



**Grupo
Educacional**

Professor: Dr. Alessandro Ferreira Alves

Disciplina: Cálculos e Estatística Básica

**Material de Apoio: Aspectos Introdutórios da Estatística Aplicada à Área
Computacional e TI, Conceitos Fundamentais e Aplicações Diversas.**

Semestre/Año: Segundo Semestre de 2024.

**Instituição credenciada pelo MEC
Centro Universitário do Sul de Minas**



Estatística ...



O que é **Estatística?**

Estatística e Engenharia



Qual a razão de estudarmos a
Estatística no contexto da **Área de**
TI e na **Área Computacional**?

Motivação

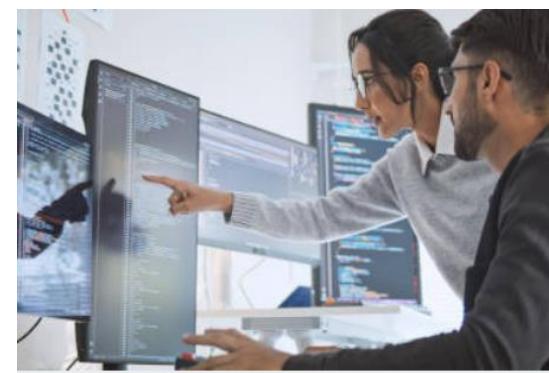
“Dados são como filhos: você não sabe no que eles vão se tornar”.

Aldous Huxley



Contextualizando ...

- A **estatística** aplicada à área **computacional** e à **Tecnologia da Informação (TI)** desempenha um papel fundamental na análise de dados, modelagem preditiva, otimização de processos, e tomada de decisões.
- Com o **aumento do volume de dados (*Big Data*)** e o crescimento de tecnologias como **Inteligência Artificial (IA)** e aprendizado de máquina (***Machine Learning***), o uso de estatísticas tornou-se crucial para entender e melhorar o desempenho de sistemas e serviços.



Algumas Aplicações

1. Análise de Desempenho de Sistemas:

- . **Monitoramento de redes e sistemas:** A estatística é usada para analisar o desempenho de redes, servidores, bancos de dados e sistemas operacionais, identificando padrões de uso, gargalos de desempenho e anomalias.
- . **Medição de tempo de resposta:** Técnicas como análise de variância (ANOVA) ou distribuições probabilísticas são aplicadas para avaliar o tempo de resposta e a carga de trabalho de sistemas e servidores, otimizando a alocação de recursos.



Algumas Aplicações

2. Análise de Big Data:

- Em TI, grandes volumes de dados são coletados em tempo real, especialmente em áreas como logs de sistema, dados de sensores e redes sociais. Técnicas de estatística descritiva e inferencial são essenciais para organizar, resumir e interpretar esses dados.
- Inferência estatística é usada para fazer previsões ou generalizações sobre um conjunto de dados, baseando-se em amostras extraídas de grandes volumes de dados.



Algumas Aplicações

3. Machine Learning e Inteligência Artificial:

- A **Estatística** está no núcleo de algoritmos **de aprendizado de máquina**. Técnicas estatísticas como **regressão linear**, **regressão logística**, **análise discriminante**, **árvores de decisão** e **redes neurais** são amplamente utilizadas para criar modelos preditivos e classificatórios.
- **Análise Bayesiana** e **modelos probabilísticos** são frequentemente usados para criar sistemas de recomendação, reconhecimento de padrões e algoritmos de aprendizado supervisionado e não supervisionado.



Algumas Aplicações

4. Mineração de Dados (*Data Mining*):

- A estatística fornece as bases para técnicas de mineração de dados, usadas para identificar padrões ocultos em grandes volumes de dados. Técnicas como *clusters*, **classificação**, **análise de associação** e **análise de componentes principais (PCA)** são usadas para extrair insights valiosos de conjuntos de dados complexos.
- **Exemplo:** Identificar padrões de comportamento de clientes em transações on-line para melhorar algoritmos de recomendação ou detectar fraudes.



Algumas Aplicações

5. Segurança Cibernética:

- A estatística é utilizada para detectar **anomalias** em padrões de tráfego de rede, possíveis sinais de ataques cibernéticos ou comportamentos suspeitos. Técnicas como **análise de séries temporais**, **análise de regressão** e **modelos probabilísticos** podem prever comportamentos de ataque e proteger sistemas críticos.
- **Análise estatística de eventos discretos** pode ser aplicada na detecção de intrusões e na análise de padrões de malware.



Algumas Aplicações

6. Simulação e Modelagem:

- Técnicas de **simulação estatística** são usadas para testar o desempenho de sistemas complexos, como redes de computadores, infraestrutura de TI, ou arquiteturas de software, antes de serem implementados em um ambiente real.
- Ferramentas de **Monte Carlo** e **modelagem probabilística** permitem prever como diferentes variáveis afetarão o desempenho dos sistemas sob várias condições.



Algumas Aplicações

7. Otimização de Algoritmos:

- A estatística é usada para avaliar o desempenho de diferentes algoritmos e otimizar sua eficiência. Métodos como **análise de regressão** podem ser usados para identificar quais variáveis afetam mais o tempo de execução de um algoritmo.
- **Algoritmos evolutivos** e **otimização estocástica** aplicam conceitos estatísticos para encontrar soluções aproximadas para problemas de otimização.



Algumas Aplicações

8. Previsão de Demanda e Gestão de Capacidade:

- A estatística é aplicada para prever o uso de recursos de TI, como processamento, armazenamento ou largura de banda, a fim de otimizar a alocação de recursos em infraestruturas de computação em nuvem.
- **Séries temporais** são usadas para modelar e prever a demanda futura com base em padrões passados, permitindo ajustes proativos na capacidade da infraestrutura de TI.



Algumas Aplicações

9. Processamento de Linguagem Natural (NLP):

- A estatística está no núcleo de técnicas de NLP, usadas para interpretar e gerar linguagem humana. Modelos como **modelos ocultos de Markov** e **modelos n-grama** são usados para prever palavras ou frases em textos.
- **Estatísticas bayesianas** são amplamente usadas para análise de sentimento, tradução automática e sistemas de diálogo.

Nos campos de linguística computacional e probabilidade, um **n-grama** é uma sequência contígua de n itens de uma determinada amostra de texto ou fala.

Os itens podem ser fonemas, sílabas, letras, palavras ou pares de bases de acordo com a aplicação.



Algumas Aplicações

10. Visualização de Dados:

- Ferramentas estatísticas são usadas para criar **visualizações de dados** que ajudam analistas e desenvolvedores a entenderem grandes volumes de dados de forma clara e intuitiva.
- Através de gráficos de dispersão, histogramas, **boxplots** e **heatmaps** (**mapas de calor para sites**), **insights** sobre o desempenho dos sistemas e comportamento dos usuários são extraídos.



Sorte ... Ao acaso?



Sorte é Isso ...



Algumas Reflexões ...



UM POUCO MAIS ...

QUIZ CASE PRÁTICO ...



QUIZ Introdutório ...

A empresa **AFA Data Science** estudando o comportamento dos consumidores do município de Varginha-MG, elegeu vários indicadores que considerava relevantes para os empresários do comércio varejista. Entre esses indicadores, foram destaque:

- Taxa de Comprometimento da Renda do Consumidor:** diz respeito à parcela da renda dos consumidores que está comprometida com contas ou dívidas tais como cheques pré-datados, cartões de crédito, carnês de lojas, empréstimo pessoal, compra de imóvel e prestações de carro e de seguros.

- Taxa de Inadimplência em Potencial:** é a taxa que reflete o número de consumidores que não terão condições de pagar contas ou dívidas atrasadas no próximo mês, no que se referem a cheques pré-datados, cartões de crédito, carnês de lojas, empréstimo pessoal, compra de imóvel e prestações de carro e de seguros.

QUIZ Introdutório ...

Durante os últimos doze meses, os resultados obtidos pela empresa AFA *Data Science* são mostrados no Quadro 1 a seguir.

	Meses											
	1	2	3	4	5	6	7	8	9	10	11	12
X	21,86	22,14	24,82	28,48	21,63	21,09	20,75	21,96	27,26	23,70	25,80	23,29
Y	4,36	4,74	6,56	9,03	3,39	5,03	3,80	3,61	8,01	5,80	6,22	5,52

Onde:

X: Taxa de Comprometimento da Renda do Consumidor

e

Y: Taxa de Inadimplência em Potencial

Pede-se determinar o **coeficiente de correlação** para os indicadores e interpretá-lo. Que outras variáveis podemos citar que contribuem para a formação da inadimplência dos consumidores nos dias atuais e, que são importantes para a tomada de decisão a nível organizacional e empresarial?

CONCLUSÃO ...



Aplicação 1

Para uma empresa manter-se competitiva no mercado globalizado dos dias atuais, gastos em Pesquisas e Desenvolvimentos (P&D) são essenciais, ou seja, são estrategicamente fundamentais. Para determinar o nível ótimo de gastos em P&D e seu efeito sobre o valor da empresa, aplicamos a Análise de Regressão Linear Simples, onde:

Y = razão entre preços e ganhos

X = razão entre gastos com P&D e vendas

Uma determinada pesquisa estudou 20 empresas e os dados estão dispostos no Quadro a seguir.

Aplicação 1

Empresas	Y	X
1	5,6	0,003
2	7,2	0,004
3	8,1	0,009
4	9,9	0,021
5	6,0	0,023
6	8,2	0,030
7	6,3	0,035
8	10,0	0,037
9	8,5	0,044
10	13,2	0,051
11	8,4	0,058
12	11,1	0,058
13	11,1	0,067
14	13,2	0,080
15	13,4	0,080
16	11,5	0,083
17	9,8	0,091
18	16,1	0,092
19	7,0	0,064
20	5,9	0,028

Aplicação 1

A partir da descrição dos resultados acima, pede-se:

- a) Encontrar a reta de regressão, ou seja, a descrição da equação $Y = A \cdot X + B$.
- b) Use a equação obtida na letra (a) para prever o valor de Y , quando $X = R\$0,080$.
- c) Qual seria o valor estimado para $X = 0,095$?
- d) Qual é o valor do intercepto? Para tal, basta fazer $X = 0$ na equação da letra (a).
- e) Qual o valor esperado de Y , quando $X = R\$0,095$?

A RELAÇÃO FORMAL DA ESTATÍSTICA COM A ÁREA COMPUTACIONAL E TI



QUAL A RAZÃO DA ESTATÍSTICA PARA A ÁREA COMPUTACIONAL E TI?

- ❑ A Estatística aplicada à computação e TI permite a extração de valor a partir de dados, otimização de processos e melhoria de sistemas, possibilitando a criação de modelos preditivos e a descoberta de padrões e tendências.
- ❑ Ela oferece a base para muitas tecnologias inovadoras, como aprendizado de máquina, *big data* e segurança cibernética, desempenhando um papel fundamental no avanço dessas áreas.



QUAL A RAZÃO DA ESTATÍSTICA PARA A ÁREA COMPUTACIONAL E TI?

- ❑ Na área computacional, tanto a **estatística descritiva** quanto a **estatística inferencial** desempenham **papéis essenciais** no processamento de dados, aprendizado de máquina, inteligência artificial e análise de grandes volumes de informação.
- ❑ Vamos explorar os conceitos fundamentais, propriedades e métodos dessas abordagens.



A Relação Estatística e Computação

É um conjunto de métodos e processos quantitativos que serve para estudar e medir os fenômenos coletivos

- Visão Antiga

É o ramo da Matemática Aplicada que está preocupada com a variabilidade e seu impacto na tomada de decisão

- Visão Moderna

CONCEITOS FUNDAMENTAIS DA ESTATÍSTICA APLICADA A ÁREA COMPUTACIONAL E TI

Hum ... Hum ...



População e Amostra

Recenseamento e Censo

Variáveis e Dados

Amostragem

Tipos de Amostragem

Hum ... Hum ...



Medidas Descritivas

Regressão Linear

Testes de Hipóteses

Correlação

Outros Métodos

Dados e Fontes de Dados



Dados e Fontes de Dados

POPULAÇÃO DO BRASIL

203.062.512 habitantes

(dados do IBGE — Censo Demográfico de 2022)



**Distribuição da população
brasileira por região**

Taxa de crescimento
anual da população:

0,52%

(menor índice desde 1872)

Classificação dos Dados

Primários

Os **dados primários** são coletados em primeira mão por quem está fazendo ou auxiliando na pesquisa/estudo.

Secundários

Os **dados secundários** são dados que já foram coletados anteriormente em outras pesquisas e que podem ser utilizados para auxiliarem em novos estudos ou pesquisas.



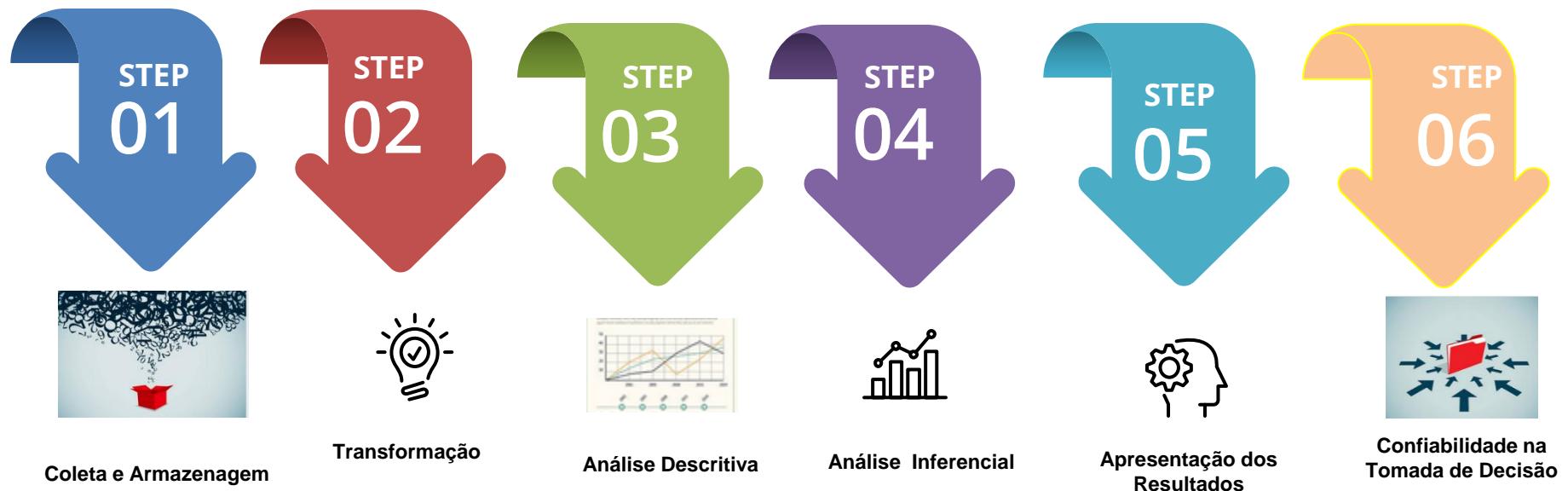
Classificação dos Dados

Classificações de Dados - Origem

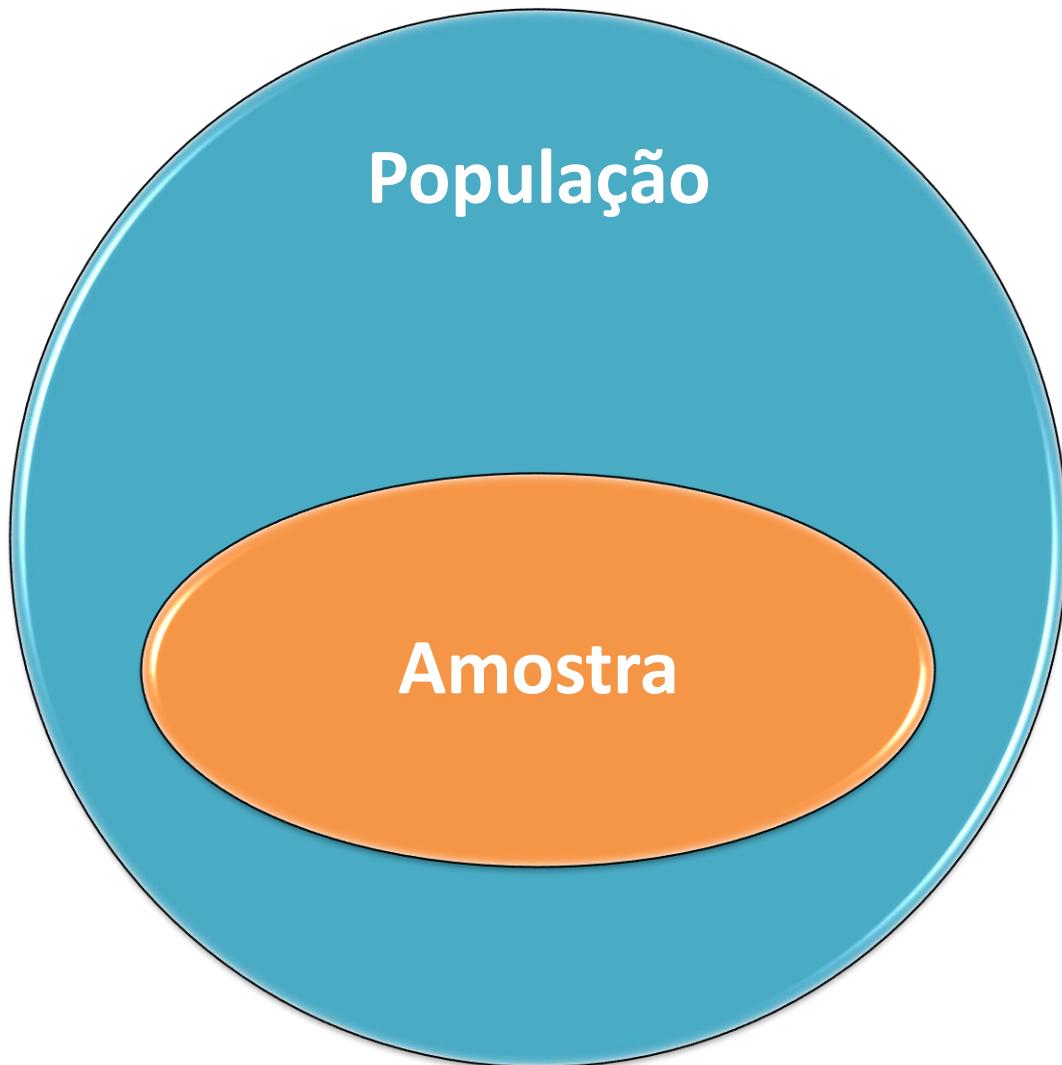
Dados Primários e Secundários



Implementação Numérica de Dados

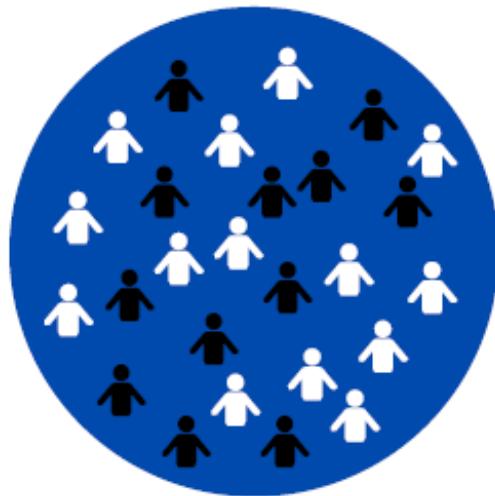


Conceitos Fundamentais



Conceitos Fundamentais

POPULAÇÃO →



Totalidade dos elementos que compõem um determinado conjunto



AMOSTRA →

Uma parte dos elementos que compõem a população

Interpretando: População e Amostra

- ❑ É por intermédio da amostra que podemos inferir sobre os parâmetros populacionais.
- ❑ A amostra deve ter como propriedade fundamental a representatividade.
- ❑ Se o tamanho dessas amostras cresce **mais precisas** são as conclusões obtidas.
- ❑ Experimentos com amostras muito grandes se aproximam de um **censo**.



Conceitos Fundamentais

População

Amostra

Parâmetro

Estimador

Dados

Elementos

Variável

Observação

Unidade
Estatística

Conceitos Fundamentais

Conceito	Descrição
População	É o conjunto formado pelas medidas que se fazem sobre elementos do Universo.
Amostra	Qualquer subconjunto não vazio de uma população.
Parâmetro	É uma característica numérica estabelecida para toda uma população.
Estimador	É uma característica numérica estabelecida para uma amostra. Alguns autores usam o termo estatística .
Dados	São os fatos e números coletados, analisados e sintetizados para apresentação e interpretação. Juntos, os dados coletados em um estudo particular são denominados o conjunto de dados para o estudo.
Elementos	São as entidades as quais os dados são coletados.
Variável	É uma característica de interesse para os elementos.
Observação	É o conjunto de medidas coletadas para um determinado elemento.
Unidade Estatística	É cada elemento da população.

Processos Estatísticos de Abordagem



PresenterMedia

**Levantamento por
Recenseamento**

Levantamento por Amostragem

Processos Estatísticos de Abordagem



PresenterMedia

Levantamento por Recenseamento

Censo – é o conjunto de dados obtidos através de recenseamento, ou seja, quando o estudo é realizado tomando como base toda a população.

Levantamento por Amostragem

Amostra – é tanto a parte retirada da população para estudo como, também, o conjunto de dados obtidos nessa parte da população.

Recenseamento versus Amostragem?



Vantagens?

Desvantagens?

Vantagens da Amostragem sobre o Recenseamento

Vantagem	Descrição
Tempo	Quando utilizamos a amostragem ao invés do censo, gastamos pouco tempo para se concluir o estudo, pois se está trabalhando com menos elementos.
Custo	O custo da pesquisa é reduzido quando se realiza uma amostragem ao invés do censo.
Aprofundamento	A pesquisa amostral pode ser mais aprofundada , visto que, examinamos uma menor quantidade de elementos.
Erros	Como se examinam menos elementos, o número de possíveis pesquisadores (coletores, entrevistadores) diminui, diminuindo com isso os erros que poderiam ocorrer, visto que poderemos treinar melhor os elementos envolvidos na pesquisa.
Testes Destrutivos	Quando a pesquisa envolve testes destrutivos, não é possível realizar o censo, pois teríamos resultados exatos de uma população que já não existe mais, como seria o caso do teste de durabilidade de lâmpadas.

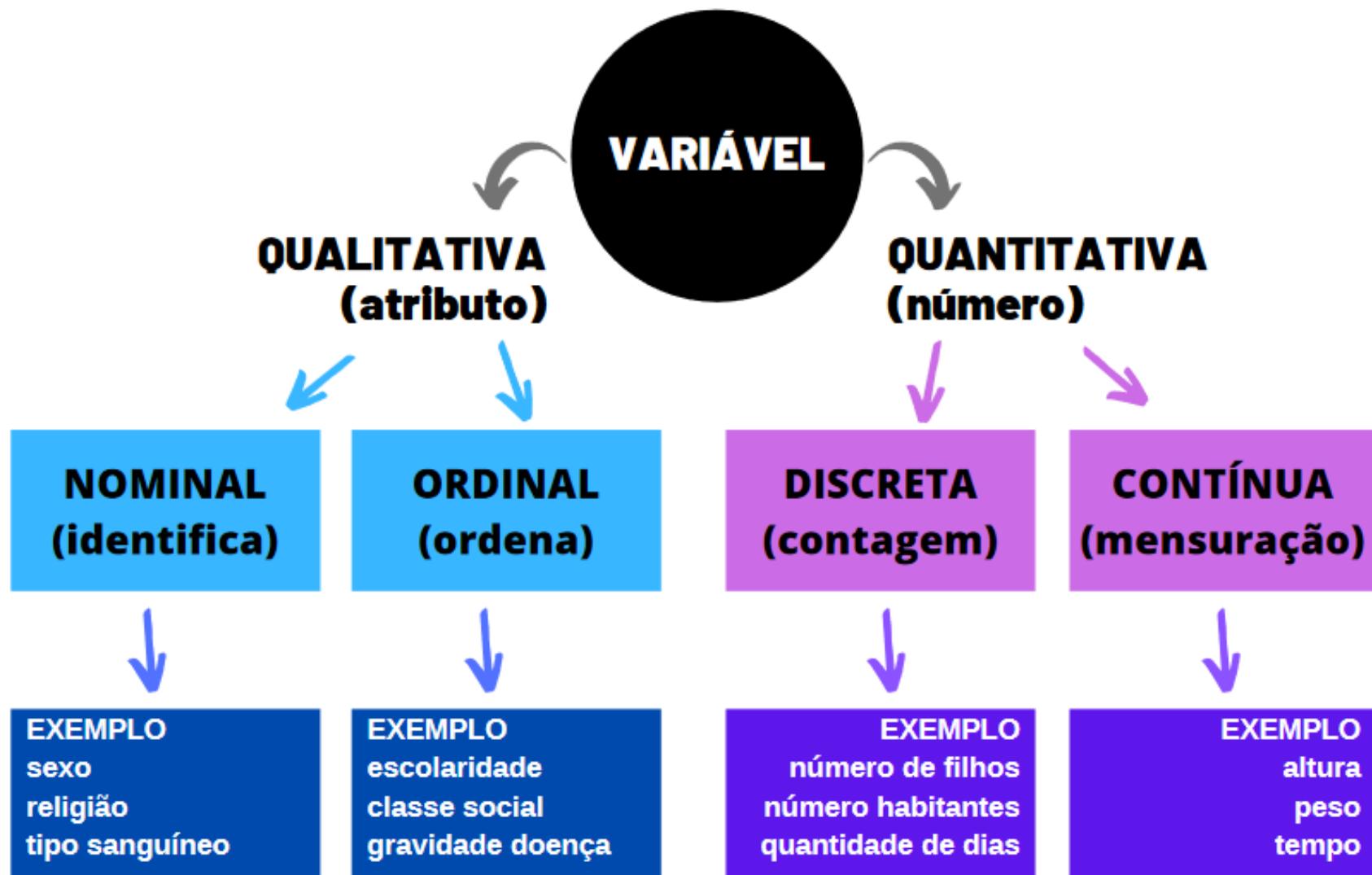
Vantagens do Recenseamento sobre a Amostragem

Vantagem	Descrição
Precisão Completa	Quando há interesse na precisão completa do levantamento não se pode realizar um levantamento por amostragem.
População Pequena	Quando a população é relativamente pequena , não é vantajoso trabalhar com amostra, pois com um pouquinho mais de trabalho a precisão seria completa.

Tipologia das Variáveis



Tipos de Variáveis: Fundamental!



Variáveis Quantitativas

Variável	Descrição
Variável Quantitativa Discreta	Os possíveis valores formam um conjunto finito ou enumerável de números (contagem) . Ex: idade, número de filhos, número de patas de um cavalo, número de pétalas das flores , etc.
Variável Quantitativa Contínua	Os possíveis valores pertencem a um intervalo de números reais e que resultam de uma mensuração . Ex: estatura, massa de um indivíduo, altura de uma árvore , etc.

Variáveis Qualitativas

Variável	Descrição
Variável Qualitativa Nominal	Não existe nenhuma ordenação nos possíveis resultados. Ex: região de procedência dos funcionários de uma clínica veterinária, sexo de animais.
Variável Qualitativa Ordinal	Existe uma ordem nos seus resultados . Ex: grau de instrução (ensino fundamental, médio e superior), classe social (alta, média e baixa), nível de intensidade de dor.

Tipos de Dados



**Dados
Discretos**

**Dados
Contínuos**

**Dados
Nominais**

**Dados
Ordinais**

Processos Estatísticos de Abordagem



População?

Amostra?

Um Pouco Mais ...



Dados
Brutos



Rol



Quiz ...



Quiz Introdutório ... (Adaptado ENADE)

Numa fábrica produtora de iogurtes, produzem-se dez mil unidades diariamente. Para efetuar a gestão da qualidade analisam-se cem iogurtes medindo o peso líquido e observando a consistência. Num estudo estatístico referente a esta situação, o que seria?

- ✓ A população?
- ✓ A unidade estatística?
- ✓ A amostra?
- ✓ As variáveis observadas?
- ✓ Que tipos de dados temos aqui?



CONCLUSÃO



RESPOSTA DO QUIZ



- ✓ População – os 10 000 iogurtes.
- ✓ Unidade estatística – cada iogurte.
- ✓ Amostra – os 100 iogurtes analisados.
- ✓ Variáveis observadas – peso líquido (variável quantitativa contínua); consistência (variável qualitativa ordinal (possíveis modalidades – elevada, alta, média, baixa, etc.).
- ✓ Tipos de dados – dados contínuos e dados ordinais.

Processos Estatísticos de Abordagem



População?

Amostra?

Técnicas de Amostragem



Amostragem Probabilística

Amostragem Não probabilística

Amostragem Probabilística

Na amostragem probabilística, todos os indivíduos da população têm uma chance conhecida e diferente de zero de serem selecionados. Isso permite fazer inferências estatísticas mais confiáveis sobre a população.

Principais Tipos de Amostragem Probabilística

Amostragem Aleatória Simples

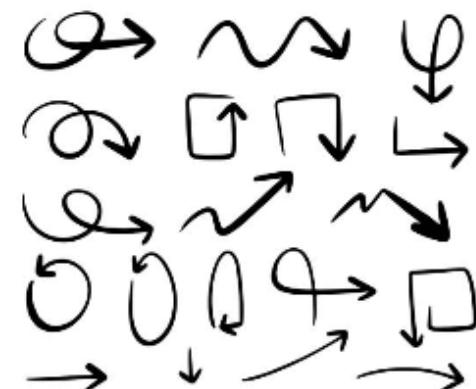
Amostragem Sistemática

Amostragem Estratificada

Amostragem por Conglomerados

Amostragem Aleatória Simples

- **Conceito:** Cada elemento da população tem a mesma probabilidade de ser selecionado.
- **Exemplo:** Sorteio de números para uma pesquisa, onde cada número tem a mesma chance de ser escolhido.
- **Aplicação:** Usada quando a população é relativamente pequena e homogênea.



Amostragem Aleatória Simples

Tipo de Amostragem	Descrição
Amostragem Aleatória Simples	<p>Consiste em enumerarmos os elementos de uma população (N elementos) e escolhermos os n elementos dessa sequência, que comporão a amostra, através de um dispositivo aleatório qualquer (sorteio ou tabela de números aleatórios, etc.).</p>
Exemplos	<ul style="list-style-type: none"><input type="checkbox"/> Sorteio aleatório de colaboradores na AFA Negócios Internacionais.<input type="checkbox"/> Computadores para gerar números aleatórios.<input type="checkbox"/> Entrevistar pessoas no centro de uma cidade para uma dada pesquisa eleitoral.

Amostragem Sistemática

- . **Conceito:** Seleciona-se um ponto de partida aleatório e, em seguida, escolhe-se cada k -ésimo elemento da população.
- . **Exemplo:** Se há uma lista de 1.000 pessoas e deseja-se uma amostra de 100, escolhe-se uma pessoa inicial e, a cada 10 pessoas ($1000/100 = 10$), uma nova pessoa é selecionada.
- . **Aplicação:** Utilizada quando se tem uma lista organizada ou uma sequência de elementos (ordenação).

Amostragem Sistemática

Tipo de Amostragem	Descrição
Amostragem Sistemática	<p>Em verdade, representa uma abreviação do caso anterior. É normalmente usada quando os elementos da população se apresentam ordenados e a retirada dos elementos da amostra é feita periodicamente.</p>
Exemplos Ilustrativos	<ul style="list-style-type: none"><li data-bbox="764 688 1875 961"><input type="checkbox"/> Se a AFA Negócios Internacionais quisesse fazer uma pesquisa sobre seus 107.000 empregados, poderia partir de uma relação completa dos mesmos e selecionar cada 100º empregado, obtendo uma amostra de 1.070 elementos.<li data-bbox="764 1033 1875 1249"><input type="checkbox"/> Se estivermos interessados numa amostra de 5% das fichas cadastrais dos 200 clientes de uma agência de publicidade, poderíamos pegar cada 20º ficha cadastral.

Amostragem Estratificada

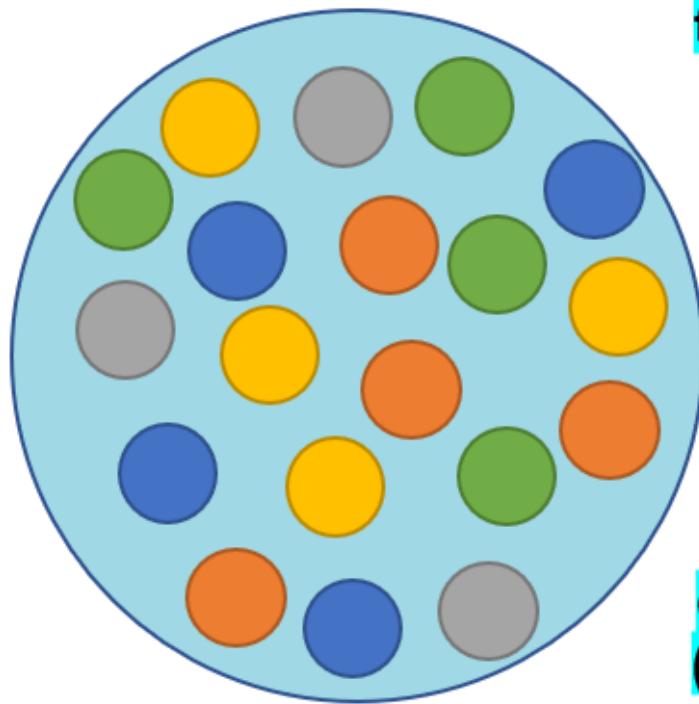
- . **Conceito:** A população é dividida em estratos (subgrupos homogêneos) e uma amostra aleatória simples é retirada de cada estrato.
- . **Exemplo:** Dividir uma população por faixa etária e, em seguida, selecionar aleatoriamente dentro de cada faixa.
- . **Aplicação:** Usada para garantir que subgrupos específicos estejam adequadamente representados na amostra.

Amostragem Estratificada

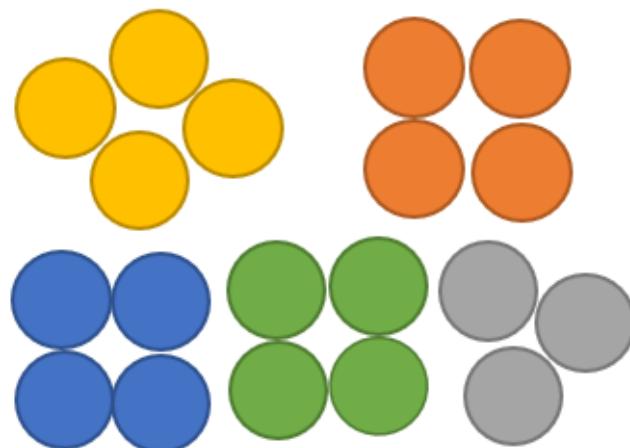
Tipo de Amostragem	Descrição
Amostragem Estratificada	<p>Subdividimos a população em, no mínimo, duas subpopulações (ou estratos) que compartilham das mesmas características (como sexo) e, em seguida, extraímos uma amostra de cada estrato.</p>
Exemplos Ilustrativos	<ul style="list-style-type: none">□ Em uma pesquisa sobre a Emenda Constitucional da Igualdade de Direitos, poderíamos utilizar o sexo como base para a criação de dois estratos. Após obter uma relação dos homens e uma relação das mulheres, aplicamos um método conveniente (como amostragem aleatória) para escolher determinado número de elementos de cada relação.□ Dividir a linha de produção de uma fábrica por turnos, cada turno seria um estrato.□ Para obtermos uma amostra de pessoas de Varginha, dividimos por bairro (estrato) e depois reunimos informações numa amostra estratificada.

Amostragem Estratificada

População

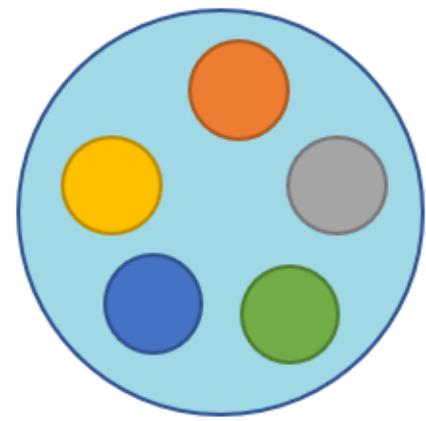


Subpopulações são formadas por características



Cada pessoa ganha um número
(que se repete em cada estrato)

Amostra é feita



Eles são “sorteados”
em seus estratos

Amostragem por Conglomerados

- . **Conceito:** A população é dividida em grupos naturais (**conglomerados**), e alguns desses conglomerados são selecionados aleatoriamente. Em seguida, todos os indivíduos dentro dos conglomerados escolhidos são analisados.
- . **Exemplo:** Uma pesquisa escolar que seleciona algumas escolas ao acaso e entrevista todos os alunos dessas escolas.
- . **Aplicação:** Utilizada quando a população está naturalmente dividida em grupos, como cidades, escolas, bairros, etc.

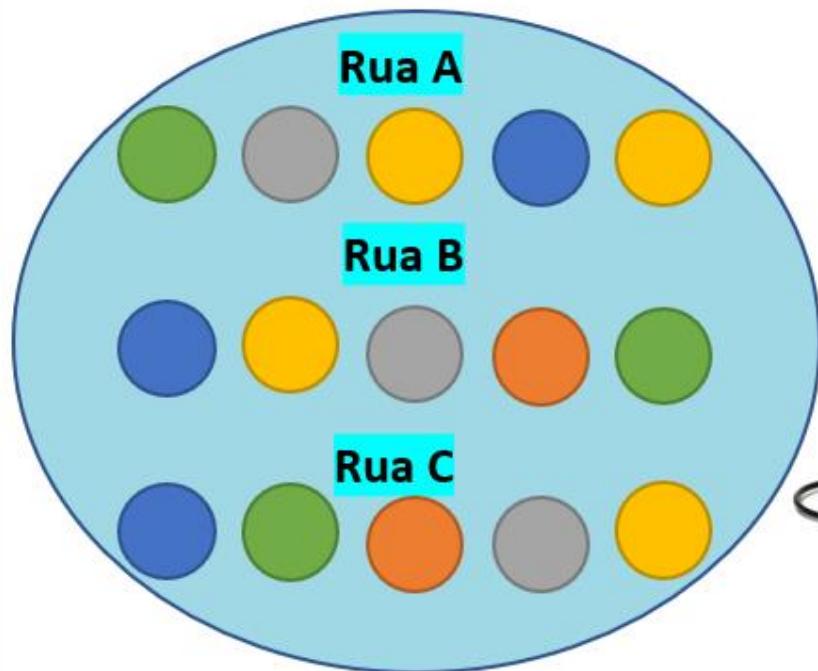


Amostragem por Conglomerados

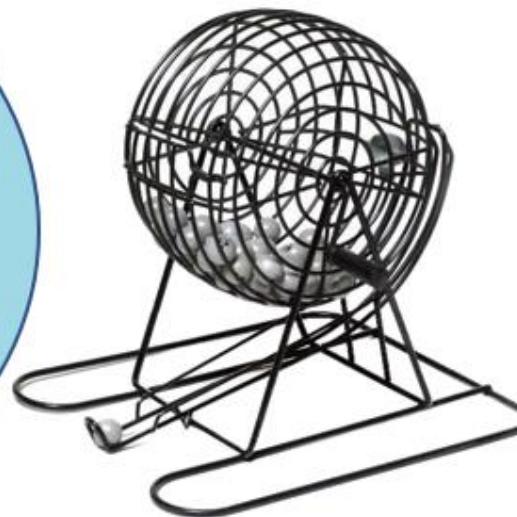
Tipo de Amostragem	Descrição
Amostragem por Conglomerados	<p>Começamos dividindo a área da população em seções (ou conglomerados); em seguida escolhemos algumas dessas seções e, finalmente, tomamos todos os elementos das seções escolhidas. É extensamente utilizada pelo governo e por organizações particulares de pesquisa.</p>
Exemplos Ilustrativos	<ul style="list-style-type: none"><input type="checkbox"/> Em uma pesquisa pré-eleitoral, onde escolhemos aleatoriamente 30 zonas eleitorais e pesquisamos todos os elementos de cada uma das zonas escolhidas.<input type="checkbox"/> Estimar o número de funcionários com alguma determinada característica de uma determinada região administrativa; neste caso serão selecionados alguns municípios dessa região para compor a amostra.

Amostragem por Conglomerados

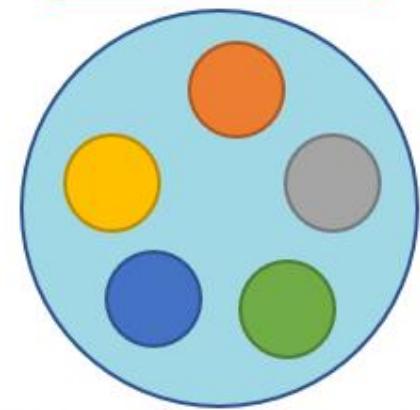
População Cada conglomerado ganha um número



Eles são “sorteados”



Amostra é feita



Todos os participantes
do conglomerado
são acessados

Amostragem Não Probabilística

Na amostragem não probabilística, a seleção dos indivíduos não segue critérios aleatórios, ou seja, não é possível garantir que todos os elementos da população tenham a mesma chance de serem escolhidos. Isso reduz a capacidade de fazer inferências precisas sobre a população.

Um Exemplo Típico de Amostragem Não Probabilística

Amostragem por Conveniência

Descrição da Amostragem por Conveniência

- . **Conceito:** A amostra é formada pelos indivíduos que estão mais facilmente acessíveis.
- . **Exemplo:** Entrevistar as pessoas disponíveis em uma rua para uma pesquisa de opinião.
- . **Aplicação:** Usada quando há limitações de tempo ou recursos, mas tem baixa representatividade.

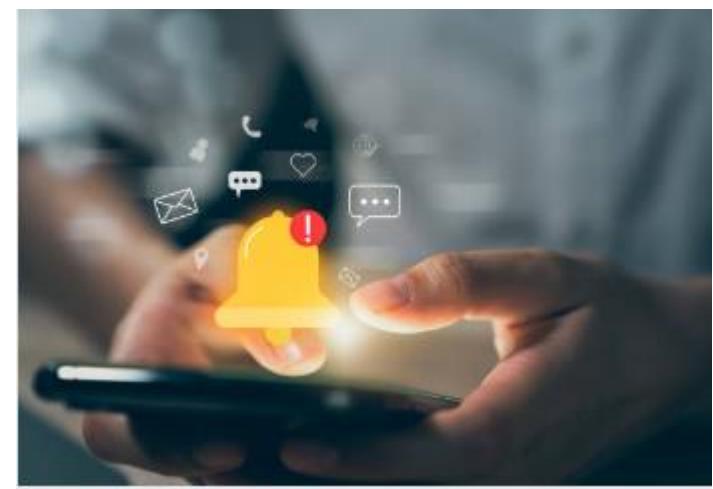


Descrição da Amostragem por Conveniência

Tipo de Amostragem	Descrição
Amostragem por Conveniência	<p>É aquele processo que se baseia no que é mais conveniente para o pesquisador. É mais prático e econômico, mas em muitos casos pode ser tendenciosa.</p> <p>(acessibilidade)</p>
Exemplos Ilustrativos	<ul style="list-style-type: none">□ Ao fazer uma pesquisa sobre pessoas canhotas, seria conveniente um estudante pesquisar seus próprios colegas de classe, porque estão ao seu alcance imediato (resultados podem ser bem satisfatórios). □ Poderia ser muito conveniente (e talvez menos lucrativo) para a Rede Globo de Televisão fazer uma pesquisa pedindo aos espectadores que liguem para um número de telefone “300” para registrar suas opiniões, mas essa pesquisa seria auto selecionada e os resultados seriam provavelmente tendenciosos.

Importante

A escolha entre amostragem probabilística e não probabilística depende do objetivo da pesquisa, da disponibilidade de recursos e da necessidade de fazer inferências sobre a população geral.





Problemas Simulados

Problemas Simulados

1) Caracterizar o tipo de amostragem em cada exemplo abaixo:

- a) Um psicólogo da Motorola faz uma pesquisa sobre todos os funcionários de cada uma de 20 turmas selecionadas aleatoriamente.
- b) Um sociólogo na AFA Logística seleciona 15 homens e 15 mulheres de cada uma de quatro turmas de uma linha de produção de parafusos.
- c) A empresa SONY seleciona cada 10º CD de sua linha de produção e faz um teste de qualidade rigoroso.
- d) Um cabo eleitoral escreve o nome de cada senador do Brasil, em cartões separados, mistura-os e extrai 12 nomes.

Problemas Simulados

- e) O gerente comercial da AFA Ltda testa uma nova estratégia de vendas selecionando aleatoriamente 150 consumidores com renda inferior a R\$5.000,00 e 150 consumidores com renda de ao menos R\$5.000,00.
- f) O programa de Planejamento Familiar do governo federal pesquisa 500 homens e 500 mulheres sobre seus pontos de vista sobre o uso de anticoncepcionais.
- g) Um pesquisador industrial entrevista todos os colaboradores da área da qualidade, em cada uma de cinco filiais de uma multinacional na área automotiva.
- h) Um pesquisador de mercado da TAM entrevista todos os passageiros de cada um de 20 voos selecionados aleatoriamente.

Problemas Simulados

- i) Um repórter da revista VEJA entrevista todo trigésimo gerente geral constante da relação das 1.000 empresas na área de varejo com maior cotação de suas ações.
- j) Um repórter da revista ISTO É obtém uma relação numerada das 1000 empresas com maiores cotações de ações na bolsa, utiliza um computador para gerar 25 números aleatórios e então entrevista os gerentes gerais das empresas correspondentes aos números extraídos.

Estatística Descritiva

1. Estatística Descritiva

A **estatística descritiva** se refere a métodos que descrevem, organizam e resumem os dados. Ela é especialmente útil para transformar grandes volumes de dados em informações compreensíveis, destacando suas características mais importantes.

Conceitos Fundamentais:

- **Medidas de Tendência Central:** Fornecem um valor "representativo" dos dados.
 - Média, mediana e moda.
- **Medidas de Dispersão:** Avaliam a variabilidade dos dados.
 - Variância, desvio padrão, amplitude, quartis.
- **Distribuição de Frequências:** Representação de dados em tabelas ou gráficos.
 - Histogramas, gráficos de barras, *boxplots*.

Estatística Descritiva

Aplicações na Computação:

- **Pré-processamento de Dados:** Normalização e padronização de dados antes de alimentá-los em algoritmos de aprendizado de máquina.
- **Visualização de Dados:** Uso de gráficos para melhor compreensão dos dados.
- **Sumarização de Grandes Bases de Dados:** Redução de dimensões sem perder a essência dos dados, ajudando em análises mais rápidas e eficientes.

Propriedades:

- **Resistência a *outliers*:** Medidas como a mediana são menos sensíveis a valores extremos.
- **Simplicidade:** Usar poucos números para descrever o comportamento geral de um conjunto de dados.

Estatística Inferencial

2. Estatística Inferencial

A **estatística inferencial** vai além de descrever os dados e faz suposições ou inferências sobre uma população com base em uma amostra.

Conceitos Fundamentais:

- **Estimação de Parâmetros:** Uso de amostras para estimar características de uma população (como a média populacional ou a proporção).
 - Intervalos de confiança.
- **Testes de Hipóteses:** Avaliação de suposições sobre os dados por meio de testes estatísticos.
 - Testes t, ANOVA, testes qui-quadrado.
- **Distribuições de Probabilidade:** Modelos que descrevem como os dados se comportam.
 - Distribuição normal, binomial, Poisson, etc.

Estatística Inferencial

Aplicações na Computação:

- **Algoritmos de Machine Learning:** A estatística inferencial é usada para avaliar a performance dos modelos com base em amostras (como validação cruzada e *bootstrap*). (*Bootstrap* é um *framework front-end* que fornece estruturas de CSS para a criação de sites e aplicações responsivas de forma rápida e simples).
- **Análise de Algoritmos:** A estatística inferencial permite avaliar o desempenho de algoritmos de aprendizado e ajustar modelos preditivos.
- **Inferência de Dados em Grandes Volumes:** Ao lidar com Big Data, é comum realizar inferências sobre um conjunto completo de dados usando apenas amostras.
- **Estatística Bayesiana:** Um ramo da inferência, que se baseia na teoria de probabilidade bayesiana, é amplamente usado em aprendizado de máquina e IA para atualizar probabilidades com base em novas evidências.

Estatística Inferencial

Propriedades:

- **Generalização:** Permite que as conclusões sejam aplicadas a populações maiores a partir de uma amostra.
- **Controle de Erros:** Métodos como intervalos de confiança e testes de hipótese controlam a probabilidade de erros (como erros do tipo I e II).

Métodos Diversos

- **Análise de Regressão:** Para prever uma variável com base em uma ou mais variáveis preditoras.
- **Métodos de Amostragem:** Técnicas de amostragem como amostragem aleatória, estratificada, sistemática, entre outras.
- **Modelos Probabilísticos:** Uso de distribuições de probabilidade para modelar fenômenos computacionais.

Sumarizando ...

A **estatística descritiva** ajuda a entender os dados, enquanto a **estatística inferencial** permite fazer previsões e tomar decisões baseadas nesses dados.

Ambas são **fundamentais na ciência da computação e TI** para melhorar algoritmos, avaliar dados, modelar problemas e tomar decisões baseadas em informações.



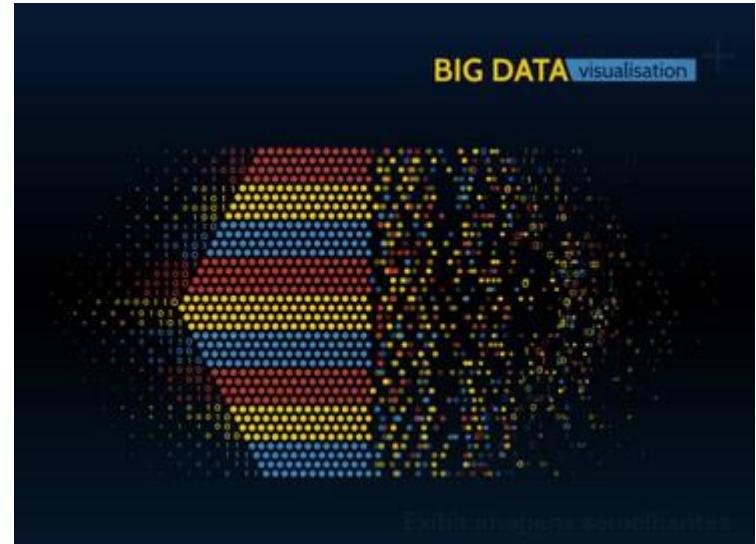
Assunto Abordado

Distribuição de Frequências,
Medidas de Centralidade e
Aplicações Diversas

Motivação

“Deus não joga dados”.

Einstein



Motivação ...



MÉDIA?



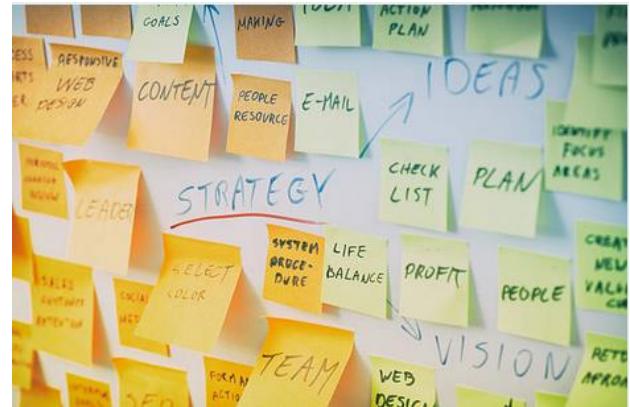
MEDIANA?

MODA?

QUARTIS?

Contextualizando ...

- Sumarização de dados?
 - Distribuição de Frequências
 - Construindo uma Distribuição de Frequências
 - Medidas Descritivas
 - Medidas de Centralidade
 - Cálculos Diversos
 - Aplicações Diversas



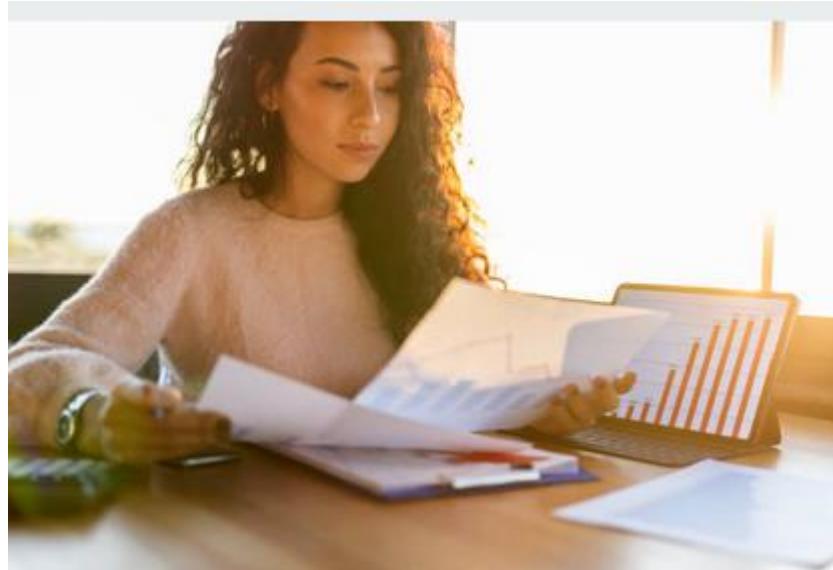
EXERCÍCIO – DESAFIO

É sabido que a maneira da Gestão da Qualidade encarar os processos se modificou ao longo do tempo, evoluindo da busca dos erros para sua prevenção e a tomada de medidas para evitá-los. Assim sendo, quando queremos verificar a amostra de um processo que apresentou maior número de erros, utilizamos qual medida descritiva?



Assunto Abordado

Distribuição de Frequências



DISTRIBUIÇÃO DE FREQUÊNCIAS

Uma **distribuição de frequência** é um agrupamento de dados em **classes**, exibindo o número ou percentagem de observações em cada classe. Uma distribuição de frequência pode ser apresentada sob a forma gráfica ou tabular.

DADOS BRUTOS E ROL



DADOS BRUTOS

ROL

EXEMPLIFICANDO ...

Tempos de serviço de 50 funcionários da AFA Engenharia. Suponhamos que cada número descrito a seguir represente o tempo de serviço em meses.

41 – 32 – 33 – 34 – 35 – 36 – 24 – 25 – 58 – 26 –
27 – 29 – 29 – 65 – 30 – 31 – 31 – 36 – 37 – 37 –
37 – 37 – 47 – 38 – 38 – 40 – 43 – 53 – 44 – 45 –
45 – 45 – 46 – 38 – 48 – 49 – 50 – 51 – 44 – 54 –
54 – 18 – 20 – 20 – 21 – 22 – 56 – 25 – 62 – 30

Dados Brutos

EXEMPLIFICANDO

Tempos de serviço de 50 funcionários (colocados em **ordem crescente**) da AFA Engenharia. Suponhamos que cada número descrito a seguir represente o tempo de serviço em meses.

18 – 20 – 20 – 21 – 22 – 24 – 25 – 25 – 26 – 27 –
29 – 29 – 30 – 30 – 31 – 31 – 32 – 33 – 34 – 35 –
36 – 36 – 37 – 37 – 37 – 37 – 38 – 38 – 38 – 40 –
41 – 43 – 44 – 44 – 45 – 45 – 45 – 46 – 47 – 48 –
49 – 50 – 51 – 53 – 54 – 54 – 56 – 58 – 62 – 65

Rol

CONSTRUINDO A DISTRIBUIÇÃO DE FREQUÊNCIAS

Passos	Descrição
01	<p>Construir o rol (dados em ordem crescente) e determinar a Amplitude Total:</p> $R = \text{Maior medida} - \text{Menor medida}$
02	<p>Como os dados serão agrupados em classe, é preciso escolher o número de classes – K, bem como o tamanho do intervalo de classes – h.</p> <p>1 <u>Critério: Fórmula de Sturges</u></p> $K = 1 + 3,33 \cdot \log n$ $K = 1 + 3,33 \cdot \log 50 = 1 + 3,33 \cdot (1,698970004) = 6,657570114, \text{ logo}$ <p>k = 7 classes</p> <p>2 <u>Critério: Regra Empírica</u></p>

CONSTRUINDO A DISTRIBUIÇÃO DE FREQUÊNCIAS

Passos	Descrição
03	<p>Quanto ao tamanho dos intervalos (iguais) das classes h:</p> $h = R \div K$ <p>No exemplo: $h = 47 \div 7 = 6,714285714$, logo h = 7.</p>
04	<p>Quanto aos limites das classes, utilizaremos o seguinte critério: (incluiremos 18 + 7 25. nesta classe todos os elementos maiores ou iguais a 18 e menores do que 25). Ou ainda, podemos representar da seguinte maneira: 18 ---- 25.</p>

DEFININDO AS FREQUÊNCIAS

Frequência Absoluta

- É o número de vezes que uma determinada característica ou valor numérico é observada.

DEFININDO AS FREQUÊNCIAS

Frequência Relativa

- É a proporção, do total, em que é observada uma determinada característica.
- Sob determinadas condições, as frequências relativas podem ser usadas para estimar quantidades importantes como por exemplo, em epidemiologia.
- Este conceito está associado com a definição clássica de probabilidade.

DEFININDO AS FREQUÊNCIAS

Frequência Acumulada

- Para um determinado valor numérico ou dado ordinal, é a soma das frequências dos valores menores ou iguais ao referido valor.

DEFININDO AS FREQUÊNCIAS

F_i: Frequência Absoluta da classe i

f_i: Frequência Relativa da classe i

$$f_i = \frac{F_i}{n} \quad \text{ou} \quad f_i = \frac{F_i}{N}$$

Fac: Frequência Acumulada

Onde:

n = tamanho da amostra

N = tamanho da população

A TABELA DE FREQUÊNCIAS DO EXEMPLO

Classes	Intervalos de Classes	Fi	fi	Fac
1	18 ----- 25	6	0,12	6
2	25 ----- 32	10	0,20	16
3	32 ----- 39	13	0,26	29
4	39 ----- 46	8	0,16	37
5	46 ----- 53	6	0,12	43
6	53 ----- 60	5	0,10	48
7	60 ----- 67	2	0,04	50
Somas		50	1	

CONCLUSÃO

- ✓ A maior quantidade de funcionários tem tempo de serviço entre 32 e 39 meses.
- ✓ Apenas 4% dos funcionários possuem tempos de serviço iguais ou superiores a 60 meses, sendo 65 meses o maior tempo de serviço do grupo.
- ✓ Cinquenta e oito por cento dos funcionários da amostra têm tempos de serviço inferiores a 39 meses, sendo 18 meses o menor tempo de serviço do grupo.

RESUMINDO

- ✓ A tabela de distribuição das frequências sintetiza e organiza uma coleção de dados, facilitando a compreensão e análise da variável sob estudo.
- ✓ A distribuição de frequências é a primeira ferramenta da Estatística a fim de organizar e caracterizar um conjunto de dados de uma forma bem simples.
- ✓ A distribuição de frequências é a primeira ferramenta da Estatística a fim de organizar e caracterizar um conjunto de dados de uma forma bem simples.