

Marcelo Colunno
Cientista de Dados
Bayer Crop Science

Formulação e
Levantamento de
Hipóteses

AGENDA

- **Bloco 1: Pensamento Científico e Incerteza**
- **Bloco 2: Princípio da Falseabilidade (Karl Popper)**
+ Intervalo - 10 min
- **Bloco 3: Levantamento de Hipóteses**
- **Bloco 4: População e Amostra**
- **Dúvidas e reflexões finais**
- **Como foi?**



Pensamento Científico e Incerteza

O que é Ciência?

- Todo o conhecimento adquirido através do estudo, **pesquisa ou da prática, baseado em princípios certos**. Esta palavra deriva do latim scientia, cujo significado é "conhecimento" ou "saber".
- Em geral, a ciência, que é muito ampla, comporta vários conjuntos de saberes nos quais são elaboradas as suas **teorias baseadas nos seus próprios métodos e pesquisas científicas**.
- Conhecimento de certas coisas que servem à **condução da vida ou à dos negócios**.
- Conjunto dos **conhecimentos adquiridos pelo estudo ou pela prática**.
- Hierarquização, organização e síntese dos conhecimentos através de **modelos e princípios gerais** (teorias, leis, etc.).

Método Científico

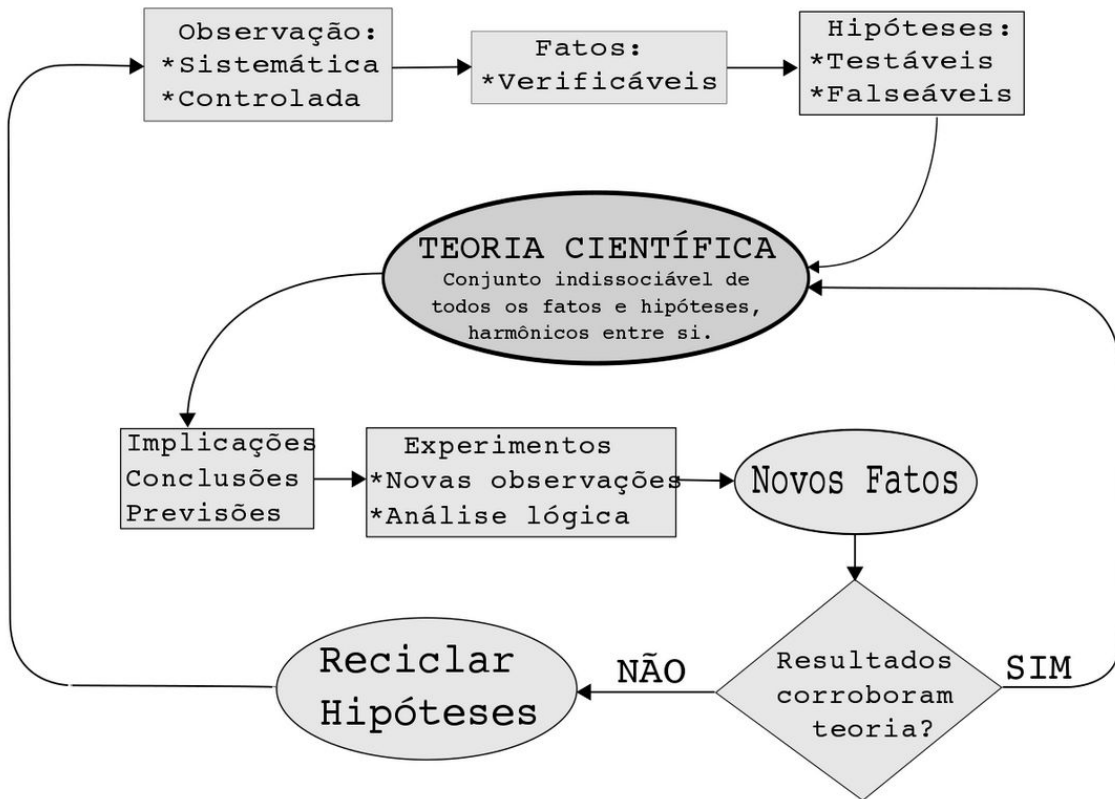
- **Aglomerado de regras básicas** dos procedimentos que produzem o conhecimento científico, quer um novo conhecimento, quer uma correção (evolução) ou um aumento na área de incidência de conhecimentos anteriormente existentes.
- Juntar **evidências empíricas verificáveis** baseadas na observação sistemática e controlada, geralmente resultantes de experiências ou pesquisa de campo, e analisá-las com o uso da lógica.
- **Bases lógicas ao conhecimento científico** são: método indutivo, método dedutivo, método hipotético-dedutivo, método dialético, método fenomenológico, etc
- [Vídeo: Método Científico](#)

Método Científico

O método é cíclico, girando em torno do que se denomina teoria científica, a união indissociável do conjunto de todos os fatos científicos conhecidos e de um **conjunto de hipóteses testáveis** e testadas capaz de explicá-los.

MÉTODO CIENTÍFICO

(Esboço)




Incerteza na Ciência

- É um dos elementos integrantes do processo de conhecimento e **sua avaliação faz parte do método científico**
- A incerteza pode ser categorizada e abordada de vários pontos de vista, lógicos, matemáticos, filosóficos ou **estatísticos**, e é especialmente relevante **quando a ciência faz previsões**
- É um elemento complexo que se refere **à probabilidade de um evento ocorrer** dentro de certos parâmetros
- A incerteza vem se tornando um campo autônomo de estudos científicos em anos recentes. Por quê? (**Análise de Riscos**)
- Exemplos? Como lidar com a **causalidade dos eventos**?



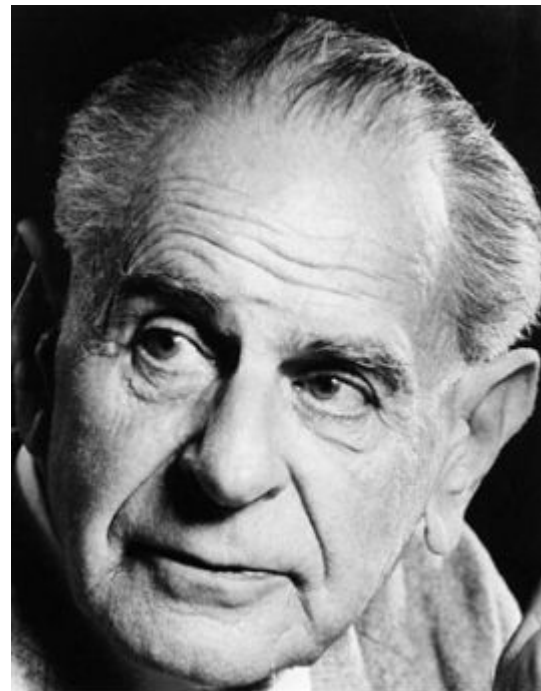
**ALGUMA DÚVIDA ATÉ
AQUI?**



Princípio da Falseabilidade (Karl Popper)

Karl Popper

- Karl Raimund Popper nasceu na Áustria, em 1902. Filho de judeus
- Em 1935, publicou a obra *“Lógica da Investigação Científica”*, considerada uma das obras mais importantes de **filosofia da ciência**
- Emigrou para a Nova Zelândia em 1937, onde publicou, em 1945, a obra de **filosofia política** *“A sociedade Aberta e Seus Inimigos”*.
- Um dos maiores filósofos da ciência do século XX, conhecido por sua **rejeição das visões indutivistas** clássicas sobre o método científico em favor do falsificacionismo.



Princípio da Falseabilidade

- **Método indutivo** seleciona os fenômenos que serão investigados para a comprovação de algo que já se supõe
- Por essa razão, o **critério de verificabilidade** nem sempre será válido.
- O **princípio proposto por Popper**: em vez de buscar a verificação de experiências empíricas que confirmassem uma teoria, buscava **fatos particulares que**, depois de verificados, **refutariam a hipótese**.
- Em vez de se preocupar em **provar que uma teoria é** verdadeira, ele se preocupava em provar que ela **é falsa**.
- Quando a teoria resiste à refutação pela experiência, **pode ser considerada comprovada**.

Conjecturas e Refutações

**Caráter
Racional da
Ciência**

tradição
racionalista
histórica
(civilização
grega)

X

**Caráter
Hipotético
das Teorias
Científicas**

novas
conjecturas e
hipóteses
ousadas

Método Hipotético-Dedutivo

- O debate livre e crítico também aponta para o **caráter hipotético das teorias científicas**, pois elas sempre estão sujeitas a serem falseadas – ou não podem ser consideradas teorias científicas.
- A ciência é valorizada pela influência liberalizadora que exerce – uma das forças mais poderosas que **contribuiu para a liberdade humana**. (POPPER, 1972, p. 129)
- Nos dias de hoje, verifica-se que o falsificacionismo popperiano não é princípio de exclusão, mas tão somente de atribuição de **graus de confiança** ao objeto passível do crivo científico

A vertical bar with a gradient from light green at the top to light blue at the bottom.

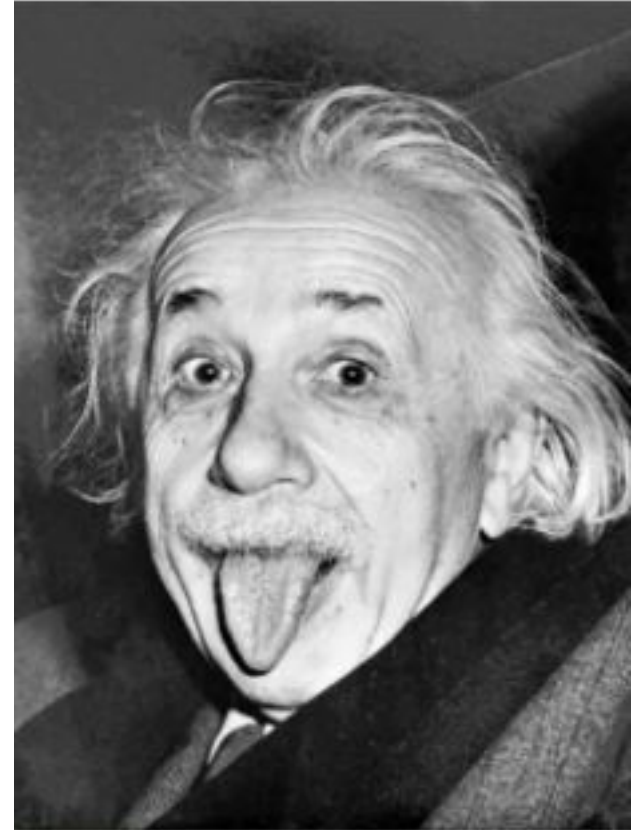
**ALGUMA DÚVIDA ATÉ
AQUI?**



Levantamento de Hipóteses

Levantamento de Hipóteses

- **hipótese é uma afirmação que introduz uma questão de pesquisa e propõe um resultado esperado.**
- É parte integrante do método científico que forma a **base de experimentos científicos.**
- Uma **hipótese testável** é uma hipótese que pode ser provada ou refutada como resultado da experimentação





Como trabalhar com Hipóteses?

- Para planejar e realizar um experimento usando o método científico, você precisa ter certeza de que **sua hipótese é testável**
- Para ser considerado testável, alguns **critérios essenciais** devem ser atendidos:
 - Deve existir a possibilidade de **provar que a hipótese é verdadeira**
 - Deve existir a possibilidade de **provar que a hipótese é falsa**
 - Os **resultados da hipótese devem ser reproduzíveis**
 - Mas como?
 - Um **set de dados** é o que possibilita trabalhar com hipóteses



Hipóteses em Ciência de Dados

- Em experimentos científicos, uma hipótese propõe e examina a relação entre uma **variável independente** e uma **variável dependente**
- O efeito sobre a **variável dependente** (a ideia que está sendo testada) depende ou é determinado pelo que acontece quando se altera a **variável independente** (o fator sendo alterado)
- Isso nos ajuda a fazer **previsões precisas com base em pesquisas anteriores**. Assim, formular uma hipótese é de grande valor para a pesquisa.
- Esse é o princípio de funcionamento de um algoritmo de **Machine Learning**

T Hipóteses na Prática

Aqui estão algumas **perguntas importantes** a serem feitas:

1. A **linguagem** é clara e focada?
2. A hipótese introduz o **tópico** de pesquisa?
3. A hipótese inclui uma **variável independente e dependente**? Eles são fáceis de identificar?
4. A hipótese pode ser testada através da **experimentação**?
5. A hipótese explica o que você **espera** que aconteça durante o **experimento**?

T Hipóteses na Prática

- A baixa na produtividade está associada a alimentação inadequada.
- Considere **x como sendo a variável independente** que representa a má alimentação e **y a variável dependente** que representa a baixa produtividade.
- Nesse caso, só é possível identificar se **existe ou não relação entre as duas variáveis**.
- Mas **não é possível determinar** qual delas poderia produzir alteração na outra – causalidade!
- Com essa hipótese, no máximo confirmaríamos se alimentação inadequada **está ou não associada** com a baixa produtividade – correlação!

T Exemplos Práticos

- **Projetos** em desenvolvimento na turma?



**Indústria
Financeira (2)**



Educação (2)



Saúde (2)



**Mercado de
Trabalho (1)**



Indústria (1)



Entretenimento (1)

- Vocês tem **dados** para aceitar ou refutar as hipóteses?



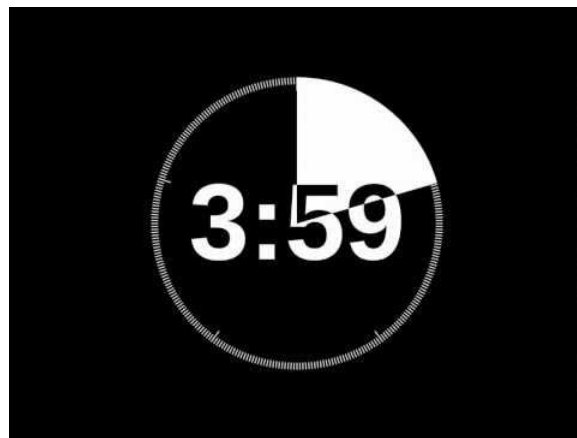
Tipos de Testes de Hipóteses

- Inferência estatística: **p-valor** de **t-test**
- Teste **Chi-quadrado**
- Teste **Tukey**
- Teste **Game Showell**
- Teste **ANOVA**
- Teste **F**



**ALGUMA DÚVIDA ATÉ
AQUI?**

INTERVALO 10 MIN



Aproveite para:

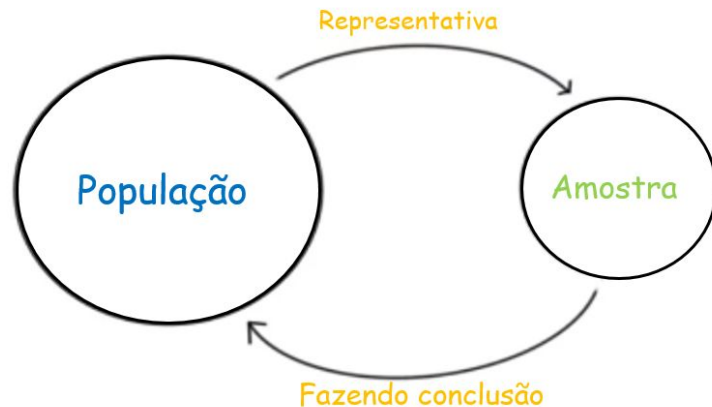
- Fazer anotações do que viu até agora (aprendizados, insights, dúvidas)
- Levantar-se, esticar os braços e as pernas
- Ir ao toalete
- Pegar uma água, um chá, um snack

A vertical bar with a gradient from green at the top to blue at the bottom.

População e Amostra

Definições

- **População** é o conjunto de **todos** os elementos ou resultados sob investigação
- **Amostra** é qualquer **subconjunto** da população
- A **amostra** pode ser definida dentro do universo como “**os indivíduos que responderam à pesquisa**”.
- É através desses indivíduos selecionados que iremos tirar conclusões válidas para **REPRESENTATIVIDADE**.



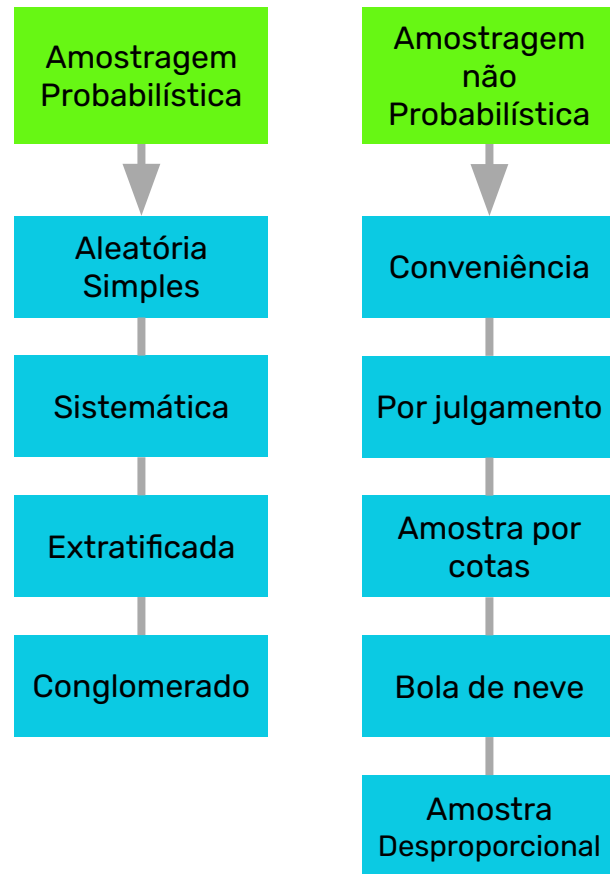
Parâmetros Estatísticos

- **População** é o conjunto de **todos** os elementos ou resultados sob investigação.
- **Amostra** é qualquer **subconjunto** da população.
- Uma **estatística** (θ) é uma característica da **amostra**.
- Um **parâmetro**, (T) é uma medida usada para descrever uma **característica** da população.

Média		
Variância		
Tamanho		

T Tipos de Amostragem

A amostra pode ser definida dentro do universo como “**os indivíduos que responderam à pesquisa**”. É através desses indivíduos selecionados que iremos tirar conclusões válidas para todo o grupo em estudo. O critério mais importante é o da **REPRESENTATIVIDADE**.



Amostragens Probabilísticas

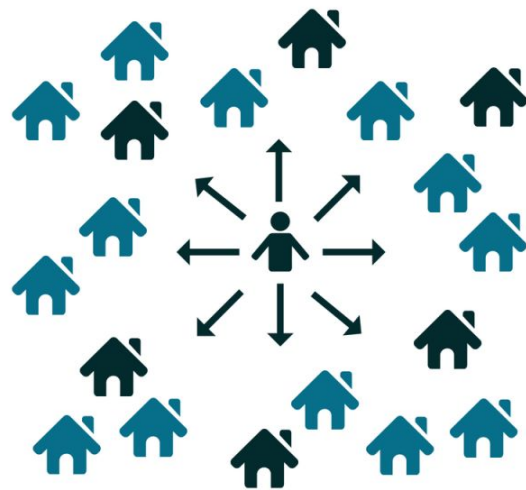
- Amostragem Aleatória Simples
- Amostragem Sistemática
- Amostragem Extratificada
- Amostragem por Conglomerado





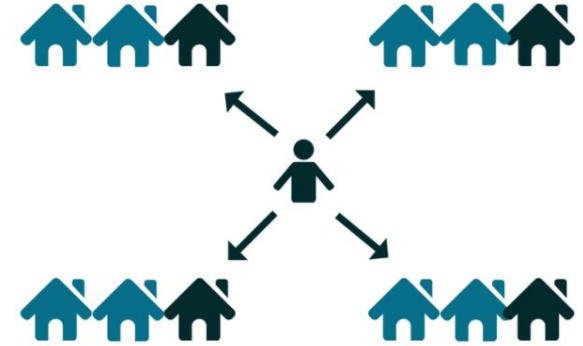
Aleatória Simples (Probabilística)

- Os indivíduos de uma população têm uma **chance igual** de serem selecionados
- a seleção de elementos é feita **em forma de sorteio**, dessa forma, não há critério ou filtro
- é um dos métodos de amostra probabilística **mais utilizados**
- **pode gerar qualquer combinação** de elementos presentes em um universo, isso pode ser bom ou ruim
- pesquisas que contêm um **universo muito grande**, se torna quase impossível obter uma listagem atualizada de todos



Amostragem Sistemática

- os elementos do universo a ser pesquisado são **divididos em grupos numericamente iguais**
- após essa segmentação é definido um “ponto de partida”, de modo a **estabelecer um número que se repetirá, em sequência**
- se decidirmos que o 4º domicílio de um grupo será o selecionado, vamos **obter uma sequência** (4, 14, 24, 34... até 494)
- o fator **representatividade** se torna mais efetivo
- se por acaso cada elemento dos grupos selecionados tenha características ou **opiniões que coincidam**, devemos **estudar previamente os grupos e separá-los com um peso proporcional**

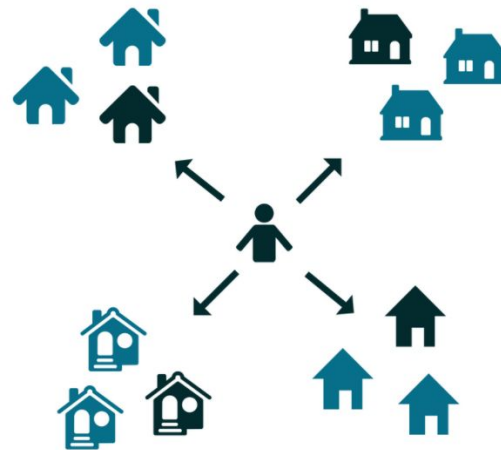


Amostragem Extratificada

- Separa a população em **grupos e subgrupos**, buscando assim, uma amostra mais representativa
- Primeiro, deve-se dividir a população em grupos distintos, que devem ser **segmentados com características da população que auxiliem o tema estudado**, podendo ser idade, sexo, trabalho, nível de escolaridade, entre outros
- Segundo, são utilizadas outras formas para eleger os entrevistados dentro de cada grupo, **podendo adotar critérios aleatórios ou não**
- Para selecionar uma amostra de forma **não enviesada** podemos utilizar a **amostra aleatória simples ou sistemática**
- a principal vantagem da amostra estratificada é o **aumento da representatividade** que ela gera por possibilitar uma estratificação

Amostragem por Conglomerados

- Primeiro, **os grupos (ou conglomerados) são definidos**
- Segundo, **os indivíduos** que participarão da entrevista **serão sorteados**
- O **número de etapas** varia com o tipo de pesquisa e o quanto o universo estudo deve ser dividido
- O número de etapas deve auxiliar o estudo (**quanto maior e mais heterogênea uma população, mais divisões se tornam necessárias**)
- **A maior vantagem é a relação custo benefício**



Amostragens não Probabilísticas

- Amostragem por Conveniência
- Amostragem por Julgamento
- Amostragem por Cotas
- Amostragem Bola de Neve
- Amostragem Desproporcional



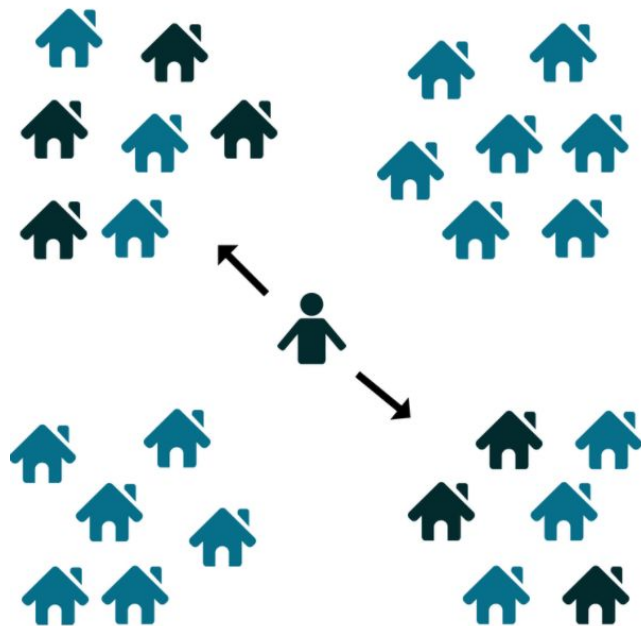


Amostragem por Conveniência

- **Não exige tanto critério na pré-seleção** do público a ser pesquisado, ou seja, o universo da pesquisa não precisa estar totalmente definido para que essa seja efetuada.
- Um **pesquisador que vá em um lugar público**, como uma praça, e lá mesmo faz a sua pesquisa, **sem muito critério de controle** com o perfil da amostra
- As **avaliações por cliente oculto** também são uma forma de pesquisa por conveniência
- A amostra por conveniência **pode gerar um resultado enviesado**
- Devem utilizadas em **pesquisas que buscam conclusões gerais**, com um perfil exploratório como o pré-teste de questionários, ou nas quais, não haja uma certeza prévia do perfil de seus respondentes.

Amostragem por Julgamento

- A escolha dos respondentes é feita partir do **juízo do pesquisador**
- O pesquisador busca por indivíduos que possuem **características definidas previamente** para sua amostra
- pessoas que têm comportamentos **que se encaixam às características pré-selecionadas**
- Tem uma **função mais exploratória** em uma pesquisa de opinião ou mercado
- Pode ser utilizada para pesquisas menores, ou como um **pré-pesquisa** para outras que buscarão dados mais aprofundados



Amostragem por Cotas

- Muito utilizada em **pesquisas de mercado, eleitorais e de opinião pública**
- Seleciona **proporcionalmente** pessoas com semelhantes características de uma população
- Primeiro segmentar o universo estudado em características, como por exemplo **dividir a população de uma cidade em cotas** como idade, sexo e escolaridade
- Segundo, os pesquisadores devem **escolher características**, nessa população, que sejam relevantes para a pesquisa
- A **precisão** dos resultados em amostra por cotas é dada a partir da **quantidade de cotas selecionadas para a pesquisa**

Amostragem Bola de Neve

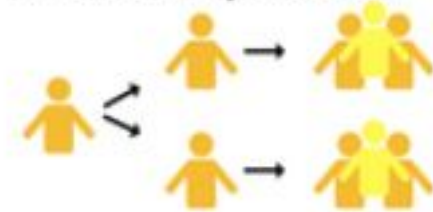
- A última pessoa entrevistada **indica ou convida uma próxima** para participar do questionário, fazendo com que a amostragem se comporte como uma bola de neve
- Boa técnica para **encontrar subgrupos ou segmentos de uma população** que são desconhecidos ou dificilmente encontrados

Amostra Linear



Amostra linear: cada indivíduo indica um único participante para a próxima pesquisa

Amostra Exponencial



Amostra exponencial: cada indivíduo indica 2 ou mais pessoas para a próxima entrevista

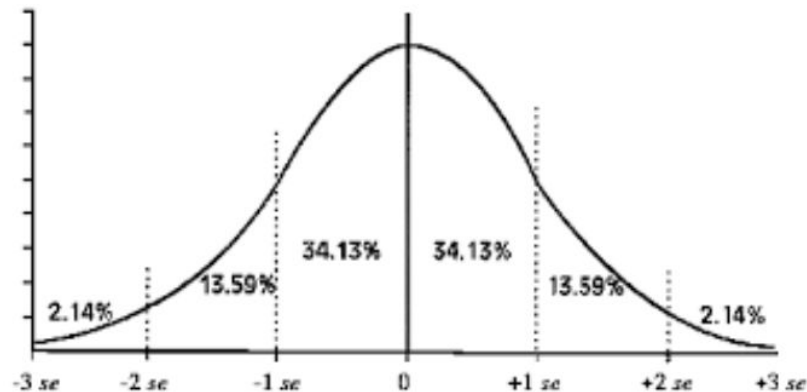
Amostragem Desproporcional

- É utilizada quando há grupos e subgrupos que geram **resultados com pesos dessemelhantes** em uma pesquisa
- Diferente da amostra por cotas, **não há a preocupação em ter uma exata proporcionalidade da população** estudada, o importante na amostra desproporcional é quanto um grupo dessa população é importante para o estudo
- **Grupos minoritários são priorizados** em um estudo, para que pequenos grupos obtenham um mínimo de representatividade nos resultados da pesquisa

Municípios	População da região em %	Nº de amostras proporcionais	Nº de amostras desproporcionais. (Reorganização)
Município 1	61%	610	580
Município 2	34%	340	320
Município 3	5%	50	100

Como definir uma Amostra

- margem de erro
- desvio-padrão
- nível de confiança





**ALGUMA DÚVIDA ATÉ
AQUI?**

A vertical bar with a gradient from green at the top to blue at the bottom.

Going Deeper: Teorema do Limite Central

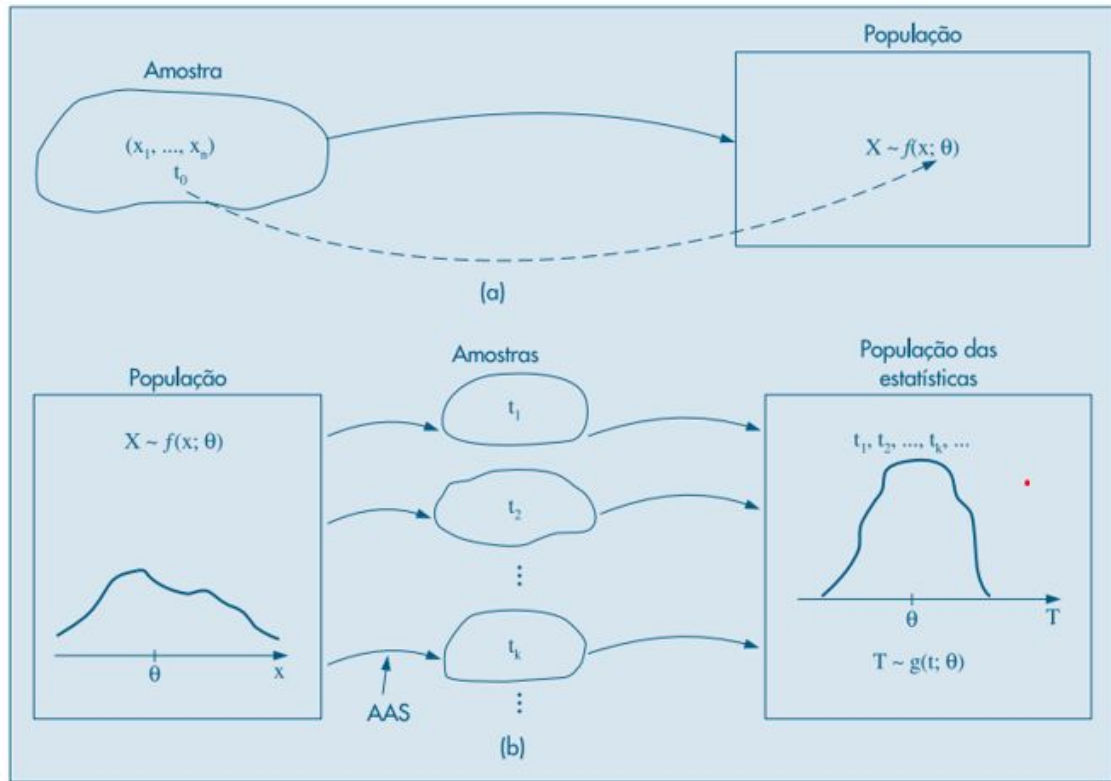


Teorema do Limite Central (TLC)

- O TLC afirma que, se você tem uma população com média, μ e desvio padrão σ e coleta amostras aleatórias suficientemente grandes da população com substituição, a **distribuição da média da amostra será distribuída aproximadamente normalmente**.
- Formalmente, O TLC afirma que **a população de todas as amostras** possíveis de tamanho n de uma população com μ e variância σ^2 **aproxima-se de uma distribuição normal** com μ e σ^2/n quando n se aproxima do infinito.

Teorema do Limite Central (TLC)

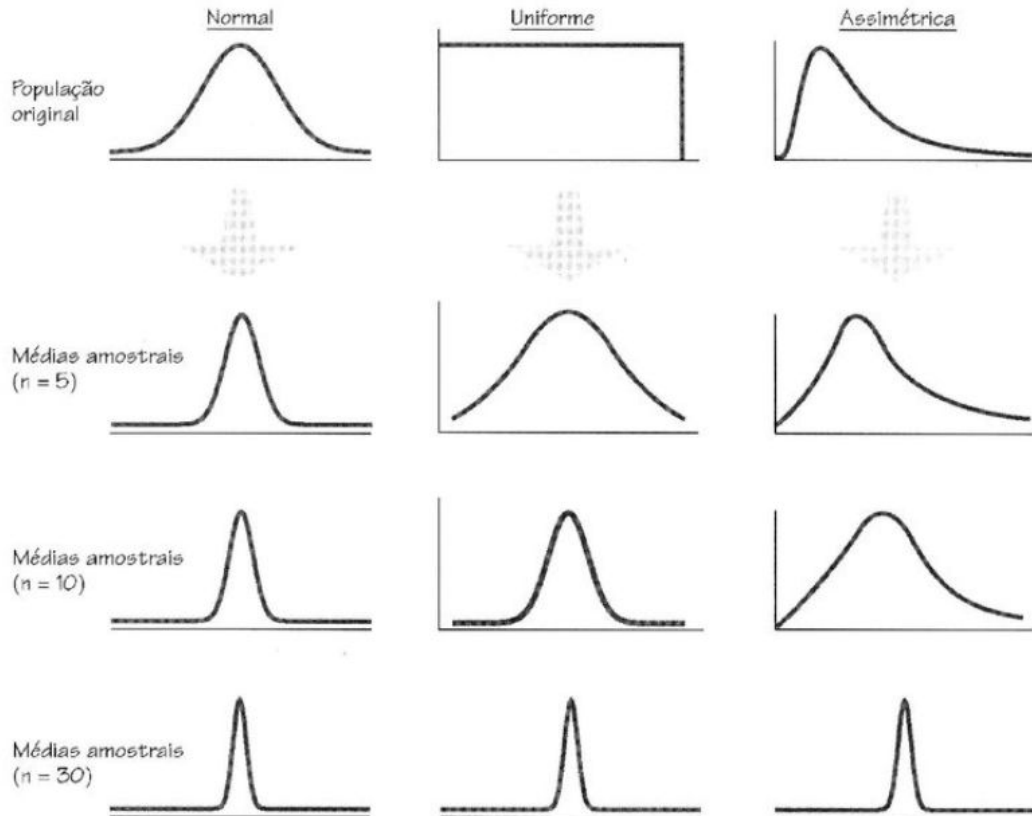
O Teorema do Limite Central nos permite realizar testes, resolver problemas e fazer inferências usando a distribuição normal, **mesmo quando a população não está normalmente distribuída**.



Teorema do Limite Central (TLC)


Quando maior o tamanho das amostras, a distribuição das médias será mais próxima de uma distribuição normal.

Para $n > 30$, a distribuição das médias amostrais pode ser aproximada satisfatoriamente por uma distribuição normal.





DÚVIDAS FINAIS



COMO FOI?

