AMOSTRAGEM

Unidade 3

Amostragem – Visão Geral

Definições e Notação para População de Pesquisa

$U = \{1, 2, \ldots, i, \ldots, N\}$ → conjunto de N rótulos distintos

N = tamanho da população de pesquisa = #U

i → rótulo para unidade genérica da população

y → variável de pesquisa / de interesse

y_i → valor da variável y para unidade i

$Y_U = \{y_1, y_2, \ldots, y_N\}$ vetor populacional

Parâmetros-Alvo (de Interesse)

Total populacional → $Y = \sum_{i=1}^{N} y_i = \sum_{i \in U} y_i$

Média populacional → $\bar{Y} = Y / N = \sum_{i \in U} y_i / N$

Variância populacional →

$$S_y^2 = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{Y})^2 = \frac{1}{N-1} \left[\sum_{i \in U} y_i^2 - N\bar{Y}^2 \right]$$

Parâmetros-Alvo (de Interesse)

Seja z outra variável de pesquisa, tomando valores z_i, $i \in U$.

Razão populacional → $R = \sum_{i \in U} y_i / \sum_{i \in U} z_i$

Covariância populacional

$$S_{yz} = \frac{1}{N-1} \sum_{i \in U} (y_i - \bar{Y})(z_i - \bar{Z}) = \frac{1}{N-1} \left[\sum_{i \in U} y_i z_i - N\bar{Y}\bar{Z} \right]$$

Coeficiente de correlação populacional

$$\rho_{yz} = S_{yz} / S_y S_z$$

Amostra

Uma **amostra** $s=\{i_1, i_2, ..., i_n\}$ é qualquer **subconjunto** não vazio de unidades da população U ($s \subset U$) selecionadas para observação visando estimar os parâmetros de interesse.

Uma amostra de tamanho n é uma amostra contendo n **unidades distintas** tiradas da população U.

$i \in s$ representa um rótulo de unidade incluída na amostra.

Dados e Estatísticas Amostrais

Dados amostrais → $y_{i_1}, y_{i_2}, \ldots, y_{i_n}$

Total (soma) amostral

$$t(s) = t = \sum_{i \in s} y_i$$

Média amostral

$$\bar{y} = t / n = \frac{1}{n} \times \sum_{i \in s} y_i$$

Estimação

Suponha que o parâmetro-alvo é o total populacional Y.

O objetivo principal é usar os dados amostrais $y_{i_1}, y_{i_2}, \ldots, y_{i_n}$ para estimar $Y = \sum_{i \in U} y_i$.

Um objetivo secundário é conseguir medir / estimar também a precisão / margem de erro da estimativa produzida para Y.

Estimador Linear

Um estimador linear $\hat{Y}_w$ é uma combinação linear dos valores amostrais y_i com pesos w_i, isto é:

$$\hat{Y}_w = \sum_{i \in s} w_i y_i$$

Problema: como determinar os pesos w_i?

Precisamos de critérios!

Amostragem Probabilística

É um procedimento de amostragem satisfazendo as condições enumeradas a seguir.

1. O espaço amostral *S*, o conjunto de todas as amostras s possíveis, é bem definido.

2. Uma probabilidade p(s) conhecida (ou calculável) é associada a cada amostra $s \in S$, e $\sum_{s \in S} p(s) = 1$.

Amostragem Probabilística

3. Cada unidade $i \in U$ tem uma probabilidade não nula de ser selecionada para a amostra, isto é: $P(i \in s) > 0 \ \forall i \in U$.

4. Uma única amostra s ($s \in S$) é selecionada para observação usando um mecanismo de aleatorização (sorteio) tal que a amostra s é escolhida com probabilidade p(s).

Exemplo 3.1

População de 4 unidades (N=4) mulheres, de quem foi indagado o número de filhos tidos vivos (y).

Rótulo da unidade (i)	1	2	3	4	Total
Valor y_i	0	0	2	1	3

Existem $\binom{4}{2} = 6$ amostras possíveis de tamanho n=2.

Conjunto de Todas as Amostras Possíveis

S = { (1;2) ; (1;3); (1;4); (2;3) ; (2;4) ; (3;4) }.

Amostras selecionadas com igual probabilidade →

Cada amostra tem probabilidade de ser selecionada = 1/6 → p(s)=1/6 $\forall s \in S$.

Conjunto de Todas as Amostras Possíveis

Amostras	Unidades na Amostra	Soma Amostral (t)	Probabilidades p(s)
1	(1;2)	0,0	1/6
2	(1;3)	2,0	1/6
3	(1;4)	1,0	1/6
4	(2;3)	2,0	1/6
5	(2;4)	1,0	1/6
6	(3;4)	3,0	1/6
Total		9,0	1,0

Distribuição da Soma Amostral

Valores possíveis de t	0,0	1,0	2,0	3,0
com probabilidade p(s)	1/6	2/6	2/6	1/6

O valor esperado de t é:

$$E_p(t) = \sum_{s \in S} t(s)\, p(s)$$

$$= 0{,}0 \times \frac{1}{6} + 1{,}0 \times \frac{2}{6} + 2{,}0 \times \frac{2}{6} + 3{,}0 \times \frac{1}{6}$$

$$= \frac{9}{6}$$

$$= 1{,}5$$

Exemplo 3.1

Porém o total populacional é:

$$Y = \sum_{i=1}^{4} y_i = 3$$

Como $1{,}5 = E_p(t) \neq Y = 3$, dizemos que t é um estimador viciado de Y sob o plano amostral p(s) adotado.

Como podemos "corrigir" o estimador de modo que fique não viciado?

Resposta: multiplicando por 2 o valor total amostral t.

Exemplo 3.1

Novo estimador do total populacional: $\hat{Y} = 2 \times t$

Estimador na forma linear →

$$\hat{Y} = 2 \times t = \sum_{i \in s} 2 \times y_i = \hat{Y}_w$$

Valor de $\hat{Y} = 2 \times t$	0,0	2,0	4,0	6,0
com probabilidade p(s)	1/6	2/6	2/6	1/6

Exemplo 3.1

O valor esperado de $\hat{Y} = 2 \times t$ é:

$$E_p(\hat{Y}) = \sum_{s \in S} \hat{Y}_s \; p(s)$$

$$= 0{,}0 \times \frac{1}{6} + 2{,}0 \times \frac{2}{6} + 4{,}0 \times \frac{2}{6} + 6{,}0 \times \frac{1}{6}$$

$$= \frac{18}{6} = 3$$

Como $E_p(\hat{Y}) = 3 = Y$, dizemos que $\hat{Y} = 2 \times t$ é um estimador não viciado de Y sob o plano amostral p(s) adotado.

Exemplo 3.1 – Lição Importante

É essencial ter algum critério para escolha de estimadores.

Critério 1

Estimadores devem ser não viciados, ou ao menos aproximadamente não viciados.

A Distribuição de Aleatorização

- A função p(s) definida no conjunto *S* de todas as amostras possíveis é uma **distribuição de probabilidades.**
- A distribuição de probabilidades p(s), s$\in$*S*, é chamada **distribuição de aleatorização.**
- Na amostragem ***probabilística***, inferências são feitas considerando a distribuição de ***aleatorização***.

A Distribuição de Aleatorização

- Tais inferências são baseadas no plano amostral, onde a fonte de variação ou incerteza é a repetição *hipotética* do processo de amostragem utilizando p(s), que resultaria em diferentes amostras s_1, s_2, ... $\in$ *S*.
- A distribuição de $\hat{Y} = 2 \times t = \sum_{i \in s} 2 \times y_i = \hat{Y}_w$ determinada por p(s) é chamada de distribuição amostral do estimador.
- Suas propriedades é que vamos estudar para avaliar se é um bom estimador para estimar o total populacional Y.

Obtenção de Estimadores Não Viciados para o Total

- Trabalhar com a distribuição p(s) é complicado.
- O número total de amostras possíveis cresce muito rapidamente com N e com n.
- Por exemplo, o número de amostras sem reposição de tamanho n de uma população com N unidades é $\binom{N}{n}$.
- A saída é usar propriedades simplificadoras desta distribuição.

Uma Propriedade Importante

$$\Pr(i \in s) = \pi_i = \sum_{s \supset i} p(s)$$

Se tomarmos o valor do inverso de sua probabilidade de seleção ($1 / \pi_i$) como peso (w_i) de uma unidade amostrada, é fácil verificar que o estimador dado por

$$\hat{Y}_w = \sum_{i \in s} w_i \, y_i = \sum_{i \in s} \frac{1}{\pi_i} y_i = \sum_{i \in s} \pi_i^{-1} \, y_i$$

é não viciado para o total populacional Y.

Exemplo 3.1 - Continuação

População de 4 unidades (N=4) mulheres, de quem foi indagado o número de filhos tidos vivos (y).

Rótulo da unidade (i)	1	2	3	4	Total
Valor y_i	0	0	2	1	3
Probabilidade de inclusão π_i	3/6 =½	3/6 =½	3/6 =½	3/6 =½	--

Pesos Amostrais no Exemplo 3.1

$w_i = 1/\pi_i = 1 / ½ = 2$

Estimador ponderado do total

$$\hat{Y}_w = \sum_{i \in s} w_i y_i = \sum_{i \in s} \pi_i^{-1} y_i = \sum_{i \in s} 2 y_i = 2t$$

E já se mostrou que este estimador é não viciado para Y.

Exemplo 3.2

Considere a mesma população fictícia do exemplo 3.1.

Considere agora o plano amostral que retira amostras de tamanho 2 dessa população segundo o plano amostral dado na tabela a seguir.

Amostras	Unidades na Amostra	Soma Amostral (t)	Probabilidade p(s)
1	(1;2)	0,0	0,00
2	(1;3)	2,0	0,20
3	(1;4)	1,0	0,15
4	(2;3)	2,0	0,20
5	(2;4)	1,0	0,15
6	(3;4)	3,0	0,30
Total		9,0	1,00

Exemplo 3.2

Chamaremos este plano amostral de plano 2.

Use as informações acima para:

1. Verificar que o estimador baseado na soma amostral (t) é viciado para estimar o total populacional Y;
2. Obter / definir um estimador não viciado para o total populacional Y;
3. Comente sobre o uso de um plano amostral em que as diferentes amostras têm probabilidades desiguais de serem selecionadas. Surpresas? Dificuldades?

Distribuição do Total Amostral sob Plano 2

Valor de t	0,0	1,0	2,0	3,0
com probabilidade p(s)	0,0	0,3	0,4	0,3

O valor esperado de t é:

$$E_p(t) = \sum_{s \in S} t(s)\, p(s)$$

$$= 0{,}0 \times 0 + 1{,}0 \times 0{,}30 + 2{,}0 \times 0{,}40 + 3{,}0 \times 0{,}30$$

$$= 2{,}0 \quad < \quad 3 = Y$$

Exemplo 3.2

Para obter estimador não viciado, devemos calcular pesos adequados para unidades amostrais.

Estes requerem calcular as probabilidades de inclusão na amostra.

Unidade (i)	1	2	3	4
Probabilidade π_i	0,35	0,35	0,70	0,60
Peso w_i	20/7= 2,857	20/7= 2,857	10/7= 1,429	5/3= 1,667

Estimador do Total com Pesos Adequados

Amostra	Unidades na Amostra	Total Amostral Ponderado ($\hat{Y}_W$)	Probabilidade p(s)	Valor do produto
1	(1;2)	0,0	0,00	0,00
2	(1;3)	$2,0 \times 10/7$	0,20	4/7
3	(1;4)	$1,0 \times 5/3$	0,15	1/4
4	(2;3)	$2,0 \times 10/7$	0,20	4/7
5	(2;4)	$1,0 \times 5/3$	0,15	1/4
6	(3;4)	$2,0 \times 10/7 + 1,0 \times 5/3$	0,30	6/7 + 1/2
Total				3

Notas

- Estimador $\hat{Y}_w$ tem valor esperado igual ao total populacional Y → logo é NÃO VICIADO.
- O fato de que a amostra (1;2) tem probabilidade nula de ser selecionada viola os critérios definidos para que o plano de amostragem 2 seja chamado de amostragem probabilística? Sim ou não? Porquê?

Notas

- Temos agora duas opções para selecionar amostras (de tamanho 2) da população U, e estimar o total populacional Y sem vício.
- Qual das duas é melhor?

Estratégia 1: seleção equiprovável de pares (amostras) com estimador ponderado

Valor de $\hat{Y}_w = 2 \times t$	0,0	2,0	4,0	6,0
com probabilidade p(s)	1/6	2/6	2/6	1/6

Estratégia 2: seleção de amostras com probabilidades desiguais, e estimador ponderado

Valor de $\hat{Y}_w$	5/3	20/7	20/7 + 5/3
com probabilidade p(s)	0,30	0,40	0,30

Como Escolher a Melhor Estratégia?

Medindo o ***afastamento esperado*** entre o valor do estimador e o valor do total populacional desconhecido (Y).

Para isso, usamos a variância do estimador, dada por:

$$V_p(\hat{Y}) = \sum_{s \in S} (\hat{Y} - Y)^2 \times p(s)$$

ou o desvio padrão do estimador, dado por

$$DP_p(\hat{Y}) = \sqrt{V_p(\hat{Y})} = \sqrt{\sum_{s \in S} (\hat{Y} - Y)^2 \times p(s)}$$

Variâncias dos Estimadores sob Duas Estratégias

Amos -tra	Unidades na Amostra	Estimador sob E2	Probabilidade p(s) sob E2	Estimador sob E1	Probabilidade p(s) sob E1
1	(1;2)	0,0	0,00	0,0	1/6
2	(1;3)	2,0 × 10/7	0,20	4,0	1/6
3	(1;4)	1,0 × 5/3	0,15	2,0	1/6
4	(2;3)	2,0 × 10/7	0,20	4,0	1/6
5	(2;4)	1,0 × 5/3	0,15	2,0	1/6
6	(3;4)	2,0×10/7 + 1,0×5/3	0,30	6,0	1/6
Var.	--	1,24	--	3,67	--

Conclusões

- Ambas as estratégias permitem usar **estimadores não viciados** do total Y.
- A estratégia 2 tem o **estimador com menor variância**, e deve ser preferida à estratégia 1, pois o tamanho das amostras é o mesmo.
- **Minimizar a variância** é o critério de desempate para escolha entre estratégias não viciadas de amostragem e estimação de igual custo total.

Teoria Básica

Sejam δ_1, δ_2, ..., δ_N variáveis aleatórias indicadoras, tal que

$$\delta_i = I(i \in s) = \begin{cases} 1 & \text{se} \quad i \in s \\ 0 & \text{se} \quad i \notin s \end{cases} \text{ para qualquer } i \in U.$$

As δ_i são variáveis indicadoras do evento inclusão da unidade i na amostra s.

Exemplo 3.1

Para N=4 e n=2, as amostras possíveis podem ser representadas por:

s	Rótulos	δ_1	δ_2	δ_3	δ_4
1	1;2	1	1	0	0
2	1;3	1	0	1	0
3	1;4	1	0	0	1
4	2;3	0	1	1	0
5	2;4	0	1	0	1
6	3;4	0	0	1	1

Teoria Básica

Cada amostra fica univocamente determinada pelas variáveis indicadoras correspondentes.

As variáveis indicadoras dependem da amostra s, apesar de não termos indicado isto explicitamente em nossa notação.

Então as probabilidades de seleção ou inclusão na amostra, denotadas π_i, são definidas como:

$$\pi_i = P(\delta_i = 1) = E_p(\delta_i) = \sum_{s \supset i} p(s) \quad \forall i \in U.$$

Teoria Básica

As probabilidades de inclusão π_i são ditas de primeira ordem.

Precisaremos também definir as probabilidades de inclusão de segunda ordem, denotadas π_{ij}, dadas por:

$$\pi_{ij} = P(\delta_i \delta_j = 1) = E_p(\delta_i \delta_j) = \sum_{s \supset i,j} p(s) \quad \forall i, j \in U.$$

Note que quando i=j, $\pi_{ij}=\pi_{ii}=\pi_i \ \forall i \in U$.

Logo:

$$V_p(\delta_i) = \pi_i (1 - \pi_i)$$

$$COV_p(\delta_i ; \delta_j) = \pi_{ij} - \pi_i \pi_j$$

Um Método Geral de Prova em Amostragem

Este método se baseia nas variáveis indicadoras $\delta_1, \delta_2, ..., \delta_N$.

Uma propriedade importante das variáveis indicadoras é que:

$$\sum_{i \in s} \delta_i(s) = \sum_{i \in U} \delta_i(s)$$

Segue também que $\sum_{i \in s} y_i = \sum_{i \in s} \delta_i y_i = \sum_{i \in U} \delta_i y_i$.

Note que o truque é converter a soma amostral em uma soma na população.

Seja $Y = \sum_{i \in U} Y_i$ (total populacional) o parâmetro alvo.

Estimador Linear do Total

Um estimador linear de Y é sempre da forma

$$\hat{Y}_w = \sum_{i \in s} w_i y_i = \sum_{i \in U} w_i y_i \delta_i$$

onde w_i é o peso da unidade i.

Para que o estimador linear de Y seja não viciado, é preciso que:

$$E_p(\hat{Y}) = Y \Leftrightarrow \sum_{i \in U} w_i y_i E_p(\delta_i) = \sum_{i \in U} y_i \Leftrightarrow \sum_{i \in U} w_i \pi_i y_i = \sum_{i \in U} y_i$$

Um Método Geral de Prova em Amostragem

Esta relação só será válida para quaisquer valores populacionais y_i da variável de pesquisa caso

$$w_i \times \pi_i = 1 \ \forall i \in U.$$

Portanto a condição para que o estimador de total $\hat{Y}_w = \sum_{i \in s} w_i y_i$ seja SEMPRE não viciado é que os pesos das unidades na amostra sejam iguais ao inverso das respectivas probabilidades de inclusão:

$$w_i = \pi_i^{-1} \ \forall i \in U.$$

Um Método Geral de Prova em Amostragem

Logo o estimador não viciado de total fica dado por

$$\hat{Y}_w = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} \pi_i^{-1} y_i = \hat{Y}_{HT}$$

$\Rightarrow$ Estimador de **Horvitz-Thompson**

Este estimador está definido para qualquer plano amostral, desde que $\pi_i > 0 \ \forall i \in U$.

Propriedades do Estimador de Horvitz-Thompson

$$E_p\left(\hat{Y}_{HT}\right) = Y$$

Prova:

$$E_p\left(\hat{Y}_{HT}\right) = E_p\left[\sum_{i\in U}\delta_i y_i/\pi_i\right] = \sum_{i\in U} E_p(\delta_i) y_i/\pi_i = \sum_{i\in U} y_i = Y$$

Esta propriedade vale para qualquer população, variável de interesse y e plano amostral, desde que $\pi_i>0\ \forall i\in U$.

Variância do Estimador Horvitz-Thompson

$$V_p\left(\hat{Y}_{HT}\right) = \sum_{i\in U}\pi_i(1-\pi_i)\left(\frac{y_i}{\pi_i}\right)^2 + \sum_{i\in U}\sum_{j\neq i\in U}\left(\pi_{ij}-\pi_i\pi_j\right)\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right)$$

Esta é a chamada forma de Horvitz-Thompson da variância.

Existe uma outra forma para esta variância, que vamos conhecer mais tarde.

Prova

$$V_p\left(\hat{Y}_{HT}\right) = V_p\left[\sum_{i\in U}\left(\frac{y_i}{\pi_i}\right)\delta_i\right]$$

$$= \sum_{i\in U}\left(\frac{y_i}{\pi_i}\right)^2 V_p(\delta_i) + \sum_{i\in U}\sum_{j\neq i}\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right)COV_p\left(\delta_i;\delta_j\right)$$

$$= \sum_{i\in U}\left(\frac{y_i}{\pi_i}\right)^2 \pi_i(1-\pi_i) + \sum_{i\in U}\sum_{j\neq i}\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right)\left(\pi_{ij}-\pi_i\pi_j\right)$$

Estimador da Variância do Estimador de Total

Um estimador não viciado da variância do estimador de Horvitz-Thompson do total é dado por

$$\hat{V}_1\left(\hat{Y}_{HT}\right) = \sum_{i\in s}\frac{\pi_i(1-\pi_i)}{\pi_i}\left(\frac{y_i}{\pi_i}\right)^2 + \sum_{i\in s}\sum_{j\neq i}\frac{\left(\pi_{ij}-\pi_i\pi_j\right)}{\pi_{ij}}\left(\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}\right)$$

Uma Forma Alternativa para a Variância

Para planos amostrais de tamanho pré-fixado, pode-se demonstrar que uma forma equivalente da variância do estimador de Horvitz-Thompson é dada pela expressão de Sen-Yates-Grundy a seguir.

$$V_p\left(\hat{Y}_{HT}\right) = \sum_{i \in U} \sum_{j>i} \left(\pi_i \pi_j - \pi_{ij}\right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

Note a troca do sinal da diferença de probabilidades de inclusão em relação à fórmula anterior.

Outro Estimador da Variância do Estimador de Total

$$\hat{V}_{SYG}\left(\hat{Y}_{HT}\right) = \sum_{i \in s} \sum_{j>i} \frac{\left(\pi_i \pi_j - \pi_{ij}\right)}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2$$

Obtido / motivado a partir da forma de Sen-Yates-Grundy para a variância do estimador de total.

Não coincide com o estimador de variância derivado a partir da expressão de Horvitz-Thompson.

Notas

- ✓ Fácil derivar estimadores de total e da variância do estimador de total como casos especiais para distintos planos amostrais.
- ✓ Fórmulas de variância disponíveis para permitir avaliar qualidade do estimador de total sob distintas situações (população, variável e plano amostral).
- ✓ Um **total populacional** sempre pode ser estimado sem vício por uma soma amostral π-ponderada.