Foundations

Noções de Inferência Estatística

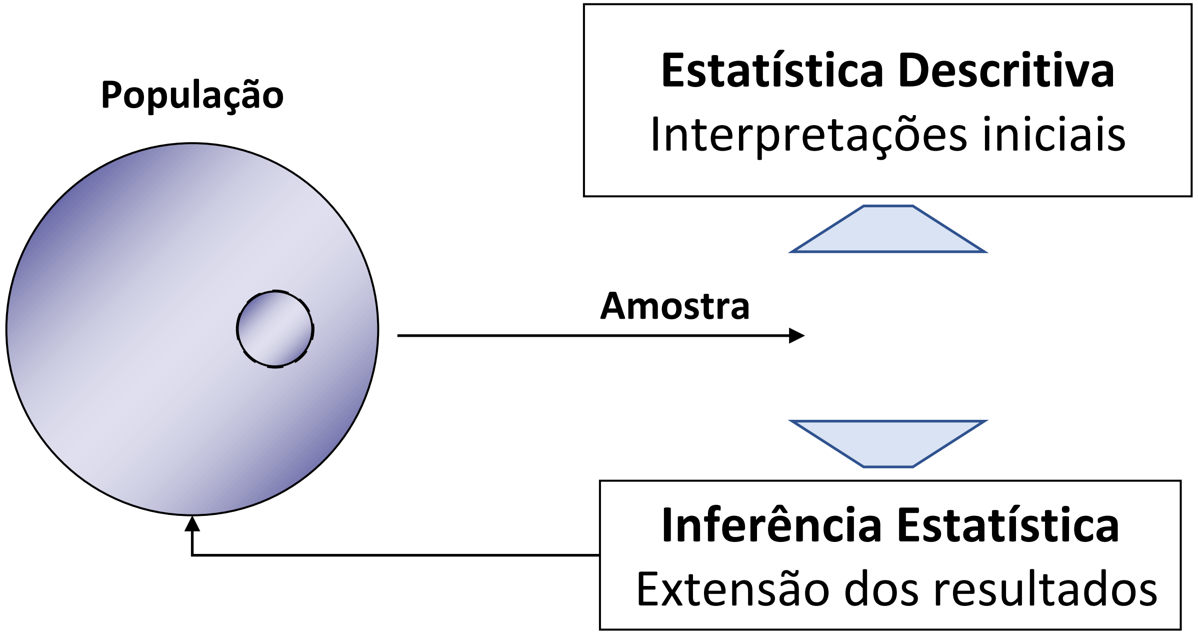
Noções de Inferência Estatística

O que é?

Preditiva.ai

A **Inferência** é o conjunto de técnicas que possibilitam a “extensão”, a um grande conjunto de dados, das informações e conclusões obtidas a partir de uma pequena quantidade de dados, ou seja, de uma **amostra**.

Assim, podemos dizer, que é a técnica a ser utilizada quando não temos acesso aos dados completos, ou seja, quando não temos acesso à **população** de interesse.



Noções de Inferência Estatística

O que é?

Em quais situações não temos acesso à população de interesse? Vejamos alguns exemplos:

1. **Problemas econômicos:**
   - É muito custoso entrevistar todas as pessoas de interesse. Ex.: Habitantes do Brasil.
   - Nem todas as pessoas estão acessíveis. Ex.: Clientes de um restaurante as 19h.

2. **Problemas físicos:**
   - Imagine o problema de se testar a durabilidade das lâmpadas. O teste poderia queimar todas elas.

3. **Problemas éticos:**
   - Testes de remédios em seres vivos pode causar grandes efeitos colaterais. Testar em toda a população seria prudente?

Noções de Inferência Estatística
Conceitos Fundamentais

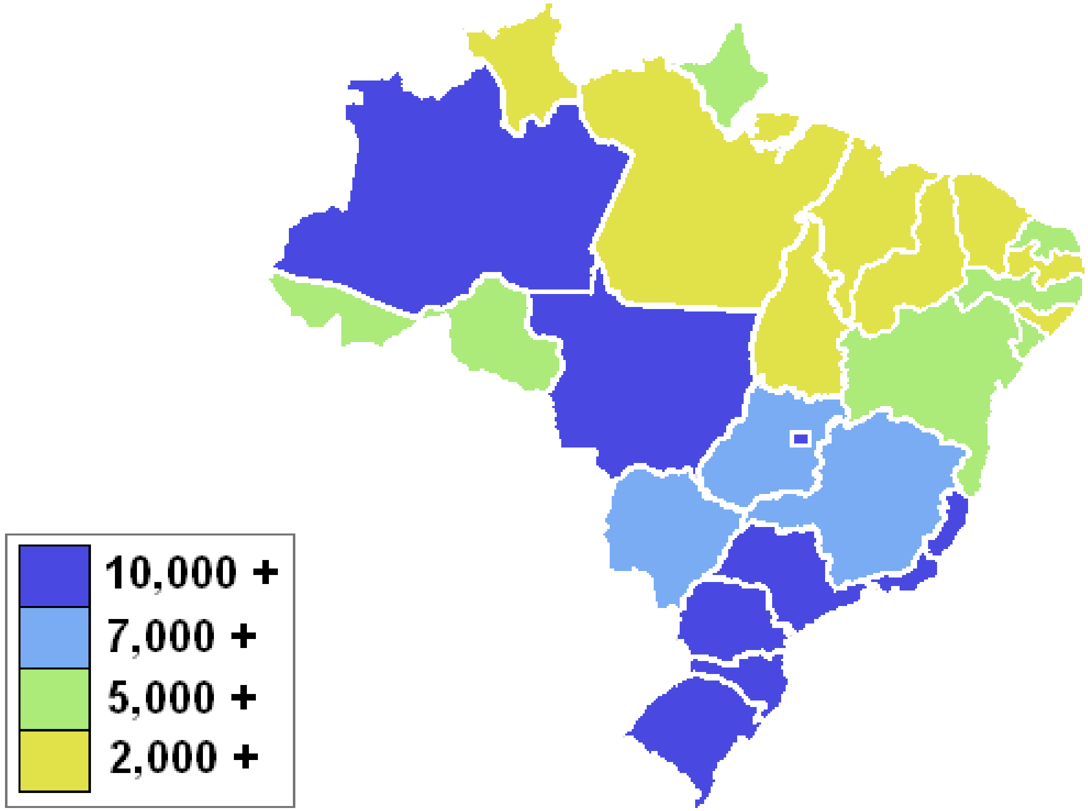
Agora que sabemos que extrair uma amostra é uma alternativa interessante para o problema de acesso à população, podemos nos perguntar: **Qualquer amostra representa bem uma população?**

Vejamos em um exemplo:

Suponha que queremos **estudar a renda per capita do Brasil**. Para isso, entrevistamos as pessoas de São Paulo e perguntamos sua renda anual.

Pergunta:

Essa amostra representa bem a população ?



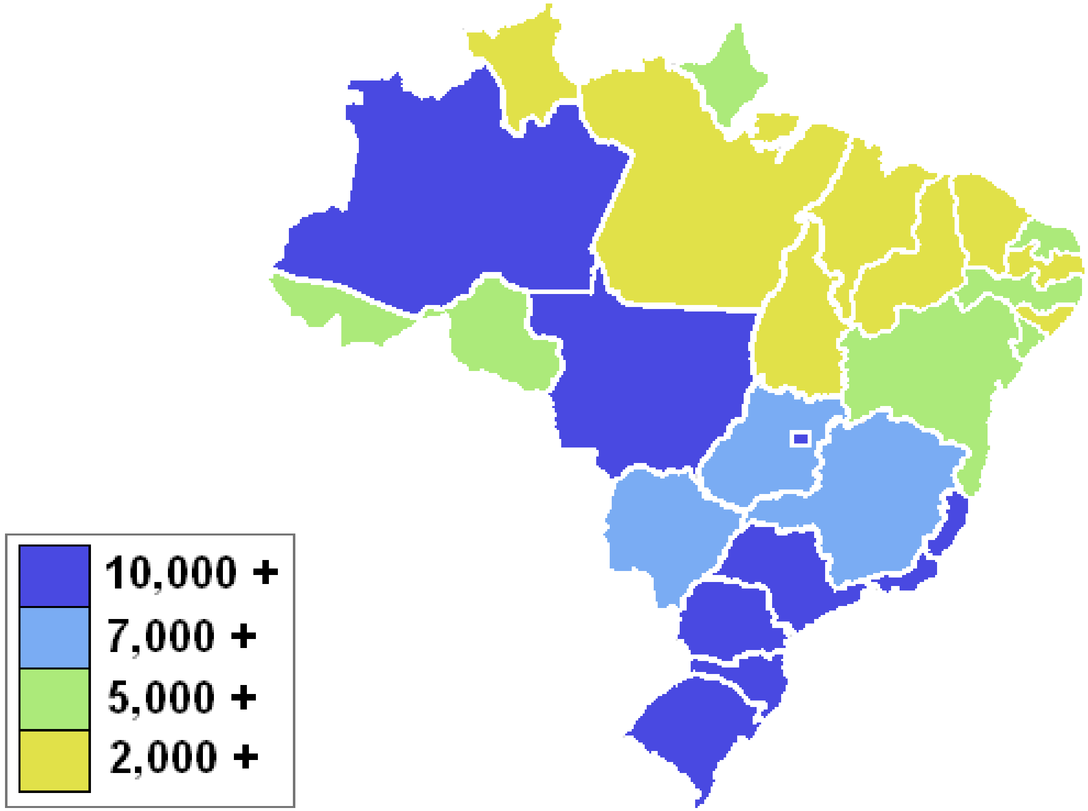
Fonte: https://pt.wikipedia.org/wiki/Ficheiro:Brazilian_States_by_GDP_per_capita.PNG

Noções de Inferência Estatística
Conceitos Fundamentais

Preditiva.ai

O problema dessa abordagem é que a renda per capita **não** é **homogênea** no país. Considerar apenas um estado brasileiro causa um **viés de seleção**.

Para tentar resolver, o pesquisador ligou para cada um de seus amigos que **moram em cada estado do país**. Com isso, perguntou a renda anual de seus conhecidos.



Pergunta:

Essa amostra representa bem a população ?

Fonte: https://pt.wikipedia.org/wiki/Ficheiro:Brazilian_States_by_GDP_per_capita.PNG

Noções de Inferência Estatística
Conceitos Fundamentais

O problema dessa outra abordagem é que, ao entrevistar apenas seus amigos, o pesquisador selecionou uma **amostra por conveniência**. Com isso ele pode ter causado um outro **viés de seleção**.

Uma abordagem mais satisfatória seria sortear aleatoriamente um conjunto de pessoas do Brasil, independente do estado. Essa técnica de amostragem é chamada **Amostragem Aleatória Simples (AAS).**

A **Amostragem Aleatória Simples (AAS)** tem como principal benefício a garantia de obter amostras que tenham **a mesma probabilidade de ocorrência**. Com isso, é mais seguro extrapolar suas conclusões para a população de interesse. Temos basicamente dois tipos de AAS. São eles:

1. **AAS sem reposição:** Quando temos que remover a unidade sorteada para não correr o risco de ser sorteada novamente na próxima coleta. Ex: Bolas em uma urna. Ao sortear uma bola, devemos removê-la da urna.

2. **AAS com reposição:** Quando não removemos a unidade sorteada. Ex: Ao sortear uma bola, devemos devolve-la para a urna, de modo a garantir a possiblidade que ela seja sorteada novamente.

Noções de Inferência Estatística

Conceitos Fundamentais

Preditiva.ai

Voltando para o estudo da **renda per capita do Brasil**, quais **outros problemas poderiam prejudicar a qualidade da amostra** do estudo?

Tamanho da amostra

Segundo o IBGE*, o Brasil tem mais de **208 milhões de habitantes com diversas características**. Com isso, é de se esperar que a amostra tenha um tamanho adequado.

Por mais que tenhamos o cuidado de ser aleatória, **não é razoável** entrevistar apenas um ou dois habitantes de cada estado brasileiro.

Sabendo disso, **qual o tamanho de amostra adequado?** Iremos estudar esse assunto mais adiante no curso.

Erro de leitura/coleta

Muito comum de acontecer, os erros de coleta são difíceis de resolver. Se o entrevistador anotar valores errados durante a entrevista ou ainda não ter cuidado com a forma de realizar a pergunta, pode **aumentar muito a imprecisão do estudo**. Vejamos um exemplo:

Um entrevistador pergunta o que o cliente acha de uma operadora de celular. Para isso, realiza a seguinte pergunta: **Por que você acha que a operadora X é melhor que a operadora Y ?**

* https://g1.globo.com/economia/noticia/2018/08/29/brasil-tem-mais-de-208-milhoes-de-habitantes-segundo-o-ibge.ghtml

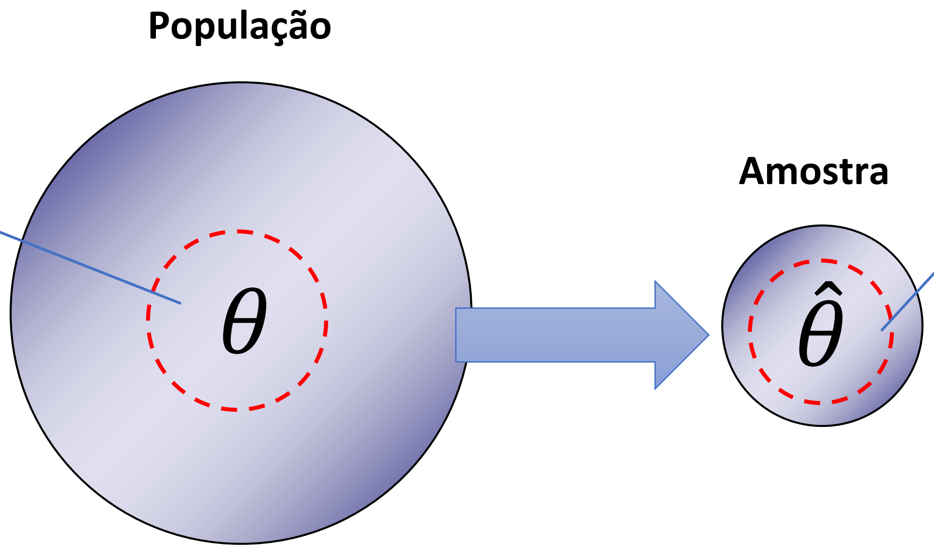
Noções de Inferência Estatística

Conceitos Fundamentais

Agora que vimos os principais cuidados a serem tomados na escolha de uma boa amostra, **o que vem a seguir para continuar o processo de Inferência**? A escolha de um bom **estimador**. Vejamos algumas definições:

A população de interesse guarda o **parâmetro de interesse**, geralmente denotado por letras gregas θ, μ , σ etc.

No estudo anterior, a **população** dos habitantes do Brasil têm o **parâmetro de interesse** "Renda Per Capita".



Usando os valores da amostra, escolhemos um **estimador** para o parâmetro de interesse. Este estimador geralmente é denotado por letras gregas com "chapéu". Ex: $\hat{\theta}$, $\hat{\mu}$, $\hat{\sigma}$.

Através do estimador, calculamos uma **estimativa** para o parâmetro de interesse.

Noções de Inferência Estatística

Conceitos Fundamentais

São muitos os estimadores de algum parâmetro populacional. A questão que devemos responder é: **Qual o melhor estimador que podemos escolher?** Essa escolha depende de 3 conceitos fundamentais:

Viés	O viés de um estimador é a diferença entre sua média e o parâmetro de interesse. Ou seja, um estimador é dito **não-viesado** quando sua média coincide com seu parâmetro populacional.
Variância	O estimador que deve ser escolhido é aquele que tem a menor variância possível. Quando um estimador tem variância mínima, dizemos que ele é **preciso**.
Consistência	Além do estimador ser não-viesado e de variância mínima, queremos que o estimador seja consistente. Um estimador é dito **consistente**, quando: $\lim_{n\to\infty} Média(\hat{\theta}) = Parâmetro$ e $\lim_{n\to\infty} Variância(\hat{\theta}) = 0$ Ou seja, a medida que o tamanho da amostra aumenta, o estimador perde o viés e sua variância fica próxima de 0 (zero).

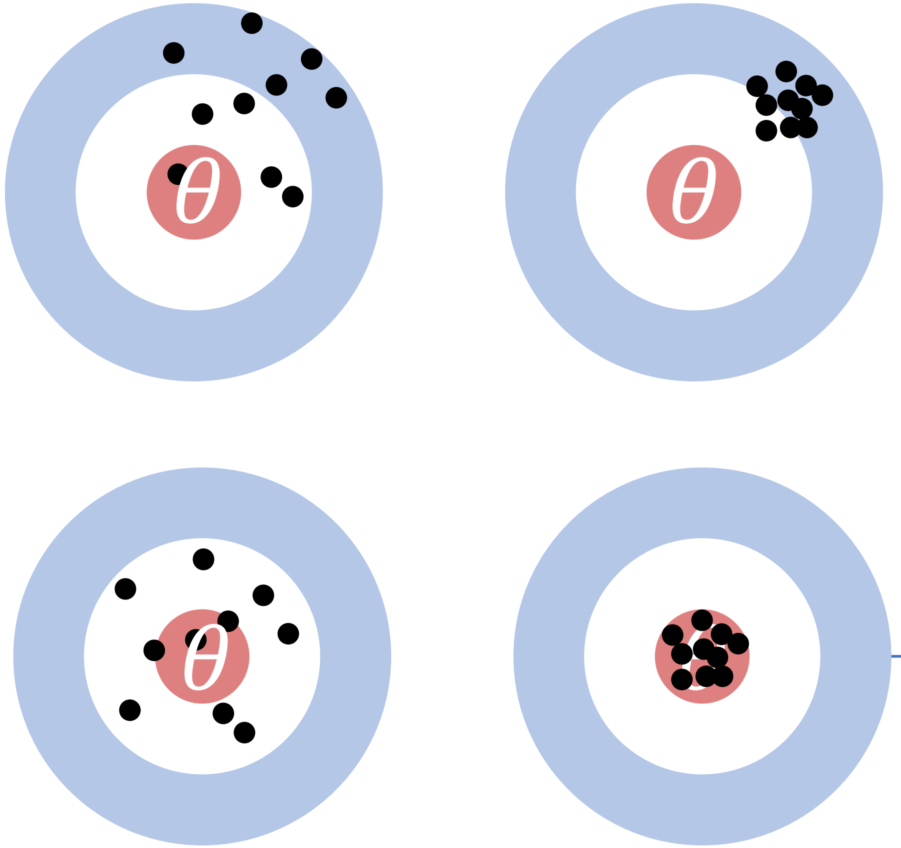
Noções de Inferência Estatística

Conceitos Fundamentais

Preditiva.ai

Vejamos a seguir exemplos dos tipos de estimadores. **Qual o melhor estimador**?

	Alta Variância	**Baixa Variância**
Viesado		
Não Viesado		



Este estimador **parece ser o melhor**, pois é não-viesado e tem a menor variância possível.

Noções de Inferência Estatística

Conceitos Fundamentais

Preditiva.ai

Quadro **resumo dos estimadores** para os principais parâmetros de interesse:

Parâmetro de interesse	Estimador	Propriedades
μ (Média populacional)	$\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n}$	Não-viesado e consistente
p (Proporção populacional)	$\hat{p} = \frac{frequência\ amostral}{n}$	Não-viesado e consistente
σ^2 (Variância populacional)	$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{X})^2$	Não-viesado e consistente

Noções de Inferência Estatística

Conceitos Fundamentais

Portanto, para um bom processo de Inferência Estatística, **precisamos tomar os seguintes cuidados**:

- ❑ Escolher uma amostra que seja **uma boa representante da população** de interesse. Para isso, devemos:
  - Sortear a amostra aleatoriamente ou com o menor viés de seleção possível
  - Calcular um tamanho de amostra adequado
  - Tomar cuidado com erros de escrita/leitura dos dados
  - Mensurar os dados sem influenciar as respostas (Ex: perguntas tendenciosas)

- ❑ Escolher **um bom estimador para o parâmetro** de interesse

Como vimos nesta aula, **são muitos os fatores que levam a erros metodológicos** no processo de Inferência Estatística.

Quem nunca se perguntou **como os institutos de pesquisa erram** ao tentar estimar parâmetros de interesse como as intenções de votos em um país.

Sem dúvida alguma, boa parte dos problemas devem-se aos erros explicados nesta aula.

Desta forma, lembre-se sempre: **Ao ouvir falar sobre algum estudo que usa o processo de Inferência Estatística, sempre verifique as premissas utilizadas no estudo. Só assim você pode confiar (ou não) nos resultados apresentados.**

Revisão

Nesta aula aprendemos que **Inferência Estatística** é a área da **Estatística** responsável pelas técnicas que nos permitem extrapolar os resultados de uma amostra para toda a população.

Vimos também que devemos tomar diversos cuidados na **seleção de uma amostra da população**, pois **vieses de seleção**, **tamanho inadequado**, **parcialidade** na realização de perguntas e **erros de leitura / escrita** dos dados gerarão **resultados distorcidos**.

E por último, que além de tomar todos esses cuidados com a amostra, devemos também utilizar **estimadores com boas propriedades** de **viés, variância e consistência.**

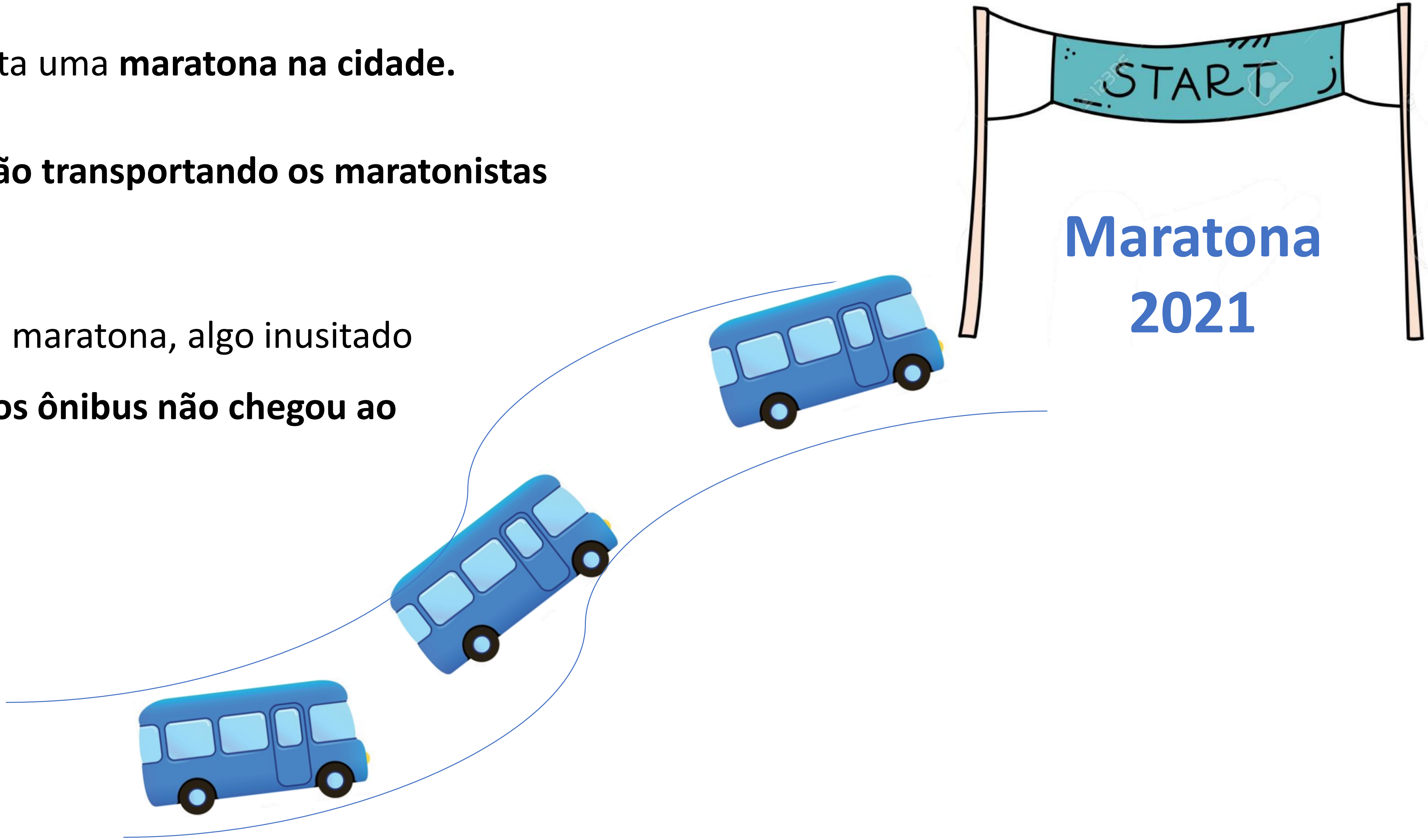
Noções de Inferência Estatística
Teorema do Limite Central

Noções de Inferência Estatística

Teorema do Limite Central

Preditiva.ai

- Suponha que exista uma **maratona na cidade.**
- Vários **ônibus estão transportando os maratonistas** para o evento.
- Antes do início da maratona, algo inusitado aconteceu: **Um dos ônibus não chegou ao destino...**



Noções de Inferência Estatística

Teorema do Limite Central

- A polícia foi acionada para buscar os desaparecidos.

- Após algumas horas de busca, **um ônibus apareceu na estrada.**

- Você pede para que os passageiros saiam do ônibus para serem identificados. Eis os passageiros:

Noções de Inferência Estatística

Teorema do Limite Central

Preditiva.ai

Rapidamente você **estima que a idade média** dos passageiros do ônibus é de **8 anos.**

Com base nessa estimativa, pergunta-se:

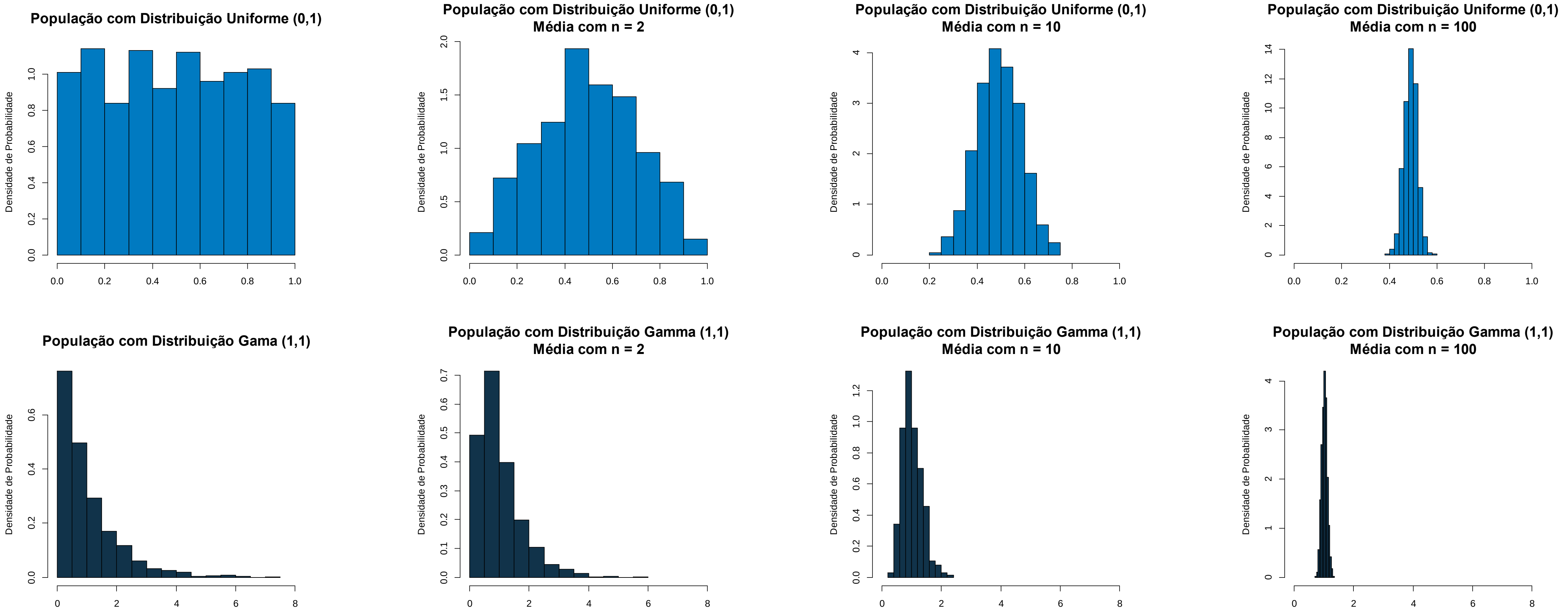
Esse é o ônibus perdido que você estava procurando?

Se você acha que é **muito pouco provável que um ônibus (a amostra)** de passageiros tão jovens sejam maratonistas, parabéns! Você entendeu a ideia principal do **Teorema do Limite Central (TLC)**.

O princípio essencial do TLC é que uma amostra grande, adequadamente escolhida, será muito próxima da população que foi retirada.

Noções de Inferência Estatística
Teorema do Limite Central

Verificação do **Teorema do Limite Central** utilizando simulações:



Noções de Inferência Estatística

Teorema do Limite Central

Preditiva.ai

Além disso, o TLC nos diz que:

1. Se você obter amostras grandes, aleatórias, de qualquer população, as médias dessas amostras serão distribuídas normalmente em torno da média da população.

2. A maioria das médias de amostras estará razoavelmente perto da média da população. O desvio padrão é o que define “razoavelmente perto”.

3. Quanto menos provável for um resultado observado, mais confiantes podemos estar em presumir que a população de origem é bem diferente da amostra que estamos analisando.

Noções de Inferência Estatística

Teorema do Limite Central

O **Teorema do Limite Central** é muito importante porque ao caracterizar a distribuição de probabilidades da **média amostral,** podemos:

- Calcular uma estimativa da **média populacional**
- Calcular o desvio padrão da estimativa da **média populacional**
- Calcular percentis para avaliar a estimativa da **média populacional**

Revisão

Nesta seção aprendemos:

- O que é o **Teorema do Limite Central** e qual sua importância.
- Como utilizar o resultado do **Teorema do Limite Central** para caracterizar a distribuição da média amostral.
- Calcular alguns percentis para analisar a distribuição da média amostral.

Próximos passos

Na próxima seção aprenderemos como fazer **estimativas pontuais e por intervalo**, ou seja, além da estimativa, conheceremos uma forma de calcular a **incerteza** sobre essa própria estimativa.

Conhecer as **incertezas das estimativas é algo extremamente útil** no mundo de Analytics! Essa informação é fundamental na análise dos dados e deve sempre ser considerada na tomada de decisões.

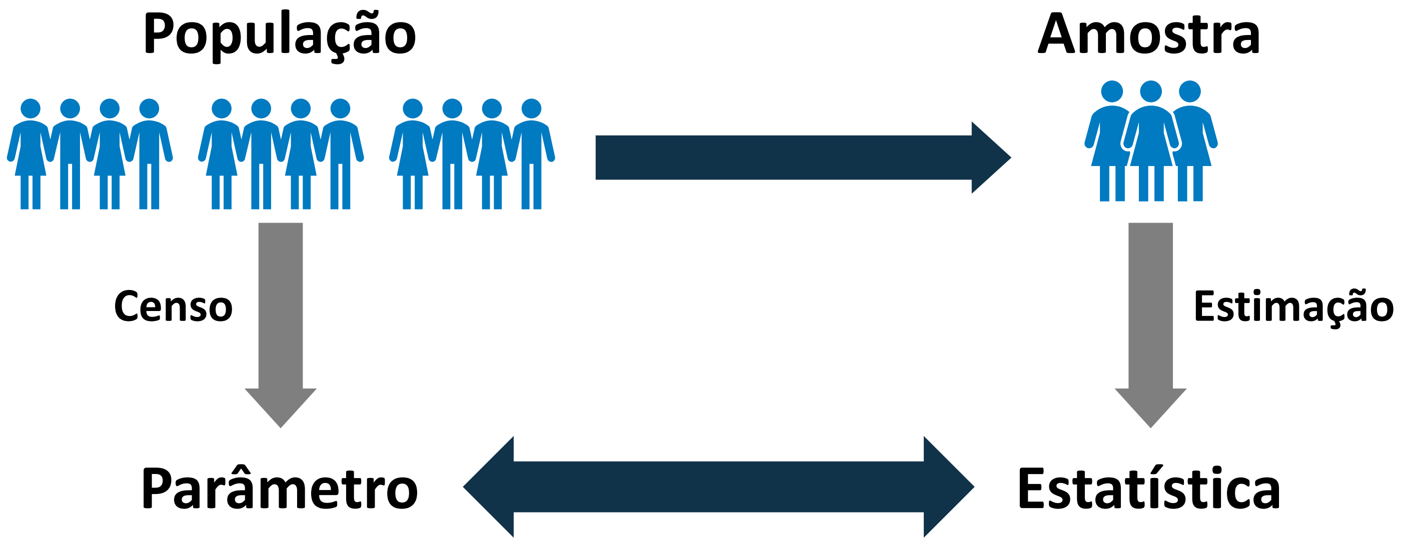
Inferência Estatística
Estimação Pontual e por Intervalo

- Média Populacional

Noções de Inferência Estatística

Estimação Pontual

Como vimos anteriormente, a **Inferência Estatística** busca **generalizar** informações sobre uma **População** a partir de dados de uma **Amostra**. Essa **generalização** é obtida através de **métodos de estimação dos parâmetros** de uma população.



Noções de Inferência Estatística
Estimação Pontual

Vamos iniciar esse tema com o seguinte exemplo:

Considere que um fabricante de automóveis deseja saber **a média de idade** dos seus clientes que compraram seu último modelo. Para isso, realizou uma pesquisa com **30 pessoas** em todo o Brasil e calculou a média de idade em **31** anos.

Considerando que a **amostra representa bem a população**, vimos que um estimador não-viesado e de baixa variância para a **média populacional** de idade é:

$$Média\ amostral = \hat{\mu} = 31$$

Algumas perguntas que podemos fazer são:

- Essa é uma boa estimativa?
- Qual é o tamanhdo **do erro** na estimação da idade média?

Noções de Inferência Estatística

Estimação por Intervalo

Para quantificarmos com maior precisão o **erro na estimação da média populacional**, utilizamos os **Intervalos de Confiança**. Vamos calculá-lo para a **média populacional** μ com **90% de confiança**.

$$IC\ (\mu; 90\%) = \left(\hat{\mu} - t_{n-1} \cdot \sqrt{\frac{S^2}{n}};\ \hat{\mu} + t_{n-1} \cdot \sqrt{\frac{S^2}{n}}\right)$$

$\hat{\mu}$: média amostral

S^2: variância amostral

n: número de observações

t_{n-1}: valor da distribuição t-Student

Noções de Inferência Estatística

Estimação por Intervalo

Preditiva.ai

Além dos dados informados anteriormente, o fabricante calculou também a **variância amostral** e obteve o valor de **25**. Dessa forma, temos todos os dados para o cálculo do **Intervalo de Confiança**:

$\hat{\mu}$: média amostral = 31

S^2: variância amostral = 25

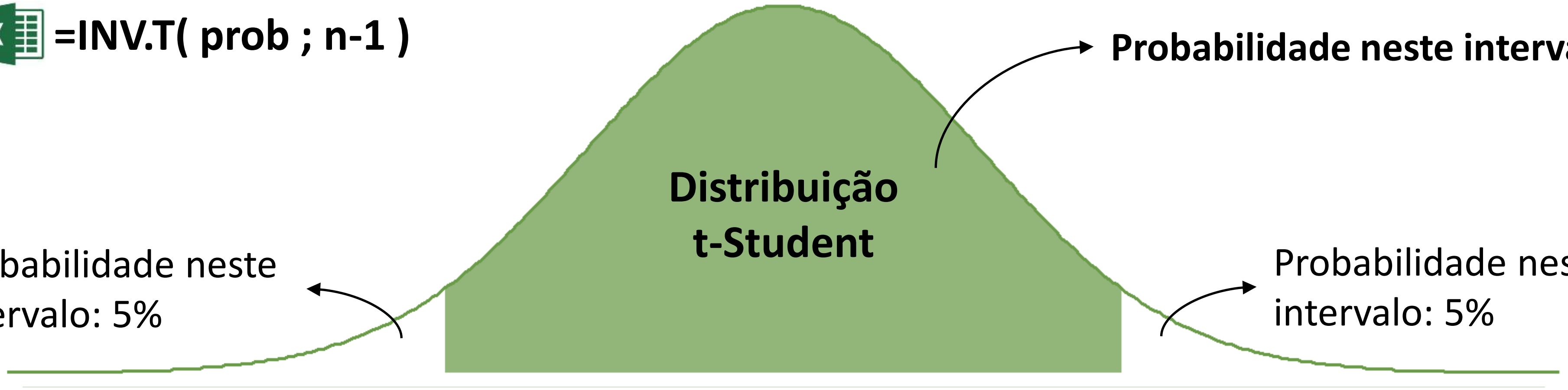
n: número de observações = 30

t_{n-1}: valor da distribuição t-Student = 1,70

=INV.T(prob ; n-1)

$$IC\,(\mu; 90\%) = (\underbrace{31 - 1,70 \cdot \sqrt{\frac{25}{30}}}_{29,4}; \underbrace{31 + 1,70 \cdot \sqrt{\frac{25}{30}}}_{32,6})$$

Probabilidade neste intervalo: 90%



Probabilidade neste intervalo: 5%

Probabilidade neste intervalo: 5%

INV.T(0,05 ; 29) → **-1,70**

0

+1,70 ← **INV.T(0,95 ; 29)**

Noções de Inferência Estatística

Estimação por Intervalo

Preditiva.ai

Podemos dizer então que o **Intervalo de Confiança** para a média populacional (idade dos clientes da empresa), com um **Nível de Confiança (γ) de 90%** é:

$$\text{IC}\ (\mu\ ;\ 90\%) = (\ 29{,}4\ ;\ 32{,}6\ )$$

Pergunta: Como devemos interpretar o **Intervalo de Confiança** acima?

a) Existe 90% de probabilidade da média de idade na população estar em 29,4 e 32,6 ?

b) A real média da população irá cair entre 29,4 e 32,6 em 90% das vezes.

c) O intervalo de confiança de 29,4 e 32,6 pode ser um dos 90 intervalos calculados a cada 100 intervalos diferentes que pode conter a real média de idade da população.

Noções de Inferência Estatística

Estimação por Intervalo

Preditiva.ai

Podemos dizer então que o **Intervalo de Confiança** para a média populacional (idade dos clientes da empresa), com um **Nível de Confiança (γ) de 90%** é:

$$\textbf{IC} \ (\mu \ ; 90\%) = (\ 29{,}4 \ ; \ 32{,}6 \ )$$

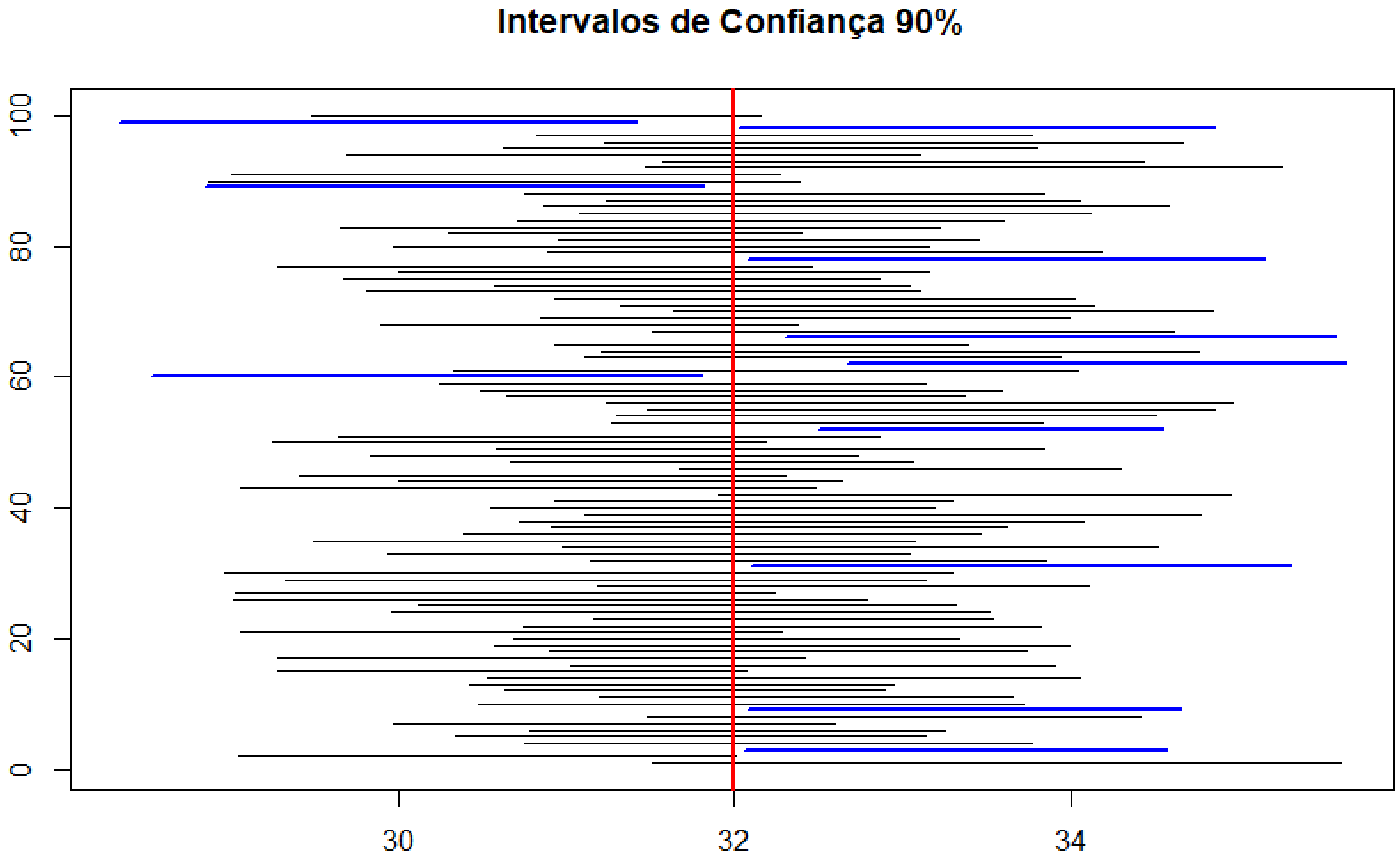
O **Intervalo de Confiança** deve ser interpretado como:

"Se realizássemos um **grande número de amostras aleatórias de tamanho n** e calculássemos o **Intervalo de Confiança** para todas elas, **90% desses intervalos conteriam a real média da população (parâmetro μ)** "

Noções de Inferência Estatística

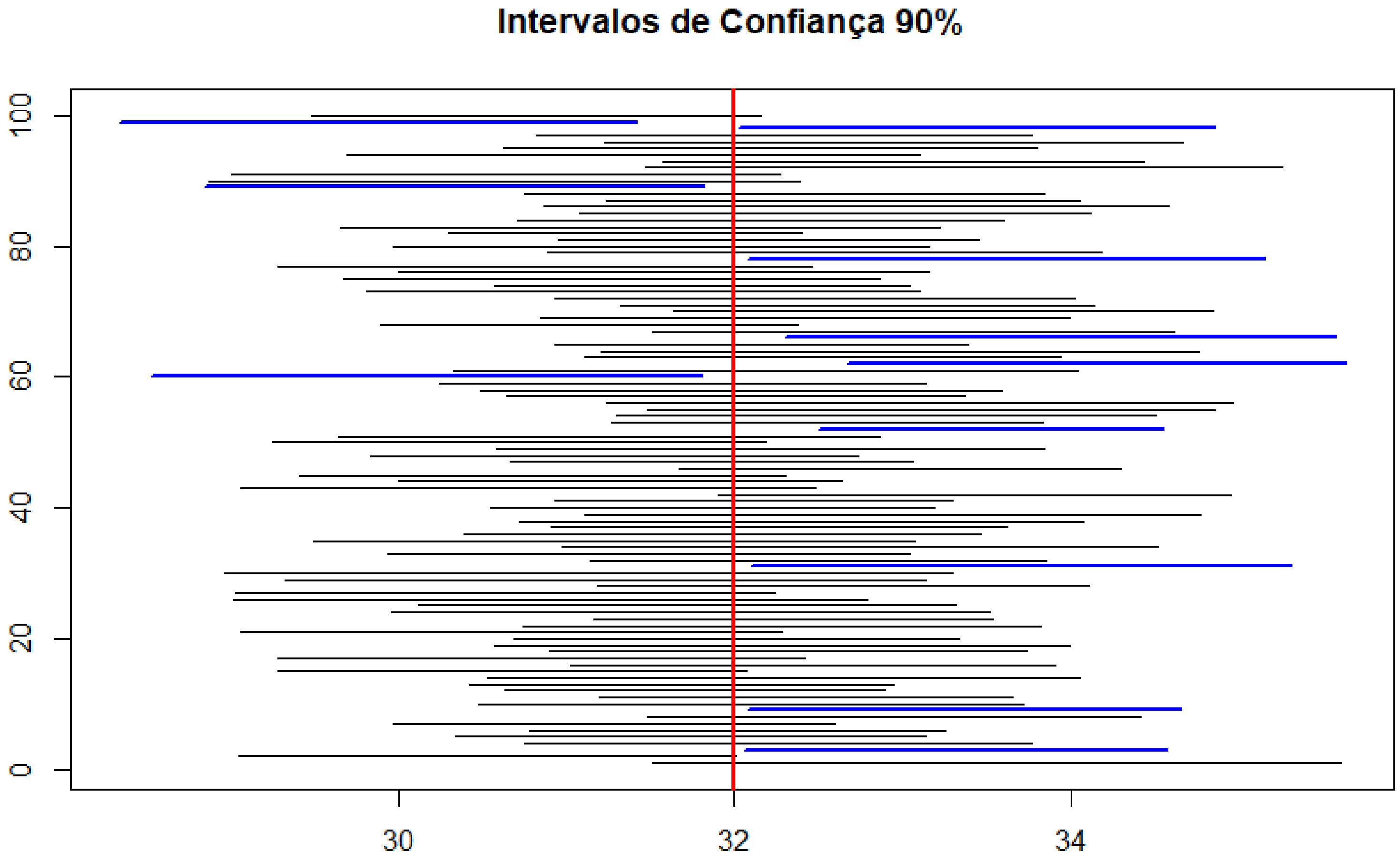
Estimação por Intervalo

Calculando 100 **Intervalos de Confiança** utilizando os parâmetros deste exemplo, temos o gráfico abaixo:



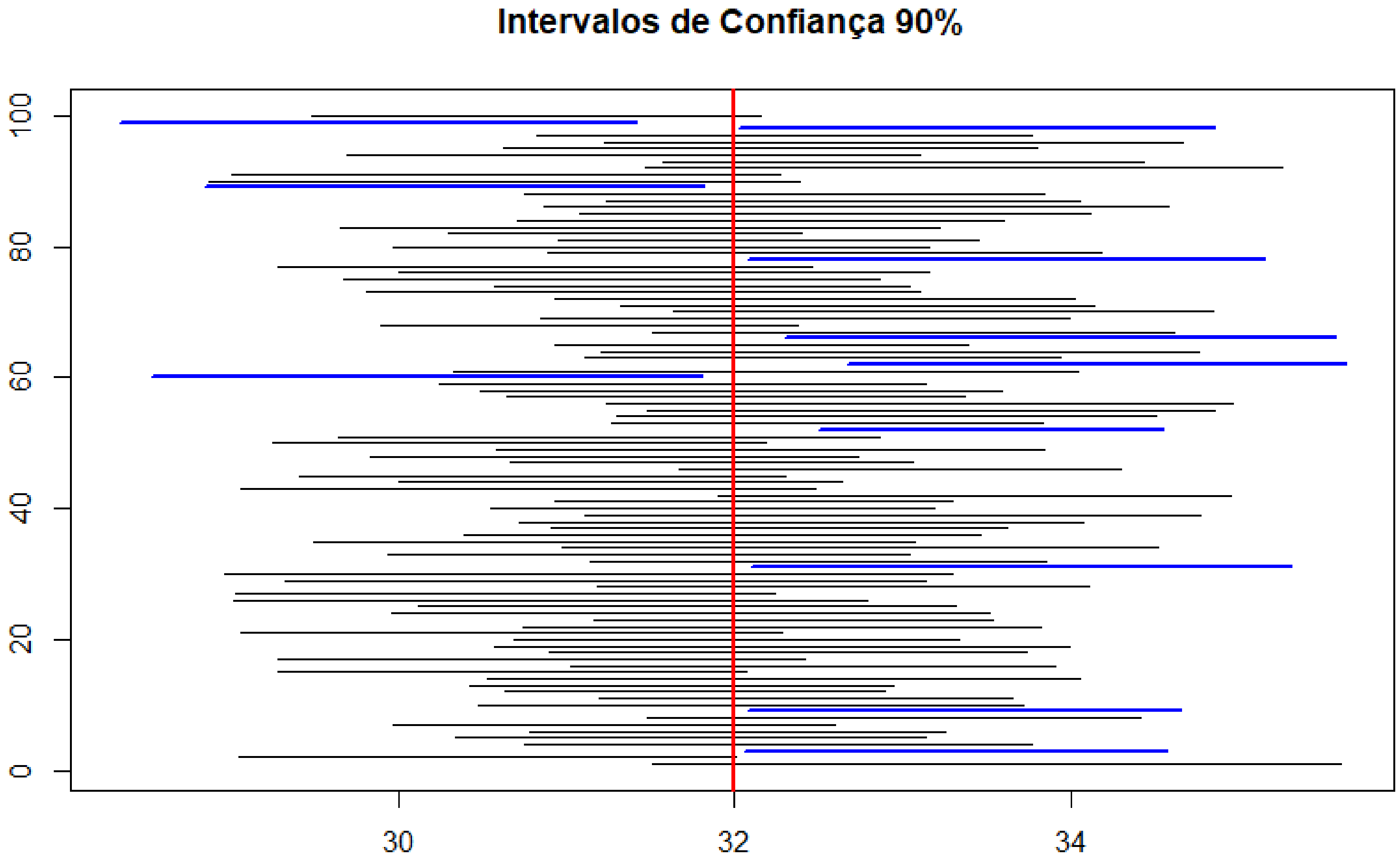
Noções de Inferência Estatística
Estimação por Intervalo

Supondo que a idade populacional seja 32. Perceba que as linhas destacadas em azul essa idade não está contida no **Intervalo de Confiança.**



Noções de Inferência Estatística
Estimação por Intervalo

Podemos perceber que dos 100 **Intervalos de Confiança** calculados, 89 contém o parâmetro populacional (89%).



Noções de Inferência Estatística

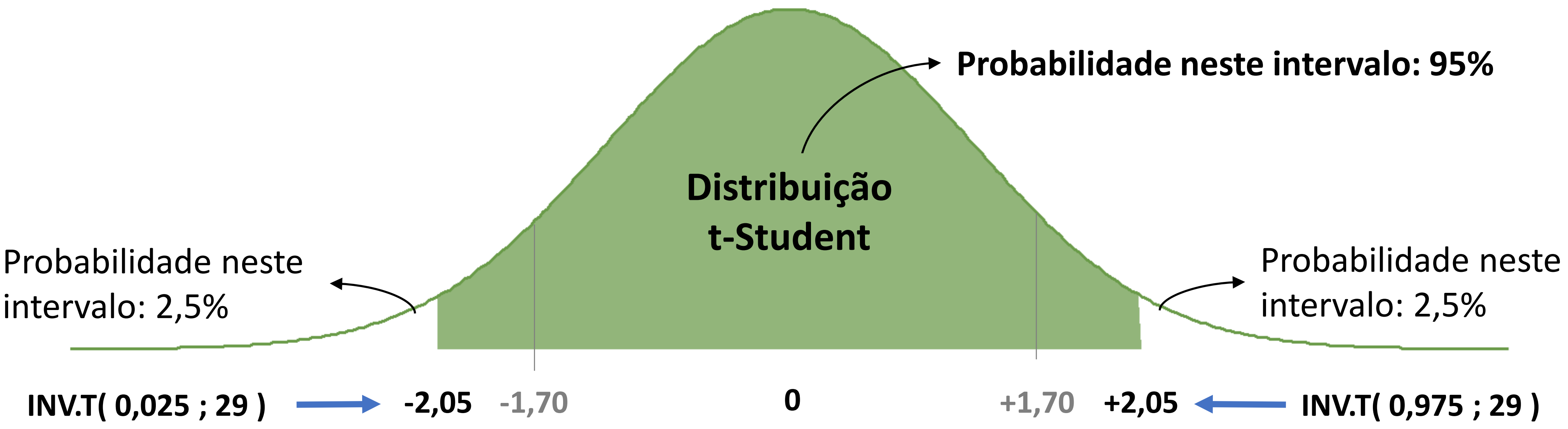
Estimação por Intervalo

Preditiva.ai

Se aumentarmos o **Nível de Confiança** para **95%**, o Intervalo de Confiança será:

- INV.T(0,025 ; 30-1) = -2,05
- INV.T(0,975 ; 30-1) = +2,05

IC (μ ; 95%) = (29,1 ; 32,9)



Noções de Inferência Estatística

Estimação por Intervalo

Dessa forma, com **30 entrevistados**, o fabricante de automóveis calcula que com **90% de confiança**, a média de idade dos compradores de todos os veículos do último modelo está entre **29,4** e **32,6**:

$$IC\ (\mu\ ;\ 90\%) = (\ 29{,}4\ ;\ 32{,}6\ )$$

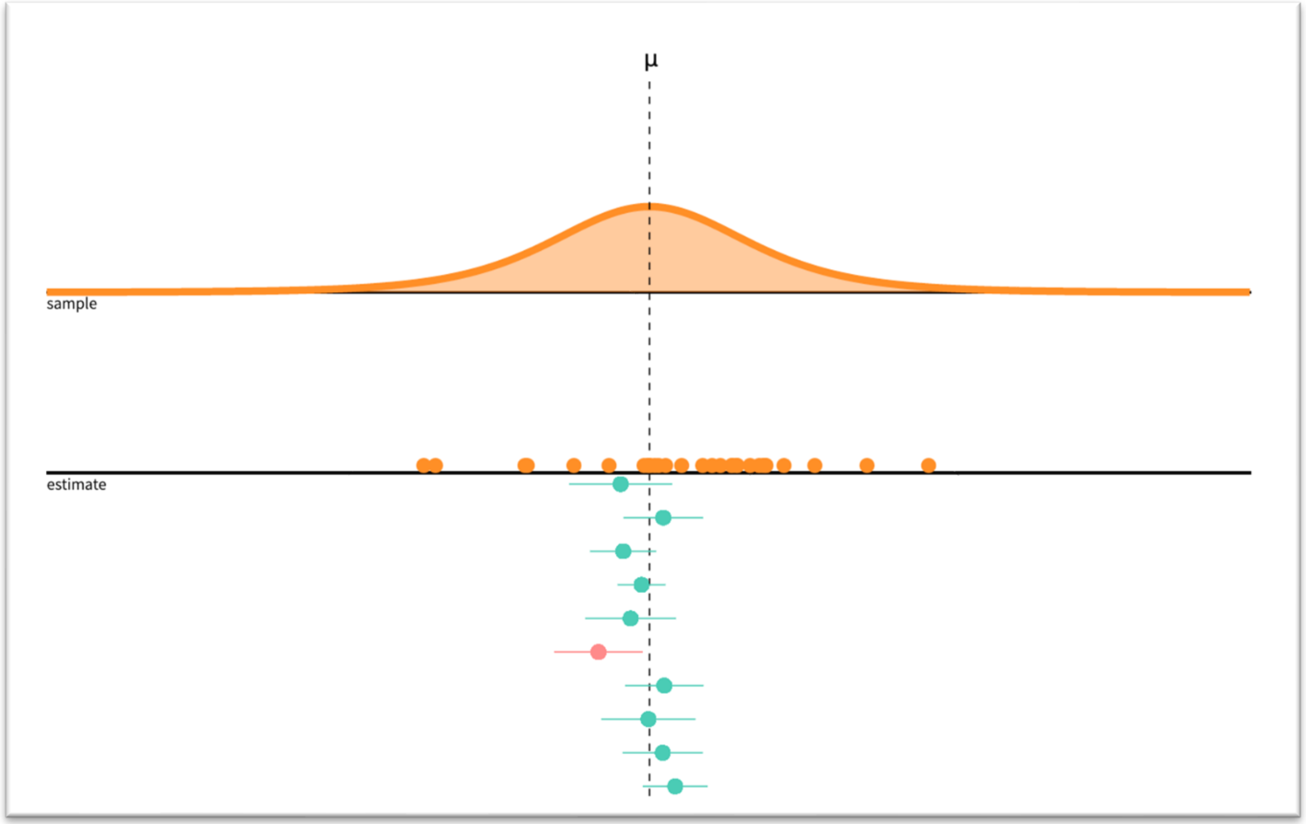
E com **95% de confiança**, a média de idade dos compradores está entre **29,1 e 32,9:**

$$IC\ (\mu\ ;\ 95\%) = (\ 29{,}1\ ;\ 32{,}9\ )$$

Noções de Inferência Estatística

Estimação por Intervalo

Exemplo visual:



https://seeing-theory.brown.edu/frequentist-inference/index.html#section3

Inferência Estatística

Estimação Pontual e por Intervalo

- Proporção Populacional

Preditiva.ai

Noções de Inferência Estatística
Estimação por Intervalo

Em problemas relacionados com **percentual** ou **proporção populacional** (em vez da média como vimos nos exemplos), procedemos da mesma forma, exceto pelas seguintes mudanças:

$$\boldsymbol{IC\ (p;90\%) = \left[\hat{p} - t_{n-1} \cdot \sqrt{\frac{\sigma^2_{\hat{p}}}{n}}\ ;\hat{p} + t_{n-1} \cdot \sqrt{\frac{\sigma^2_{\hat{p}}}{n}}\right]}$$

Intervalo de confiança para a proporção populacional

Sendo,

$$\hat{p} = \frac{\text{Frequência do evento de interesse}}{\text{número total de eventos}}$$

Proporção amostral

$$\sigma^2_{\hat{p}} = \hat{p}(1 - \hat{p})$$

Variância da proporção amostral

Preditiva.ai

Inferência Estatística

Margem de Erro e Tamanho na Amostra

Noções de Inferência Estatística

Margem de Erro

Preditiva.ai

Vimos anteriormente que a **Estimação Pontual e por Intervalo** visa **extrapolar** os resultados obtidos de uma **amostra** para a **população**.

Aprendemos que por utilizar uma **amostra**, temos um **erro intrínseco ao cálculo dos estimadores**. E por esse motivo, é importante calcular além da estimação pontual, a estimação por intervalo, os chamados **Intervalos de Confiança**.

Vamos entender agora como podemos **controlar o erro nas estimações variando o tamanho da amostra**.

Noções de Inferência Estatística

Margem de Erro

Para apresentar o conceito de **Margem de Erro**, vamos relembrar o exemplo da seção anterior.

Com **30 entrevistados**, o fabricante de automóveis calculou que a média de idade dos compradores de veículo foi de **31**, e com **90% de confiança**, o percentual populacional estaria entre **29,4 e 32,6.**

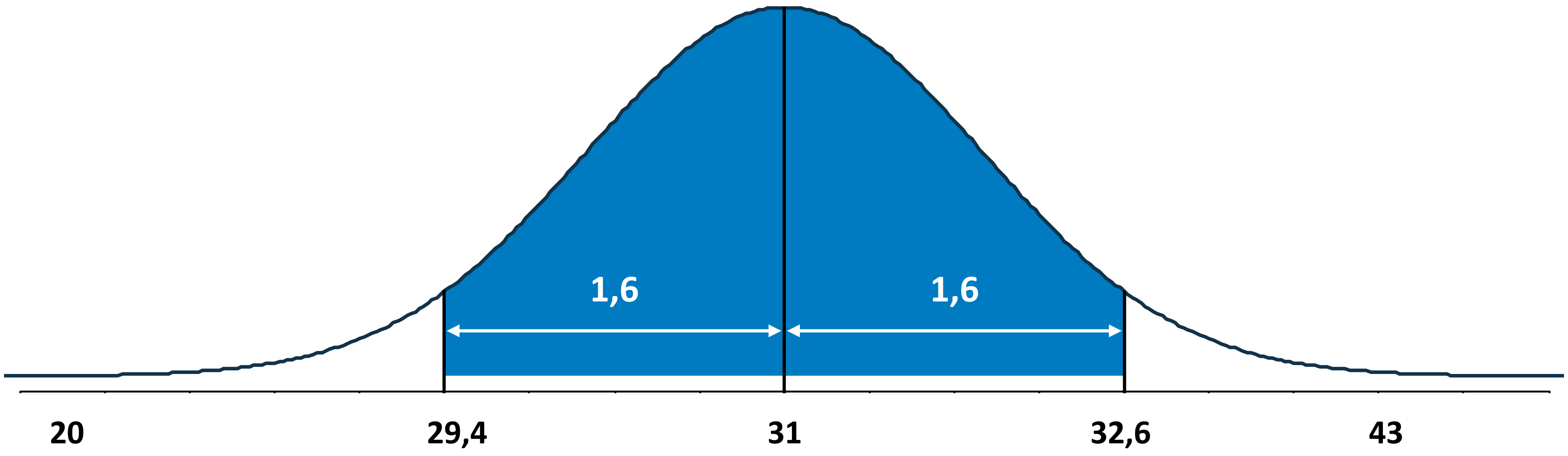
$$IC(\mu\ ;\ 90\%) = (\ 29{,}4\ ;\ 32{,}6\ )$$

Noções de Inferência Estatística

Margem de Erro

Nesse contexto, dizemos que a **Margem de Erro** na estimação do parâmetro μ com **Nível de Confiança de 90%** é de **1.6 ,** pois essa é a distância entre a **média amostral** e os **limites do Intervalo de Confiança**.

Distribuição da Média Amostral



Noções de Inferência Estatística

Cálculo do Tamanho da Amostra

Preditiva.ai

Digamos que o fabricante de automóveis não tenha ficado muito satisfeito com a **Margem de Erro** dessa estimação. Como podemos fazer para reduzir essa incerteza na estimação da média de idade dos compradores do último modelo?

"Calculando um **Tamanho de Amostra** compatível com a **Margem de Erro desejada pelo fabricante**"

Noções de Inferência Estatística

Cálculo do Tamanho da Amostra

Para calcular o **Tamanho da Amostra** utilizamos a equação a seguir:

$$n = \frac{S^2 z_\gamma^2}{\varepsilon^2}$$

Onde:

- **n**: tamanho da amostra que teremos após o cálculo
- S^2: variância amostral
- z_γ: quantil do Nível de Confiança usando a distribuição Normal
  - Função no Excel: INV.NORMP.N((1- $NIVEL\ DE\ CONFIANCA$) / 2)
- ε: Margem de Erro desejada

Noções de Inferência Estatística
Cálculo do Tamanho da Amostra

O fabricante de automóveis deseja que a **Margem de Erro** na estimação da média de compradores seja **inferior a 1,0**.

Dessa forma, o **Tamanho da Amostra** mínimo necessário para obter uma **Margem de Erro** inferior a solicitada pelo fabricante é (supondo que a variância amostral da idade das 30 pessoas entrevistadas seja **25**):

$$n = \frac{S^2 z_\gamma^2}{\varepsilon^2} = \frac{25 \cdot 1,65^2}{1^2} = 68$$

Ou seja, para obter uma Margem de Erro **inferior a 1** precisamos ter uma amostra com **pelo menos 68** pessoas entrevistadas.

Próximos passos

Nesta seção aprendemos como calcular a **Margem de Erro** e como calcular o **Tamanho da Amostra** de forma a controlar a **Margem de Erro**.

Nas próximas aulas veremos como usar a técnica **Testes de Hipóteses** para comparar grupos diferentes de dados. Muito importante em vários projetos de Dados no nosso dia a dia.

Até lá !

Preditiva.ai