Teoria da Amostragem por Conglomerados e o Estimador de Horvitz-Thompson

Uma Abordagem Teórica para Planos Amostrais Complexos

Material de Aula para Pós-Graduação

2025-06-30

Sumário

1 1. Fundamentos da Amostragem por Conglomerados

1.1 1.1. Motivação: A Limitação do Cadastro Populacional

A teoria clássica de amostragem frequentemente assume a existência de um **cadastro** (ou *sampling frame*), uma lista completa de todas as N_0 unidades elementares da população U. Em muitas aplicações reais—sociais, econômicas ou de saúde pública—tal cadastro é inexistente ou sua construção é logisticamente inviável e financeiramente proibitiva.

A **amostragem por conglomerados** surge como uma solução elegante e prática para este problema. A ideia central é agrupar as unidades elementares em **conglomerados** (ou *clusters*), que servem como unidades primárias de amostragem.

Definição Formal: Uma população U é particionada em N subconjuntos disjuntos $U_1, U_2, \dots, U_N$ tais que $U = \bigcup_{i=1}^{N} U_i$. Cada U_i é um conglomerado. O plano amostral consiste em selecionar uma amostra de conglomerados, e não de unidades elementares.

Principais Vantagens:

1. **Redução de Custo:** Exige apenas o cadastro dos conglomerados (e.g., quarteirões de uma cidade, escolas de um estado), que é muito mais fácil de obter.
2. **Eficiência Logística:** Os custos de deslocamento e coleta são drasticamente reduzidos, pois as unidades observadas estão geograficamente concentradas.

Desvantagem Principal:

- Unidades dentro de um mesmo conglomerado tendem a ser mais homogêneas entre si do que com o resto da população (correlação intra-classe positiva). Isso geralmente leva a uma perda de precisão (aumento da variância) em comparação com uma Amostra Aleatória Simples (AAS) de mesmo número de unidades elementares.

1.2 1.2. Notação Formal

Para desenvolver a teoria, estabelecemos a seguinte notação:

- N: Número total de conglomerados (UPAs) na população.
- n: Número de conglomerados selecionados na amostra.
- M_i: Número de unidades secundárias (elementos) no conglomerado U_i.
- $M_0 = \sum_{i=1}^{N} M_i$: Número total de unidades elementares na população.
- y_{ij}: Valor da variável de interesse y para a unidade j no conglomerado i.
- $\tau_i = \sum_{j=1}^{M_i} y_{ij}$: Total populacional do conglomerado i.
- $\tau = \sum_{i=1}^{N} \tau_i$: Total populacional (o parâmetro que mais frequentemente queremos estimar).
- $\mu = \tau / M_0$: Média populacional por unidade elementar.

2 2. Amostragem em Um Estágio e a Necessidade de uma Teoria Geral

No desenho em **um estágio**, após selecionar uma amostra S de n conglomerados, **todas** as M_i unidades dentro de cada conglomerado $i \in S$ são investigadas.

2.1 2.1. O Cenário Idealizado: Conglomerados de Tamanhos Iguais

Vamos considerar um caso inicial onde $M_i = M$ para todo i. Se selecionarmos os conglomerados via AAS, a probabilidade de inclusão de qualquer conglomerado i é $\pi_i = n/N$.

Estimador para o Total (τ):

$$\hat{\tau}_{clu} = N \cdot \frac{1}{n} \sum_{i \in S} \tau_i$$

Este estimador é não-viesado para τ, e sua variância é:

$$\text{Var}(\hat{\tau}_{clu}) = N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma_b^2}{n}, \quad \text{onde } \sigma_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\tau_i - \bar{\tau})^2$$

A variância depende exclusivamente da variabilidade **entre** os totais dos conglomerados.

2.2 2.2. A Realidade: Tamanhos Desiguais e Probabilidades de Inclusão

Na prática, os conglomerados quase sempre possuem tamanhos M_i desiguais. Uma AAS de conglomerados neste cenário implica que a probabilidade de inclusão de uma unidade elementar j no conglomerado i, denotada por $\pi_{(ij)}$, é:

$$\pi_{(ij)} = P(\text{unidade } (ij) \text{ é incluída}) = P(\text{conglomerado } i \text{ é incluído}) = \frac{n}{N}$$

Embora simples, este esquema é sub-ótimo. A intuição sugere que conglomerados maiores, que contribuem mais para o total populacional τ, deveriam ter uma maior chance de serem selecionados. Este é o princípio da **amostragem com probabilidades desiguais**.

Para lidar com essa generalidade, precisamos de uma ferramenta teórica robusta: o estimador de Horvitz-Thompson.

3 3. O Estimador de Horvitz-Thompson (HT)

O estimador HT é uma estrutura geral para estimar totais populacionais sob qualquer plano amostral probabilístico. No nosso caso, as “unidades” da teoria HT são os conglomerados.

3.1 3.1. Probabilidades de Inclusão

A teoria requer o conhecimento das probabilidades de inclusão dos conglomerados:

- $\pi_i = P(i \in S)$: Probabilidade de o conglomerado i ser incluído na amostra.
- $\pi_{ij} = P(i \in S \text{ e } j \in S)$: Probabilidade conjunta de os conglomerados i e j serem incluídos.

Condição: O plano amostral deve garantir que $\pi_i > 0$ para todos os $i \in \{1, \dots, N\}$.

3.2 3.2. O Estimador e seu Não-Viesamento

O **estimador de Horvitz-Thompson** para o total populacional $\tau = \sum_{i=1}^{N} \tau_i$ é definido como:

$$\hat{\tau}_{HT} = \sum_{i \in S} \frac{\tau_i}{\pi_i}$$

A intuição é que cada conglomerado amostrado τ_i é ponderado pelo inverso de sua chance de ser selecionado.

Teorema: O estimador $\hat{\tau}_{HT}$ é não-viesado para o total populacional τ.

Prova: Seja I_i uma variável aleatória indicadora tal que $I_i = 1$ se o conglomerado i está em S, e $I_i = 0$ caso contrário. Por definição, $E[I_i] = P(I_i = 1) = \pi_i$. Podemos reescrever o estimador como uma soma sobre a população de conglomerados:

$$\hat{\tau}_{HT} = \sum_{i=1}^{N} I_i \frac{\tau_i}{\pi_i}$$

Usando a linearidade da esperança:

$$\begin{aligned} E[\hat{\tau}_{HT}] &= E\left[\sum_{i=1}^{N} I_i \frac{\tau_i}{\pi_i}\right] \\ &= \sum_{i=1}^{N} \frac{\tau_i}{\pi_i} E[I_i] \\ &= \sum_{i=1}^{N} \frac{\tau_i}{\pi_i} \pi_i = \sum_{i=1}^{N} \tau_i = \tau \quad \blacksquare \end{aligned}$$

3.3 3.3. A Variância do Estimador HT

A variância é a medida central da precisão de um estimador. Sua derivação é um resultado clássico e fundamental.

Teorema: A variância de $\hat{\tau}_{HT}$ é dada por:

$$\text{Var}(\hat{\tau}_{HT}) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i}\tau_i^2 + \sum_{i=1}^{N}\sum_{j\neq i} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\tau_i\tau_j$$

Prova: A variância de uma soma de variáveis aleatórias é $\text{Var}(\sum X_i) = \sum \text{Var}(X_i) + \sum_{i\neq j} \text{Cov}(X_i, X_j)$. Aplicando isso a $\hat{\tau}_{HT} = \sum_{i=1}^{N} I_i(\tau_i/\pi_i)$:

$$\begin{aligned}
\text{Var}(\hat{\tau}_{HT}) &= \sum_{i=1}^{N} \text{Var}\left(I_i \frac{\tau_i}{\pi_i}\right) + \sum_{i\neq j} \text{Cov}\left(I_i \frac{\tau_i}{\pi_i}, I_j \frac{\tau_j}{\pi_j}\right) \\
&= \sum_{i=1}^{N} \left(\frac{\tau_i}{\pi_i}\right)^2 \text{Var}(I_i) + \sum_{i\neq j} \left(\frac{\tau_i\tau_j}{\pi_i\pi_j}\right) \text{Cov}(I_i, I_j)
\end{aligned}$$

Os termos de variância e covariância das indicadoras são:

- $\text{Var}(I_i) = E[I_i^2] - (E[I_i])^2 = \pi_i - \pi_i^2 = \pi_i(1-\pi_i)$.
- $\text{Cov}(I_i, I_j) = E[I_iI_j] - E[I_i]E[I_j] = \pi_{ij} - \pi_i\pi_j$.

Substituindo de volta na equação:

$$\begin{aligned}
\text{Var}(\hat{\tau}_{HT}) &= \sum_{i=1}^{N} \left(\frac{\tau_i^2}{\pi_i^2}\right) \pi_i(1-\pi_i) + \sum_{i\neq j} \left(\frac{\tau_i\tau_j}{\pi_i\pi_j}\right)(\pi_{ij} - \pi_i\pi_j) \\
&= \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i}\tau_i^2 + \sum_{i\neq j} \frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\tau_i\tau_j \quad \blacksquare
\end{aligned}$$

3.3.1 Forma Alternativa da Variância (Sen-Yates-Grundy)

Para planos amostrais de tamanho **fixo** n, uma forma alternativa e muitas vezes preferível da variância é a de Sen-Yates-Grundy (SYG):

$$\text{Var}_{SYG}(\hat{\tau}_{HT}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}(\pi_i\pi_j - \pi_{ij})\left(\frac{\tau_i}{\pi_i} - \frac{\tau_j}{\pi_j}\right)^2$$

Implicação Estratégica: Esta fórmula revela a chave para um desenho amostral eficiente. A variância será minimizada se conseguirmos fazer com que a razão τ_i/π_i seja o mais constante possível para todos os conglomerados i.

4 3. Amostragem com Probabilidade Proporcional ao Tamanho (PPS)

A fórmula da variância de Sen-Yates-Grundy (SYG) nos fornece a principal intuição para a construção de planos amostrais eficientes:

$$\text{Var}_{SYG}(\hat{\tau}_{HT}) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j\neq i}(\pi_i\pi_j - \pi_{ij})\left(\frac{\tau_i}{\pi_i} - \frac{\tau_j}{\pi_j}\right)^2$$

Para minimizar esta variância, devemos escolher as probabilidades de inclusão π_i de modo a tornar a razão τ_i/π_i o mais constante possível. Embora τ_i seja desconhecido, frequentemente temos acesso a uma variável auxiliar x_i (uma "medida de tamanho"), como o tamanho do conglomerado M_i, que é altamente correlacionada com τ_i.

4.1 3.1. O Princípio PPS

Um plano de **amostragem com Probabilidade Proporcional ao Tamanho (PPS)** é aquele em que a probabilidade de selecionar uma unidade (conglomerado) é diretamente proporcional à sua medida de tamanho x_i.

Definição: Um plano é PPS se a probabilidade de seleção no primeiro sorteio é $p_i = x_i/X$, onde $X = \sum_{k=1}^{N} x_k$.

Quando a amostragem é realizada **com reposição**, a probabilidade de inclusão de 1ª ordem é simplesmente $\pi_i = n \cdot p_i$. Embora mais simples de analisar, a amostragem sem reposição é universalmente mais eficiente.

Para **amostragem sem reposição (PPS-SR)**, o cálculo das π_i é mais complexo, mas a intuição central permanece: π_i deve ser o mais próximo possível de $n \cdot p_i$. Vários algoritmos existem para implementar a amostragem PPS-SR (e.g., método de Brewer, amostragem sistemática PPS).

4.2 3.2. Por Que a Amostragem PPS Funciona?

A eficiência do PPS repousa na suposição de que $\tau_i \approx \beta x_i$. Se esta relação linear através da origem se mantém, então:

$$\frac{\tau_i}{\pi_i} \approx \frac{\beta x_i}{n(x_i/X)} = \frac{\beta X}{n} = \text{constante}$$

Quando esta condição é satisfeita, a variância do estimador HT se aproxima de zero, resultando em uma precisão notavelmente alta para um dado tamanho de amostra.

Cenários:

1. **Relação Perfeita ($\tau_i/x_i = \beta$):** A variância de SYG é zero. Obtemos uma estimativa perfeita.
2. **Relação Forte ($\tau_i/x_i \approx \beta$):** A variância será pequena. O plano PPS é muito eficiente.
3. **Relação Fraca ou Inexistente:** O plano PPS pode ser *menos* eficiente que uma AAS de conglomerados. A escolha da variável de tamanho x_i é, portanto, uma decisão crítica no planejamento amostral.

5 4. Eficiência do Desenho e o Efeito de Desenho (`deff`)

Como podemos quantificar a perda (ou ganho) de precisão de um plano amostral complexo, como o de conglomerados, em relação a uma Amostra Aleatória Simples (AAS)? A resposta é o **Efeito de Desenho (`deff`)**.

Definição: O efeito de desenho de um estimador $\hat{\theta}$ sob um plano amostral p é a razão entre a variância do estimador sob o plano p e a variância que seria obtida com um estimador análogo sob uma AAS de **mesmo número de unidades elementares**.

$$\text{deff}(\hat{\tau}_{clu}) = \frac{\text{Var}_p(\hat{\tau}_{clu})}{\text{Var}_{AAS}(\hat{\tau}_{AAS})}$$

onde $\text{Var}_{AAS}(\hat{\tau}_{AAS}) = M_0^2(1 - m/M_0)\frac{S_y^2}{m}$, com m sendo o número total de elementos na amostra de conglomerados.

5.1 4.1. `deff` na Amostragem por Conglomerados com Tamanhos Iguais

Para o caso simplificado de n conglomerados de tamanho M selecionados por AAS:

$$\text{deff} \approx 1 + (M-1)\rho_{int}$$

onde ρ_{int} é o **coeficiente de correlação intra-classe**, que mede a homogeneidade média das unidades dentro dos conglomerados.

- $\rho_{int} = \frac{1}{M-1} \frac{\sum_{i=1}^{N} \sum_{j \neq k} (y_{ij}-\mu)(y_{ik}-\mu)}{M_0 \sigma^2}$

Interpretação:

- **Se** $\rho_{int} > 0$ **(cenário comum):** Unidades dentro do mesmo cluster são parecidas. `deff` > 1, significando que a amostragem por conglomerados é **menos precisa** que uma AAS de mesmo tamanho. Cada unidade adicional dentro de um cluster fornece menos "informação nova".
- **Se** $\rho_{int} \approx 0$**:** Não há correlação. `deff` ≈ 1. A precisão é similar à de uma AAS.
- **Se** $\rho_{int} < 0$ **(raro):** Unidades dentro de um cluster são mais diferentes entre si do que a média. `deff` < 1. O desenho é **mais preciso** que uma AAS.

O `deff` é a "taxa de câmbio" da informação: se `deff`=2, precisamos de uma amostra de 2000 unidades em um desenho por conglomerados para obter a mesma precisão de uma AAS com 1000 unidades.

5.2 Exemplo Numérico (para ser resolvido "no quadro")

Considere uma população pequena com $N = 4$ conglomerados, da qual selecionaremos uma amostra de $n = 2$ conglomerados.

Conglomerado (i)	Tamanho (M_i)	Total (τ_i)
1	10	50
2	20	105
3	30	150
4	40	195
Total	**100**	**500**

O verdadeiro total populacional é $\tau = 500$. Vamos comparar dois planos amostrais.

5.2.1 Plano 1: Amostra Aleatória Simples de Conglomerados

- Neste caso, $\pi_i = n/N = 2/4 = 0.5$ para todos os i.
- **Amostra possível:** $S = \{1, 3\}$.
- **Estimativa HT:** $\hat{\tau}_{HT} = \frac{\tau_1}{\pi_1} + \frac{\tau_3}{\pi_3} = \frac{50}{0.5} + \frac{150}{0.5} = 100 + 300 = 400$.
- **Amostra possível:** $S = \{3, 4\}$.
- **Estimativa HT:** $\hat{\tau}_{HT} = \frac{\tau_3}{\pi_3} + \frac{\tau_4}{\pi_4} = \frac{150}{0.5} + \frac{195}{0.5} = 300 + 390 = 690$.
- **Observação:** As estimativas variam muito, pois não levamos em conta os tamanhos muito diferentes dos conglomerados.

5.2.2 Plano 2: Amostragem com Probabilidade Proporcional ao Tamanho (PPS)

Vamos definir a probabilidade de seleção $p_i = M_i/M_0$. * $p_1 = 0.1, p_2 = 0.2, p_3 = 0.3, p_4 = 0.4$. * Usaremos um método onde π_i é aproximadamente $n \cdot p_i$. Para simplificar, vamos assumir $\pi_i = np_i$: * $\pi_1 = 0.2, \pi_2 = 0.4, \pi_3 = 0.6, \pi_4 = 0.8$. (Nota: esta é uma aproximação; um método PPS exato teria valores ligeiramente diferentes, mas a lógica se mantém).

- **Amostra possível:** $S = \{1, 3\}$.

- **Estimativa HT:** $\hat{\tau}_{HT} = \frac{50}{0.2} + \frac{150}{0.6} = 250 + 250 = 500$.
- **Amostra possível:** $S = \{3, 4\}$.
- **Estimativa HT:** $\hat{\tau}_{HT} = \frac{150}{0.6} + \frac{195}{0.8} = 250 + 243.75 = 493.75$.
- **Observação:** As estimativas são muito mais estáveis e próximas do valor real (500). Isso ocorre porque os valores τ_i/π_i são quase constantes:
  - $\tau_1/\pi_1 = 250$, $\tau_2/\pi_2 = 262.5$, $\tau_3/\pi_3 = 250$, $\tau_4/\pi_4 = 243.75$.
  - A condição $\tau_i/M_i \approx$ constante é bem satisfeita nesta população, tornando o plano PPS muito eficiente.

6 5. Determinação do Tamanho da Amostra

A questão mais prática no planejamento amostral é: "Quantos conglomerados (n) devo selecionar?". A resposta depende da precisão desejada e do orçamento disponível.

A abordagem geral é:

1. Definir uma margem de erro aceitável para o estimador. Por exemplo, queremos que a variância do estimador do total seja no máximo um valor V^*: $\text{Var}(\hat{\tau}) \leq V^*$.
2. Isolar o tamanho da amostra n na fórmula da variância.

6.1 5.1. Tamanho da Amostra para Conglomerados de Tamanhos Iguais (via AAS)

Usando a fórmula da variância, temos:

$$V^* = N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma_b^2}{n}$$

Resolvendo para n:

$$V^* n = N^2 \sigma_b^2 - nN\sigma_b^2 \implies n(V^* + N\sigma_b^2) = N^2 \sigma_b^2$$

$$n = \frac{N^2 \sigma_b^2}{V^* + N\sigma_b^2}$$

Se a fração de amostragem for pequena ($n/N \approx 0$), a correção para população finita pode ser ignorada, e a fórmula se simplifica para:

$$n_0 = \frac{N^2 \sigma_b^2}{V^*}$$

E o tamanho final da amostra é ajustado por:

$$n = \frac{n_0}{1 + n_0/N}$$

Problema Prático: A fórmula requer o conhecimento de σ_b^2, a variância entre os totais dos conglomerados, que é um parâmetro populacional desconhecido. Soluções: * Usar dados de uma pesquisa piloto. * Usar dados de uma pesquisa anterior similar. * Fazer suposições embasadas sobre a estrutura da população.

6.2 5.2. Tamanho da Amostra Usando o Efeito de Desenho (`deff`)

Uma abordagem mais prática e geral envolve o `deff`. 1. Primeiro, calcula-se o tamanho de amostra necessário para uma **Amostra Aleatória Simples** (m_{aas}) para atingir a precisão desejada.

$$m_{aas} = \frac{M_0^2 S_y^2}{V^* + M_0 S_y^2}$$

onde S_y^2 é a variância populacional das unidades elementares.

2. Em seguida, inflaciona-se este número pelo `deff` estimado para obter o tamanho de amostra total necessário (m_{clu}) no desenho por conglomerados.

$$m_{clu} = m_{aas} \cdot \text{deff}$$

3. Finalmente, determina-se o número de conglomerados (n) a serem amostrados. Se os conglomerados têm tamanho médio $\bar{M}$:

$$n = \frac{m_{clu}}{\bar{M}} = \frac{m_{aas} \cdot \text{deff}}{\bar{M}}$$

Vantagem desta abordagem: Muitas vezes é mais fácil estimar ou encontrar na literatura valores plausíveis para a correlação intra-classe (ρ_{int}), e assim estimar o `deff`, do que estimar diretamente a variância entre os totais dos conglomerados (σ_b^2).

6.3 Exemplo Numérico de Planejamento (para ser resolvido "no quadro")

Cenário: Queremos estimar o número total de horas de estudo semanais dos alunos de uma universidade (τ). A universidade tem 10.000 alunos, distribuídos em aproximadamente $N = 200$ cursos (conglomerados) de tamanho médio $\bar{M} = 50$.

Objetivo: Obter uma estimativa com um erro padrão de no máximo 5.000 horas, o que implica uma variância máxima $V^* = (5.000)^2 = 25 \times 10^6$.

Informação prévia: 1. Uma pesquisa anterior em outra universidade sugeriu que o desvio padrão das horas de estudo por aluno é de cerca de $S_y = 10$ horas ($S_y^2 = 100$). 2. Para pesquisas em contextos educacionais, a correlação intra-classe para variáveis de comportamento (como horas de estudo) dentro de cursos é tipicamente moderada, em torno de $\rho_{int} = 0.1$.

Passo a Passo do Cálculo:

1. **Tamanho de amostra sob AAS (m_{aas}):**
   - Ignorando a correção para população finita para simplificar: $m_0 = \frac{M_0^2 S_y^2}{V^*} = \frac{(10.000)^2 \cdot 100}{25 \times 10^6} = \frac{10^8 \cdot 100}{25 \cdot 10^6} = \frac{10^{10}}{25 \cdot 10^6} = \frac{10^4}{25} = 400$ alunos.
   - Com correção: $m_{aas} = \frac{400}{1+400/10000} = \frac{400}{1.04} \approx 385$ alunos. Vamos usar $m_{aas} = 385$.
2. **Estimar o Efeito de Desenho (`deff`):**
   - $\text{deff} \approx 1 + (\bar{M} - 1)\rho_{int} = 1 + (50 - 1)(0.1) = 1 + 49 \cdot 0.1 = 1 + 4.9 = 5.9$.
   - *Interpretação:* Precisaremos de uma amostra quase 6 vezes maior no desenho por conglomerados para atingir a mesma precisão de uma AAS.
3. **Calcular o tamanho de amostra total para o desenho por conglomerados (m_{clu}):**
   - $m_{clu} = m_{aas} \cdot \text{deff} = 385 \cdot 5.9 \approx 2272$ alunos.
4. **Determinar o número de conglomerados a amostrar (n):**
   - $n = \frac{m_{clu}}{\bar{M}} = \frac{2272}{50} \approx 45.4$.
   - **Conclusão:** Devemos arredondar para cima. Seria necessário selecionar uma amostra de $n = 46$ **cursos** (conglomerados) e entrevistar todos os alunos dentro desses cursos para atingir a precisão desejada.

7 6. Conclusões Finais

A amostragem por conglomerados é indispensável na prática, mas sua implementação e análise exigem um ferramental teórico mais sofisticado. O estimador de Horvitz-Thompson oferece uma estrutura unificada e não-viesada, enquanto a amostragem PPS emerge como a estratégia chave para maximizar a eficiência. O conceito de **efeito de desenho** é a ponte crítica entre a teoria e o planejamento, permitindo quantificar a perda de precisão e, fundamentalmente, determinar um tamanho de amostra adequado para atingir os objetivos da pesquisa. "'