# COMO SELECIONAR UMA AMOSTRA

## AMOSTRAGEM ALEATÓRIA SIMPLES

Essa é a maneira mais fácil para selecionarmos uma amostra probabilística de uma população. Nesse tipo de seleção de amostra cada indivíduo tem a mesma chance de ser sorteada para fazer parte da amostra. Por exemplo: Temos uma sala com 100 pessoas e quero uma amostra de n = 10, então escrevo o nome de todos, coloco numa caixa e sorteio 10 nomes.

Esses 10 nomes são nossa amostra aleatória.

## AMOSTRAGEM ESTRATIFICADA

Nesse método a população é dividida em subgrupos e então é feita uma amostragem aleatória simples em cada grupo. Por exemplo: pode-se dividir aquela sala do exemplo anterior em pessoas com mais de 30 anos e menos de 30 anos e então fazer uma amostragem aleatória simples em cada um dos grupo.

## AMOSTRAGEM POR CONGLOMERADOS

Tal qual a amostragem estratificada, a amostragem por conglomerado também divide a população em vários grupos, por exemplo bairros de uma cidade. Porém agora você irá sortear o grupo em que irá fazer os estudos. E essa será sua amostra.

## AMOSTRAGENS NÃO PROBABILÍSTICAS

Existem também amostragens que não são probabilísticas, ou seja, que não são obtidas de forma aleatória. Como por exemplo:

## AMOSTRAGEM POR CONVENIÊNCIA

Sua amostra são aqueles elementos que estão mais fáceis mesmo. Antes de soltar uma pesquisa de mercado você pode fazer um pré-teste com seus colegas de trabalho, por exemplo, para identificar um erro e ver se o teste já pode ser realizado com uma amostra maior.

quem vai participar. Por exemplo, fazer uma pesquisa musical mas só selecionar para entrevista as pessoas na rua que estiverem usando camiseta de alguma banda.

## E AGORA?

Agora, como podemos estimar os parâmetros de uma população através de amostras?

Vamos construir uma base teórica antes.

## AMOSTRA COM REPOSIÇÃO

Dentro das metodologias de amostras é possível também fazer essa seleção com reposição, por exemplo, numa sala com 100 pessoas, ao selecionar 10 pessoas para a minha amostra, uma mesma pessoa pode ser selecionada mais de uma vez.

Essa técnica é muito utilizada é ao fazer reamostragens (**Bootstrapping**).

Se temos apenas uma amostra da população, ou talvez uma amostra com n menor do que gostaríamos, podemos criar mais amostras à partir dessa que nós temos. Numa amostra de n=30, podemos fazer uma nova amostragem de n=30 porém com reposição, isto é, cada observação pode ser sorteada mais de uma vez. E podemos fazer isso diversas vezes e ter um conjunto de diferentes amostras.

---

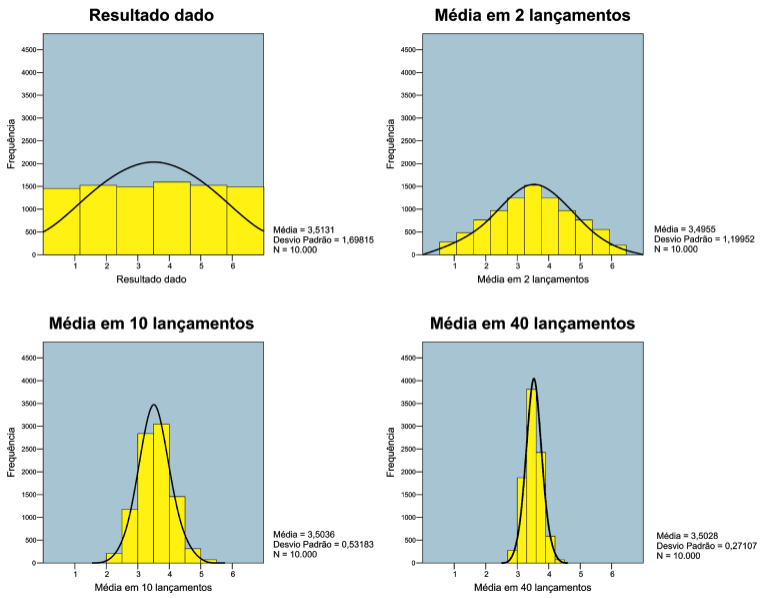
AVANÇAR

ESTATÍSTICA INFERENCIAL PARA DATA SCIENCE

# TEOREMA DO LIMITE CENTRAL

Muito da inferência estatística depende deste teorema que afirma que, independentemente de qual seja a distribuição original, a distribuição das médias da variável da amostra se aproxima de uma Normal. Quanto maior o *n* amostral, mais se aproximará de uma normal (aquela distribuição em formato de sino que você viu no material de Estatística Descritiva).

Veja na imagem abaixo que, mesmo para uma distribuição uniforme como o lançamento de um dado, na qual todas as faces têm igual probabilidade de sair, conforme aumentamos a amostra (número de lançamentos) a distribuição das médias de lançamento vai se aproximando de uma normal.




*Curiosidade: **Neste site** você pode brincar com os tipos de distribuição, tamanho e quantidade de amostras e ver como isso afeta a distribuição de médias das amostras.*

Assista esse vídeo em que Josh Starmer, host do canal StatQuest, faz uma explicação bem visual sobre o tema:

**https://www.youtube.com/watch?v=YAlJCEDH2uY**

# ESTATÍSTICA INFERENCIAL PARA DATA SCIENCE

ESTATÍSTICA INFERENCIAL PARA DATA SCIENCE

# INTERVALO E NÍVEL DE CONFIANÇA

Imagine novamente aquela sala com 100 pessoas. Agora eu preciso saber qual a média de altura dessas pessoas. Fiz uma amostra aleatória e a média da minha amostra é de 1,70.

É provável que a média do restante da população esteja de fato por volta desse número, mas não podemos afirmar baseado apenas nisso. É para acomodar essa incerteza que vamos entender como funciona o **Intervalo de confiança**.

Para estabelecer esse intervalo, precisamos saber qual será nosso **nível de confiança,** ou seja, qual a margem que daremos ao redor da média para que a gente possa considerar a média da nossa amostra igual a média populacional.

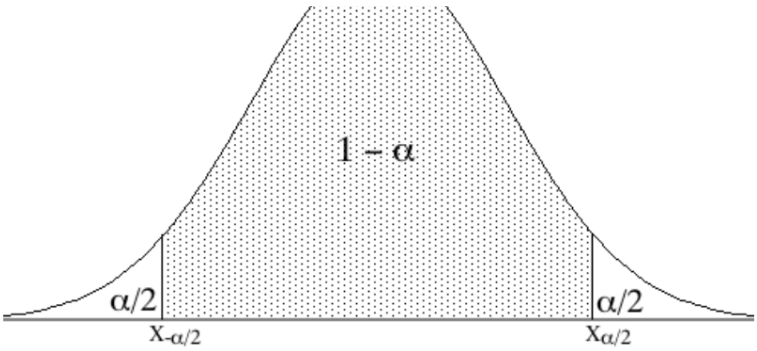
Esse nível vai variar muito de acordo com o problema de negócio que você estará resolvendo. Se estiver fazendo um estudo sobre um novo remédio, você vai querer estar bem seguro sobre os seus testes, podendo escolher um nível de confiança de 99% ou até maior. Caso seja só um experimento com 100 amigos numa sala imaginária, talvez 90% já seja o suficiente.

O mais comum e o que você vai encontrar na maioria das vezes é um nível de confiança de 95%.

**Nível de confiança** = 1 - $\alpha$

Agora veja abaixo como fica nosso nível de confiança dada uma distribuição normal. O intervalo representado pelos limites da área pontilhada representa o Intervalo de confiança.

ESTATÍSTICA INFERENCIAL PARA DATA SCIENCE




Uma maneira de entender esse gráfico é pensando que o parâmetro que eu estou estimando tem 95% de probabilidade (a área pontilhada) de estar dentro deste intervalo. Já não é uma estimativa pontual e sim intervalar.

Neste vídeo você verá como o que vimos até agora pode ser resumido na resolução de um problema (15 min):

**https://www.youtube.com/watch?v=ivXW8WLnzdU**)

---

AVANÇAR

ESTATÍSTICA INFERENCIAL PARA DATA SCIENCE

## FORMULAÇÃO DAS HIPÓTESES H0 E H1

Podemos pensar num cenário em que estamos fazendo um experimento, então como podemos verificar se os resultados que obtivemos está ou não de acordo com o que esperávamos encontrar? Em outras palavras, como que podemos comparar nossa com a da população que estamos estudando?

Para realizar o teste de hipótese, a primeira coisa que devemos fazer é colocar da maneira mais clara possível o que estamos testando. Sempre quebramos nosso teste em duas premissas:

O H0 também chamado de hipótese nula é o valor que é atualmente estabelecido para o parâmetro. No caso do exemplo anterior poderíamos dizer que nossa hipótese nula é que a média da população é igual a 15 km/l.

Para o H1 , ou hipótese alternativa, tem relação com a nova premissa que estamos testando. Neste caso podemos desafiar que este modelo de carro não tem média de 18 km/l. Assim sendo poderíamos formular nossas hipóteses da seguinte maneira:

**H0: $\mu = 15$ km/l**

**H1: $\mu \neq 18$ km/l**

Os possíveis resultados desse teste são:

- **Rejeitar H0**
- **Não rejeitar H0**

Leia atentamente os possíveis resultados. Veja que o contrário de Rejeitar H0 não é aceitar H0. Mas simplesmente **não rejeitar H0**.

> É VERDADE, E ASSIM DE TESTE EM TESTE, DE ESTUDO EM ESTUDO, TEMOS CADA VEZ MAIS UM ENTENDIMENTO MELHOR DO MUNDO.

## TESTE DE HIPÓTESES

Flashback do começo deste material: *"É isso que possibilita entrevistar algumas centenas de pessoas para fazer uma pesquisa eleitoral que representa milhões de eleitores ou fazer uma pesquisa com algumas dezenas de pessoas para verificar se um medicamento é seguro para uso humano e eficaz contra a doença que está sendo tratada."*

Agora que temos o embasamento teórico vamos ver a aplicação destes conceitos.

Teste de hipótese é um processo que usa estatísticas das amostras para testar uma afirmação sobre o valor de um parâmetro populacional. Por exemplo: Uma montadora afirma que seu carro faz em média 15 km por litro de gasolina. Como saber se esta afirmação é verdadeira ou não através da análise uma amostra?

---

AVANÇAR

# CONTINUANDO...

Neste capítulo veremos:

- **ESCOLHA DA DISTRIBUIÇÃO AMOSTRAL**
- **NÍVEL DE CONFIANÇA**
- **ERRO AMOSTRAL**
- **ÁREA DE ACEITAÇÃO E REJEIÇÃO**
- **CÁLCULO DA ESTATÍSTICA-TESTE**
- **CONHECENDO O P-VALOR**

## ESCOLA DA DISTRUIBÇÃO AMOSTRAL

Depois de feita a hipótese nula nós temos que escolher qual distribuição amostral. Lembra do Teorema do Limite Central? Conforme vamos tirando amostras da população, a médias das nossas amostras tende a formar uma curva normal.

Mas dependendo do tamanho do nosso n ou se conhecemos ou não o desvio padrão populacional, a distribuição amostral que iremos comparar será diferente. Como por exemplo a t de Student para uma amostra de n=20. Se nesse caso eu tivesse acesso à apenas 20 carros, eu deveria escolher a distribuição t de Student. Por outro lado, se eu pudesse coletar uma amostra de 200 carros, a distribuição escolhida seria a normal.

🔴 No final desta seção deixarei um link de um vídeo com um teste de hipótese feito de ponta a ponta para você ver como ele é realizado e seus cálculos.

## NÍVEL DE CONFIANÇA

Pode ser que você veja isso colocado de duas maneiras: Podem te pedir para fazer um teste com um nível de confiança de 95% ou um nível de significância de 5%. As duas coisas querem dizer a mesma coisa. Dentre 100

Imagine que você irá medir a média de km/l que aquele certo modelo de carro que estamos estudando faz. Para um teste com 95% de confiança, podemos assumir que, de cada 100 amostras que você irá coletar, pelo menos 95 estarão dentro desse intervalo que nós estamos considerando.

## ERRO AMOSTRAL

Lembre-se que estamos lidando com algum nível de incerteza, então é claro que às vezes a amostra não estará dentro daquele nosso intervalo de confiança. Quando confrontamos a realidade com o nosso teste chegamos no quadro abaixo.

| Decisão | Situação Real | |
|---|---|---|
| | **A hipótese $H_0$ é verdadeira** | **A hipótese $H_0$ é falsa** |
| **Rejeita-se $H_0$** | Erro do tipo I | sem erro |
| **Não se Rejeita-se $H_0$** | sem erro | Erro do tipo II |

O Erro tipo I é quando Rejeitamos H0 mas na realidade H0 era verdade. Por exemplo: a média amostral deu 5km/l, mas a média populacional estava próxima dos 15km/l e por acaso nossa amostra não refletiu isso.

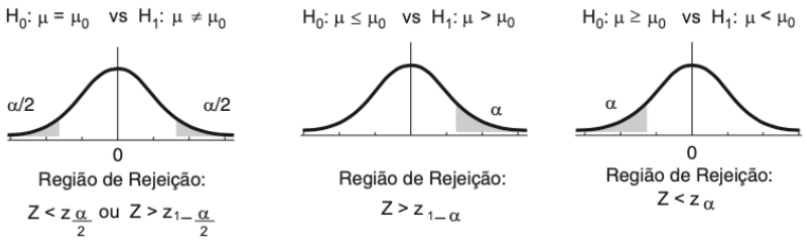
O Erro tipo II é quando Não rejeitamos H0 mas na realidade deveríamos ter rejeitado. Se nossa média amostral tivesse apontado algo próximo a 15km/l e, na realidade a média de todos os carros do modelo era de 5km/l e por acaso nossa amostra refletia nossa hipótese nula (H0).

A probabilidade de cometermos o erro do Tipo I é o nível de significância que acabamos de ver. São aqueles 5% que caso a hipótese nula fosse verdadeira, mas, por conta da amostra que selecionamos, acabamos por rejeitar H0.

## ÁREA DE ACEITAÇÃO E REJEIÇÃO

Abaixo estão os gráficos para as três possibilidades de região de aceitação e rejeição. Se estamos desafiando a média populacional e achamos que ela é **diferente** da nossa amostra, escolhemos o teste bi-caudal, ou seja, rejeitamos os valores extremos dos dois lados da distribuição.

Também podemos fazer testes para ver se a média da amostra é maior do que a populacional, então rejeitamos a área indicada mais a direita. E se



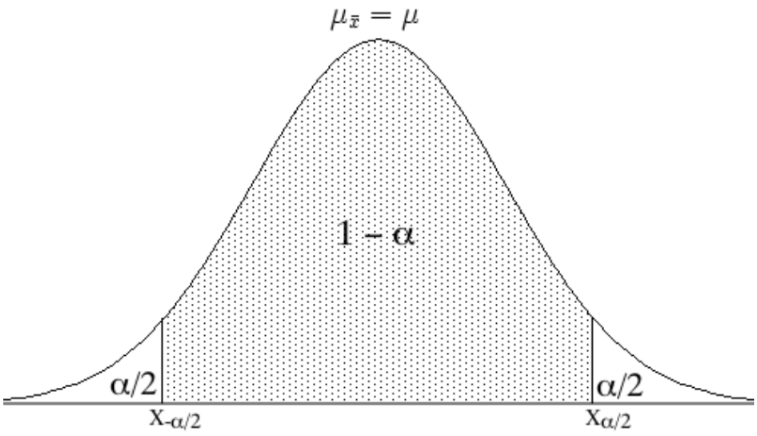

**Exemplos:**

Uma fábrica de shampoos precisa encher cada pote com 100ml de produto. O ideal é que não erre para menos para, os consumidores não serem prejudicados, e nem para mais, para que a empresa não tenha prejuízo. Nesse caso, faço um teste bi-caudal, tendo área de rejeição dos dois lados.

Outro caso: será que os cachorros bebem mais água que os gatos? Para isso vejo se a média da quantidade de água que os cachorros bebem esté dentro da área de rejeição à direita.

## CÁLCULO DA ESTATÍSTICA-TESTE

Dependendo da característica da sua distribuição você terá uma 'fórmula' diferente para fazer esse cálculo. Mas em essência você terá um número que te dirá onde na distribuição a sua média amostral caiu. Caso ela esteja dentro da área de rejeição (abaixo representada pela área em branco) você **Rejeita H0**. No caso do carro por exemplo, pode-se dizer que, pela amostra, aquele modelo de carro não faz uma média de 15 km/l. E caso esteja na área pontilhada você **Não rejeita H0.**




Os diferentes testes estatísticos te dão um número que é relacionado ao desvio-padrão da sua amostra ou da sua população e, com esse número,

Esses números geralmente não nos dizem muita coisa por si só, e é aí que entra o p-valor.

## CONHECENDO O P-VALOR

O p-valor pode ser descrito como a probabilidade de se obter uma estatística de teste igual ou mais extrema que aquela observada em uma amostra, sob a hipótese nula. Ou seja, se o nosso nível de significância é de 5%, então para p-valor menor que 5% deve-se rejeitar a hipótese nula.

**Em outras palavras quanto menor o p-valor maior a probabilidade de eu rejeitar a hipótese nula.**

Em geral, quando você se deparar com artigos científicos, autores vão se referir ao p-valor para descrever a conclusão de seus testes.

Existe até um termo chamado **p-hacking**, que é quando quem escreve manipula o cenário, ou faz diferentes testes até chegar num p-valor que comprove o que ele(a) está buscando, mais na base da manipulação e tentativa e erro do que de fato rigor científico.

Esse é um vídeo que faz um belo resumo sobre o assunto (11 min, em inglês):

**https://www.youtube.com/watch?v=Gx0fAjNHb1M**

## TUDO BEM ATÉ AQUI?

---

AVANÇAR

ESTATÍSTICA INFERENCIAL PARA DATA SCIENCE

# VAMOS VER EXEMPLOS?

A principal ideia deste material foi mais te apresentar os principais conceitos da Inferência Estatística e entender qual a intuição por trás desses conceitos do que realizar cálculos propriamente.

Por incrível que pareça é mais fácil achar conteúdo para aprender a fazer os cálculos do que para entender os conceitos de maneira mais intuitiva. Por isso esse material é tão relevante.

Mas a parte matemática também é muito importante para execução da Estatística Inferencial, por isso deixo abaixo alguns exemplos de testes de hipótese sendo feitos junto com seus cálculos.

Este vídeo é bem legal pois consegue apresentar um teste de hipótese enquanto vai te apresentando muito da intuição por trás também:

**https://www.youtube.com/watch?v=VK-rnA3-41c**

Neste outro vídeo, você vai ver as fórmulas que são usadas para fazer o cálculo do teste estatístico:

**https://www.youtube.com/watch?v=dluicq-hlm4**

---

AVANÇAR

ESTATÍSTICA INFERENCIAL PARA DATA SCIENCE

# MATERIAIS

Os materiais para aprofundamento são os mesmos de **Estatística Descritiva**, pois todos eles contém também a parte de **Estatística Inferencial**.

Mas aqui vai uma sugestão de uma série de vídeos do Crash Course sobre estatística. São Excelentes e muito visuais!

**Confira:**

**https://www.youtube.com/watch?v=zouPoc49xbk**

**Além de...**

Curso de estatística da PennState University - **Welcome to STAT 200! | STAT 200**

**StatQuest** | Diversos vídeos que explicam os conceitos estatísticos

## AULA AVALIADA COM SUCESSO!

Obrigada pelo seu feedback!

**REFAZER**

# ESTATÍSTICA INFERENCIAL PARA DATA SCIENCE

# ESTATÍSTICA INFERENCIAL

O que possibilita entrevistar algumas centenas de pessoas para fazer uma pesquisa eleitoral que representa milhões de eleitores ou fazer uma pesquisa com algumas dezenas de pessoas para verificar se um medicamento é seguro para uso humano e eficaz contra a doença que está sendo tratada?

A Estatística Inferencial permite que você faça inferências, tire conclusões de uma **população** à partir de uma **amostra** de dados aleatórias provenientes desta população.

Quase sempre, seria muito caro (para não dizer impossível) ter acesso a informação sobre toda a população que você quer estudar e é por isso que a estatística inferencial é muito útil.

O principal objetivo desse material vai ser apresentar os conceitos mais importantes do tema enquanto construímos uma boa base intuitiva, por isso não entraremos na matemática e nas fórmulas. No final, nós deixaremos alguns links caso queira se aprofundar no tema. :)

---

AVANÇAR