Estatística Avançada - Aula 04

Fundamentos da Inferência Estatística: A Amostra

Kaique Matias de Andrade Roberto

Ciências Atuariais - Ciências Econômicas

HECSA - Escola de Negócios

FIAM-FAAM-FMU

Conteúdo

Conceitos que aprendemos em Aulas anteriores

Conceitos que aprendemos em Aulas anteriores

- variáveis aleatórias discretas e contínuas;
- distribuições de probabilidades discretas e contínuas;
- distribuição normal;
- vários cálculos envolvendo distribuições.

População e Amostra

População e Amostra

Vimos, na Aula-00, como resumir descritivamente variáveis associadas a um ou mais conjuntos de dados.

População e Amostra

Nas Aulas 01-03, construímos modelos teóricos (probabilísticos), identificados por parâmetros, capazes de representar adequadamente o comportamento de algumas variáveis.

População e Amostra

Agora vamos apresentar os argumentos estatísticos para fazer afirmações sobre as características de uma população, com base em informações dadas por amostras.

População e Amostra

O uso de informações de uma amostra para concluir sobre o todo faz parte da atividade diária da maioria das pessoas.

População e Amostra

Basta observar como uma cozinheira verifica se o prato que ela está preparando tem ou não a quantidade adequada de sal. Ou, ainda, quando um comprador, após experimentar um pedaço de laranja numa banca de feira, decide se vai comprar ou não as laranjas.

População e Amostra

Essas são decisões baseadas em procedimentos amostrais.

População e Amostra

Nosso objetivo na aula de hoje é procurar dar a conceituação formal a esses princípios intuitivos do dia-a-dia para que possam ser utilizados **cientificamente** em situações mais complexas.

População e Amostra

Nas aulas anteriores, tomamos conhecimento de alguns modelos probabilísticos que procuram medir a variabilidade de fenômenos casuais de acordo com suas ocorrências: as distribuições de probabilidades de variáveis aleatórias (qualitativas ou quantitativas).

População e Amostra

Na prática, frequentemente o pesquisador tem alguma ideia sobre a forma da distribuição, mas não dos valores exatos dos parâmetros que a especificam.

População e Amostra

Por exemplo, parece razoável supor que a distribuição das alturas dos brasileiros adultos possa ser representada por um modelo normal (embora as alturas não possam assumir valores negativos).

População e Amostra

Mas essa afirmação não é suficiente para determinar qual a distribuição normal correspondente; precisaríamos conhecer os parâmetros (média e variância) dessa normal para que ela ficasse completamente especificada.

População e Amostra

O propósito do pesquisador seria, então, descobrir (*estimar*) os parâmetros da distribuição para sua posterior utilização.

População e Amostra

Se pudéssemos medir as alturas de todos os brasileiros adultos, teríamos meios de obter sua distribuição exata e, daí, produzir os correspondentes parâmetros.

População e Amostra

Mas nessa situação não teríamos necessidade de usar a inferência estatística!

População e Amostra

Raramente se consegue obter a distribuição exata de alguma variável, ou porque isso é muito dispendioso, ou muito demorado ou às vezes porque consiste num processo destrutivo.

População e Amostra

Por exemplo, se estivéssemos observando a durabilidade de lâmpadas e testássemos todas até queimarem, não restaria nenhuma para ser vendida.

População e Amostra

Assim, a solução é selecionar parte dos elementos (amostra), analisá-la e inferir propriedades para o todo (população).

População e Amostra

Outras vezes estamos interessados em explorar relações entre variáveis envolvendo experimentos mais complexos, para a obtenção dos dados.

População e Amostra

Por exemplo, gostaríamos de obter resposta para a seguinte indagação: a altura que um produto é colocado na gôndola de um supermercado afeta a sua venda?

População e Amostra

Observe que para responder a questão precisamos obter dados de vendas com o produto oferecido em diferentes alturas, e que essas vendas sejam controladas para evitar interferências de outros fatores que não a altura.

População e Amostra

Nesse caso não existe claramente um conjunto de todos os elementos para os quais pudéssemos encontrar os parâmetros populacionais. Recorrer a modelos para descrever o todo (população) facilita a identificação e solução do problema.

População e Amostra

Nesse exemplo, supondo que as vendas V_h do produto oferecido na altura h ($h = 1$ representando baixo, $h = 2$ representando meio e $h = 3$ representando alto) segue uma distribuição próxima a normal, ou seja, $V_h \sim N(\mu_h, \sigma^2)$, o nosso problema passa a ser o de verificar, por meio de dados coletados do experimento (amostra), se existe evidência de igualdade das médias μ_1, μ_2 e μ_3.

População e Amostra

Note que, em nossa formulação do problema, supusemos que as três situações de alturas resultam observações com a mesma variância σ^2. Essa suposição poderia ser modificada.

População e Amostra

Soluções de questões como as apresentadas acima são o objeto da Inferência Estatística.

População e Amostra

Dois conceitos básicos são, portanto, necessários para o desenvolvimento da Inferência Estatística: população e amostra.

População e Amostra

Definição 2.1

População é o conjunto de todos os elementos ou resultados sob investigação. **Amostra** é qualquer subconjunto da população.

Exemplo 2.2 (Salários da empresa Gantois)

Consideremos uma pesquisa para estudar os salários dos 500 funcionários da Companhia Gantois. Seleciona-se uma amostra de 36 indivíduos, e anotam-se os seus salários.

População e Amostra

Salários da empresa Gantois

A variável aleatória a ser observada é "salário". A população é formada pelos 500 funcionários da companhia. A amostra é constituída pelos 36 indivíduos selecionados.

População e Amostra

Salários da empresa Gantois

Na realidade, estamos interessados nos salários, portanto, para sermos mais precisos, devemos considerar como a população os 500 salários correspondentes aos 500 funcionários. Consequentemente, a amostra será formada pelos 36 salários dos indivíduos selecionados.

População e Amostra

Salários da empresa Gantois

Podemos estudar a distribuição dos salários na amostra, e esperamos que esta reflita a distribuição de todos os salários, desde que a amostra tenha sido escolhida com cuidado.

Exemplo 2.3 (Quilombo da Boa Esperança)

Queremos estudar a proporção de indivíduos no quilombo da Boa Esperança que são favoráveis a certo projeto estrutural. Uma amostra de 200 pessoas é sorteada, e a opinião de cada uma é registrada como sendo a favor ou contra o projeto.

População e Amostra

Quilombo da Boa Esperança

A população consiste de todos os moradores do quilombo, e a amostra é formada pelas 200 pessoas selecionadas. Podemos definir a variável X, que toma o valor 1, se a resposta de um morador for favorável, e o valor 0, se a resposta for contrária ao projeto.

Quilombo da Boa Esperança

Assim, nossa população pode ser reduzida à distribuição de X, e a amostra será constituída de uma sequência de 200 zeros e uns.

Exemplo 2.4 (Vida-útil de Lâmpadas)

O interesse é investigar a duração de vida de um novo tipo de lâmpada, pois acreditamos que ela tenha uma duração maior do que as fabricadas atualmente.

População e Amostra

Vida-útil de Lâmpadas

Então, 100 lâmpadas do novo tipo são deixadas acesas até queimarem. A duração em horas de cada lâmpada é registrada. Aqui, a variável é a duração em horas de cada lâmpada.

População e Amostra

Vida-útil de Lâmpadas

A população é formada por todas as lâmpadas fabricadas ou que venham a ser fabricadas por essa empresa, com o mesmo processo. A amostra é formada pelas 100 lâmpadas selecionadas.

População e Amostra

Vida-útil de Lâmpadas

Note-se que nesse caso não podemos observar a população, ou seja, a distribuição da duração de vida das lâmpadas na população, pois isso corresponderia a queimar todas as lâmpadas.

População e Amostra

Vida-útil de Lâmpadas

Assim, em alguns casos, não podemos observar a população toda, pois isso significaria danificar (ou destruir) todos os elementos da população.

Vida-útil de Lâmpadas

Esse problema geralmente é contornado atribuindo-se um modelo teórico para a distribuição da variável populacional.

População e Amostra

Em alguns casos, fazemos suposições mais precisas sobre a população (ou sobre a variável definida para os elementos da população).

Exemplo 2.5 (Pacotes de Café)

Digamos que X represente o peso real de pacotes de café, enchidos automaticamente por uma máquina. Sabe-se que a distribuição de X pode ser representada por uma normal, com parâmetros μ e σ^2 desconhecidos.

População e Amostra

Pacotes de Café

Sorteamos 100 pacotes e medimos seus pesos. A população será o conjunto de todos os pacotes enchidos ou que virão a ser enchidos pela máquina, e que pode ser suposta como normal.

População e Amostra

Pacotes de Café

A amostra será formada pelas 100 medidas obtidas dos pacotes selecionados, que pode ser pensada como constituída de 100 observações feitas de uma distribuição normal.

Exemplo 2.6 (Honestidade da Moeda)

Para investigar a "honestidade" de uma moeda, nós a lançamos 50 vezes e contamos o número de caras observadas. A população pode ser considerada como tendo a distribuição da variável X, assumindo o valor 1, com probabilidade p, se ocorrer cara, e assumindo o valor 0, com probabilidade $1-p$, se ocorrer coroa.

Honestidade da Moeda

Ou seja, a população pode ser considerada como tendo distribuição de Bernoulli com parâmetro p. A variável ficará completamente especificada quando conhecermos p. A amostra será uma sequência de 50 números zeros ou uns.

Exemplo 2.7 (Tempo de Reação)

Há razões para supor que o tempo Y de reação a certo estímulo visual dependa da idade do indivíduo. Suponha, ainda, que essa dependência seja linear.

População e Amostra

Tempo de Reação

Para verificarmos se essa suposição é verdadeira, obtiveram-se 20 dados da seguinte maneira: 20 pessoas foram selecionadas, sendo 10 homens e 10 mulheres. Dentro de cada grupo de homens e mulheres foram selecionadas duas pessoas das seguintes faixas de idade: 20, 25, 30, 35 e 40 anos.

População e Amostra

Tempo de Reação

Cada pessoa foi submetida ao teste e seu tempo de reação y foi medido. A população poderia ser considerada como formada por todas aquelas pessoas que viessem a ser submetidas ao teste, segundo o sexo e a idade. A amostra é formada pelas 20 medidas.

Amostragem (revisitando)

Amostragem (revisitando)

Em problemas envolvendo amostras, antes de tomarmos uma decisão, queremos de responder a quatro perguntas:

Amostragem (revisitando)

- Qual a população a ser amostrada?
- Como obter os dados (a amostra)?
- Que informações pertinentes (estatísticas) serão retiradas da amostra?
- Como se comporta(m) a(s) estatística(s) quando o mesmo procedimento de escolher a amostra é usado numa população conhecida?

Amostragem (revisitando)

Vamos tentar responder essas perguntas nessa e nas próximas aulas.

Amostragem (revisitando)

As observações contidas em uma amostra são tanto mais informativas sobre a população quanto mais conhecimento explícito ou implícito tivermos dessa mesma população.

Amostragem (revisitando)

Exemplo 3.1

A análise da quantidade de glóbulos brancos obtida de algumas gotas de sangue da ponta do dedo de um paciente dará uma ideia geral da quantidade dos glóbulos brancos no corpo todo, pois sabe-se que a distribuição dos glóbulos brancos é homogênea, e de qualquer lugar que se tivesse retirado a amostra ela seria "representativa".

Amostragem (revisitando)

Mas nem sempre a escolha de uma amostra adequada é imediata.

Amostragem (revisitando)

Exemplo 3.2

Voltando ao Exemplo do Quilombo da Boa Esperança, para o qual queríamos obter uma amostra de habitantes para saber a opinião sobre um projeto estrutural, escolhendo intencionalmente uma amostra de 200 indivíduos moradores de certa região beneficiada pelo projeto, saberemos de antemão que o resultado conterá um viés de seleção.

Amostragem (revisitando)

A maneira de se obter a amostra é tão importante, e existem tantos modos de fazê-lo, que esses procedimentos constituem especialidades dentro da Estatística, sendo Amostragem e Planejamento de Experimentos as duas mais conhecidas.

Amostragem (revisitando)

Poderíamos dividir os procedimentos científicos de obtenção de dados amostrais em três grandes grupos:

Amostragem (revisitando)

- Levantamentos Amostrais;
- Planejamento de Experimentos;
- Levantamentos Observacionais.

Amostragem (revisitando)

Nesta disciplina iremos nos concentrar principalmente em levantamentos amostrais e, mais ainda, num caso simples de amostragem probabilística, a amostragem aleatória simples, com reposição, a ser designada por AAS.

Amostragem (revisitando)

A amostragem aleatória simples é a maneira mais fácil para selecionarmos uma amostra probabilística de uma população.

Amostragem (revisitando)

Além disso, o conhecimento adquirido com esse procedimento servirá de base para o aprendizado e desenvolvimento de outros procedimentos amostrais, planejamento de experimentos, estudos observacionais etc.

Amostragem (revisitando)

Comecemos introduzindo o conceito de AAS de uma população finita, para a qual temos uma listagem de todas as N unidades elementares. Podemos obter uma amostra nessas condições, escrevendo cada elemento da população num cartão, misturando-os numa urna e sorteando tantos cartões quantos desejarmos na amostra.

Amostragem (revisitando)

Esse procedimento torna-se inviável quando a população é muito grande. Nesse caso, usa-se um processo alternativo, no qual os elementos são numerados e em seguida sorteados por meio de uma tabela de números aleatórios ou por meio do uso de computadores, que podem gerar números aleatórios.

Amostragem (revisitando)

Utilizando-se um procedimento aleatório, sorteia-se um elemento da população, sendo que todos os elementos têm a mesma probabilidade de ser selecionados. Repete-se o procedimento até que sejam sorteadas as n unidades da amostra.

Amostragem (revisitando)

Podemos ter uma AAS com reposição, se for permitido que uma unidade possa ser sorteada mais de uma vez, e sem reposição, se a unidade sorteada for removida da população.

Amostragem (revisitando)

Do ponto de vista da quantidade de informação contida na amostra, amostrar sem reposição é mais adequado.

Amostragem (revisitando)

Contudo, a amostragem com reposição conduz a um tratamento teórico mais simples, pois ela implica que tenhamos independência entre as unidades selecionadas. Essa independência facilita o desenvolvimento das propriedades dos estimadores que serão considerados.

Amostragem (revisitando)

Portanto, para o restante desta disciplina, o plano amostral considerado será o de amostragem aleatória simples com reposição, que denotaremos simplesmente por AAS.

Amostragem (revisitando)

Definição 3.3

Uma amostra aleatória simples de tamanho n de uma variável aleatória X, com dada distribuição, é o conjunto de n variáveis aleatórias independentes $X_1, X_2, ..., X_n$, cada uma com a mesma distribuição de X.

Amostragem (revisitando)

Quando a população é caracterizada por uma distribuição de probabilidades, o modo mais simples para sortear uma AAS é usar algum procedimento de simulação.

Amostragem (revisitando)

O processo de simular uma observação de uma distribuição especificada por seus parâmetros nada mais é do que retirar uma AAS de tamanho um da população.

Amostragem (revisitando)

Desse modo, para retirar uma AAS (com reposição) de n indivíduos da população X, basta gerar n números aleatórios independentes dessa distribuição.

Amostragem (revisitando)

Obtida uma amostra, muitas vezes desejamos usá-la para produzir alguma característica específica.

Amostragem (revisitando)

Por exemplo, se quisermos calcular a média da amostra $(X_1, X_2, ..., X_n)$, esta será dada por

$$\overline{X} = \frac{1}{n}(X_1 + ... + X_n).$$

Amostragem (revisitando)

Veja que $\overline{X}$ **também é uma variável aleatória**. Podemos também estar interessados em qualquer outra característica da amostra, que será sempre uma função do vetor aleatório $(X_1, ..., X_n)$.

Amostragem (revisitando)

Definição 3.4

Um **parâmetro** é uma medida usada para descrever uma característica da população.

Amostragem (revisitando)

Definição 3.5

Uma **estatística** é uma característica da amostra, ou seja, uma estatística T é uma função de $X_1, X_2, ..., X_n$.

Amostragem (revisitando)

As estatísticas mais comuns para uma amostra $\{X_1, ..., X_n\}$ são:

- Média amostral

$$\overline{X} = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right);$$

- Variância amostral

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)$$

- Desvio padrão amostral

$$S = \sqrt{S^2}.$$

Amostragem (revisitando)

Denominação	População	Amostra
Média	$\mu = E(X)$	$\bar{X} = \sum X_i / n$
Mediana	$Md = Q_2$	$md = q_2$
Variância	$\sigma^2 = Var(X)$	$S^2 = \sum (X_i - \bar{X})^2/(n-1)$
Nº de elementos	N	n
Proporção	p	$\hat{p}$
Quantil	Q(p)	q(p)
Quartis	Q_1, Q_2, Q_3	q_1, q_2, q_3
Intervalo inter-quartil	$d_Q = Q_3 - Q_1$	$d_q = q_3 - q_1$
Função densidade	f(x)	histograma
Função de distribuição	F(x)	$F_e(x)$

Distribuição Amostral

Distribuição Amostral

Vimos que o problema da inferência estatística é fazer uma afirmação sobre os parâmetros da população através da amostra.

Distribuição Amostral

Digamos que nossa afirmação deva ser feita sobre um parâmetro θ da população (por exemplo, a média, a variância ou qualquer outra medida).

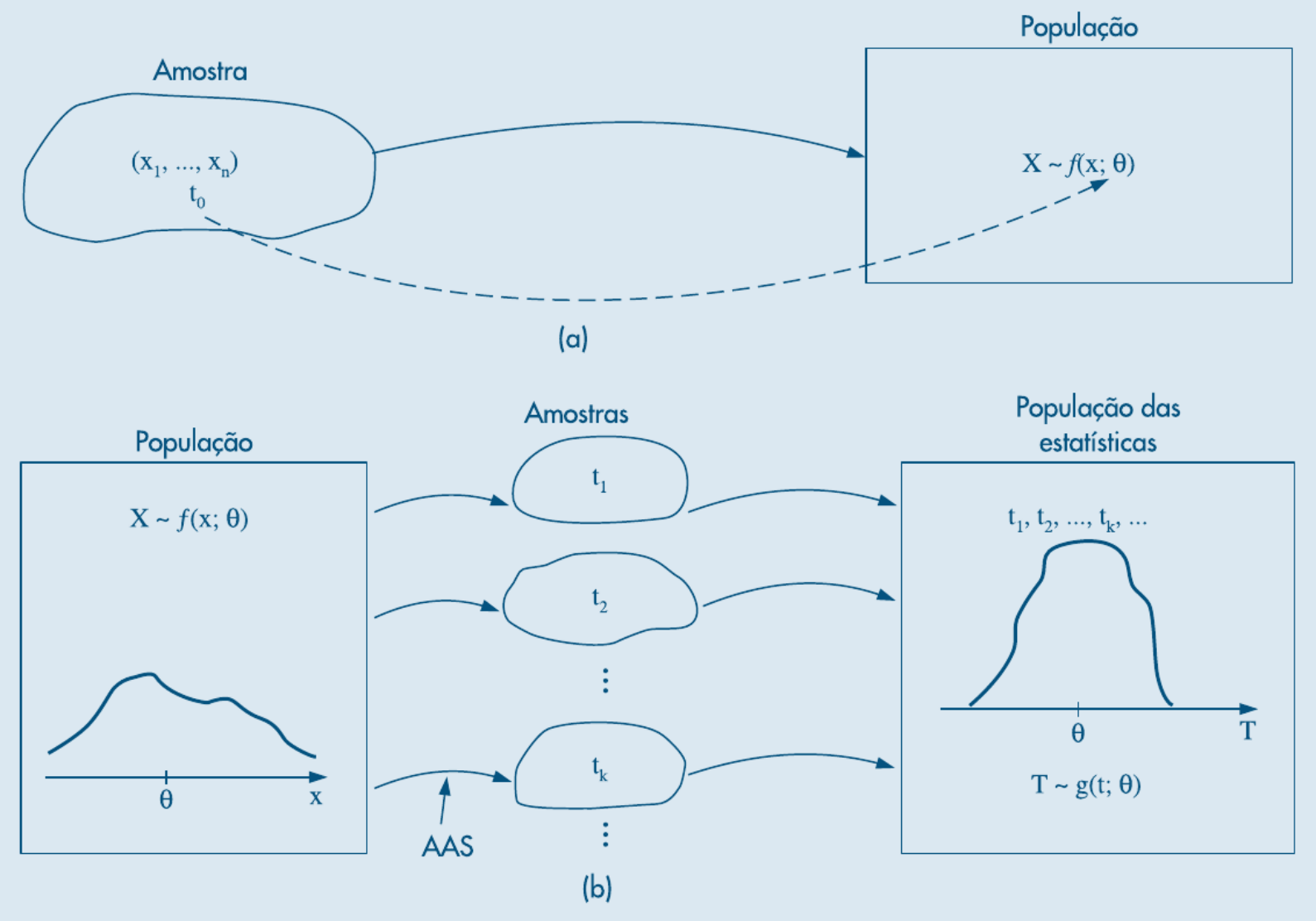
Distribuição Amostral

Decidimos que usaremos uma AAS de n elementos sorteados dessa população. Nossa decisão será baseada na estatística T, que será uma função da amostra $(X_1, X_2, ..., X_n)$, ou seja, $T = f(X_1, ..., X_n)$.

Distribuição Amostral

Colhida essa amostra, teremos observado um particular valor de T, digamos t_0, e baseados nesse valor é que faremos a afirmação sobre θ, o parâmetro populacional.

Distribuição Amostral



Distribuição Amostral

A validade da nossa resposta seria melhor compreendida se soubéssemos o que acontece com a estatística T, quando retiramos todas as amostras de uma população conhecida segundo o plano amostral adotado.

Distribuição Amostral

Isto é, qual a distribuição de T quando $(X_1, ..., X_n)$ assume todos os valores possíveis. Essa distribuição é chamada **distribuição amostral** da estatística T e desempenha papel fundamental na teoria da inferência estatística.

Distribuição Amostral

Esquematicamente temos:

- uma população X, com determinado parâmetro de interesse θ;
- todas as amostras retiradas da população, de acordo com certo procedimento;
- para cada amostra, calculamos o valor t da estatística T;
- os valores t formam uma nova população, cuja distribuição recebe o nome de distribuição amostral de T.

Distribuição Amostral

Vejamos alguns exemplos simples para aclarar um pouco mais o conceito de distribuição amostral de uma estatística.

Distribuição Amostral

Exemplo 4.1

Considere os dados a seguir:

Distribuição Amostral

Unidade	Nome do Chefe	Sexo	Idade	Fumante	Renda Bruta Familiar	Número de Trabalhadores
1	Ada	0	20	0	12	1
2	Beto	1	30	1	30	3
3	Ema	0	40	1	18	2

Distribuição Amostral

Considere como parâmetro

$$\theta = \text{Idade.}$$

Com plano amostral AAS para amostras de tamanho 2, calcule a distribuição amostral da estatística $\overline{X}$ (média amostral).

Distribuição Amostral

Faça as mesmas contas com a estatística S^2 (variância amostral).

Distribuição Amostral

É evidente que a distribuição de T irá depender da distribuição de X e do plano amostral, em nosso caso reduzido a AAS.

Distribuição Amostral

Vamos estudar agora a distribuição amostral da estatística $\overline{X}$, a média da amostra.

Distribuição Amostral

Consideremos uma população identificada pela variável X, cujos parâmetros média populacional $E(\overline{X}) = \mu$ e variância populacional $\text{Var}(X) = \sigma^2$ são supostos conhecidos.

Distribuição Amostral

Vamos retirar todas as possíveis AAS de tamanho n dessa população, e para cada uma calcular a média $\overline{X}$. Em seguida, consideremos a distribuição amostral e estudemos suas propriedades.

Teorema 4.2

Seja X uma variável aleatória com média μ e variância σ^2, e seja $(X_1, ..., X_n)$ uma AAS de X. Então,

$$E(\overline{X}) = \mu \text{ e } \text{Var}(X) = \frac{\sigma^2}{n}.$$

Teorema 4.3 (Limite Central)

Para amostras aleatórias simples $(X_1, ..., X_n)$, retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média $\overline{X}$ aproxima-se, para n grande, de uma distribuição normal, com média μ e variância σ^2/n.

Corolário 4.4

Se $(X_1, ..., X_n)$ for uma amostra aleatória simples da população X, com média μ e variância σ^2 finita, e $\overline{X} = (X_1 + X_2 + ... + X_n)/n$, então

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Distribuição Amostral

O TLC afirma que $\overline{X}$ aproxima-se de uma normal quando n tende para o infinito, e a rapidez dessa convergência depende da distribuição da população da qual a amostra é retirada.

Distribuição Amostral

Se a população original tem uma distribuição próxima da normal, a convergência é rápida; se a população original se afasta muito de uma normal, a convergência é mais lenta, ou seja, necessitamos de uma amostra maior para que $\overline{X}$ tenha uma distribuição aproximadamente normal.

Distribuição Amostral

Para amostras da ordem de 30 ou 50 elementos, a aproximação pode ser considerada boa.

Distribuição Amostral

Também estudaremos a Distribuição Amostral de uma Proporção (tema que detalharemos algumas aulas adiante).

O Tamanho da Amostra

O Tamanho da Amostra

Em nossas considerações anteriores fizemos a suposição que o tamanho da amostra, n, era conhecido e fixo.

O Tamanho da Amostra

Podemos, em certas ocasiões, querer determinar o tamanho da amostra a ser escolhida de uma população, de modo a obter um erro de estimação previamente estipulado, com determinado grau de confiança.

O Tamanho da Amostra

Não vamos entrar em grandes detalhes sobre esse tema (pois há uma disciplina inteira dedicada a isso).

O Tamanho da Amostra

Aqui, em geral, vamos partir da hipótese de que o procedimento amostral já foi realizado previamente.

O Tamanho da Amostra

Apenas para registro, segue uma fórmula para o tamanho da amostra em um caso específico.

O Tamanho da Amostra

Fórmula para o Tamanho da Amostra

Suponha que estejamos estimando a média μ populacional e para tanto usaremos a média amostral, $\overline{X}$, baseada numa amostra de tamanho n. Suponha que se queira determinar o valor de n de modo que

$$P(|\overline{X} - \mu| \leq \varepsilon) \geq \gamma,$$

com $0 < \gamma < 1$ e ε é o erro máximo que podemos suportar, ambos os valores fixados. Temos que

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

onde z_γ é tal que $P(-z_\gamma < Z < z_\gamma) = \gamma$.

Comentários Finais

Comentários Finais

Em resumo, começamos a estudar Inferência Estatística. Em particular, falamos de:

- população e amostra;
- amostragem;
- distribuição amostral;
- tamanho da amostra.

Comentários Finais

Nas próximas aulas nós vamos focar em:

- distribuição amostral da proporção;
- estimadores e estimativas;
- propriedades dos estimadores;
- intervalo de confiança.

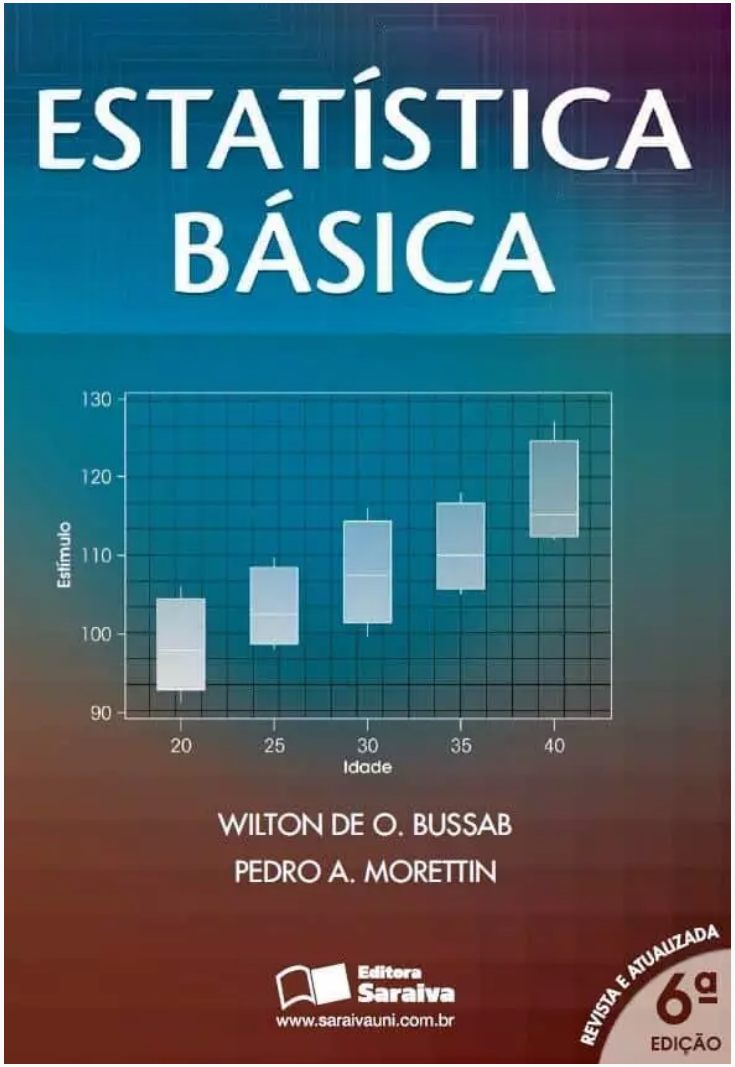
Comentários Finais

ATIVIDADE PARA ENTREGAR (E COMPOR A NOTA N1)

Em grupos de até 4 integrantes resolva três dentre os Exercícios 4.1-4.7.

Referências

Referências



Bons Estudos!