Estatística Descritiva

Stefano Mozart

12/02/2025

Sumário

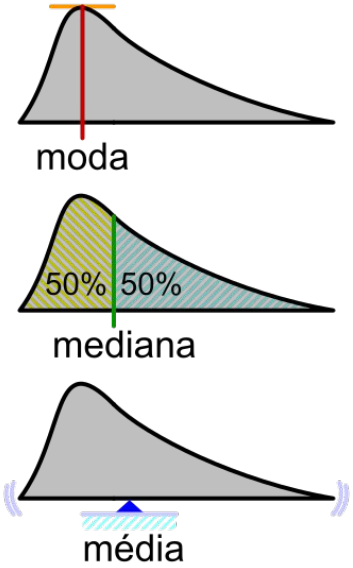
- Medidas de tendência central
- Medidas separatrizes

Tendência central

Tendência central

A tendência central é um conceito estatístico que procura representar uma amostra através de um único valor que sintetize o "centro" ou o ponto ao redor do qual os valores observados se distribuem. Esse valor serve como um resumo que facilita a compreensão e a comparação de diferentes amostras ou subconjuntos.

- Entre as medidas de tendência central, a mediana é um caso especial, pois é também uma medida de posição: ela indica o valor que separa a metade inferior da metade superior dos elementos (o 50º percentil) de uma amostra ordenada.



Média

É a medida de tendência central mais comum, intuitiva e amplamente conhecida. No entanto, é bastante sensível a valores extremos (*outliers*), o que pode comprometer sua qualidade como representação do centro dos valores da amostra.

- **Média Aritmética:**
  - A soma de todos os valores observados, dividida pelo número de elementos na amostra.
  - Apropriada para valores em escala natural e sem a presença de *outliers* (extremos) que possam distorcer seu valor.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Média

- **Geométrica:**
  - A raiz *n*-ésima do produto dos *n* valores;
  - Indicada para valores em escalas crescentes, ou proporções e percentuais;
  - Mais resistente a *outliers*;
  - Exemplos: aumento médio de preços (inflação), crescimento médio da dívida pública.

$$\left(\prod_{i=1}^{n} a_i\right)^{1/n} = \sqrt[n]{a_1 a_2 \cdots a_n}.$$

- **Harmônica:**
  - Inverso da média aritmética dos inversos dos valores;
  - Indicada para amostras de taxas ou proporções.
  - Exemplos: rendimento médio de aplicação, proporção média entre área construída e área total de imóveis.

$$\frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

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.

Moda

É a observação (ou classe) com maior frequência (maior número de ocorrências na amostra).

- Caso duas ou mais classes apresentem a mesma frequência, a amostra pode ser considerada bimodal ou multimodal.
- Caso todas as classes tenham a mesma frequência, a amostra é considerada amodal.
- Em muitos casos, no entanto, a amostra é dita ‘amodal’ pela inexistência de uma classe com frequência absoluta maior que todas as outras;

Mediana

É o valor que ocupa a posição central de uma amostra ordenada. De modo que 50% dos valores observados na amostra estão abaixo da mediana, e 50% acima.

- Se o número de observações for ímpar, a mediana é o valor central;
- Se for par, é a média dos dois valores centrais.

Média

Exemplo de fragilidade da média.

- Média da receita: R$ 77.545,45;
- Média de anos: 4,81 anos;

Investidor	Anos de pregão	Receita último período (R$)
Obadias	1	21.000,00
Ageu	2	38.000,00
Naum	3	47.000,00
Jonas	4	48.000,00
Sofonias	3	53.000,00
Malaquías	4	55.000,00
Habacuque	3	56.000,00
Joel	3	73.000,00
Miquéias	7	105.000,00
Amós	9	146.000,00
Zacarías	14	211.000,00

Média

Exemplo de fragilidade da média.

- Média da receita: R$ 1.785.500,00;
- Média de anos: 9.91 anos;

Investidor	Anos de pregão	Receita último período (R$)
Obadias	1	21.000,00
Ageu	2	38.000,00
Naum	3	47.000,00
Jonas	4	48.000,00
Sofonias	3	53.000,00
Malaquías	4	55.000,00
Habacuque	3	56.000,00
Joel	3	73.000,00
Miquéias	7	105.000,00
Amós	9	146.000,00
Zacarías	14	211.000,00
Isaías	66	1.292.000,00

Média

Exemplo de fragilidade da média.

- Receita:
  - Média: R$ 1.785.500,00;
  - Média geométrica: R$ 81.792,76
  - Média harmônica: R$ 58.101,04
  - Mediana: R$ 55.500,00;
  - Moda: indefinida/amodal;
- Anos:
  - Média: 9.91 anos;
  - Média geométrica: 4.8 anos;
  - Média harmônica: 3.26 anos;
  - Mediana: 3.5 anos;
  - Moda: 3 anos;

Investidor	Anos de pregão	Receita último período (R$)
Obadias	1	21.000,00
Ageu	2	38.000,00
Naum	3	47.000,00
Jonas	4	48.000,00
Sofonias	3	53.000,00
Malaquías	4	55.000,00
Habacuque	3	56.000,00
Joel	3	73.000,00
Miquéias	7	105.000,00
Amós	9	146.000,00
Zacarías	14	211.000,00
Isaías	66	1.292.000,00

Obs.: MA > MG > MH

Medidas populacionais *vs* amostrais

- Medidas Populacionais: Calculadas com dados de toda a população.
  - Exemplo: Média populacional (μ), Variância populacional (σ^2).
  - Vantagem: Precisão, pois utiliza todos os dados.
  - Limitação: Muitas vezes impraticável ou custosa de se obter.

- Medidas Amostrais: Calculadas com base nos dados da amostra.
  - Exemplo: Média amostral ($\bar{x}$), Variância amostral (s^2).
  - Vantagem: Viabilidade prática e redução de custos.
  - Limitação: Estimativas sujeitas a erro de amostragem e variabilidade.

Erro amostral

É a discrepância entre um parâmetro populacional (como a média real de uma população) e a estimativa desse parâmetro obtida a partir de uma amostra.

- Reflete a variabilidade natural do processo de amostragem;

- **Erro tolerável (E_0)**: a margem máxima de erro que se aceita em um processo de medição ou estimativa sem comprometer a qualidade, a confiabilidade ou a tomada de decisão. Essa tolerância é definida com base em critérios técnicos, normativos ou práticos e serve como um limite para avaliar se um.
- Tamanho mínimo da amostra:
  - Aproximação n_0 (sem conhecimento do tamanho da população): $n_0 = \frac{1}{E_0^2}$
  - Refinamento n (com tamanho da população N definido): $n = \frac{N.n_0}{N + n_0}$

Erro médio

Medida que representa a diferença, em média, entre os valores observados (ou medidos) e os valores esperados (ou reais). Em outras palavras, ele indica o viés ou a precisão média de um conjunto de medições ou estimativas.

- **Erro médio absoluto**: utiliza o valor absoluto das diferenças, evitando que erros positivos e negativos se cancelem;

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - p_i|$$

- **Erro quadrático médio:** média aritmética do desvio quadrático de cada observação.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Prática

Exercícios:

1. Escolha uma variável contínua em sua amostra e identifique sua moda, média e mediana.
    1.1. Anote no caderno de análise se essas medidas são coincidentes e qual a interpretação associada (tanto para o caso afirmativo, quanto negativo);
    1.2. Identifique também as médias geométrica e harmônica, discorrendo a respeito de qual delas apresenta uma melhor informação acerca da amostra.

2. Exiba um histograma da variável escolhida, acrescentando:
    2.1. Uma linha vertical azul para indicar a moda;
    2.2. Uma linha vertical verde para marcar a média;
    2.3. E uma linha vertical cinza para indicar a mediana.

Medidas Separatrizes

Medidas separatrizes

Valores limítrofes das posições que dividem uma amostra ordenada em partes iguais, auxiliando na compreensão da distribuição dos dados.

- Ajudam a identificar a posição relativa de cada observação dentro da distribuição, facilitando a análise da dispersão e a identificação de assimetrias e outliers.

- **Quartis:** dividem uma amostra ordenada em quatro partes iguais.
  - Primeiro quartil (Q1): valor que delimita os 25% menores valores observados
  - Segundo quartil (Q2): coincide com a mediana, representando 50% dos elementos
  - Terceiro quartil (Q3): valor que delimita os 75% dos dados

- **Percentis:** dividem uma amostra ordenada em 100 partes iguais. Por exemplo, quando um valor está no percentil 75, significa que ele é maior que 75% dos dados da distribuição e menor que 25% dos valores restantes.

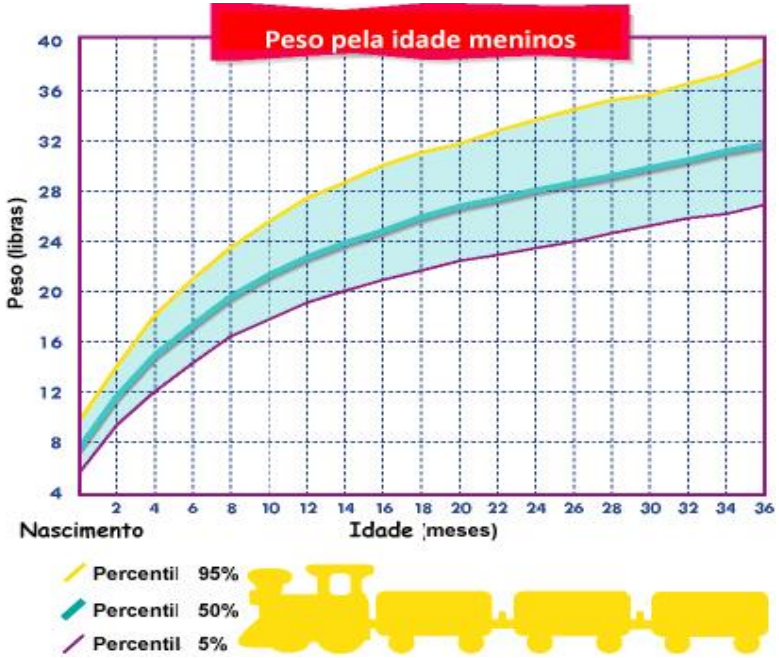
Medidas separatrizes

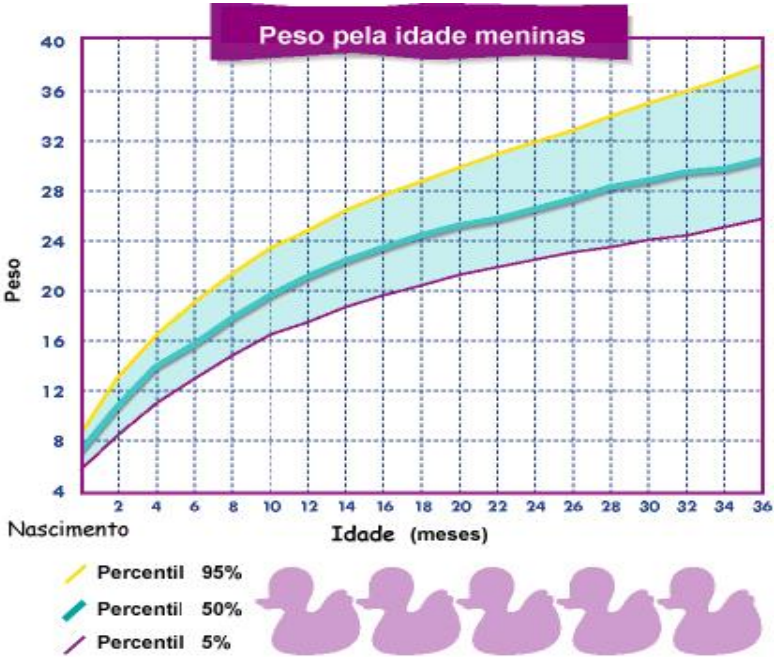
- **Quantis:** generalização dos quartis e percentis, sendo valores que dividem um conjunto ordenado de dados em partes iguais. Assim:
  - Quartis são quantis que dividem em 4 partes;
  - Percentis são quantis que dividem em 100 partes;
  - Decis são quantis que dividem em 10 partes;

- **Aplicações:**
  - Análise da distribuição de elementos na amostra;
  - Identificação de valores extremos;
  - Comparação de posições relativas;

Medidas Separatrizes

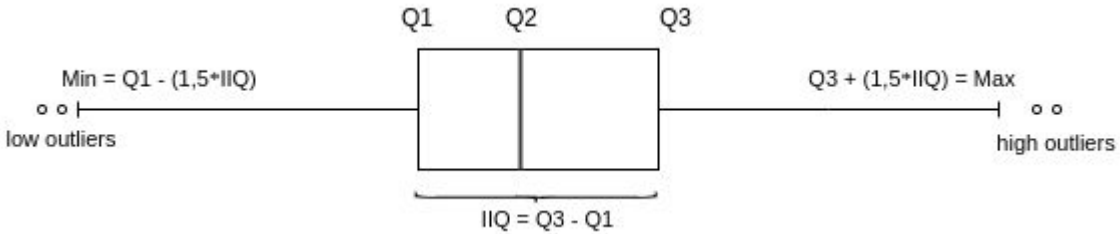
Exemplo de uso do percentil





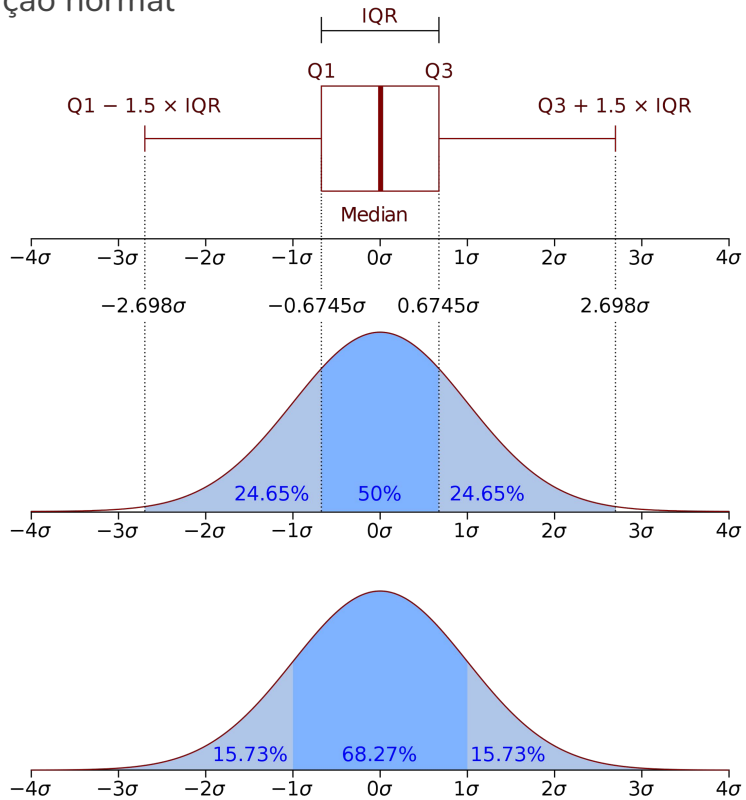
Boxplot

- O *boxplot* é uma ferramenta visual, utilizada para informar a distribuição das observações de uma amostra ao longo dos quartis.



Medidas separatrizes

Relação entre boxplot e distribuição normal



Prática

Tópicos para discussão:

- Escolha uma variável quantitativa em seu dataset e exiba um boxplot.
- Identifique a existência, ou não, de *outliers*
- Compare o primeiro e terceiro quartil de sua amostra, discorra em seu cader sobre os achados

Fontes de dados:

- https://dados.gov.br/dados/conjuntos-dados
  (exige autenticação com Gov.br)
- https://sidra.ibge.gov.br/
- https://basedosdados.org
- Sugestões?

Obrigado

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