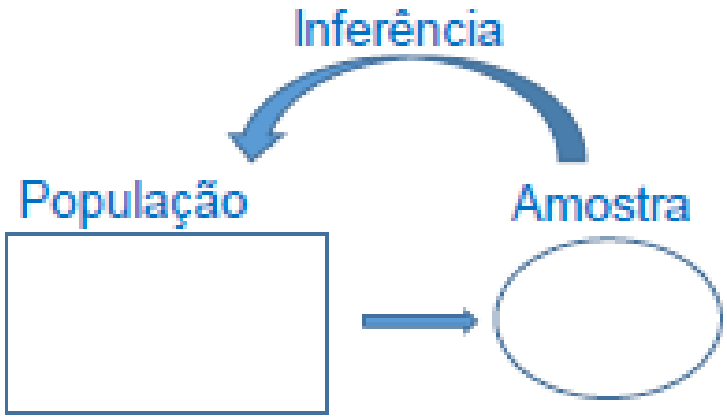
INFERÊNCIA ESTATÍSTICA (EST0035)

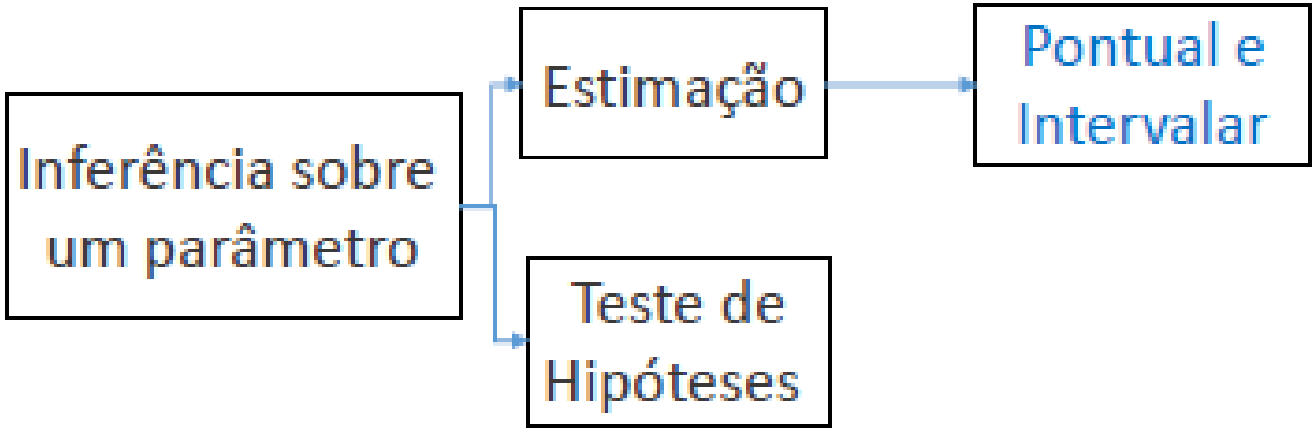
AULA01: AMOSTRA ALEATÓRIA

Frederico Machado Almeida
frederico.almeida@unb.br

Departamento de Estatística
Instituto de Exatas
Universidade de Brasília (UnB)

Revisitando Alguns Conceitos





Revisitando Alguns Conceitos

Definição 1 (Inferência Estatística)

É a área da Estatística que engloba um conjunto de técnicas que tem por objetivo, estudar a população através de evidências fornecidas por uma amostra.

- Ou seja, uma amostra é a parte do todo (população), que contêm os elementos que podem ser observados, e a partir daí, quantidades de interesse (parâmetros) podem ser mensurados.
- O objetivo é inferir (conjecturar a respeito do verdadeiro (ou verdadeiros) valor(es) de θ possivelmente um vetor), ou seja, a respeito do(s) parâmetro(s) de interesse.

Revisitando Alguns Conceitos

- Em linhas gerais, a Inferência Estatística pode ser classificada em *dedutiva* e *indutiva*.
- **A inferência indutiva** é baseada em uma sequência de premissas, chega-se a conclusões de interesse, através de nexos causais lógicas.
- **A inferência dedutiva** tem com base em uma parte da população, obter conclusões para o todo (esta é a que será abordada no curso).

Amostra Aleatória

Definição 2 (Amostra aleatória)

A coleção de variáveis aleatórias (v.a.) $X_1, \cdots, X_n$ é dita ser uma amostra aleatória (a.a.) se (i) os $X_i's$ forem variáveis aleatória (v.a's) independentes, e (ii) cada X_i tiver a mesma distribuição de probabilidades.

- Simbolicamente escrevemos $X_i \overset{iid}{\sim} F$ para representar v.a's independentes e identicamente distribuídas (*iid*).
- No caso em que a amostra não é aleatória, mas sim baseada em julgamentos ou qualquer outra maneira, então os métodos estatísticos não funcionarão de forma apropriada e levarão a decisões incorretas.
- Considere $X_1, X_2, \cdots, X_n$ uma a.a. proveniente de uma população descrita pela função $f(x|\theta)$ e caracterizada pelo parâmetro $\theta \in \mathbf{\Theta}$.

Amostra Aleatória

- Além do que foi dito, a coleção $X_1, \cdots, X_n$ denota uma a.a. porque o modelo de amostragem que foi utilizado para obter tal coleção foi baseado em "*processo de amostragem aleatória simples*".
- Se $x_1, \cdots, x_n$ denota uma amostra observada de $X_1, \cdots, X_n$, com função densidade de probabilidade (fdp) ou função de probabilidade (fp), $f(x|\theta)$, segue imediatamente que,

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{1}$$

- Em suma, a a.a. que será objeto de estudo nessa disciplina, fornece uma ligação importante entre os dados observados, e a distribuição na população de interesse.

Amostra Aleatória

É importante destacar que, todas as quantidades de interesse (amostrais) que serão estudadas nessa disciplina, serão baseadas (ou expressas), como função da a.a.

Definição 3 (Estatística)

Seja $X_1, \cdots, X_n$ uma a.a. de tamanho n extraída de uma população com fdp (ou fp), $f(x|\theta)$. A função $T(X_1, \cdots, X_n)$ da a.a. (ou amostra observada), cujo o domínio inclui o espaço amostral $\mathcal{X}$, é dita ser uma Estatística.

Amostra Aleatória

- Por ser função de uma a.a. $T_n = T(X_1, \cdots, X_n)$ é uma v.a. e portanto, ela apresenta uma distribuição de probabilidades (distribuição amostral da estatística).

Definição 4 (Estimador)

Qualquer estatística T_n que assume valores no espaço paramétrico Θ, é chamado de estimador de θ, ou $g(\theta)$, se ela não depender do parâmetro de interesse.

Observação 1

O conjunto $\Theta \subset \mathbb{R}^d$ em que θ toma valores é denominado espaço paramétrico.

Amostra Aleatória

Existem vários tipos de estatísticas, a saber:

i. $T_1 = \sum_{i=1}^{n} X_i$.

ii. $T_2 = \sum_{i=1}^{n} X_i/n = T_1/n$.

iii. $T_3 = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 / (n-1)$.

iv. $T_4 = \frac{X_1+(n-1)X_n}{n}$.

v. $T_5 = \frac{X_{(1)}+X_{(n)}}{2}$.

vi. $T_6 = \frac{X_1+X_3+2X_4}{3}$, etc.

Onde $X_{(1)} = \text{mínimo}\{X_1, \cdots, X_n\}$ e $X_{(n)} = \text{máximo}\{X_1, \cdots, X_n\}$, denotam as estatísticas extremais da a.a.

Amostra Aleatória

Observação 2

Denotemos por $\epsilon = T_n - \theta$ o erro amostral que cometemos ao estimar o parâmetro θ da distribuição da v.a. X pelo estimador T_n.

Observação 3

Ao valor numérico que o estimador T_n assume depois que a amostra $x_1, \cdots, x_n$ for observada, isto é, $T_n = T(\mathbf{x})$ recebe o nome de estimativa.

- Por exemplo, $T_n = 1,87$ *metros* pode ser uma estimativa da altura média dos alunos matriculados em Inferência Estatística.

Amostra Aleatória

Exemplo 1: Suponha que $X_1, \cdots, X_n$ é uma a.a. de uma v.a. X proveniente de uma população caracterizada pela fdp (ou fp), $f(x|\theta)$. Portanto, possíveis estimadores pontuais são: (i) a média aritmética, $\bar{X}_n = \sum_{i=1}^{n} X_i/n$, e (ii) a variância amostral $S^2 = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2/(n-1)$.

Exemplo 2: Assuma agora que a v.a. X segue uma distribuição de Bernoulli(θ), tal que, $\theta = P(X = 1)$ e $1 - \theta = P(X = 0)$, com $0 < \theta < 1$. Portanto, supondo que o experimento foi executado n vezes de forma independente, então, um estimador natural para θ seria a proporção amostral, $\hat{\theta}_n = \sum_{i=1}^{n} X_i/n$.

Amostra Aleatória

- Nos slides anteriores vimos que, vários estimadores para uma mesma quantidade θ ou para uma função $g(\theta)$, podem ser definidos com base na a.a. $X_1, \cdots, X_n$.
- Portanto, um dos grandes problemas da estatística é o de encontrar um estimador razoável para a quantidade de interesse.
- Duas perguntas de interesse podem ser feitas. A saber: (i) qual, entre os estimadores apresentados no SLIDE 9, é o melhor? E, (ii) qual(is) é/são o(s) procedimentos para escolher o melhor o estimador que melhor representa a quantidade de interesse?

Qualidade dos Estimadores

- O procedimento comumente utilizado para avaliar a qualidade de um estimador, inclui avalia a:
  - Tendenciosidade (não-viesamento).
  - Consistência.
  - Eficiência.
  - Deve apresentar uma variância (ou erro quadrático médio), menor possível.

Exemplo 3: Considere a Figura abaixo, em que o alvo representa o valor do parâmetro e os "tiros", indicados pelos símbolo x, representam os diferentes valores amostrais da estatística de interesse.

Qualidade dos Estimadores

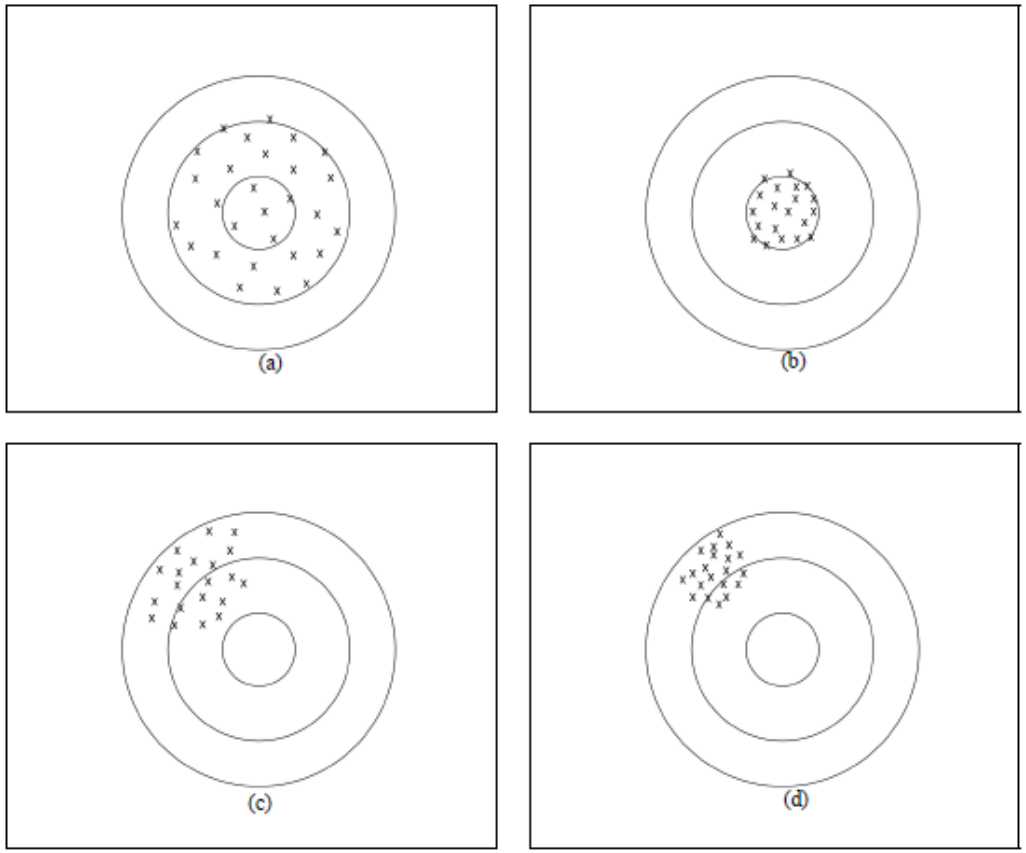


Figura: Nos painéis acima temos estimadores: (a) muito acurado mas pouco preciso; (b) muito acurado e muito preciso; (c) pouco acurado e pouco preciso, e (d) muito preciso, mas pouco acurado.

Qualidade dos Estimadores

Definição 5 (Estimador Não-Viesado)

Um estimador $T_n = T(X_1, \cdots, X_n)$, de θ é dito ser não-viesado (ou não-tendencioso), se $\mathbb{E}_\theta(T_n) = \theta$, para todo $\theta \in \boldsymbol{\Theta}$.

- A definição 5 diz que, se o processo de estimação for repetido várias vezes para diferentes amostras extraídas na mesma população, e para a mesma característica de interesse, a média das médias amostrais deve estar próxima do alvo, θ.
- Caso a definição 5 não seja verificada, o estimador é dito ser viesado (ou tendencioso).

Qualidade dos Estimadores

Exemplo 4: Seja $X_1, \cdots, X_n$ uma a.a. de uma v.a. X com $\mathbb{E}(X) = \mu$ e $Var(X) = \sigma^2$. Se $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ e $\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ denotam os estimadores pontuais para a média e variância, o que se pode dizer quanto a tendenciosidade ou não dos estimadores?

Qualidade dos Estimadores

Observação 4

Um estimador T_n de θ é dito ser **assintoticamente não-viesado**, *se:*

$$\lim_{n\to\infty} \mathbb{E}_\theta\left(T_n\right) = \theta, \quad \textit{para todo}\, \theta \in \boldsymbol{\Theta}.$$

- Neste caso, temos que $\lim_{n\to\infty} b\left(T_n\right) = 0$. Com $b\left(T_n\right) = \mathbb{E}_\theta\left(T_n\right) - \theta$ denotando o viés (vício) do estimador T_n.
- A *observação 4* indica que, quando o tamanho de amostra aumenta, o viés de qualquer estimador não centrado pode ser considerado desprezível.

Qualidade dos Estimadores

Definição 6 (Consistência)

Seja $X_1, \cdots, X_n$ uma a.a. proveniente da distribuição de uma v.a. X. Uma sequência de estimadores $\{T_n\}_{n\geq 1}$ é dita ser consistente para o parâmetro de interesse θ, se:

i. $\lim_{n\to\infty} \mathbb{E}_\theta(T_n) = \theta.$

ii. $\lim_{n\to\infty} Var_\theta(T_n) = 0.$

Observação 5

A definição clássica de consistência de uma sequência de estimadores $\{T_n\}_{n\geq 1}$ segue da definição de convergência em probabilidade, i.e., a sequência $\{T_n\}_{n\geq 1}$ é consistente para θ, se $T_n \xrightarrow{p} \theta$, quando $n \to \infty$.

Qualidade dos Estimadores

Definição 7 (Eficiência)

Se T_n e T_n^* denotam dois estimadores não-viesados do parâmetro θ, a definição mais simplista de eficiência assume que, o estimador T_n é mais eficiente que T_n^*, se $Var_\theta(T_n) < Var_\theta(T_n^*)$.

- Uma definição mais abrangente de eficiência consiste em comparar a variância do estimador T_n, com o limite inferior limite inferior da variância dos estimadores não viciados de θ.

Qualidade dos Estimadores

Definição 8 (Erro Quadrático Médio)

O erro quadrático médio (EQM) é um dos procedimentos comumente utilizado para avaliar a qualidade de um estimador. Assim, o EQM de um estimador T_n de $\theta \in \boldsymbol{\Theta}$ é dado por:

$$EQM\left(T_n\right) = \mathbb{E}_\theta\left[\left(T_n - \theta\right)^2\right] = Var_\theta\left(T_n\right) + \left[b\left(T_n\right)\right]^2. \tag{2}$$

- Observe que, na expressão 2 a quantidade $\left[b\left(T_n\right)\right]^2$ denota um termo de penalidade no EQM.
- Assim, se T_n for um estimador não-viesado, então, $EQM\left(T_n\right) = Var_\theta\left(T_n\right)$. Consequentemente, a consistência de um estimador T_n pode ser aferida usando o EQM.

Qualidade dos Estimadores

Exemplo 5: Seja $X_1, \cdots, X_5$ uma a.a. da v.a. X, com $\mathbb{E}_\theta(X) = \theta$ e $Var_\theta(X) = 1$. Considere os seguintes estimadores, $T_{1n} = \sum_{i=1}^{5} X_i/5$; $T_{2n} = (X_1 + 2X_2 + X_3 + 3X_4 + X_5)/8$ e $T_{3n} = (X_2 + 3X_4 + X_5)/3$.

a. Qual(is) dos estimadores é/são não-viesados?
b. Obtenhas os EQM's dos três estimadores.
c. Qual dos três estimadores é o mais eficiente?

Qualidade dos Estimadores

Definição 9

Seja $X_1, \cdots, X_n$ denota uma aa proveniente de uma população caracterizada por uma fdp (ou fp) $f(x|\theta)$. Assim, a esperança matemática da va X é dada por:

$$\mathbb{E}_\theta(X) = \begin{cases} \int_{\mathcal{A}} xf(x|\theta) & \text{se } X \text{ for contínua,} \\ \\ \sum_{x \in \mathcal{A}} xP(X = x|\theta) & \text{se } X \text{ for discreta,} \end{cases}$$

com $\mathcal{A}$ denotando o suporte da va X (equivalente ao espaço amostral $\mathcal{X}$).

Qualidade dos Estimadores

Lema 1

Seja $X_1, \cdots, X_n$ uma aa obtida a partir de uma população caracterizada pela fdp (ou fp) $f(x|\theta)$. Seja igualmente $g(x)$ uma função, tal que, $\mathbb{E}(g(X_1))$ e $Var(g(X_1))$ existam. Então,

a. $\mathbb{E}_\theta\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathbb{E}_\theta(g(X_1))$

b. $Var_\theta\left(\sum_{i=1}^{n} g(X_i)\right) = nVar_\theta(g(X_1))$.

Qualidade dos Estimadores

Teorema 2

Se $X_1, \cdots, X_n$ denota uma aa da va X, seja $T_n = \sum_{i=1}^{n} a_i X_i$ uma estatística qualquer. Supondo que $\mathbb{E}(|X_i|) < \infty$ e $\mathbb{E}(X_i^2) < \infty$, para todo $i = 1, \cdots, n$. Então,

a. $\mathbb{E}(T_n) = \sum_{i=1}^{n} a_i \mathbb{E}(X_i)$

b. $Var(T_n) = \sum_{i=1}^{n} a_i^2 Var(X_i)$.

Prova: a prova do Teorema 1 é imediata, e segue das propriedades da esperança matemática, e da independência das va's

Função Geradora de Momentos

Definição 10 (Função Geradora de Momentos)

Seja X uma va com função de distribuição $F_X(x)$, ou fdp/fp $f_X(x)$. A função geradora de momentos (fgm), com a notação $M_X(t)$, é dada por:

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \begin{cases} \int\limits_{\mathcal{A}} e^{tX} dF_X, & \text{se } X \text{ for contínua,} \\ \sum\limits_{x \in \mathcal{A}} e^{tX} P(X = x), & \text{se } X \text{ for discreta,} \end{cases}$$

Teorema 3

Suponha que a fgm de uma va X exista para $|t| < t_0$, com $t_0 > 0$. Então, $\mathbb{E}\left(X^{(s)}\right)$ existe para todo $s = 1, 2, \cdots$ e temos:

$$\mathbb{E}(X^s) = \frac{d^{(s)}}{dt^s} M_X(t) = \frac{d^{(s)}}{dt^s} \int\limits_{\mathcal{A}} e^{tX} dF_X, \ \forall_{|t|<t_0}.$$