ACH2053
Introdução à Estatística

Inferência Estatística

Prof. Marcelo S. Lauretto
marcelolauretto@usp.br
www.each.usp.br/lauretto

Referência:
W.O.Bussab, P.A.Morettin. Estatística Básica, 6ª Edição.
São Paulo: Saraiva, 2010 – Capítulo 10

1. Inferência Estatística: Introdução

- Definição de Inferência Estatística:
  - Processo de aprender (inferir/generalizar) as características de uma população a partir de uma amostra
  - As características da população são denominadas parâmetros
    - Usualmente, não são observáveis diretamente
  - As características da amostra são denominadas estatísticas
    - São observadas / computadas a partir dos dados
- Contraste com a Estatística Descritiva:
  - Estatística descritiva foca exclusivamente nas propriedades dos dados observados
  - Não se assume que os dados vieram de uma população maior
  - Não há preocupação com a generalização para a população

1. Inferência Estatística: Introdução

- Principais problemas da inferência estatística:
  - Estimação
    - Derivação de estimativas pontuais e intervalos estatísticos para os parâmetros
  - Testes de hipóteses
  - Previsão

2. População e Amostra

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

- Exemplo 10.1 (adaptado): Salários dos moradores de um bairro
  - Consideremos uma pesquisa para estudar as remunerações dos *moradores de um bairro em São Paulo* (população)
  - Seleciona-se uma *amostra* de 2000 moradores daquele bairro
  - Esperamos que a distribuição observada dos salários na amostra reflita a distribuição de todos os salários – desde que a amostra tenha sido escolhida com cuidado.

2. População e Amostra

- Exemplo 10.2: Opinião sobre um projeto
  - Queremos estudar a proporção de indivíduos na cidade A que são favoráveis a um certo projeto governamental
  - Uma amostra de 200 moradores é sorteada, e a opinião (contrário/favorável) é registrada
  - Podemos definir a variável X da seguinte forma:
    - X=1 se o morador for favorável; X = 0 se for contrário
  - A amostra pode ser sintetizada como a sequência de 0's e 1's obtidos
  - Inferências de interesse:
    - Qual a proporção (*estimada*) de moradores favoráveis ao projeto?
      - Estimativa *pontual* e *intervalar*
    - Assumindo que uma pesquisa similar tenha sido conduzida na cidade B (com outra amostra, naturalmente), será que as taxas de aprovação ao projeto são as mesmas para as duas cidades?

2. População e Amostra

- Exemplo 10.3: Duração de lâmpadas
  - Suponha que o interesse seja investigar a duração de um novo tipo de lâmpada
  - Uma *amostra* de 100 lâmpadas do novo tipo são deixadas acesas até queimarem, e a duração (h) de cada lâmpada é registrada
  - População: universo de todas as lâmpadas fabricadas ou a serem fabricadas por essa empresa sob o mesmo processo
    - Impossível observar toda a população:
      - Ensaio destrutivo
      - Não é possível conhecer todas as lâmpadas que ainda serão produzidas

2. População e Amostra

- Exemplo 10.5: Moeda
  - Suponha que, no lançamento de uma moeda específica, consideramos a variável aleatória *X* definida como:
    - X = 1 se a moeda der cara; X = 0 se der coroa
  - A probabilidade da moeda dar cara, denotada por *p*, é desconhecida. Ou seja, $\Pr(X = 1) = \theta$, $\Pr(X = 0) = 1 - \theta$
  - Para poder conhecer melhor a moeda e podermos fazer algumas inferências sobre θ, lançamos a moeda 50 vezes e contamos o número de caras observadas.
  - A população pode ser considerada como tendo distribuição de Bernoulli com parâmetro θ.
  - A amostra será uma sequência de 50 números 0's e 1's.

2. População e Amostra

- Exemplo 10.6: Tempo de reação a estímulos visuais
  - Suponha que um investigador deseja verificar se o tempo *Y* de reação a certo estímulo visual depende da idade do indivíduo
  - Para verificar se a suposição é verdadeira, obteve-se uma amostra de 20 pessoas
    - 10 homens e 10 mulheres
    - Dentro de cada grupo de homens e de mulheres, foram selecionadas duas pessoas das seguintes faixas de idades: 20, 25, 30, 35 e 40 anos
  - Cada pessoa foi submetida ao teste e seu tempo de reação *y* foi medido
  - População: todas as pessoas que viessem a ser submetidas ao teste, segundo o sexo e a idade
  - Observação sobre notação:
    - Variável aleatória desconhecida: Y (maiúscula)
      Valores observados: y_1, y_2, ..., y_{20} (minúsculas)

3. Problemas de Inferência

- Exemplos de formulações e problemas de inferência:

- Retornando ao Exemplo 10.5 – moeda
  - Indicando por Y o número de caras obtidas depois de lançar a moeda 50 vezes, se pudermos assumir que os lançamentos são independentes e realizados aproximadamente sob as mesmas condições, sabemos que Y segue uma distribuição binomial, $Y \sim Bin(50, \theta)$
  - Suponha que, após os lançamentos da moeda, tenham ocorrido 36 caras.
    - Podemos concluir que a moeda é “honesta”?
      - Problema de teste de hipótese
    - Supondo que tenhamos concluído que a moeda não é honesta (ou seja, concluímos que $\theta \neq 1/2$), qual é a melhor estimativa para θ?
      - Problema de estimação

3. Problemas de Inferência

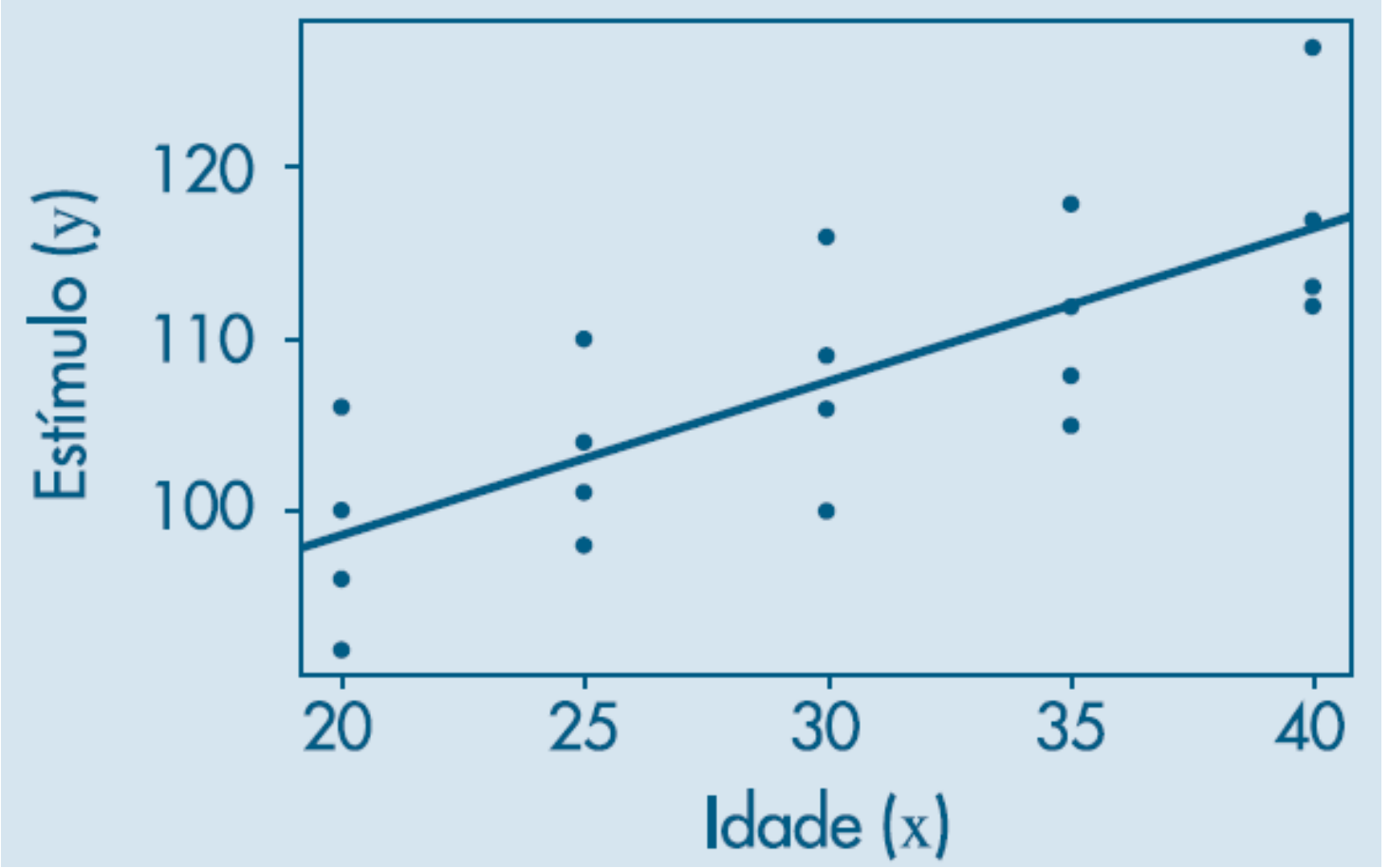
- Retornando ao Exemplo 10.6 – Tempo de reação a estímulos visuais
  - Um investigador deseja verificar se o tempo *Y* de reação a certo estímulo visual dependa da idade do indivíduo; para isso, tomou uma amostra de pessoas de diferentes idades
  - Suponha que o tempo Y, para uma dada idade x, seja uma variável aleatória com distribuição normal, com média dependendo da idade x, ou seja,

    $$Y \sim N(\mu(x), \sigma^2), \text{ onde } \mu(x) = \alpha + \beta x.$$

  - Problemas de interesse:
    - *Estimar* os parâmetros α e β (e assim explicar melhor a relação entre idade e tempo de reação)
    - *Testar* se $\beta = 0$ (uma forte evidência de que $\beta \neq 0$ indica que há uma associação, causal ou não, entre as duas variáveis)
    - *Prever* o tempo de reação para um indivíduo com uma certa idade

3. Problemas de Inferência

- Exemplo 10.6 (cont): Tempo de reação a estímulos visuais



Este exemplo é um problema de *regressão*, que não será abordado nesta disciplina

3. Problemas de Inferência

- Perguntas importantes antes de aplicar um plano para selecionar amostras:
  a) Qual a população a ser amostrada?
  b) Quais são os parâmetros de interesse sobre essa população?
     - E quais as inferências de interesse?
  c) Como obter os dados (a amostra)?
  d) Que informações pertinentes (estatísticas) serão retiradas da amostra?
  e) Como se comportariam as estatísticas se o mesmo procedimento de escolher a amostra fosse usado numa população (distribuição) conhecida?
     - (Qual a distribuição amostral da estatística)?

4. Como selecionar uma amostra

- As observações contidas em uma amostra são tanto mais informativas sobre a população quanto mais conhecimento explícito ou implícito houver sobre essa população
  - Para estimar a quantidade de glóbulos brancos no sangue de uma pessoa, algumas gotas colhidas na ponta do dedo fornecem uma amostra “representativa”
    - Distribuição de glóbulos brancos é homogênea
  - Já para o exemplo 2 (opinião sobre projeto governamental), entrevistar pessoas apenas em um bairro pode não ser representativo
    - Viés de seleção: bairros beneficiados tendem a ser mais favoráveis
    - Se a opinião tiver associação com fatores socioeconômicos, entrevistar apenas um bairro dará uma ideia apenas das subpopulações com mesmas características daquele bairro

4. Como selecionar uma amostra

- Procedimentos científicos de obtenção de dados amostrais podem ser divididos em três grandes grupos
  1. Levantamentos amostrais:
     - A amostra é obtida de uma população bem definida, por meio de processos bem protocolados e controlados pelo pesquisador
     - Podem ser subdivididos em três subgrupos:
       - Levantamentos probabilísticos: usam mecanismos aleatórios de seleção dos elementos de uma amostra, atribuindo a cada um deles uma probabilidade, conhecida a priori, de pertencer à amostra
       - Levantamentos não-probabilísticos: incluem outros grupos, como:
         - » amostras intencionais, obtidas por otimização ou com o auxílio de especialistas
         - » amostras de voluntários (também chamadas amostras por conveniência), como ocorre em ensaios clínicos, análise de crédito, etc.
       - Procedimentos híbridos (semi-intencionais)

4. Como selecionar uma amostra

2. Planejamento de experimentos:
   - Objetivo principal é o de analisar o efeito de uma variável sobre outra
   - Requer interferências do pesquisador sobre o ambiente em estudo (população), bem como o controle de fatores externos, com o intuito de medir o efeito desejado.
   - Ex 1: considere a seguinte pergunta: a altura em que um produto é colocado na gôndola de um supermercado afeta sua venda?
     Para responder a essa pergunta, é necessário
     - obter dados de vendas do produto sob diferentes alturas
     - que essas vendas sejam controladas para evitar interferências de outros fatores que não a altura (p.ex. sazonalidade)
   - Ex 2: ensaios clínicos para teste de eficácia de novos medicamentos
     - necessidade de grupo controle para comparação
     - fatores associados com o desfecho da doença (ex. sexo, idade, etnia, gravidade da doença etc) precisam ser controlados
       - » grupos de controle e de tratamento precisam ser razoavelmente similares em relação a esses fatores

4. Como selecionar uma amostra

3. Levantamentos observacionais:
   - Os dados são coletados sem que o pesquisador tenha controle sobre as informações obtidas, exceto eventualmente sobre possíveis erros grosseiros ou condições anômalas
   - Ao contrário de um planejamento de experimentos, no qual os indivíduos são alocados aos grupos, em um levantamento observacional os indivíduos da amostra não foram designados aos grupos, mas já pertenciam previamente aos respectivos grupos
   - Ex 1: Comparação de certos fenômenos entre alcoólatras e não alcoólatras; ou ainda entre homens e mulheres
     - Os indivíduos já pertenciam aos respectivos grupos *antes* do levantamento
   - Ex 2: Previsão de vendas de uma empresa em função de vendas passadas
     - Pesquisador não pode selecionar dados: esses são as vendas efetivamente ocorridas.

4. Como selecionar uma amostra

- Problemas:

1. Dê sua opinião sobre os tipos de problemas que surgiriam nos seguintes planos amostrais:
   (a) Para investigar a proporção dos operários de uma fábrica favoráveis à mudança do início das atividades das 7h para as 7h30, decidiu-se entrevistar os 30 primeiros operários que chegassem à fábrica na quarta-feira.
   (b) Mesmo procedimento, só que o objetivo é estimar a altura média dos operários.
   (c) Para estimar a porcentagem média da receita municipal investida em lazer, enviaram-se questionários a todas as prefeituras, e a amostra foi formada pelas prefeituras que enviaram as respostas.
   (d) Para verificar o fato de oferecer brindes nas vendas de sabão em pó, tomaram-se quatro supermercados na zona sul e quatro na zona norte de uma cidade. Nas quatro lojas da zona sul, o produto era vendido com brinde, enquanto nas outras quatro era vendido sem brinde. No fim do mês, compararam-se as vendas da zona sul com as da zona norte.

5. Amostragem aleatória simples

- A amostragem aleatória simples (AAS) é a maneira mais fácil para seleção de uma amostra probabilística de uma população
- Para populações finitas, onde se tem uma listagem de todas as *N* unidades elementares, pode-se atribuir um número sequencial para cada elemento, e em seguida sortear-se *n* desses números (por métodos manuais ou por rotinas computacionais)
- Todos os elementos têm a mesma probabilidade de ser selecionados

5. Amostragem aleatória simples

- Exemplo 10.7 (adaptado):
  - Uma urna (população) contém cinco tiras de papel, numeradas 1,3,5,5,7
  - Defina a variável *X* como sendo o valor assumido por um elemento retirado ao acaso da população. A distribuição de *X* é dada pela Tabela 10.1 abaixo

Tabela 10.1: Distribuição da v.a. X para o Problema 2.

x	1	3	5	7
$P(X = x)$	1/5	1/5	2/5	1/5

  - Suponha que duas tiras sejam retiradas ao acaso da urna, com reposição.
  - Denote por X_1 e X_2 os números sorteados na 1ª e na 2ª extração

5. Amostragem aleatória simples

- Exemplo 10.7 (cont):
  - A distribuição conjunta do par (X_1, X_2) pode ser calculada diretamente por $\Pr(X_1, X_2) = \Pr(X_1)\Pr(X_2)$, já que X_1 e X_2 são independentes. Exemplos:
    - $\Pr(1,1) = \Pr(1)\Pr(1) = \frac{1}{5}\frac{1}{5} = \frac{1}{25}$
    - $\Pr(1,5) = \Pr(1)\Pr(5) = \frac{1}{5}\frac{2}{5} = \frac{2}{25}$
    - Tabela 10.2 apresenta as probabilidades de todos os pares
  - Além disso, as distribuições marginais de X_1 e X_2 (somas das linhas e das colunas na tabela anterior) são independentes e iguais às distribuições de *X* (ver tabela 10.2 abaixo).

5. Amostragem aleatória simples

- Exemplo 10.7 (cont):
  - Desse modo, cada uma das 25 possíveis amostras de tamanho 2 que podemos extrair dessa população corresponde a observar uma realização particular da variável aleatória conjunta (X_1, X_2), com X_1 e X_2 independentes e $\Pr(X_1 = x) = \Pr(X_2 = x) = \Pr(X = x)$, para todo x.
  - Essa é a caracterização de amostra casual simples que usaremos nesta disciplina.

5. Amostragem aleatória simples

- Exemplo 10.7 (cont):

Tabela 10.2: Distribuição das probabilidades das possíveis amostras de tamanho 2 que podem ser selecionadas com reposição da população $\{1, 3, 5, 5, 7\}$.

X_2 \ X_1	1	3	5	7	Total
1	1/25	1/25	2/25	1/25	1/5
3	1/25	1/25	2/25	1/25	1/5
5	2/25	2/25	4/25	2/25	2/5
7	1/25	1/25	2/25	1/25	1/5
Total	1/5	1/5	2/5	1/5	1

5. Amostragem aleatória simples

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

  - A amostra será a n-upla ordenada $(X_1, X_2, \ldots, X_n)$, onde X_i indica a observação do i-ésimo elemento sorteado.

5. Amostragem aleatória simples

- Note que amostras aleatórias obtidas sem reposição não satisfazem à definição acima.
  - Tomando o Exemplo 10.7 (urna com cinco tiras), suponha que X_1 e X_2 sejam retirados sem reposição
  - Note que X_1 e X_2 não são independentes, e a distribuição de probabilidades de X_2 após a retirada de X_1 não é igual à distribuição original. P.ex.
    - $\Pr(X_2 = 1 | X_1 = 1) = 0 \neq \Pr(X_2 = 1) = \frac{1}{5}$
    - $\Pr(X_2 = 3 | X_1 = 1) = \frac{1}{4} \neq \Pr(X_2 = 3) = \frac{1}{5}$
    - $\Pr(X_2 = 5 | X_1 = 1) = \frac{1}{2} \neq \Pr(X_2 = 5) = \frac{2}{5}$

5. Amostragem aleatória simples

- Problemas:

3. A distribuição do número de filhos, por família, de uma zona rural está no quadro abaixo.

Nº de filhos	Porcentagem
0	10
1	20
2	30
3	25
4	15
Total	100

(a) Sugira um procedimento para sortear uma observação ao acaso dessa população.

(b) Dê, na forma de uma tabela de dupla entrada, as possíveis amostras do número de filhos de duas famílias que podem ser sorteadas e as respectivas probabilidades de ocorrência.

(c) Se fosse escolhida uma amostra de tamanho 4, qual seria a probabilidade de se observar a quádrupla ordenada $(2, 3, 3, 1)$?

(d) Responder ao item (b) por meio de simulação (gerando 10.000 pares de famílias e calculando suas probabilidades através das respectivas frequências); comparar os resultados com os do item (b)

6. Estatísticas e parâmetros

- Obtida uma amostra, quase sempre desejamos usá-la para produzir alguma característica específica

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

$$\overline{X} = \frac{1}{n}\{X_1 + X_2 + \ldots + X_n\}.$$

  – Note que $\bar{X}$ também uma variável aleatória!
  (Pois só é conhecida após a observação da amostra)

- Outras características da amostra também serão funções do vetor $(X_1, X_2, \ldots, X_n)$.

6. Estatísticas e parâmetros

- **Definição:**
  - Uma *estatística* é uma característica da amostra, ou seja, uma estatística T é uma função de $X_1, X_2, \ldots, X_n$.
    Notação: $r(X_1, X_2, \ldots, X_n)$

- Algumas estatísticas comuns são:

$$\overline{X} = 1/n \sum_{i=1}^{n} X_i : \text{ média da amostra,}$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2 : \text{variância da amostra,}$$

$$X_{(1)} = \min (X_1, X_2, \ldots, X_n) : \text{o menor valor da amostra,}$$

6. Estatísticas e parâmetros

- Algumas estatísticas comuns são (cont):

$$X_{(n)} = \max\,(X_1, X_2, \ldots, X_n) : \text{o maior valor da amostra,}$$

$$W = X_{(n)} - X_{(1)} : \text{amplitude amostral,}$$

$$X_{(i)} = \text{a } i\text{-ésima maior observação da amostra.}$$

- Em inferência estatística, usamos nomenclaturas distintas para as características da amostra e da população.

6. Estatísticas e parâmetros

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

- Assim, se estivermos colhendo amostras de uma população, identificada pela variável aleatória X, seriam parâmetros a média $E(X)$ e a variância $\mathrm{Var}(X)$.

6. Estatísticas e parâmetros

- Símbolos mais comuns para parâmetros e estatísticas:

Denominação	População	Amostra
Média	$\mu = E(X)$	$\bar{X} = \sum X_i / n$
Mediana	$\mathrm{Md} = Q_2$	$\mathrm{md} = q_2$
Variância	$\sigma^2 = \mathrm{Var}(X)$	$S^2 = \sum (X_i - \bar{X})^2/(n-1)$
Nº de elementos	N	n
Proporção	p	$\hat{p}$

6. Estatísticas e parâmetros

- Símbolos mais comuns para parâmetros e estatísticas:

Denominação	População	Amostra
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	$F(x)$	$F_e(x)$

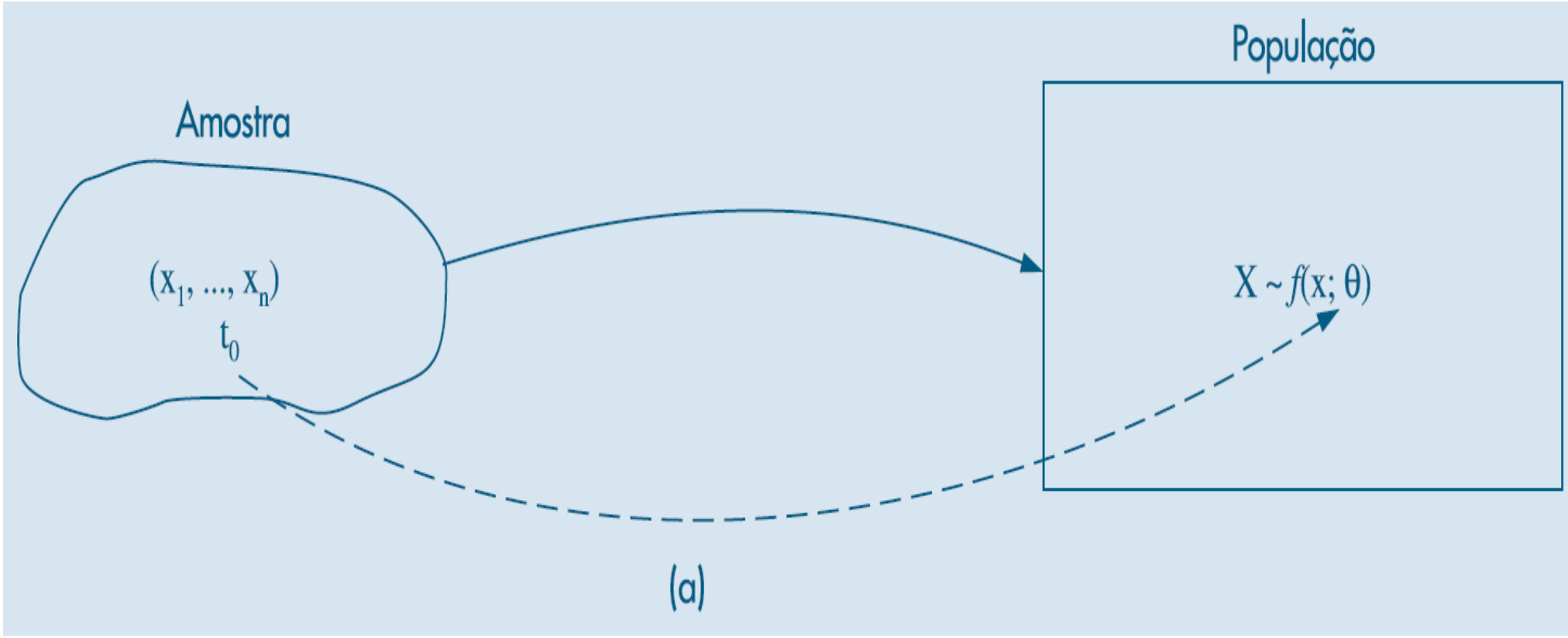
7. Distribuições amostrais

- O problema da inferência é fazer uma afirmação sobre os parâmetros da população através da amostra
- Digamos que nossa afirmação deva ser feita sobre um parâmetro θ da população (p.ex. média, variância ou qualquer outra medida)
- Suponha que foi adotada uma AAS (amostragem aleatória simples) de *n* elementos sorteados dessa população.
- Nossa decisão será baseada na estatística *T*, que será uma função da amostra $(X_1, X_2, \ldots, X_n)$, isto é, $T = r(X_1, X_2, \ldots, X_n)$
- Colhida a 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

7. Distribuições amostrais

- Esquema de inferência sobre θ

Figura 10.1 (a): Esquema de inferência sobre θ



7. Distribuições amostrais

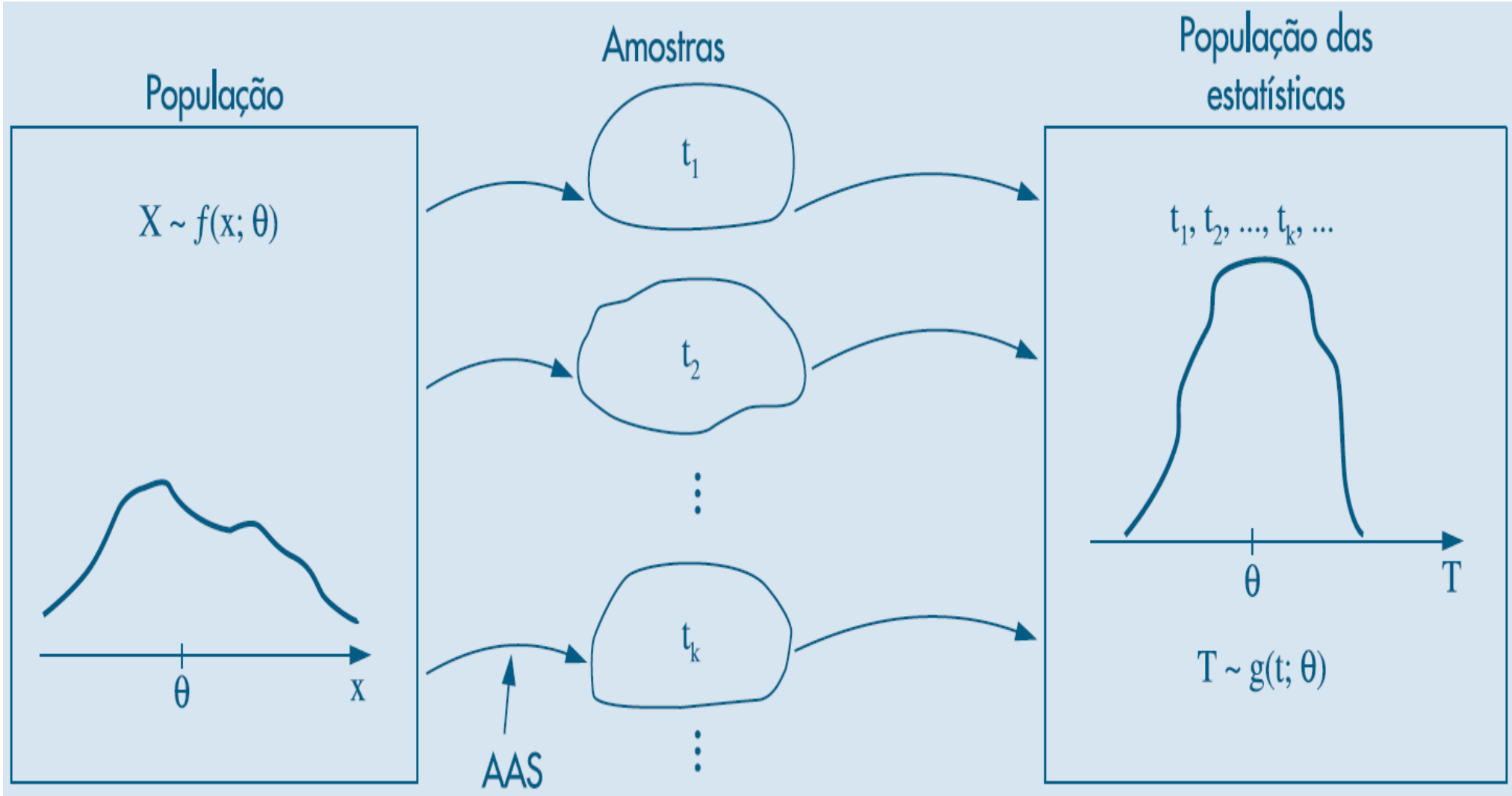
- A validade da afirmação (inferência) sobre θ depende de conhecermos o que ocorreria com a estatística *T*, se pudéssemos retirar todas as amostras da população (usando o mesmo plano amostral).
- Ou seja, deveríamos conhecer qual seria a distribuição de *T* se pudéssemos calcular *T* para todos os valores possíveis de $(X_1, X_2, \ldots, X_n)$.
- Essa distribuição é chamada *distribuição amostral da estatística T* e desempenha papel fundamental na teoria da inferência estatística frequentista.
  - Obs: Na inferência Bayesiana, busca-se conhecer a distribuição do próprio parâmetro populacional θ sem necessidade do conceito de distribuição amostral; todavia, essa abordagem não será estudada nesta disciplina.

7. Distribuições amostrais

- A distribuição amostral pode ser compreendida esquematicamente conforme figura 10.1 (b), onde se tem:
  - Uma variável aleatória X na população segue uma distribuição de probabilidade $X \sim f(x|\theta)$, onde θ é o parâmetro de interesse
  - Todas as amostras retiradas da população, de acordo com um procedimento de amostragem pré-definido
  - Para cada amostra, calcula-se o valor t da estatística T
  - Os valores t formam uma nova população, cuja distribuição recebe o nome de distribuição amostral de T

7. Distribuições amostrais

Figura 10.1 (b): Distribuição amostral da estatística *T*



7. Distribuições amostrais

- Exemplo 10.9:
  - Voltemos ao exemplo 10.7, no qual consideramos a seleção de amostras de tamanho 2, com reposição, da população {1, 3, 5, 5, 7}
  - Consideremos a distribuição da estatística *média amostral*

$$\bar{X} = \frac{X_1 + X_2}{2}$$

Note que assumimos amostras de tamanho 2

  - Essa distribuição é obtida com o auxílio da Tabela 10.2
    - Por exemplo, $\bar{X} = 1$ somente ocorre o par (1,1) e portanto

$$\Pr(\bar{X} = 1) = 1/25.$$

    - $\bar{X} = 3$ ocorre para os pares {(1,5), (3,3), (5,1)} e, portanto,

$$\Pr(\bar{X} = 3) = \frac{2}{25} + \frac{1}{25} + \frac{2}{25} = \frac{5}{25} = \frac{1}{5}\ .$$

7. Distribuições amostrais

- Exemplo 10.9 (cont):

Tabela 10.2: Distribuição das probabilidades das possíveis amostras de tamanho 2 que podem ser selecionadas com reposição da população $\{1, 3, 5, 5, 7\}$.

X_2 \ X_1	1	3	5	7	Total
1	1/25	1/25	2/25	1/25	1/5
3	1/25	1/25	2/25	1/25	1/5
5	2/25	2/25	4/25	2/25	2/5
7	1/25	1/25	2/25	1/25	1/5
Total	1/5	1/5	2/5	1/5	1

7. Distribuições amostrais

- Exemplo 10.9 (cont):
  - Procedendo de maneira análoga para os demais valores que $\bar{X}$ pode assumir, obtemos a Tabela 10.3

Tabela 10.3: Distribuição amostral da estatística $\bar{X}$.

$\bar{x}$	1	2	3	4	5	6	7	Total
$P(\bar{X} = \bar{x})$	1/25	2/25	5/25	6/25	6/25	4/25	1/25	1,00

7. Distribuições amostrais

- Exemplo 10.9 (cont):
  - Distribuições amostrais de outras estatísticas de interesse podem ser obtidas:
  - P.ex: W=amplitude $W = \max(X) - \min(X)$
    S^2=desvio quadrático médio $S^2 = \sum(X_i - \overline{X})^2/(n-1)$

Tabela 10.4: Distribuição amostral de W.

w	0	2	4	6	Total
$P(W = w)$	7/25	10/25	6/25	2/25	1,00

Tabela 10.5: Distribuição amostral de S^2.

s^2	0	2	8	18	Total
$P(S^2 = s^2)$	7/25	10/25	6/25	2/25	1,00

7. Distribuições amostrais

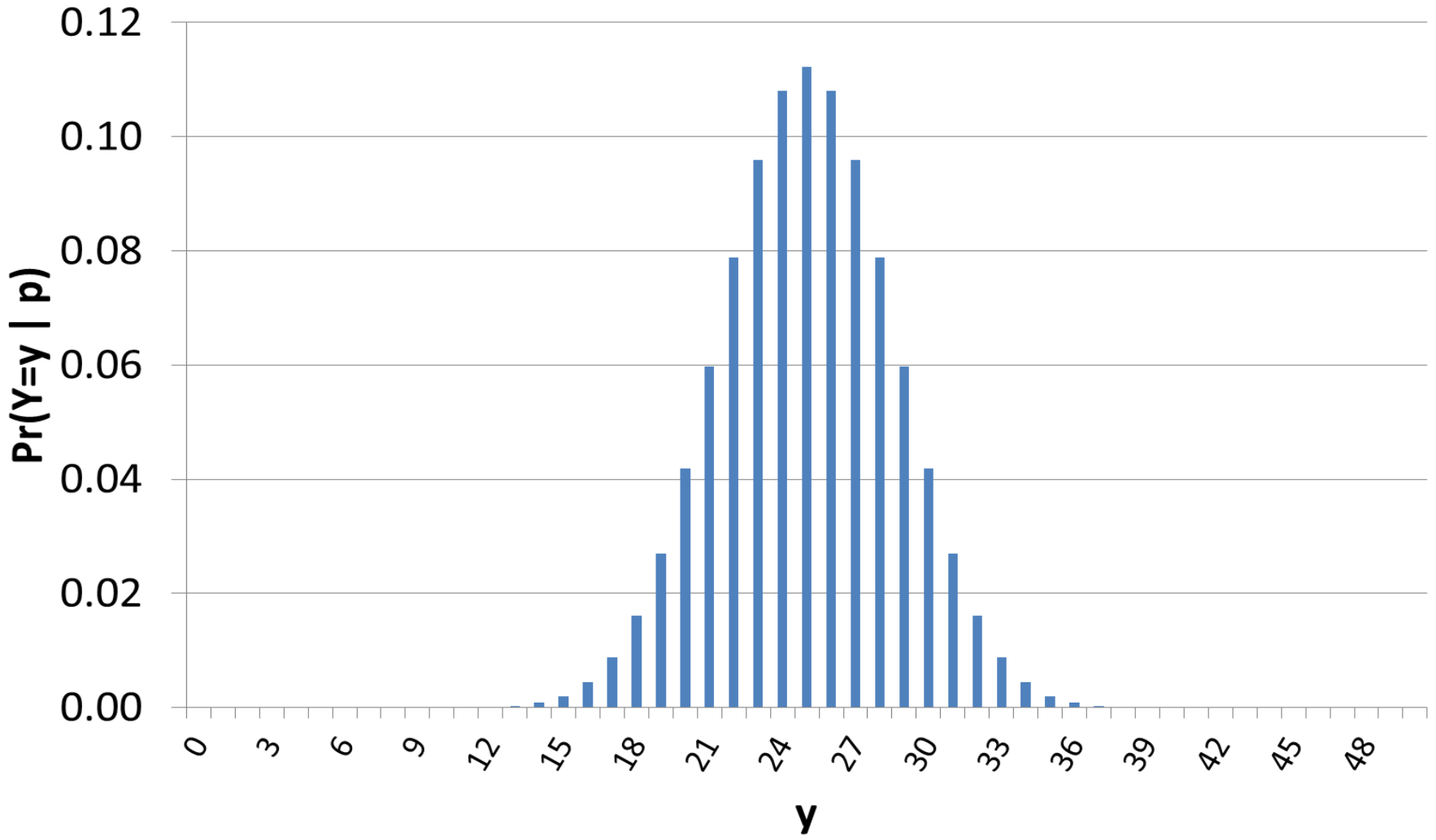
- Exemplo 10.5 (cont):
  - No caso do lançamento de uma moeda 50 vezes, usando como estatística

    Y = número de caras obtidas,

    a obtenção da distribuição amostral, que já foi vista, é feita por meio do modelo binomial $Bin(50, p)$, onde p denota a probabilidade de ocorrência de cara em um lançamento, $0 < p < 1$.
  - Suponha que, na realização do experimento, obtivemos Y = 36 caras
  - Se a moeda fosse honesta, a probabilidade de se obterem 36 ou mais caras em 50 lançamentos seria da ordem de 1/1000
  - Ou seja, se a moeda fosse honesta, o resultado observado (36 caras) seria muito pouco provável, o que indica que a probabilidade de cara é maior do que meio, ou seja, $p > 0{,}5$.

7. Distribuições amostrais

- Exemplo 10.5 (cont):



7. Distribuições amostrais

- Exemplo 10.5 (cont):
  - Outra forma de calcular $\Pr(Y \geq 36|p = 0.5)$: Simulação
    1. Sorteie *M* valores $Y_1, Y_2, \ldots, Y_M$, cada qual com distribuição $Bin(50, p)$. *M* deve ser um número moderado (p.ex. *M*=10000)
    2. A probabilidade $\Pr(Y \geq 36|p = 0.5)$ será estimada por

$$\Pr(Y \geq 36|p = 0.5) = \frac{|A|}{M}, \text{ em que } |A| = \text{quant. valores } Y_i \geq 36$$

  - Exemplo de script em R:

```
M = 10000
Y = rbinom(n=M, size=50, prob=0.5)
hist(Y, breaks=100)

prY = length(which(Y>=36)) / M
print(prY)
```

7. Distribuições amostrais

- Exemplo 10.8:
  - Considere a retirada de uma AAS de 5 alturas (em cm) de uma população de mulheres cujas alturas *X* seguem a distribuição normal $\mathrm{N}(167; 25)$ ($\mu = 167, \sigma^2 = 25, \sigma = 5$)
  - Qual seria a distribuição amostral da mediana das 5 alturas retiradas da população?
  - Como não podemos gerar todas as possíveis amostras de tamanho 5 da população, é possível simular um conjunto grande de amostras de tamanho 5, calcular a mediana de cada amostra e em seguida calcular algumas estatísticas de interesse sobre as medianas calculadas
    - Este procedimento é usualmente denominado ***Bootstrap paramétrico*** (veremos o conceito de *Bootstrap* mais adiante na disciplina)

7. Distribuições amostrais

- Exemplo 10.8 (cont):
  - No exemplo do livro, os autores geraram 200 amostras de tamanho 5 (denotadas por $X_1, X_2, \ldots, X_{200}$) e obtiveram os seguintes resultados:

$$E(\text{md}) = 166{,}88, \quad \text{Var}(\text{md}) = 7{,}4289, \quad \text{dp}(\text{md}) = 2{,}72,$$

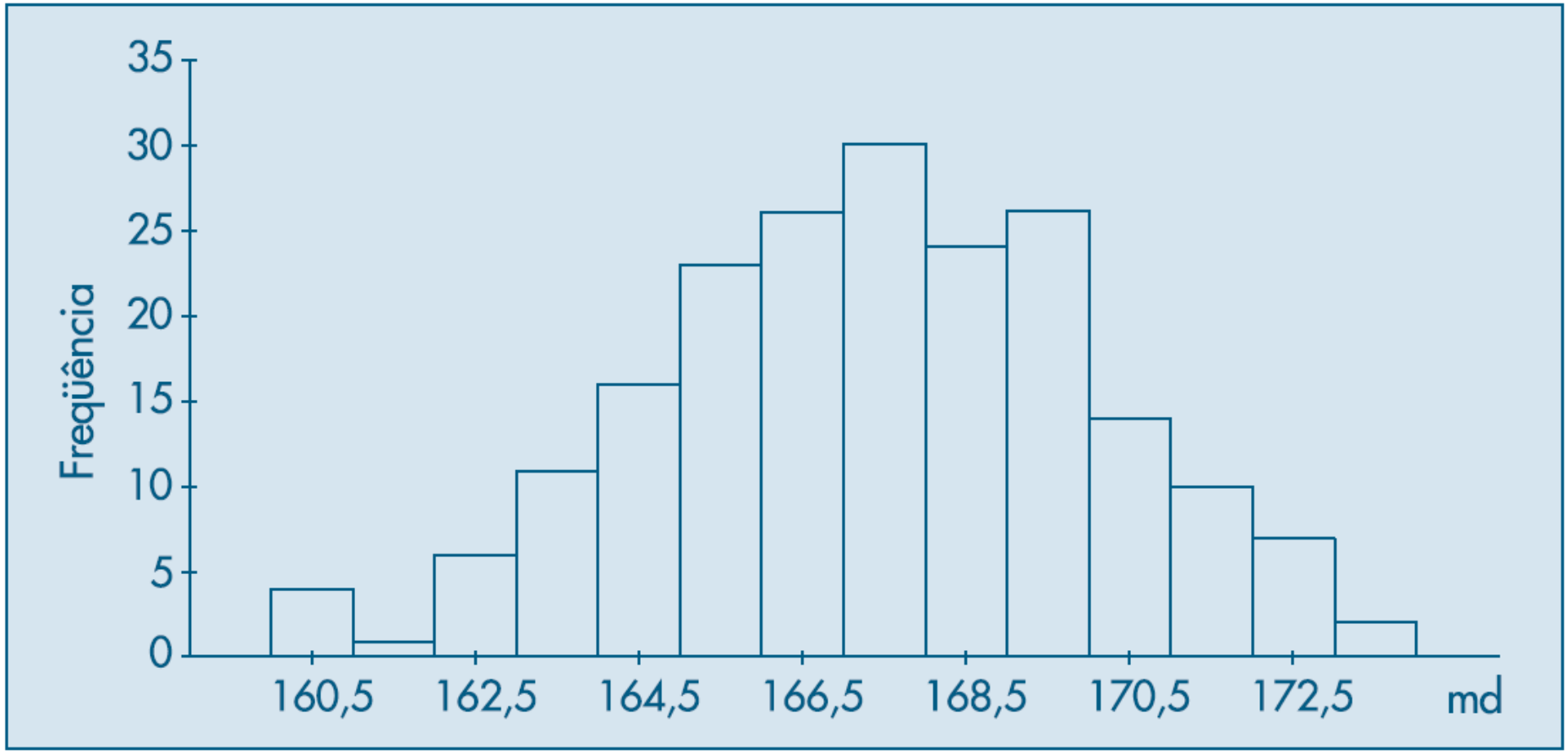
$$x_{(1)} = \min(X_1, \ldots, X_{200}) = 160, \quad x_{(200)} = \max(X_1, \ldots, X_{200}) = 173.$$

  - Os resultados indicam que a distribuição amostral de md deve ser próxima de uma normal, com média próxima de $\mu = 167$ e desvio padrão menor do que $\sigma = 5$.
  - Figura 10.3 apresenta o histograma dos valores das medianas obtidos nas 200 amostras

7. Distribuições amostrais

- Exemplo 10.8 (cont):

Figura 10.3: Distribuição amostral da mediana, obtida de 200 amostras de tamanho 5 de $X \sim N(167; 25)$.



7. Distribuições amostrais

- Exemplo 10.8 (cont):
  - Script em R para simulação:

```
mu = 167    # media da populacao
sigma = 5   # desvio padrao da populacao
M=4000      # Numero de amostras simuladas
smpsz = 5   # Tamanho de cada amostra

alturas = rnorm(n=smpsz*M, mean=mu, sd=sigma)
X = matrix(alturas, ncol=smpsz)

medias = apply(X, 1, mean)
print(c(mean(medias), var(medias), sd(medias),
        min(medias), max(medias)))
hist(medias, breaks=100)

medianas = apply(X, 1, median)
print(c(mean(medianas), var(medianas), sd(medianas),
        min(medianas), max(medianas)))
hist(medianas, breaks=100)
```

7. Distribuições amostrais

- Exemplo 10.8 (cont):
  - Resultados:
    - Média:

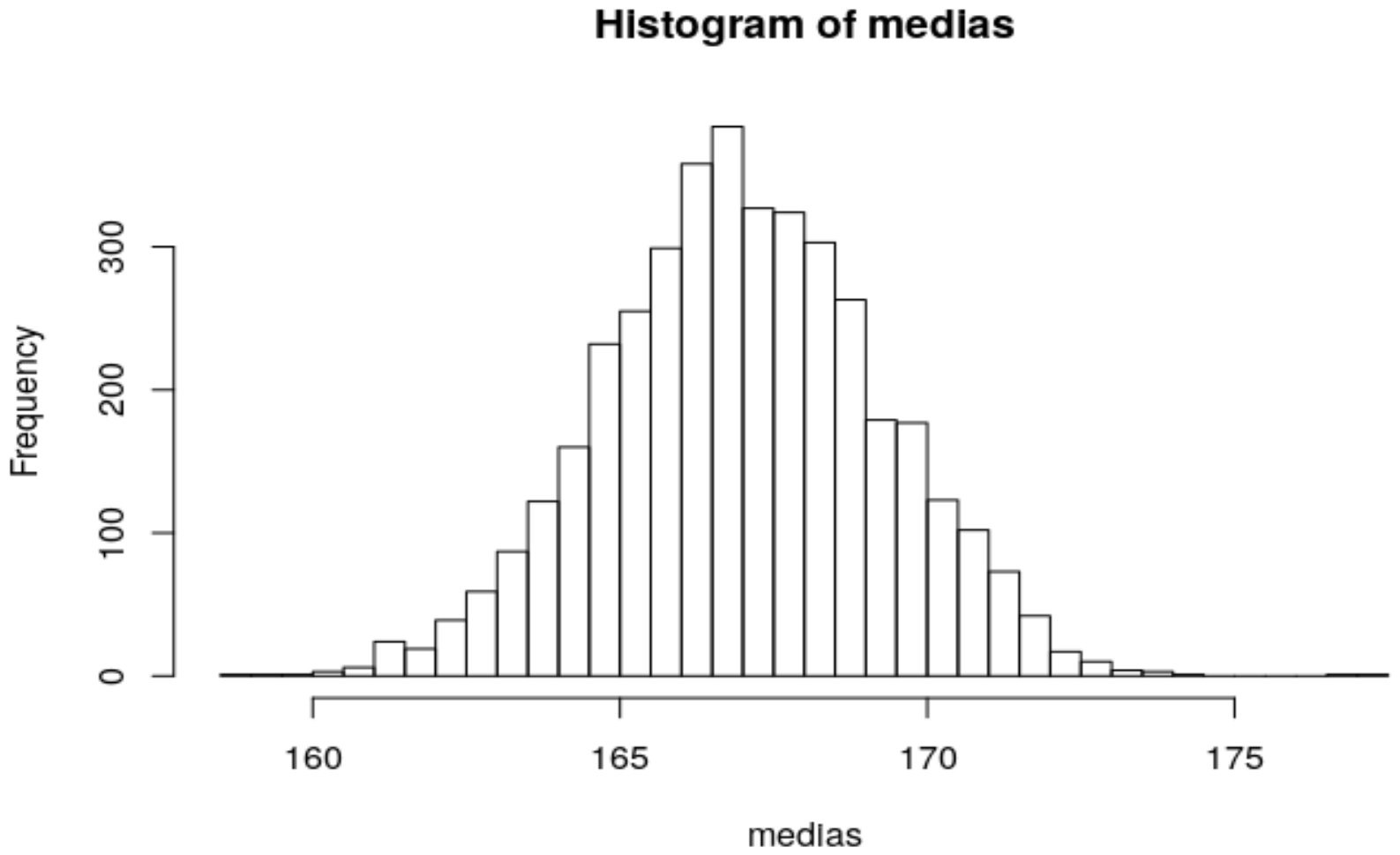
$$E(\bar{X}) = 167.0;\ \ Var(\bar{X}) = 5.05;\ \ dp(\bar{X}) = 2.25;$$
$$\min(\bar{X}) = 158.7;\ \ \max(\bar{X}) = 177.0$$

    - Mediana:

$$E(md) = 167.0;\ \ Var(md) = 7.21;\ \ dp(md) = 2.68;$$
$$\min(md) = 158.0;\ \ \max(md) = 178.0$$

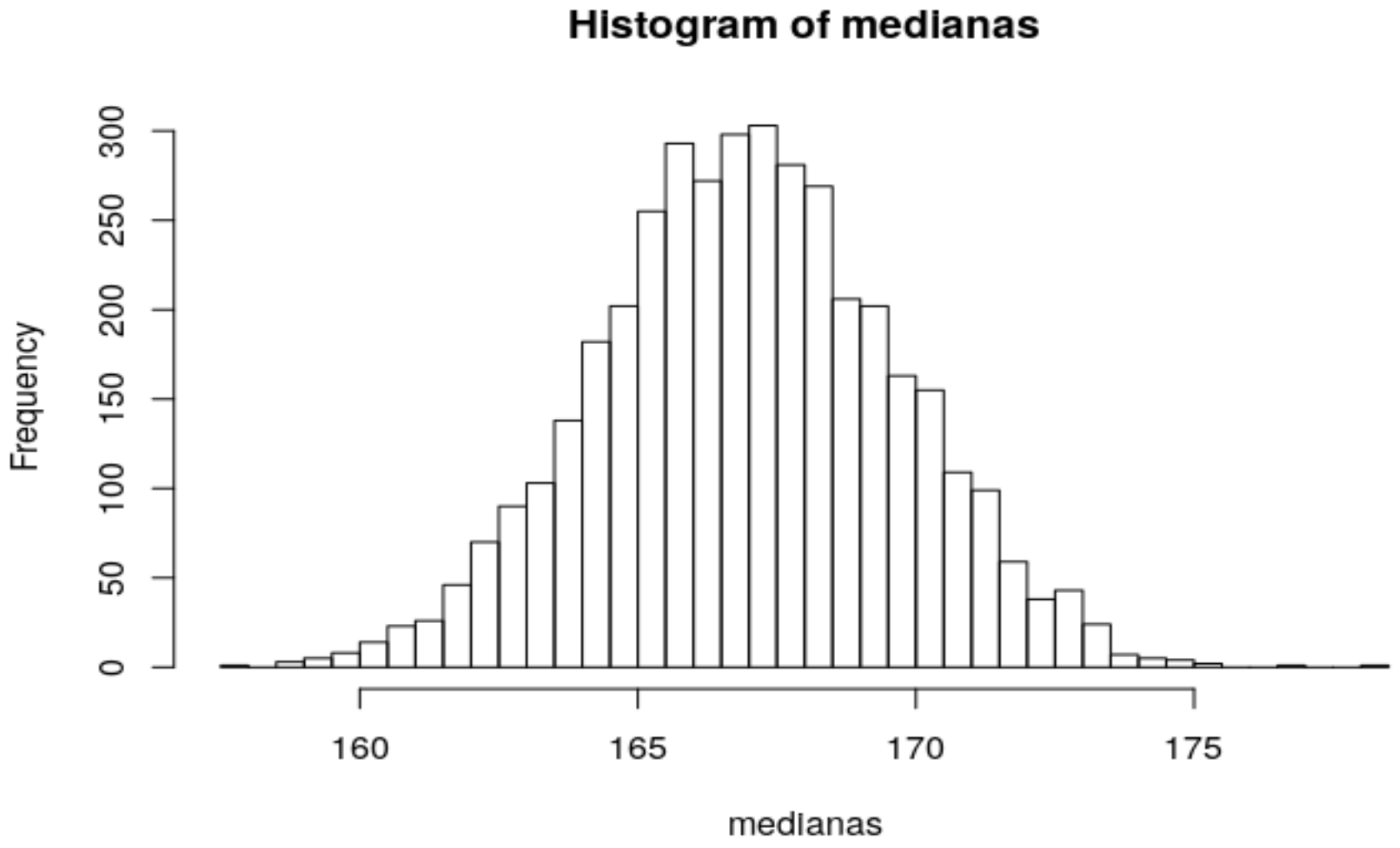
7. Distribuições amostrais

- Exemplo 10.8 (cont):



7. Distribuições amostrais

- Exemplo 10.8 (cont):



7. Distribuições amostrais

- Problemas

4. Usando os dados da Tabela 10.2, construa a distribuição amostral da estatística $\hat{\sigma}^2 = \dfrac{\sum (X_i - \overline{X})^2}{n}$.

5. No Problema 3, se X indicar o número de filhos na população, X_1 o número de filhos observados na primeira extração e X_2 na segunda:
   (a) calcule a média e a variância de X;
   (b) calcule $E(X_i)$ e $\text{Var}(X_i)$, $i = 1, 2$;
   (c) construa a distribuição amostral de $\overline{X} = \dfrac{(X_1 + X_2)}{2}$;
   (d) calcule $E(\overline{X})$ e $\text{Var}(\overline{X})$;
   (e) faça num mesmo gráfico os histogramas de X e de $\overline{X}$;
   (f) construa as distribuições amostrais de $S^2 = \sum_{i=1}^{2} (X_i - \overline{X})^2$ e $\hat{\sigma}^2 = \sum_{i=1}^{2} (X_i - \overline{X})^2/2$;
   (g) baseado no resultado de (f), qual dos dois estimadores você usaria para estimar a variância de X? Por quê?
   (h) calcule $P(|\overline{X} - \mu| > 1)$.

7. Distribuições amostrais

- Problemas

6. Ainda com os dados do Problema 3, e para amostras de tamanho 3:
   (a) determine a distribuição amostral de $\overline{X}$ e faça o histograma;
   (b) calcule a média e variância de $\overline{X}$;
   (c) calcule $P(|\overline{X} - \mu| > 1)$.
   (d) se as amostras fossem de tamanho 4, a $P(|\overline{X} - \mu| > 1)$ seria maior ou menor do que a probabilidade encontrada em (c)? Por quê?

8. Distribuição amostral da média

- Estudaremos a distribuição amostral da estatística $\bar{X}$, a média da amostra
- Consideremos uma população identificada pela variável X, cujos parâmetros média populacional $\mu = E(X)$ e variância populacional $\sigma^2 = Var(X)$ são supostamente conhecidos
- Vamos retirar todas as possíveis AAS de tamanho *n* dessa população, e para cada uma calcular a média $\bar{X}$
- Em seguida, consideremos a distribuição amostral e estudemos suas propriedades

8. Distribuição amostral da média

- Exemplo 10.10:
  - Voltemos ao Exemplo 10.7
  - A população {1,3,5,5,7} tem média $\mu = 4{,}2$ e variância $\sigma^2 = 4{,}16$. A distribuição amostral de $\bar{X}$ está na Tabela 10.3, da qual obtemos

$$E(\bar{X}) = \sum_i \bar{x}_i\, p_i = 1 \times \frac{1}{25} + 2 \times \frac{2}{25} + 3 \times \frac{5}{25} + 4 \times \frac{6}{25} + 5 \times \frac{6}{25}$$

$$+ 6 \times \frac{4}{25} + 7 \times \frac{1}{25} = 4{,}2.$$

  - Analogamente, encontramos $Var(\bar{X}) = \sum_i p_i\,(\bar{x}_i - E(\bar{X}))^2 = \sum_i p_i\,\bar{x}_i^{\,2} - E(\bar{X})^2$, resultando em

$$\mathrm{Var}(\bar{X}) = 2{,}08.$$

8. Distribuição amostral da média

- Exemplo 10.10:
  - Verificamos dois fatos:
    - A média das médias amostrais coincide com a média populacional
    - A variância de $\bar{X}$ é igual à variância de *X* dividida pelo tamanho da amostra ($n = 2$)
  - Esses fatos não são casos isolados.
  - O resultado a seguir mostra que isso vale no caso geral

8. Distribuição amostral da média

Teorema 10.1. Seja X uma v.a. com média μ e variância σ^2, e seja $(X_1, \ldots, X_n)$ uma AAS de X. Então,

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

Prova. Pelas propriedades vistas no Capítulo 8, temos:

$$E(\bar{X}) = (1/n)\{E(X_1) + \ldots + E(X_n)\}$$

$$= (1/n)\{\mu + \mu + \ldots + \mu\} = n\mu/n = \mu.$$

De modo análogo, e pelo fato de $X_1, \ldots, X_n$ serem independentes, temos

$$\text{Var}(\bar{X}) = (1/n^2)\{\text{Var}(X_1) + \ldots + \text{Var}(X_n)\}$$

$$= (1/n^2)\{\sigma^2 + \ldots + \sigma^2\} = n\sigma^2/n^2 = \sigma^2/n.$$

Obs: As propriedades mencionadas na prova acima são que, se $X_1, \ldots, X_n$ são variáveis aleatórias independentes, então a média de suas somas é igual à soma de suas médias, e a variância de suas variâncias é igual à soma de suas variâncias

8. Distribuição amostral da média

- O desvio padrão da distribuição amostral de uma estatística $T = r(X_1, X_2, \ldots, X_n)$ é usualmente denominado **erro padrão**
  - Termo adotado para evitar confusão entre o desvio padrão de X e o desvio padrão de T
  - Por essa razão, é usual nos referirmos a

$$\sqrt{\sigma^2/n} \quad = \sigma/\sqrt{n}$$

(o desvio padrão de $\bar{X}$) como o *erro padrão* de $\bar{X}$

8. Distribuição amostral da média

- Exemplo 10.10 (cont):
  - Para a população {1,3,5,5,7}, vamos construir os histogramas das distribuições de $\bar{X}$ para $n = 1,2$ e 3

(i) Para $n = 1$, vemos que a distribuição de $\overline{X}$ coincide com a distribuição de X, com $E(\overline{X}) = E(X) = 4{,}2$ e $\text{Var}(\overline{X}) = \text{Var}(X) = 4{,}16$ (Figura 10.4(a)).

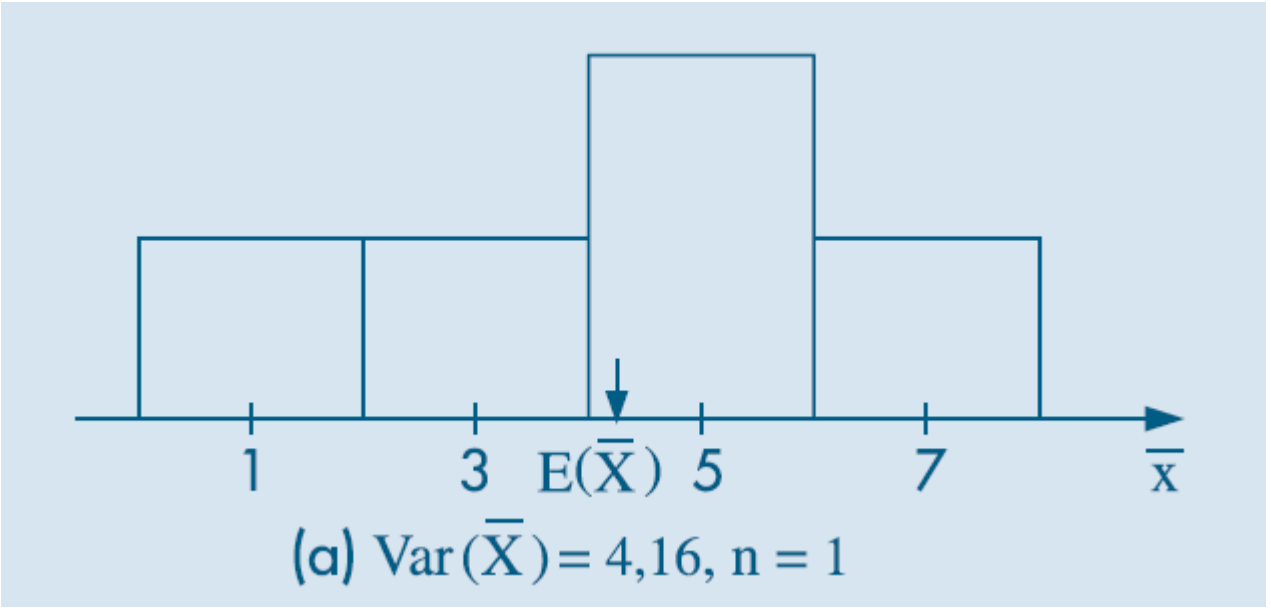
(ii) Para $n = 2$, baseados na Tabela 10.3, temos a distribuição de $\overline{X}$ dada na Figura 10.4(b), com $E(\overline{X}) = 4{,}2$ e $\text{Var}(\overline{X}) = 2{,}08$.

(iii) Finalmente, para $n = 3$, com os dados da Tabela 10.6, temos a distribuição de $\overline{X}$ na Figura 10.4 (c), com $E(\overline{X}) = 4{,}2$ e $\text{Var}(\overline{X}) = 1{,}39$.

8. Distribuição amostral da média

- Exemplo 10.10 (cont):

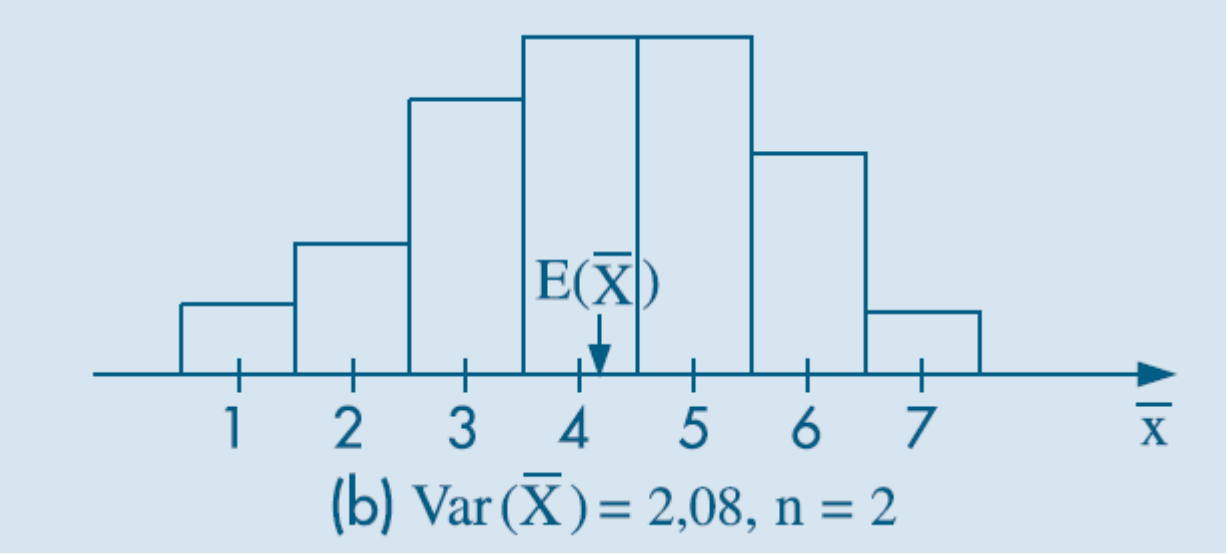
Figura 10.4: Distribuição de $\overline{X}$ para amostras de $\{1, 3, 5, 5, 7\}$.



8. Distribuição amostral da média

- Exemplo 10.10 (cont):

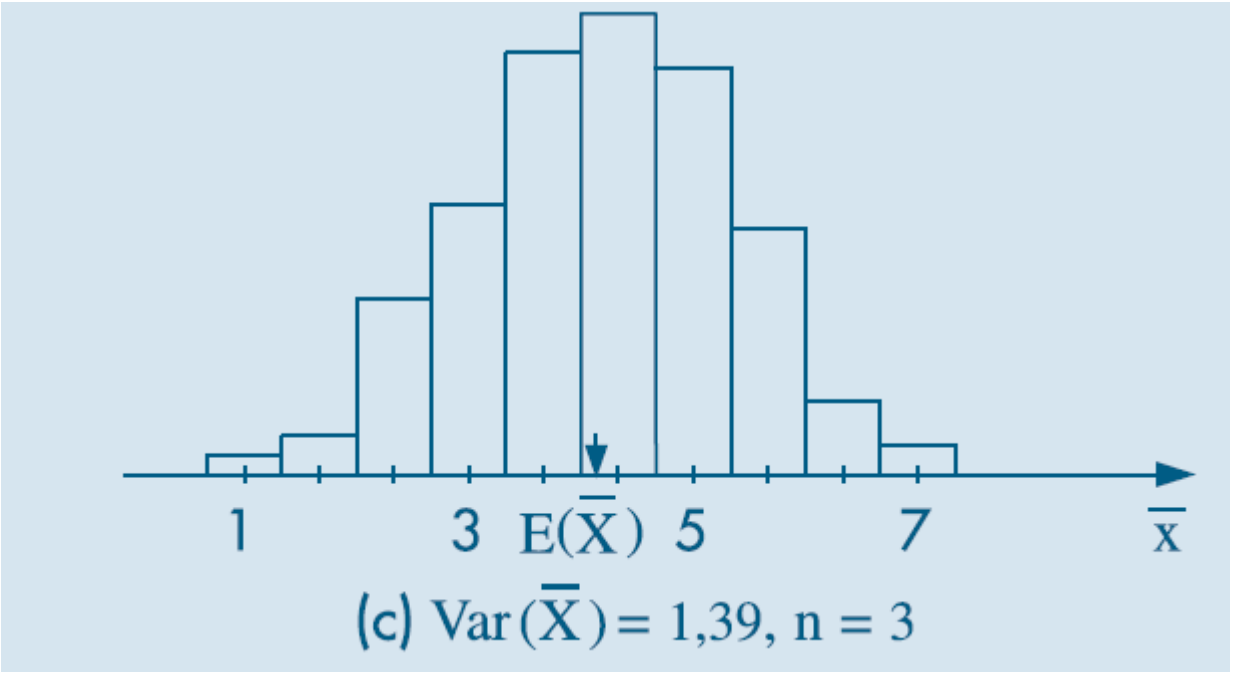
Figura 10.4: Distribuição de $\overline{X}$ para amostras de $\{1, 3, 5, 5, 7\}$.



8. Distribuição amostral da média

- Exemplo 10.10 (cont):

Figura 10.4: Distribuição de $\overline{X}$ para amostras de $\{1, 3, 5, 5, 7\}$.

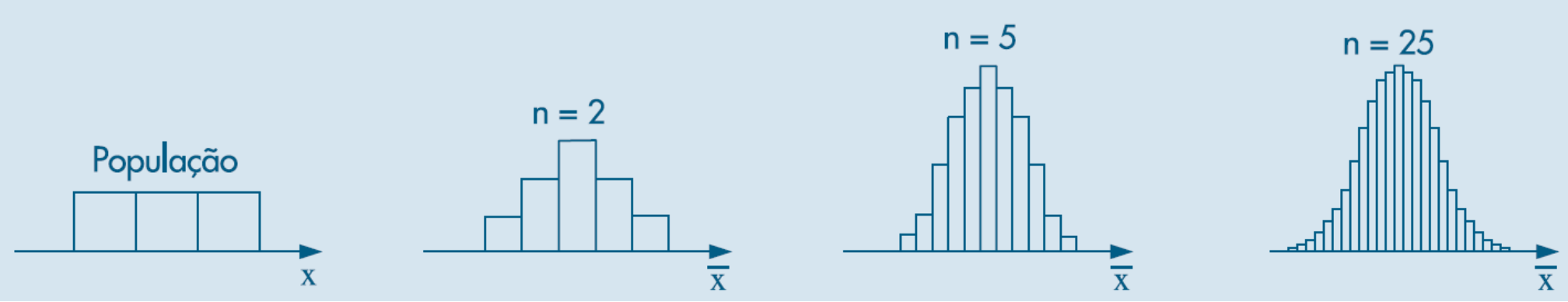


8. Distribuição amostral da média

- Exemplo 10.10 (cont):
  - Observe nos histogramas que, conforme o tamanho da amostra (*n*) vai aumentando, o histograma tende a se concentrar cada vez mais em torno de $\mathrm{E}(\bar{X}) = E(X) = 4.2$, já que a variância vai diminuindo.
  - Quando *n* for suficientemente grande, o histograma alisado aproxima-se de uma distribuição normal.
  - Essa aproximação pode ser verificada analisando-se os gráficos da Figura 10.5 (próximos slides), que mostram o comportamento do histograma de $\bar{X}$ para várias formas da distribuição da população e vários valores do tamanho da amostra *n*.

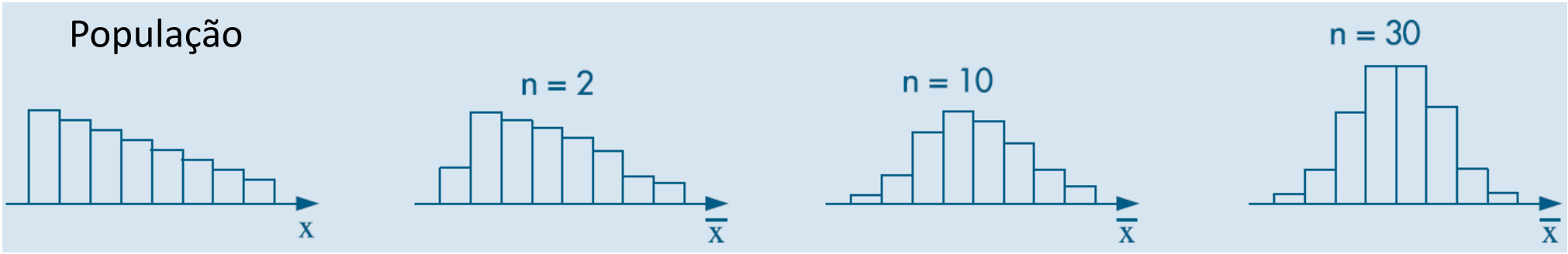
8. Distribuição amostral da média

Figura 10.5: Histogramas correspondentes às distribuições amostrais de $\overline{X}$ para amostras extraídas de algumas populações.



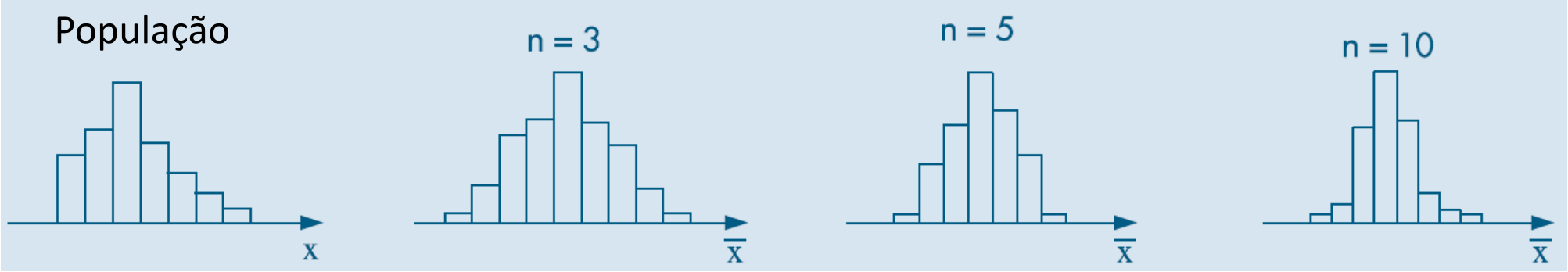
8. Distribuição amostral da média

Figura 10.5: Histogramas correspondentes às distribuições amostrais de $\overline{X}$ para amostras extraídas de algumas populações.



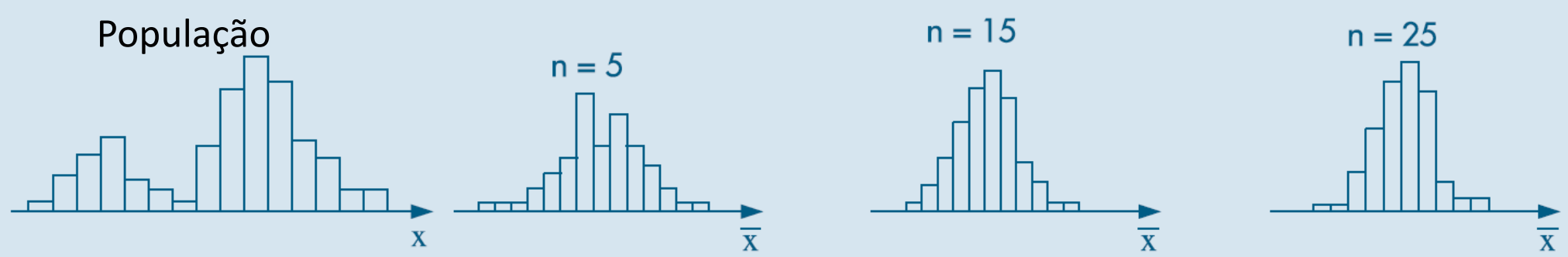
8. Distribuição amostral da média

Figura 10.5: Histogramas correspondentes às distribuições amostrais de $\overline{X}$ para amostras extraídas de algumas populações.



8. Distribuição amostral da média

Figura 10.5: Histogramas correspondentes às distribuições amostrais de $\overline{X}$ para amostras extraídas de algumas populações.



8. Distribuição amostral da média

- As observações empíricas nos gráficos anteriores de que $\bar{X}$ se aproxima de uma distribuição normal para valores grandes de *n* são consequência do *Teorema do Limite Central* (TLC), apresentado abaixo:

Teorema 10.2. (TLC) 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 $\bar{X}$ aproxima-se, para n grande, de uma distribuição normal, com média μ e variância σ^2/n.

- Embora não seja apresentada a demonstração desse teorema, o importante é saber como esse resultado pode ser usado

8. Distribuição amostral da média

- Exemplo 10.11 (Teorema do Limite Central):
  - Suponha que uma máquina empacotadora de café está regulada para encher pacotes cujos pesos *X* (em gramas) devem seguir uma $X \sim N(500{,}100)$ – média de 500g e desvio padrão de 10g
  - Denotamos por *X* o peso de um pacote enchido pela máquina
  - Suponha que nosso interesse seja avaliar, a partir de uma amostra de $n = 100$ pacotes, se essa máquina está regulada
  - Pelo Teorema do Limite Central, $\bar{X}$ deverá ter uma distribuição normal com média 500g e variância $\sigma^2/n = 100/100 = 1$, e portanto seu erro padrão será $\sigma/\sqrt{n} = 1\text{g}$.

8. Distribuição amostral da média

- Exemplo 10.11 (cont):
  - Queremos calcular a probabilidade de obtermos uma amostra de 100 pacotes com média diferindo de 500g por uma diferença menor que dois gramas:
    $P(|\bar{X} - 500| < 2) = P(-2 < \bar{X} - 500 < 2) = P(498 < \bar{X} < 502)$
  - Usamos a seguinte padronização sobre os limites e sobre $\bar{X}$:
    $Z_{498} = \frac{498-\mu}{\sigma/\sqrt{n}} = \frac{498-500}{1} = -2;$
    $Z_{502} = \frac{502-\mu}{\sigma/\sqrt{n}} = \frac{502-500}{1} = 2;$
    $Z_{\bar{X}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-500}{1};$
  - Assim, a probabilidade $P(Z_{498} < Z_{\bar{X}} < Z_{502}) = P(-2 < Z_{\bar{X}} < 2)$ pode ser calculada pela distribuição normal padrão:

8. Distribuição amostral da média

- Exemplo 10.11 (cont):
  - $P(-2 < Z_{\bar{X}} < 2) \approx 0.95$

  - Ou seja, a probabilidade de uma amostra de 100 pacotes ter uma média fora do intervalo (498, 502) é de aproximadamente 5%, considerada baixa (dependendo do critério adotado).
  - Se observarmos uma média fora desse intervalo, podemos considerar como um evento raro, e será razoável supor que a máquina esteja desregulada.

8. Distribuição amostral da média

Corolário 10.1. Se $(X_1, ..., X_n)$ for uma amostra aleatória simples da população X, com média μ e variância σ^2 finita, e $\bar{X} = (X_1 + ... + X_n)/n$, então

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1). \qquad (10.2)$$

– No corolário acima, basta notar que se usou a transformação de padronização de $\bar{X}$

- A variável aleatória $e = \bar{X} - \mu$ é chamada *erro amostral da média*; o resultado abaixo é imediato.

Corolário 10.2. A distribuição de e aproxima-se de uma distribuição normal com média 0 e variância σ^2/n, isto é,

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

8. Distribuição amostral da média

- O Teorema do Limite Central afirma que a distribuição de $\bar{X}$ aproxima-se da distribuição normal quando n tende a infinito
- A rapidez dessa convergência depende da distribuição original da população da qual a amostra é retirada (ver Figura 10.5)
  - Se a população original tem uma distribuição próxima da normal, a convergência é rápida
  - Se a população original se afasta muito de uma distribuição normal, a convergência é mais lenta
    - precisamos de amostras maiores
  - Na literatura, considera-se que, para amostras da ordem de 30 elementos ou mais, a aproximação pode ser considerada boa
    - Mas cuidado! o exemplo 10.12 (adiante) é um contra-exemplo

8. Distribuição amostral da média

- Problemas:

7. Uma v.a. X tem distribuição normal, com média 100 e desvio padrão 10.
   (a) Qual a $P(90 < X < 110)$?
   (b) Se $\overline{X}$ for a média de uma amostra de 16 elementos retirados dessa população, calcule $P(90 < \overline{X} < 110)$.
   (c) Represente, num único gráfico, as distribuições de X e $\overline{X}$.
   (d) Que tamanho deveria ter a amostra para que $P(90 < \overline{X} < 110) = 0{,}95$?

8. A máquina de empacotar um determinado produto o faz segundo uma distribuição normal, com média μ e desvio padrão 10 g.
   (a) Em quanto deve ser regulado o peso médio μ para que apenas 10% dos pacotes tenham menos do que 500 g?
   (b) Com a máquina assim regulada, qual a probabilidade de que o peso total de 4 pacotes escolhidos ao acaso seja inferior a 2 kg?

Dicas: Para o item (a), calcule $F^{-1}(0.1\,|\mu, \sigma = 10)$, para valores de $\mu = 500, 501, 502, \ldots$ onde F^{-1} denota a função quantil (inversa) da distribuição normal; escolha o menor valor de μ para o qual $F^{-1}(0.1 \mid \mu, \sigma = 10) \geq 500$; para o item (b), calcule $\Pr(\bar{X} < 500)$

8. Distribuição amostral da média

- Problemas:

9. No exemplo anterior, e após a máquina estar regulada, programou-se uma carta de controle de qualidade. De hora em hora, será retirada uma amostra de quatro pacotes e esses serão pesados. Se a média da amostra for inferior a $495\ \mathrm{g}$ ou superior a $520\ \mathrm{g}$, encerra-se a produção para reajustar a máquina, isto é, reajustar o peso médio.
   (a) Qual é a probabilidade de ser feita uma parada desnecessária?
   (b) Se o peso médio da máquina desregulou-se para $500\ \mathrm{g}$, qual é a probabilidade de continuar a produção fora dos padrões desejados?

10. A capacidade máxima de um elevador é de $500\ \mathrm{kg}$. Se a distribuição X dos pesos dos usuários for suposta $N(70, 100)$:
   (a) Qual é a probabilidade de sete passageiros ultrapassarem esse limite?
   (b) E seis passageiros?

9. Intervalos de confiança

- Intervalos de confiança são intervalos estatísticos baseados na distribuição amostral de um estimador pontual

- Exemplo 11.12:
  - Suponha que queiramos estimar a média μ de uma população qualquer, e para tanto usamos a média $\bar{X}$ de uma amostra de tamanho *n*. Do TLC,

$$e = (\bar{X} - \mu) \sim N(0, \sigma^2_{\bar{x}}), \qquad (11.32)$$

  com $Var(\bar{X}) = \sigma_{\bar{X}}{}^2 = \sigma^2/n$.

  - Desse resultado podemos determinar a probabilidade de cometermos erros de determinadas magnitudes

9. Intervalos de confiança

- Exemplo 11.12 (cont):
  - Por exemplo,

$$P(|e| < 1{,}96\sigma_{\bar{X}}) = 0{,}95$$

ou

$$P(|\bar{X} - \mu| < 1{,}96\sigma_{\bar{X}}) = 0{,}95,$$

que é equivalente a

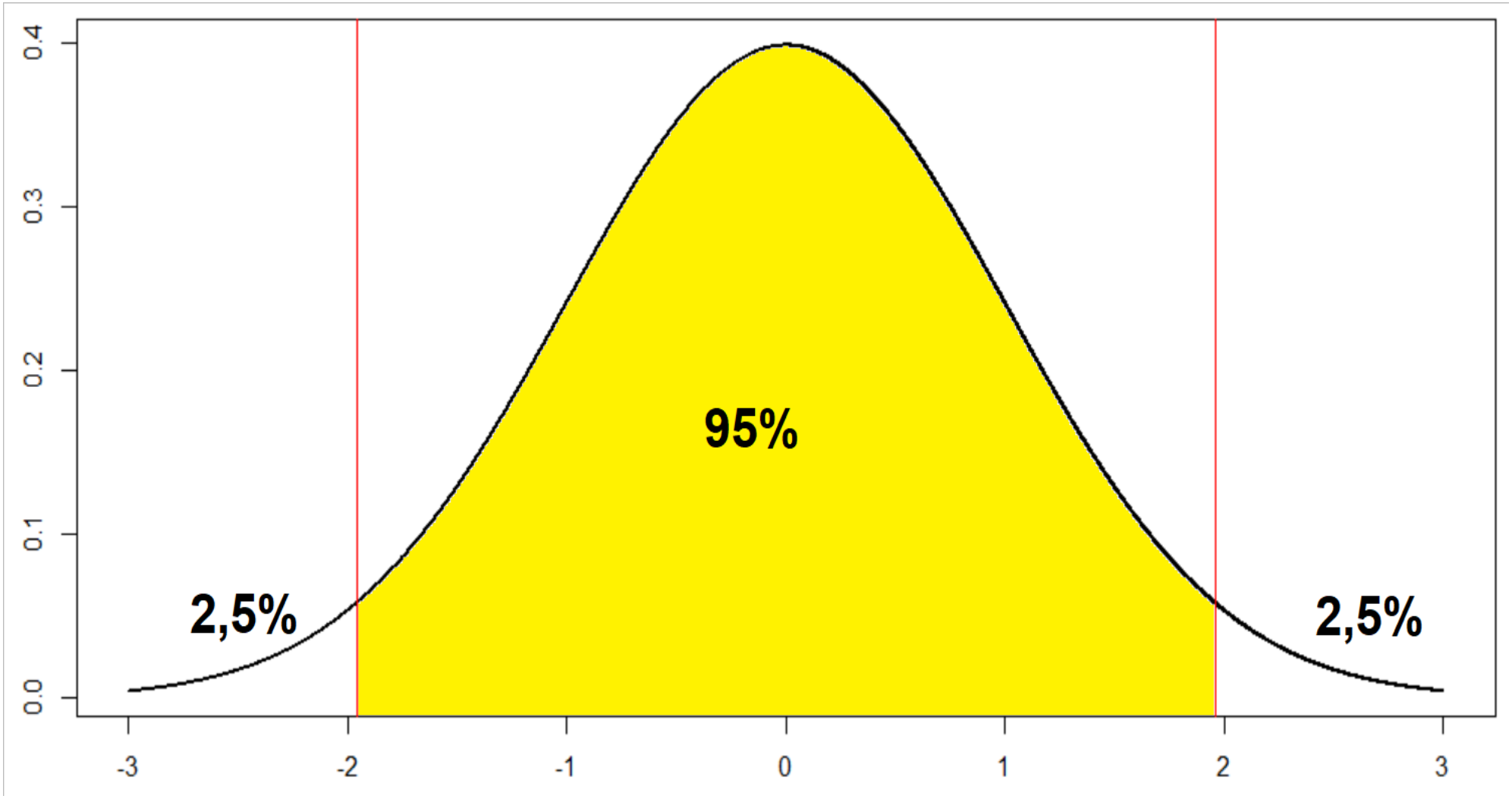
$$P(-1{,}96\sigma_{\bar{X}} < \bar{X} - \mu < 1{,}96\sigma_{\bar{X}}) = 0{,}95,$$

e, finalmente,

$$P(\bar{X} - 1{,}96\sigma_{\bar{X}} < \mu < \bar{X} + 1{,}96\sigma_{\bar{X}}) = 0{,}95. \qquad (11.33)$$

9. Intervalos de confiança

- Exemplo 11.12 (cont):



9. Intervalos de confiança

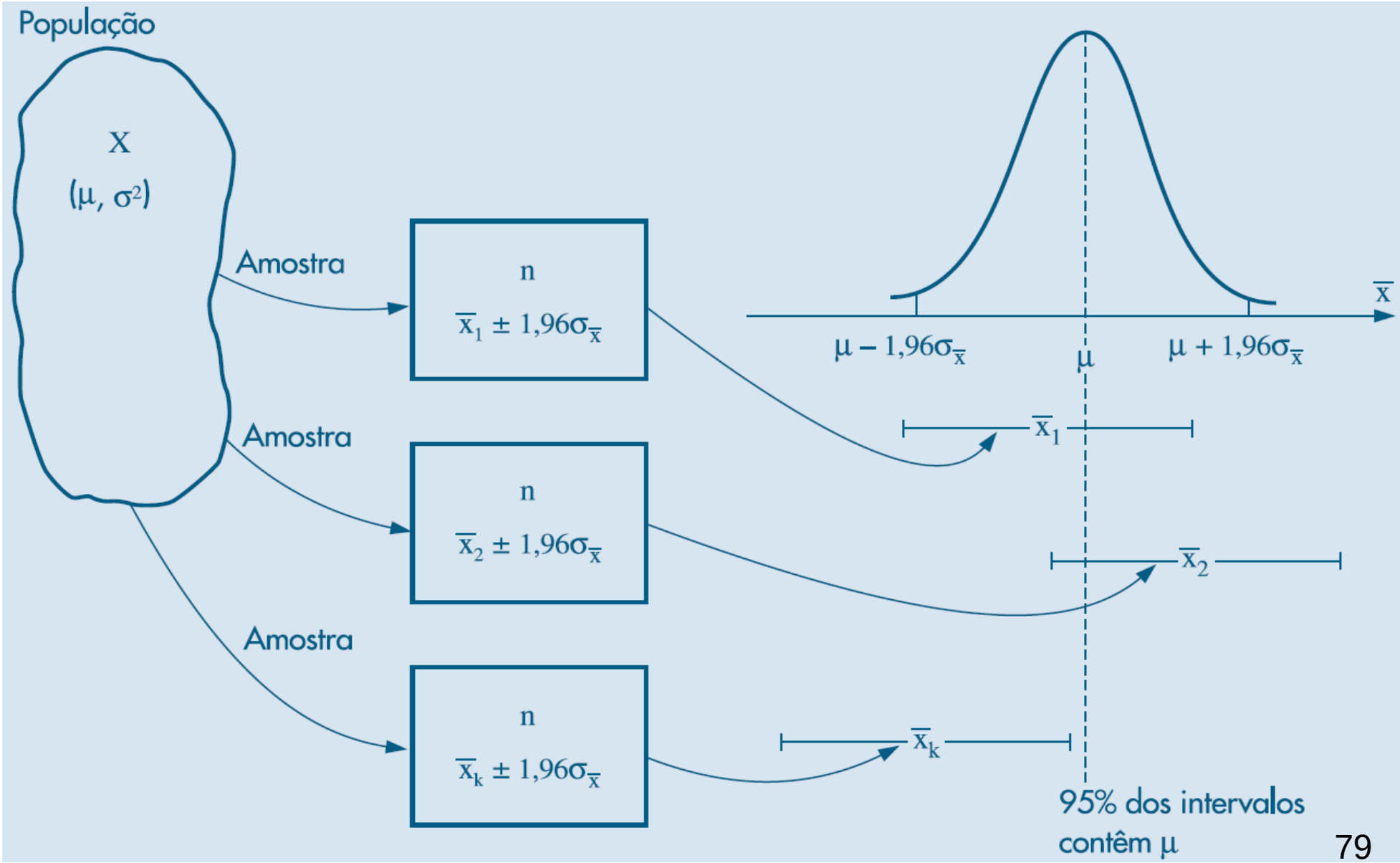
- É importante lembrar que, sob o ponto de vista frequentista, μ não é uma variável aleatória e sim um parâmetro, e a expressão (11.33) deve ser interpretada da seguinte maneira:
- Se pudéssemos construir uma quantidade grande de intervalos (aleatórios) da forma $(\bar{X} - 1.96\,\sigma_{\bar{X}}\,, \bar{X} + 1.96\,\sigma_{\bar{X}})$, todos baseados em amostras de tamanho *n*, 95% deles conteriam o parâmetro μ.
  - Figura 11.13 mostra o funcionamento e o significado de um intervalo de confiança (IC) para μ, com $\gamma = 0.95$ e σ^2 conhecido
- Dizemos que γ é o **coeficiente de confiança**
- Escolhida uma amostra e encontrada sua média $\bar{x}_0$, e admitindo-se σ^2 conhecido, podemos construir o intervalo

$$(\bar{x}_0 - 1.96\,\sigma_{\bar{X}}\,, \bar{x}_0 + 1.96\,\sigma_{\bar{X}}) \qquad (11.34)$$

- Esse intervalo pode ou não conter o parâmetro μ, mas pelo exposto acima temos 95% de confiança de que contenha.

9. Intervalos de confiança

- Figura 11.3: significado de um IC para μ, com $\gamma = 0.95$



9. Intervalos de confiança

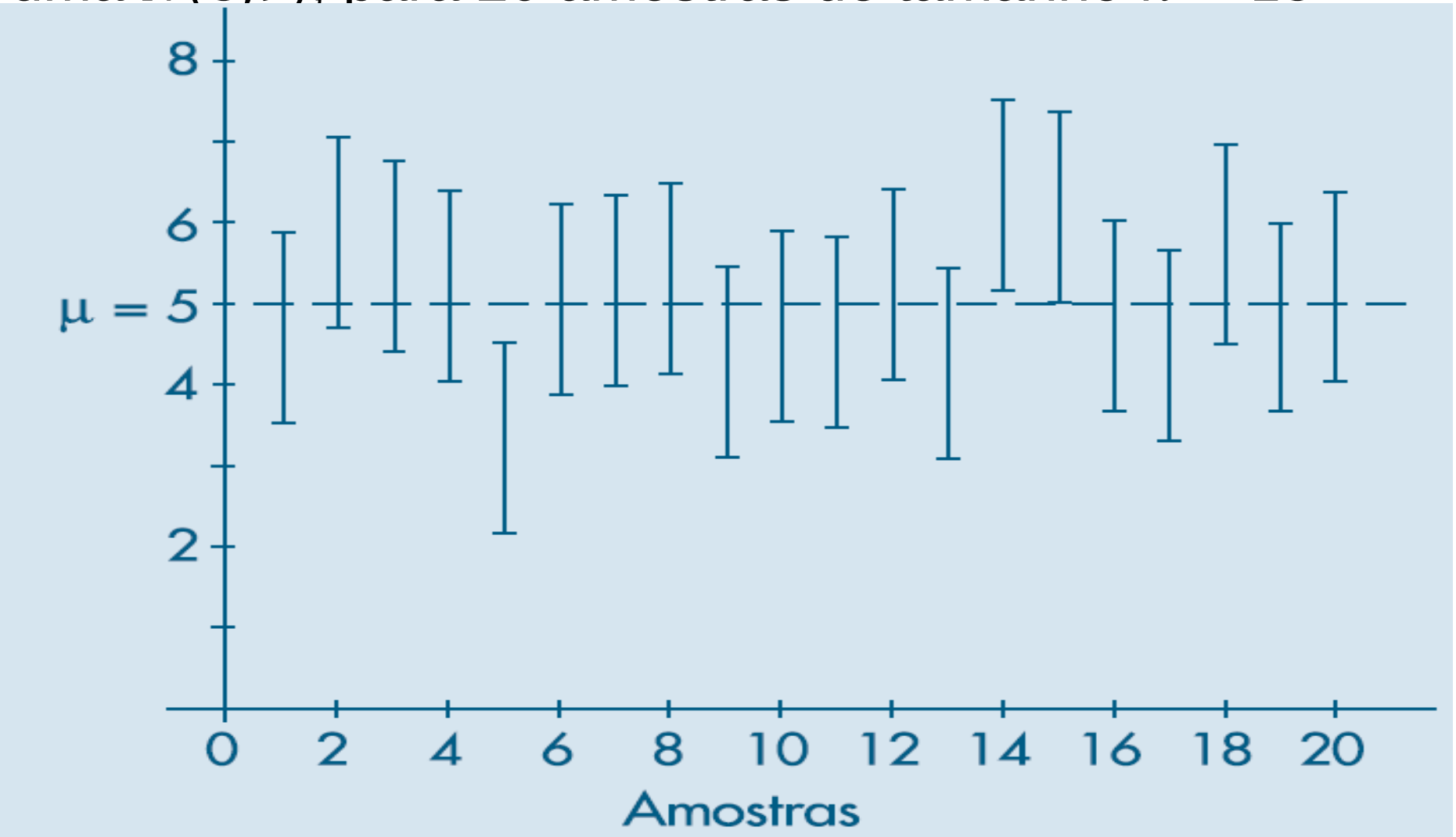
- Para ilustrar, consideremos o seguinte experimento de simulação:
  - Foram geradas 20 amostras de tamanho $n = 25$ de uma distribuição normal com média $\mu = 5$ e desvio padrão $\sigma = 3$
  - Para cada amostra foi construído o intervalo de confiança para μ, com coeficiente de confiança $\gamma = 0.95$:

$$\sigma_{\bar{X}} = \frac{3}{\sqrt{25}} = \frac{3}{5} = 0.6 \Longrightarrow 1.96\ \sigma_{\bar{X}} = 1.176$$

    ou seja, cada intervalo é da forma $\bar{X} \pm 1.176$
  - Figura 11.4 apresenta os intervalos obtidos, dentre os quais 3 intervalos (amostras 5, 14 e 15) não contêm a média $\mu = 5$
- Exercício:
  - Repetir a simulação descrita acima, com 1000 amostras de tamanho $n = 25$ e demais parâmetros similares ao exemplo
  - Calcular a proporção de intervalos contendo a média $\mu = 5$

9. Intervalos de confiança

- Figura 11.4: Intervalos de confiança para a média de uma $N(5,9)$, para 20 amostras de tamanho $n = 25$



9. Intervalos de confiança

- Usando a aproximação pela distribuição normal (Eq. 11.34), para um coeficiente de confiança qualquer γ, teremos que usar o valor z_γ tal que $\mathrm{P}(-z_\gamma \leq Z \leq z_\gamma) = \gamma$, com $Z \sim N(0,1)$. O valor de z_γ é calculado por

$$z_\gamma = -\Phi^{-1}\left(\frac{1-\gamma}{2}\right)$$

  - Outra forma de calcular z_γ: $\alpha = (1-\gamma); z_\gamma = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$

- O intervalo fica então:

$$\left(\bar{X} - z_\gamma\, \sigma_{\bar{X}}\, , \bar{X} + z_\gamma\, \sigma_{\bar{X}}\right) \qquad (11.37)$$

- Observe que a amplitude do intervalo (11.37) é $L = 2z_\gamma\, \sigma/\sqrt{n}$, que é uma constante, independente de $\bar{X}$
  - Se construirmos vários intervalos de confiança para o mesmo valor de n, σ e γ, estes terão extremos aleatórios, mas todos terão a mesma amplitude L

9. Intervalos de confiança

- Se o desvio padrão populacional σ não for conhecido, podemos substituir em (11.37) $\sigma_{\bar{X}}$ pelo desvio padrão amostral s, dado por:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\,, \qquad s = \sqrt{s^2}$$

Assim, podemos estimar $\sigma_{\bar{X}}$ por:

$$\hat{\sigma}_{\bar{X}} = {}^{s}/_{\sqrt{n}}$$

  - Para n grande, da ordem de 100 ou superior, o intervalo (11.37) pode ser usado com $\hat{\sigma}_{\bar{X}}$ no lugar de $\sigma_{\bar{X}}$
    - Em simulação, utiliza-se normalmente $n \gg 100$, e portanto essa aproximação pode ser usada sem problemas.
    - Quando n é pequeno, é mais adequado construir o intervalo de confiança usando a distribuição t de Student ao invés da distribuição normal padrão (assunto não abordado nesta disciplina, pela razão exposta no item acima).

11. Determinação do tamanho de uma amostra

- Em nossas considerações anteriores, fizemos a suposição de que o tamanho da amostra, n, era conhecido e fixo
- Podemos, em certas situações, 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
- Por exemplo, suponha que estejamos estimando a média μ populacional e para tanto usaremos a média amostral, $\bar{X}$, baseada numa amostra de tamanho n

11. Determinação do tamanho de uma amostra

- Suponha que se queira determinar o valor de n de modo que

$$P(|\bar{X} - \mu| \leqslant \varepsilon) \geqslant \gamma, \qquad (10.5)$$

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

- Pelo TLC, podemos considerar que $\bar{X} \sim N(\mu, \sigma^2/n)$, logo $\bar{X} - \mu \sim N(0, \sigma^2/n)$, e portanto (10.5) pode ser escrita

$$P(-\varepsilon \leqslant \bar{X} - \mu \leqslant \varepsilon) = P\left(\frac{-\sqrt{n}\varepsilon}{\sigma} \leqslant Z \leqslant \frac{\sqrt{n}\varepsilon}{\sigma}\right) \approx \gamma,$$

com $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$

11. Determinação do tamanho de uma amostra

- Dado γ, podemos obter z_γ da $N(0,1)$, tal que $\mathrm{P}\left(-z_\gamma \leq Z \leq z_\gamma\right) = \gamma$, de modo que

$$\frac{\sqrt{n}\,\varepsilon}{\sigma} = z_\gamma,$$

do que obtemos finalmente

$$n = \frac{\sigma^2 z_\gamma^2}{\varepsilon^2}. \qquad (10.6)$$

- Veremos a seguir como obter z_γ

11. Determinação do tamanho de uma amostra

- Obtendo z_γ:
  - Consideremos a restrição

$$\mathrm{P}\left(-z_\gamma \leq Z \leq z_\gamma\right) = \gamma \qquad (1)$$

  - Por outro lado, note que:

$$\mathrm{P}\left(-z_\gamma \leq Z \leq z_\gamma\right) = 1 - \left[P\left(Z < -z_\gamma\right) + P\left(Z > z_\gamma\right)\right] =$$
$$= 1 - 2P\left(Z < -z_\gamma\right) = 1 - 2\Phi\left(-z_\gamma\right) \qquad (2)$$

    onde Φ é a função de distribuição acumulada normal padrão; a 3ª igualdade decorre da simetria da distribuição Normal em torno da média $\mu = 0$ (ver figura no próximo slide)

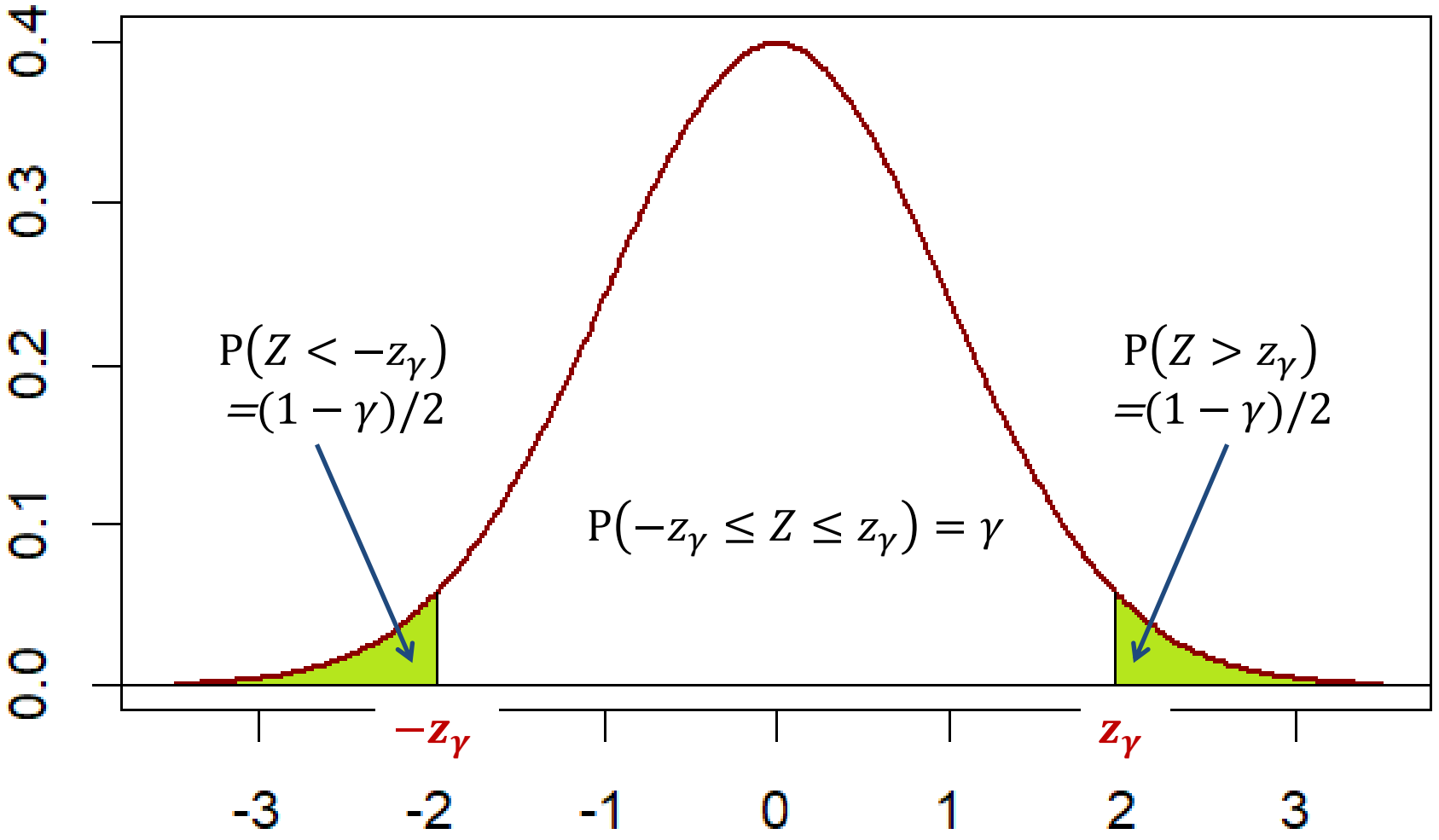
  - Unindo as duas igualdades (1) e (2):

$$1 - 2\Phi\left(-z_\gamma\right) = \gamma \Longrightarrow \Phi\left(-z_\gamma\right) = \frac{1-\gamma}{2} \Longrightarrow z_\gamma = -\Phi^{-1}\left(\frac{1-\gamma}{2}\right)$$

  - P.ex: para $\gamma = 0.95$, $z_\gamma = -\Phi^{-1}(0.025) = -(-1.96) = 1.96$

11. Determinação do tamanho de uma amostra

- Obtendo z_γ (cont):



11. Determinação do tamanho de uma amostra

- Note que em (10.6) conhecemos z_γ e ε, mas σ^2 é a variância desconhecida da população
- Para podermos ter uma ideia sobre n devemos ter alguma informação prévia sobre σ^2 ou, então, usar uma pequena amostra para estimar este parâmetro

11. Determinação do tamanho de uma amostra

- Exemplo 10.13:
  - Suponha que uma pequena amostra piloto de $n = 10$, extraída de uma população, forneceu os valores $\bar{X} = 15$ e

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 16.$$

  - Fixando-se $\varepsilon = 0.5$ e $\gamma = 0.95$, temos

$$n = \frac{16 \times (1{,}96)^2}{(0{,}5)^2} \approx 245.$$

11. Determinação do tamanho de uma amostra

- Tamanho amostral para proporções:
- No caso de proporções, usando a aproximação normal para $\hat{p}$, é fácil ver que (10.6) resulta

$$n = \frac{z_\gamma^2 p(1-p)}{\varepsilon^2} . \qquad (10.7)$$

- Como não conhecemos *p*, a verdadeira proporção populacional, podemos usar o fato de que $p(1-p) \leq 1/4$, para todo *p*, e portanto (10.7) fica

$$n \approx \frac{z_\gamma^2}{4\varepsilon^2} . \qquad (10.8)$$

- Por outro lado, se tivermos alguma informação sobre p ou pudermos estimá-lo usando uma amostra piloto, podemos substituir esse valor estimado em (10.7)

11. Determinação do tamanho de uma amostra

- Exemplo 10.14:
  - Suponha que numa pesquisa de mercado estima-se que no mínimo 60% das pessoas entrevistadas preferirão a marca A de um produto – informação baseada em pesquisas anteriores
  - Determine o tamanho da amostra, n, tal que o erro amostral de $\hat{p}$ seja no máximo menor do que $\varepsilon = 0.03$ com probabilidade $\gamma = 0.95$.
  - **Resposta 1** (aproximação pela normal):
    - Se quisermos que o erro amostral de $\hat{p}$ seja menor do que $\varepsilon = 0.03$, com probabilidade $\gamma = 0.95$, teremos

$$n \approx \frac{\left(z_\gamma\right)^2 p(1-p)}{\varepsilon^2} = \frac{(1.96)^2(0.6)(0.4)}{(0.03)^2} = \mathbf{1024}$$

    - Como sabe-se que $p \geq 0.6$, na equação acima usamos a igualdade $p = 0.6$, que resulta na maior variância possível, e consequentemente no maior valor de n (tamanho de amostra mais conservador)

11. Determinação do tamanho de uma amostra

- Exemplo 10.14 (cont):
  - **Resposta 2** (usando a distribuição binomial):
    - Não há solução analítica para calcular n, mas pode-se testar diversos valores de n_1, n_2, n_3, ... e determinar o menor n_j tal que $\mathrm{P}(\mathrm{p} - \varepsilon \leq \hat{p} \leq p + \varepsilon) \geq \gamma$.
    - **Procedimento:**

    1. Fixe $n_1 = 10$, $n_2 = 20$, $n_3 = 30$, etc (intervalos de 10 em 10)
    2. Para cada n_j, calcule:
       - $y_{min} = (p - \varepsilon)n_j$, $y_{max} = (p + \varepsilon)n_j$
       - $P_{n_j}(y_{min} \leq Y \leq y_{max}) = F\left(y_{max}|n_j, p\right) - F\left(y_{min}(0.999)|n_j, p\right)$ onde $F\left(y|n_j, p\right)$ denota a função de distribuição acumulada da binomial com parâmetros (n_j, p)
    3. Escolha o menor dentre os n_j tal que $P_{n_j}(y_{min} \leq Y \leq y_{max}) \geq \gamma$

       » Excel: pode-se usar a função `PROCV`

    - Em nosso exemplo, o valor encontrado foi $n = \mathbf{1000}$

11. Determinação do tamanho de uma amostra

Problemas

17. Suponha que uma indústria farmacêutica deseja saber a quantos voluntários se deva aplicar uma vacina, de modo que a proporção de indivíduos imunizados na amostra difira de menos de 2% da proporção verdadeira de imunizados na população, com probabilidade 90%. Qual o tamanho da amostra a escolher? Use (10.8).

18. No problema anterior, suponha que a indústria tenha a informação de que a proporção de imunizados pela vacina seja $p \geq 0{,}80$. Qual o novo tamanho de amostra a escolher? Houve redução?

Dicas: Para a questão 17, como não há informação sobre a proporção de imunizados pela vacina (p), assuma p=0.50

10.13 Problemas e complementos

21. Uma v.a. X tem distribuição normal com média 10 e desvio padrão 4. Aos participantes de um jogo é permitido observar uma amostra de qualquer tamanho e calcular a média amostral. Ganha um prêmio aquele cuja média amostral for maior que 12.

    (a) Se um participante escolher uma amostra de tamanho 16, qual é a probabilidade de ele ganhar um prêmio?

    (b) Escolha um tamanho de amostra diferente de 16 para participar do jogo. Qual é a probabilidade de você ganhar um prêmio?

    (c) Baseado nos resultados acima, qual o melhor tamanho de amostra para participar do jogo?

    (d) Construa um gráfico em curva da probabilidade de ganho do prêmio em função do tamanho da amostra

24. A distribuição dos comprimentos dos elos da corrente de bicicleta é normal, com média 2 cm e variância $0{,}01$ cm^2. Para que uma corrente se ajuste à bicicleta, deve ter comprimento total entre 58 e 61 cm.

    (a) Qual é a probabilidade de uma corrente com 30 elos não se ajustar à bicicleta?

    (b) E para uma corrente com 29 elos?

    [Observação: suponha que os elos sejam selecionados ao acaso para compor a corrente, de modo que se tenha independência.]

10.13 Problemas e complementos

26. Um professor dá um teste rápido, constante de 20 questões do tipo certo-errado. Para testar a hipótese de o estudante estar adivinhando a resposta, ele adota a seguinte regra de decisão: "Se 13 ou mais questões estiverem corretas, ele não está adivinhando". Qual é a probabilidade de rejeitarmos a hipótese, sendo que na realidade ela é verdadeira?

27. Um distribuidor de sementes determina, por meio de testes, que 5% das sementes não germinam. Ele vende pacotes com 200 sementes com garantia de 90% de germinação. Qual é a probabilidade de que um pacote não satisfaça à garantia?

32. Distribuição amostral da diferença de duas médias. Consideremos duas populações X com parâmetros μ_1 e σ_1^2 e Y com parâmetros μ^2 e σ_2^2. Sorteiam-se duas amostras independentes: a da primeira população de tamanho n e a da segunda de tamanho m. Calculam-se as médias amostrais $\overline{X}$ e $\overline{Y}$.
    (a) Qual a distribuição amostral de $\overline{X}$? E de $\overline{Y}$? (Usando o TLC)
    (b) Defina $D = \overline{X} - \overline{Y}$. O que você entende por distribuição amostral de D?
    (c) Calcule $E(D)$ e $\mathrm{Var}(D)$.
    (d) Como você acha que será a distribuição de D? Por quê?