UFRRJ
UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

Especialização em Análise de Dados como Método de Apoio às Políticas Públicas

Análise Exploratória e Visualização de Dados

Marcello Montillo Provenza

Bibliografia Recomendada

Bibliografia Recomendada

LEVIN, Jack. **Estatística Aplicada a Ciências Humanas.** Harbra, 2ª ed. 1987.

VIEIRA, Sônia. **Elementos de Estatística.** São Paulo: Atlas, 5ª ed. 2012.

DIETZ, Thomas; KALOF, Linda. **Introdução a Estatística Social: a Lógica do Raciocínio Estatístico.** Rio de Janeiro: LTC, 2015.

Conceitos Básicos

Conceitos Básicos

Estatística: utiliza-se através das teorias probabilísticas para explicar a frequência de fenômenos e possibilitar a previsão desse acontecimentos no futuro.

Conceitos Básicos

População (N): é o conjunto de todos os valores (indivíduos ou objetos) que descrevem o fenômeno que interessa ao pesquisador.

Censo: é uma pesquisa na qual todos os elementos populacionais são investigados, de tal forma que os valores da variável de interesse tornem-se conhecidos para toda a população.

Amostra (n): é um subconjunto (uma parcela) de elementos pertencentes a uma população. Uma amostra precisa de ser representativa, não viciada, aleatória e ampla.

Conceitos Básicos

Parâmetro: grandeza mensurável que permite apresentar as características principais de uma população (média, variância, desvio padrão etc.).

Estimativa: é o processo que consiste no uso de dados amostrais para estimar valores de parâmetros populacionais desconhecidos, tais como média, desvio padrão, proporções etc.

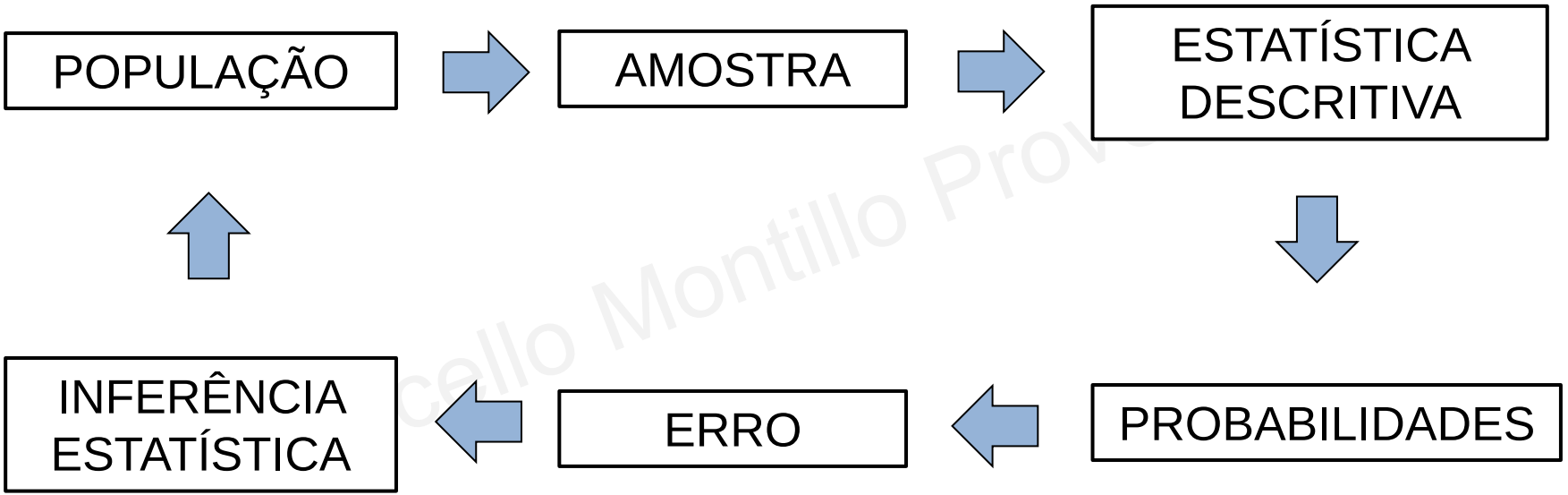
Conceitos Básicos

Na estatística existem dois ramos:

1) Estatística Indutiva: que a partir de uma amostra da população, permite estender os resultados à população inteira. Trata de estabelecer conclusões relativas a população com base na amostra.

2) Estatística Descritiva: que visa descrever o real de forma a permitir entendê-lo melhor. Trata da recolha, organização e tratamento dos dados. O seu objetivo é informar, prevenir e esclarecer.

Conceitos Básicos



Conceitos Básicos

Representação:

A representação das medidas diferem da população para amostra. Na população são utilizadas letras gregas, já para amostra são utilizadas letras latinas.

Medidas	População	Amostra
Média	μ	x
Desvio padrão	σ	s
Proporção	π	p

Conceitos Básicos

Uma amostra precisa ser:

- ✓ Representativa: deve conter indivíduos de todos os extratos da população;
- ✓ Não viciada: o número de elementos de cada extrato deve ser proporcional à população desse extrato;
- ✓ Aleatória: em cada extrato os indivíduos devem ser escolhidos ao acaso;
- ✓ Ampla: deve ser bastante alargada para poder representar características semelhantes às da população total que pretende representar.

Coleta e Organização de Dados

Tipos de Dados

Tipos de Dados

Dados são informações coletadas por meio de observações, medições ou pesquisas.

Eles são a base para análise estatística e tomada de decisão em diversas áreas, tais como saúde, educação, segurança, economia, entre outras.

Tipos de Dados

1. Dados Qualitativos (ou Categóricos): são aqueles que expressam qualidades, características ou categorias, e não podem ser medidos numericamente.

a) Nominais: não possuem ordem natural, por exemplo: cor dos olhos (azul, verde, castanho); estado civil (solteiro, casado, divorciado).

b) Ordinais: possuem uma ordem ou hierarquia, mas não têm intervalos numéricos definidos, por exemplo: nível de escolaridade (fundamental, médio, superior); grau de satisfação (ruim, regular, bom, ótimo).

Tipos de Dados

2. Dados Quantitativos (ou Numéricos): são expressos por números e podem ser medidos.

a) Discretos: assumem valores inteiros, geralmente resultados de contagem, por exemplo: número de filhos; número de habitantes em uma cidade; chamadas atendidas por dia.

b) Contínuos: podem assumir qualquer valor dentro de um intervalo, geralmente resultados de medição, por exemplo: altura; peso; temperatura; tempo de reação em segundos.

Tipos de Dados

Tipo de Dado	Subtipo	Exemplos
Qualitativo	Nominal	Gênero, cor, profissão
	Ordinal	Grau de instrução, nível de dor
Quantitativo	Discreto	Número de filhos, votos em uma urna
	Contínuo	Peso corporal, velocidade, altura

Exemplos

1. Educação: Coletar dados sobre o nível de escolaridade (ordinal) da população ajuda a planejar políticas de alfabetização e expansão do ensino superior.

2. Saúde: Em campanhas de vacinação, dados como idade (quantitativo contínuo) e tipo sanguíneo (qualitativo nominal) são essenciais para grupos prioritários.

3. Segurança Pública: A contagem de ocorrências criminais (quantitativo discreto) e a classificação de áreas por grau de vulnerabilidade (qualitativo ordinal) orientam ações estratégicas.

Banco de Dados

Banco de Dados

Alguns termos típicos:

- Dados: fatos que podem ser armazenados;

  Ex: nomes, telefones, endereços etc.;

- Base de dados: coleção de dados inter-relacionados logicamente;

  Ex: Censo, Serasa, agenda de telefones etc.;

- Sistema de Gerência de Bases de Dados (SGBD): coleção de programas que permite a criação e gerência de bases de dados ou Sistema de Banco de Dados;

  Ex: SQL-Server, Oracle, MySQL.

Banco de Dados

- **Banco de dados** (ou **base de dados**) é um conjunto de registros dispostos em estrutura regular que possibilita a reorganização dos mesmos e produção de informação. Um banco de dados normalmente agrupa registros utilizáveis para um mesmo fim;

- Coleção de dados relacionados, onde dados significam fatos conhecidos que podem ser logicamente armazenados e que possuem significado implícito;

- É um conjunto de informações organizadas sobre pessoas em geral, fatos, objetos, empresas, mercadorias, serviços, domicílios etc.

Banco de Dados

- Pode ser gerado e mantido por meio manual ou magnético;
- Restrito quando gerado por meios manuais;
- Formado por variáveis qualitativas e quantitativas;
- Microdados e dados agregados;
- Alocados por casos ou por associação de casos, que formam um subconjunto;
- Diferença entre banco de dados e surveys;

Banco de Dados

Survey é uma pesquisa que permite a obtenção de dados ou informações sobre características, ações e opiniões de um determinado grupo de pessoas. Pode ser um levantamento, uma sondagem ou um diagnóstico.

1. Estudo de pesquisa de mercado (market research) que formula perguntas a fim de receber informação sobre atitudes, motivos e opiniões. Esses estudos podem ser feitos frente a frente, pelo telefone ou pelo correio.

2. Coleta de dados por amostragem, por exemplo, questionando um grande grupo de cônjuges sobre os fatores da felicidade conjugal; sondagem de opinião pública.

Banco de Dados

Propriedades implícitas de um banco de dados:

- Representa algum aspecto do mundo real, por vezes chamado de universo ou população;
- Coerentemente organizado;
- Deve permitir a recuperação automática da informação;
- É construído e povoado com dados para uma proposta específica;
- É constituído, sobretudo, para gerar controle sobre determinadas coisas, processos ou pessoas.

Banco de Dados

Qual a utilidade de um banco de dados?

- Controle;
- Gerar soluções;
- Distribuir efetivos;
- Alocar ou realocar recursos;
- Construir aproximações da realidade;
- Subsídio para políticas públicas e pesquisas científicas;
- Controle de estoques e serviços;
- Acompanhamento de metas por meio de indicadores;
- Gerar relatórios diversos, entre outros.

- Não interessam os casos individuais e sim os casos em conjunto;
- Casos isolados não revelam nada.

Banco de Dados

Microdado: é a menor unidade de observação de um conjunto de dados. Neste tipo de arranjo é possível separar um caso de seu conjunto e observá-lo linearmente, independente dos demais. Pode ser trabalhado desagregadamente.

Variável: *sujeito a variar, mudável, volúvel, inconstante.*

- Em estatística, é um atributo, mensurável ou não, sujeito à variação quantitativa ou qualitativa, no interior de um conjunto;
- Em pesquisa, são atributos lógicos de um conceito operacional.

Variável aleatória: é uma variável tal que não sabemos ao certo que valor tomará, mas para qual podemos calcular a probabilidade de tomar determinado valor.

Banco de Dados

Exemplos do dia-a-dia

- Agenda telefônica
- Lista de compras

Exemplos consolidados

- Censo
- PNAD
- POF
- SPC
- Serasa
- Cadastros de clientes e fornecedores

Precisão e Arredondamento

Precisão

Precisão: é grau de detalhamento ou exatidão com que um valor numérico é registrado. Isso depende do instrumento de medida utilizado e da forma como o dado é coletado e registrado.

Exemplo: se medimos uma régua com um instrumento que mostra milímetros, podemos registrar 23,76 cm (mais preciso). Já com uma régua comum, podemos registrar apenas 24 cm (menos preciso).

Arredondamento

Por que Arredondar?

Nem sempre é necessário ou útil apresentar os dados com muitos dígitos. O **arredondamento** serve para:

- Facilitar a leitura e a comunicação dos dados;
- Evitar falsa impressão de exatidão;
- Tornar os cálculos mais práticos (especialmente em situações onde a precisão extrema não é necessária).

Arredondamento

Para arredondar um número decimal, observamos o dígito à direita da casa decimal desejada:

Dígito a ser descartado	Ação
0, 1, 2, 3 ou 4	Arredonda para baixo (mantém o número)
5, 6, 7, 8 ou 9	Arredonda para cima (acrescenta 1)

Arredondamento

Exemplo 1: arredondar 3,746 para duas casas decimais.

- Observa-se o terceiro dígito (6) ⇒ Arredonda para cima → 3,75

Exemplo 2: arredondar 2,432 para uma casa decimal.

- Observa-se o segundo dígito (3) ⇒ Arredonda para baixo → 2,4

Arredondamento e Truncamento

Arredondamento: considera o próximo dígito para decidir se sobe ou mantém.

Truncamento: simplesmente corta os dígitos sem considerar os seguintes.

Exemplo: o valor 3,786.

- Arredondando para 2 casas → 3,79
- Truncando para 2 casas → 3,78

Exemplo 1

Planejamento Orçamentário em Educação

Durante a elaboração de políticas para a distribuição de verbas escolares, a média de alunos por turma em uma rede municipal foi calculada como 28,73 alunos.

Para facilitar a comunicação com gestores e a alocação de recursos, o valor foi arredondado para 29 alunos por turma, respeitando a margem aceitável de variação.

Exemplo 2

Análise de Taxas de Criminalidade

Em um estudo de segurança pública, uma cidade apresentou uma taxa de homicídios de 12,457 por 100 mil habitantes.

Para fins de divulgação e comparação com metas do plano de segurança, o dado foi arredondado para 12,5 homicídios por 100 mil, mantendo precisão suficiente sem sobrecarregar a leitura técnica.

Exemplo 3

Monitoramento da Taxa de Desemprego

Uma política de geração de emprego está sendo avaliada com base na taxa de desemprego, que foi registrada em uma pesquisa como 8,049%.

Para apresentação em relatórios executivos e na mídia, o índice foi arredondado para 8,0%, facilitando a compreensão pública sem comprometer a análise da tendência.

Tabelas

Elementos e Normas de Tabelas

Tabelas

As tabelas são instrumentos fundamentais para a apresentação organizada de dados, facilitando a análise, a comparação e a comunicação de informações de forma clara e objetiva.

Uma tabela estatística completa e bem estruturada deve conter os seguintes elementos: número da tabela; título (ou cabeçalho); cabeçalho das colunas; corpo da tabela (ou campo/área de dados); fonte; e, se necessário, notas e observações.

Elementos de Tabelas

Número da Tabela: sempre posicionado acima da tabela. Usado para identificação e referência no texto (por exemplo: Tabela 1, Tabela 2 etc.).

Título: posicionado logo após o número da tabela. Deve ser claro, conciso e explicativo, indicando o conteúdo, a localização, o tempo e a unidade de medida (se aplicável).

Cabeçalho das Colunas: indica o tipo de dado apresentado em cada coluna. Deve conter unidades de medida (quando necessário) e deve ser alinhado com o conteúdo da coluna.

Elementos de Tabelas

Corpo da Tabela: é o conjunto de linhas e colunas onde os dados estão organizados. As informações devem ser precisas, legíveis e ordenadas logicamente.

Fonte: deve ser informada abaixo da tabela e indica a origem dos dados.

Notas e Observações: explicações adicionais, siglas ou símbolos os quais, se necessário, devem ser usados na tabela.

Normas para Apresentação de Tabelas

As tabelas devem seguir padrões formais, principalmente quando usadas em trabalhos acadêmicos ou relatórios técnicos.

Normas da ABNT (NBR 14724 e NBR 6029)

- A tabela deve ser autoexplicativa;
- Não se usa traços verticais (|);
- O espaçamento interno deve ser suficiente para garantir a leitura clara;
- Todas as tabelas devem ser citadas no texto;
- Usar notação padronizada de números (ponto para milhar e vírgula para decimal no Brasil, conforme ABNT).

Tabelas Eficientes

Mantenha o layout limpo e organizado.

Use destaque (negrito, sombreado leve) apenas quando necessário para chamar atenção a dados importantes.

Evite excesso de informação (prefira dividir em mais de uma tabela se houver muitos dados).

Exemplo

Tabela 1 – Número de matrículas em cursos de graduação, por modalidade e turno, na Universidade XYZ – 2023.

Modalidade	Turno	Matrículas
Presencial	Diurno	1.200
Presencial	Noturno	1.450
EaD	Virtual	980

Fonte: Universidade XYZ, Sistema Acadêmico, 2023.

Séries Estatísticas

Séries Estatísticas

Distribuição de Frequência

Em Estatística, distribuição de frequência é um arranjo de valores que uma ou mais variáveis tomam em uma amostra. Cada entrada na tabela contém a frequência ou a contagem de ocorrências de valores dentro de um grupo ou intervalo específico, e deste modo, a tabela resume a distribuição dos valores da amostra.

Séries Estatísticas

Distribuição de Frequência

Dados Brutos é o conjunto dos dados que ainda não foram numericamente organizados.

Dados brutos da taxa de Colesterol (mg/dl) em pacientes internados									
248	157	124	124	215	312	254	156	132	145
214	256	258	298	189	178	186	231	301	265
298	178	196	152	144	185	132	289	264	256

Séries Estatísticas

Distribuição de Frequência

Rol é o conjunto organizado dos dados brutos por ordem de valor, podendo ser crescente ou decrescente.

ROL da taxa de Colesterol (mg/dl) em pacientes internados									
124	124	132	132	144	145	152	156	157	178
178	185	186	189	196	214	215	231	248	254
256	256	258	264	265	289	298	298	301	312

Séries Estatísticas

Distribuição de Frequência Agrupada

Taxa de colesterol (mg/dl) em pacientes internados

Colesterol (mg/dl)	PM_i	F_i	Fa_i	Fr_i (%)	Far_i (%)
122 ⊢ 154	138	7	7	23,3	23,3
154 ⊢ 186	170	5	12	16,7	40,0
186 ⊢ 218	202	5	17	16,7	56,7
218 ⊢ 250	234	2	19	6,7	63,3
250 ⊢ 282	266	6	25	20,0	83,3
282 ⊢ 314	298	5	30	16,7	100,0
T O T A L	-	30	-	100,0	-

Séries Estatísticas

Elementos da Distribuição de Frequência Agrupada

Classe (i): é o intervalo.

Ponto Médio (PM_i): é o ponto que divide o intervalo de classe em duas partes iguais.

Frequência Simples Absoluta (F_i): é a quantidade de observações em cada uma das classes.

Frequência Acumulada Absoluta (Fa_i): é a soma das Frequências Simples Absolutas (F_i).

Séries Estatísticas

Elementos da Distribuição de Frequência Agrupada

Frequência Relativa (Fr_i): apresenta a relação percentual de cada quantidade observada na Frequência Simples Absoluta (F_i) em relação ao total de dados observados, ou seja, é a relação de cada grupo de observações da classe para com o total de observações (n).

$$Fr_i = \frac{f_i}{n}$$

Frequência Acumulada Relativa (Far_i): é a soma das Frequências Relativas (Fr_i).

Séries Estatísticas

Para dados qualitativos (nominais ou ordinais), as tabelas de frequência permitem resumir os dados por categorias.

Para dados quantitativos (discretos ou contínuos), a tabela de distribuição de frequências agrupa os dados em valores únicos (dados discretos) ou em **classes** (dados contínuos).

Séries Estatísticas

Exemplo: Uma prefeitura deseja avaliar a demanda por atendimento médico em uma Unidade Básica de Saúde (UBS) ao longo de um trimestre. Para isso, coletou dados sobre o número de consultas realizadas por paciente durante o período. Os dados são quantitativos discretos, pois representam contagens inteiras (0, 1, 2, ...).

Dados coletados

0	1	2	2	3	1	0	1	2	2
4	3	1	1	0	2	2	3	4	1
2	1	2	3	2	2	0	1	2	3

Séries Estatísticas

Número de Consultas Médicas por Pessoa em uma Unidade de Saúde

Nº de Consultas	Frequência Absoluta (F_i)	Frequência Relativa (Fr_i (%))	Frequência Acumulada (Fa_i)
0	3	10,0	3
1	8	26,7	11
2	11	36,6	22
3	6	20,0	28
4	2	6,7	30
Total	30	100,0	—

Séries Estatísticas

Interpretação:

- A maioria dos pacientes (36,6%) realizou duas consultas no período;
- 10,0% não realizaram nenhuma consulta, o que pode indicar ausência de necessidade ou dificuldade de acesso.

Esses dados podem orientar políticas de agendamento, ampliação de equipe médica ou ações de busca ativa para pacientes que não estão acessando o serviço.

Análise Exploratória de Dados

Análise Exploratória de Dados

Às vezes observamos ou coletamos dados com um objetivo específico em vista. Outras vezes, não há qualquer objetivo específico; apenas desejamos explorar os dados para ver o que eles nos revelam.

- ✓ Explorar os dados em um nível preliminar;
- ✓ Poucas (ou talvez nenhuma) hipóteses são feitas sobre os dados;
- ✓ Costuma exigir cálculos gráficos relativamente simples.

Análise Exploratória de Dados

Compreendendo a análise percentual

Como organizar os dados e transformá-los em um conjunto-resumo fácil de entender?

- ✓ Distribuir os dados por frequências;
- ✓ Método percentual de padronização das frequências quanto ao tamanho;
- ✓ Cálculo (Borges & Dirk, 2006: 123-4):

$$\% = \frac{Frequência}{N} * 100$$

Análise Exploratória de Dados

Exemplo:

Tabela 1: População residente por região do Brasil - 2010.

Região	População	%
Norte	15.864.454	8,3
Nordeste	53.081.950	27,8
Sudeste	80.364.410	42,1
Sul	27.386.891	14,4
Centro-Oeste	14.058.094	7,4
Total	190.755.799	100,0

Fonte: IBGE, Censo Demográfico 2010.

Análise Exploratória de Dados

Compreendendo as taxas

- ✓ Razão entre duas quantidades, onde o numerador está contido no denominador;
- ✓ Método padronizador para análises comparativas;
- ✓ Cálculo:
  - ✓ Pegue o indicador principal (crimes, nascimentos, mortes, doença etc.);
  - ✓ Divida pela população (ou outra variável de interesse);
  - ✓ Multiplique o resultado por “K” (1.000, 10.000, 100.000 etc.).

$$\text{Taxa} = \frac{Indicador}{Variável} * K$$

Análise Exploratória de Dados

Exemplo:

Tabela 2 – Número de Vítimas de Homicídio Doloso nas Cidades do Rio de Janeiro e Vitória – valores absolutos e taxa por 100 mil habitantes - 2001

Municípios	Homicídios	População	Taxa por 100.000 hab.
Rio de Janeiro	2.098	5.893.258	35,6
Vitória	187	295.886	63,2

Fonte: SENASP - MJ

Análise Exploratória de Dados

Gráficos: são representações visuais dos dados estatísticos que devem corresponder aos elementos e variáveis tratados nas análises:

- ✓ Bem elaborado, o gráfico informa de maneira rápida e eficiente o leitor ou analista;
- ✓ Se mal construído, pode desinformar mais do que esclarecer;
- ✓ Por sua qualidade intuitiva e visual, que permite instruir de maneira sucinta e estética, os gráficos são muito mais comuns no cotidiano do que as tabelas, que requerem leitura mais retida e demorada sobre o assunto em questão.

Análise Exploratória de Dados

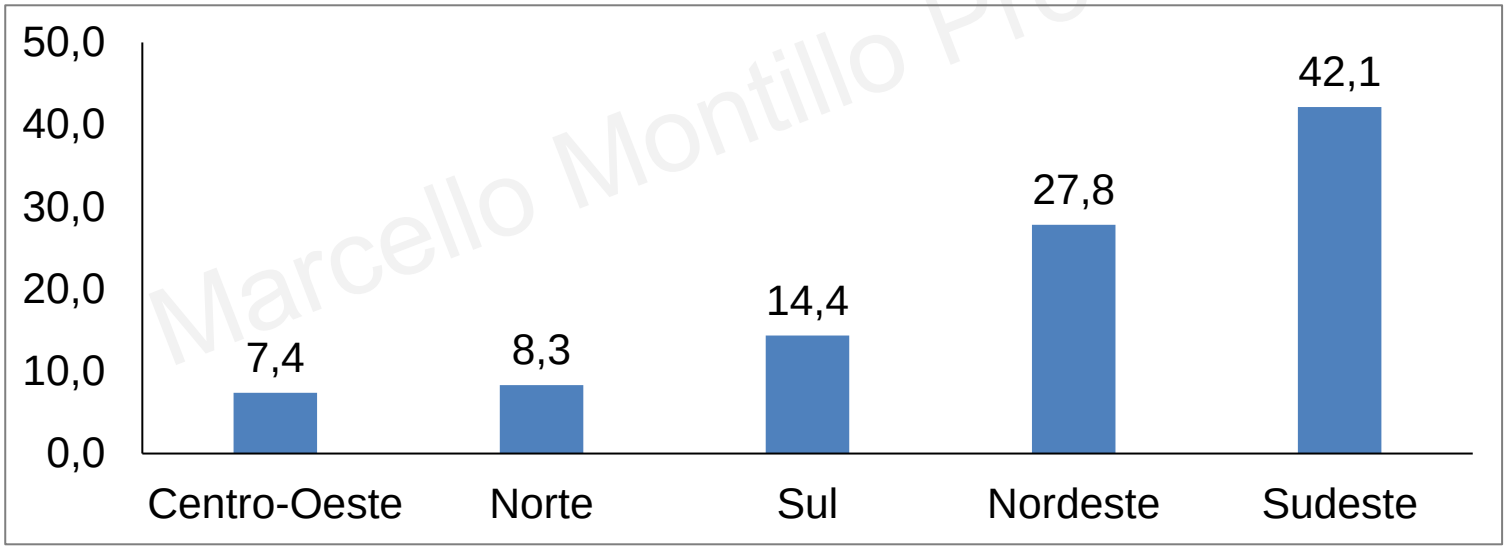
Gráficos de informação: são gráficos destinados principalmente ao público em geral, objetivando proporcionar uma visualização rápida e clara. São gráficos tipicamente expositivos, dispensando comentários explicativos adicionais.

Gráficos de análise: são gráficos que se prestam melhor ao trabalho estatístico, fornecendo elementos úteis à fase de análise dos dados, sem deixar de ser também informativos.

Análise Exploratória de Dados

Gráfico de Colunas: adequado para o uso de séries específicas e geográficas. Para cada rótulo da série existe um valor a ser representado, com uma coluna de tamanho igual ao valor observado.

Gráfico 1: População residente por região do Brasil - 2010 - valores percentuais.

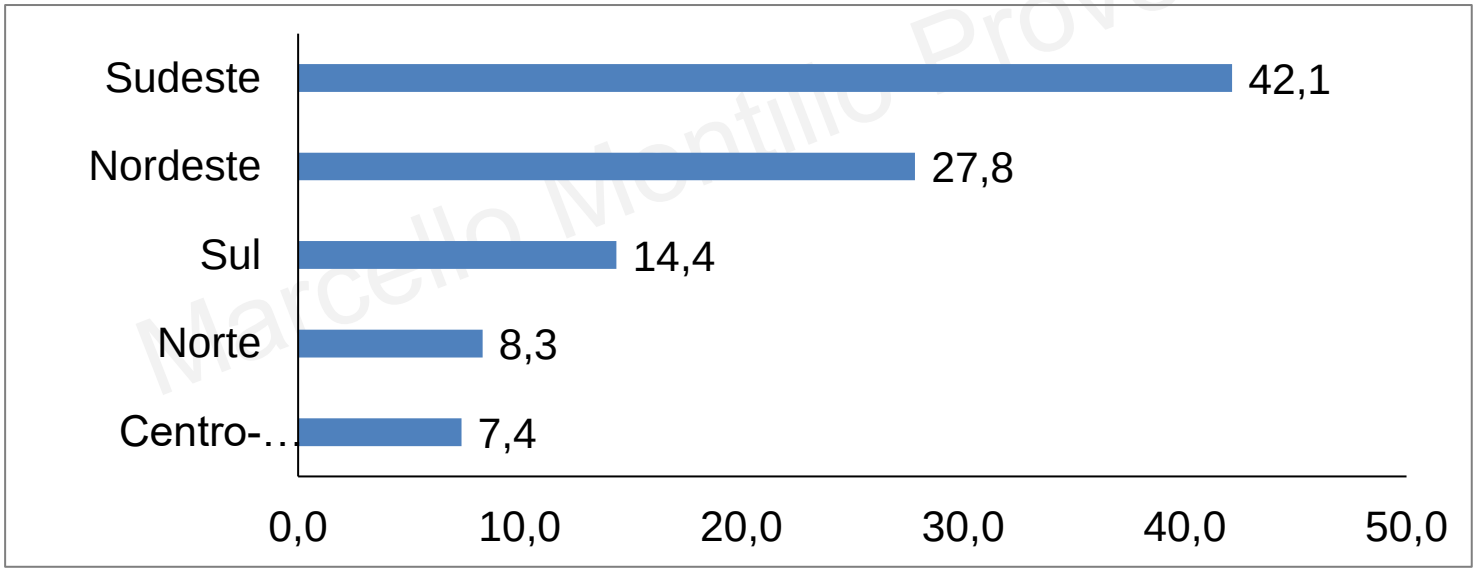


Fonte: IBGE, Censo Demográfico 2010.

Análise Exploratória de Dados

Gráfico de Barras: análogo ao de colunas, apenas invertendo o que estava no eixo horizontal para o eixo vertical.

Gráfico 2: População residente por região do Brasil - 2010 - valores percentuais.

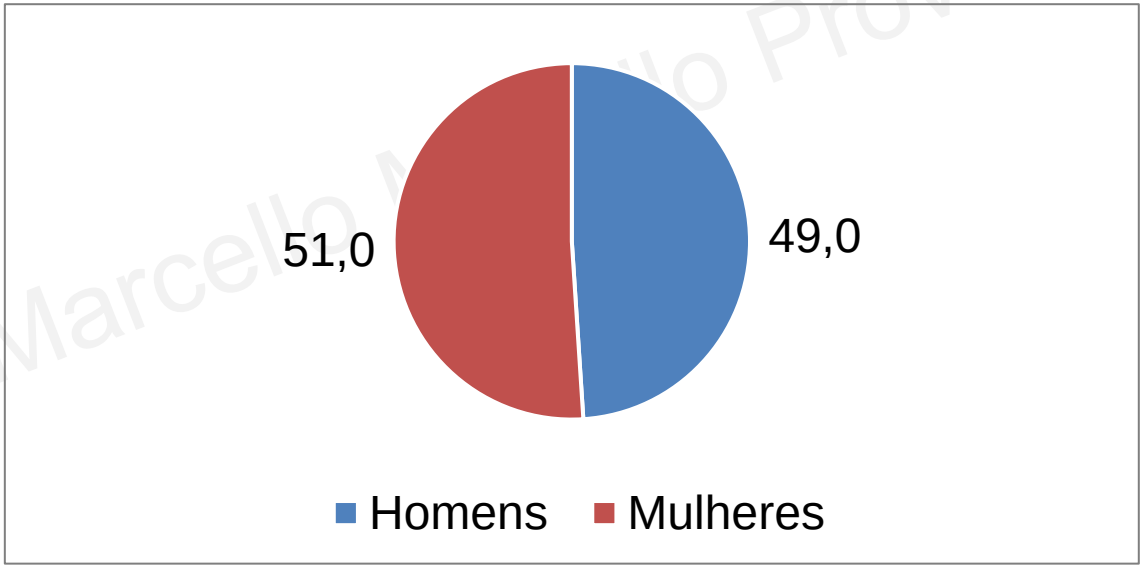


Fonte: IBGE, Censo Demográfico 2010.

Análise Exploratória de Dados

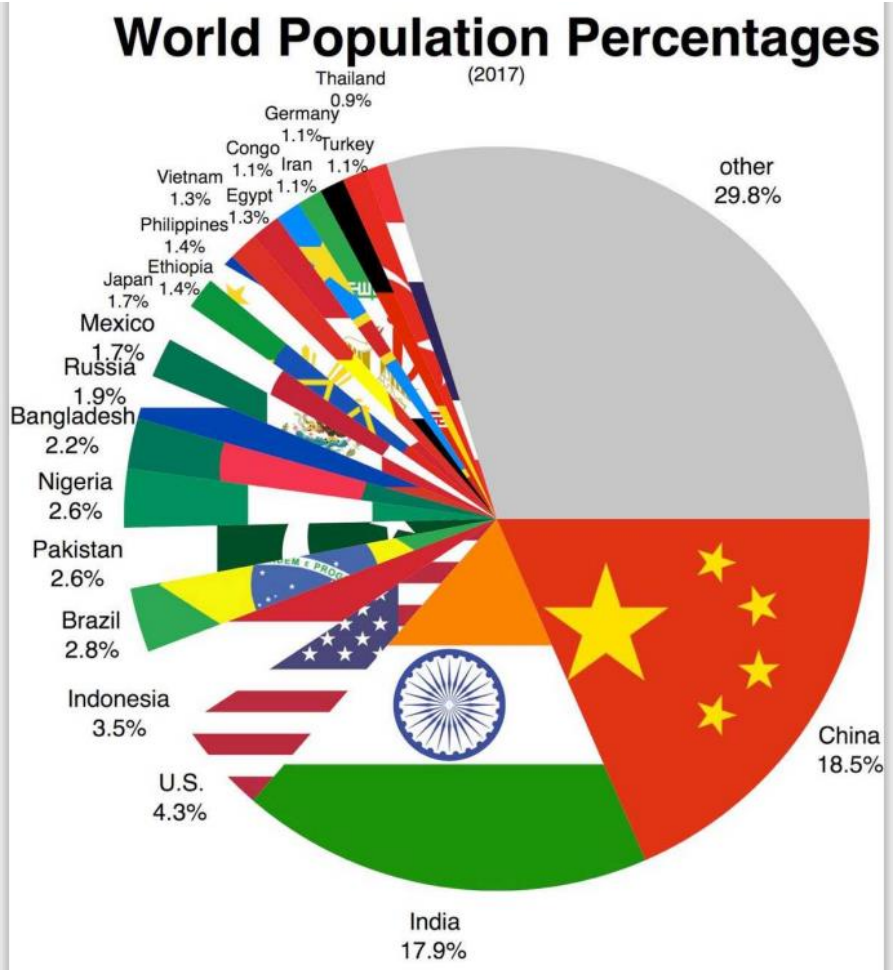
Gráfico em Setores (ou Pizza): é dividido em fatias proporcionais à contribuição de cada rótulo ou categoria. Não é recomendado para muitas categorias.

Gráfico 3: População residente por sexo do Brasil - 2010 – valores percentuais.



Fonte: IBGE, Censo Demográfico 2010.

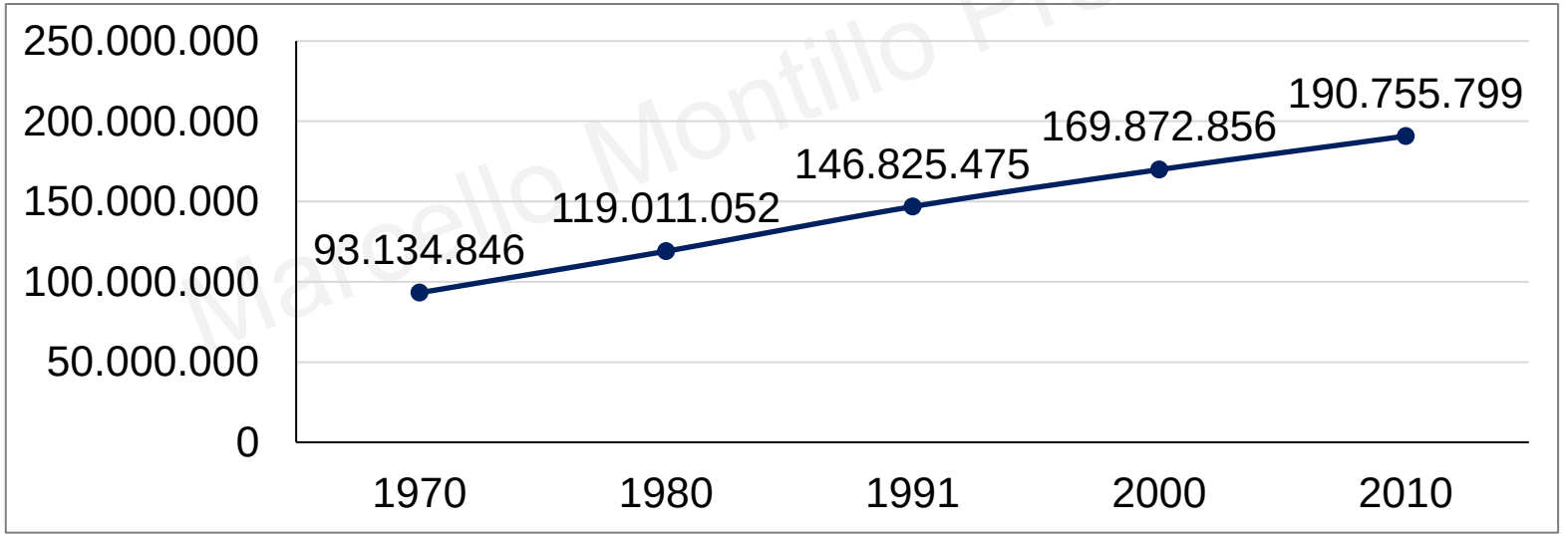
Mau Exemplo



Análise Exploratória de Dados

Gráfico em Linha: recomendado para uso em séries temporais. No eixo horizontal são representadas as unidades de tempo consideradas no gráfico. No eixo vertical são considerados os valores medidos pela unidade de tempo.

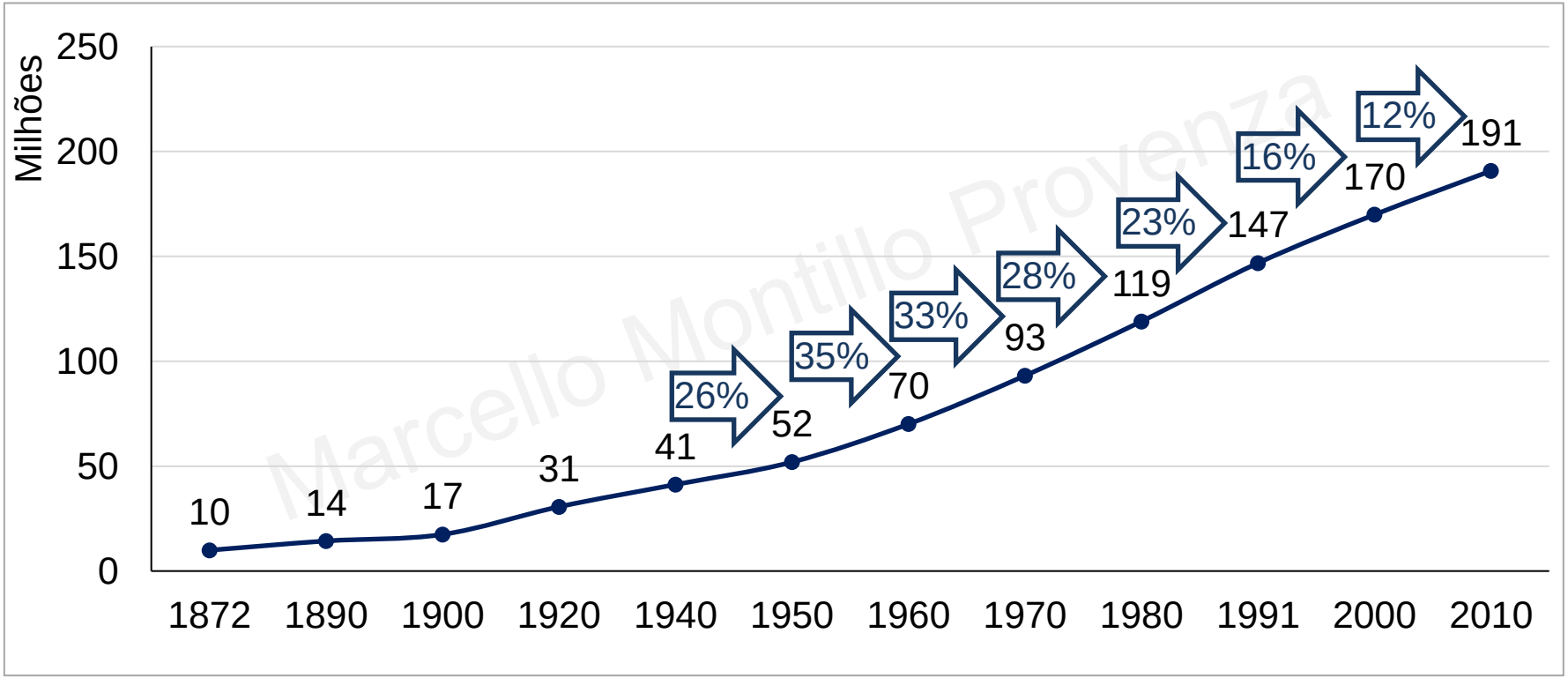
Gráfico 4: Evolução da população residente no Brasil entre 1970 e 2010 - valores absolutos.



Fonte: IBGE, Censo Demográfico 1970, 1980, 1991, 2000 e 2010.

Análise Exploratória de Dados

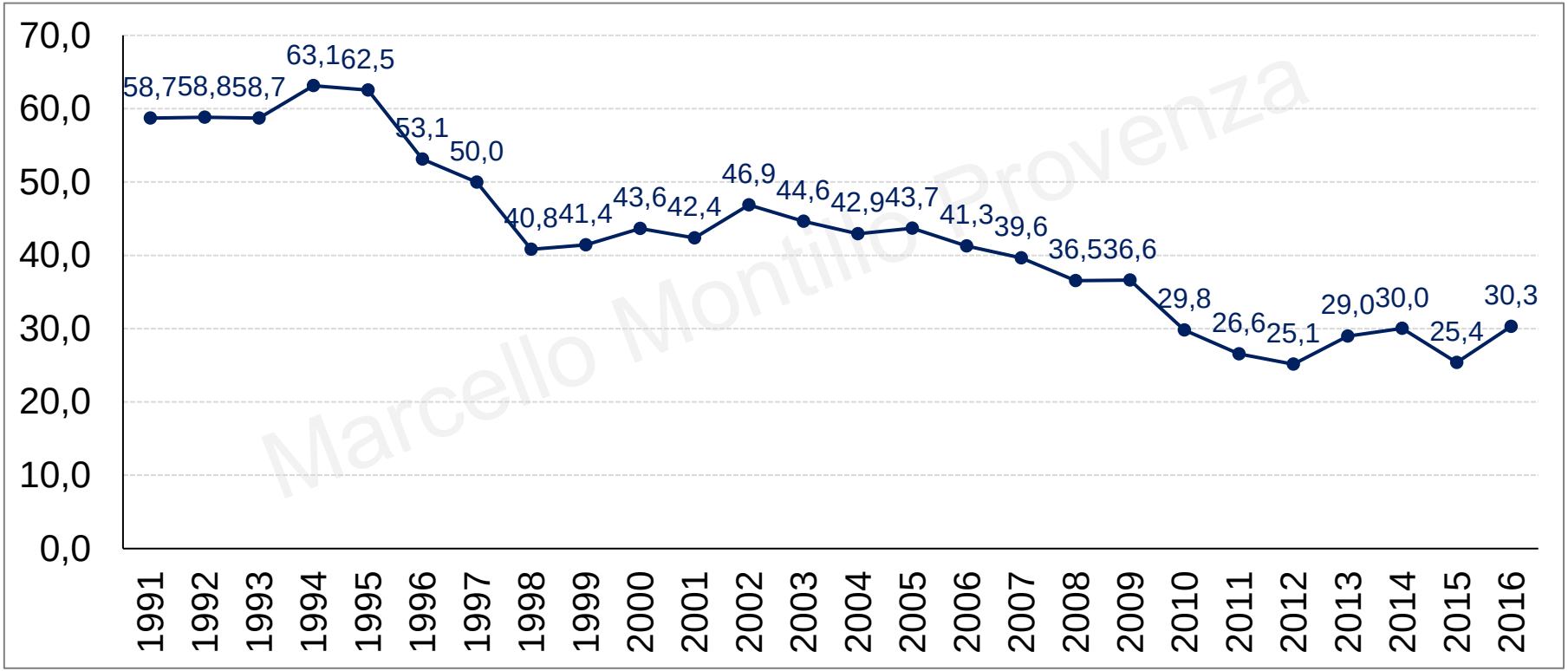
Gráfico 5: Evolução da população residente no Brasil entre 1872 e 2010 - valores absolutos.



Fonte: IBGE, Censo Demográfico 1872 - 2010.

Análise Exploratória de Dados

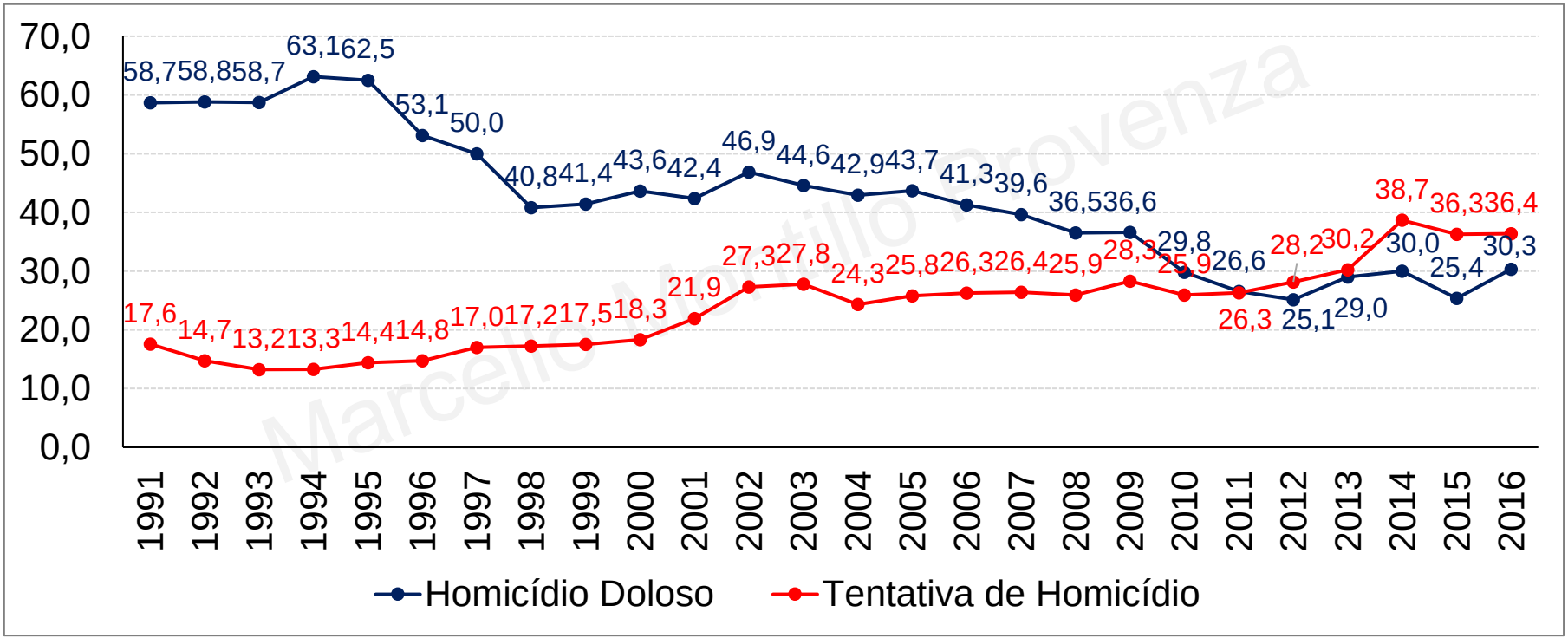
Gráfico 6: Homicídio doloso no estado do Rio de Janeiro no período entre 1991 e 2016 - taxas por 100.000 habitantes.



Fonte: Instituto de Segurança Pública, Registros de Ocorrência da PCERJ.

Análise Exploratória de Dados

Gráfico 7: Homicídio doloso e tentativa de homicídio no estado do Rio de Janeiro no período entre 1991 e 2016 – taxas por 100.000 habitantes.

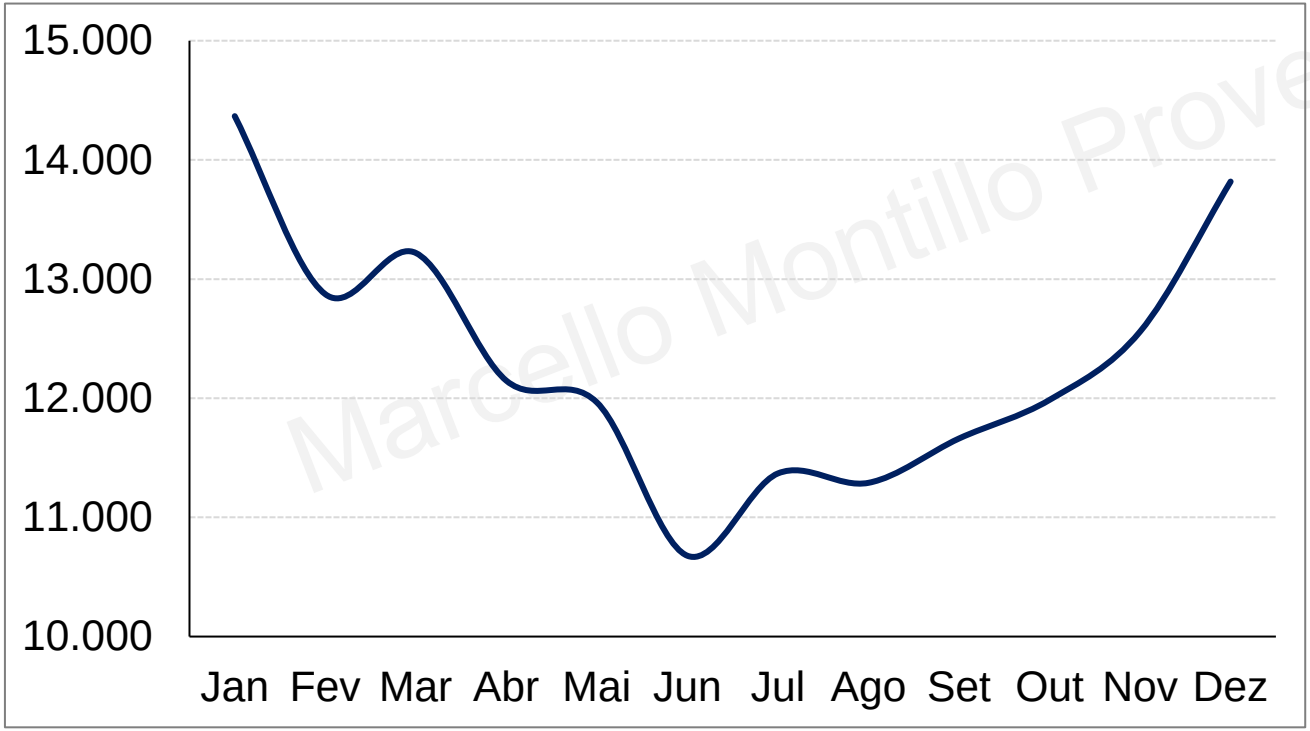


Fonte: Instituto de Segurança Pública, Registros de Ocorrência da PCERJ.

Análise Exploratória de Dados

Mau exemplo:

Gráfico 8: Homicídio doloso - 2015.



i) Título incompleto;

ii) Valores dos meses;

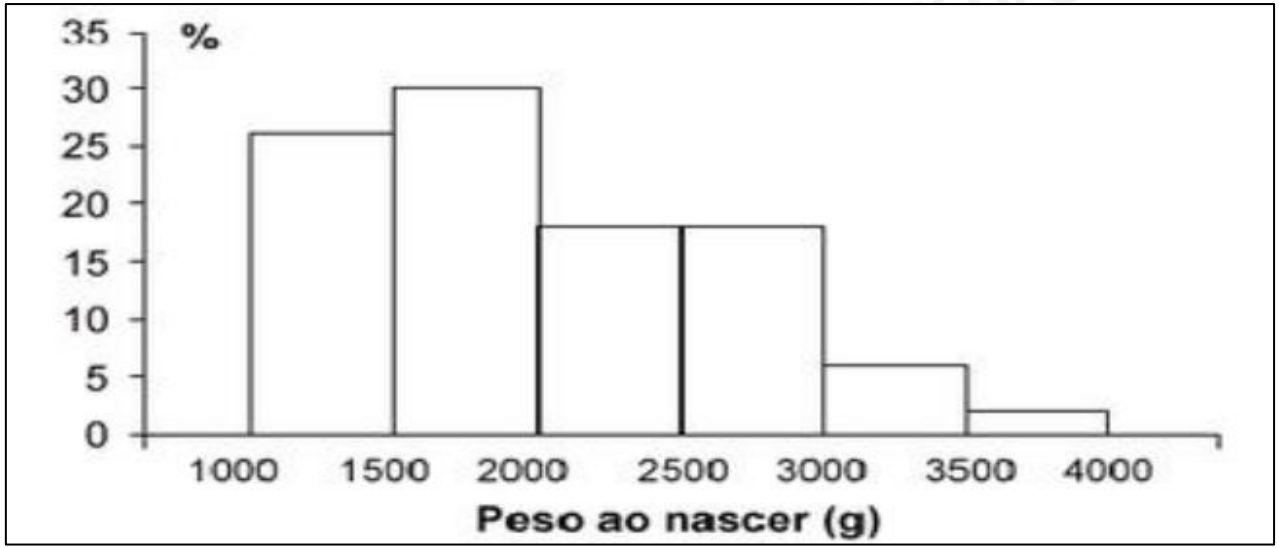
iii) Sem fonte;

iv) Escala enganosa.

Análise Exploratória de Dados

Histograma: são retângulos contíguos com base nas faixas de valores da variável e com área igual à frequência relativa da respectiva faixa.

Gráfico 9: Distribuição de recém-nascidos acometidos de síndrome de desconforto idiopático grave segundo peso ao nascer (g).

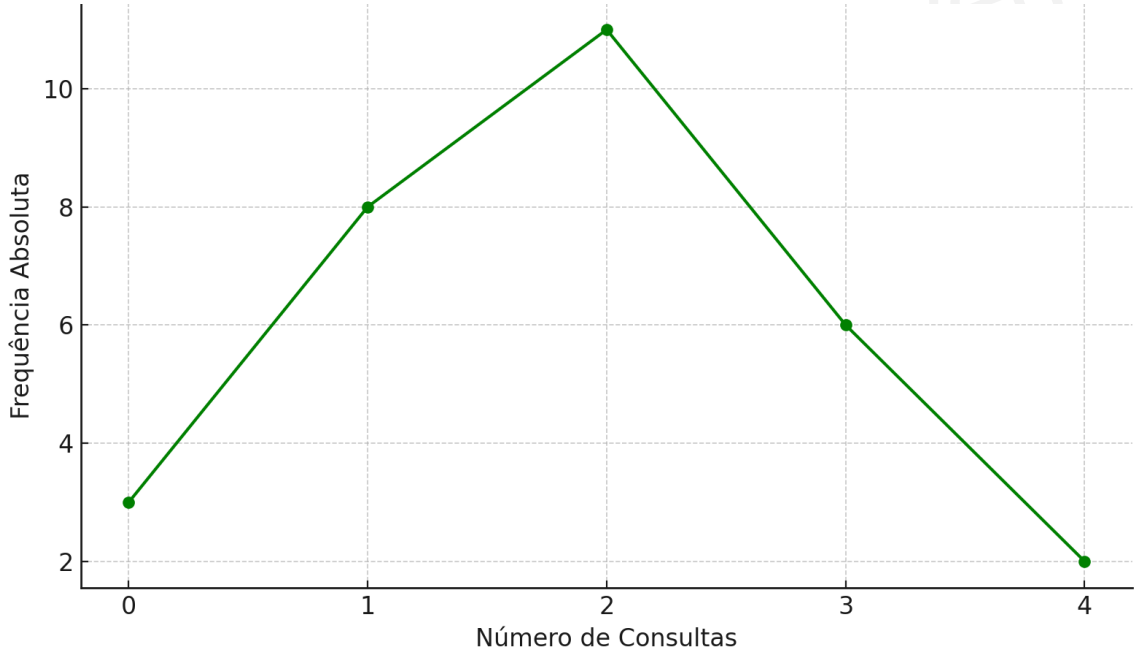


Fonte: van Vliet PKJ, Gupta JM (1973).

Análise Exploratória de Dados

Polígono de frequências: une através de segmento de reta as ordenadas correspondentes aos pontos médios das bases superiores dos retângulos correspondentes a cada uma das classes do histograma.

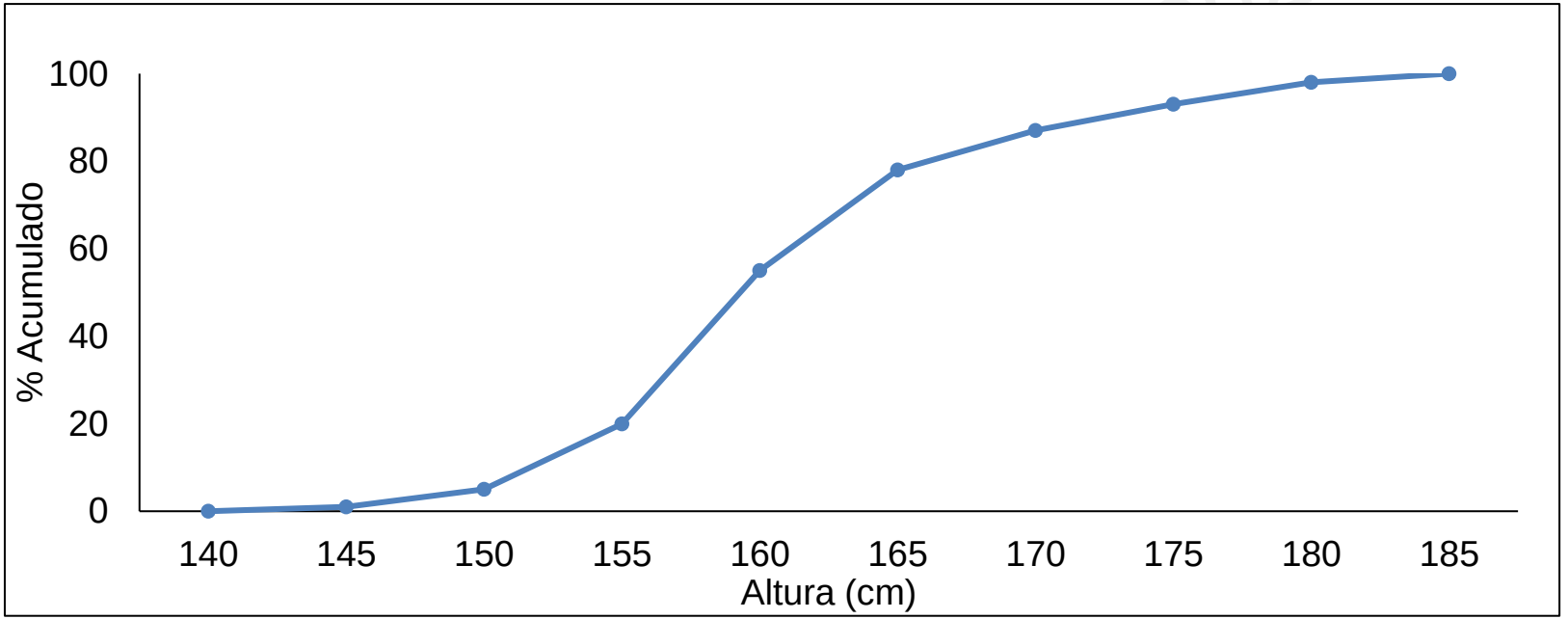
Gráfico 10: Número de consultas em UBS (dados fictícios).



Análise Exploratória de Dados

Gráfico de ogiva: Apresenta uma distribuição de frequências acumuladas, utiliza uma poligonal ascendente utilizando os pontos extremos.

Gráfico 11: Distribuição de mulheres idosas segunda a altura (dados fictícios).



Análise Exploratória de Dados

Diagrama de Ramo-e-folhas: útil para visualizar a distribuição dos dados sem perder os valores originais, e também com dados numéricos pequenos e discretos.

Exemplo: Idade de Pessoas Atendidas em um Programa Social

Um município quer analisar o perfil etário das pessoas atendidas em um programa de transferência de renda. A seguir, temos uma amostra com as idades de 20 beneficiários:

Idades (em anos): 18, 19, 21, 22, 22, 23, 24, 25, 25, 26, 27, 27, 28, 30, 31, 32, 33, 35, 36 e 38.

Análise Exploratória de Dados

Neste tipo de diagrama, o **ramo** representa a dezena da idade e a **folha** representa a unidade.

Ramo | Folhas

1 | 8 9
2 | 1 2 2 3 4 5 5 6 7 7 8
3 | 0 1 2 3 5 6 8

Esse tipo de gráfico é útil para **visualizar rapidamente a concentração de valores** e identificar possíveis assimetrias ou valores atípicos.

Análise Exploratória de Dados

A maioria dos beneficiários está concentrada na faixa dos 20 a 29 anos. Isso sugere que o programa tem forte impacto sobre jovens adultos, possivelmente no início da vida produtiva.

A faixa dos 30 a 39 anos conta com 7 pessoas, indicando que o programa também atende a uma quantidade relevante de adultos em idade produtiva avançada.

Apenas 2 pessoas estão na faixa dos 18-19 anos, o que pode indicar menor demanda ou elegibilidade nessa faixa e/ou falta de conhecimento ou barreiras de acesso ao programa por jovens recém-saídos da adolescência.

Não há valores extremos (outliers), o que mostra consistência no público-alvo.

Medidas de Posição

Medidas de Posição

Dentre todas as informações da população ou da amostra, podemos retirar valores que representem, de algum modo, todo o conjunto.

As medidas de posição (ou medidas de tendência central) buscam identificar valores típicos de uma determinada distribuição.

As mais comuns são: média aritmética, mediana e moda.

Medidas de Posição

Média Aritmética: é o valor que aponta para onde mais se concentram os dados de uma distribuição.

Populacional

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \ldots + X_N}{N}$$

Amostral

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

Média Aritmética Ponderada: onde se tem a atribuição de pesos aos valores observados de uma distribuição.

$$\mu = \frac{\sum_{i=1}^{N} X_i \times P_i}{\sum_{i=1}^{N} P_i} = \frac{(X_1 \times P_1) + (X_2 \times P_2) + \ldots + (X_N \times P_N)}{P_1 + P_2 + \ldots + P_N}$$

Medidas de Posição

Ex. 1: Do conjunto de valores dados abaixo de uma **população** de seis elementos, calcule a média.

X = { 3, 5, 8, 4, 8, 6 }

Solução:

$$\mu_{(X)} = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{\sum_{i=1}^{6} X_i}{6} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6}{6}$$

$$\mu_{(X)} = \frac{3+5+8+4+8+6}{6} = \frac{34}{6} = 5{,}67$$

Medidas de Posição

Ex. 2: Um candidato obteve, nas diversas provas de um concurso, as seguintes notas com seus respectivos pesos descritos abaixo. Calcule a média das notas.

Matéria	Nota	Peso
Português	66	3
Contabilidade	63	3
Estatística	74	2
Direito	79	2

Solução:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i \times p_i}{\sum_{i=1}^{n} p_i} = \frac{\sum_{i=1}^{4} x_i \times p_i}{\sum_{i=1}^{4} p_i} = \frac{(x_1 \times p_1) + (x_2 \times p_2) + (x_3 \times p_3) + (x_4 \times p_4)}{p_1 + p_2 + p_3 + p_4}$$

$$\overline{X} = \frac{(66 \times 3) + (63 \times 3) + (74 \times 2) + (79 \times 2)}{3 + 3 + 2 + 2} = \frac{693}{10} = 69{,}3$$

Exercício Relâmpago

Sabendo-se que o nível de colesterol (mg/100 ml) de um conjunto de sete pacientes clínicos foi de 10, 14, 13, 15, 16, 18 e 12, qual o colesterol médio desses pacientes?

Medidas de Posição

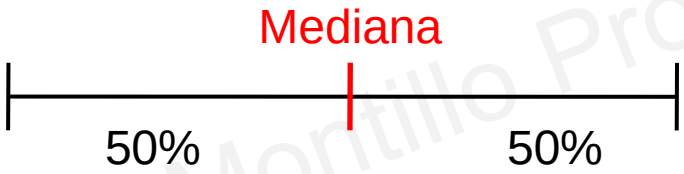
Solução:

$$\overline{X} = \frac{\sum_{i=1}^{7} x_i}{7} = \frac{10 + 14 + 13 + 15 + 16 + 18 + 12}{7}$$

$$\overline{X} = 14 \text{ m g}/100 \text{ ml}$$

Medidas de Posição

Mediana: É o valor que ocupa a posição central da série de observações, quando estão em ordem crescente ou decrescente. Se o conjunto de observações for par, usa-se como mediana a média aritmética das duas observações centrais. Se o conjunto for ímpar, é o valor central da série.



Moda: É o valor mais frequente do conjunto dos elementos observados.

- Amodal: quando não existe moda;
- Unimodal: quando há uma única moda;
- Bimodal: quando há duas modas;
- Multimodal: quando há mais de duas modas.

Medidas de Posição

Ex. 3: Do conjunto de valores observados abaixo, calcule a mediana e moda.

a) X = { 3, 5, 8, 4, 8, 6 }

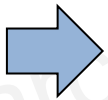
b) Y = { 3, 6, 8, 8, 6, 6, 10, 20, 12 }

Solução:

Letra a.

n = 6 (par)

3
4
5
6
8
8



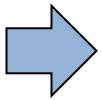
$$Md = \frac{5+6}{2} = 5{,}5$$

$$Mo = 8$$

Letra b.

n = 9 (ímpar)

3
6
6
6
8
8
10
12
20



$$Md = 8$$

$$Mo = 6$$

Medidas de Posição

Exercício Relâmpago

Sabendo-se que o nível de colesterol (mg/100 ml) de um conjunto de oito pacientes clínicos foi de 10, 10, 13, 15, 16, 16, 12 e 18, qual a mediana e a moda do colesterol desses pacientes?

Medidas de Posição

Solução

Colocando os dados em ordem crescente, temos:

X = { 10, 10, 12, 13, 15, 16, 16, 18 } => n = 8 (par)

$$\mathrm{Md(X)} = \frac{13 + 15}{2} = \frac{28}{2} = 14$$

$$\mathrm{Mo(X)} = 10 \text{ e } 16 \Rightarrow \text{Distribuição Bimodal}$$

Medidas de Posição

A medida de posição mais usada é a média aritmética, que apresenta em relação à mediana e à moda vantagens apreciáveis, tais como:

- É facilmente calculável;
- É a que melhor se presta a imediatas análises estatísticas;
- Depende de todos os valores da série;
- É uma medida de tendência central particularmente estável, variando o menos possível de amostra para amostra extraídas da mesma população;
- Pode ser tratada algebricamente.

Medidas de Posição

Então, qual a vantagem de utilizar a mediana ao invés da média?

A principal vantagem é que a mediana não é influenciada por valores extremos.

Ex. 4: Do conjunto de valores observados abaixo por um médico em relação as idades de seus pacientes, calcule a média e a mediana.

$$X = \{ 3, 4, 5, 8, 100 \}$$

$$\bar{X} = \frac{3 + 4 + 5 + 8 + 100}{5} = \frac{120}{5} = 24 \qquad Md(X) = 5$$

Logo, a mediana expressa melhor o conjunto das idades do que a média. O valor extremo (100) pode ter ocorrido por diversos motivos, inclusive, por um simples erro de digitação.

Medidas de Posição

Separatrizes

São números que dividem uma sequência ordenada de dados em partes que contêm a mesma quantidade de elementos da série.

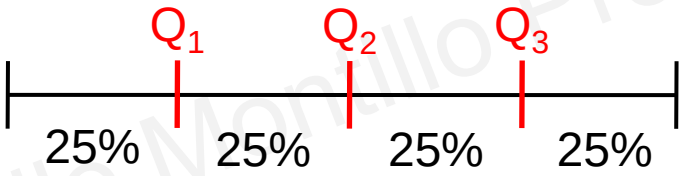
Desta forma, a mediana também é uma separatriz, pois divide a série ordenada em dois grupos.

Além da mediana, as outras medidas separatrizes mais conhecidas são: quartil, decil e percentil.

Medidas de Posição

Quartil

Os quartis dividem a série ordenada em 4 partes iguais, contendo cada uma delas 1/4 (ou 25%) das observações.



Amplitude Interquartil: é definida como a diferença entre o Quartil 3 e o Quartil 1.

$$AIQ = Q_3 - Q_1$$

Medidas de Posição

Decil

Os decis dividem a série ordenada em 10 partes iguais, contendo cada uma delas 1/10 ou 10% das observações.

D_1 = 1º decil (corresponde ao quantil de ordem p=1/10)

D_2 = 2º decil (corresponde ao quantil de ordem p=2/10)

. .

. .

. .

D_9 = 9º decil (corresponde ao quantil de ordem p=9/10)

Medidas de Posição

Percentil

Os percentis dividem a série ordenada em 100 partes iguais, contendo cada uma delas 1/100 ou 1% das observações.

P_1 = 1º percentil (corresponde ao quantil de ordem p=1/100)

P_2 = 2º percentil (corresponde ao quantil de ordem p=2/100)

. .

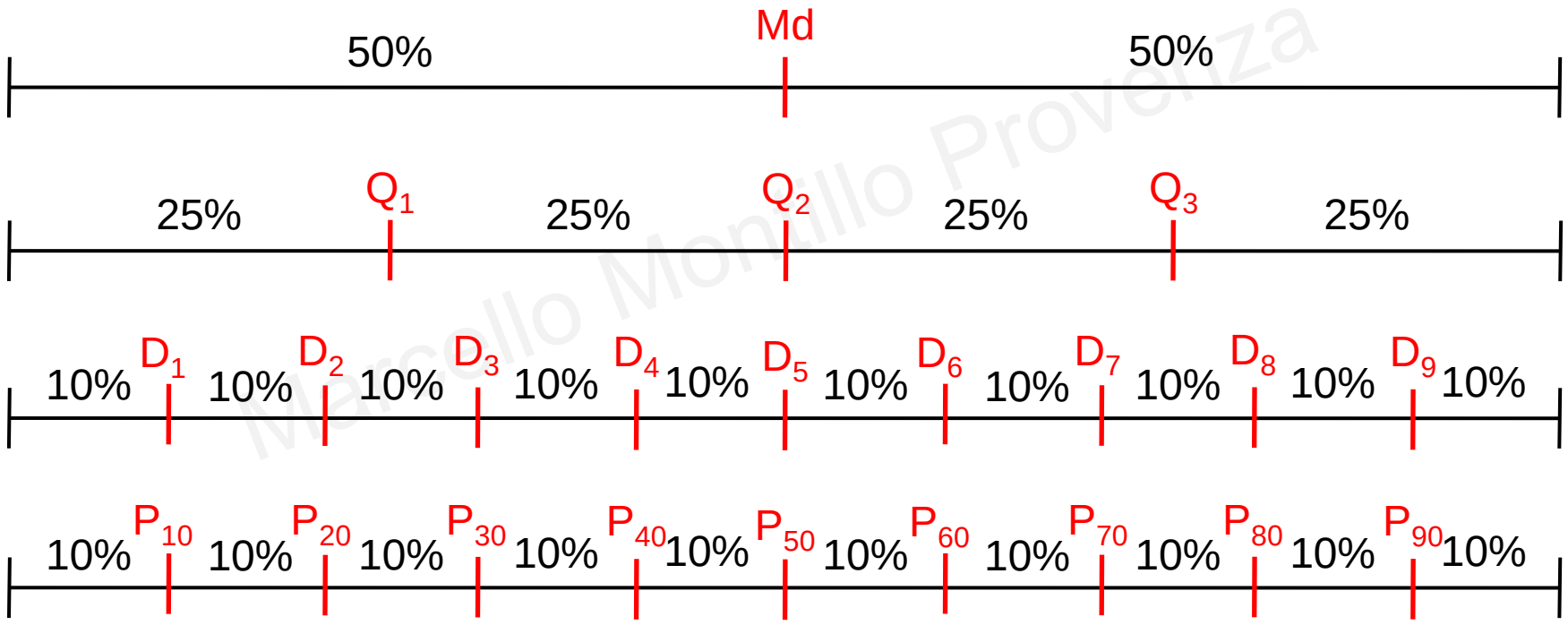
. .

. .

P_{99} = 99º percentil (corresponde ao quantil de ordem p=99/100)

Medidas de Posição

Equivalência das separatrizes



Outlier e Boxplot

Outlier

Outliers são valores extremos (positivos e/ou negativos) que se encontram muito distantes da maioria dos outros valores em um conjunto de dados.

Esses valores podem ser causados por erros de medição ou podem representar situações incomuns ou excepcionais.

Também podem ser chamados de valores discrepantes ou valores atípicos.

Outlier

Para lidar com outliers, é preciso primeiro pesquisar se são erros de medição ou valores reais dos dados.

Se forem erros de medição, é possível corrigi-los ou removê-los do conjunto de dados. Caso contrário, é importante avaliar se os outliers são relevantes para a análise e se devem ser mantidos ou removidos.

Em alguns casos, os outliers podem fornecer informações valiosas sobre o conjunto de dados e não devem ser descartados sem uma análise cuidadosa.

Boxplot

Também chamado de **diagrama de caixa**, é utilizado para representar dados estatísticos de maneira visual, muito útil para identificar padrões nos dados e para comparar conjuntos de dados diferentes.

Composto por um retângulo que representa os quartis do conjunto de dados, uma linha vertical que indica a mediana e dois segmentos de reta que se estendem a partir do retângulo até os limites superior e inferior dos dados.

Além disso, permite identificar facilmente valores discrepantes, conhecidos como *outliers*, que podem afetar a interpretação dos dados.

Boxplot

Como interpretar um boxplot?

- Um boxplot é um gráfico utilizado para representar dados estatísticos de forma visual.

- Ele é composto por cinco partes: limite inferior, primeiro quartil, mediana, terceiro quartil e limite superior.

- O limites inferior e superior são os menores e maiores valores calculados para os dados.

    - $\text{Limite Inferior} = \ Q_1 - 1{,}5 * AIQ$
    - $\text{Limite Superior} = \ Q_3 + 1{,}5 * AIQ$

Boxplot

Como interpretar um boxplot?

- O primeiro quartil é o valor que divide os dados em 25% inferiores e 75% superiores, enquanto o terceiro quartil é o valor que divide os dados em 75% inferiores e 25% superiores.

- A mediana é o valor que divide os dados em duas partes iguais, sendo que metade dos dados está acima dela e metade está abaixo.

- Os outliers são valores extremos que estão muito distantes do restante dos dados, e estão representados por pontos fora do intervalo entre os limites inferior e superior.

Boxplot

O boxplot é uma ferramenta útil para representar dados estatísticos quando se deseja ter uma visão geral da distribuição dos dados.

Ele permite identificar os quartis, a mediana e os outliers de forma clara e objetiva.

Uma das principais vantagens do boxplot em relação a outros tipos de gráficos é que ele apresenta informações sobre a assimetria e a presença de valores extremos na distribuição dos dados.

Boxplot

Além disso, o boxplot é capaz de lidar com dados discrepantes sem afetar a interpretação da mediana e dos quartis.

No entanto, em algumas situações, o boxplot pode não ser a melhor opção para representar dados estatísticos, especialmente quando há muitos valores repetidos ou quando a distribuição dos dados é muito complexa.

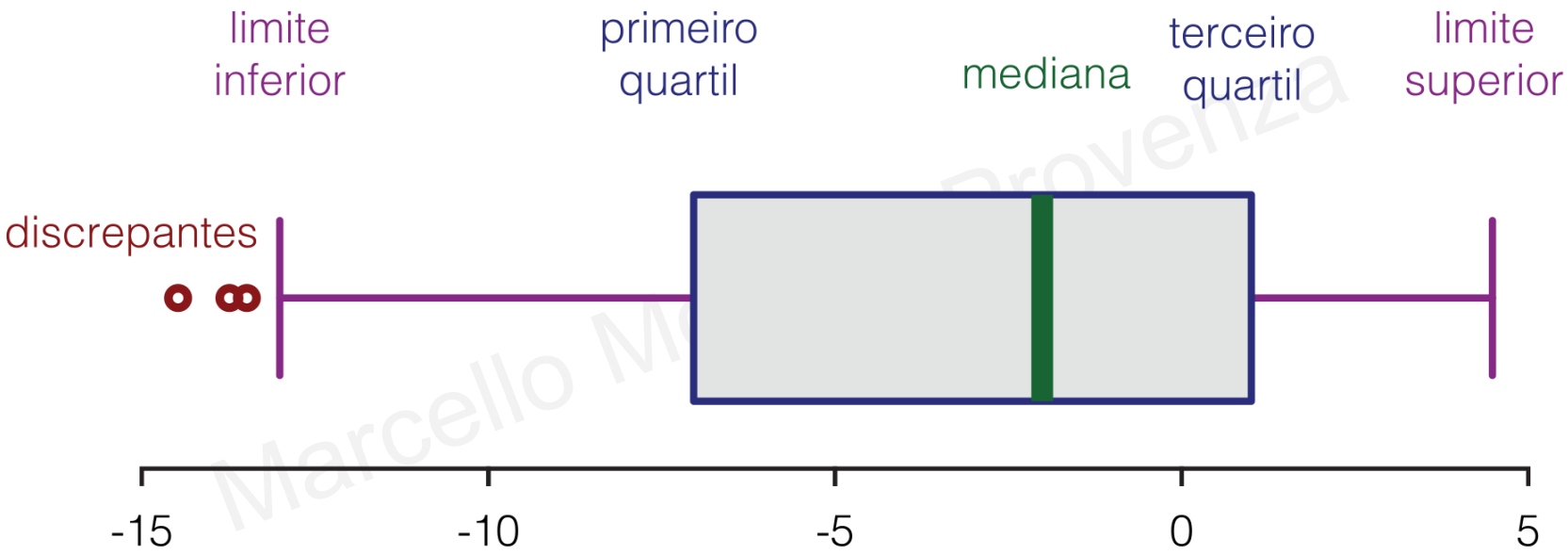
Boxplot

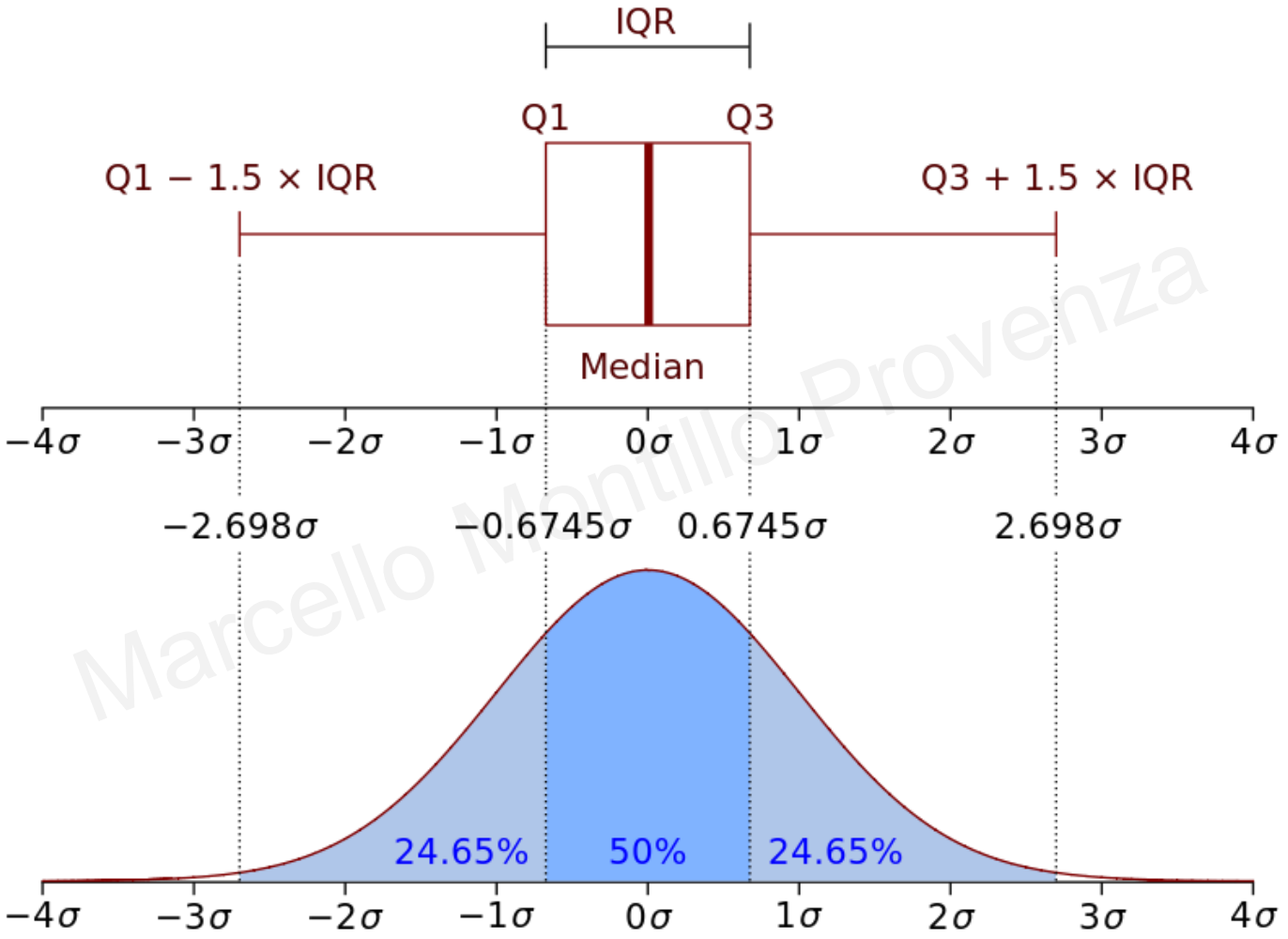
Exemplos:

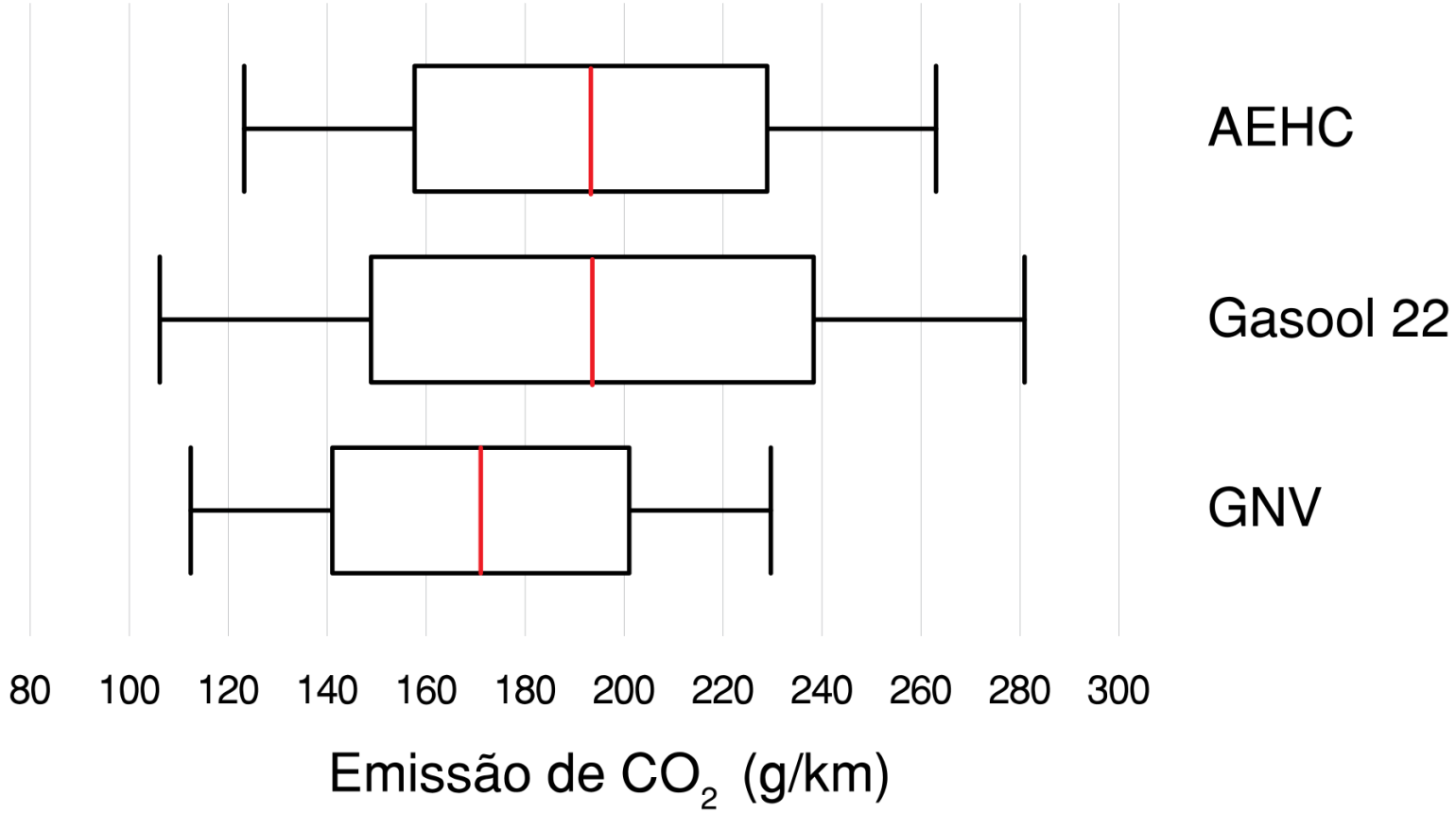
- **Representar a distribuição dos salários:** nesse caso, os quartis indicam a faixa salarial da maioria dos funcionários, enquanto os outliers representam os funcionários com salários muito altos ou muito baixos em relação à média.

- **Representar a distribuição das notas de um teste:** nesse caso, os quartis indicam a faixa de notas da maioria dos alunos, enquanto os outliers representam os alunos com notas muito altas ou muito baixas em relação à média.

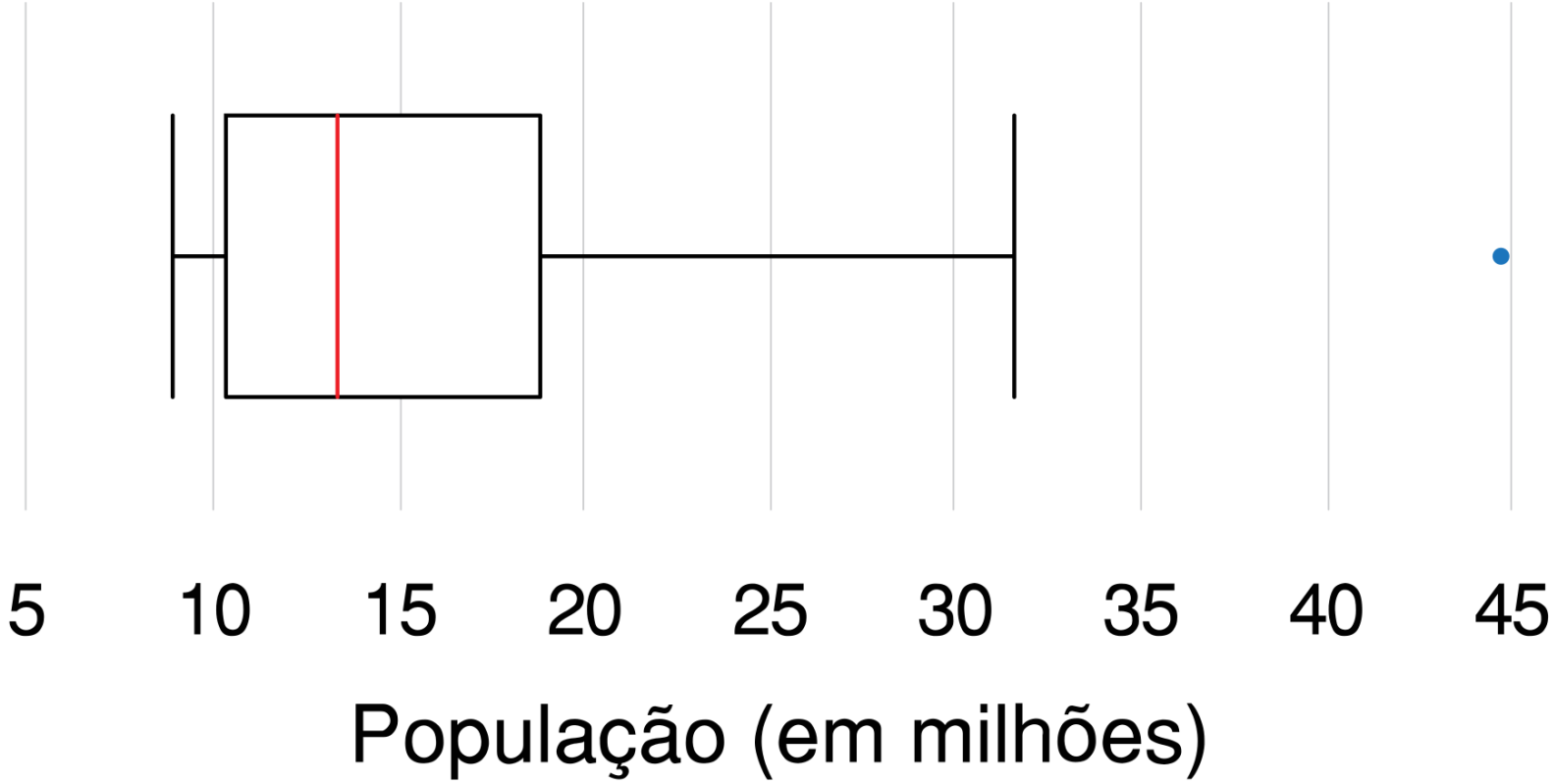
Boxplot



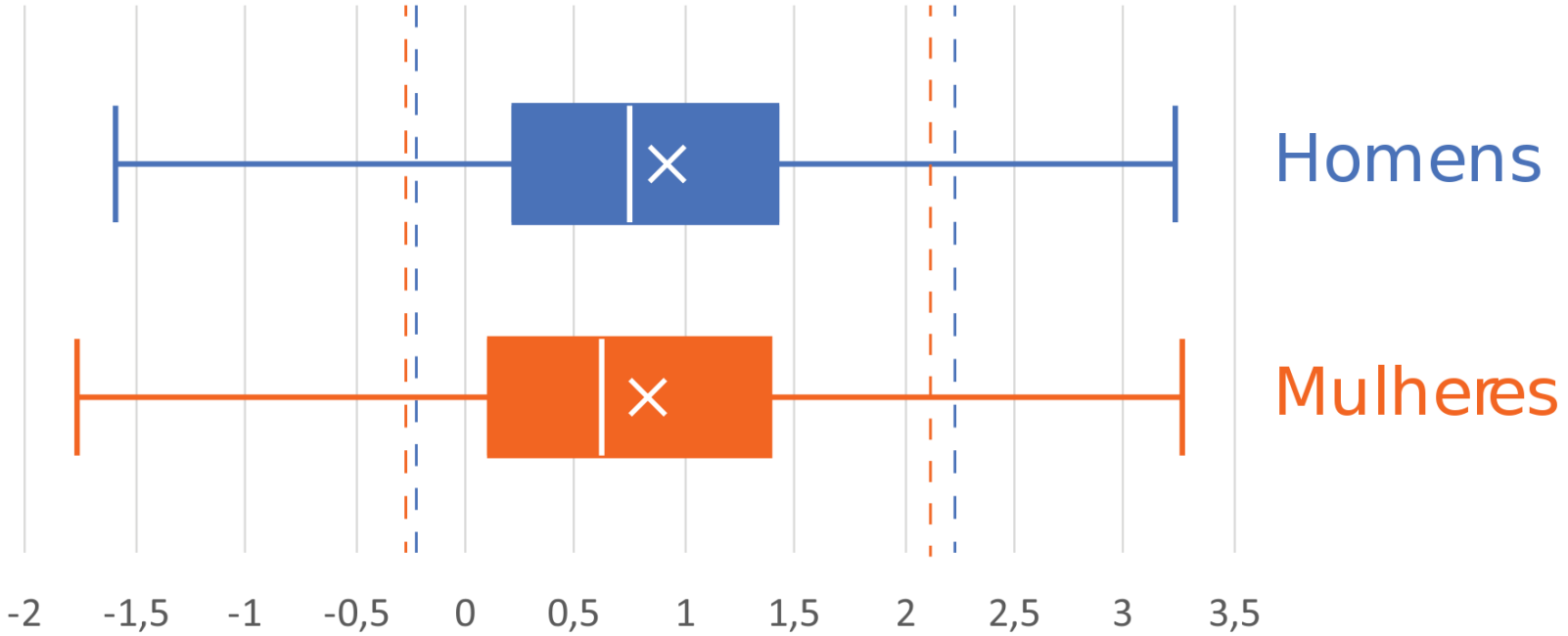




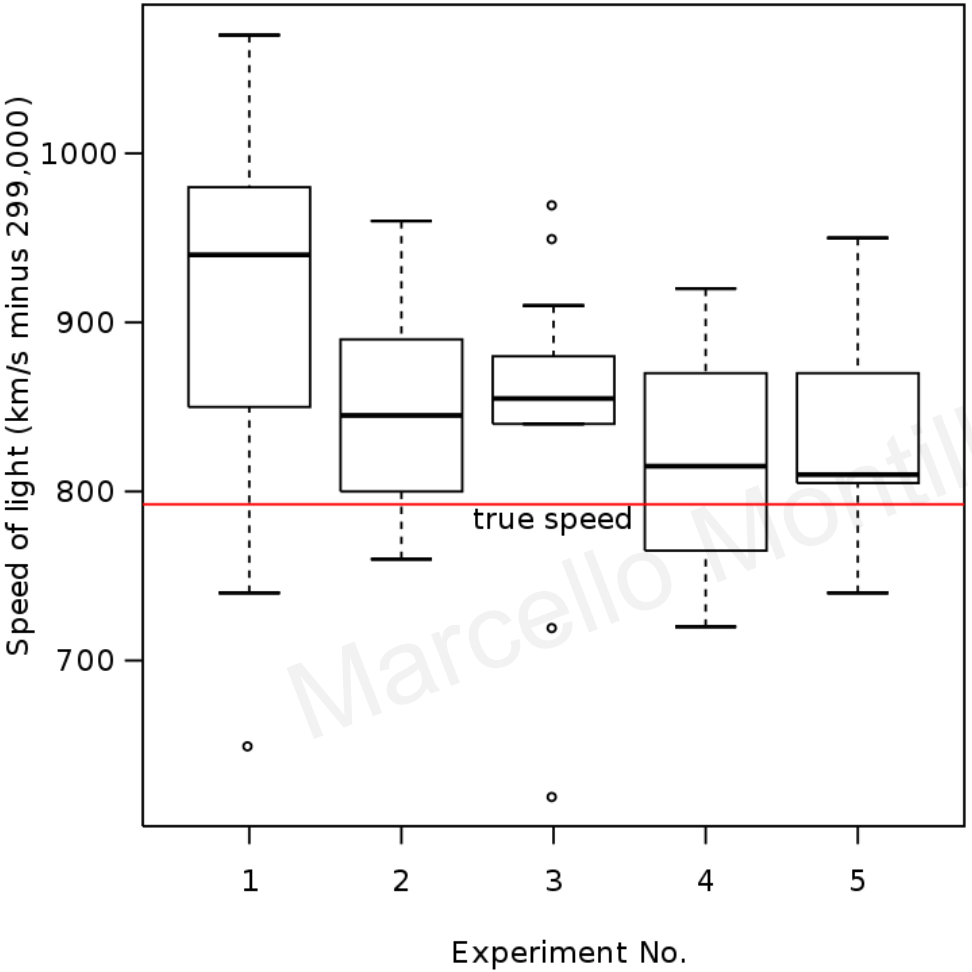
Exemplo 1: Caracterização das emissões de gases de efeito estufa por veículos automotores leves no Estado de São Paulo (as linhas vermelhas representam as medianas dos dados).



Exemplo 2: Estimativas da população residente no Brasil e UF, 2016, IBGE. A linha vermelha representa a mediana e o ponto azul representa o valor discrepante (estado de São Paulo).



Exemplo 3: Rendimentos-hora de homens e mulheres. As linhas tracejadas à esquerda representam o percentil 10 e as linhas tracejadas à direita representam o percentil 90. As barras brancas representam a mediana das observações e os x's brancos representam a média.



Exemplo 4: Velocidade da luz, dados do experimento de Michelson–Morley.

Medidas de Dispersão

Exercício Relâmpago

Vamos considerar o exemplo de um conjunto de dados que representa o tempo de espera no atendimento em uma delegacia (em minutos). Suponha que os dados sejam os seguintes: 3, 5, 7, 10, 15, 20, 30, 60. Construa um boxplot e identifique possíveis outliers.

Medidas de Dispersão

Solução

Ao construir o boxplot desses dados, podemos ver que a mediana é 12,5, o primeiro quartil é 6 e o terceiro quartil é 25. O limite inferior é -22,5 e o superior é 53,5. Além disso, há um outlier em 60, que representa um tempo de espera muito acima do restante dos dados.

Medidas de Dispersão

Medidas de Dispersão

Uma medida de posição não nos dá, só por si, uma informação exaustiva da distribuição considerada. Devemos observar outras medidas que revelem o grau de variabilidade dos dados.

As medidas de dispersão mais comuns são: variância, desvio padrão e o coeficiente de variação.

Medidas de Dispersão

O que é **desvio**?

Como queremos avaliar a dispersão dos dados em torno da média, esse valor estará relacionado com a distância dos dados em relação à média. Essa distância será chamada de desvio.

$$d_i = X_i - \overline{X}$$

Medidas de Dispersão

Desvio médio: mede o "afastamento" dos dados em relação a média. É a média de todos os desvios.

$$DM(X) = \frac{\sum_{i=1}^{N} \left| X_i - \overline{X} \right|}{N} = \frac{\left| X_1 - \overline{X} \right| + \left| X_2 - \overline{X} \right| + ... + \left| X_N - \overline{X} \right|}{N}$$

Ex. 1: Em duas rodas de amigos, foi pesquisado as notas que tinham obtido em um trabalho do curso. Cada grupo tinha três pessoas e as notas estão descritas abaixo. Calcule o desvio médio.

Roda 1 = { 3, 10, 8 }

Roda 2 = { 7, 6, 8 }

Medidas de Dispersão

Solução:

Calculando a média das duas rodas temos: $\overline{X}_{R1} = 7$ e $\overline{Y}_{R2} = 7$

E, calculando o desvio médio das as rodas 1 e 2, temos:

$$DM_{(R1)} = \frac{|3-7|+|10-7|+|8-7|}{3} = \frac{4+3+1}{3} = \frac{8}{3} = 2{,}67$$

$$DM_{(R2)} = \frac{|7-7|+|6-7|+|8-7|}{3} = \frac{0+1+1}{3} = \frac{2}{3} = 0{,}67$$

Note que, mesmo com médias iguais, o desvio médio é diferente. Enquanto na roda 1 o desvio médio é 2,67, na roda 2 é 0,67. Podemos concluir que, mesmo com médias iguais, as notas da roda 2 se “afastam” menos da média do que as notas da roda 1.

Medidas de Dispersão

Variância: quantifica a variabilidade ou o espalhamento dos dados ao redor da média. A variância mostra o quão distante cada valor desse conjunto está do valor média.

Populacional

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + ... + (X_N - \overline{X})^2}{N}$$

Amostral

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_N - \overline{x})^2}{n-1}$$

Medidas de Dispersão

Desvio padrão: é a raiz quadrada da variância. Usa a mesma unidade de medida da média.

Populacional

$$\sigma = \sqrt{\sigma^2}$$

Amostral

$$s = \sqrt{s^2}$$

Ex: Se tenho uma unidade de medida em metros, a média calculada é em metros, a variância é em metros² e o desvio padrão é em metros.

Medidas de Dispersão

Coeficiente de variação: Útil para a comparação em termos relativos do grau de concentração em torno da média de séries distintas.

$$CV(X) = \frac{s_x \times 100}{\overline{X}}$$

Interpretação do coeficiente de variação:

CV < 15% ⟹ Baixa dispersão

15% ≤ CV ≤ 30% ⟹ Média dispersão

CV > 30% ⟹ Alta dispersão

Medidas de Dispersão

Ex. 2: Voltando ao exemplo das rodas entre amigos (exemplo 1), calcule a variância, o desvio padrão e o coeficiente de variação. Relembrando: R1 = {3,10,8} e R2 = {7,6,8}

Solução:

Calculando os dados para roda 1:

$$\overline{X}_{R1} = 7$$

$$s^2_{(R1)} = \frac{(3-7)^2 + (10-7)^2 + (8-7)^2}{3-1} = \frac{4^2 + 3^2 + 1^2}{2} = \frac{26}{2} = 13$$

$$s_{(R1)} = \sqrt{13} = \pm 3{,}61$$

$$CV_{(R1)} = \frac{3{,}61 \times 100}{7} = 51{,}57\% \ \ (\textit{alta dispersão})$$

Medidas de Dispersão

Solução:

Calculando os dados para roda 2:

$$\bar{Y}_{R2} = 7$$

$$s^2_{(R2)} = \frac{(7-7)^2 + (6-7)^2 + (8-7)^2}{3-1} = \frac{0^2 + 1^2 + 1^2}{2} = \frac{2}{2} = 1$$

$$s_{(R2)} = \sqrt{1} = \pm 1$$

$$CV_{(R2)} = \frac{1 \times 100}{7} = 14{,}29\% \ \ (\textit{baixa dispersão})$$

Nas duas rodas de amigos, a média das notas foi 7. Contudo, analisando a dispersão dos dados, as notas da roda 2 teve dados mais homogêneos do que as notas da roda 1.

Medidas de Dispersão

Exercício Relâmpago

Um professor de educação física registrou o tempo de corrida de três alunos na esteira: o primeiro correu 40 minutos, o segundo 45 e o terceiro 50. Calcule a variância e o desvio padrão do tempo.

Medidas de Dispersão

Solução

Primeiramente devemos calcular a média de tempo em que os três alunos correram na esteira.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{3} x_i}{3} = \frac{40 + 45 + 50}{3} = \frac{135}{3} = 45$$

Posteriormente, calculamos a variância. Nesse caso, utilizamos a fórmula da variância amostral.

$$s^2(x) \frac{\sum_{i=1}^{3} (x_i - 45)^2}{3 - 1} = \frac{(40 - 45)^2 + (45 - 45)^2 + (50 - 45)^2}{3 - 1} = \frac{50}{2} = 25$$

Medidas de Dispersão

Solução

Esses resultados permitem afirmar que os alunos demoraram 45 minutos, em média, correndo na esteira, com variância de 25 minutos ao quadrado. Como a unidade de medida da variância (minutos ao quadrado) não tem sentido prático, devemos calcular o desvio padrão, o qual é a raiz da variância. Fazendo-se isso, o resultado fica na mesma unidade de medida da média.

$$s = \sqrt{s^2(x)} = \sqrt{25} = \pm 5$$

Os tempos variaram, mas, tipicamente, a diferença em relação à média foi de 5 minutos.

Medidas de Dispersão

Exercício Relâmpago

Um médico descreveu as medidas das estaturas e dos pesos de um mesmo grupo de indivíduos. Calcule o coeficiente de dispersão para a estatura e para o peso.

Medidas	$\overline{X}$	s
Estaturas	175 cm	5,0 cm
Pesos	68 kg	2,0 kg

Medidas de Dispersão

Solução

Como já foram fornecidos os valores da média e do desvio padrão para as variáveis estatura e peso, basta calcular os coeficientes de variação com base na fórmula.

$$CV(E) = \frac{s(E) * 100}{\bar{x}} = \frac{5 * 100}{175} = 2{,}85\%$$

$$CV(P) = \frac{s(P) * 100}{\bar{x}} = \frac{2 * 100}{68} = 2{,}94\%$$

Logo, nesse grupo de indivíduos, os pesos apresentam maior grau de dispersão que as estaturas.

Medidas de Formato

Medidas de Formato

As medidas de formato indicam o padrão da distribuição dos valores ao longo do intervalo que contém o total dos dados. Elas tentam captar, em um número, características da distribuição dos dados como assimetria e “achatamento“.

As medidas de formato mais conhecidas são a assimetria e a curtose.

Medidas de Formato

Assimetria

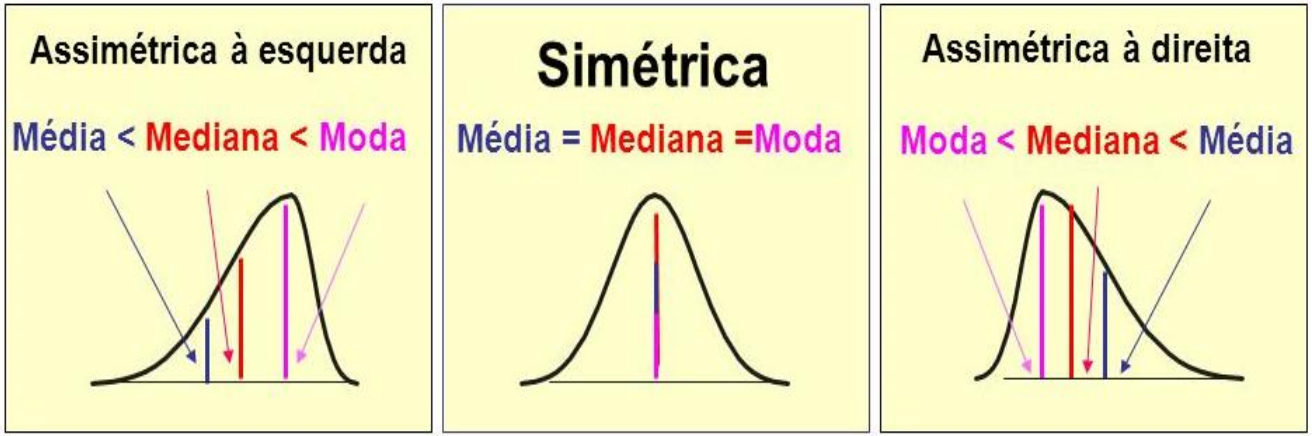
É o grau de desvio ou afastamento da simetria de uma distribuição. Existem várias fórmulas para o cálculo do coeficiente de assimetria, dentre elas, destaca-se o coeficiente de Pearson que pode ser calculado de duas maneiras distintas:

$$\Delta_S = \frac{\bar{x} - Mo}{S_{(x)}} \quad \text{ou} \quad \Delta_S = \frac{Q_1 + Q_3 - 2 * Md}{Q_3 - Q_1}$$

Medidas de Formato

Assimetria

- $\Delta_S = 0 \rightarrow$ Distribuição simétrica $(\bar{x} = Md = Mo)$
- $\Delta_S > 0 \rightarrow$ Distribuição assimétrica positiva (à direita) $(\bar{x} > Md > Mo)$
- $\Delta_S < 0 \rightarrow$ Distribuição assimétrica negativa (à esquerda) $(\bar{x} < Md < Mo)$



Medidas de Formato

Exemplo: Em uma pesquisa sobre distribuição de renda em uma comunidade, observou-se que a maioria das famílias ganha entre R$ 400 e R$ 800, mas algumas poucas famílias têm renda acima de R$ 5.000.

- A distribuição apresenta assimetria positiva (à direita).
- Isso significa que a cauda da distribuição está esticada para os valores mais altos.
- A média é maior que a mediana, e os dados estão concentrados nos valores mais baixos.

Essa assimetria pode indicar desigualdade de renda. Pode justificar políticas de redistribuição, como ampliação de transferências de renda (ex: Bolsa Família) ou isenções fiscais para os mais pobres.

Medidas de Formato

Curtose

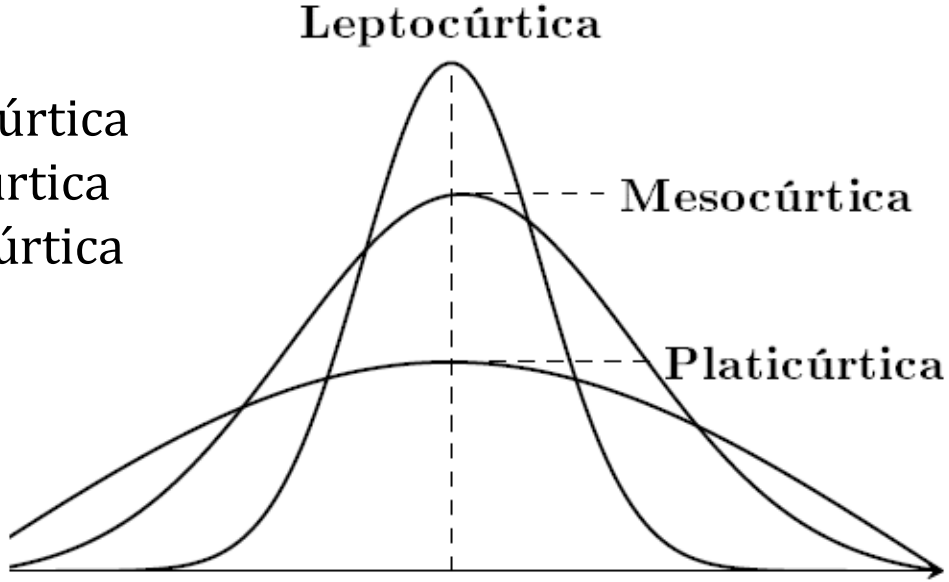
Dá-se o nome de curtose ao grau de “achatamento” da distribuição. Para medir o grau de curtose pode-se utilizar o seguinte coeficiente a seguir (também chamado de **coeficiente percentílico de curtose**):

$$K = \frac{Q_3 - Q_1}{2 * (P_{90} - P_{10})}$$

Medidas de Formato

Curtose

- $K = 0{,}263 \rightarrow$ Distribuição mesocúrtica
- $K > 0{,}263 \rightarrow$ Distribuição platicúrtica
- $K < 0{,}263 \rightarrow$ Distribuição leptocúrtica



Medidas de Formato

Exemplo: Em um estudo sobre inserção juvenil no mercado, a idade média de entrada é de 16 anos, e a maioria entra entre 15 e 17 anos, com poucos casos fora dessa faixa.

- A distribuição tem alta curtose (leptocúrtica): é mais "pontuda", com muitos valores concentrados próximos à média e caudas mais pesadas que o normal.
- Isso indica baixa dispersão em torno da média, mas a presença de alguns valores extremos (ex: entrada precoce com 12 anos ou tardia com 20 anos).

Pode indicar a necessidade de intervenção pontual: controle do trabalho infantil (casos extremos para menos) e programas de qualificação para retardar a entrada precoce. O formato da distribuição sugere que as políticas devem focar na faixa central, mas sem negligenciar os extremos.

Análise Bivariada

Análise Bivariada

A análise bivariada é uma etapa da análise estatística que estuda a relação entre duas variáveis simultaneamente.

O **objetivo** é identificar se há associação entre essas variáveis, bem como caracterizar essa relação (por exemplo, sua intensidade e direção).

Associação entre Variáveis Qualitativas

Associação entre Variáveis Qualitativas

Quando duas variáveis categóricas (qualitativas) são analisadas juntas, buscamos verificar se os valores de uma variável influenciam ou estão associados aos valores da outra.

Tabelas de Contingência: são tabelas que cruzam duas variáveis qualitativas, mostrando as frequências conjuntas de suas categorias.

Qui-Quadrado (χ^2): verifica se as diferenças observadas entre as frequências são estatisticamente significativas.

Tabelas de Contingência

Exemplo: Uma prefeitura implementou um programa de policiamento comunitário em alguns bairros da cidade e deseja avaliar se houve impacto na percepção de segurança da população. Foi realizada uma pesquisa com moradores de diferentes bairros, categorizando-os por participação no programa (Sim/Não) e percepção de segurança (Alta/Baixa).

Participação no Programa	Percepção Alta	Percepção Baixa	Total
Sim	120	30	150
Não	80	70	150
Total	200	100	300

Tabelas de Contingência

Participação no Programa	Percepção Alta	Percepção Baixa	Total
Sim	120	30	150
Não	80	70	150
Total	200	100	300

Interpretação:

- Entre os que participaram do programa, 80% (120/150) relataram alta percepção de segurança.
- Entre os que **não** participaram, apenas 53,3% (80/150) relataram alta percepção de segurança.

Qui-Quadrado (χ^2)

O tipo de tabela anterior permite aplicar o teste do qui-quadrado para verificar se a diferença na percepção de segurança é estatisticamente significativa, ou seja, se pode ser atribuída à participação no programa e não ao acaso.

As hipóteses são:

- Hipótese nula (H_0): As variáveis são independentes.
- Hipótese alternativa (H_1): Existe associação entre as variáveis.

Qui-Quadrado (χ^2)

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Onde:

- O: frequência observada
- E: frequência esperada (calculada assumindo independência)

Qui-Quadrado (χ^2)

Calculando o resultado Teste Qui-Quadrado para os dados da Tabela de Contingência entre participação no programa e percepção de segurança.

Hipóteses do Teste Qui-Quadrado:

- H_0: As variáveis são independentes (a percepção de segurança não está associada à participação no programa).
- H_1: As variáveis são dependentes (a percepção de segurança está associada à participação no programa).

Qui-Quadrado (χ^2)

Cálculo das Frequências Esperadas para cada célula:

$$E_{ij} = \frac{(Total\ da\ Linha) \times (Total\ da\ Coluna)}{Total\ Geral}$$

Para Sim / Alta:

$$E = \frac{150 \times 200}{300} = 100$$

Para Não / Alta:

$$E = \frac{150 \times 200}{300} = 100$$

Para Sim / Baixa:

$$E = \frac{150 \times 100}{300} = 50$$

Para Não / Baixa:

$$E = \frac{150 \times 100}{300} = 50$$

Qui-Quadrado (χ^2)

Cálculo da Estatística Qui-Quadrado:

Célula	O	E	$(O-E)^2/E$
Sim / Alta	120	100	$(20)^2/100$
Sim / Baixa	30	50	$(20)^2/50$
Não / Alta	80	100	$(20)^2/100$
Não / Baixa	70	50	$(20)^2/50$

$$\chi^2 = 4 + 8 + 4 + 8 = 24$$

Qui-Quadrado (χ^2)

Graus de Liberdade e Valor Crítico:

$$gl = (n_{linhas} - 1) \times (n_{colunas} - 1) = (2 - 1) \times (2 - 1) = 1$$

Usando uma tabela do qui-quadrado:

- Valor crítico para $\chi^2 = 24$ com $gl = 1$ e $\alpha = 0{,}05$;
- Valor crítico = 3,84;

Como $\chi^2 = 24 > 3{,}84$, rejeitamos H_0.

Qui-Quadrado (χ^2)

Conclusão:

Há evidências estatísticas significativas de que existe uma associação entre participação no programa de policiamento comunitário e percepção de segurança.

Ou seja, a percepção de segurança está associada à implementação do programa, sugerindo que o programa pode estar cumprindo seu objetivo.

Coeficiente de Contigência (C)

Mede a força da associação entre variáveis qualitativas. Vai de 0 (sem associação) a um valor máximo menor que 1 (dependente do número de categorias):

- Próximo de 0 → associação fraca ou inexistente;
- Próximo do máximo → associação forte.

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Coeficiente de Contigência (C)

Mede a força da associação entre variáveis qualitativas. Vai de 0 (sem associação) a um valor máximo menor que 1.

$$C = \sqrt{\frac{24}{24 + 300}} = \sqrt{0{,}0741} = 0{,}272$$

Embora o teste qui-quadrado tenha indicado que a associação é estatisticamente significativa, o valor C = 0,272 sugere que a força dessa associação é moderada, ou seja, existe relação entre participar do programa e sentir-se mais seguro, mas outros fatores podem estar envolvidos.

Associação entre Variáveis Quantitativas

Associação entre Variáveis Quantitativas

Quando analisamos duas variáveis numéricas, o objetivo é entender a forma, a intensidade e o sentido da relação entre elas.

Gráfico de Dispersão: Também conhecido como "scatter plot", mostra visualmente a relação entre duas variáveis quantitativas.

Correlação Linear: a correlação de Pearson mede o grau de associação linear entre duas variáveis quantitativas, e varia de entre -1 e 1.

Gráfico de Dispersão

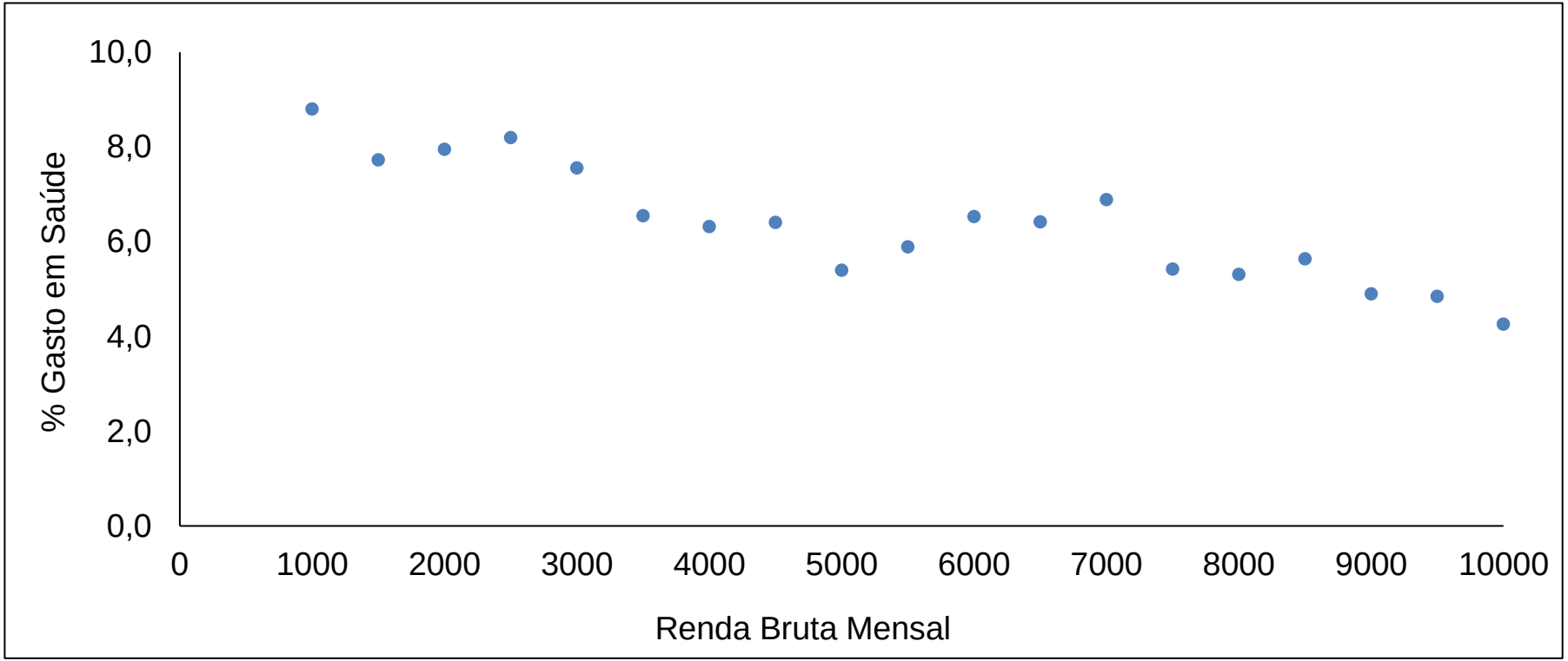
O gráfico de dispersão, também conhecido como diagrama de dispersão ou *scatter plot*, mostra visualmente a relação entre duas variáveis quantitativas.

Cada ponto do gráfico representa um par de valores observados:

- Pontos que seguem uma linha reta indicam forte correlação;
- Padrões positivos, negativos ou inexistentes podem ser observados.

Gráfico de Dispersão

Exemplo: Renda bruta mensal familiar e percentual da renda gasto em saúde - Brasil (2015)



Fonte: Dados fictícios (elaborado para entendimento da relação entre variáveis quantitativas).

Correlação Linear

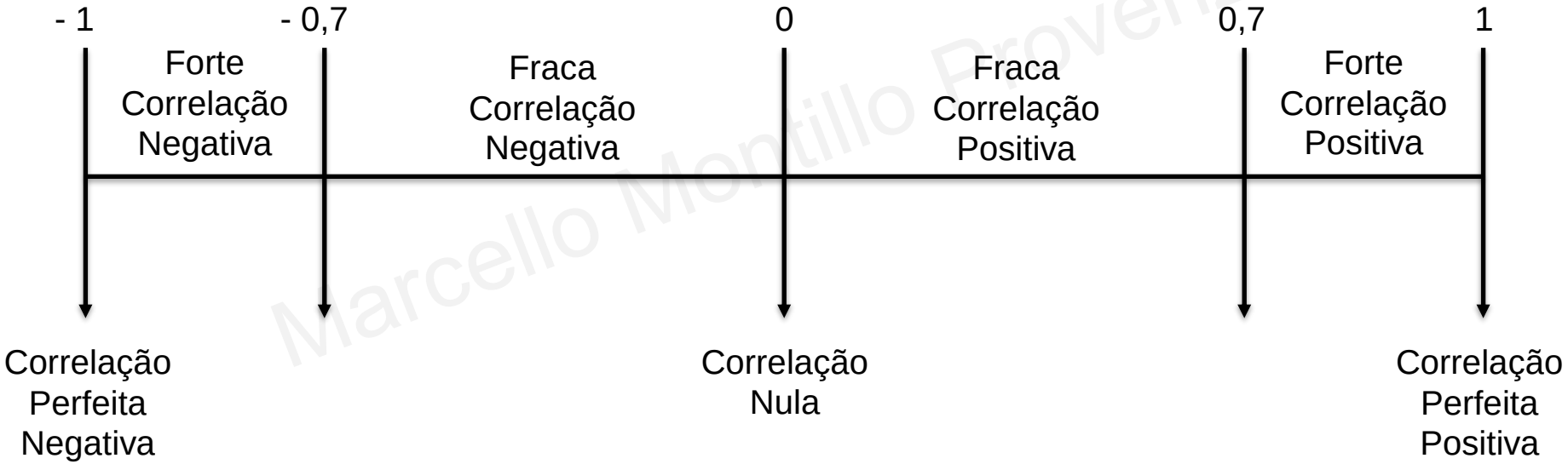
A correlação linear é uma análise estatística que mede a relação entre duas variáveis quantitativas. Ela pode ser positiva, negativa, inexistente ou perfeita.

O coeficiente de correlação linear de Pearson (r) mede a força e a direção da correlação. O valor de r varia entre -1 e +1.

- Um valor de r = 0 indica que não há associação entre as variáveis;
- Um valor de r = 1 indica uma correlação perfeita positiva;
- Um valor de r = -1 indica uma correlação perfeita negativa.

Correlação Linear

Intensidade do coeficiente de correlação linear de Pearson (r)



Correlação Linear

Cálculo do coeficiente de correlação linear de Pearson (r):

$$r = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}$$

Correlação Linear

Exemplo: Suponha que um pesquisador deseja analisar se existe uma associação linear entre o investimento público por aluno e a nota média no ENEM em diferentes estados brasileiros.

Estado	Investimento por Aluno	Nota Média no ENEM
A	4.000	520
B	5.000	540
C	6.000	560
D	7.000	580
E	8.000	590

Correlação Linear

Estado	Invest. (x)	Nota (y)	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
A	4.000	520	76.000	4.000.000	1.444
B	5.000	540	18.000	1.000.000	324
C	6.000	560	0	0	4
D	7.000	580	22.000	1.000.000	484
E	8.000	590	64.000	4.000.000	1.024
Média	**6.000**	**558**	**-**	**-**	**-**
Total	**-**	**-**	**180.000**	**10.000.000**	**3.280**

Correlação Linear

Resolvendo os cálculos:

$$\bar{x} = 6.000$$
$$\bar{y} = 558$$
$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = 180.000$$
$$\Sigma(x_i - \bar{x})^2 = 10.000.000$$
$$\Sigma(y_i - \bar{y})^2 = 3.280$$

Resolvendo o coeficiente de Pearson:

$$r = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}} = \frac{180.000}{\sqrt{10.000.000 \times 3.280}} = \frac{180.000}{181.107,70} = 0,99$$

Correlação Linear

O valor de r = 0,99 indica forte correlação linear positiva entre o investimento por aluno e a nota no ENEM, ou seja, à medida que o investimento aumenta, as notas também aumentam.

É importante destacar que correlação não implica causalidade. Um alto investimento pode estar associado a melhores notas, mas outros fatores como qualificação de professores, infraestrutura, condições sociais, etc. também influenciam.

Associação entre uma Variável Quantitativa e outra Qualitativa

Associação entre uma Variável Quantitativa e outra Qualitativa

Nesses casos, investiga-se como a variável quantitativa se comporta em diferentes grupos definidos pela variável qualitativa.

- **Gráfico de Médias:** mostra a média da variável quantitativa em cada grupo da variável qualitativa.
- **Gráfico de Perfis:** linhas conectam as médias entre grupos, facilitando a visualização de padrões.
- **Boxplot Comparativo:** permite comparar a distribuição de uma variável quantitativa entre dois ou mais grupos categóricos.

Gráfico de Médias

Gráficos de Médias (também chamados de gráficos de barras de médias) são uma forma de representar graficamente a associação entre uma variável quantitativa e uma qualitativa.

- Uma variável qualitativa que define grupos ou categorias (sexo, região, tipo de escola).
- Uma variável quantitativa que você quer comparar entre esses grupos (salário, nota, tempo de espera).

O gráfico de médias mostra, para cada categoria da variável qualitativa, a média da variável quantitativa correspondente.

Gráfico de Médias

Usado quando se pretende comparar médias entre dois ou mais grupos ou quando se deseja resumir informações numéricas por categorias.

É útil em estudos de impacto de políticas públicas, por exemplo:

- Média de crimes por região.
- Média de internações por tipo de hospital.
- Média de renda por nível de escolaridade.

Gráfico de Médias

Exemplo: Suponha que você quer analisar o desempenho médio de alunos no ENEM por tipo de escola.

Tipo de Escola	Nota Média no ENEM
Pública	510
Privada	620

Gráfico de Médias

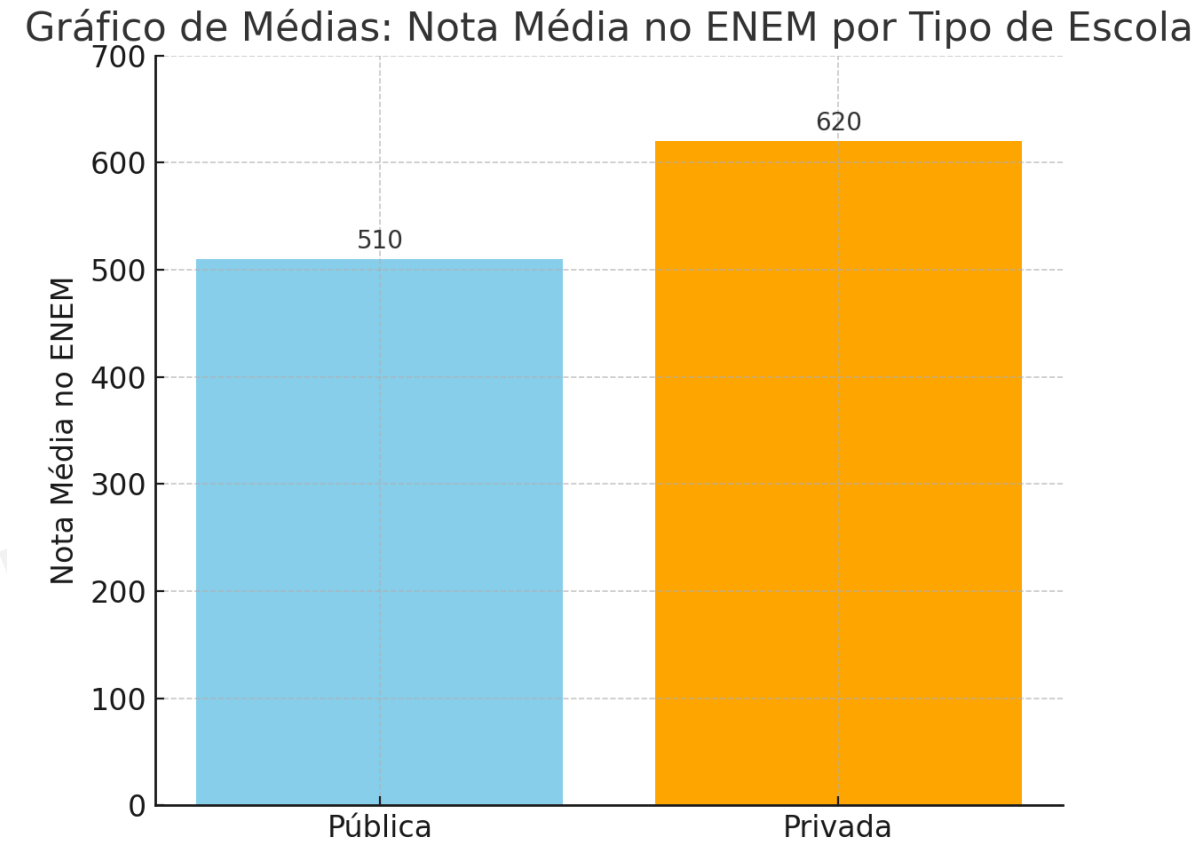


Gráfico de Médias

O gráfico mostra que alunos de escolas privadas obtiveram, em média, notas mais altas no ENEM do que alunos de escolas públicas.

Essa diferença de desempenho sugere possíveis desigualdades no acesso à qualidade de ensino entre as redes de educação, podendo refletir diferenças em infraestrutura, recursos didáticos e apoio pedagógico.

Gráfico de Perfis

Os gráficos de perfis são uma extensão dos gráficos de médias, usados especialmente quando:

- Você tem duas ou mais variáveis qualitativas, ou;
- Uma variável qualitativa com mais de dois níveis, e;
- Deseja comparar como as médias de uma variável quantitativa variam entre os grupos.

Em vez de barras, como nos gráficos de médias, os gráficos de perfis usam linhas conectando os pontos médios das categorias - um ponto para cada categoria da variável qualitativa, e uma linha para cada grupo comparado. É muito útil quando se quer comparar padrões ou tendências entre grupos.

Gráfico de Perfis

Interpretação:

- Se as linhas forem paralelas, os grupos têm padrões semelhantes.
- Se as linhas se cruzarem ou divergirem, isso indica interação ou diferença nos padrões.

Esses gráficos são muito usados em análise de variância (ANOVA) e estudos de políticas públicas, por exemplo:

- Avaliar se o efeito de um programa educacional muda conforme o ano escolar.
- Verificar se a evolução da renda média difere entre regiões.

Gráfico de Perfis

Exemplo: Imagine que você está analisando a nota média no ENEM (variável quantitativa) por ano escolar (1º, 2º, 3º ano) e por tipo de escola (pública e privada).

Ano Escolar	Pública	Privada
1º Ano	500	580
2º Ano	520	600
3º Ano	540	620

Gráfico de Perfis

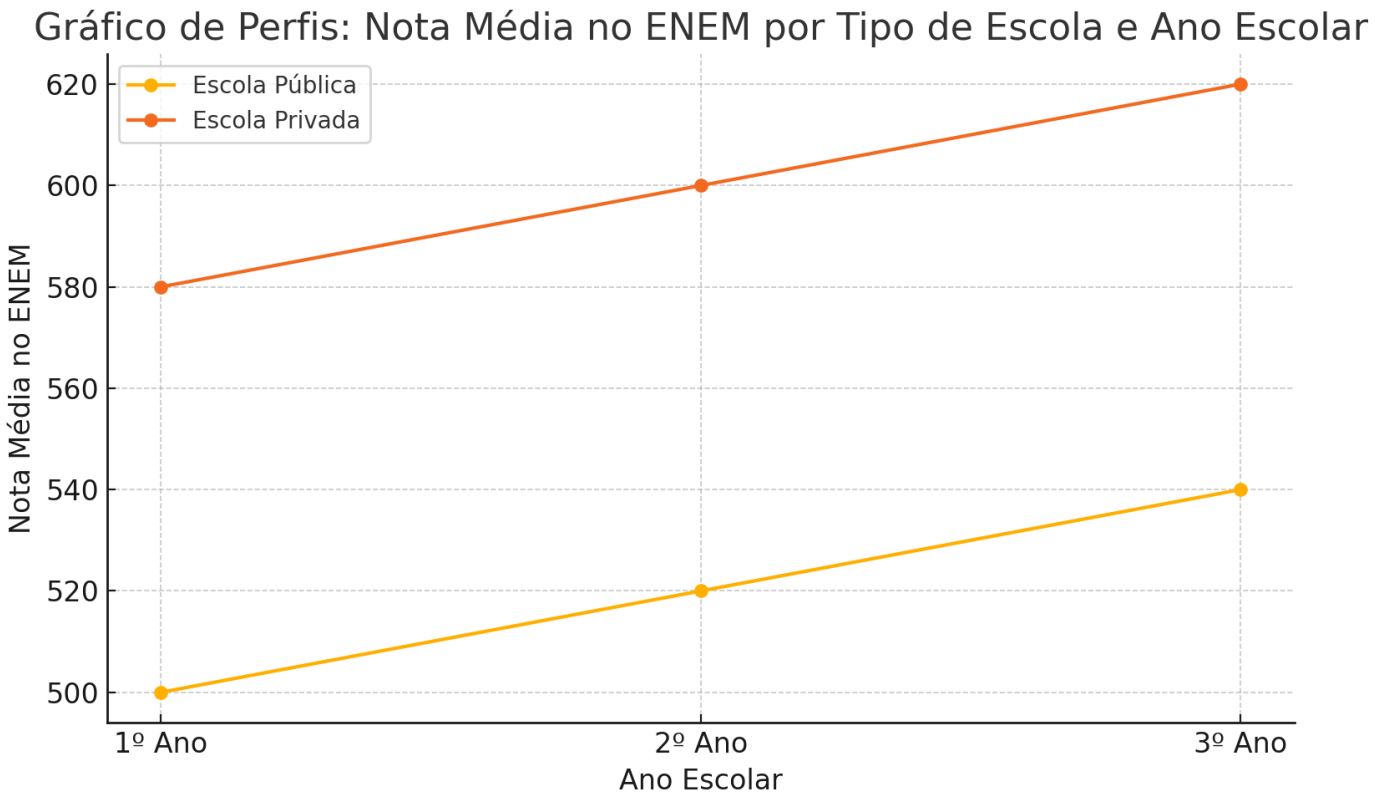


Gráfico de Perfis

O gráfico de perfis evidencia que:

- O ano escolar está associado a um aumento nas notas médias no ENEM, o que é esperado.
- O tipo de escola tem impacto importante sobre o desempenho dos estudantes, com vantagem para a rede privada.
- As trajetórias paralelas indicam que as duas redes evoluem de forma semelhante, mas com níveis médios diferentes.