



Preditiva.ai

Estatística Introdução

Com a **crescente quantidade de dados disponíveis** em todo o mundo, é cada vez maior também o interesse em utilizar esses dados para responder questões importantes de forma **objetiva e robusta**.

A “**quantificação**” de nossas vidas acontecem nos mais variados meios, como por exemplo:

- Utilização de redes sociais
- Compras em e-commerce
- Utilização de serviços de streaming
- Entre outros

Com a **evolução da computação**, fomos conseguindo tabular e resumir esses dados de forma muito mais eficiente do que antes.

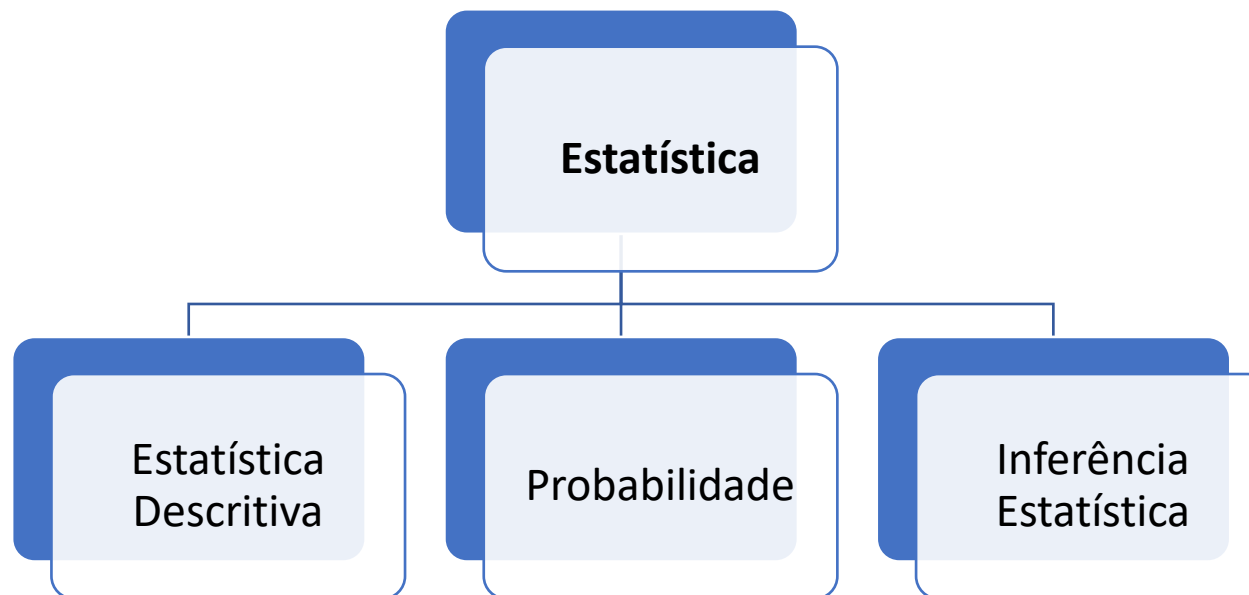
No entanto, **essa facilitação trouxe o risco da má interpretação** desses dados, pois pessoas sem o devido conhecimento dos **conceitos** utilizados nas **técnicas analíticas** podem ser induzidas ao erro.



Este conjunto de **técnicas analíticas** pode ser definido como:

O conjunto de técnicas que permite, **organizar, descrever, analisar e interpretar os dados** oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento, é chamado de **Estatística**.

A **Estatística** pode ser dividida em três áreas:



O primeiro passo para tirar conclusões sobre os dados envolve o seu resumo e descrição. Portanto, a **Estatística Descritiva** pode ser interpretada como o **conjunto de técnicas utilizadas para descrever e resumir um conjunto de dados**.

Um exemplo de utilização da **Estatística Descritiva**, é a construção de uma tabela de frequência com as notas de 311 alunos de uma escola. Nessa tabela, temos as faixas de notas e a frequência observada.

ID	Aluno	Nota
1	João	17
2	Mariana	89
3	Carlos	50
4	Juliana	3
5	George	69
6	Luiz	90
7	Rosa	41
...
311	Bianca	99



Faixa de nota	Frequência Observada
De 0 a 70	30
De 71 a 89	80
De 90 a 99	154
100	47
Total	311

A **Probabilidade** pode ser pensada como a **teoria matemática utilizada** para se **estudar a incerteza oriunda de fenômenos de caráter aleatório**. Denominamos fenômeno aleatório a situação ou acontecimento cujos resultados **não podem ser previstos com certeza**.

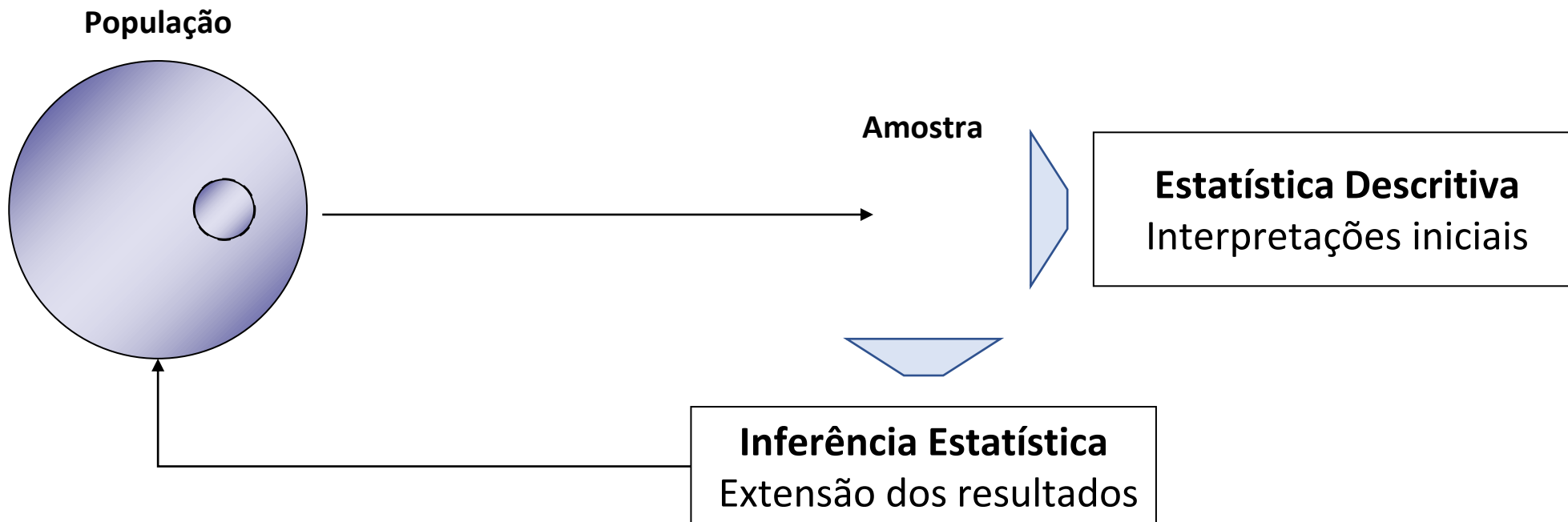
Exemplos:

- ☐ Previsão de Tempo para hoje e próximos dias.
- ☐ Quantidade de pessoas que utilizarão o metro em São Paulo em uma 2ª-feira comum.
- ☐ As dezenas sorteadas da Mega Sena da Virada.

No entanto, **podemos utilizar técnicas que nos permitem quantificam nossa incerteza**.

Exemplos:

- ☐ 23% de probabilidade de chover na Zona Sul de São Paulo.
- ☐ 85% de probabilidade de que mais de 3 milhões de pessoas utilizarão o metro de SP em uma 2ª-feira.
- ☐ 0,0001% de probabilidade da Mega Sena da Virada sair para apenas um ganhador.



Inferência é a área da Estatística responsável pelo conjunto de técnicas que possibilitam que as conclusões obtidas a partir de uma **amostra**, sejam estendidos para a **população**.

Essas técnicas são fundamentais quando **não temos acesso à população** de interesse, como por exemplo:

- Novo medicamento
- Fábrica de lâmpadas

Revisão

Nesta introdução vimos que diversos eventos do nosso dia a dia são **quantificados**, gerando muitas oportunidades para a **análise desses dados**.

Apresentamos também as **3 principais áreas da estatística** que estão diretamente relacionadas com os **fundamentos de análise de dados**.

Por fim, a Estatística é um assunto fantástico e uma área bem extensa, por isso exige alguns cuidados:

- O **desafio é crescente**, ou seja, é importante entender bem um tema antes de passar para o próximo.
- Tenha **paciência e disposição** para **rever conceitos** que podem ter vindo do trabalho ou da graduação **distorcidos**. Aproveite a oportunidade para **tirar todas as dúvidas!**





Preditiva.ai

Estatística Descritiva


Tipos de Variáveis


Um conjunto de dados pode ser organizado em formato de tabela. Quando isso acontece, os dados são ditos “**estruturados**”. Vejamos um exemplo:


Observações (linhas da tabela)	ID	Nome	Escolaridade	Idade	Salário
	2	Luis	Ensino Médio	49	5.130,80
	4	Helena	Mestrado	33	4.193,52
	5	João	Ensino Técnico	27	3.468,35
Variáveis (colunas da tabela)					


Nesta tabela exemplo existem diferentes tipos de variáveis:

ID	Nome	Escolaridade	Idade	Salário
2	Luís	Ensino Médio	49	5.130,80
4	Helena	Mestrado	33	4.193,52
5	João	Ensino Técnico	27	3.468,35


Variável Qualitativa Nominal

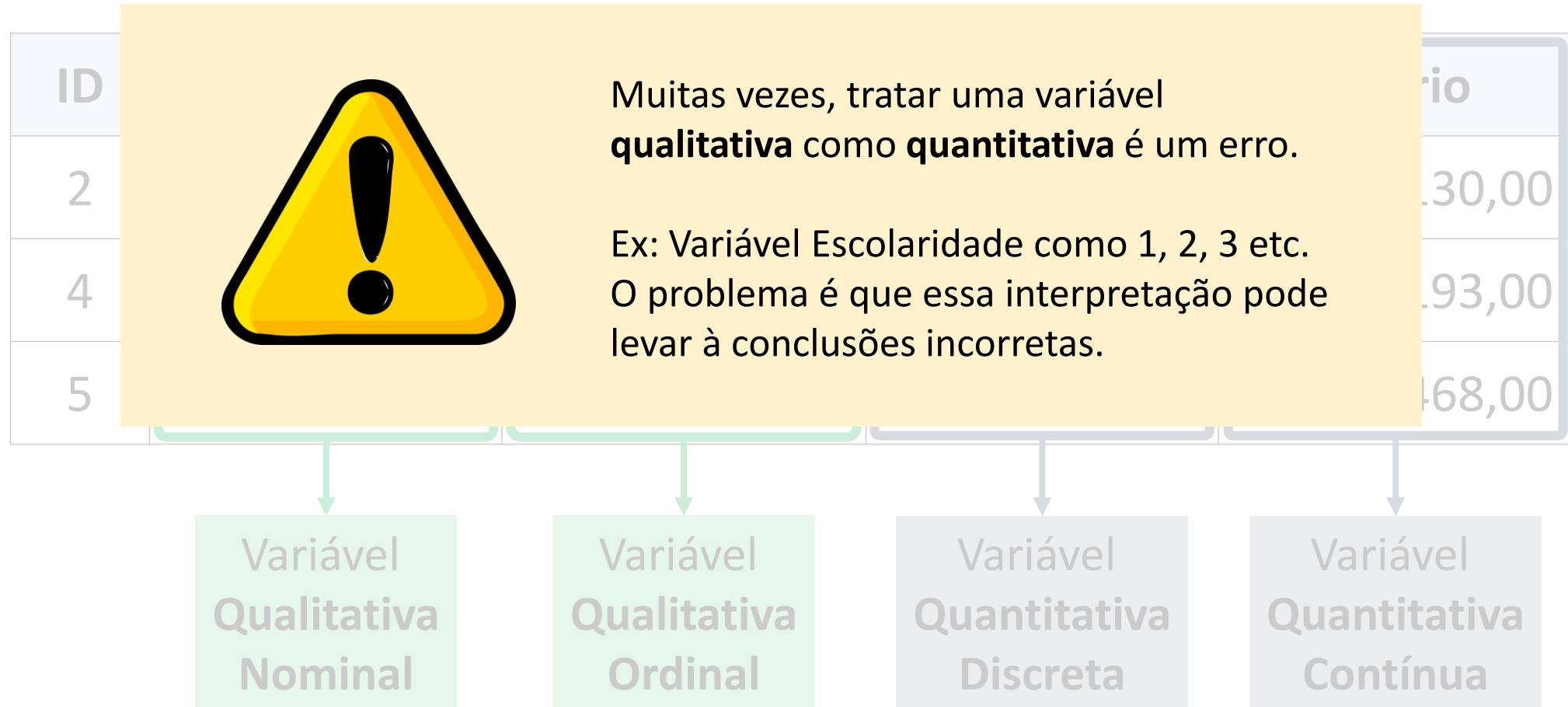

Variável Qualitativa Ordinal


Variável Quantitativa Discreta

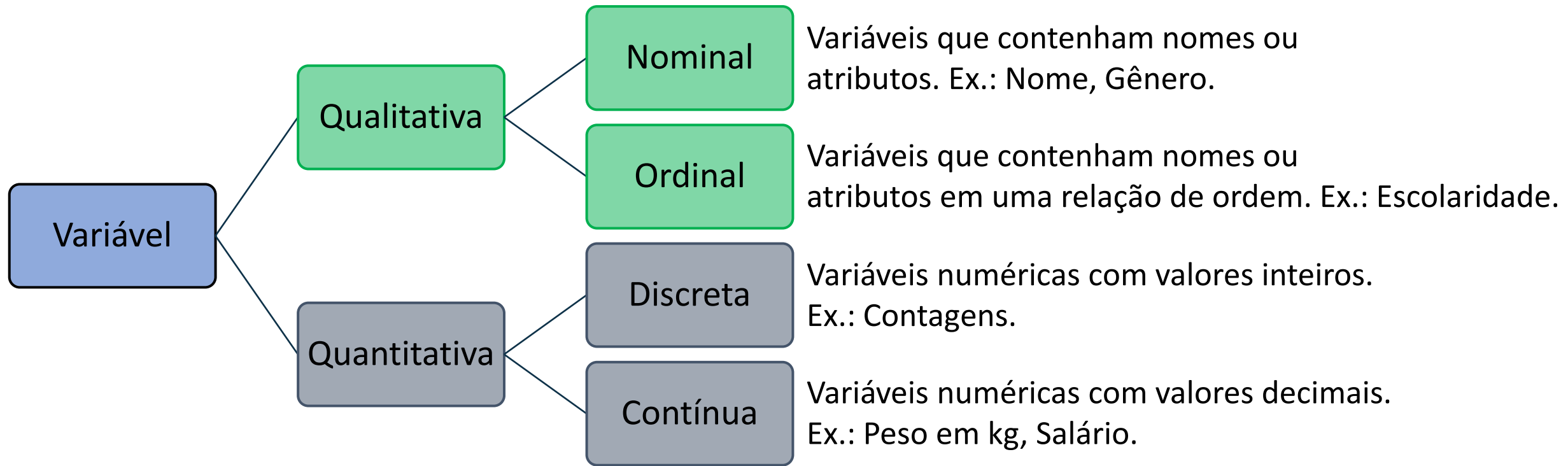

Variável Quantitativa Contínua



Nesta tabela exemplo existem diferentes tipos de variáveis:



Resumindo, as variáveis de um conjunto de dados diferem em relação a sua “natureza”. Basicamente, temos **2 tipos de variáveis, com 2 subtipos cada**. São elas:

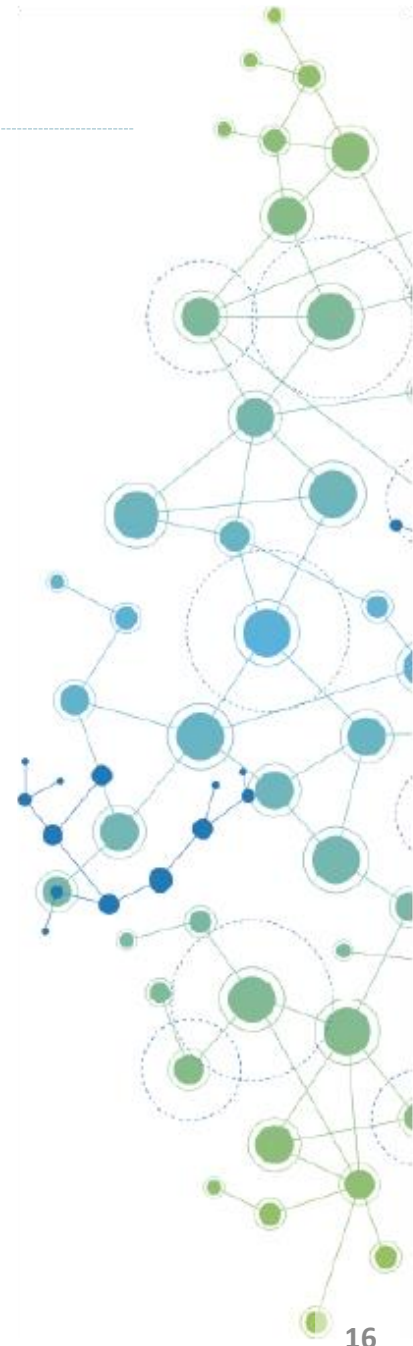


Revisão

Vimos que as variáveis são classificadas em 2 tipos principais: **Qualitativas** e **Quantitativas**. E que cada um desses tipos possui 2 subtipos.

Além disso, vimos que tratar uma variável **Qualitativa** como uma variável **Quantitativa** pode gerar **distorções** que irão comprometer os resultados da análise.

A seguir, veremos como selecionamos as técnicas mais adequadas para analisar cada **Tipo de Variável**.





Preditiva.ai

Estatística Descritiva

Tabelas de Frequência

Estatística Descritiva

Tabelas de Frequência



Uma das funções da **Estatística Descritiva** é a de **resumir um conjunto de dados** para **facilitar sua análise**. E as **Tabelas de Frequência** são muito úteis para atingir este objetivo. Veja o exemplo abaixo:

Nome	Escolaridade	Idade	Salário
Luis	Ensino Médio	49	5.130,00
Helena	Ensino Médio	33	4.193,00
João	Ensino Médio	27	3.468,00
Julio	Ensino Técnico	32	3.068,00
Mariana	Ensino Superior	59	2.670,00
Augusto	Ensino Médio	30	2.693,00
Gustavo	Ensino Superior	38	9.526,00
Cecilia	Ensino Técnico	29	3.068,00
Kaian	Ensino Superior	36	5.237,00
Ana	Mestrado	29	9.980,00

Resumo da variável
Escolaridade



Escolaridade	Frequência absoluta
Ensino Médio	4
Ensino Técnico	2
Ensino Superior	3
Mestrado	1
Total	10



A **Frequência absoluta** é a contagem de vezes que o valor de uma variável ocorre em um conjunto de dados.

Estatística Descritiva

Tabelas de Frequência

A **Tabela de Frequência** pode ser aprimorada com a inclusão de outras duas frequências: **Frequência Relativa** e **Frequência Acumulada**. Veja o exemplo abaixo:

Escolaridade	Frequência absoluta	Frequência relativa	Frequência acumulada
Ensino Médio	4	40%	40%
Ensino Técnico	2	20%	60%
Ensino Superior	3	30%	90%
Mestrado	1	10%	100%
Total	10	100%	

A **Frequência Relativa** mostra a quantidade de vezes que um valor aparece no conjunto dos dados em relação ao total de observações.
No exemplo acima, **40% das pessoas têm o Ensino Médio** ($40\% = 4/10$).

A **Frequência Acumulada** mostra a soma das frequências relativas até um determinado valor do conjunto de dados. No exemplo ao lado, **60% das pessoas têm escolaridade ATÉ o Ensino Técnico**.

Ou seja, $60\% = 40\% \text{ (Ensino Médio)} + 20\% \text{ (Ensino Técnico)}$.

Estatística Descritiva

Tabelas de Frequência

E se quiséssemos resumir a variável **Salário**? Seria interessante simplesmente criar uma **Tabela de Frequência**? Vejamos no exemplo abaixo:

Nome	Escolaridade	Idade	Salário
Luis	Ensino Médio	49	5.130,00
Helena	Ensino Médio	33	4.193,00
João	Ensino Médio	27	3.468,00
Julio	Ensino Técnico	32	3.068,00
Mariana	Ensino Superior	59	2.670,00
Augusto	Ensino Médio	30	2.693,00
Gustavo	Ensino Superior	38	9.526,00
Cecilia	Ensino Técnico	29	3.068,00
Kaian	Ensino Superior	36	5.237,00
Ana	Mestrado	29	9.980,00

Resumo da variável
Salário



Salário	Frequência absoluta
2.670,00	1
2.693,00	1
3.068,00	2
3.468,00	1
4.193,00	1
5.130,00	1
5.237,00	1
9.526,00	1
9.980,00	1
Total	10

Estatística Descritiva

Tabelas de Frequência

Como podemos observar, o **resumo** dessa variável **não ficou** muito **informativo**. Isso acontece com **variáveis que contêm muitos valores possíveis**, como as **variáveis quantitativas** em geral.

Salário	Frequência absoluta
2.670,00	1
2.693,00	1
3.068,00	2
3.468,00	1
4.193,00	1
5.130,00	1
5.237,00	1
9.526,00	1
9.980,00	1
Total	10

Para melhorar o resumo de variáveis desse tipo, podemos criar **“Faixas de Valores”**.
Veja:



Salário	Frequência absoluta
2.000,00 -- 4.000,00	5
4.000,00 -- 6.000,00	3
8.000,00 -- 10.000,00	2
Total	10

A diferença entre o valor inicial e final da faixa de valor é chamado de **amplitude do intervalo**. No exemplo, a amplitude é de 2.000,00 reais.

Revisão

Vimos que as **Tabelas de Frequência** são simples e bastante úteis para realizarmos o primeiro passo em uma análise de dados.

Vimos também que quando a variável for **quantitativa** é necessário **criar faixas** para que as informações sejam resumidas **adequadamente** na tabela. A **quantidade de faixas** e a sua **amplitude** devem ser escolhidas de forma a facilitar o atingimento do **objetivo da análise**.



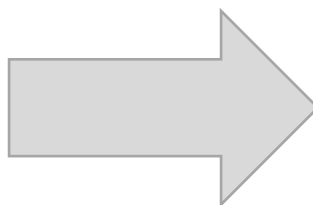
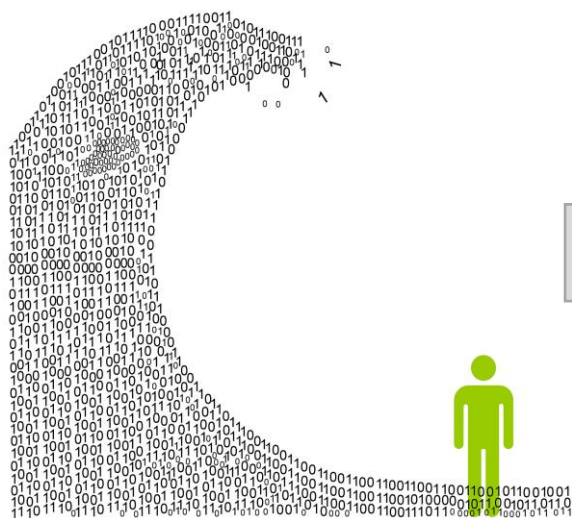


Preditiva.ai

Estatística Descritiva

Medidas Resumo

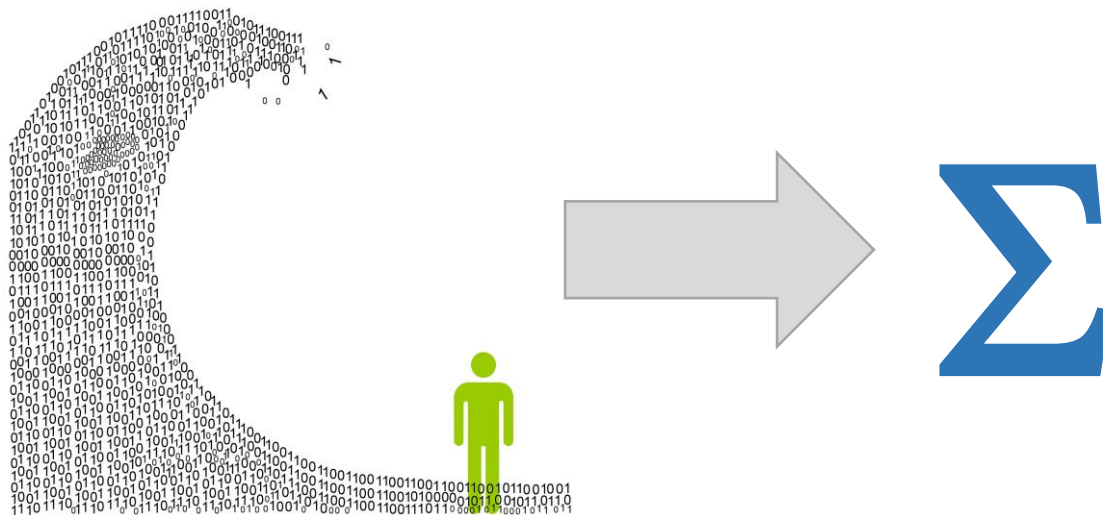
Em grandes bases de dados, com **centenas de variáveis** e **milhares ou mesmo milhões de observações**, qual é a **técnica** que pode nos ajudar a compreender esse grande volume de dados?



Acertou quem disse **Estatística Descritiva!**

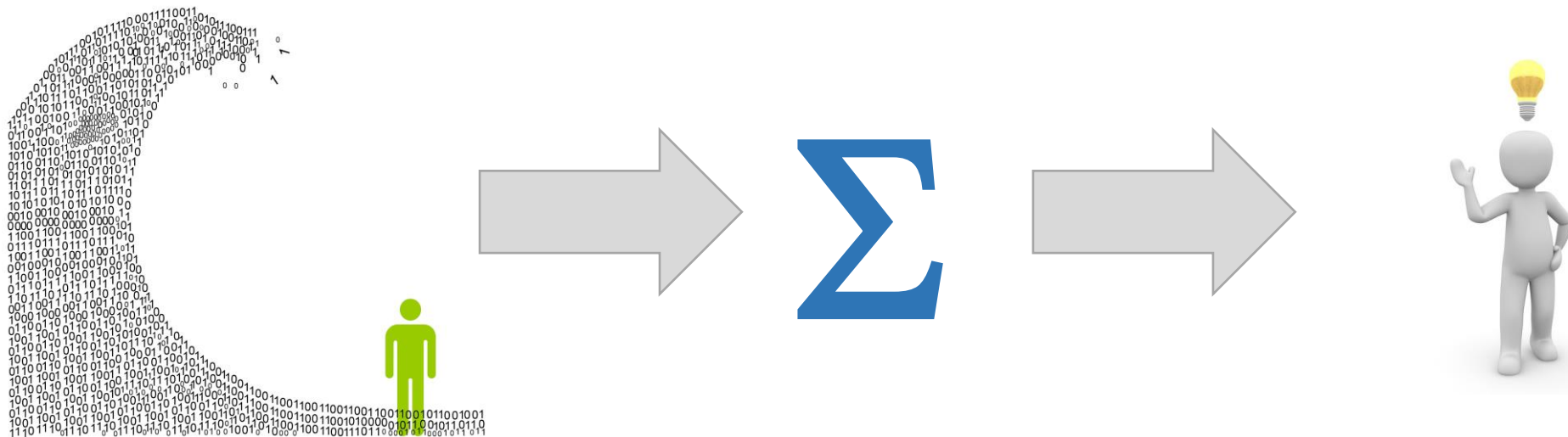
Essa é a técnica que nos permitirá, principalmente:

Resumir um grande volume de observações em diferentes valores
utilizando as chamadas **Medidas Resumo**



Essas **Medidas Resumo** são bastante úteis para:

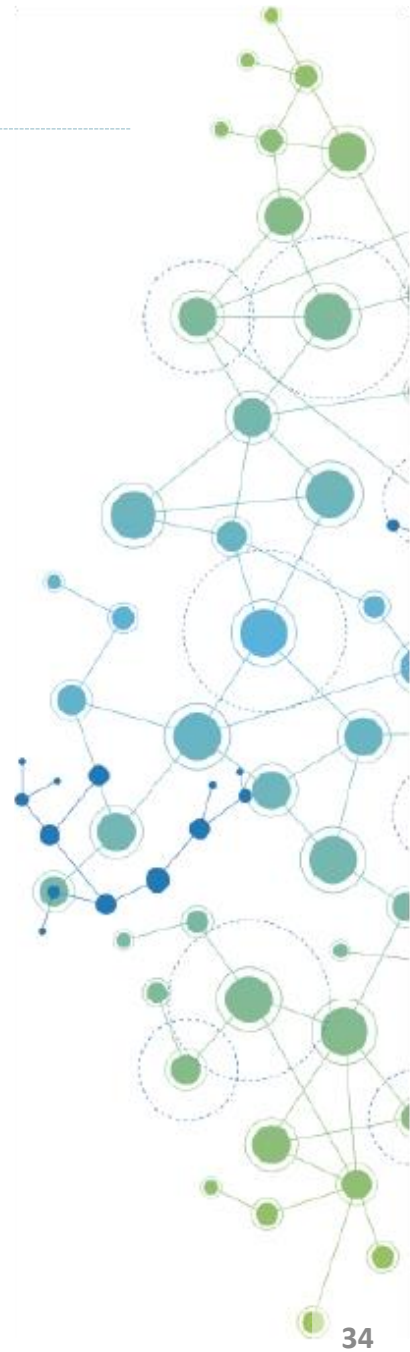
- **Caracterizar** o conjunto de dados.
- **Realizar comparações** entre diferentes conjuntos de dados ou grupos dentro do mesmo conjunto de dados.
- **Compreender** quais informações estão contidas nesse conjunto de dados.



As **Medidas Resumo** são agrupadas em 2 principais grupos:

- **Medidas de Posição:** indicam posições de referência
 - Média
 - Mínimo e Máximo
 - Moda
 - Mediana
 - Quartis
- **Medidas de Dispersão:** indicam a variabilidade
 - Variância
 - Desvio Padrão

Medidas de Posição



A **Média** é a medida resumo mais utilizada e conhecida para representar um conjunto de dados. Com certeza você já ouviu algo como:

- A média de salários em uma empresa de Tecnologia é de R\$ 5.000,00.
- O tempo médio de duração de um filme é de 2 horas.

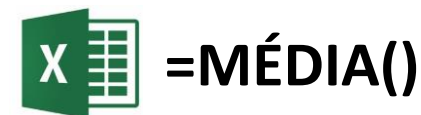
Forma de cálculo da **Média**: somar todos os valores e dividir essa soma pela quantidade de valores:

$$\text{Média} = \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

ID	Salário
1	5.130,00
2	4.193,00
3	3.468,00
4	3.068,00
5	2.670,00
6	2.693,00
7	9.526,00
8	3.068,00
9	5.237,00
10	9.980,00
11	2.426,00
12	2.911,00

A **Média** representa qual seria o valor do salário se o valor **total** de remuneração desses colaboradores fosse **distribuído uniformemente** entre eles.

$$\text{Média} = \frac{54.370,00}{12} = 4.530,83$$



Porém, muitas vezes **ela não é a medida resumo mais indicada**. Veremos isso mais adiante.



O que você faria na seguinte situação?

Você é convidado para trabalhar em uma startup com 15 funcionários e, segundo o RH o salário **médio** dos funcionários é R\$4.200,00.

Você atualmente ganha R\$1.000,00.

Funcionário	Salário
1	1.000,00
2	1.000,00
3	1.000,00
4	1.000,00
5	1.000,00
6	1.000,00
7	1.000,00
8	1.000,00
9	1.000,00
10	1.000,00
11	1.000,00
12	1.000,00
13	1.000,00
14	10.000,00
15	40.000,00
Média	4.200,00

ID	Salário
11	2.426,00
5	2.670,00
6	2.693,00
12	2.911,00
4	3.068,00
8	3.068,00
3	3.468,00
2	4.193,00
1	5.130,00
9	5.237,00
7	9.526,00
10	9.980,00

Diagram illustrating the range (Amplitude) of the salary variable. A vertical double-headed arrow labeled "Amplitude" spans from the minimum value (2.426,00) to the maximum value (9.980,00). Arrows point from the highlighted minimum and maximum values in the table to the labels "Mínimo" and "Máximo" respectively.

Ao ordenar a variável Salário, teremos o **Mínimo** na 1ª posição e o **Máximo** na última.

Nesse exemplo, além do salário **mínimo** de R\$ 2.426,00 e do salário **máximo** de R\$ 9.980,00, temos a informação de que a **amplitude da variável** salário é de R\$ 7.554,00, que é o valor máximo menos o valor mínimo.



=MÍNIMO()



=MÁXIMO()

ID	Salário
11	2.426,00
5	2.670,00
6	2.693,00
12	2.911,00
4	3.068,00
8	3.068,00
3	3.468,00
2	4.193,00
1	5.130,00
9	5.237,00
7	9.526,00
10	9.980,00

Mediana
3.268,00

A **Mediana** representa a posição central do conjunto de dados.

Ou seja, 50% dos valores são inferiores à **Mediana** e 50% dos valores são superiores à **Mediana**.

Quando o número de observações é par, a **Mediana** é a média entre os 2 valores centrais.

 =MED()

Estatística Descritiva

Medidas Resumo: Medidas de Posição - Mediana



Mediana = 3.268,00

Média = 4.530,83

ID	Salário
11	2.426,00
5	2.670,00
6	2.693,00
12	2.911,00
4	3.068,00
8	3.068,00
3	3.468,00
2	4.193,00
1	5.130,00
9	5.237,00
7	9.526,00
10	9.980,00

Mediana = 3.468,00

Média = 6.105,38

ID	Salário
11	2.426,00
5	2.670,00
6	2.693,00
12	2.911,00
4	3.068,00
8	3.068,00
3	3.468,00
2	4.193,00
1	5.130,00
9	5.237,00
7	9.526,00
10	9.980,00
13	25.000,00

A **Mediana** também é bastante utilizada como alternativa a **Média** por ser menos sensível a valores extremos.

Adicionando 1 colaborador com salário de R\$25.000,00, vemos que a **Mediana** aumentou R\$200,00 enquanto a **Média** aumentou R\$1.574,55.

O que você faria na seguinte situação?

Você está doente e só existe um remédio que pode te ajudar. Segundo a bula, o tempo de sobrevida **mediano** é de 8 semanas. Todos que tomam o remédio tem severos efeitos colaterais. Vale a pena tomar o remédio?

Tempo de Sobrevida (em semanas)	% Relativa
2	15%
4	15%
8	20%
16	5%
32	5%
64	5%
128	5%
256	10%
512	20%

ID	Salário
11	2.426,00
5	2.670,00
6	2.693,00
12	2.911,00
4	3.068,00
8	3.068,00
3	3.468,00
2	4.193,00
1	5.130,00
9	5.237,00
7	9.526,00
10	9.980,00

Diagram illustrating the distribution of salaries into four quartiles (25% each):

- 1º Quartil: 2.856,50
- Mediana: 3.268,00
- 3º Quartil: 5.156,75

Os **Quartis** representam posições específicas que permitem identificar como os dados estão distribuídos. Os Quartis têm esse nome porque dividem o conjunto de dados em 4 partes:

- **1º Quartil**: indica que 25% das observações têm valores inferiores a ele.
- **2º Quartil** ou **Mediana**: indica que 50% das observações têm valores inferiores a ele e 50% valores superiores a ele.
- **3º Quartil**: indica que 75% das observações têm valores inferiores a ele.

 =QUARTIL.INC()

Estatística Descritiva

Medidas Resumo: Medidas de Posição



Conjunto A
1
3
5
7
9

Mediana = 5
Média = 5

Conjunto B
5
5
5
5
5

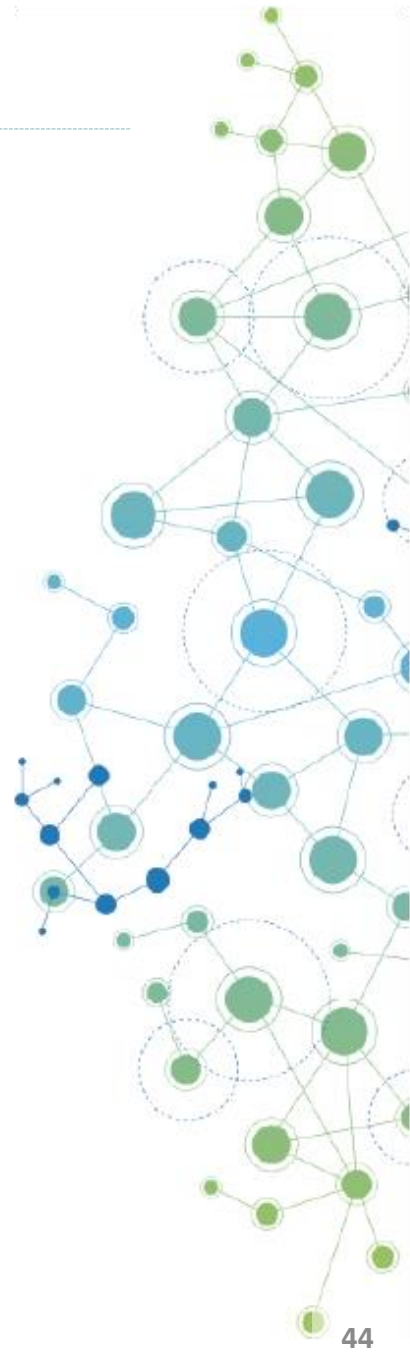
Mediana = 5
Média = 5

Agora tente resumir os conjuntos de dados A e B ao lado usando a **Média** e a **Mediana**.

Neste exemplo observamos que as **medidas de posição nem sempre conseguem resumir** todas as características de um conjunto de dados.

Desta forma, precisamos avaliar outra característica dos dados: **a variabilidade** ou dispersão.

Medidas de Dispersão

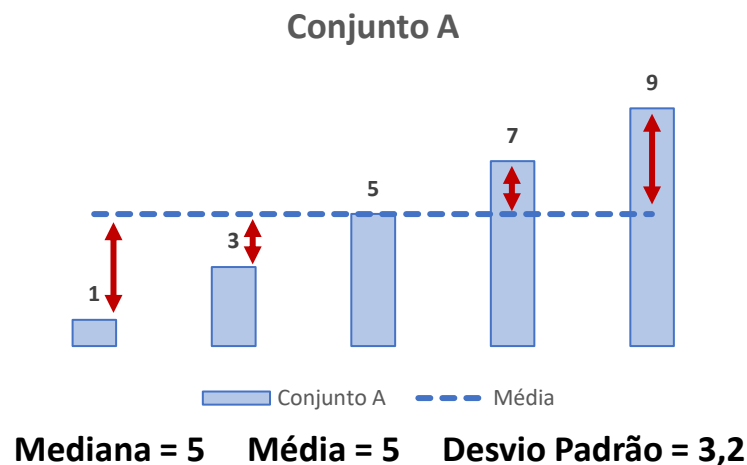


Estatística Descritiva

Medidas Resumo: Medidas de Dispersão - Desvio Padrão

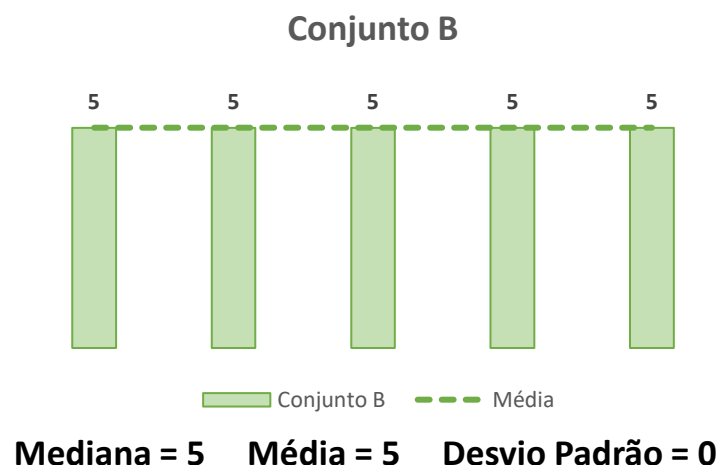


Preditiva.ai



O **Desvio Padrão** representa o quão dispersas estão as observações do conjunto de dados.

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



Ou seja, quanto maior for a diferença entre a média e cada valor, mais dispersas estarão as observações e por consequência, maior será o **Desvio Padrão**.



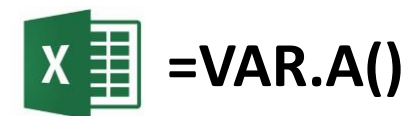
=DESVPAD.A()

A **Variância** é o **Desvio Padrão** ao quadrado, e também representa a dispersão das observações no conjunto de dados, porém em na escala ao quadrado.

$$\text{Variância} = \text{Desvio Padrão}^2$$

Ou seja, se estamos medindo valores em reais, a **Variância** será uma medida em reais ao **quadrado**.

Na prática, a **Variância** não é tão utilizada na Estatística Descritiva, porém é fundamental em diversos métodos estatísticos que veremos mais adiante.



Revisão

Nesta aula aprendemos quais são as principais **Medidas Resumo** de **posição** e **dispersão** utilizadas em uma **análise de dados**.

Vimos também que em algumas situações, apenas as **Medidas Resumo** de **posição** não são **suficientes** para descrever os conjuntos de dados de forma a identificar suas diferenças.

E que uma forma bastante interessante para **representar graficamente a distribuição dos dados**.





Preditiva.ai

Estatística Descritiva

Visualizando um conjunto de dados

Estatística Descritiva

Visualização de dados

Embora a Tabela de Frequência seja um recurso eficaz no resumo de uma grande quantidade de dados, o **uso de gráficos facilita ainda mais a interpretação** do “comportamento” desses dados. Vejamos a seguir, as alternativas gráficas mais utilizadas na Estatística Descritiva.

Escolaridade	Frequência absoluta	Frequência relativa
Ensino Médio	4	40%
Ensino Técnico	2	20%
Ensino Superior	3	30%
Mestrado	1	10%
Total	10	100%

Resumo gráfico
da variável
Escolaridade



Percentual de Colaboradores por
Escolaridade

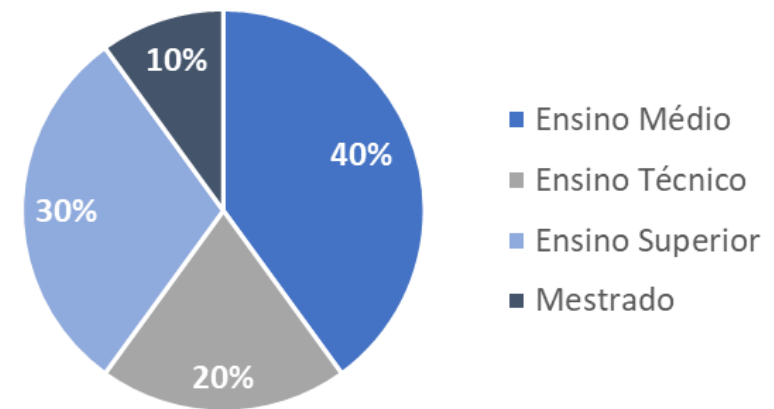


Gráfico de Pizza

Uso: Frequências relativas de variáveis qualitativas com poucas categorias.

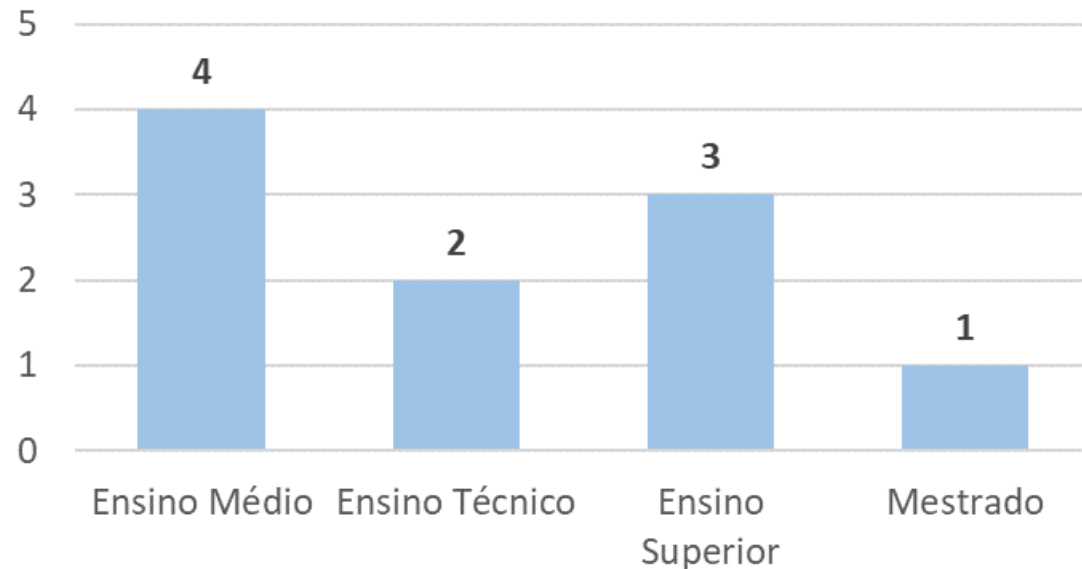
Estatística Descritiva

Visualização de dados



Preditiva.ai

Quantidade de Colaboradores por Escolaridade



Salário Médio por Escolaridade

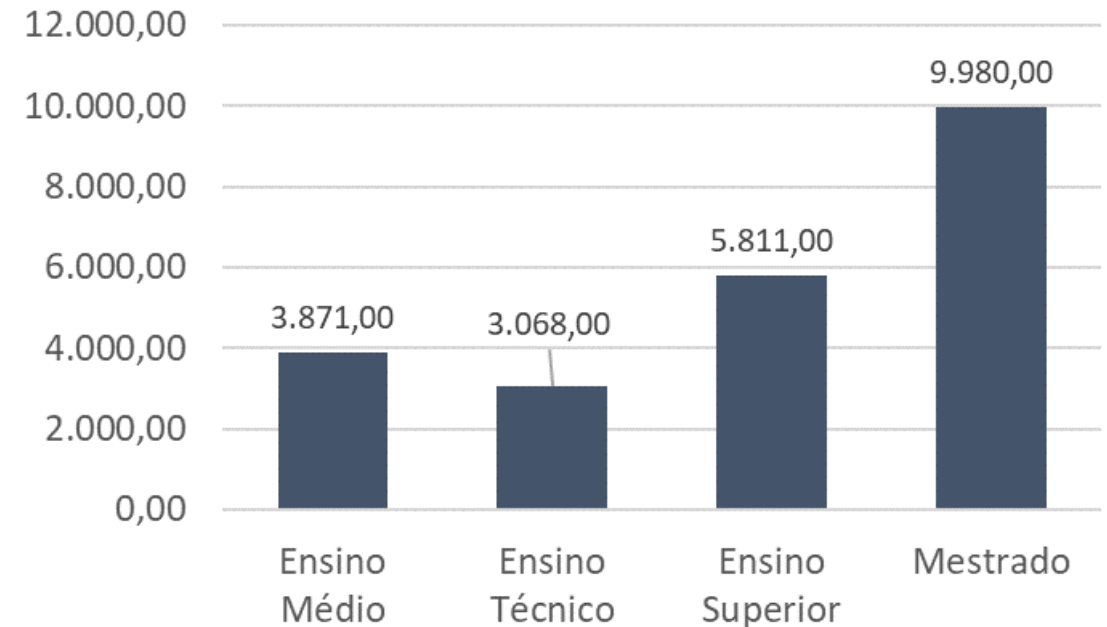


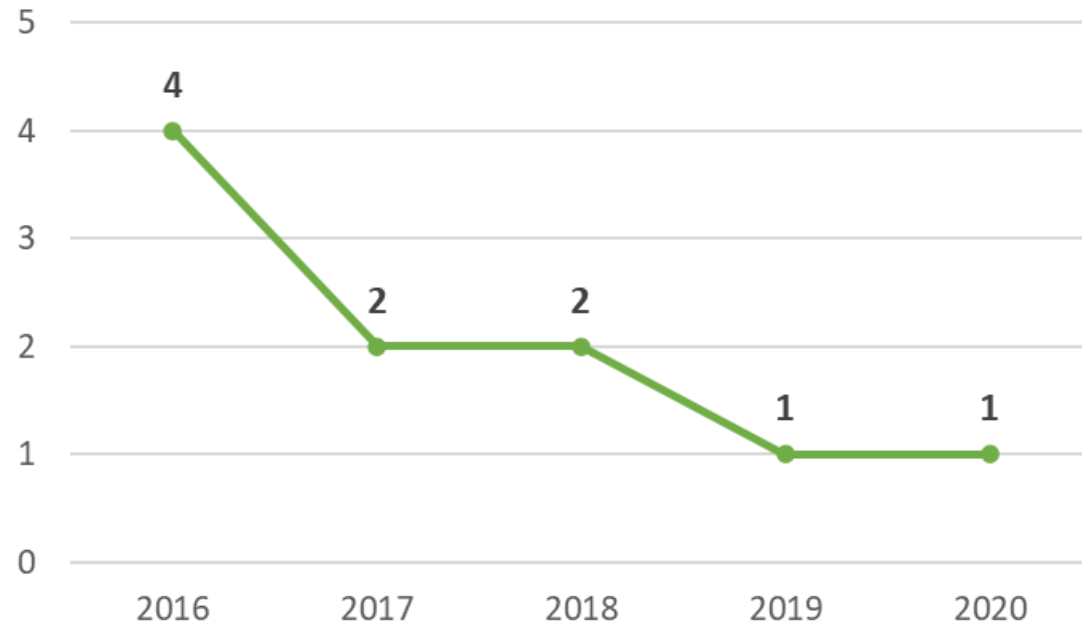
Gráfico de Barras

Uso: Resumo de variáveis quantitativas (contagem, média etc.) por categorias de variáveis qualitativas.

Estatística Descritiva

Visualização de dados

Número de Contratações por Ano



Salário Médio por Escolaridade

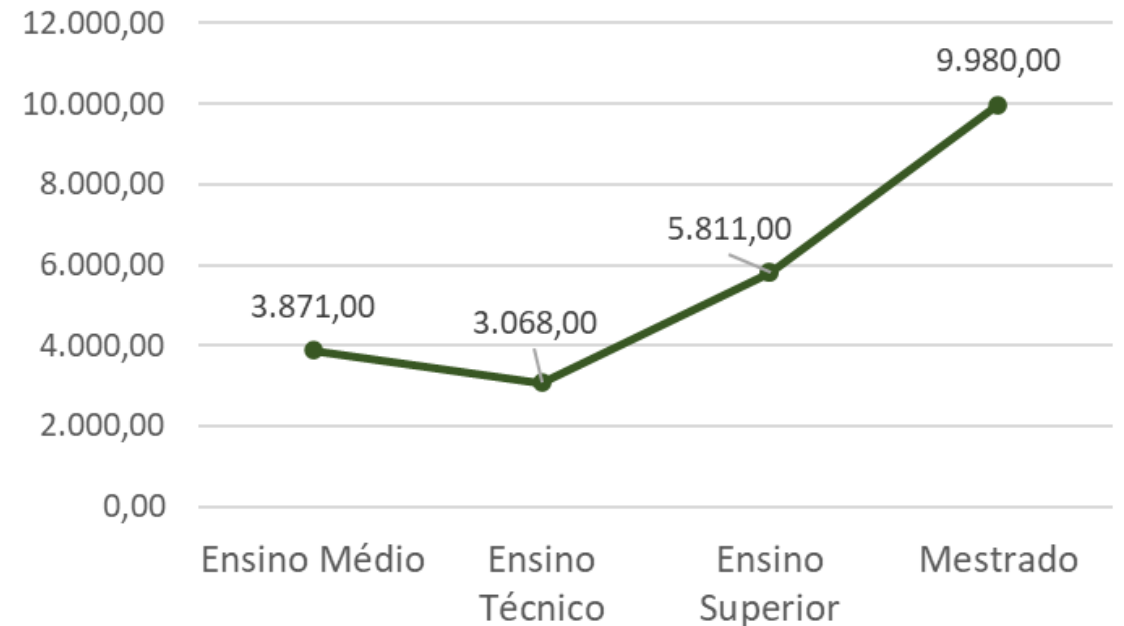


Gráfico de Linhas

Uso: Resumo de **variáveis quantitativas** (contagem, média etc.) por categorias de **variáveis qualitativas ordinais** - indica continuidade.

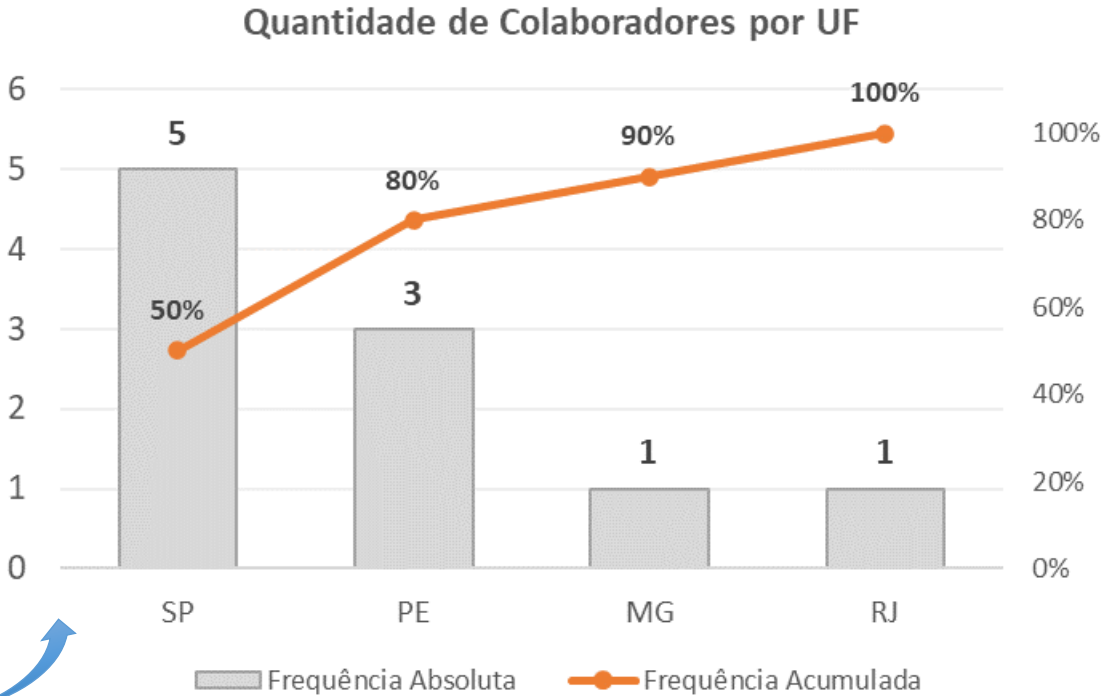
Estatística Descritiva

Visualização de dados



UF	Frequência absoluta	Frequência relativa	Frequência acumulada
SP	5	50%	50%
PE	3	30%	80%
MG	1	10%	90%
RJ	1	10%	100%
Total	10	100%	

Para resumir graficamente uma variável utilizando suas frequências relativas e acumuladas, podemos usar o **Gráfico de Pareto**.

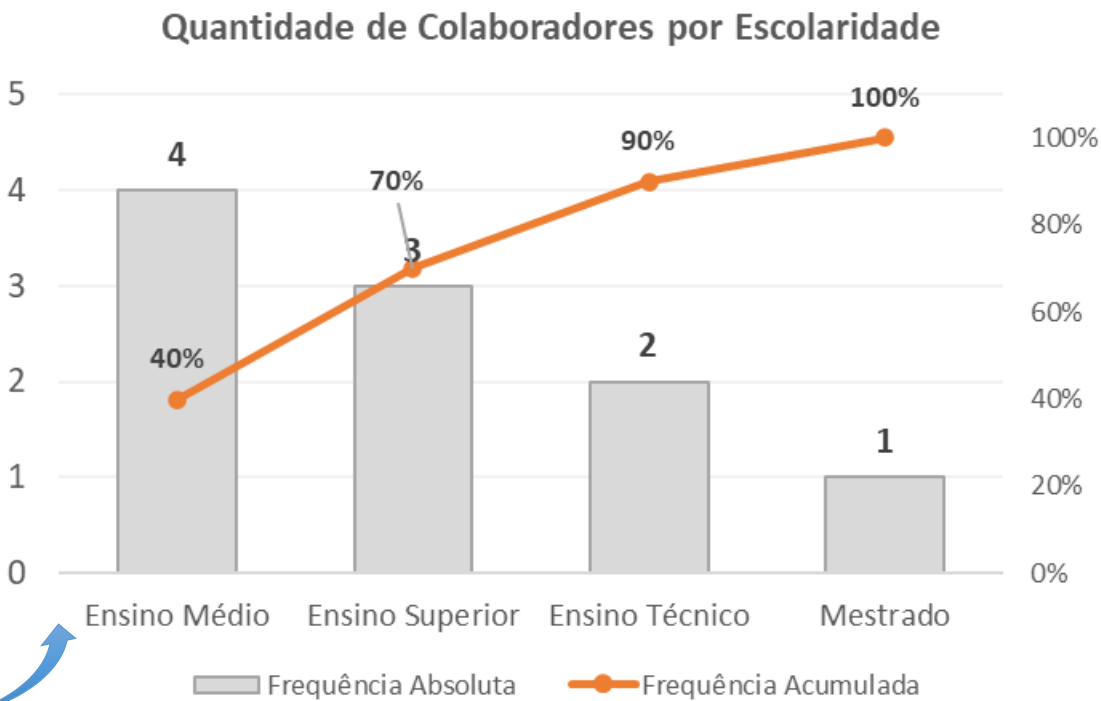


O **Princípio de Pareto***, também conhecido como regra do 80/20, afirma que, para muitos eventos, aproximadamente 80% dos efeitos vêm de 20% das causas.

* Fonte: https://pt.wikipedia.org/wiki/Princ%C3%ADpio_de_Pareto

Escolaridade	Frequência absoluta	Frequência relativa	Frequência acumulada
Ensino Médio	4	40%	40%
Ensino Superior	3	30%	70%
Ensino Técnico	2	20%	90%
Mestrado	1	10%	100%
Total	10	100%	

Para resumir graficamente uma variável utilizando suas frequências relativas e acumuladas, podemos usar o **Gráfico de Pareto**.



O **Princípio de Pareto***, também conhecido como regra do 80/20, afirma que, para muitos eventos, aproximadamente 80% dos efeitos vêm de 20% das causas.

* Fonte: https://pt.wikipedia.org/wiki/Princ%C3%ADpio_de_Pareto

Estatística Descritiva

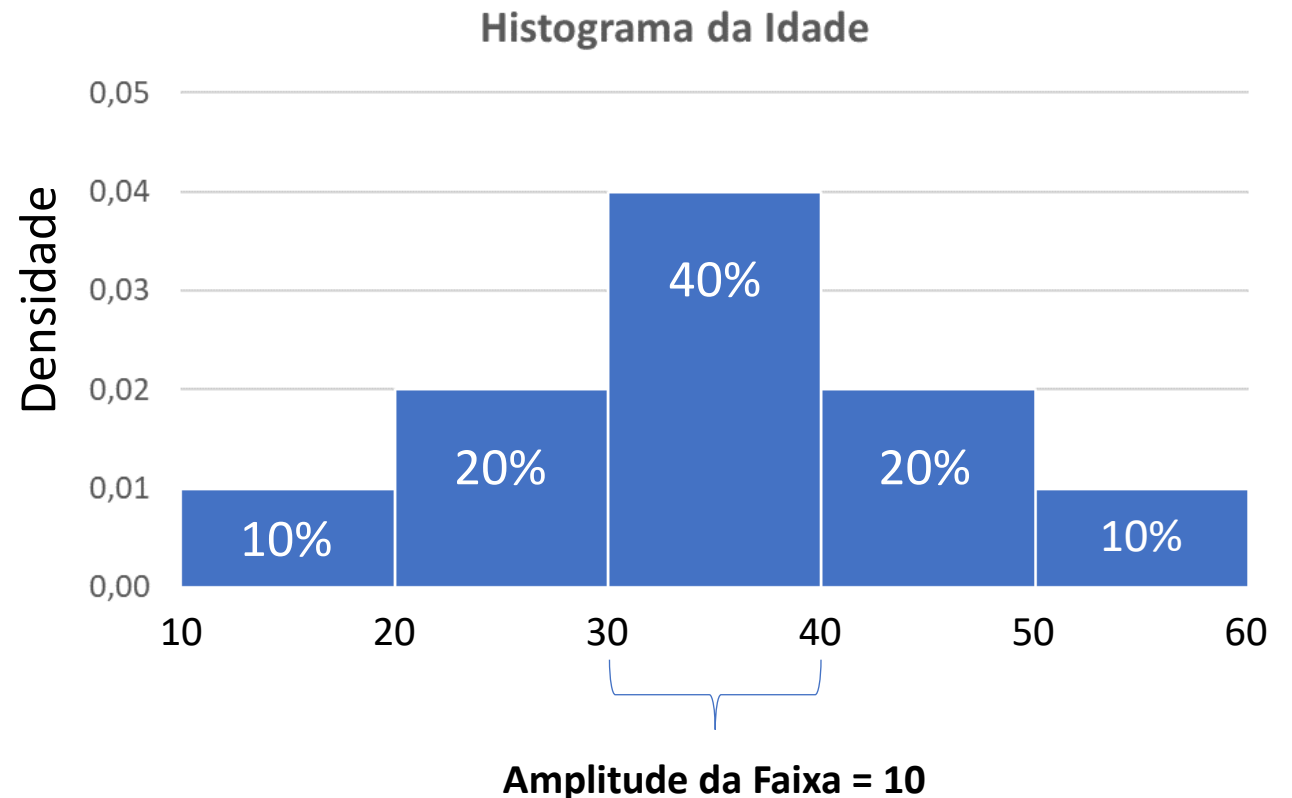
Visualização de dados

Outro gráfico muito importante em Estatística Descritiva é o **Histograma**. Ele é utilizado para resumir graficamente **variáveis quantitativas**. Veja no exemplo:

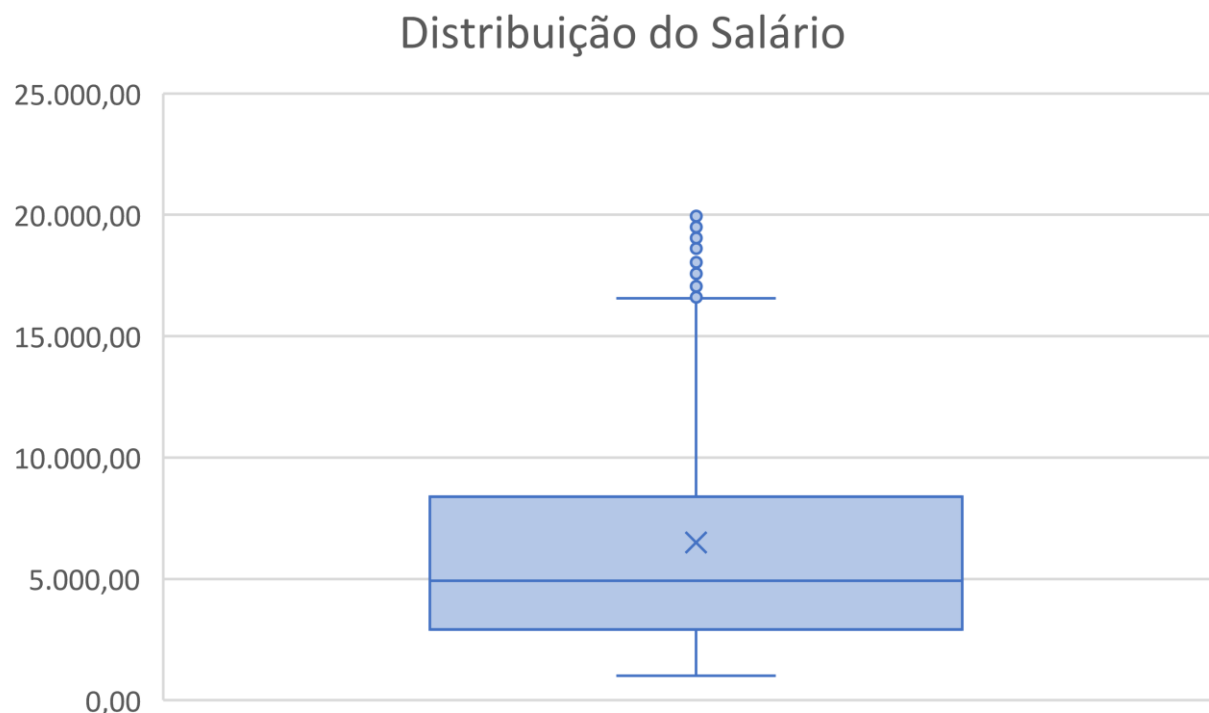
Faixa Etária	Frequência absoluta	Frequência relativa	Densidade
10 - 20	5	10%	0,01
20 - 30	10	20%	0,02
30 - 40	20	40%	0,04
40 - 50	10	20%	0,02
50 - 60	5	10%	0,01
Total	50	100%	

Em um **Histograma**, a **área da barra equivale à Frequência Relativa** da faixa de valor.

No exemplo, 40% das pessoas estão na faixa etária de 30 a 40 anos.

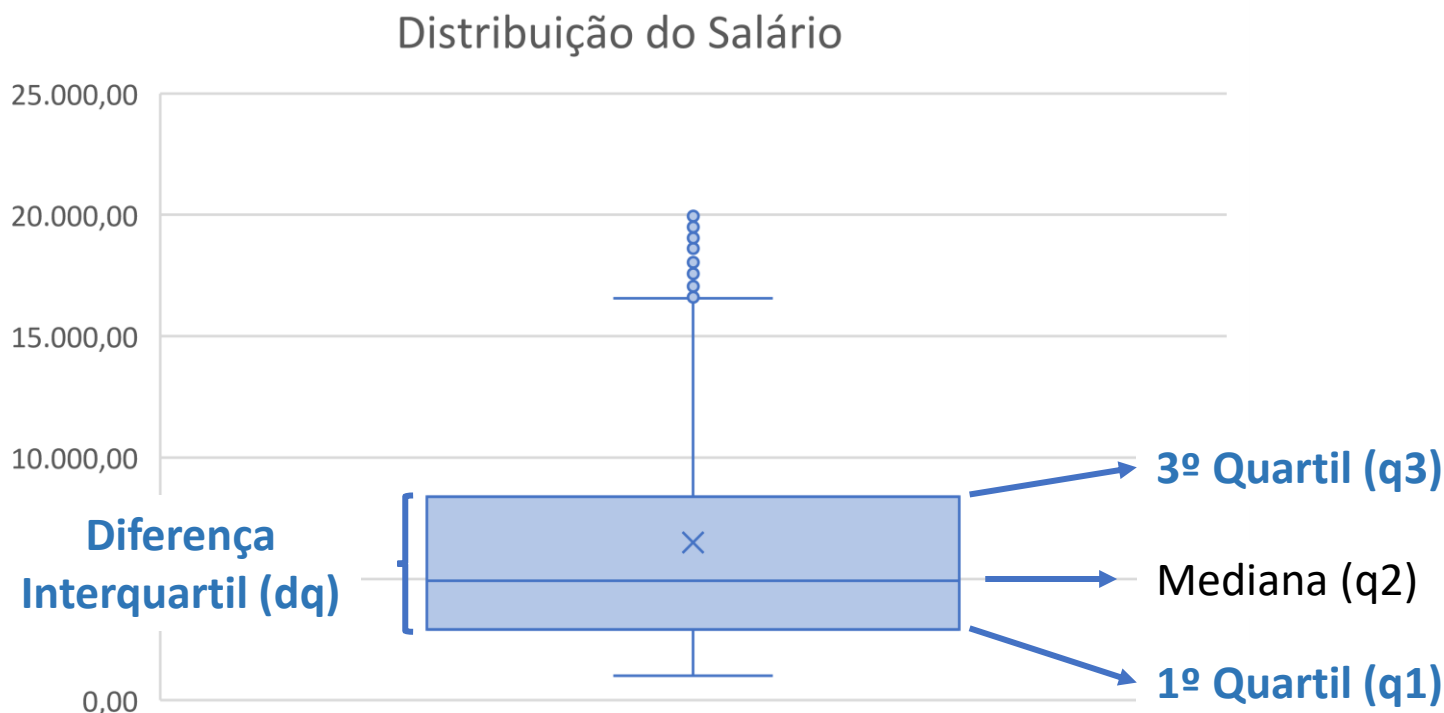


Para visualizar essas medidas e ter uma ideia da posição e dispersão dos dados de uma maneira bastante prática e intuitiva utilizamos o *boxplot*.



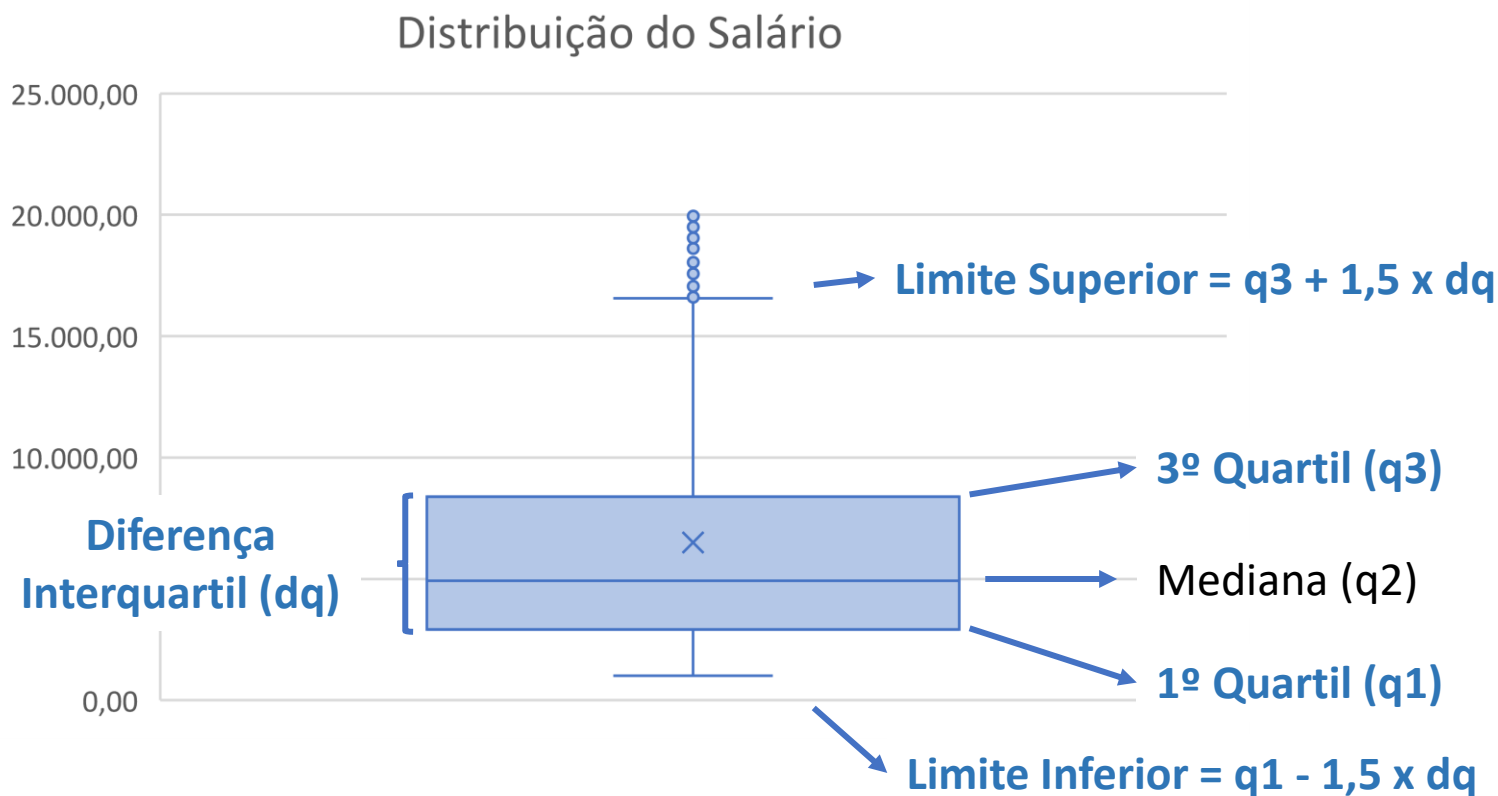


A **Diferença Interquartil** é a diferença entre o **1º Quartil** e o **3º Quartil** e fornece uma medida de variabilidade, pois indica como 50% das observações estão dispersas.

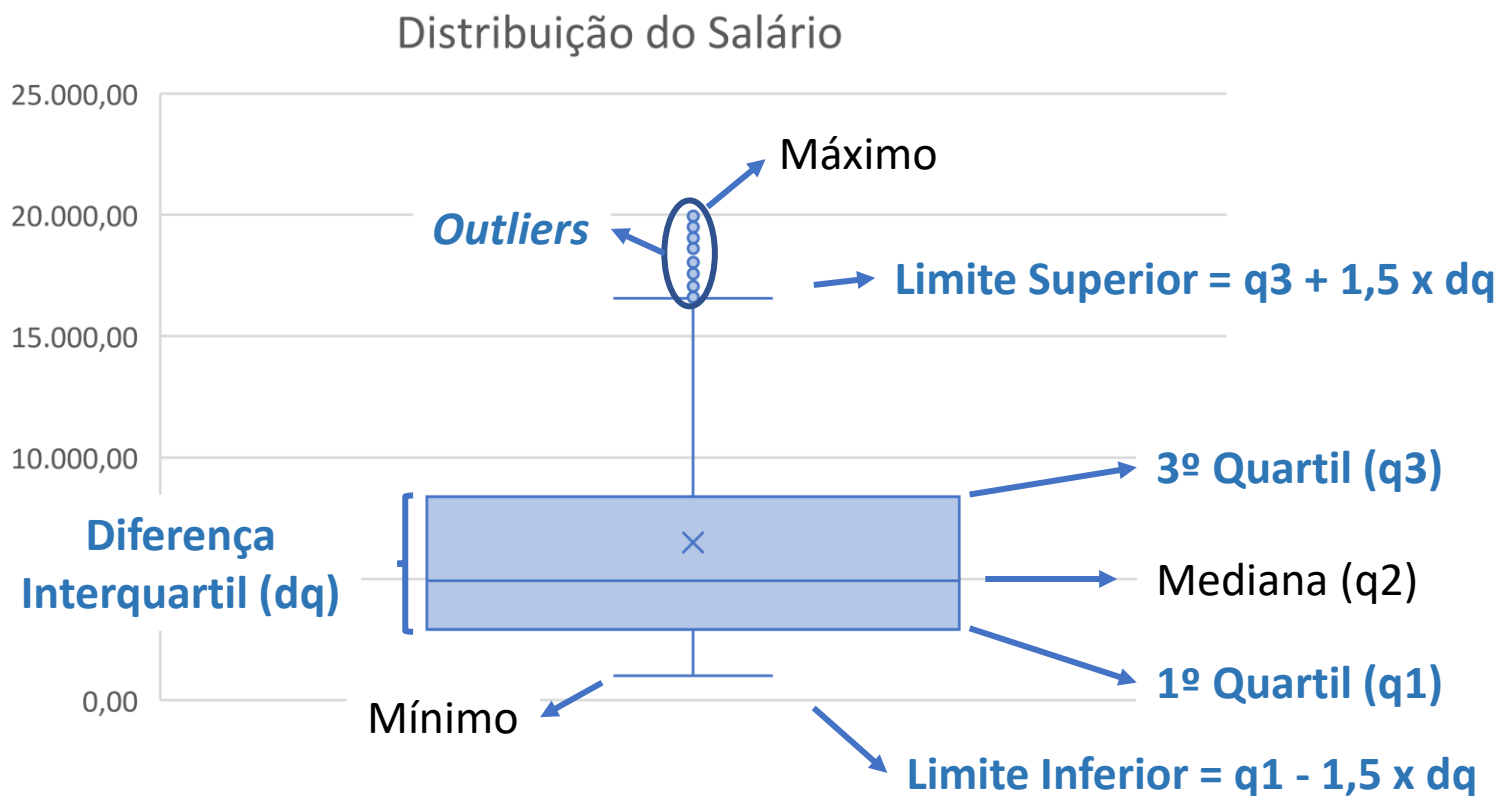


Baseado na **Diferença Interquartil (dq)** são calculados **Limite Inferior** e **Limite Superior**.

Quanto maior o valor de **dq**, maiores serão esses limites.



Os valores além desses limites são chamados de **outliers**, ou **valores atípicos**. Podem ser produto de erro nas medições ou de fato terem **valores bastante diferentes dos demais**.



Revisão

Vimos quais são os principais **gráficos** utilizados nas **análises de dados** e qual tipo de gráfico é mais adequado para cada tipo de variável.

Além do tipo de gráfico, é importante também manter uma **coerência no código de cores** utilizado. Se utilizar mais de 1 gráfico para representar as mesmas variáveis, utilize as mesmas cores para cada categoria.

Outro aspecto importante nos gráficos é o **título**, que **deve expressar claramente seu conteúdo**. Os títulos dos eixos, sempre que necessário, devem ser definidos de forma clara, assim que receber o gráfico não precisará se esforçar para absorver suas informações.

