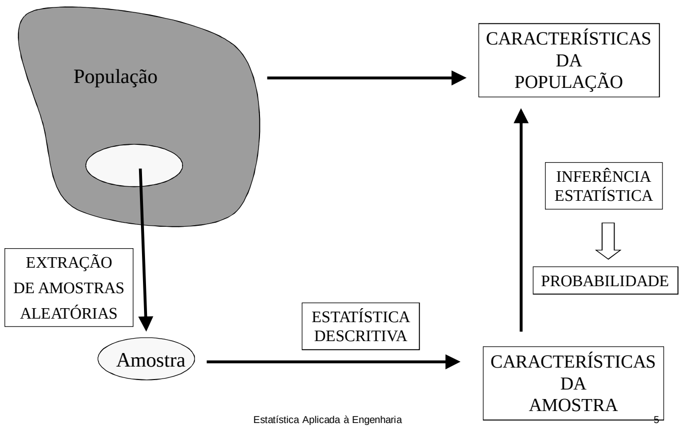
Distribuição Amostral e Estimação Pontual de Parâmetros

Roteiro

Introdução

População e Amostra

- População:
  - √ Conjunto de elementos que apresentam pelo menos uma característica em comum
- População Alvo:
  - √ População de interesse da pesquisa
- Amostra:
  - √ Qualquer subconjunto não vazio da população



Técnicas de Amostragem

- Procedimento a ser adotado na seleção dos elementos da amostra
- O principal objetivo central é obter uma amostra representativa
  - √ Amostra que representa toda a população da melhor maneira possível
- A representatividade depende de:
  - √ Metodologia adotada para seleção da amostra
  - √ Tamanho da amostra

Problema Fundamental da Estatística

A partir da observação de amostras, COMO podemos tirar CONCLUSÕES sobre a POPULAÇÃO ?

Erro Amostral

- Diferença entre um resultado amostral e o verdadeiro resultado populacional
  - √ são resultantes de flutuações amostrais aleatórias.

Erro Não–amostral

- Incorreção na coleta, registro ou análise de dados amostrais
  - √ Ex.
    - Coleta tendenciosa de amostra
    - Utilização de instrumento descalibrado
    - Registro incorreto de dados amostrais

Inferência Estatística

- Definição:
  - √ Procedimentos generalizar características de população a partir da informação contida na amostra.
- Baseia-se na Teoria de Probabilidades
- Áreas:
  - √ Estimação de parâmetros
  - √ Testes de hipóteses.

Estimação de Parâmetros

- Estimação Pontual
- Estimação Intervalar
  - √ Intervalos de Confiança

Teste de Hipóteses

- Hipótese:
  - √ Afirmação (alegação) sobre característica populacional
- Teste de Hipóteses:
  - √ Procedimento padrão para se testar uma afirmativa sobre característica populacional

Conceitos Fundamentais

- Amostra aleatória:
  - √ As variáveis aleatórias $X_1, X_2, ..., X_n$ são uma amostra aleatória de tamanho *n*, se:
    - Forem independentes
    - Cada X_i tiver mesma distribuição de probabilidades

- Parâmetro:
  - √ Quantidades de interesse da população
  - √ Em geral, desconhecidas
    - Média de uma população (μ)
    - Desvio-padrão de uma população (σ)
  - √ Representadas por letras gregas
  - √ Notação para estimador qualquer: θ

- Estatística:
  - √ Qualquer função da amostra que não dependa de parâmetros desconhecidos
  - √ Exemplo : Algumas estatísticas da amostra aleatória $X_1, X_2, ..., X_n$:

$X_{(1)} = \text{mín}(X_1, X_2, ..., X_n)$

$X_{(n)} = \text{máx}(X_1, X_2, ..., X_n)$

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

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

- Distribuição amostral:
  - √ Distribuição de probabilidades de uma estatística
  - √ Exemplo:
    - Distribuição amostral da média
    - Parâmetros da distribuição amostral da média

$$\mathrm{E}\left(\overline{X}_n\right) = \mu_X$$

$$\mathrm{Var}\left(\overline{X}_n\right) = \frac{\sigma_X^2}{n}$$

- Estimador de θ
  - √ Qualquer estatística que assuma valores em Θ.
  - √ Notação: $\hat{\theta}$
  - √ Exemplo:

    Alguns estimadores para a média μ de uma população
    - Média da amostra
    - Mediana da amostra
    - X_1
    - Etc.

- Estimativa de parâmetro populacional:
  - √ é um valor específico, ou um intervalo de valores, usado para estimar parâmetro populacional
- Estimativa pontual:
  - √ é um único valor numérico de uma estatística $\hat{\theta}$

Teorema Central do Limite

Teorema Central do Limite

- Seja X_1, X_2, ..., X_n uma amostra aleatória de tamanho *n* de uma população (finita ou infinita), com média μ e variância finita σ^2. Então

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

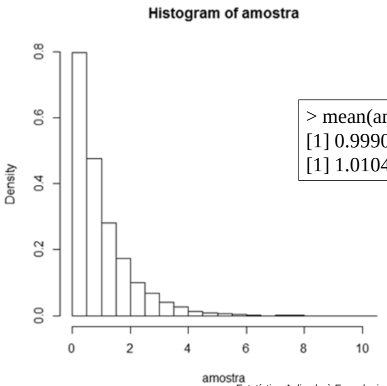
quando $n \to \infty$.

- Comentários:
  - √ A aproximação normal para a média amostral depende do tamanho da amostra
  - √ Com população contínua, unimodal e simétrica, na maioria dos casos, o TCL trabalha bem para pequenas amostras (n = 4, 5).
  - √ Em muitos casos de interesse prático, a aproximação normal será satisfatória para $n \geq 30$
  - √ Se $n < 30$, o TCL funcionará se a distribuição da população não for muito diferente da normal

Exemplo – Simulação

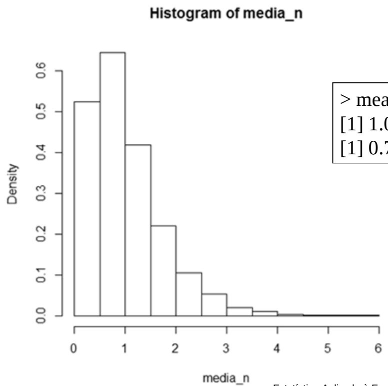
- População exponencial com média 1:
  - √ $\lambda = 1$
  - √ Geração de 10.000 valores dessa população
  - √ Amostra de tamanho 1 (n = 1)

- Amostra n =1



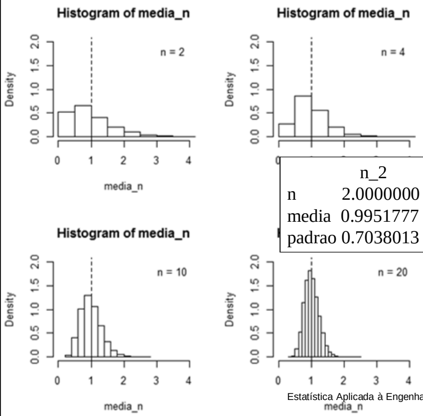

```
> mean(amostra); sd(amostra)
[1] 0.9990838
[1] 1.010478
```


- Amostra n = 2




```
> mean(media_n); sd(media_n)
[1] 1.012711
[1] 0.7129089
```

- Amostras de tamanhos 2, 4, 10 e 20

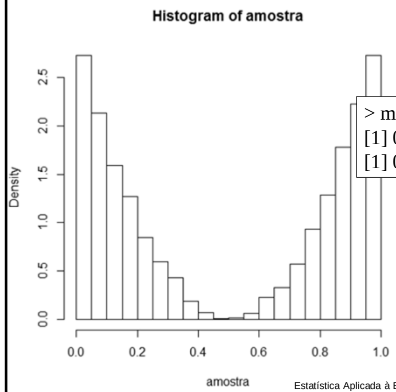


	n_2	n_4	n_10	n_20
n	2.0000000	4.0000000	10.0000000	20.0000000
media	0.9951777	1.0004503	0.9963200	0.9997876
padrao	0.7038013	0.5018562	0.3190797	0.2223573

Exemplo – Simulação

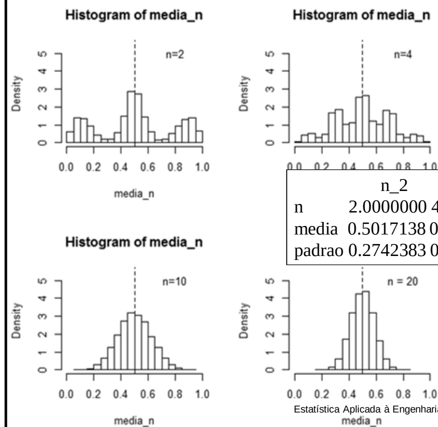
- População com densidade em U:
  - √ f(x) = 12 (x – 0,5)^2
  - √ Geração de 10.000 valores dessa população
  - √ Amostra de tamanho 1 (n = 1)

- Amostra n =1




```
> mean(amostra); sd(amostra)
[1] 0.5061657
[1] 0.3877259
```


- Amostras de tamanhos 2, 4, 10 e 20



	n_2	n_4	n_10	n_20
n	2.0000000	4.0000000	10.0000000	20.00000000
media	0.5017138	0.4986490	0.4994313	0.49993781
padrao	0.2742383	0.1957755	0.1221512	0.08599018

Comparação de Populações

- Considere duas populações:
  - √ População 1: média μ_1 e variância σ_1^2
  - √ População 2: média μ_2 e variância σ_2^2
- Amostras aleatórias das duas populações?
  - √ Amostra da população 1 de tamanho n_1: $\bar{X}_1$
  - √ Amostra da população 2 de tamanho n_2: $\bar{X}_2$

- Caso 1: As duas populações são normais
  - √ Distribuição amostral da diferença

$$\bar{X}_1 - \bar{X}_2 \sim \mathrm{N}(\mu_{\text{dif}}, \sigma^2_{\text{dif}})$$

  - √ Média da diferença de médias amostrais:

$$\mu_{\text{dif}} = \mathrm{E}\left[\bar{X}_1 - \bar{X}_2\right] = \mathrm{E}\left[\bar{X}_1\right] - \mathrm{E}\left[\bar{X}_2\right] = \mu_1 - \mu_2$$

  - √ Variância da diferença de médias amostrais:

$$\sigma^2_{\text{dif}} = \mathrm{Var}\left[\bar{X}_1 - \bar{X}_2\right] = \mathrm{Var}\left[\bar{X}_1\right] + \mathrm{Var}\left[\bar{X}_2\right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- Caso 2: populações não normais com tamanhos amostrais maiores que 30
  - √ Pode-se usar o TCL para aproximar a distribuição amostral da diferença:

Distribuição Amostral Aproximada de Diferença de Médias Amostrais

- Suponha:
  - √ Duas populações independentes, com médias μ_1 e μ_2 e variâncias σ_1^2 e σ_2^2
  - √ Amostras aleatórias independentes de tamanhos n_1 e n_2 dessas populações

$$Z = \frac{\bar{X}_{1,n_1} - \bar{X}_{2,n_2} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

se as condições do TCL se aplicarem

Exemplo – Vida de Motor

- Motor de turbina de aeronave a jato
  - √ Vida de componente é variável aleatória com média 5000 h e desvio-padrão 40 horas
  - √ Melhoria no componente: média 5050 h e desvio-padrão 30 horas
  - √ Suponha amostra de $n_1 = 16$ componente do processo antigo e $n_2 = 25$ do processo aprimorado
- Qual a probabilidade de que a diferença das média amostrais seja no mínimo 25 horas?

Estatística Aplicada à Engenharia 36

Estimação Pontual – Conceitos

Propriedades de um Estimador

- Alguma propriedades importantes:
  - √ Vício
  - √ Consistência
  - √ Eficiência

Estatística Aplicada à Engenharia 38

Vício

- Vício de um estimador:
  - √ Vício($\hat{\theta}$) = $E[\hat{\theta}] - \theta$
- Um estimador $\hat{\theta}$ é não viciado (não viesado, não tendencioso) para um parâmetro θ se $E[\hat{\theta}] = \theta$
- A esperança de um estimador está relacionada com sua **exatidão**

Estatística Aplicada à Engenharia 39

- Exemplos:
  - √ A média amostral é não viciada para estimar a média verdadeira (populacional):

$$\mathrm{E}(\bar{X}_n) = \mu_X$$

  - √ X_1 (primeiro item coletado da amostra) é não viciado para estimar a verdadeira média

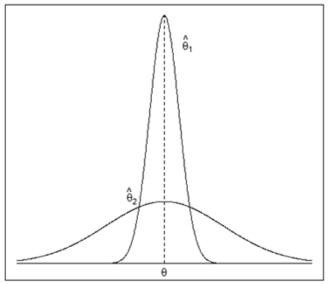
$$\mathrm{E}(X_1) = \mu_X$$

- A variância amostral é não viciada para estimar a variância populacional (σ^2)?

$$\begin{aligned}
\mathrm{E}\left[S^2\right] &= \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left[\left(X_i-\bar{X}\right)^2\right] \\
&= \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}\left(X_i-\mu+\mu-\bar{X}\right)^2\right] \\
&= \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-\frac{1}{n-1}\mathrm{E}\left[n\left(\bar{X}-\mu\right)^2\right] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\mu)^2\right]-\frac{n}{n-1}\mathrm{E}\left[(\bar{X}-\mu)^2\right] \\
&= \frac{n}{n-1}\sigma^2-\frac{1}{n-1}\sigma^2 \\
&= \sigma^2
\end{aligned}$$

Variância de Estimador

- $\hat{\theta}_1$ e $\hat{\theta}_2$ estimadores não-viciados de θ
  - √ Variâncias diferentes



  - √ $\mathrm{Var}(\hat{\theta}_1) < \mathrm{Var}(\hat{\theta}_2)$
    - É mais provável que $\hat{\theta}_1$ produza uma estimativa mais próxima do valor verdadeiro de θ

Estimador de Variância Mínima

- Se considerarmos todos os estimadores não-tendenciosos de θ, aquele com a menor variância será chamado de **estimador não-tendencioso de variância mínima**
  - √ Esse estimador é o mais provável, dentre todos os não-viciados, para produzir uma estimativa que seja próxima do valor verdadeiro

Erro Padrão

- O erro padrão de um estimador $\hat{\theta}$ é o seu desvio-padrão.

$$\sigma_{\hat{\theta}} = \sqrt{\text{Var}\left[\hat{\theta}\right]}$$

√ O erro padrão (ou da variância) do estimador está relacionada com sua **precisão**

√ Se o erro-padrão envolver parâmetros desconhecidos que possam ser estimados, então a substituição daqueles valores produz um erro-padrão estimado

– Exemplo: O erro padrão da média amostral é: $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$

– Se não conhecermos s, mas substituímos pelo desvio-padrão amostral, então o erro-padrão amostral estimado da média amostral é: $\hat{\sigma}_{\bar{X}} = \dfrac{S}{\sqrt{n}}$

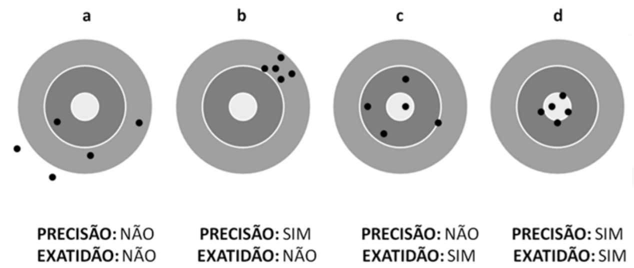
√ Quando o estimador seguir uma distribuição normal, podemos estar confiantes que o valor verdadeiro do parâmetro estará entre dois erros-padrão da estimativa

– Para grandes valores de n este é um resultado útil

√ Nos casos em que o estimador é não-viciado e não normalmente distribuído

– Estimativa do parâmetro, em no máximo 6% das vezes, se desviará do valor verdadeiro tanto quanto 4 erros-padrão

- Quadro comparativo:



Consistência

- Um estimador $\hat{\theta}$ é consistente se à medida em que o tamanho amostral aumenta, seu valor esperado converge para o parâmetro de interesse e sua variância converge para zero.
  - √ O estimador é consistente se $\lim_{n \to \infty} \mathrm{E}\left[\hat{\theta}\right] = \theta$ e

    $$\lim_{n \to \infty} \mathrm{Var}\left[\hat{\theta}\right] = 0$$
  - √ Consistência é uma propriedade assintótica (grandes amostras)

- Exemplos:
  - √ A média amostral é consistente para estimar a média verdadeira
  - √ O primeiro item coletado da amostra não é consistente para estimar a média populacional.

Erro Quadrático Médio

- O erro quadrático médio de um estimador $\hat{\theta}$ do parâmetro θ é definido como:

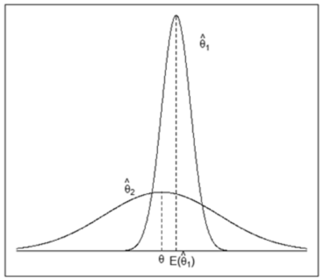
$$\mathrm{EQM}\left[\hat{\theta}\right] = \mathrm{E}\left[\hat{\theta} - \theta\right]^2$$

- EQM – Vício e erro-padrão

$$\mathrm{EQM}[\hat{\theta}] = \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta})]^2 + [\theta - \mathrm{E}(\hat{\theta})]^2$$
$$= \mathrm{Var}[\hat{\theta}] + [\mathrm{Vício}(\hat{\theta})]^2$$

- O EQM é um critério importante para comparar dois estimadores

- √ Estimadores tendenciosos podem ser preferíveis a estimadores não-tendenciosos se tiverem EQM menor



- √ Estimativa baseada em $\hat{\theta}_1$ estaria provavelmente mais próxima do valor verdadeiro do que a baseada em $\hat{\theta}_2$

- Estimador ótimo de θ:
  - √ Tem EQM menor ou igual ao EQM de qualquer outro estimador, para todos os valores de θ
  - √ Estimadores ótimos raramente existem

EQM de Estimadores Não-viciados

- No caso em que $\hat{\theta}$ é um estimador não viciado para um parâmetro θ, então:

$$\begin{aligned}\mathrm{EQM}[\hat{\theta}] &= \mathrm{Var}[\hat{\theta}] + [\mathrm{Vício}(\hat{\theta})]^2 \\ &= \mathrm{Var}[\hat{\theta}]\end{aligned}$$

Eficiência

- Dados dois estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$, não viciados para um parâmetro θ, dizemos que $\hat{\theta}_1$ é mais eficiente que θ_2 se $\mathrm{Var}[\hat{\theta}_1] < \mathrm{Var}[\hat{\theta}_2]$.

- Exemplo:
  - √ No caso de amostra proveniente de distribuição Normal.
    - Média amostral e mediana amostral são não viciadas para estimar a média populacional:

$$\mathrm{E}\left[\bar{X}\right] = \mu \quad \text{e} \quad \mathrm{E}\left[\tilde{X}\right] = \mu$$

    - Média amostral e mediana amostral são consistentes para estimar a média verdadeira

$$\mathrm{Var}\left[\bar{X}\right] = \frac{\sigma^2}{n} \quad \text{e} \quad \mathrm{Var}\left[\tilde{X}\right] = \frac{\pi\sigma^2}{2\,n}$$

    - A média amostral é mais eficiente que a mediana amostral para estimar a média populacional

$$\frac{\mathrm{Var}\left[\bar{X}\right]}{\mathrm{Var}\left[\tilde{X}\right]} = \frac{\frac{\sigma^2}{n}}{\frac{\pi\sigma^2}{2\,n}} = \frac{2}{\pi} = 0,63 < 1$$

Referências

Bibliografia Recomendada

- Montgomery, D. C. (LTC)

  Estatística Aplicada e Probabilidade para Engenheiros

- Pinheiro, J. I. D et al. (Campus)

  Probabilidade e Estatística: Quantificando a Incerteza