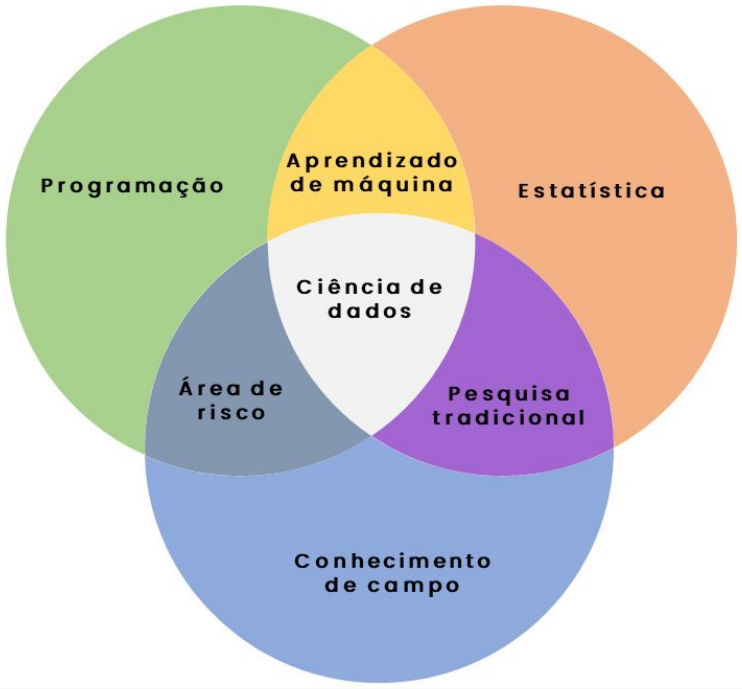
Sistemas Inteligentes 2024/1

Aula 2

@omiguelsma

Introdução - Estatística



Fonte: O que é Ciência de Dados?

Estatística: É a ciência que oferece uma coleção de métodos para produzir e obter dados, organizá-los, resumi-los, analisá-los, interpretá-los e deles extrair conhecimento (Adaptado de Triola (1999)).

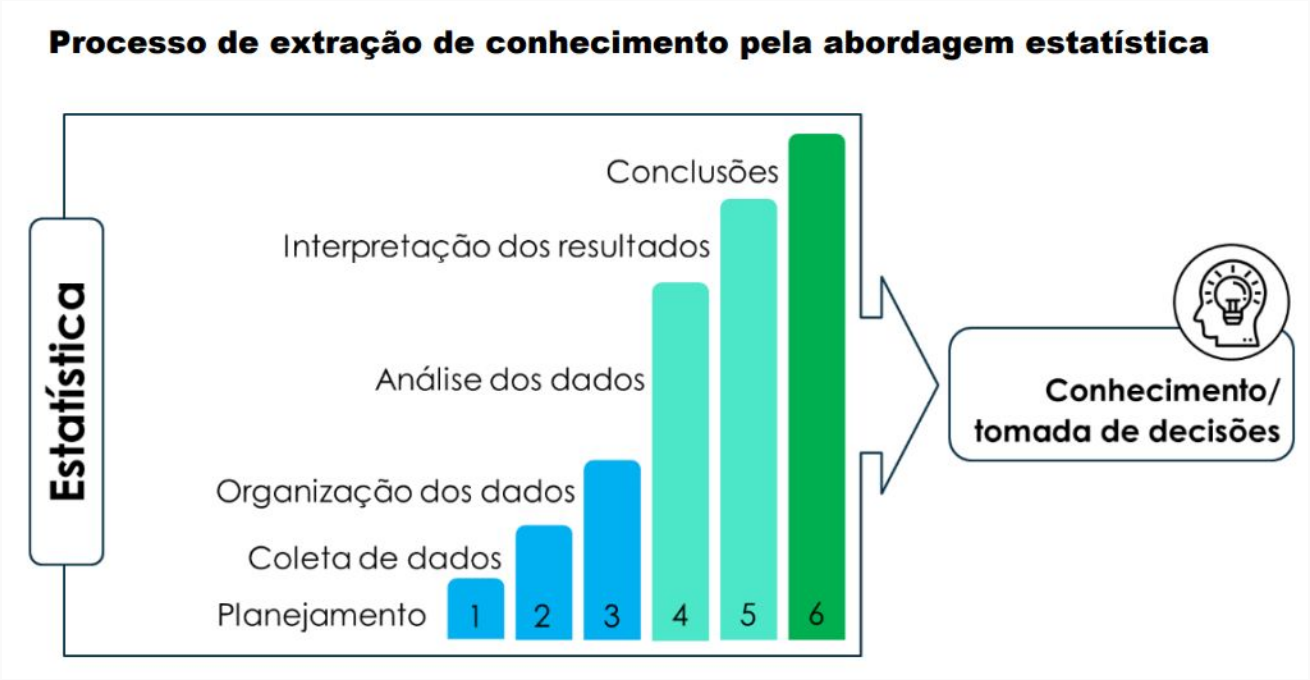
Deste modo, a Estatística contribui para que dados gerem conhecimento e, como tal, deve ter como objetivo não só a produção de dados, como também a interpretação de dados já existentes, utilizando a combinação de gráficos, tabelas e medidas numéricas que permitam interpretar o que esses dados significam.

Pensamento estatístico:
Raciocínio analítico que enfoca em: processo, variação e dados, considerando: a presença inequívoca da variação nos processos de interesse; a necessidade da medição para a quantificação da variação e identificação das fontes de variação; o planejamento da coleta de dados com a variação em mente

Métodos estatísticos:
Os métodos estatísticos são usados para facilitar o entendimento da variabilidade existente nas variáveis investigadas, a obtenção de evidências fundamentadas em dados; e a explicação do fenômeno.

Fonte

@omiguelsma


Processo de extração de conhecimento pela abordagem estatística
Estatística
Planejamento
Coleta de dados
Organização dos dados
Análise dos dados
Interpretação dos resultados
Conclusões
1
2
3
4
5
6
Conhecimento/
tomada de decisões

Fonte

Conceitos básicos:

População: é uma coleção completa de todos os elementos (valores, pessoas, medidas, etc) a serem estudados.

Censo: é uma coleção de dados relativos a todos os elementos de uma população. Amostra: é uma sub-coleção de elementos extraídos de uma população.

Parâmetro: é uma medida numérica que descreve uma característica de uma população.

Estatística: é uma medida numérica que descreve uma característica de uma amostra.

Variável: Qualquer conjunto de dados contém informações sobre algum grupo de indivíduos. As informações são organizadas em variáveis. Uma variável é uma característica, propriedade ou atributo de uma unidade da população, cujo valor pode variar entre as unidades da população.

Variação: O padrão de variação de uma variável constitui a sua distribuição. A distribuição de uma variável quantitativa registra seus valores numéricos e a frequência de ocorrência de cada valor.

Fonte

O que são dados?

Os dados são a matéria prima da Estatística.

Definido o assunto de interesse, os dados são obtidos da medição de determinada característica ou propriedade de um objeto, pessoa ou coisa.

Os dados em geral são números, mas não são ”apenas números”. Os dados são números com um contexto.

➢ Procure entender o que os dados dizem em cada contexto específico.
Todos os métodos estatísticos nada mais são do que instrumentos que nos ajudam a entender os dados.

➢ Deixe para uma calculadora ou um computador o máximo possível dos cálculos e gráficos e procurar concentrar-se no que fazer? E por que fazer?

➢ Enfoque as grandes ideias da estatística, e não apenas regras e receitas.

Fonte

O que são dados?

Os **dados discretos** resultam de um conjunto finito de valores possíveis, ou de um conjunto enumerável desses valores.

Os **dados contínuos** resultam de um número infinito de valores possíveis que podem ser associados a pontos em uma escala contínua de tal maneira que não haja lacunas ou interrupções.

Os **dados quantitativos** consistem em números que representam contagens ou medidas.

Os **dados qualitativos** podem ser separados em diferentes categorias que se distinguem por alguma característica não-numérica.

Fonte

Como devemos coletar dados?

Os dados são parte crucial no estudo da variabilidade, pois, assim como são geradores de resultados podem também ser geradores de incerteza, estimulando e motivando mais aprofundamento e estudo sobre seu comportamento e sobre sua distribuição. Aos dados também estão associadas às fontes de erros que interferem diretamente na aplicação dos métodos estatísticos.

Fonte

Produção de dados: parte essencial para embasar a inferência estatística, para responder questões específicas formuladas antes dos dados serem produzidos.

Como podemos produzir dados?

➢ **Estudos observacionais ou levantamentos:**
Visam retratar a população o menos distorcida possível pelo ato da coleta de informações. Não tentamos manipular, influenciar ou modificar as respostas dos elementos a serem estudados.

➢ **Experimentos: decorre da aplicação de determinado tratamento para posterior observação de seus efeitos.**
Preocupação com a relação de causa e efeito sobre os elementos pesquisados. Impõe deliberadamente algum tratamento aos indivíduos, a fim de observar sua reação e medir respostas. Requer planejamento apropriado para serem válidos cientificamente.

➢ **Simulação: Uso de modelo matemático ou físico para reproduzir as condições de uma situação ou processo, usando métodos computacionais.**
A simulação é uma alternativa quando é impraticável ou mesmo perigoso estudar os fenômenos de interesse em condições reais.

Fonte

@omiguelsma

Planejamento amostral: parte crucial numa coleta de dados que vise um estudo estatístico.

Os processos ou padrões definidos para coletar dados são chamados de **planejamentos**.

Os planejamentos devem abordar principalmente:

- ➢ **Como vamos selecionar os indivíduos a serem estudados** (tipo de amostragem ou de delineamento experimental);
- ➢ **Quantos indivíduos devemos estudar** (tamanho da amostra);
- ➢ **Se há necessidade de composição de grupos e como eles devem ser formados para que possam ser comparados** (alinhamento da amostra com os objetivos e restrições);
- ➢ **Como serão feitas as medições** (procedimentos e instrumentos de medição); etc.

O planejamento sistemático para gerar dados é um dos primeiros passos para a realização de um estudo com base científica. A falta de planejamento pode levar a tendenciosidades, à falta de dados ou a resultados confusos e imprecisos.

Fonte

Medidas de posição: representantes do conjunto de dados

• **Média** ($\bar{x}$): é a média aritmética, ou seja, é a soma das observações dividida pelo número total das mesmas.
• **Mediana** (Md): é o ponto do meio de uma distribuição, considerando a série de observações ordenada de forma crescente. O número tal que metade das observações são menor do que ele e metade maior. Se o número de observações n for ímpar, a mediana Md será a observação central na lista ordenada. Se o número n de observações for par, a mediana Md será a média das duas observações centrais na lista ordenada. • Você sempre pode localizar a mediana na lista ordenada das observações contando até (n + 1)/2 observações a partir do menor valor da lista.
• **Moda** (Mo): é o dado mais frequente observado em um conjunto de dados. Em uma distribuição, pode haver mais de uma moda.

Fonte

Medidas de dispersão: sumarizam a variabilidade do conjunto de dados

• **Variância (s^2):** é uma média dos quadrados dos desvios das observações a partir de sua média.

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

• **Desvio médio (dm):** é uma média dos desvios das observações em relação à média em valor absoluto.

$$dm = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$$

• **Desvio padrão (s):** é a raiz quadrada da variância • Amplitude ou intervalo: diferença entre os valores máximo ($x(n)$) e mínimo ($x(1)$) do conjunto de dados.

$$a = x_{(n)} - x_{(1)}$$

• **Coeficiente de variação (cv):** é a razão entre o desvio padrão (s) e a média $\bar{x}$. É uma medida de dispersão relativa.

$$cv = {}^{s}/_{\bar{x}}$$

Fonte

@omiguelsma

Quartis: delimitam a metade central do conjunto de dados.

O primeiro quartil **($Q1$)** é o ponto central entre o mínimo e a mediana.
O terceiro quartil **($Q3$)** é o ponto central entre a mediana e o máximo.

Intervalo interquartílico: é a diferença entre o terceiro e o primeiro quartis, ou seja, $IQ = Q3 - Q1$.

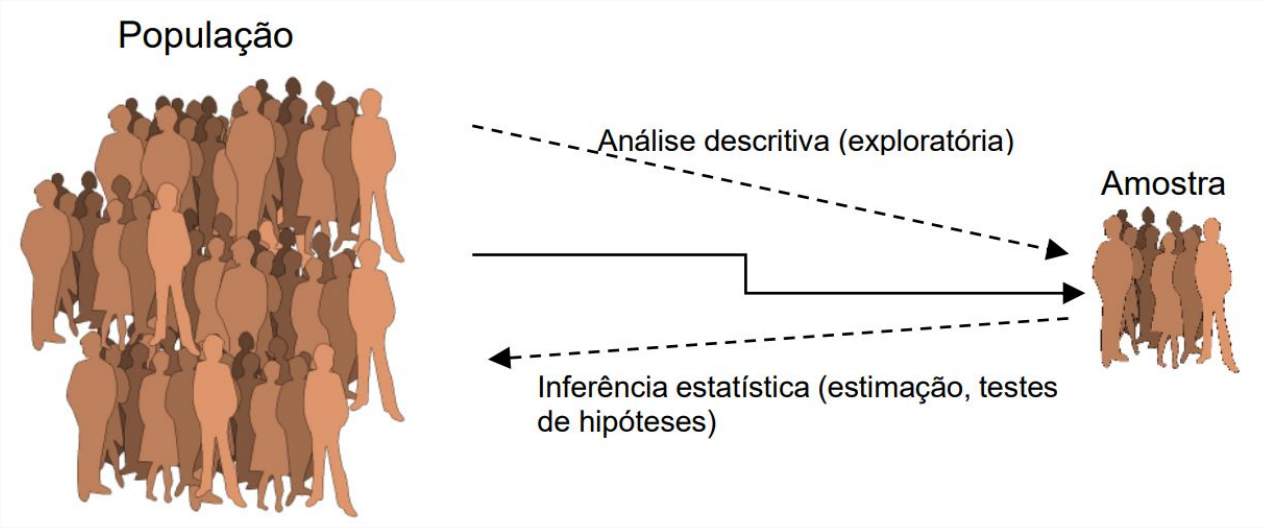
Resume a distribuição focando na metade central dos dados.

Resumo de cinco números:

O resumo de cinco números oferece uma descrição razoavelmente completa de centro e dispersão e serve para construção de um gráfico, **o boxplot.**

Os cinco números são: Mínimo; $Q1$, Md, $Q3$, Máximo.

Fonte


População
Análise descritiva (exploratória)
Amostra
Inferência estatística (estimação, testes de hipóteses)

Fonte

Distribuição amostral:
É a distribuição de probabilidades associada à estatística, considerando todas as amostras possíveis de mesmo tamanho tomadas da mesma população.

Como selecionar amostras?

1. **Por levantamentos amostrais:** a amostra é obtida de uma população bem definida, por meio de processos bem protocolados e controlados pelo pesquisador.
2. **Planejando experimentos:** o objetivo é analisar o efeito de uma variável sobre outra. Requer interferência do pesquisador sobre as condições experimentais, bem como o controle de fatores externos, com o intuito de medir o efeito desejado.
3. **Por estudos observacionais:** os dados são coletados sem que o pesquisador tenha controle sobre as informações obtidas, exceto sobre possíveis erros grosseiros.

Problemas:

- Falta de informação sobre os parâmetros;
- Falta de informação sobre a distribuição, ou seja, sobre como os dados se comportam em termos da forma como se distribuem;
- Faltam tanto os parâmetros quanto a curva de distribuição.

Fonte

@omiguelsma

Conceitos de probabilidade importantes para a inferência estatística:

Lei dos grandes números
Selecione observações ao acaso de qualquer população com média μ. À medida que o número de observações selecionadas aumenta, a média da amostra se aproxima cada vez mais da média populacional μ.

Distribuição amostral da média
Suponha que $\bar{X}$ é a média de n observações independentes da variável X, ou seja, uma amostra aleatória X1, ..., Xn . Se X tem média μ e desvio padrão , então, a média da distribuição amostral de $\bar{X}$é μ e o desvio padrão $\sigma/\sqrt{n}$.

$$E(\bar{X}) = \mu \text{ e } V(\bar{X}) = \sigma^2/n$$

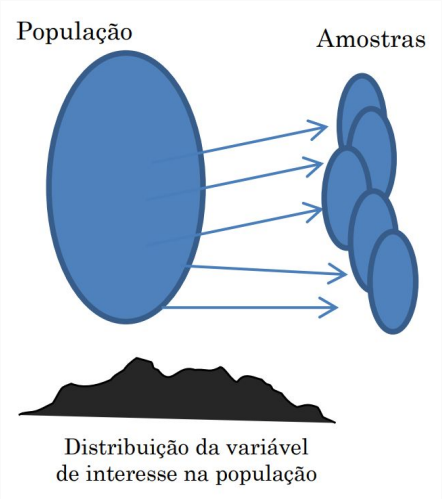
Distribuição amostral da proporção
Suponha que $\hat{p}$ é a proporção amostral de elementos com certa característica e que n observações independentes da variável Xi, que assume valores 0 ou 1, são selecionadas de modo que $\hat{p} = \sum Xi/n$. Então, $\hat{p}$ será uma variável aleatória cuja média é p e o desvio padrão $\sqrt{pq/n}$.

$$E(\hat{p}) = p \text{ e } V(\hat{p}) = pq/n$$

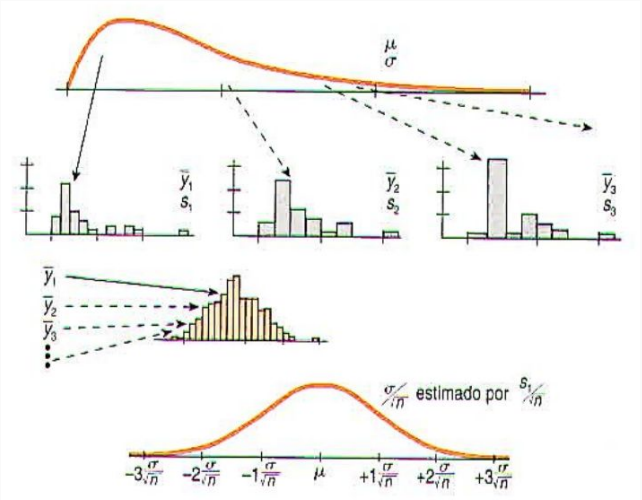
Fonte

Teorema do Limite Central

Considere uma amostra aleatória simples de tamanho n extraída de uma população qualquer com média μ e desvio padrão. Quando n é grande, a distribuição amostral da média $\boldsymbol{X}$ se aproxima da distribuição normal com média μ e desvio padrão $\boldsymbol{\sigma/\sqrt{n}}$.



Fonte



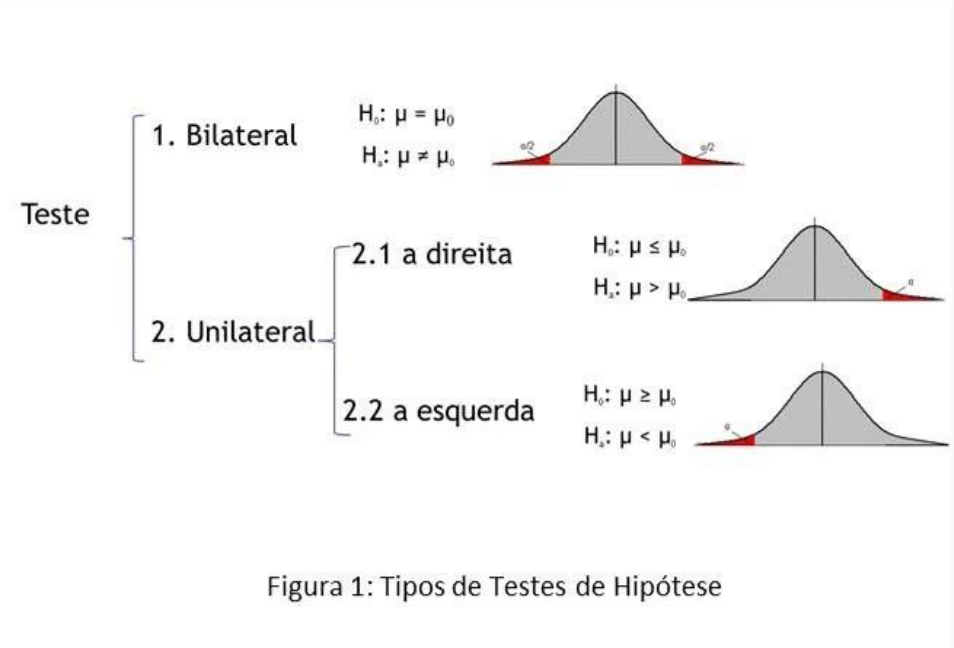
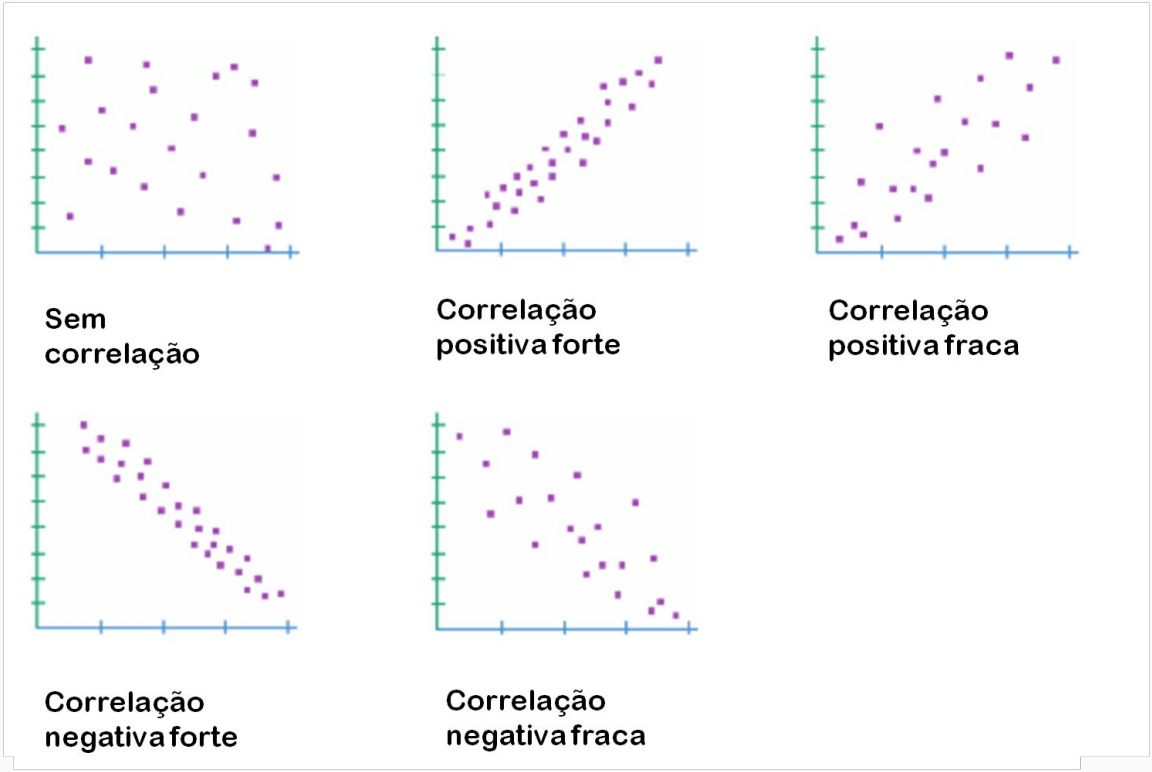


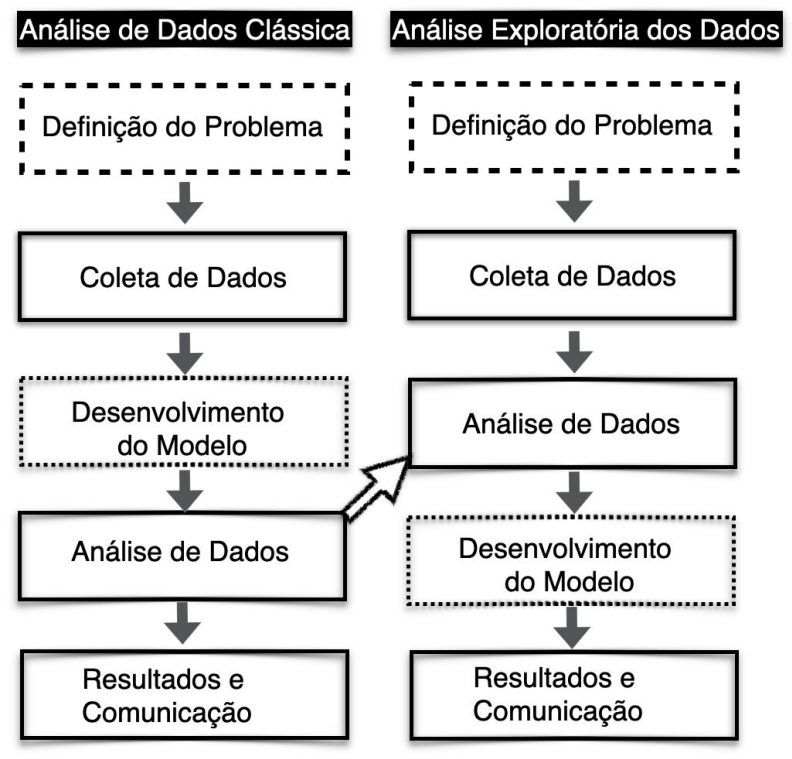
Figura 1: Tipos de Testes de Hipótese

Fonte


Sem correlação
Correlação positiva forte
Correlação positiva fraca
Correlação negativa forte
Correlação negativa fraca

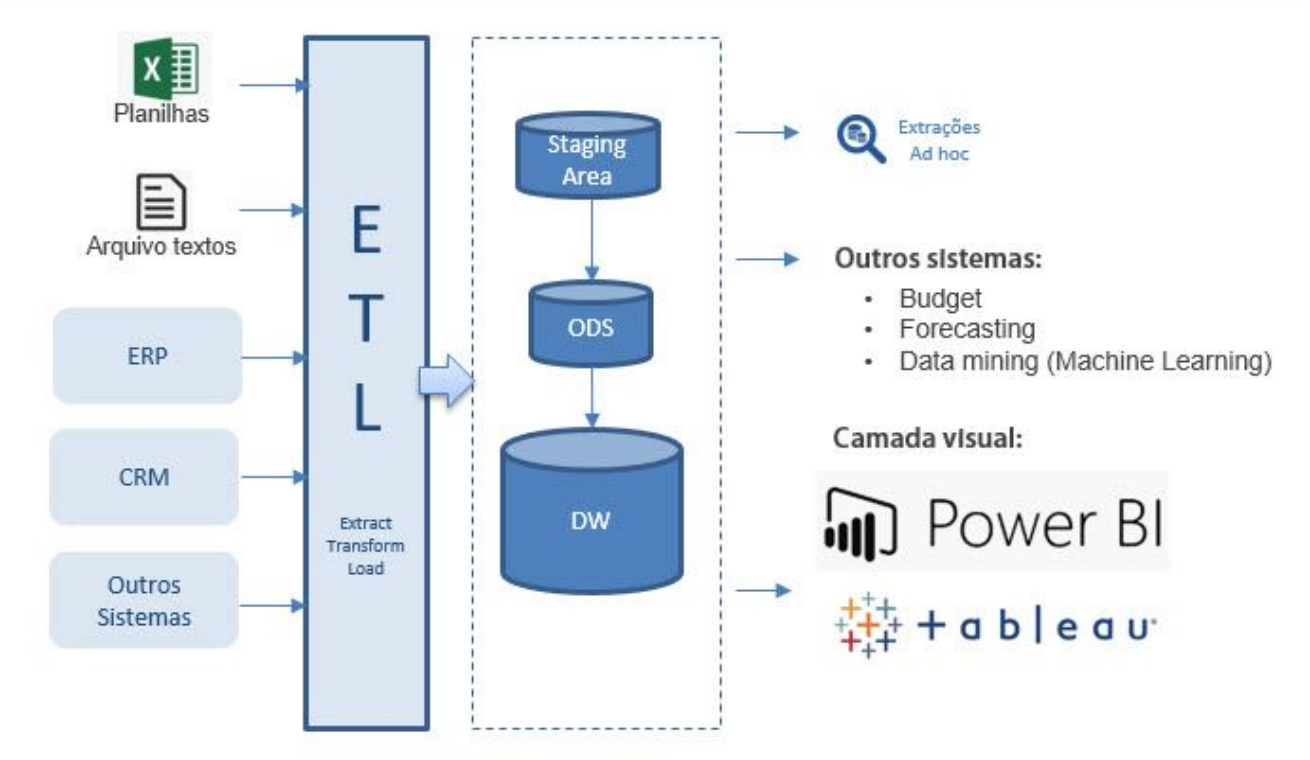
Fonte

Introdução - Análise Exploratória de Dados (AED).



Fonte

Introdução - Extração, Transformação e Carregamento (ETL)



Fonte

Saiba mais:

História da Estatística

Breve História da Estatística

Quem foi Gertrude Mary Cox, mais conhecida como a "The First Lady of Statistics"?