valores atípicos no conjunto de dados
- Uma alta curtose indica maior probabilidade de valores extremos na distribuição
- Distribuições com curtose elevada sugerem maior risco de eventos extremos

Normalidade dos Dados:

- Auxilia na avaliação da normalidade quando combinada com outras métricas estatísticas
- Influencia a escolha de métodos estatísticos apropriados para análise
- Pode indicar necessidade de normalização dos dados através de transformações específicas

Curtose



Curtose *vs* Assimetria

A curtose e a assimetria são medidas complementares que descrevem a forma de uma distribuição de dados, cada uma focando em aspectos diferentes:

Assimetria:

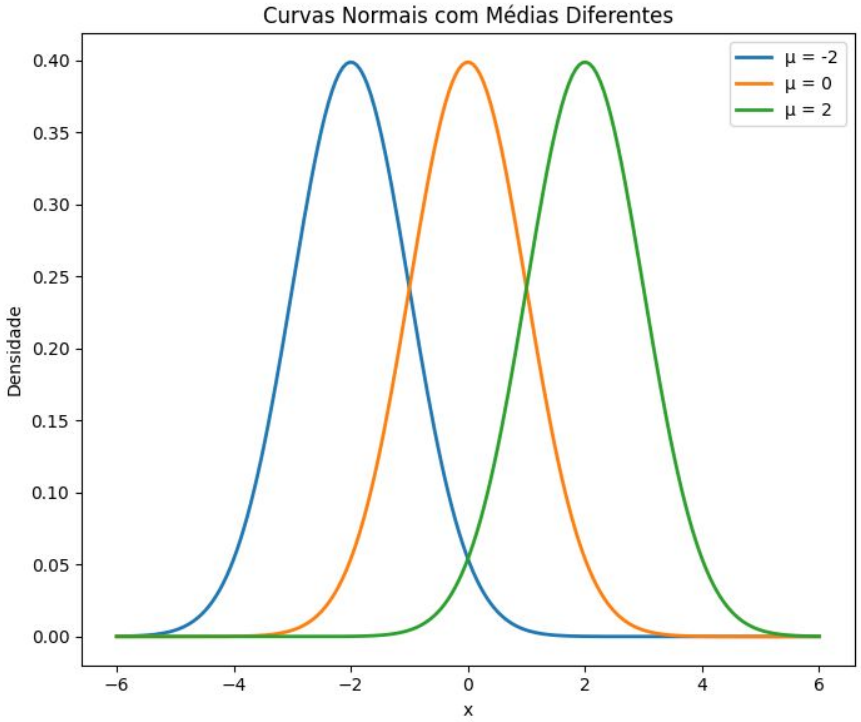
- Mede o grau de desvio lateral da distribuição
- Indica se os dados tendem mais para um lado ou outro da média
- Foca na direção do alongamento da distribuição

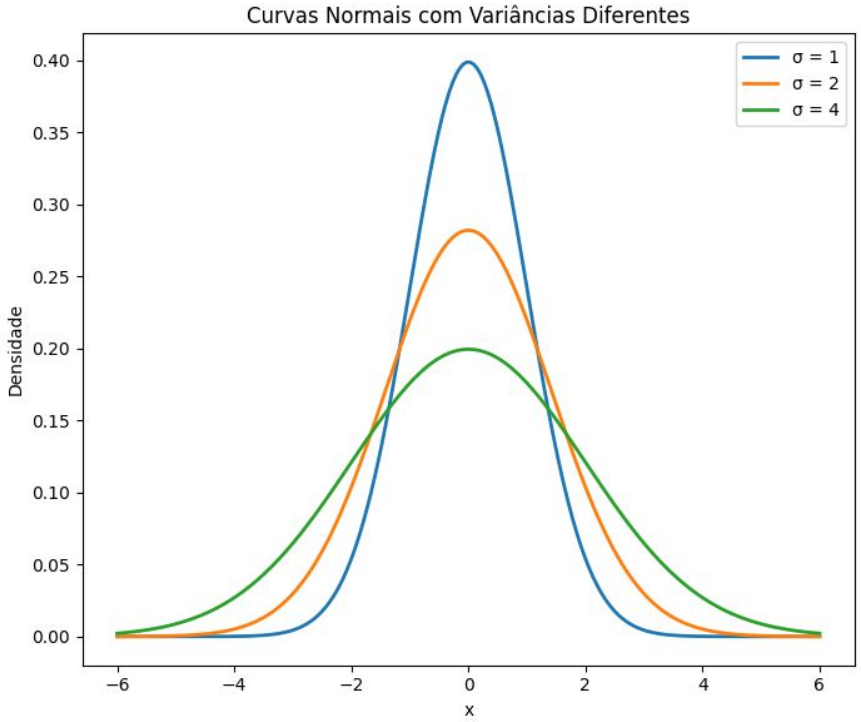
Curtose:

- Avalia a concentração de dados nas caudas da distribuição
- Mede o "achatamento" ou "alongamento" vertical da curva
- Analisa especificamente a quantidade de observações nas extremidades, independente do lado em que se encontram

Análise paramétrica

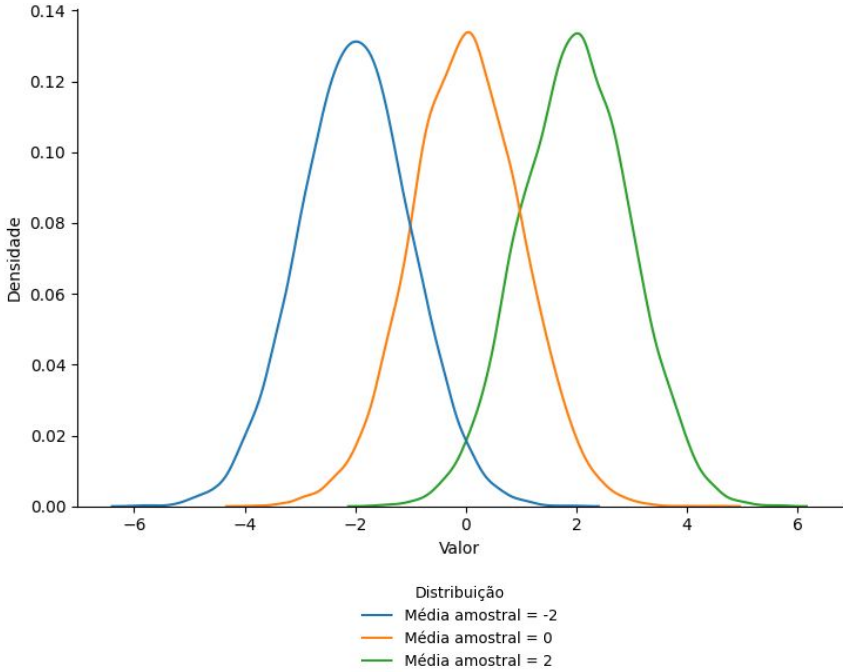
Curvas normais

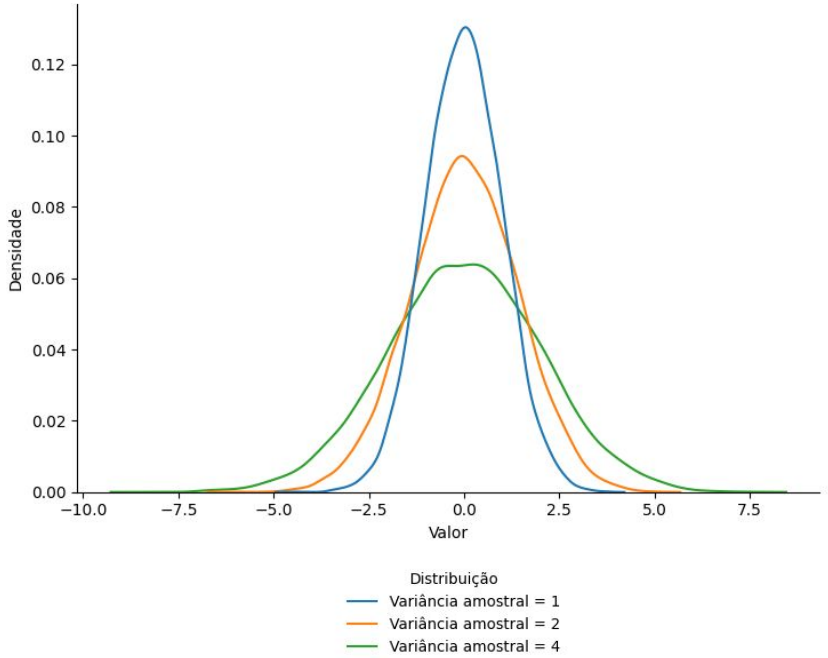




Análise paramétrica

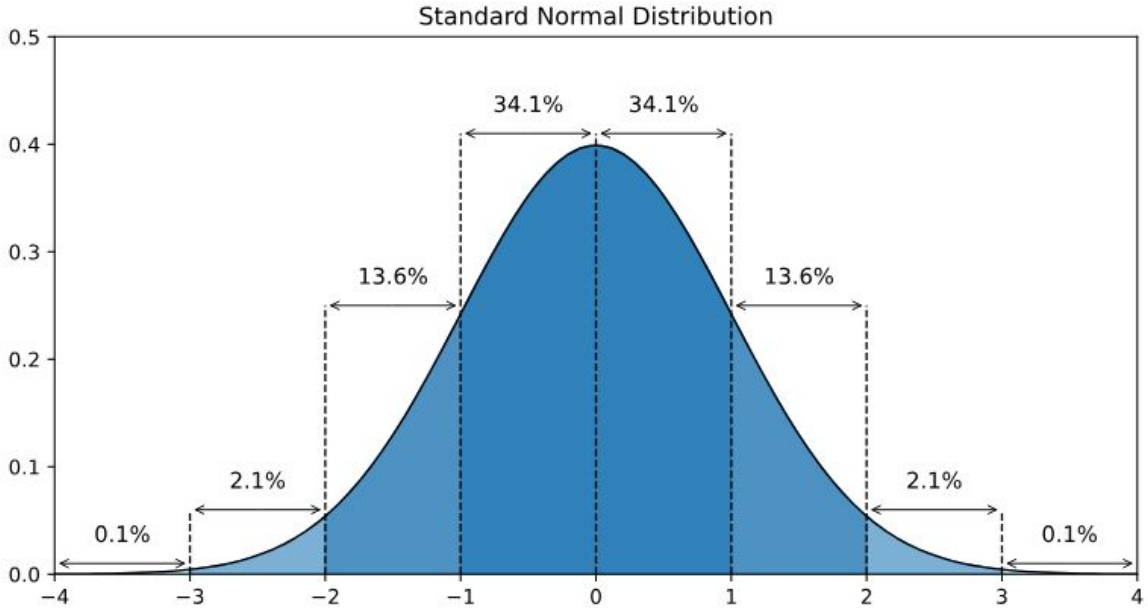
Amostras aproximadamente normais





Análise paramétrica

Distribuição padrão: trata-se de um caso particular de distribuição normal, onde **μ**=0 e **σ**=1



Prática

Exercícios

1. 4. Agora, calcule as medidas de assimetria e curtose da variável analisada;

2. Anote em seu caderno as respectivas interpretações das medidas encontradas;

3. Discorra sobre a existência, ou não, de outliers e sobre uma forma de eliminá-los com base nas medidas de dispersão e forma disponíveis;

Obrigado

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