1 - ESTATÍSTICA DESCRITIVA

ANÁLISE ESTATÍSTICA DE DADOS

- **Definindo Estatística: É a ciência que nos permite aprender a partir dos DADOS.**

Fornece técnicas de análise de dados que auxiliam o processo de tomada de decisão nos problemas onde existe incerteza.

A Estatística permite:

- Coletar dados (técnicas de amostragem)
- Organizar os dados (tabular, calc freq..)
- Apresentar os dados (gráficos)
- Descrever os dados (media, mediana, máx e min, distr normal...)
- Interpretação dos dados (para fazer inferências)

ÁREAS ONDE APLICAMOS ESTATÍSTICA

- Onde há dados!

- DADOS: A matéria prima da quarta revolução industrial!

- Com as técnicas de Análise de Dados, como as fornecidas pela Estatística, podemos obter informação, conhecimento e inteligência a partir dos dados!

ESTATÍSTICA E BIG DATA ANALYTICS

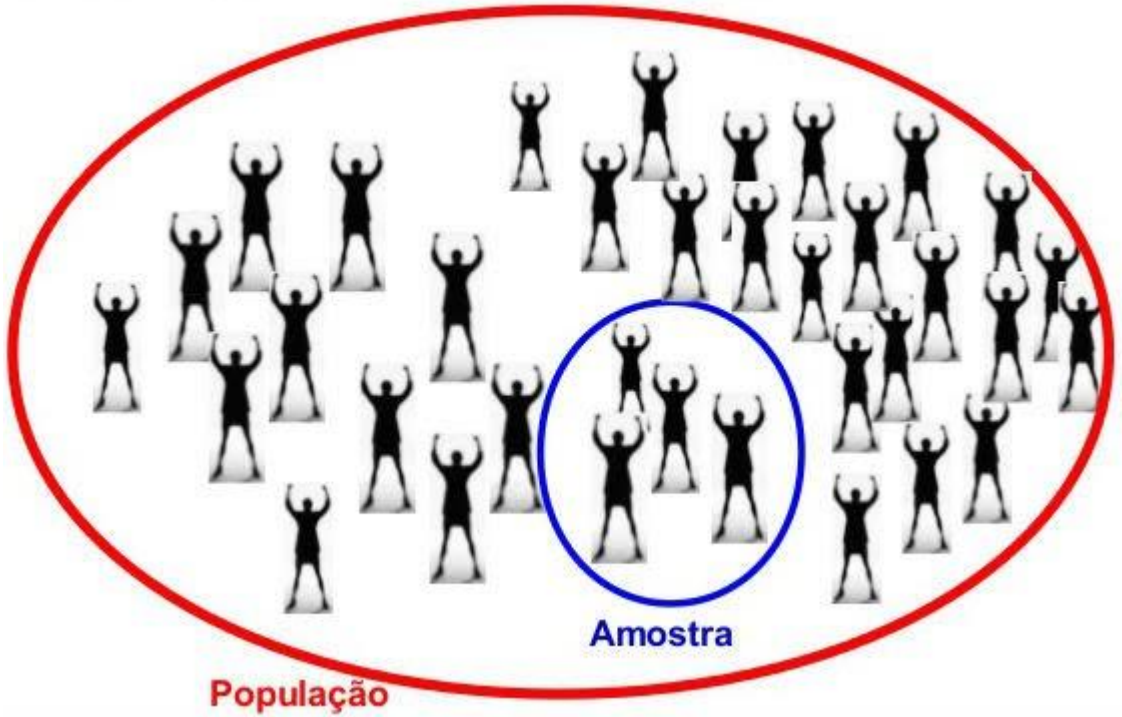
- Estatística: área de conhecimento, parte da Matemática Aplicada, que fornece métodos para coletar, descrever, apresentar e interpreter dados, para utilização dos mesmos na tomada de decisões.

- Big Data Analytics: é o termo que se refere a análise estatística de grandes quantidades de dados, para que se possa extrair informações relevantes para a compreensão da situação atual e a tomada de decisões.

POPULAÇÃO E AMOSTRA

Garantir que a amostra represente fielmente a população!

Para isso, devemos coletar a amostra de forma randomizada, sem escolher exatamente quem fará parte da amostra usando as técnicas estatísticas para o processo de amostragem.



TÉCNICAS DE AMOSTRAGEM

- **Amostragem**: usa a coleta, organização, apresentação e análise dos dados como meio de estudar os parâmentros de uma população.
- **Censo:** é a técnica que seleciona e avalia todos os elementos da população quando se realiza uma pesquisa.
- Técnicas de amostragem:

Probabilística x **Não Probabilística**

- Amostragem a Esmo
- Amostragem Intencional
- Amostragem Por Voluntários

- Amostragem Simples ao Acaso
- Amostragem Sistemática
- Amostragem Por Conglomerado
- Amostragem Estratificada
- Reamostragem (Bootstrap)

PARÂMETRO X ESTATÍSTICA

- Parâmetro – Característica sobre a população.

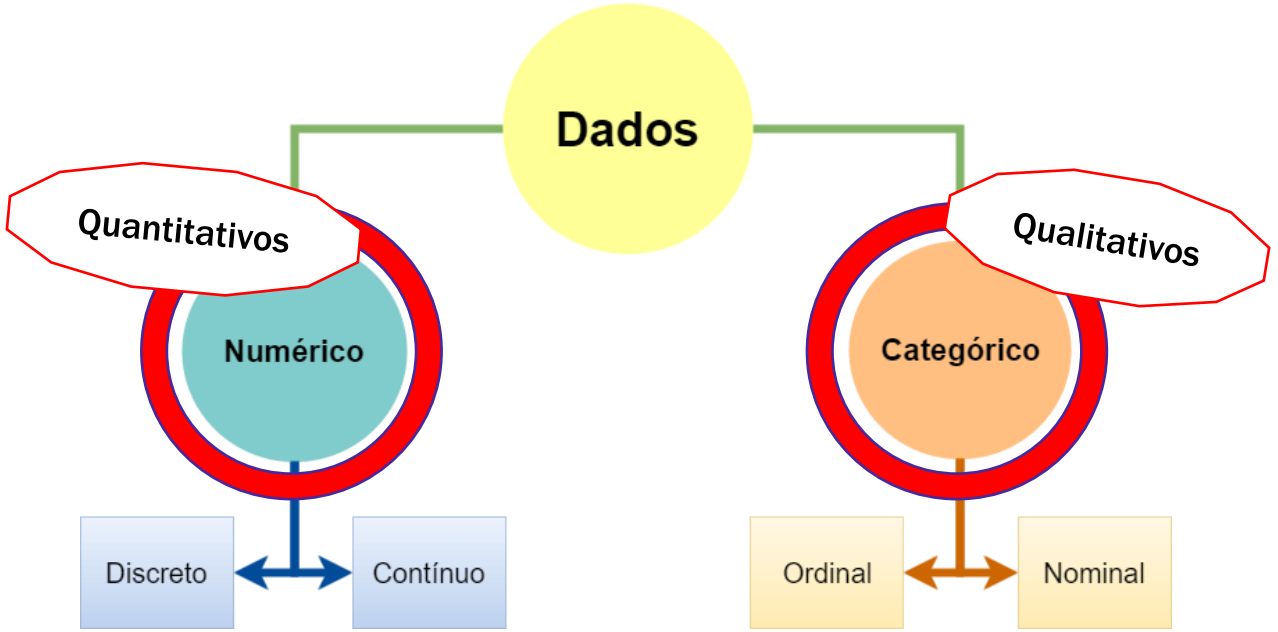
Valores calcaulados usando **dados da população** são chamados de **parâmetros.**

- Estatística – Característica sobre a amostra.

Valores calculados usando **dados da amostra** são chamados de **estatísticas.**

Estatística Inferencial realiza deduções e conclusões sobre **a população**, baseados nos resultados obtidos da análise da **amostra.**

TIPOS DE DADOS



- Número de filhos
- Número de carros
- Número de acessos

(contagem, valores inteiros)

contar

- Altura
- Peso
- Salário

(medido, range)

medir

- Escolaridade
- Classe social
- Fila
- (tem uma hierarquia,
- Exceção range de idade)

- Profissão
- Sexo
- Religião
- (não tem uma hierarquia)

OBSERVAÇÃO X EXPERIMENTAÇÃO

- Há dois tipos de estudos estatísticos:

- Observacional: os dados são recolhidos e observados.
- Experimental: cada indivíduo é aleatoriamente atribuído a um grupo de tratamento, em seguida, os dados específicos são observados e coletados.

A Análise de Dados é o meio através do qual usamos a Estatística para apresentar e demonstrar os resultados dos dados que foram avaliados.

ÁREAS DA ESTATÍSTICA

1

Estatística Descritiva

(usada no processo de análise)

2

Probabilidade

(usada nos algoritmos de machine learning)

3

Estatística Inferencial

(usada para fazer inferencias sobre uma população/amostra)

1 ESTATÍSTICA DESCRITIVA

- Tem por objetivo sumarizar e mostrar os dados, de forma que possa rapidamente obter uma visão geral da informação que está sendo analisada.
- Essa técnica estatística utiliza métodos para coleta, organização, apresentação, análise e síntese de dados obtidos em uma população ou amostra.
- Descreve as principais características dos dados, são 3:

1. Um valor representative do conjunto de dados. Ex: a media.

2. Uma medida de dispersão ou variação. Ex: variância, desvio padrão.

3. A natureza ou forma da distribuição dos dados. Ex: sino, uniforme, assimétrica.

TABELA DE FREQUÊNCIA

Um dos meios mais simples de descrever dados é através de tabelas de frequência, que refletem as observações feitas nos dados.

Classe	Frequência Absoluta	Frequência relativa %	Frequência absoluta acumulada	Frequência relativa acumulada %
0	5	25	5	25
1	8	40	13	65
2	5	25	18	90
3	2	10	20	100
Total	20	100		

DISTRIBUIÇÃO DE FREQUÊNCIA

5 etapas para construir uma distribuição de frequência:

1. Criar o Rol
2. Definir a Amplitude
3. Determinar o Número de Classes
4. Determinar o Tamanho do Intervalo de Classes
5. Construir a Distribuição de Frequência

A Distribuição de Frequência mostra o número de observações de dados que estão em um intervalo específico.

JNB -01 DISTRIBUIÇÃO DE FREQUÊNCIA

- **Solução com Series em Pandas**
- **Solução com Dataframes em Pandas**

Distribuição de Frequência Simples ou Absoluta – fi

São os valores que representam o número de dados de cada classe. A soma das frequências simples é igual ao número total dos dados.

Distribuição de Frequência Relativa Simples – fri (%)

Permite visualizar os valores das razões entre as Frequências Simples e a Frequência Total.

Distribuição de Frequência Acumulada – Fi (.cumsum())

Permite visualizar o total das Frequências de todos os valores inferiores ao limite superior do intervalo de uma dada classe.

Distribuição de Frequência Relativa Acumulada – Fri

Permite visualizar a frequência acumulada da classe, dividida pela frequência total da distribuição.

FERRAMENTAS OFERECIDAS PELA ESTATÍSTICA DESCRITIVA

Análise Univariada
(1 variável)

- Tabela de Frequência
- Gráfico de Barras
- Gráfico de Pareto (cada barra representa uma classe)
  (LINHA, do lado esquerdo a principal causa do problema, do lado direito as causas menos relevantes)
- Gráfico de Pizza
- Gráfico de Linha (evolução de uma variável ao longo do tempo)
- Gráfico Caule e Folha (em forma de tabela)
- Histograma (distribuição dos dados)

Análise Bivariada
(2 variáveis)

- Tabela de Contingência (relação númerica entre duas variáveis)
- Gráfico de Dispersão (relação/correlação entre duas variáveis)

MÉTODOS ESTATÍSTICOS PARA ANÁLISE DE DADOS

Métodos Gráficos ou Tabulares

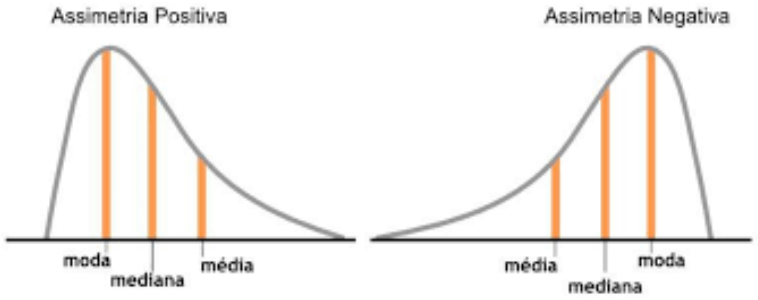
- Tabela de Frequência
- Tabela de Contingência
- Gráfico de Linhas
- Gráfico de Barras
- Gráfico de Pareto
- Histogramas
- Gráficos de Caixa (boxplots)
- Diagramas de dispersão
- Gráfico Temporal
- Ogiva (frequência cumulativa)
- Ramo e Folhas
- Gráficos de Pontos
- Gráfico de Quartis

Métodos Numéricos

- Média
- Mediana
- Moda
- Quartis
- Desvio Padrão
- Variância
- Intervalo Interquartil
- Coeficiente de Variação
- Coeficiente de Assimetria
- Curtose
- Coeficiente de Correlação Linear
- Covariância
- Coeficientes de Associação

MEDIDAS PARA INTERPRETAR OS DADOS

Medidas de Tendência Central



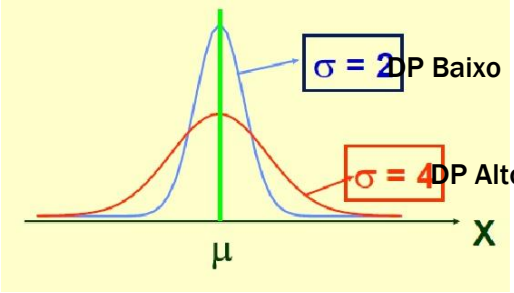
Média (forma mais simples de identificar tendências em um conj de dados)

Mediana (É um valor que divide um conj de dados em duas partes com a mesma quantidade de dados)

Moda (É o valor de maior frequência na amostra)

Medidas de Dispersão
(variabilidade dentro do conjunto de dados)

Desvio Padrão (é a distância média, da média)



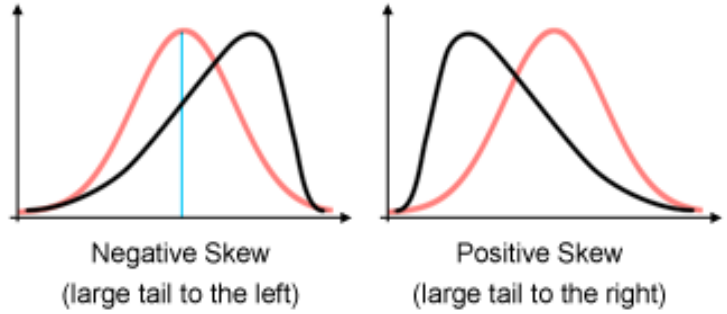
Variância $= (Desvio\ Padrão)^2$
(Mede a amplitude dos dados em relação à média)

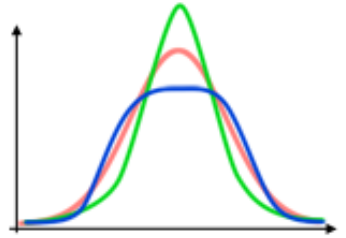
Intervalo (Range) = Max – Min

Percentil (95 th Percentile)

Quartil (4 partes de 25%)

Medidas de Forma – Skewness e kurtosis





COEFICIENTE DE VARIAÇÃO

O coeficiente de variação (CV) mede o desvio em termos de percentual da media.

- Um CV alto indica alta variabilidade dos dados, ou seja, menos consitência dos dados.
- Um CV menor indica mais consistência dentro do conjunto de dados.

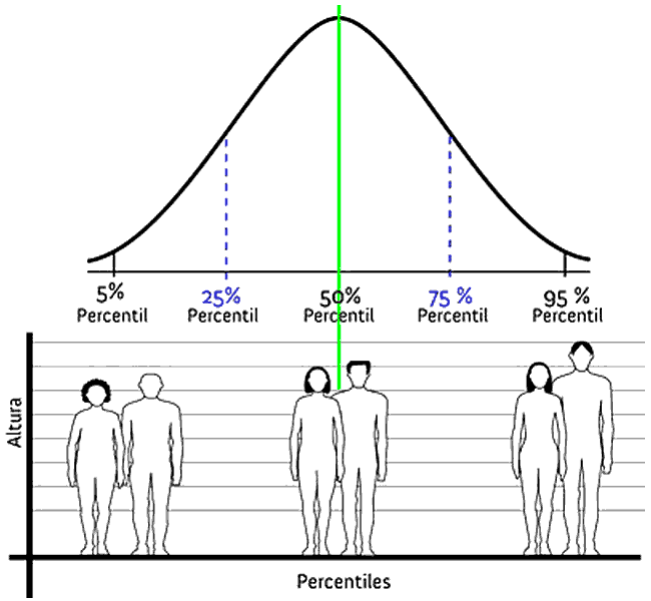
Quando comparamos a consistência entre 2 conjuntos de dados em relação a suas medias, o ideal é utilizer o **coeficiente de variação.**

$$CV = \frac{S}{x} * 100$$

S = Desvio Padrão
x = Média

MEDIDAS DE POSIÇÃO RELATIVA

- **Percentil** e **Quartil** são as medidas mais comuns de posição relativa.



- **Percentil:** é o ponto da distribuição dos resultados ordenados da amostra (por ordem crescente dos dados) em 100 partes de igual amplitude. Por ex, um resultado no percentil 90 significa que 90% dos resultados se situam nesse ponto ou abaixo dele.

Ex: um aluno conseguiu nota 36 em um exame de admissão cujo valor máximo era 45. Sabendo que esse aluno ficou no 97º percentil, isso significa que o aluno foi melhor que 97% dos outros alunos que prestaram o mesmo exame.

MEDIDAS DE POSIÇÃO RELATIVA

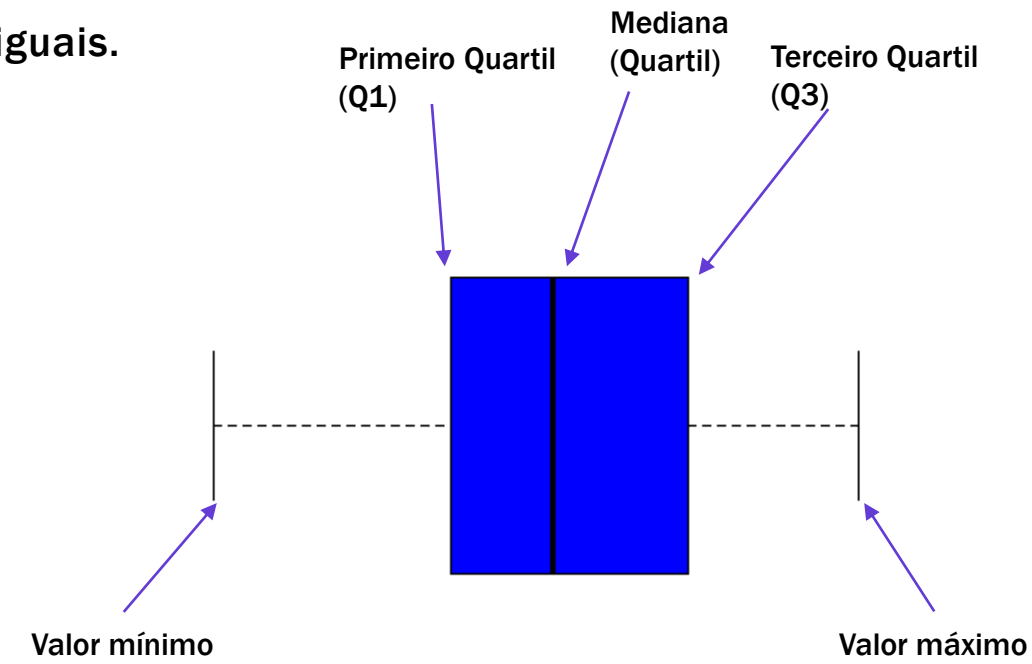
Quartil: é simplismente um percentil específico de interesse.

Quartis são valores que dividem uma tabela de dados em 4 partes iguais.

- O **primeiro quartil** é o valor que constitui **25% percentil.**
- O **segundo quartil** é o valor que constitui **50% percentil**.
- O **terceiro quartil** é o valor que constitui **75% percentil.**
- O **quarto quartil** é o valor que constitui **100% percentil.**

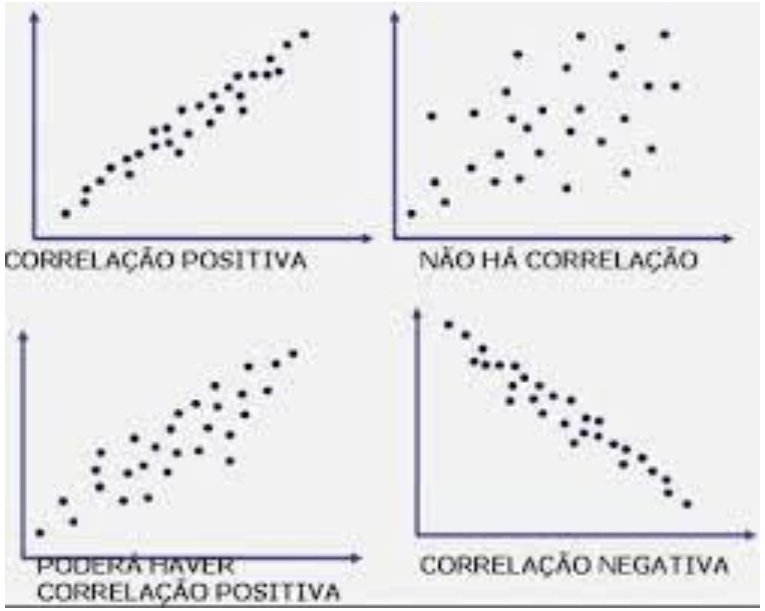
Temos ainda os intervalos interquartis:

- Intervalo Interquartil -> Q3 –Q1
- Intervalo Semi-interquartil -> (Q3-Q1)/2
- Quartil Médio -> (Q3+Q1)/2



COEFICIENTE DE CORRELAÇÃO

- Relação entre duas variáveis.
- A Correlação permite determiner quão fortemente os pares de variáveis estão relacionados.



JNB -02 ESTATÍSTICA DESCRITIVA

Descrição dos dados:

- Média
- Mediana
- Moda
- Contagem
- Valor máx e min
- Variância
- Desvio padrão
- Skewness
- Kurtosis
- Correlação
- Histograma/Dispersão

FIM