Amostragem

Prof. Adriano Mendonça Souza, Dr.
Departamento de Estatística
PPGEMQ / PPGEP - UFSM

Amostragem

É o processo de seleção de amostras de uma população com o objetivo de fazer inferências sobre a população em estudo.

Conhecendo-se algumas características dessa população, o objetivo é obter valores aproximados, ou estimativas, para as características populacionais.

Classificação e tipos de Amostragem

1) Probabilística

Quando as amostras, obtidas da população são aleatórias, todos os elementos da população tem a mesma probabilidade de serem escolhidos.

Assim toda a variabilidade da população é refletida na amostra e o conhecimento de um elemento qualquer nada nos informa quanto ao valor dos outros elementos.

1.1) Aleatória simples

Os elementos da população são enumerados, onde a ***seleção da amostra é feita por meio de um sorteio***, sem restrição, onde cada elemento da população tem a mesma probabilidade de pertencer à amostra;

1.2) Sistemática

É um processo de amostra que segue uma **sistemática** e possibilita resultados semelhantes aos da amostra aleatória simples, por um processo mais rápido, desde que a **população se encontre ordenada** e seja seguido o seguinte procedimento:

Procedimento:

Calcula-se o intervalo de amostragem **k =N/n**;

Sorteia-se um número **X** dentro do intervalo **1 a k**;

A amostra é formada pelos elementos:

x; x + k; x + 2k; ...,x + (n-1)k;

1.3) Estratificada

Dividi-se a população em **k subgrupos** denominados de estratos. Estes **estratos** devem ser **internamente** mais **homogêneos** do que a população em relação as variáveis em estudo.

$$N = N_1 + N_2 + \ldots + N_k = \sum_{i=1}^{k} N_i \, ;$$

$$n = n_1 + n_2 + \ldots + n_k = \sum_{i=1}^{k} n_i \, ;$$

$$k = \acute{e} \;\; o \;\; n\acute{u}mero \;\; de \;\; estratos \, .$$

1.4) Estratificada proporcional

Neste caso a proporcionalidade do tamanho de cada estrato da população é mantida na amostra

$$n / N = n_i / N_i$$

Este tipo de amostragem tende a gerar resultados mais precisos, quando comparada com uma AES

1.5) Estratificada uniforme

Geralmente usada em situações em que o maior interesse é obter estimativas separadas para cada estrato, quando os estratos tem aproximadamente o mesmo tamanho ou ainda, quando se deseja comparar diversos estratos. Sendo assim seleciona-se a mesma quantidade de elementos em cada estrato.

$\mathbf{n_1 = n_2 = ... = n_i}$, onde $\mathbf{n_i = n / k}$.

2) Não- probabilística

É um processo subjetivo de amostragem e depende muito do conhecimento do pesquisador a respeito da estrutura das populações e a amostra é uma parcela proporcional desta estrutura.

O erro de amostragem neste caso não é possível de ser estimado, pois não se conhece a probabilidade que cada elemento possui de pertencer a amostra

2.1) Inacessibilidade a população

A amostra é colhida na parte da população que é acessível. Então faz-se uma distinção entre população-objeto e população-amostrada. Se as duas populações tiverem as mesmas características, esse tipo de amostragem eqüivalerá a uma amostragem probabilística.

2.2) Amostragem a esmo

Não se utiliza nenhum sorteio para escolher os elementos que pertencerão a amostra, embora o amostrador procure ser aleatório.

2.3) A população é formada por material contínuo

No caso de líquidos ou gases é necessário homogeneizá-la para logo em seguida retirar a amostra a esmo, pois um sorteio neste caso é impossível.

2.4) Amostragens intencionais

Neste caso o amostrador escolhe deliberadamente certos elementos para formar a amostra, baseado em um conhecimento prévio.

Importância da utilização da amostragem

- Economia;
- Tempo;
- Confiabilidade dos dados;
- Operacionalidade;

Situações em que pode não valer a pena a amostragem

- População pequena;
- Característica de fácil mensuração;
- Necessidade de alta precisão;

Plano de amostragem

Num plano de amostragem deve constar:

- a definição da unidade de amostragem
- a forma de seleção dos elementos da população
- o tamanho da amostra.

Definição dos objetivos

- O que?
- Onde?
- Precisão exigida
- Tempo necessário
- Custo previsto

Determinação dos meios

- Tipo de amostragem
- Tamanho da amostra
- Qual o método para o levantamento dos dados
- Como os interessados serão questionados

Preparação do plano

- Elaboração do questionário
- Características das questões
- Experimentação do questionário
- Execução
- Coleta
- Crítica
- Apuração
- Apresentação dos dados

Análise dos resultados

Determinar uma característica

Estimar e verificar os parâmetros

Relatório final:

Claro e conciso

Honesto

Para se fazer um plano de amostragem deve-se ter bem definidos:

- os objetivos da pesquisa
- a população a ser amostrada
- os parâmetros a serem estimados

Elaboração de Questionários e Formulários

A elaboração de questionários e formulários constitui passo importante no planejamento da pesquisa.

Definição De Um Questionário:

Técnica estruturada para coleta de dados, que consiste de uma série de perguntas - escritas ou verbais - que um entrevistado deve responder.

Objetivos De Um Questionário:

Qualquer questionário tem três objetivos específicos:

1º) traduzir a informação desejada

2º) motivar e incentivar o entrevistado

3º) minimizar o erro a resposta

Processo de Elaboração de Questionários

1º) Especificar a Informação Buscada

2º) Especificar o Tipo do Método da Entrevista

3º) Determinar o Conteúdo das Perguntas Individuais

4º) Planejar as Questões de Modo a Superar a Incapacidade e/ou Má Vontade do Entrevistado

5º) Decidir Quanto à Estrutura das Questões

6º) Determinar o Fraseado das Questões

7º) Dispor as Questões na Ordem Adequada

8º) Identificar a Forma e o LayOut

9º) Reproduzir o Questionário

10º) Eliminar Defeitos por Meio de um Pré-Teste

Determinação do Tamanho da Amostra

Em pesquisas, uma etapa de grande importância é a determinação do tamanho da amostra que será utilizado para o levantamento dos dados.

Depende de três fatores:

1º)Nível de confiança $(\alpha - 1)$: o pesquisador é que vai determinar o valor desejado;

2º)Precisão (e_0): em toda experimentação ou pesquisa, a utilização da amostragem está condicionada a um erro amostral, que corresponde à diferença entre as estimativas amostrais e os parâmetros populacionais;

3º)Tipo de Investigação: depende das características populacionais a serem investigadas.

Cálculo para o tamanho mínimo da amostra

Um primeiro cálculo do tamanho da amostra pode ser feito, mesmo sem conhecer o tamanho da população, da seguinte forma:

$$\boxed{n_0 = \frac{1}{e_0^2}}$$

Onde:

$n_0 \longrightarrow$ **primeira aproximação para o tamanho da amostra.**

$e_0 \longrightarrow$ **erro amostral tolerável**

Quando se conhece o tamanho da população

o cálculo anterior pode ser corrigido:

$$n = \frac{N \cdot n_0}{N + n_0}$$

N ⟶ **tamanho da população**

n ⟶ **tamanho da amostra**

Tamanho da amostra para estimar a média populacional

com variância populacional conhecida:

População Infinita	População Finita
$n = \left[\dfrac{\left(z_{\alpha/2} \right) \sigma}{e_0} \right]^2$	$n = \dfrac{\left(z_{\alpha/2} \right)^2 \sigma^2 N}{e_0^2 (N-1) + \left(z_{\alpha/2} \right)^2 \sigma^2}$

com variância populacional desconhecida:

Substituí-se pela variância que é obtida através de uma amostra piloto, de tamanho n_1.

População Infinita	População Finita
$n = \left[\dfrac{\left(t_{\delta, \alpha/2} \right) S}{e_0} \right]^2$	$n = \dfrac{\left(t_{\delta, \alpha/2} \right)^2 s^2 N}{e_0^2 (N-1) + (t_{\delta, \alpha/2})^2 s^2}$

Onde δ = n1 - 1 graus de liberdade

Considerações após o cálculo do tamanho da amostra:

- Se $n < n_1$: a pré-amostra selecionada, de tamanho n_1, foi suficiente para garantir a precisão desejada.

- Se $n > n_1$: deve-se completar a pré-amostra, acrescentando elementos até atingir o valor "n" que garanta a precisão desejada.

Tamanho da amostra para estimar uma proporção populacional

População Infinita	População Finita
$n = \dfrac{\left(z_{\alpha/2}\right)^2 p^* p^*}{e_0^2}$	$n = \dfrac{\left(z_{\alpha/2}\right)^2 p^* q^* N}{e_0^2 (N-1) + (z_{\alpha/2})^2 p^* q^*}$

Onde: p* é a proporção amostral que pode ser obtida através de um pré-amostra de n_1 elementos e $q^* = (1\text{-}p)$

Observações:

1) Quando não se tem informação a respeito de p*, usa-se p* = q* = 50% o que levará a um tamanho de amostra superavaliado mas garantindo a precisão desejada, embora podendo ter como conseqüência, aumentos no custo e no tempo de amostragem e consequentemente na pesquisa.

2) Se p for próximo de 0 ou 1 corre-se o risco de se dimensionar uma amostra maior do que o necessário.

Exemplos

1) Qual o tamanho da amostra necessária para se estimar a média de uma população infinita cujo desvio padrão é 4mm, com 98% de confiança e precisão de 0,5mm?

2) Qual o número de jogadas de uma moeda é suficiente para se estimar a proporção de "caras" obtidas, com precisão de 3% e confiança de 90%?

3) Foram feitas 20 medidas do tempo gasto em minutos para se fabricar um componente industrial,obtendo-se: 13,15,12,14,17,15,16,15,14,16,17,14,16,15,15,13, 14,15,16,15.

Esses dados são suficientes para estimar o tempo médio gasto nessa fabricação, com precisão de 30 segundos e 95% de certeza? Caso negativo, qual o tamanho da amostra adicional necessária?

4) Em uma pesquisa de mercado bem conduzida, 57 das 150 pessoas entrevistadas afirmaram que seriam compradoras de certo produto a ser lançado. Essa amostra é suficiente para estimar a proporção real de futuros compradores, com uma precisão de 0,08 e confiança de 95%?

Inferência Estatística

Uma população pode ser inferida através de sua amostra através de dois modos:

- **Estimação**
- **Testes de hipóteses**

Estimação

A estimação de um parâmetro populacional pode ser de dois tipos:

estimação pontual

estimação intervalar

Testes de hipóteses

O Teste de Hipótese é uma regra de decisão para aceitar ou rejeitar uma hipótese estatística com base nos elementos amostrais.

Uma hipótese estatística é uma suposição acerca da distribuição de uma variável aleatória.

A inferência através dos testes de hipóteses se dá quando usamos os resultados extraídos da amostra para testarmos valores de certos parâmetros da população, ou mesmo testarmos a natureza da população. Podendo ser, testes **paramétricos** *ou de* **aderência**.

Exemplos:

1) Em um exame a que se submeteram estudantes de um grande número de escolas diferentes, o grau médio foi 74,5 e o desvio padrão foi 8. Em uma escola particular em que 200 estudantes foram submetidos a esse exame, o grau médio foi 75,9. Testar a hipótese de que a média é de fato 74,5 contra a alternativa de que a média é diferente de 74,5. Adotar $\alpha = 0{,}05$.

2) *Uma amostra aleatória de 36 copos de um certo vinho mostrou que tinha um conteúdo médio líquido de 220 ml, com desvio padrão de 26 ml. Testar a hipótese de que $\mu = 225$ ml contra a alternativa $\mu < 225$ ml, com o nível de significância de $\alpha = 0{,}05$.*

Parâmetros e Estatísticas

Qualquer valor calculado com base nos elementos de uma amostra é chamado uma "estatística"

Parâmetro	Estimador
μ = média σ^2 = variância σ = desvio padrão	$\bar{x}$ = média s^2 = variância s = desvio padrão

As estatísticas variam de uma para outra amostra, sendo pois, elas próprias VA. Podemos, assim, falar de população de médias amostrais, de medianas amostrais ou, equivalentemente, de distribuição das médias amostrais, das medianas amostrais, e assim por diante.

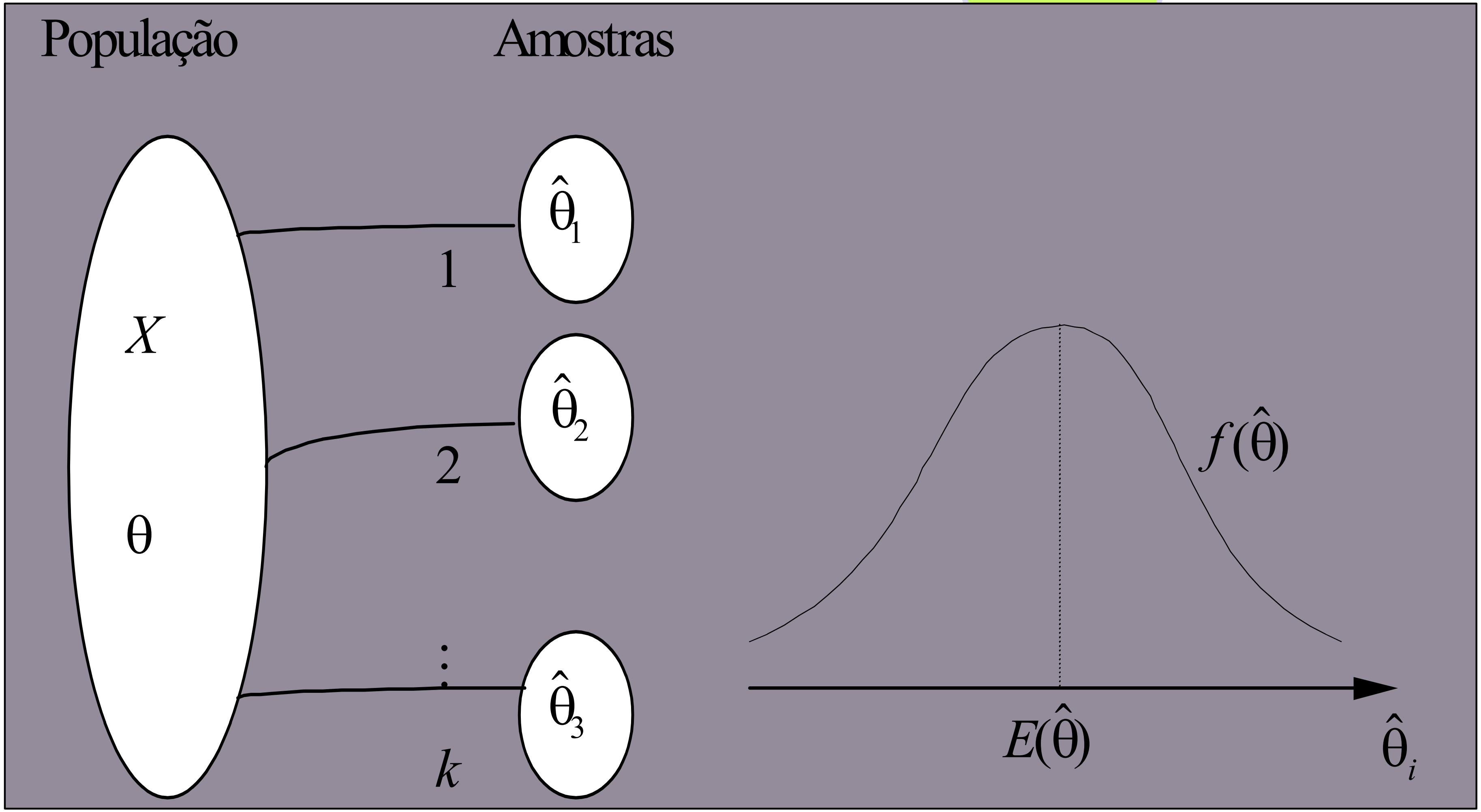
Conceitos de Estimador e Estimativa

Estimador ($\hat{\theta}$) de um parâmetro populacional teta é uma variável aleatória função dos elementos amostrais $\hat{\theta} = f(X_1, X_2, \ldots, X_n)$;

Estimativa é o valor numérico obtido pelo estimador (ou estatística em uma determinada amostra.

A estatística que estima θ é comumente $\hat{\theta}$. E como diferentes amostras originam valores distintos de $\hat{\theta}$, $\hat{\theta}$ é, ele próprio, uma variável aleatória. Nesse último caso temos uma ***Distribuição Amostral*** da estatística $\hat{\theta}$.

O gráfico seguinte ilustra tal distribuição onde para cada amostra $\hat{\theta}_i$, calculamos o valor da estatística $\hat{\theta}$. Os valores $\hat{\theta}_i$ formam uma $\hat{\theta}$ nova população, a Distribuição Amostral da estatística.



Principais qualidades de um estimador

Quando propomos $\hat{\theta}$ como uma estimativa de θ, não esperamos realmente que $\hat{\theta}$ venha a ser igual a θ. (Recordemos que é uma VA. e, por isso, pode tomar diferentes valores). Este dilema origina duas importantes questões:

1. Quais as características que desejamos que uma "boa" estimativa apresente?

2. Como decidiremos que uma estimativa é "melhor" do que a outra?

Tais questões não admitem uma análise tão simples.

1) o próprio valor do parâmetro θ não é verdadeiramente correto (ao menos em uma situação concreta). Assim, como dizer que $\hat{\theta}$ o é?

2) se tivermos duas estimativas de $\hat{\theta}$,

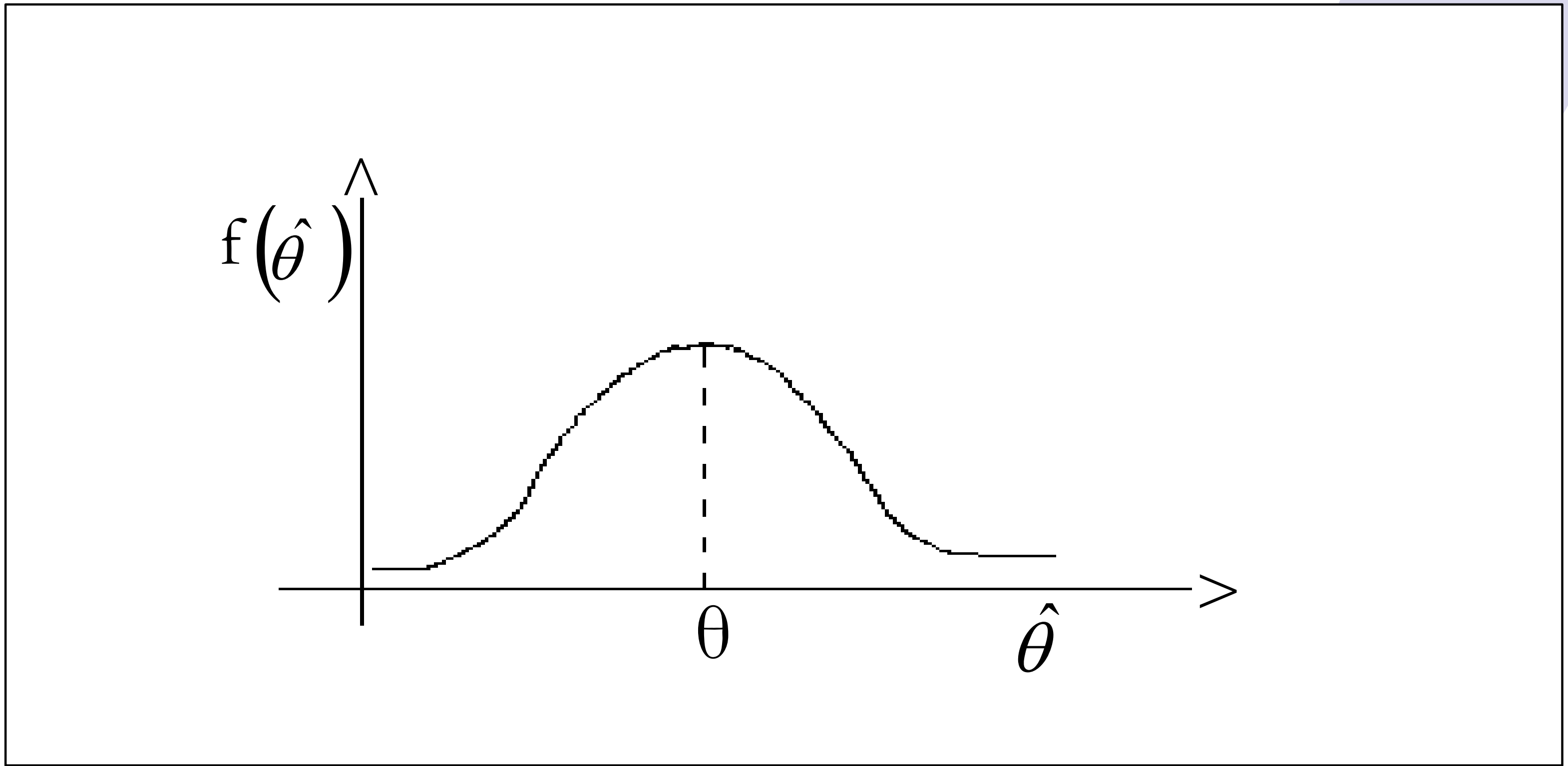
por exemplo $\hat{\theta}_0$ e $\hat{\theta}_1$, qual o critério de comparação para decidirmos entre $\hat{\theta}_0$ e $\hat{\theta}_1$?

Justeza (ou não-tendenciosidade, ou sem vício, ou sem viés)

Nós diremos que $\hat{\theta}$ é uma estimativa não tendenciosa de θ se:

$$\boxed{E(\hat{\theta}) = \theta \; \forall \theta}$$

Estimador Não-tendencioso:



$E[\hat{\theta}] = \theta \Rightarrow$ A média da V.A.

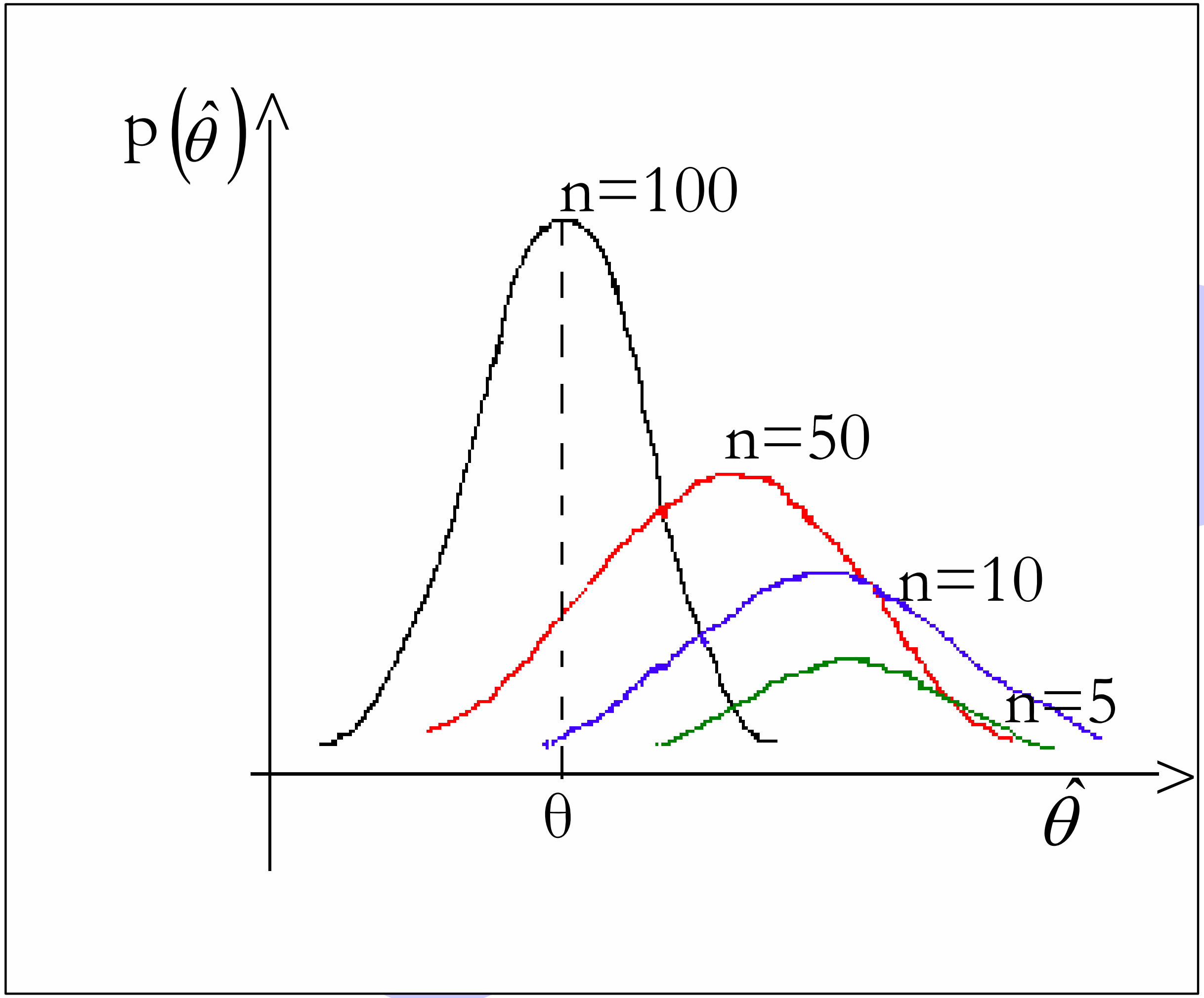
$\hat{\theta}$ coincide com θ ,o verdadeiro valor do parâmetro.

Consistência (ou coerência)

$\hat{\theta}$ é consistente se (além de ser justo) sua variância tende para zero, quando *n* é suficientemente grande (n $\rightarrow\infty$), isto é:

$$\boxed{E\left[\hat{\theta}\right]=\theta \quad e \quad Lim_{n\rightarrow\infty} Var\left[\hat{\theta}\right]=0}$$

Estimador Consistente:

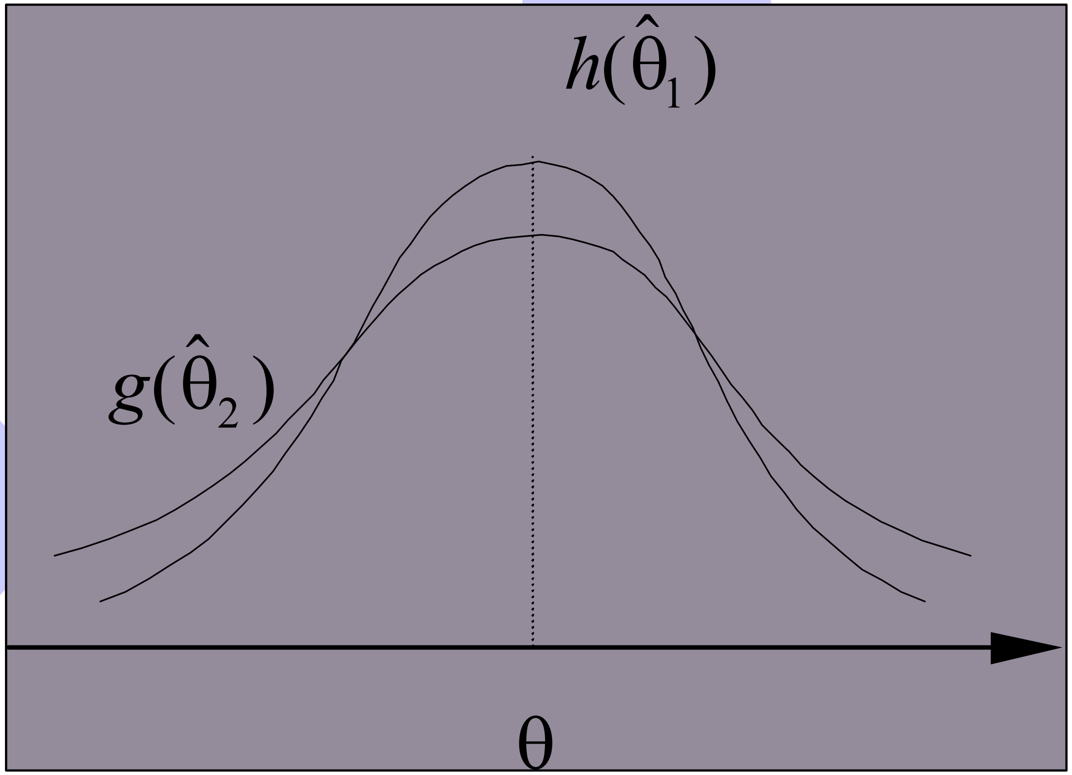


* **A propriedade da consistência garante que nossas estimativas melhoram à medida que o tamanho da amostra aumente.**

3) Eficiência

Seja $\hat{\theta}$ uma estimativa não-tendenciosa de θ. Diremos que é uma estimativa de variância mínima de θ, se para todas as estimativas $\hat{\theta}^*$ tais que $E(\hat{\theta}^*) = \theta$, tivermos $Var(\hat{\theta}) \leq Var(\hat{\theta}^*)$ para todo θ .

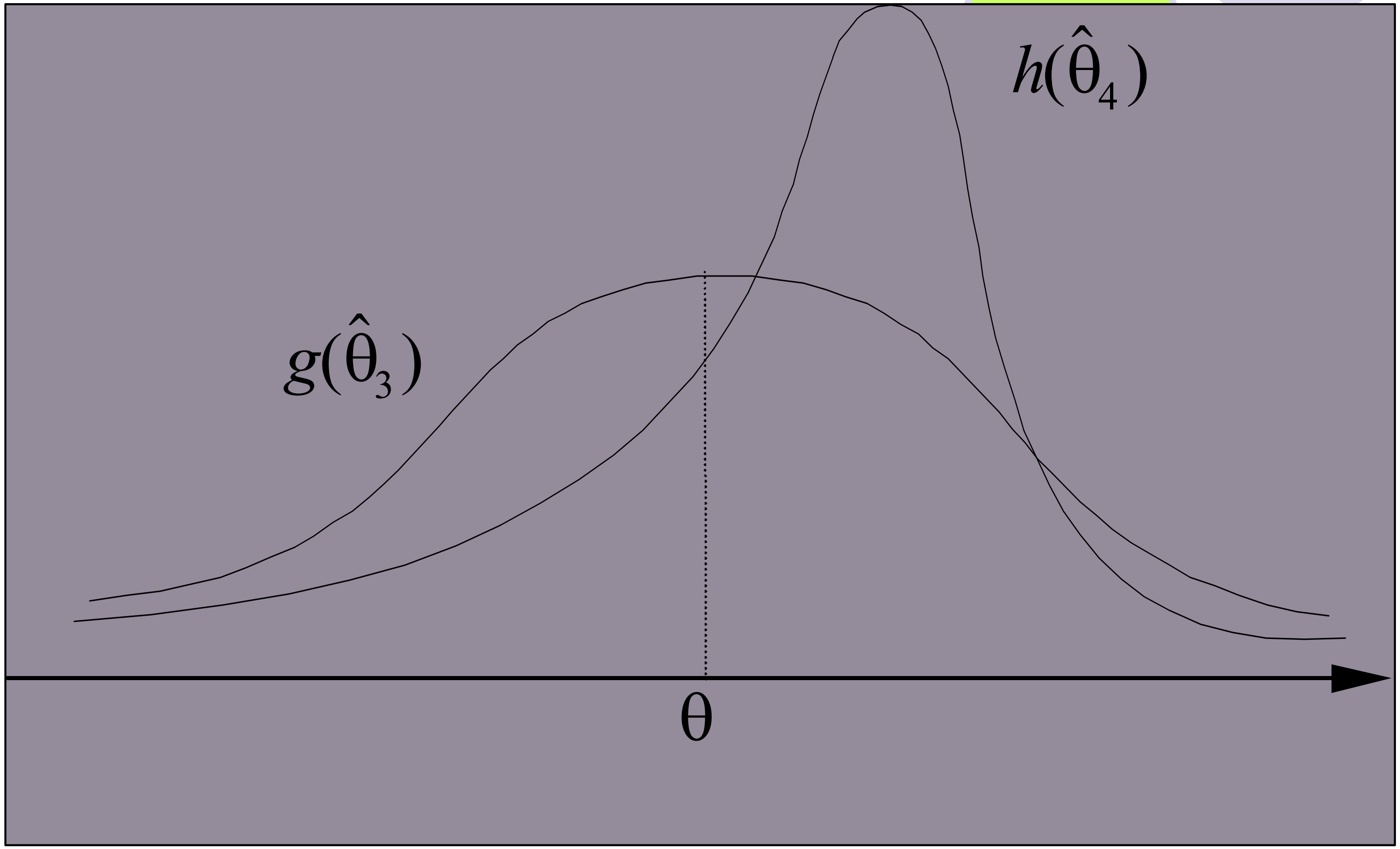
Assim, se $\hat{\theta}_1$ e $\hat{\theta}_2$ forem duas estimativas de θ, cuja *fdp* seja representada na figura seguinte, presumivelmente preferiríamos $\hat{\theta}_1$ a $\hat{\theta}_2$.



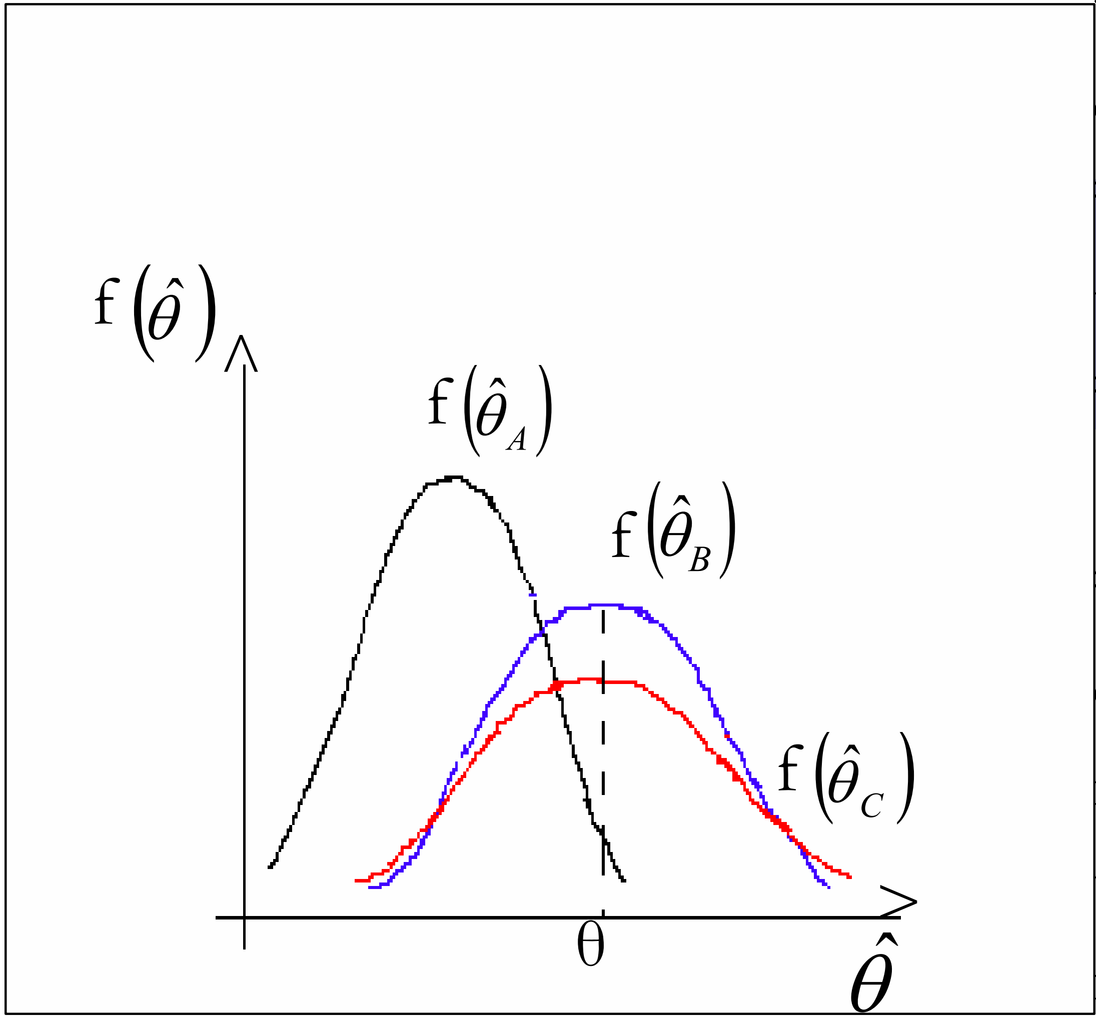
No caso de estimativas $\hat{\theta}_3$ e $\hat{\theta}_4$, (ver figura a seguir) a decisão não é tão evidente, porque $\hat{\theta}_3$ é não-tendenciosa, enquanto $\hat{\theta}_4$ é tendenciosa.

Todavia $Var(\hat{\theta}_3) > Var(\hat{\theta}_4)$. Isto significa que, enquanto em média $\hat{\theta}_3$ será próxima de θ, sua variância revela que desvios consideráveis em relação a θ não serão de surpreender. $\hat{\theta}_4$ por sua vez, tende a ser um tanto maior do que θ, em média, e no entanto poderá ser mais próxima de θ do que $\hat{\theta}_3$.

Estimativas $\hat{\theta}_3$ e $\hat{\theta}_4$:



Estimador Eficiente:



* $\hat{\theta}_A, \hat{\theta}_B, \hat{\theta}_C$ são estimadores de θ.

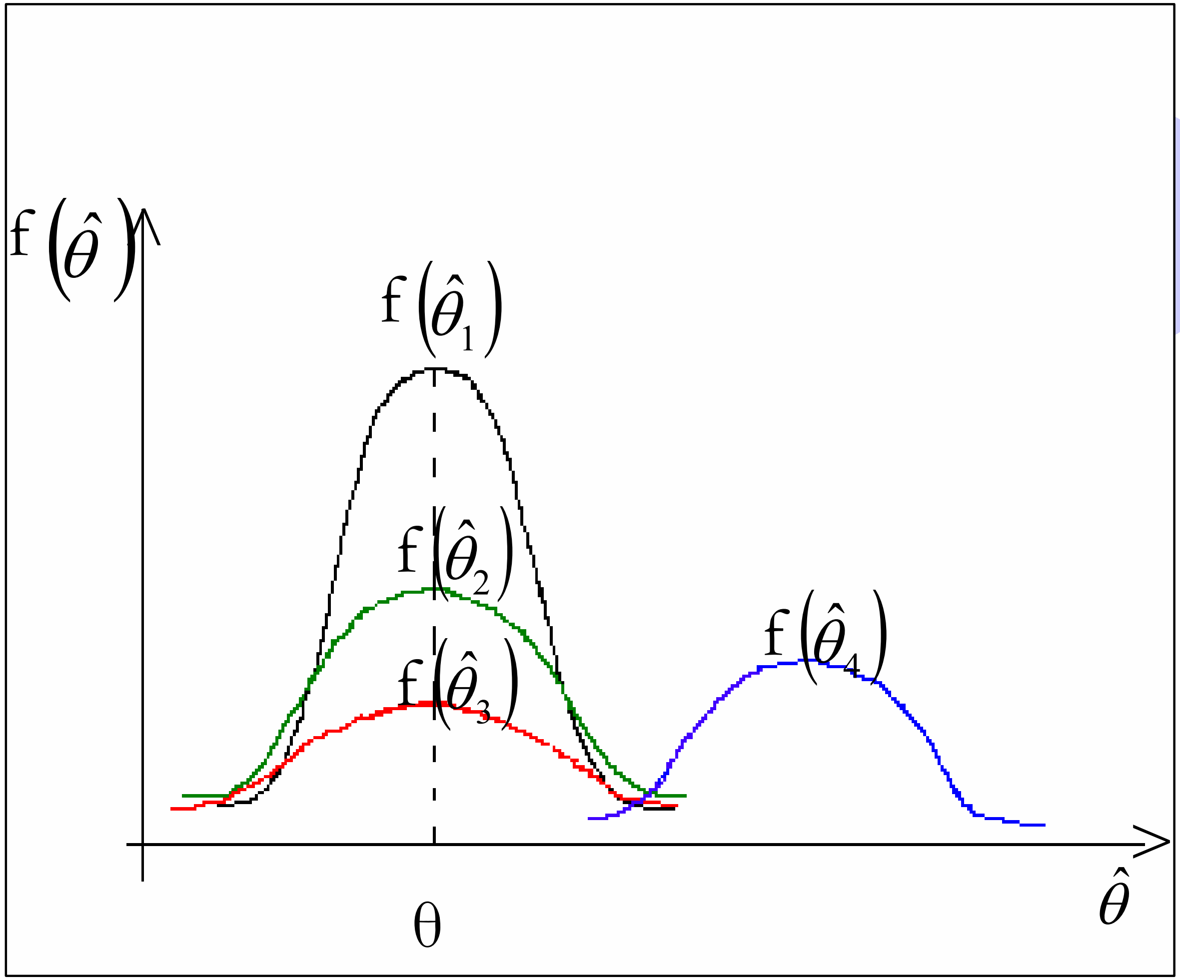
* $\hat{\theta}_B, \hat{\theta}_C$ são não tendenciosos.

* $\hat{\theta}_A$ é tendencioso, mas com variância mínima.

* $\hat{\theta}_B$ será um estimador eficiente de $\hat{\theta}$ se não houver nenhum outro estimador não tendencioso de $\hat{\theta}$ com menor variância.

Exemplo:

Observando o gráfico abaixo, responda:



a) Quais os estimadores justos de θ?

b) Quais os estimadores viezados de θ?

c) Qual o estimador de variância mínima de θ?

d) Qual o estimador de maior variância?

e) Qual deles você escolheria?

Distribuições por amostragem

Considerar todas as possíveis amostras de tamanho *n* retiradas da população. Para cada amostra calcular a estatística de interesse, obtendo-se desta maneira uma distribuição desses resultados originando uma distribuição por amostragem. Desta maneira pode-se obter as distribuições por amostragem da média, da variância, da proporção e de outras estatísticas.

Amostragem com ou sem reposição

- Se o processo de retirada for com reposição, N^n = número de amostras de tamanho *n* que poderão ser extraídas com reposição da população de tamanho N;

- Se o processo de retirada for sem reposição, o número combinatório C_N^n =número de amostras de tamanho *n* extraídas, sem reposição, da população de tamanho N.

Distribuição amostral das médias

Admita-se que todas as amostras possíveis de tamanho *n* são retiradas, **sem reposição**, de uma população finita de tamanho N.

Distribuição amostral das médias

Se a média e o desvio padrão da distribuição amostral das médias forem designados por $\mu(\bar{x})$ e $\sigma(\bar{x})$, e os valores correspondentes da população o forem por μ e σ, respectivamente, então:

$$\mu(\bar{x}) = \mu$$

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Distribuição amostral das médias

Se a população for infinita, ou se a amostragem for tomada **com reposição**, os resultados acima reduzir-se-ão a:

$$\mu(\bar{x}) = \mu$$

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Exemplos:

1) Consideremos uma população formada por 5 números: 2, 3, 6, 8, 11.

Considerar todas as amostras possíveis com 2 elementos cada uma, que possam ser retiradas desta população:

a) com reposição;

b) sem reposição.

Calcular a média da distribuição amostral das médias e o erro padrão.

2) Admite-se que as alturas de 3.000 estudantes do sexo masculino de uma universidade são normalmente distribuídas com média $\mu = 172{,}72$ cm e desvio padrão $\sigma = 7{,}62$ cm. Se forem obtidas 80 amostras de 25 estudantes cada uma, quais serão a média e o desvio padrão esperado da distribuição amostral das médias resultantes, se a amostragem for:

1º) Com reposição;

2º) Sem reposição.

Distribuição amostral da proporção

Seja X uma população infinita e p a probabilidade de sucesso de um certo evento. Logo, $1 - p = q$ será a probabilidade de não ocorrer o evento. Seja uma amostra aleatória de n elementos dessa população, e x o número de sucesso na amostra. Então X (freqüência absoluta) número de sucessos é uma variável aleatória Binomial de média np e variância npq.

Logo a distribuição de freqüência relativa $f = \frac{x}{n}$ será:

$$\mu(f) = E[f] = \left[\frac{x}{n}\right] = \frac{np}{n} = p$$

$$\sigma^2(f) = \sigma^2\left[\frac{x}{n}\right] = \frac{1}{n^2}\sigma^2[x] = \frac{npq}{n^2} = \frac{pq}{n}$$

Exemplo:

Seja uma população formada por 2 mulheres e 1 homem. Considerando-se sucesso a ocorrência de mulheres, retirar todas as amostras possíveis de tamanho 2 (n = 2), **com reposição**.