

Introdução à Tecnologia de Amostragem

Prof. Caio Azevedo

16 de agosto de 2011

- **Ponto de partida:** qual(ais) é (são) a(s) população(ões) de interesse? Quais são as questões relevantes com relação à população de interesse?
- **População:** conjunto de elementos com características semelhantes.
- **Amostra:** subconjunto da população.
- **Variável:** característica de interesse relacionada a cada elemento da população (amostra).
- **Parâmetro:** característica de interesse relacionada à população (ou parte dela) como um todo (ou da amostra).

Populações infinitas

- Obter conclusões acerca de parâmetros (de uma ou mais populações) com base em uma amostra (ou mais de uma).
- Usualmente: cada elemento tem a mesma probabilidade de ser selecionado.
 - Interesse: amostra aleatória Y_1, \dots, Y_n (conjunto de possíveis valores observados), regida por alguma distribuição de probabilidade (paramétrica).
 - Inferência a respeito dos parâmetros dessa distribuição.
 - Métodos de estimação, construção de IC's e testes de hipótese.

Populações finitas

- População composta por N elementos.
- Define-se alguma característica de interesse dessa população: média, variância, total, coeficiente de variação.
- Processo de amostragem: cada elemento pode ou não ter a mesma probabilidade de ser selecionado.
 - Interesse: amostra aleatória (selecionado segundo algum **plano amostral (PA)**) Y_1, \dots, Y_n (conjunto de possíveis valores observados). Pode se considerar, a rigor, que amostra é aleatória simples com reposição (AAS_s), embora na prática, não o seja.
 - Probabilidades de ocorrência de cada amostra: **função do plano amostral**.

Estrutura geral

- População alvo: população de interesse. **Exemplo: pessoas residentes no Brasil no mês de Agosto de 2011.**
- População referenciada: subconjunto da população alvo para a qual está disponível um sistema de referência. **Exemplo: pessoas residentes no Brasil no mês de Agosto de 2011 das quais se possui o endereço correto.**
- População amostrada: subconjunto da população referenciada da qual, efetivamente, é possível retirar uma amostra. **Exemplo: pessoas residentes no Brasil no mês de Agosto de 2011 das quais se possui o endereço correto e não residem em localidades de difícil acesso.**

Estrutura geral cont.

- Sistema de referência: banco de dados contendo informações sobre os elementos da população referenciada (organizado de modo a permitir implementar o PA).
- Unidade elementar: elemento da população portadora das informações de interesse. **Exemplo: Eleitor brasileiro.**
- Unidade amostral: entidade que será selecionada no processo de amostragem. Pode ser formada por uma ou mais de uma unidades elementares. **Exemplo: Domicílio.**
- Unidade resposta: entidade que fornece as informações de interesse relacionadas à unidade amostral. **Exemplo: Pessoa responsável pelo sustento do domicílio.**

Plano amostral (PA)

- PA: conjunto de procedimentos que definem como a amostra deve ser selecionada. Ela determina a probabilidade de cada elemento ser selecionado.
- A PA pode influenciar o comportamento do estimador: esperança, vício, erro quadrático médio.
- Ignorar a PA pode comprometer o processo de inferência, e.g., subestimar ou superestimar a variância do estimador.

Cont.

- Objetivos:
 - Definir a PA que produza uma amostra que leve a inferências apropriadas (junto com a escolha de um estimador apropriado).
 - Incorporar a estrutura da PA na construção e obtenção das características (esperança, variância) de estimadores.
- Amostra representativa : definição problemática.
- Amostra probabilística: selecionada segundo algum plano amostral (regido **totalmente** por mecanismos de sorteio aleatório).
- Amostra sistemática: selecionada segundo algum plano amostral (regido **parcialmente** por mecanismos de sorteio aleatório).

Tipos de amostra

Critério do “amostrista”	Procedimento de seleção	
	probabilístico	não-probabilístico
objetivo	amostra probabilísticas	amostras criteriosas
subjetivo	amostra quase-aleatória	amostra intencional

Classificação de amostras probabilísticas

- Probabilidade de seleção da unidade amostral: **igual** ou distinta.
- Unidade amostral: **uma unidade de resposta** (elementar) ou **elementos (conglomerado)**.
- Divisão em estratos: **não estratificada** ou **estratificada**.
- Número de estágios: **um único** ou mais de um (**somente dois**).
- Seleção de unidades: **aleatória** ou sistemática.

Erros (imprecisões) envolvidos

- Censo: analisa-se todos os elementos da população de interesse (não há erros relativos à processos de amostragem).
- Amostragem: analisa-se uma parte da população de interesse (há erros (imprecisões) relativos à processos de amostragem).
- Erros (imprecisões) amostrais: desvios existentes entre as estimativas e os verdadeiros valores dos parâmetros decorrentes do processo amostral (controláveis através do PA).
- Erros não amostrais: desvios existentes entre as estimativas e os verdadeiros valores dos parâmetros decorrentes de fatores não inerentes ao processo amostral (não-controláveis através do PA).

Definições/Notações

- **População ou universo:** $\mathcal{U} = \{1, 2, \dots, N\}$, em que N é o tamanho da população, que pode ser conhecido ou desconhecido.
- **Elemento populacional:** elemento pertencente à \mathcal{U} , i.e., $i \in \mathcal{U}$.
- **Característica(s) de interesse:** variável ou vetor de variáveis associado a cada elemento da população.
- Representação: $y_i, i \in \mathcal{U}$ ou $\mathbf{y}_i = (y_{i1}, \dots, y_{ip}), i \in \mathcal{U}$.
- Usaremos letras maiúsculas para representar variáveis aleatórias (e.g., X) e letras minúsculas para representar variáveis não aleatórias ou valores observados de variáveis aleatórias (e.g., x).

Cont.

- **Parâmetro populacional:** denota o vetor correspondente a todos os valores de uma variável de interesse (ou um vetor de variáveis).
Notação: $\mathbf{d} = (y_1, \dots, y_N)$ ou $\mathbf{d} = (\mathbf{y}_1, \dots, \mathbf{y}_N)$.
- **Função paramétrica populacional (ou parâmetro):** característica numérica qualquer de uma população. Notação: $\theta(\mathbf{d})$ ou θ .
- **Amostra ordenada:** qualquer sequência de n unidades de \mathcal{U} , ou seja, $\mathbf{s} = (k_1, \dots, k_n), k_i \in \mathcal{U}$.
- **Espaço amostral:** conjunto de todas as amostras de \mathcal{U} , $\mathcal{S}(\mathcal{U})$ e a subclasse de todas as amostras de tamanho n , $\mathcal{S}_n(\mathcal{U})$.

Exemplo

- Considere uma população formada por três domicílios, $\mathcal{U} = \{1, 2, 3\}$ e que se observam as seguintes variáveis: nome (do chefe), sexo, idade, fumante ou não, renda bruta (mensal em salários mínimos) familiar e número de trabalhadores.

Cont.

Tabela: População de três domicílios

Variável	Valores			Notação
Unidade	1	2	3	i
nome do chefe	Ada	Beto	Ema	a_i
sexo	0	1	0	x_i
idade	20	30	40	y_i
fumante	0	1	1	g_i
renda bruta familiar	12	30	18	f_i
nº de trabalhadores	1	3	2	t_i

Exemplo

- Parâmetro populacional
 - Variável idade: $\mathbf{d} = (20, 30, 40)$.
 - Variáveis (f,t): $\mathbf{d} = \begin{pmatrix} 12 & 30 & 18 \\ 1 & 3 & 2 \end{pmatrix}$
- Função paramétrica (chamaremos de parâmetros):
- Média das variáveis renda e número de trabalhadores,
 $\theta = (20, 2)$
- Renda média por trabalhador
 $\theta = 10$

Parâmetros de interesse

- Em geral, os parâmetros de interesse associados à uma variável são:

- Total populacional: $\tau = \sum_{i=1}^N y_i$

- Média populacional: $\mu = \frac{\tau}{N} = \frac{1}{N} \sum_{i=1}^N y_i$.

- Variância populacional: $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$

- Proporção populacional: $p = \frac{1}{N} \sum_{i=1}^N y_i$, quando $y_i \in \{0, 1\}, \forall i$.

Planejamento amostral

- **Planejamento amostral:** mecanismo que associa a probabilidade de ocorrência de cada amostra possível $P(\mathbf{s})$, ou seja

$$P(.) : \mathcal{S}(U) \rightarrow [0, 1]$$
$$P(\mathbf{s}) \geq 0, \forall \mathbf{s} \quad , \quad \sum_{\mathbf{s}:\mathbf{s} \in \mathcal{S}} P(\mathbf{s}) = 1$$

- Conjunto das amostras possíveis sobre o plano amostral A : \mathcal{S}_A .

Cont. do Exemplo

- Exemplos de amostra: $\mathbf{s}_1 = (1, 2)$, $\mathbf{s}_2 = (2, 1)$, $\mathbf{s}_3 = (1, 1, 3)$,
 $\mathbf{s}_4 = (3)$, $\mathbf{s}_5 = (2, 2, 1, 3, 2)$
- Conjunto de todas as amostras possíveis (com reposição):
 $\mathcal{S} = (\mathcal{U}) = \{(1), (2), (3), (1, 1), (1, 2), \dots, (2, 2, 1, 3, 2), \dots\}$
- Conjunto de todas as amostras possíveis de tamanho 2 (com reposição):
 $\mathcal{S}_2(\mathcal{U}) = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\}$.
- Exercício: obter todas as amostras possíveis para um sorteio sem reposição.

Cont. do Exemplo: planos amostrais

- Plano A: Sorteia-se com igual probabilidade um elemento de \mathcal{U} e anota-se a unidade sorteada. Este elemento é devolvido à população e sorteia-se um segundo elemento do mesmo modo.

$$\mathcal{S}_A = \{11, 12, 13, 21, 22, 23, 31, 32, 33\}.$$

$$P_A(\mathbf{s}) = \begin{cases} 1/9, & \text{se } i \in \mathbf{s} \\ 0, & \text{se } i \notin \mathbf{s} \end{cases}$$

Cont. do Exemplo: planos amostrais

- Plano B: Sorteia-se com igual probabilidade um elemento de \mathcal{U} e anota-se a unidade sorteada. Este elemento é retirado da população e sorteia-se um segundo elemento do mesmo modo.

$$\mathcal{S}_A = \{12, 13, 21, 23, 31, 32\}.$$

$$P_A(\mathbf{s}) = \begin{cases} 1/6, & \text{se } i \in \mathbf{s} \\ 0, & \text{se } i \notin \mathbf{s} \end{cases}$$

Cont. do Exemplo: planos amostrais

- Plano C: Sorteia-se um elemento de \mathcal{U} com probabilidade proporcional ao número de trabalhadores. Sem repor o domicílio selecionado, sorteia-se um segundo também com probabilidade igual ao número de trabalhadores.

$$\mathcal{S}_C = \{12, 13, 21, 23, 31, 32\}.$$

$$\begin{aligned} P(12) &= \frac{1}{6} \times \frac{3}{5} = \frac{1}{10} & ; & & P(21) &= \frac{3}{6} \times \frac{1}{3} = \frac{1}{6} \\ P(13) &= \frac{1}{6} \times \frac{2}{5} = \frac{1}{15} & ; & & P(31) &= \frac{2}{6} \times \frac{1}{4} = \frac{1}{12} \\ P(23) &= \frac{3}{6} \times \frac{2}{3} = \frac{1}{3} & ; & & P(32) &= \frac{2}{6} \times \frac{3}{4} = \frac{1}{4} \end{aligned}$$

Inferência

- Dada uma amostra $\mathbf{s} = (k_1, k_2, \dots, k_n)$ tem-se um vetor de características \mathbf{y}_{k_j}
- **Dados da amostra \mathbf{s} :** (vetor ou matriz) $\mathbf{d}_{\mathbf{s}} = (y_{k_1}, \dots, y_{k_n}), k_i \in \mathbf{s}$.
- **Conjunto de todas as amostras possíveis:** Nesse caso, teremos um vetor (matriz) aleatória:

$$\mathbf{D}_{\mathbf{s}} = \mathbf{Y} = (Y_1, \dots, Y_i, \dots, Y_n),$$

em que Y_i é a va que representa os valores possíveis de ocorrerem na i -ésima posição da amostra.

- No caso multivariado: $\mathbf{d}_{\mathbf{s}} = (\mathbf{y}_{k_i}, k_i \in \mathbf{s})$ e $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$.

Cont.

- **Estatística (estimador):** função da amostra $T = h(\mathbf{Y})$, $t = h(\mathbf{y})$ (valor observado).
- **Distribuição amostral de uma estatística segundo um PA, A:**

$$p_h = P_A(\mathbf{s} \in \mathcal{S}_A : h(\mathbf{y}) = h) = P(h) \quad (1)$$

- **Valor esperado de T sob o plano amostral A:** $\mathcal{E}_A(T) = \sum_h h p_h$.
- **Variância de T sob o plano amostral A:**

$$\mathcal{V}_A(T) = \sum_h (h - \mathcal{E}_A(T))^2 p_h = \mathcal{E}_A(T^2) - \mathcal{E}_A^2(T).$$

Cont.

- Sejam T e G duas estatísticas.
- **Covariância entre T e G** : $Cov(T, G) = \mathcal{E}_A(TG) - \mathcal{E}_A(T)\mathcal{E}_A(G)$.
- **Correlação entre T e G** : $Corre(T, G) = \frac{Cov(T, G)}{\sqrt{\mathcal{V}_A(T)\mathcal{V}_A(G)}}$.

Cont. do Exemplo: Estatísticas e distribuições amostrais

- Considere que o interesse reside nas variáveis (f,t):

$$\mathbf{d} = \begin{pmatrix} 12 & 30 & 18 \\ 1 & 3 & 2 \end{pmatrix}$$

- Considere os planos amostrais A e B e a estatística $R = h(\mathbf{D}_s)$:
razão entre o total da renda familiar e o número de trabalhadores.

Cont. do Exemplo: Estatísticas e distribuições amostrais

■ Plano amostral A:

\mathbf{s}	11	12	13	21	22	23	31	32	33
$P(\mathbf{s})$	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9
$h(\mathbf{d}_s) = r$	12	10,5	10	10,5	10	9,6	10	9,6	9

■ Distribuição amostra de R sob o plano amostral A:

h	9	9,6	10	10,5	12
p_h	1/9	2/9	3/9	2/9	1/9

Cont. do Exemplo: Estatísticas e distribuições amostrais

■ Plano amostral B:

s	12	13	21	23	31	32
$P(\mathbf{s})$	1/6	1/6	1/6	1/6	1/6	1/6
$h(\mathbf{d}_s) = r$	10,5	10	10,5	9,6	10	9,6

■ Distribuição amostra de R sob o plano amostral A:

h	9,6	10	10,5
p_h	1/3	1/3	1/3

Incorporação do planejamento amostral

- Levar em consideração a probabilidade de cada indivíduo aparecer na amostra e não a “distribuição de frequências” da população.
- Estimadores que levem em consideração, na sua fórmula, a estrutura do planejamento amostral.
- Incorporar o planejamento amostral no cálculo da esperança, variância etc do estimador.