Capítulo 13 Amostragem de Conglomerados em Um ou Mais Estágios

13.1 Conceituação básica

Nos planos amostrais já apresentados (Amostragem Aleatória Simples - com ou sem reposição -, Amostragem Binomial ou de Bernoulli, Amostragem Sistemática, Amostragem com Probabilidades Desiguais - Proporcionais ao Tamanho: com reposição, Poisson, Poisson Sequencial, Pareto - e Amostragem Estratificada Simples) alguns aspectos são comuns: requerem cadastro de unidades individuais para permitir selecionar a amostra; a seleção da amostra é feita num único estágio ou etapa; as unidades de referência são iguais às unidades de amostragem; e o tamanho total da população é conhecido com base no cadastro.

Se tal cadastro não existir ou não puder ser construído ou o custo de atualizá-lo for muito elevado, a solução pode ser através do uso de amostragem de conglomerados, onde grupos de unidades são selecionadas com probabilidades conhecidas.

A *amostragem de conglomerados* consiste num esquema de amostragem em estágios, sendo que em cada estágio a unidade amostral, para a qual é atribuída a probabilidade de seleção, é grupada em um subconjunto (*conglomerado*) de unidades populacionais.

A formação de conglomerados pode ser:

- natural (exemplos: cacho de uvas, turma de alunos, edifício, quarteirão, município); ou
- artificial, construído pelo estatístico de acordo com o objetivo da pesquisa (exemplos: conglomerados de seis pessoas, de dez peças industriais do mesmo tipo, de cinco domicílios do mesmo edifício).

A unidade populacional depende da análise que está sendo feita e é determinada pelo propósito da pesquisa e não pelo plano amostral. Pode acontecer de mais de uma unidade populacional estar envolvida na pesquisa, quando por exemplo, características de domicílios e de pessoas são investigadas no mesmo levantamento.

Não há uma definição possível para os conglomerados. Por exemplo, a turma tanto pode ser uma unidade populacional (se estivermos interessados em investigar o número de alunos por turma), como pode ser um conglomerado de alunos (se estivermos interessados em investigar o aproveitamento dos alunos).

Para exemplificar, a Tabela 13.1 apresenta algumas ilustrações de possíveis conglomerados associados com a população, a variável de interesse e a unidade de referência para análise.

Tabela 13.1: Ilustrações de possíveis conglomerados

População	**Variáveis de interesse**	**Unidade de Referência**	**Conglomerados**
Turmas de alunos	Alunos por turma	Turma	Escolas
Estudantes de escolas de 2º grau	Aproveitamento dos estudantes	Estudante	Turmas
Visitantes de parques nacionais	Facilidades do parque	Visitante de parque nacional	Veículos que entram no parque
Passageiros de avião	Propósito da viagem	Passageiro de avião	Lotação de passageiros
Domicílios	Características de domicílios	Domicílio	Setores censitários
Moradores em favelas do Rio	Características de pessoas	Morador de favela do Rio	Domicílios em favelas do Rio

Conforme ilustração na Figura 13.1 temos regras de associação com a hieraquia com vários níveis (vários para um), considerando o cadastro de setores, os domicílios como nível 1 e a população de moradores.

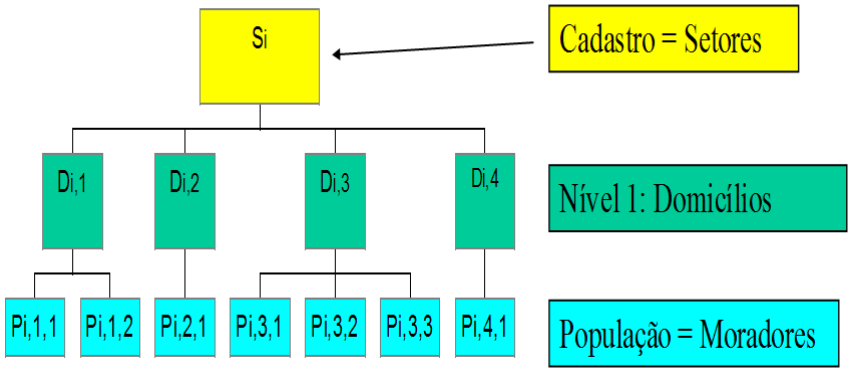

Figura 13.1: Ilustração da hierarquia com vários níveis (vários para um)

Cabe registrar que os vários planos amostrais discutidos anteriormente podem ser aplicados a amostragem de conglomerados, onde os conglomerados são as unidades amostrais.

13.2 Amostragem de áreas

O cadastro ou sistema de referência da pesquisa é a fonte que serve de guia e permite identificar a população a ser coberta para a seleção de amostras.

Os esquemas probabilísticos propostos para seleção de amostras pressupõem a existência de uma lista completa das unidades da população a ser pesquisada. Porém, uma lista pode não estar disponível, ou estar desatualizada, ou o custo de preparar uma lista atualizada pode ser proibitivo. Além disso, uma amostra selecionada de uma população dispersa geograficamente provavelmente será muito dispersa também.

Para reduzir custos é muito frequente o uso de amostragem de conglomerados definidos por áreas geográficas com limites naturais ou artificiais bem definidos. Neste caso a amostra resultante pode ser concentrada dentro de um número de áreas geográficas.

Portanto, a utilização de amostras de áreas se dá quando não existe um cadastro de boa qualidade disponível e/ou quando a população for muito dispersa e o fator custo de deslocamento for preponderante. Neste caso, a necessidade de uma lista atualizada das unidades para as quais se requer a informação é restrita às áreas que foram selecionadas para a amostra.

A grande vantagem da amostragem de conglomerados é a sua conveniência operacional vinculada a possíveis reduções de custo.

Num levantamento de população, por exemplo, é operacionalmente mais conveniente pesquisar todas as pessoas numa amostra de domicílios do que selecionar o mesmo número de pessoas espalhadas por toda a população ou mesmo pesquisar todos os domicílios de uma amostra de áreas (por exemplo, setores) do que selecionar uma amostra do mesmo número de domicílios selecionados aleatoriamente de uma lista de todos os domicílios. Tal lista nem sempre é disponível e o seu preparo torna a pesquisa bem mais cara.

Suponha-se que uma AAS de $n = 40$ deva ser selecionada de uma população de $N = 10.000$ domicílios de uma cidade. Como não dispomos de uma lista atualizada com todos os domicílios, optamos por uma amostra de domicílios localizados dentro de uma amostra de quarteirões. Isto pode ser feito dividindo a área toda da cidade em quarteirões e selecionando 1/25 quarteiros. A probabilidade de selecionar um domicílio na cidade é a probabilidade de selecionar um quarteirão, ou seja, $1/25 = 400/10.000$.

Portanto, as unidades amostrais são quarteirões selecionados de uma lista completa. A seleção da amostra de quarteirões determina a seleção dos domicílios que estão localizados nos quarteirões.

Mesmo se a lista de todos os domicílios fosse disponível, considerações na redução do custo pode ser observada na amostra de conglomerados. Pois a localização e identificação dos 400 domicílios espalhados aumentaria o custo com gastos de transporte, bem como um maior tempo para a coleta em comparação com a localização dos quarteirões e visita em todos os domicílios nestes quarteirões.

Mas para um dado tamanho de amostra, uma unidade menor em geral dá resultados mais precisos do que uma unidade maior.

Portanto, se compararmos uma amostra de conglomerados com uma amostra de unidades elementares compreendida do mesmo número de elementos, em geral na amostra de conglomerados tem-se:

- o custo por unidade elementar é mais baixo, devido ao mais baixo custo de listagem ou da localização, ou de ambos;
- a variância amostral é mais alta dependendo da homogeneidade dos elementos dos conglomerados.

Entretanto, levando em conta os aspectos operacionais e a redução de custos (devido ao possível ganho no tempo de coleta, identificação, contato, etc.) que a amostragem de conglomerados proporciona, em muitas situações práticas a perda de eficiência amostral é balaceada com essas vantagens.

Uma aplicação usual de amostragem de áreas se dá nas pesquisas domiciliares do IBGE, através da utilização da *Base Operacional Geográfica (BOG)* que tem como suas menores áreas os setores censitários, conforme descrito em 2.5.1.

13.3 Definições básicas e notação na amostragem conglomerada

A população de unidades U é particionada em M grupos *mutuamente exclusivos e exaustivos*, chamados *conglomerados*, denotados $C_1, \ldots, C_i, \ldots, C_M$, de modo que:

$$U = C_1 \cup C_2 \cup \cdots \cup C_M = \bigcup_{i=1}^{M} C_i \text{ e } C_i \cap C_k = \phi, \ i \neq k.$$

Então $C_i = \{\text{unidades pertencentes ao conglomerado } i\}$, para $i = 1, 2, \ldots, M$.

Seja N_i o tamanho de C_i. Então $N_1 + N_2 + \cdots + N_M = N$, o tamanho total da população.

Selecione uma amostra $a = \{i_1, \ldots, i_m\}$ de tamanho m $(m > 0)$, entre os rótulos de $C = \{1, \ldots, M\}$ para selecionar os conglomerados, segundo um plano amostral $p(a)$.

13.3.1 Amostragem Conglomerada em um estágio (AC1)

Para um plano *conglomerado em um estágio (AC1)*, inclua na amostra da pesquisa TODAS as unidades de U encontradas nos conglomerados selecionados em a. A amostra s de unidades da população U que serão pesquisadas é dada por:

$$s = C_{i_1} \cup C_{i_2} \cup \cdots \cup C_{i_m} = \cup_{k=1}^{m} C_{i_k}$$

Portanto, a amostragem conglomerada em um estágio ou AC1 é caracterizada pelos seguintes aspectos:

- As unidades populacionais são reunidas em *grupos* denominados *conglomerados*;
- Uma amostra de unidades é obtida selecionando uma *amostra de grupos* (conglomerados) e incluindo na amostra *todas as unidades* pertencentes aos grupos selecionados.

Segue ilustração utilizando um baralho, onde pode ser observado na Figura 13.2 que cada grupo de um mesmo número ou letra (Ás, 2, …, 7, J, D, K) representa um conglomerado com os quatro diferentes naipes (paus, ouros, copas e espadas).

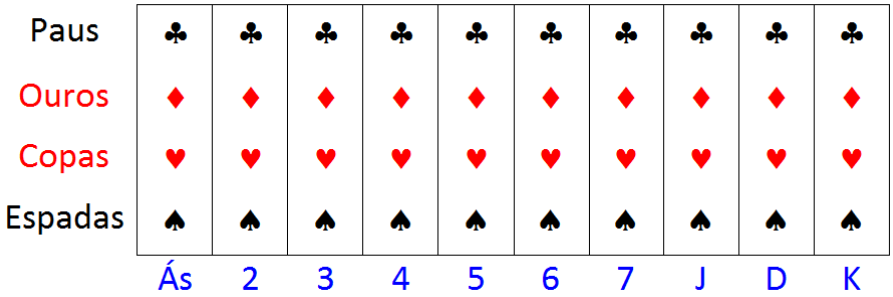

Figura 13.2: Ilustração de unidades elementares “naturalmente” agrupadas

Na ilustração da Figura 13.3 temos a seleção de amostragem conglomerada em um estágio (AC1).

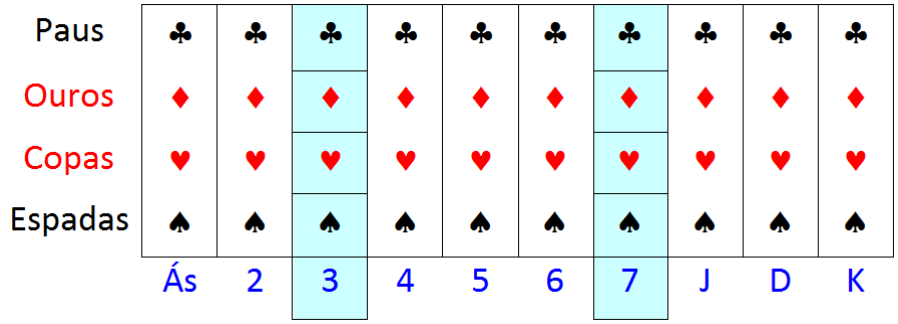

Figura 13.3: Ilustração da seleção de amostragem conglomerada em um estágio

Foram selecionados dois conglomerados (os de número 3 e 7), resultando numa amostra de 8 cartas do baralho (3 de paus, 3 de ouros, 3 de copas, 3 de espadas, 7 de paus, 7 de ouros, 7 de copas, 7 de espadas).

13.3.2 Amostragem conglomerada em vários estágios

A *amostragem conglomerada em vários estágios* é caracterizada por unidades populacionais arranjadas em grupos conforme uma hierarquia, com seleção de grupos nos vários níveis da hierarquia até chegar às unidades elementares (de referência) da pesquisa.

Na amostragem de conglomerados em três estágios adota-se a seguinte terminologia, onde em cada estágio da amostragem considera um tipo de unidade: unidades primárias de amostragem (UPAs); unidades secundárias de amostragem (USAs); e unidades elementares.

Tal estratégia consiste no arranjo da população em UPAs, as quais são divididas em USAs, que contém as unidades elementares, formando os estágios sucessivos.

Estágio 1: amostra de UPAs é selecionada.

Estágio 2: amostra de USAs é selecionada dentro de cada uma das UPAs selecionadas no primeiro estágio.

Estágio 3: amostra de unidades elementares é selecionada dentro de cada uma das USAs selecionadas.

Exemplo 13.1 Como exemplo, considere a amostragem em três estágios onde: as UPAs são os Municípios; as USAs são os setores censitários e as unidades elementares são os domicílios.

Notas

1. Sem USAs, o esquema acima se torna amostragem conglomerada em dois estágios, ou amostragem de conglomerados com subamostragem.

2. O processo acima pode ser estendido para 4 ou mais estágios.

3. Na prática é raro ver planos com mais de três ou quatro estágios de seleção.

Segue ilustração na Figura 13.4 de amostragem conglomerada em dois estágios (AC2), utilizando o baralho com os grupamentos definidos anteriormente.

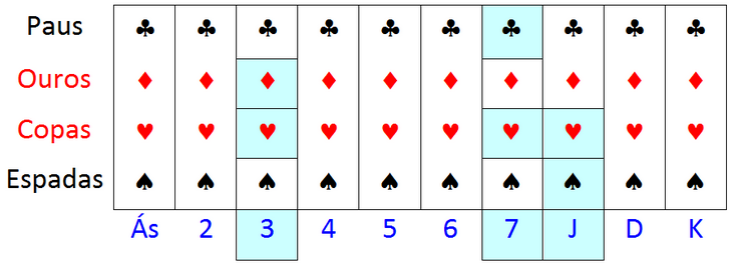

Figura 13.4: Ilustração da seleção de amostragem conglomerada em 2 estágios

Foram selecionados os conglomerados (UPAs) 3, 7 e J (valete) e selecionadas duas cartas por conglomerado selecionado, resultando numa amostra de 6 cartas do baralho (3 de ouros, 3 de copas, 7 de paus, 7 de copas, valete de copas e valete de espadas).

Populações humanas em geral são organizadas segundo uma hierarquia definida por regiões, estados, municípios, bairros, endereços, domicílios, famílias, pessoas.

Exemplo 13.2 Como outros exemplos de estruturas com hierarquia, considere:

a. Conglomerados = áreas num mapa; e unidades elementares = fazendas

b. UPAs = hospitais; USAs = enfermarias; e unidades elementares = pacientes

c. UPAs = escolas; USAs = turmas; e unidades elementares = crianças / estudantes

d. Conglomerados = carros cruzando pedágio; e unidades elementares = passageiros nos carros

e. UPAs = empresas industriais; e unidades elementares = unidades locais.

Exemplo 13.3 Plano amostral da Pesquisa Mensal de Emprego (PME) na Região Metropolitana de São Paulo realizada pelo IBGE[1]

Consiste em aplicar amostragem conglomerada em três estagios, onde os setores censitários da Região Matroplitana de São Paulo são as UPAs, os domícilios são as USAs e as unidades elementares são os moradores em domicílios particulares. O número médio de domicílios (USAs) por setor (UPA) era de 300 na zona urbana e 200 na zona rural. Todos os moradores dos domicílios selecionados eram pesquisados, sendo que somente os que tinham 10 anos ou mais de idade preenchiam a parte referente às características de ocupação e rendimento.

Eram 431 setores na amostra por mês para a Região Metropolitana de São Paulo, aproximadamente 18 domicílios selecionados em cada setor da amostra, resultando em aproximadamente 7.820 domicílios na amostra por mês.

A amostragem conglomerada em vários estágios deve ser adotada em situações em quê e por quê:

1. Não existe cadastro de unidades elementares, mas existe (ou se pode construir) um *cadastro de UPAs*.
2. *Concentrar a coleta* de dados em umas poucas localidades reduz o custo de deslocamentos entre unidades elementares.
3. O *acesso* às unidades elementares pode depender de “porteiros” nalgum nível da hierarquia.
4. A maior capacidade de supervisão do trabalho em grupos de unidades pode resultar em melhor qualidade dos dados.

13.4 Teoria básica de estimação para Amostragem Conglomerada em um estágio (AC1)

A Tabela 13.2 apresenta a notação de tamanhos da população e da amostra na AC1 para o conjunto das unidades em questão.

Tabela 13.2: Notação dos tamanhos da população e amostra na AC1 por conjunto de unidades

Conjunto	População	Amostra
Conglomerados	M	m
Unidades no conglomerado i	N_i	N_i
Todas as unidades	$N = \sum_{i \in C} N_i$	$n = \sum_{i \in a} N_i$

A Tabela 13.3 apresenta parâmetros populacionais do total e da média no conglomerado e em toda população.

Tabela 13.3: Descrição de parâmetros populacionais no conglomerado e em toda população

Descrição	Parâmetro
Valor da variável de pesquisa para unidade j do conglomerado i	y_{ij}
Total no conglomerado i	$Y_i = \sum_{j=1}^{N_i} y_{ij}$
Média no conglomerado i	$\bar{Y}_i = Y_i/N_i = \sum_{j=1}^{N_i} y_{ij}/N_i$
Total populacional	$Y = \sum_{i=1}^{M} Y_i = \sum_{i \in C} Y_i$
Média por conglomerado	$\bar{Y}_C = Y/M = \left(\sum_{i \in C} Y_i\right)/M$
Média por unidade	$\bar{Y} = Y/N = \left(\sum_{i \in C} Y_i\right)/N$

13.5 Amostragem conglomerada simples (ACS)

O esquema de seleção de uma amostra conglomerada simples consiste nos seguintes passos:

1. Selecione m conglomerados por AAS, dentre os M existentes.
2. Pesquise cada uma das unidades elementares presentes nos conglomerados selecionados.

A amostra desejada é formada por *todas as unidades elementares existentes nos m conglomerados selecionados*.

A Tabela 13.4 apresenta as informações amostrais nos conglomerados selecionados e em toda amostra.

Tabela 13.4: Descrição das informações amostrais na ACS

Descrição	Valores amostrais
Valor da variável de pesquisa para unidade j do conglomerado selecionado i	$y_{ij} \quad \forall j = 1, \ldots, N_i,$ $i \in a = \{i_1, \ldots, i_m\}$
Total no conglomerado i da amostra $i \in a = \{i_1, \ldots, i_m\}$	$Y_i = \sum_{j=1}^{N_i} y_{ij}$
Média no conglomerado i da amostra $i \in a = \{i_1, \ldots, i_m\}$	$\bar{Y}_i = Y_i / N_i$
Total amostral	$t = \sum_{i \in a} Y_i$
Média por conglomerado	$\bar{y}_C = t/m = \left(\sum_{i \in a} Y_i \right) \Big/ m$
Média por unidade	$\bar{y} = t/n = \left(\sum_{i \in a} Y_i \right) \Big/ \left(\sum_{i \in a} N_i \right)$
Probabilidade de inclusão dos conglomerados (qualquer um)	m/M
Probabilidade de inclusão das unidades elementares (todas)	m/M

13.5.1 Estimação sob plano amostral da ACS

O estimador natural, de Horvitz-Thompson, do total populacional sob plano amostral da ACS é dado por:

$$\hat{Y}_{ACS} = \frac{M}{m} \sum_{i \in a} Y_i = M \bar{y}_C = \sum_{i \in a} \sum_{j=1}^{N_i} w_{ij} \times y_{ij}$$

onde $w_{ij} = M/m$ são os pesos individuais.

A variância do estimador natural do total populacional é dada por:

$$V_{ACS}\left(\hat{Y}_{ACS}\right) = M^2 \frac{1-f}{m} S_e^2$$

onde $f = m/M$ e $S_e^2 = \frac{1}{M-1}\sum_{i \in C}\left(Y_i - \bar{Y}_C\right)^2$ é a *variância entre* os totais dos conglomerados.

O estimador da variância do estimador natural do total populacional é dado por:

$$\hat{V}_{ACS}\left(\hat{Y}_{ACS}\right) = M^2\frac{1-f}{m}s_e^2$$

onde S_e^2 é estimado por: $s_e^2 = \frac{1}{m-1}\sum_{i \in a}\left(Y_i - \bar{y}_C\right)^2$.

Há situações em que o controle da variação dos tamanhos dos conglomerados pode ser feito através da estimação por razão baseada no tamanho dos conglomerados.

O estimador de razão para o total populacional baseado no tamanho dos conglomerados é dado por:

$$\hat{Y}_R = \frac{N}{n}\sum_{i \in a} Y_i = N\bar{y} = \sum_{i \in a}\sum_{j=1}^{N_i} w_{ij}^R \times y_{ij}$$

onde $w_{ij}^R = N/n$ são pesos amostrais 'calibrados'.

Note que este estimador requer que o total de unidades elementares na população (N) seja conhecido. Portanto, em muitas situações este estimador não é viável.

A variância aproximada do estimador de razão do total pode ser obtida por:

$$V_{ACS}\left(\hat{Y}_R\right) \doteq M^2\frac{1-f}{m}\frac{1}{M-1}\sum_{i=1}^{M} N_i^2(\bar{Y}_i - \bar{Y})^2$$

Esta aproximação requer que o número de conglomerados na amostra (m) seja grande.

O estimador da variância do estimador de razão do total pode ser obtido por:

$$\hat{V}_{ACS}\left(\hat{Y}_R\right) = M^2\frac{1-f}{m}\frac{1}{m-1}\sum_{i \in a} N_i^2(\bar{Y}_i - \bar{y})^2$$

Na comparação de estimadores natural e de razão do total sob o plano amostral da ACS tem-se:

1. Se os conglomerados tiverem todos tamanhos iguais (isto é, se $N_i = \bar{N} = N/M\ \forall i = 1, \ldots, M$) então $\hat{Y}_R = \hat{Y}_{ACS}$.

2. Somente $\hat{Y}_{ACS}$ pode ser usado quando N for desconhecido.

3. $\hat{Y}_{ACS}$ é exatamente não viciado.

4. $\hat{Y}_R$ é apenas aproximadamente não viciado, para grandes amostras.

5. $\hat{Y}_R$ pode ser muito mais preciso que $\hat{Y}_{ACS}$ em certos casos.

6. Se $Y_i \doteq Y \ \forall i$ então $V_{ACS}\left(\hat{Y}_R\right) \doteq 0$ enquanto que

$$V_{ACS}\left(\hat{Y}_{ACS}\right) \propto \sum_{i=1}^{M}\left(Y_i - Y_C\right)^2 = \sum_{i=1}^{M}\left(N_i Y_i - NY\right)^2 \doteq Y^2 \sum_{i=1}^{M}\left(N_i - N\right)^2$$

Isto é, a variância do estimador natural incorpora parcela devida à variação dos tamanhos dos conglomerados. A ocorrência de variabilidade nos tamanhos dos conglomerados causa acentuada perda de precisão nos estimadores com amostragem de conglomerados em um estágio.

Na prática, as médias Y_i são menos variáveis entre conglomerados que os totais Y_i, e portanto: $V_{ACS}\left(\hat{Y}_R\right) < V_{ACS}\left(\hat{Y}_{ACS}\right)$.

Os ganhos de precisão podem ser grandes quando:

- for grande a variação dos tamanhos N_i;
- for pequena a variação entre as médias Y_i dos conglomerados.

Na prática, a formação de conglomerados com tamanhos iguais para controlar a variação de tamanho na variância do estimador, e também na variação do tamanho final da amostra nem sempre é possível, sendo que a ocorrência de conglomerados com tamahos iguais é pouco comum.

Assim, ao invés de tentar controlar artificialmente os tamanhos dos conglomerados, procura-se manter os conglomerados com os tamanhos desiguais e controlar a variação de tamanho dos conglomerados na expectativa de redução da variância e de menor perda de precisão com o uso da amostragem de conglomerados.

Os processos usuais de controle do tamanho dos conglomerados são:

a. selecionar os conglomerados com probabilidades proporcionais ao tamanho dos conglomerados;

b. estratificar os conglomerados, de modo que a variável de estratificação seja o tamanho;

c. usar um estimador de razão, com variável auxiliar definida pelo tamanho do conglomerado.

Recomendações

Em geral, prefira $\hat{Y}_R$ a menos que N seja desconhecido.

Se $\hat{Y}_{ACS}$ tiver que ser usado: *estratifique os conglomerados* por tamanho; ou *use amostragem de conglomerados com PPT*.

Em termos de eficiência não parece haver vantagem nítida de qualquer das duas alternativas, sendo bastante semelhante os resultados obtidos com ambas as técnicas em termos da precisão final das estimativas.

O estimador natural, de Horvitz-Thompson, da média por unidade $\bar{Y}$ sob plano amostral da ACS é dado por:

$$\bar{y}_N = \frac{\hat{Y}}{N} = \frac{M}{N}\frac{1}{m}\sum_{i \in a} Y_i = \bar{y}_C/\bar{N}$$

A variância do estimador natural da média pode ser obtida por:

$$V_{ACS}\left(\bar{y}_N\right) = \frac{M^2}{N^2}\frac{1-f}{m}S_e^2 = \frac{1}{N^{-2}}\frac{1-f}{m}S_e^2$$

O estimador da variância do estimador natural da média é dado por:

$$\hat{V}_{ACS}\left(\bar{y}_N\right) = \frac{M^2}{N^2}\frac{1-f}{m}s_e^2 = \frac{1}{N^{-2}}\frac{1-f}{m}s_e^2$$

Um estimador de razão da média por unidade $\bar{Y}$ sob plano amostral da ACS é dado por:

$$\bar{y}_R = \frac{\hat{Y}_R}{N} = \frac{1}{n}\sum_{i \in a} Y_i = \bar{y}_C/\bar{n} = \bar{y}$$

onde $\bar{n} = \frac{1}{m}\sum_{i \in a} N_i = \frac{n}{m}$.

A variância aproximada do estimador de razão da média é dado por:

$$V_{ACS}\left(\bar{y}_R\right) \doteq \frac{1}{n^{-2}} \frac{1-f}{m} \frac{1}{M-1} \sum_{i=1}^{M} N_i^2 \left(\bar{Y}_i - \bar{Y}\right)^2$$

Esta aproximação também é válida somente para amostras grandes.

O estimador da variância do estimador de razão da média é dado por:

$$\hat{V}_{ACS}\left(\bar{y}_R\right) = \frac{1}{n^{-2}} \frac{1-f}{m} \frac{1}{m-1} \sum_{i \in a} N_i^2 \left(\bar{Y}_i - \bar{y}\right)^2$$

com n em lugar de N quando este for desconhecido.

Notas

1. Se N (ou N) for desconhecido, só podemos usar $\bar{y}_R$.
2. As comparações de vício e variância feitas para o caso dos estimadores de total seguem válidas para a média.
3. Quase sempre é preferível usar $\bar{y}_R = \bar{y}$, a média simples por unidade elementar.

A Tabela 13.5 apresenta um resumo da estimação de parâmetros média e total da variável y sob ACS.

Tabela 13.5: Estimadores dos parâmetros média e total sob ACS

Parâmetro	Estimador ACS
$\bar{Y} = Y/N = \left(\sum_{i \in C} Y_i\right)/N$	$\bar{y}_N = \frac{\hat{Y}}{N} = \frac{M}{N}\frac{1}{m}\sum_{i \in a} Y_i = \bar{y}_C/N$ $\bar{y}_R = \frac{\hat{Y}_R}{N} = \frac{1}{n}\sum_{i \in a} Y_i = \bar{y}_C/n = \bar{y}$
$Y = \sum_{i=1}^{M} Y_i = \sum_{i \in C} Y_i$	$\hat{Y}_{ACS} = \frac{M}{m}\sum_{i \in a} Y_i = M\bar{y}_C = \sum_{i \in a}\sum_{j=1}^{N_i} w_{ij} \times y_{ij}$ $\hat{Y}_R = \frac{N}{n}\sum_{i \in a} Y_i = N\bar{y} = \sum_{i \in a}\sum_{j=1}^{N_i} w_{ij}^R \times y_{ij}$
$V_{ACS}\left(\bar{y}_N\right) = \frac{1}{N^{-2}}\frac{1-f}{m}S_e^2$	$\hat{V}_{ACS}\left(\bar{y}_N\right) = \frac{1}{N^{-2}}\frac{1-f}{m}s_e^2$
$V_{ACS}\left(\bar{y}_R\right) \doteq \frac{1}{N^{-2}}\frac{1-f}{m}\frac{1}{M-1}\sum_{i=1}^{M} N_i^2\left(\bar{Y}_i - \bar{Y}\right)^2$	$\hat{V}_{ACS}\left(\bar{y}_R\right) = \frac{1}{n^{-2}}\frac{1-f}{m}\frac{1}{m-1}\sum_{i \in a} N_i^2\left(\bar{Y}_i - \bar{y}\right)^2$
$V_{ACS}\left(\hat{Y}_{ACS}\right) = M^2\frac{1-f}{m}S_e^2$	$\hat{V}_{ACS}\left(\hat{Y}_{ACS}\right) = M^2\frac{1-f}{m}s_e^2$
$V_{ACS}\left(\hat{Y}_R\right) \doteq M^2\frac{1-f}{m}\frac{1}{M-1}\sum_{i=1}^{M} N_i^2(\bar{Y}_i - \bar{Y})^2$	$\hat{V}_{ACS}\left(\hat{Y}_R\right) = M^2\frac{1-f}{m}\frac{1}{m-1}\sum_{i \in a} N_i^2(\bar{Y}_i - \bar{y})^2$

13.6 Efeito de plano amostral (EPA)

O *efeito de plano amostral* é uma medida para comparar a *eficiência* de duas estratégias (E_1 e E_2, digamos), formadas por combinações de *plano amostral e estimador*, para um tamanho de amostra comum.

$$EPA(E_1;\, E_2) = V_{E_1}(\hat{\theta}_1)/V_{E_2}(\hat{\theta}_2)$$

O termo original em inglês é *design effect (DEFF)*, e foi sugerido por Kish (1965).

Outra medida que dá uma indicação semelhante ao EPA é o "*fator de plano amostral*" (FPA), definido como: $FPA = \sqrt{EPA}$, nome que vem do inglês *design factor*.

O FPA compara diretamente o *desvio padrão* de um estimador sob duas estratégias diferentes de amostragem.

É mais comum o uso do EPA que do FPA, mas são medidas equivalentes, embora expressas em unidades distintas.

Exemplo 13.4 Efeito do plano amostral ao estimar a média populacional por unidade elementar, através do estimador natural com amostragem conglomerada simples em relação ao uso da AAS, ou seja:

Estratégia 1: Amostragem conglomerada simples (ACS), com o estimador natural $\bar{y}_N$.

Estratégia 2: Amostragem aleatória simples (AAS) de mesmo tamanho total (n), com o estimador usual de média $\bar{y} = \sum_{k=1}^{n} y_k / n$.

O efeito do plano amostral (neste caso, conglomeração) ao estimar a média populacional por unidade elementar é:

$$EPA(ACS;\ \bar{y}_N) = V_{ACS}(\bar{y}_N) / V_{AAS}(\bar{y})$$

EPA mede o quanto a variância do estimador é maior (ou menor) por usar ACS em lugar de AAS.

$EPA > 1 \Rightarrow$ *perda de precisão* devida ao uso de amostragem conglomerada; e

$EPA < 1 \Rightarrow$ *ganho de precisão* devido ao uso de amostragem conglomerada.

Exemplo 13.5 Efeito do plano amostral ao estimar a média populacional por unidade elementar, através do estimador de razão com amostragem conglomerada simples em relação ao uso da AAS, ou seja:

Estratégia 1: Amostragem conglomerada simples (ACS), com estimador de razão $\bar{y}_R = \frac{1}{n}\sum_{i \in a} Y_i$ para a média.

Estratégia 2: Amostragem aleatória simples (AAS) de mesmo tamanho total (n), com o estimador usual de média $\bar{y} = \sum_{k=1}^{n} y_k / n$.

O efeito do plano amostral (neste caso, conglomeração) ao estimar a média populacional por unidade elementar é:

$$EPA(ACS;\ \bar{y}_R) = V_{ACS}(\bar{y}_R) / V_{AAS}(\bar{y})$$

Nota: Os estimadores pontuais são idênticos; somente os planos amostrais (e as variâncias) são diferentes.

Se os tamanhos dos conglomerados forem *iguais*, isto é, se $N_i = N\ \forall i = 1, \ldots, M$, então, de acordo com Cochran (1977), p. 252, tem-se:

$$EPA(ACS;\ \bar{y}_R) \doteq 1 + (\bar{N} - 1) \times \rho$$

onde:

$\rho = \dfrac{\sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{k \neq j=1}^{N} \left(y_{ij} - \bar{Y}\right)\left(y_{ik} - \bar{Y}\right)}{\left(\bar{N} - 1\right)\left(M\bar{N} - 1\right) S_y^2} \doteq 1 - \dfrac{S_D^2}{S_y^2}$ é o coeficiente de correlação intraconglomerado (ou intraclasse).

S_D^2 é a medida da variância *dentro* dos conglomerados, dada por:

$$S_D^2 = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{\bar{N} - 1} \sum_{j=1}^{N} \left(y_{ij} - \bar{Y}_i\right)^2 = \frac{1}{M} \sum_{i=1}^{M} S_i^2 \text{ com } S_i^2 = \frac{1}{\bar{N} - 1} \sum_{j=1}^{N} \left(y_{ij} - \bar{Y}_i\right)^2$$

S_y^2 é a medida da variância *total*, dada por:

$$S_y^2 = \frac{1}{M\bar{N} - 1} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(y_{ij} - \bar{Y}\right)^2$$

A expressão para o $EPA(ACS;\ \bar{y}_R)$ resulta do uso das expressões de acordo com Cochran (1977), p. 241:

$$V_{ACS}(\bar{y}_R) \doteq \frac{1 - f}{m\bar{N}} S_y^2 [1 + (\bar{N} - 1)\rho]$$

$$V_{AAS}(\bar{y}) = \frac{1 - f}{m\bar{N}} S_y^2$$

Cabem algumas considerações relacionadas com a variação do EPA para ACS:

1. Se os conglomerados tiverem variância dentro grande, isto é, se $S_D^2 \doteq S_y^2$, então $\rho \doteq 0$ e portanto, $EPA(ACS; \bar{y}_R) \doteq 1 + (N-1) \times 0 = 1$.

Nesse caso, não ocorreria perda de precisão devido ao uso de amostragem conglomerada.

2. Pode ser demonstrado que $-\left(\frac{M-1}{N-M}\right) \leq \rho \leq 1$, ou ainda, de forma aproximada, que $\frac{-1}{N-1} \leq \rho \leq 1$.

Usualmente $\rho > 0$, porque os conglomerados tendem a ser mais homogêneos que a população em geral.

“Birds of a feather flock together!”

Consequência: $EPA(ACS; \bar{y}_R) > 1$ na maioria das vezes.

3. Raramente $\rho < 0$, caso em que amostragem conglomerada simples seria mais eficiente que AAS.

4. Num caso extremo, $\rho = 1$ e portanto $EPA = N$ e $V_{ACS}(\bar{y}_R) = EPA \times V_{AAS}(\bar{y}) \doteq N \times \frac{S_y^2}{mN} = \frac{S_y^2}{m}$

Nesse caso a precisão da amostra conglomerada de tamanho total igual a mN é equivalente apenas àquela obtida com uma amostra aleatória simples de tamanho n!!!

A Tabela 13.6 apresenta efeitos de plano amostral sob ACS de acordo com os tamanhos dos conglomerados e do coeficiente de correlação intraclasse.

Tabela 13.6: Efeitos de plano amostral sob ACS por tamanho do conglomerado e do coeficiente de correlação intraclasse

Tamanho do conglomerado	$\rho = 0,01$	$\rho = 0,05$	$\rho = 0,1$	$\rho = 0,2$	$\rho = 0,3$	$\rho = 0,5$
2	1	1	1	1	1	2
5	1	1	1	2	2	3
11	1	2	2	3	4	6
21	1	2	3	5	7	11
31	1	3	4	7	10	16
51	2	4	6	11	16	26
101	2	6	11	21	31	51
201	3	11	21	41	61	101
301	4	16	31	61	91	151
501	6	26	51	101	151	251

Observe que com o aumento do tamanho do conglomerado, como aumenta a perda de precisão da ACS em relação à AAS e para conglomerados maiores a medida que aumenta o coeficiente de correlação intraclasse diminui a eficiência da ACS.

13.7 Amostragem conglomerada com probabilidade proporcional ao tamanho (PPT) em um estágio

A ocorrência de variabilidade nos tamanhos dos conglomerados causa acentuada perda de precisão nos estimadores naturais com amostragem de conglomerados em um estágio.

A amostragem conglomerada com probabilidade proporcional ao tamanho (PPT) em um estágio (AC1PPT) se caracteriza por:

1. Ser útil para controlar os efeitos da variação nos tamanhos dos conglomerados.
2. Ser adotada na etapa de seleção da amostra, enquanto que estimadores tipo razão podem ser considerados na etapa de estimação.

3. Selecionar conglomerados com probabilidades proporcionais ao seu tamanho (número de unidades subordinadas ou outra medida de tamanho).

A seguir, o estimador não viciado do total Y com AC1PPT, no caso de PPT com reposição.

$$\hat{Y}_{PPTC} = \frac{1}{m} \sum_{k=1}^{m} \frac{Y_{i_k}}{p_{i_k}}$$

onde:

p_{i_k} é probabilidade de seleção associada ao conglomerado i selecionado no sorteio k;

p_{i_k} é igual a algum dos $p_i = N_i/N$ $(i = 1, 2, \ldots, N)$.

O estimador não viciado do total Y com AC1PPT, no caso de PPT sem reposição é dado por:

$$\hat{Y}_{PPT} = \sum_{i \in a} \frac{Y_i}{\pi_i}$$

sendo:

π_i a probabilidade de inclusão do conglomerado i na amostra.

As respectivas variâncias de $\hat{Y}_{PPTC}$ e $\hat{Y}_{PPT}$ são dadas por:

$$V_{PPTC}(\hat{Y}_{PPTC}) = \frac{1}{M} \left(\sum_{k=1}^{M} \frac{Y_{i_k}^2}{p_{i_k}} - Y^2 \right)$$

$$V_{PPT}(\hat{Y}_{PPT}) = \sum_{i=1}^{M} \frac{(1-\pi_i)}{\pi_i} Y_i^2 + \sum_{i=1}^{M} \sum_{j \neq i}^{M} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} Y_i Y_j$$

Os respectivos estimadores das variâncias de $\hat{Y}_{PPTC}$ e $\hat{Y}_{PPT}$ são dados por:

$$\hat{V}_{PPTC}(\hat{Y}_{PPTC}) = \frac{1}{m(m-1)} \sum_{k=1}^{m} \left(\frac{Y_{i_k}}{p_{i_k}} - \hat{Y}_{PPTC} \right)^2$$

$$\hat{V}_{PPT}(\hat{Y}_{PPT}) = \sum_{i=1}^{m} \frac{(1-\pi_i)}{\pi_i} Y_i^2 + \sum_{i=1}^{m} \sum_{j \neq i}^{m} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} Y_i Y_j$$

Notas:

a. $\hat{Y}_{PPTC}$ é *mais preciso* que o estimador não viciado do total $\hat{Y}$ com ACS, quando as médias dos conglomerados não são relacionadas com os tamanhos dos conglomerados.

b. $\hat{Y}_{PPTC}$ *não se beneficia* do fator de correção de população finita.

c. Métodos para amostragem de conglomerados com *PPT sem reposição* estão disponíveis e podem ser usados em lugar de PPTC.

d. Para planos AC1PPT, os tamanhos precisam ser conhecidos exatamente para cada conglomerado da população, antes de selecionar a amostra. Caso contrário, podem ser usadas medidas aproximadas de tamanho.

e. $\hat{Y}_{PPTC}$ tem desempenho similar ao estimador de razão do total, $\hat{Y}_R$, sob ACS. Quando não for feita amostragem PPT de conglomerados, usar sempre que possível.

f. $\hat{Y}_{PPTC}$ possui propriedades exatas (não viciado, variância, estimador não viciado de variância) que valem também para amostras pequenas.

13.8 Amostragem conglomerada em dois estágios (AC2)

O plano amostral de conglomerados em dois estágios (AC2) se caracteriza pela seleção de uma amostra de conglomerados com subamostragem, assim definida:

Estágio 1: Selecione amostra a de m UPAs (conglomerados).

Estágio 2: Para cada UPA i da amostra de primeiro estágio, selecione amostra s_i de n_i unidades secundárias das N_i unidades existentes nessa UPA.

A amostra completa de unidades selecionadas é dada por: $s = s_{i_1} \cup s_{i_2} \cup \cdots \cup s_{i_m} = \cup_{k=1}^{m} s_{i_k}$

O tamanho total da amostra é $n = \sum_{i \in a} n_i$.

As principais razões para adotar Amostragem Conglomerada em dois estágios são as seguintes:

1. Geralmente não é prático pesquisar todas as unidades nos conglomerados selecionados: *conglomerados* muito *grandes*, *carga de trabalho variável* por entrevistador, etc.

2. Constatou-se que a perda de precisão da AC1 em relação à AAS para amostras de mesmo tamanho é tanto maior quanto maior for o tamanho do conglomerado. A adoção de AC2 vem reduzir a influência do tamanho dos conglomerados na eficiência da AC1.

3. Se a variância dentro dos conglomerados for pequena, as médias por conglomerados $\bar{Y}_i$ podem ser bem estimadas por amostragem.

4. *Amostragem em dois estágios* é mais complexa, porém *mais flexível*.

Na estimação sob AC2, cujo plano amostral compreende *dois estágios de seleção*, para encontrar médias e variâncias de estimadores, médias sobre todas as possíveis amostras sob o plano amostral devem ser calculadas.

Isto requer considerar todas as possíveis amostras no primeiro estágio, e todas as possíveis amostras no segundo estágio, dentro de cada amostra de UPAs do primeiro estágio.

Tudo fica mais fácil se usarmos os resultados de Cochran (1977), equação 10.1 da p.275 e equação 10.2 da p.276, respectivamente:

$$E[\hat{\theta}] = E_1[E_2(\hat{\theta})]$$

$$V[\hat{\theta}] = V_1[E_2(\hat{\theta})] + E_1[V_2(\hat{\theta})]$$

E_2, V_2 denotam, respectivamente, média e variância sobre todas as possíveis amostras de unidades dentro de um conjunto fixado de UPAs (estágio 2);

E_1, V_1 denotam, respectivamente, média e variância sobre todas as possíveis amostras de UPAs (estágio 1).

$\hat{\theta}$ é algum estimador para um parâmetro θ.

Resultados similares podem ser estendidos para planos em três ou mais estágios (veja Cochran (1977), seção 10.8).

O estimador não viciado de Horvitz-Thompson do total Y sob AC2 é dado por:

$$\hat{Y}_{HT} = \sum_{i \in a} \frac{\hat{Y}_i}{\pi_i} = \sum_{i \in a} \frac{1}{\pi_i} \sum_{j \in s_i} \frac{y_{ij}}{\pi_{j|i}} = \sum_{i \in a} \sum_{j \in s_i} w_{ij}^{HT} y_{ij}$$

π_i é a probabilidade de inclusão da UPA i;

s_i é a amostra de unidades selecionadas dentro da UPA i;

$\hat{Y}_i = \sum_{j \in s_i} \frac{y_{ij}}{\pi_{j|i}}$ é um estimador HT do total Y_i da UPA i;

$\pi_{j|i} = P(j \in s_i | i \in a)$ é a probabilidade de inclusão da unidade j dado que a UPA i está na amostra a;

$w_{ij}^{HT} = \pi_{ij}^{-1} = \pi_i^{-1}\pi_{j|i}^{-1}$ é o peso associado à unidade j da UPA i.

A variância de $\hat{Y}_{HT}$ com AC2 é dada por:

$$
\begin{aligned}
V_{AC2}\left(\hat{Y}_{HT}\right) &= V_1\left[E_2\left(\sum_{i \in a} \frac{\hat{Y}_i}{\pi_i}\right)\right] + E_1\left[V_2\left(\sum_{i \in a} \frac{\hat{Y}_i}{\pi_i}\right)\right] \\
&= V_1\left[\sum_{i \in U} \delta_i E_2\left(\hat{Y}_i\right)/\pi_i\right] + E_1\left[\sum_{i \in U} \delta_i V_2\left(\hat{Y}_i\right)/{\pi_i}^2\right] \\
&= V_1\left(\sum_{i \in a} Y_i/\pi_i\right) + \sum_{i \in U} V_2\left(\hat{Y}_i\right)/\pi_i \\
&= V_{UPA} + V_{USA}
\end{aligned}
$$

V_{UPA} é a componente de variância de $\hat{Y}_{HT}$ proveniente da amostragem de UPAs (estágio 1), isto é, variância caso amostragem de conglomerados em um estágio fosse usada (sem fazer subamostragem no segundo estágio);

V_{USA} é a componente de variância proveniente da amostragem de USAs (amostragem no estágio 2).

Um estimador não viciado da média por unidade $\bar{Y}$ (estimador de Horvitz-Thompson) é dado por:

$$
\bar{y}_N = \hat{Y}_{HT}/N = \left(\sum_{i \in a} \frac{\hat{Y}_i}{\pi_i}\right)/N
$$

Se N for conhecido, um estimador de razão para estimar o total Y é dado por:

$$
\hat{Y}_R = N \times \left(\sum_{i \in a} \frac{\hat{Y}_i}{\pi_i}\right) / \left(\sum_{i \in a} \frac{N_i}{\pi_i}\right)
$$

Um estimador de razão da média por unidade é dado por:

$$\bar{y}_R = \left(\sum_{i \in a} \frac{\hat{Y}_i}{\pi_i} \right) / \left(\sum_{i \in a} \frac{N_i}{\pi_i} \right)$$

Este estimador de razão da média pode ser calculado mesmo quando N for desconhecido.

13.8.1 Plano amostral 1 – AC2 com AAS nos 2 estágios

Trataremos agora do plano amostral AC2 com AAS nos 2 estágios de seleção, ou seja:

Estágio 1: Selecione amostra de m UPAs usando AAS.

Estágio 2: Para cada UPA i da amostra de primeiro estágio, selecione n_i unidades secundárias das N_i unidades existentes usando AAS.

Para esse plano, a probabilidade de inclusão da unidade j da UPA i é dada por:

$$\pi_{ij} = P(i \in a, j \in s) = P(i \in a) \times P(j \in s \mid i \in a) = \frac{m}{M} \times \frac{n_i}{N_i}$$

Planos amostrais são mais simples quando as probabilidades de inclusão são constantes, isto é, para quaisquer $\pi_{ij} = f = n/N \; \forall i$ e j. Nestas condições o plano amostral é dito *equiponderado* ou *autoponderado*.

Com o Plano amostral 1, isto pode ser conseguido tomando $n_i \propto N_i$.

Uma desvantagem importante desse tipo de plano seria a geração de *cargas de trabalho desiguais* por UPA ou por entrevistador, caso cada UPA seja alocada a um só entrevistador.

De acordo com Cochran (1977), equação 11.21, o estimador não viciado do total sob o Plano 1 é dado por:

$$\hat{Y}_{Plano1} = \frac{M}{m} \sum_{i \in a} \hat{Y}_i$$

com $\hat{Y}_i = \frac{N_i}{n_i} \sum_{j \in s_i} y_{ij}$ para toda UPA i.

De acordo com Cochran (1977), equação 11.22, a variância do estimador não viciado do total sob o Plano 1 é dada por:

$$V_{Plano1}\left(\hat{Y}_{Plano1}\right) = M^2\left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{M-1}\sum_{i=1}^{M}\left(Y_i-\bar{Y}_C\right)^2 + \frac{M}{m}\sum_{i=1}^{M}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2$$

onde as parcelas do 2º membro representam as “componentes” da variância devidas ao 1º e ao 2º estágios de seleção, respectivamente.

sendo: $S_i^2 = \frac{1}{N_i-1}\sum_{j=1}^{N_i}\left(y_{ij}-\bar{Y}_i\right)^2$ a variância dentro da UPA i.

Note-se que:

i. Se $m = M$ então, a 1ª componente da variância é nula, ou seja:

$$V_{Plano1}\left(\hat{Y}_{Plano1}\right) = \frac{M}{m}\sum_{i=1}^{M}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2 = V_{AES}\left(\hat{Y}_{AES}\right)$$

e este plano equivale ao de uma amostra estratificada.

ii. Se $n_i = N_i\ \forall i = 1, 2, \ldots, n)$ então, a 2ª componente da variância é nula, ou seja:

$$V_{Plano1}\left(\hat{Y}_{Plano1}\right) = M^2\left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{M-1}\sum_{i=1}^{M}\left(Y_i-\bar{Y}_C\right)^2 = V_{ACS}\left(\hat{Y}\right)$$

e este plano amostral equivale ao de uma amostra de conglomerados em um estágio.

De acordo com Cochran (1977), equação 11.24, o estimador da variância do estimador não viciado do total sob o Plano 1 é dado por:

$$\hat{V}_{Plano1}\left(\hat{Y}_{Plano1}\right) = M^2\left(\frac{1}{m}-\frac{1}{M}\right)\frac{1}{m-1}\sum_{i\in a}\left(\hat{Y}_i-\bar{y}_C\right)^2 + \frac{M}{m}\sum_{i\in a}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)\hat{S}_i^2$$

onde: $\bar{y}_C = \frac{1}{m}\sum_{i\in a}\hat{Y}_i$ estima a média por conglomerado $\bar{Y}_C$; e

$\hat{S}_i^2 = \frac{1}{n_i - 1} \sum_{j \in s_i} \left(y_{ij} - \bar{y}_i \right)^2$ estima a variância dentro da UPA i.

Um estimador não viciado da média por unidade sob o Plano 1 é dado por:

$$\bar{y}_{Plano1} = \frac{\hat{Y}_{HT}}{N} = \frac{M}{mN} \sum_{i \in a} \hat{Y}_i = \frac{M}{mN} \sum_{i \in a} \frac{N_i}{n_i} \sum_{j \in s_i} y_{ij}$$

A variância do estimador não viciado da média por unidade sob o Plano 1 é dada por:

$$V_{Plano1}\left(\bar{y}_{Plano1} \right) = \frac{1}{N^2} V_{Plano1}\left(\hat{Y}_{Plano1} \right)$$

O estimador da variância do estimador não viciado da média por unidade sob o Plano 1 é dada por:

$$\hat{V}_{Plano1}\left(\bar{y}_{Plano1} \right) = \frac{1}{N^2} \hat{V}_{Plano1}\left(\hat{Y}_{Plano1} \right)$$

13.8.2 Plano amostral 2 – AC2 com PPT com reposição no primeiro estágio e AAS no segundo estágio (AC2PPTC)

Trataremos agora do plano amostral AC2 com PPT com reposição no primeiro estágio de seleção e AAS no segundo estágio, ou seja:

Estágio 1: Selecione amostra de m UPAs usando probabilidades proporcionais a uma medida de tamanho x_i;

Estágio 2: Para cada UPA i da amostra de primeiro estágio, selecione n_i unidades secundárias das N_i unidades existentes usando AAS.

Para o Plano amostral 2, agora denominado Plano 2, a probabilidade de inclusão da unidade j da UPA i é dada por:

$$\pi_{ij} = P(i \in a, j \in s) = P(i \in a) \times P(j \in s \,|\, i \in a) = m \frac{x_i}{X} \times \frac{n_i}{N_i}$$

Este plano será *equiponderado* ou *autoponderado* quando:

a. $n_i \propto \frac{N_i}{x_i}$; ou

b. $x_i = N_i$ e $n_i = n$. Esta é a opção mais usada na prática.

A seguir, a estimação no Plano 2, considerando que no estágio 1 a seleção da amostra de n UPAs foi feita usando probabilidades proporcionais ao tamanho da UPA N_i.

De acordo com Cochran (1977), equação 11.31, um estimador não viciado do total sob o Plano 2 é dado por:

$$\hat{Y}_{Plano2} = \frac{1}{m}\sum_{k=1}^{m}\frac{\hat{Y}_{i_k}}{p_{i_k}} = \frac{1}{m}\sum_{i\in a}\frac{\hat{Y}_i}{p_i}$$

onde:

p_{i_k} é probabilidade de seleção associada à UPA i selecionada no sorteio k;

$p_i = N_i/N \; \forall i = 1, 2, \ldots, M$; p_{i_k} é igual a algum dos p_i $(i = 1, 2, \ldots, M)$; e

$\hat{Y}_{i_k} = \frac{N_i}{n_i}\sum_{j\in s_i} y_{ij} = \hat{Y}_i$ é o total estimado para a UPA i selecionada no sorteio k.

A variância de $\hat{Y}_{Plano2}$, conforme Cochran (1977), equação 11.32, é dada por:

$$\begin{aligned}
V_{Plano2}\left(\hat{Y}_{Plano2}\right) &= V_1\left[E_2\left(\hat{Y}_{Plano2}\right)\right] + E_1\left[V_2\left(\hat{Y}_{Plano2}\right)\right] \\
&= V_1\left[E_2\left(\frac{1}{m}\sum_{i\in a}\frac{\hat{Y}_i}{p_i}\right)\right] + E_1\left[V_2\left(\frac{1}{m}\sum_{i\in a}\frac{\hat{Y}_i}{p_i}\right)\right] \\
&= V_1\left[\frac{1}{m}\sum_{i\in a}\frac{E_2\left(\hat{Y}_i\right)}{p_i}\right] + E_1\left[\frac{1}{m^2}\sum_{i\in a}\frac{V_2\left(\hat{Y}_i\right)}{p_i^2}\right] \\
&= V_1\left[\frac{1}{m}\sum_{i\in a}\frac{Y_i}{p_i}\right] + E_1\left[\frac{1}{m^2}\sum_{i\in a}\frac{1}{p_i^2}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2\right] \\
&= \frac{1}{m}\sum_{i\in U}\left(\frac{Y_i}{p_i}-Y\right)^2 p_i + \frac{1}{m}\sum_{i\in U}\frac{1}{p_i}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2
\end{aligned}$$

Um estimador da variância de $\hat{Y}_{Plano2}$, de acordo com Cochran (1977), equação 11.35, é dado por:

$$\hat{V}_{Plano2}\left(\hat{Y}_{Plano2}\right) = \frac{1}{m(m-1)} \sum_{i \in a} \left(\frac{N_i \bar{y}_i}{p_i} - \hat{Y}_{Plano2} \right)^2$$

13.8.3 Plano amostral 3 – AC2 com PPT Poisson Sequencial no estágio 1 e AAS no estágio 2 (AC2PPPS)

Amostragem Conglomerada PPT Poisson Sequencial de UPAs, mais AAS de n_i USAs nas UPAs selecionadas.

Um estimador não viciado do total Y sob o Plano 3 é dado por:

$$\hat{Y}_{Plano3} = \sum_{i \in a} \frac{\hat{Y}_i}{m p_i}$$

onde:

$p_i = N_i / N \; \forall i = 1, 2, \ldots, M$; e

$\hat{Y}_i = \frac{N_i}{n_i} \sum_{j \in s_i} y_{ij}$ para toda UPA i selecionada.

A variância de $\hat{Y}_{Plano3}$ é dada por:

$$
\begin{aligned}
V_{Plano3}\left(\hat{Y}_{Plano3}\right) &= V_1\left[E_2\left(\hat{Y}_{Plano3}\right)\right] + E_1\left[V_2\left(\hat{Y}_{Plano3}\right)\right] \\
&= V_1\left[E_2\left(\sum_{i\in a}\frac{\hat{Y}_i}{mp_i}\right)\right] + E_1\left[V_2\left(\sum_{i\in a}\frac{\hat{Y}_i}{mp_i}\right)\right] \\
&= V_1\left[\sum_{i\in a}\frac{E_2\left(\hat{Y}_i\right)}{mp_i}\right] + E_1\left[\sum_{i\in a}\frac{V_2\left(\hat{Y}_i\right)}{m^2p_i^2}\right] \\
&= V_1\left[\sum_{i\in a}\frac{Y_i}{mp_i}\right] + E_1\left[\sum_{i\in a}\frac{1}{m^2p_i^2}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2\right] \\
&= \frac{1}{m}\frac{M}{M-1}\sum_{i\in U}\left(\frac{Y_i}{p_i}-Y\right)^2(1-mp_i)p_i \\
&\quad + \sum_{i\in U}\frac{1}{mp_i}N_i^2\left(\frac{1}{n_i}-\frac{1}{N_i}\right)S_i^2
\end{aligned}
$$

A Tabela 13.7 apresenta um resumo da estimação do parâmetro do total da variável y sob os Planos 1, 2 e 3 de AC2.

Tabela 13.7: Estimadores do parâmetro total sob os Planos 1, 2 e 3 de AC2

Parâmetro	Estimador AC2
Y	$\hat{Y}_{Plano1} = \frac{M}{m}\sum_{i \in a} \hat{Y}_i$ $\hat{Y}_{Plano2} = \frac{1}{m}\sum_{i \in a}\frac{\hat{Y}_i}{p_i} = \hat{Y}_{Plano3}$
$V_{Plano1}\left(\hat{Y}_{Plano1}\right)$	$\hat{V}_{Plano1}\left(\hat{Y}_{Plano1}\right) = M^2\left(\frac{1}{m} - \frac{1}{M}\right)\frac{1}{m-1}\sum_{i \in a}\left(\hat{Y}_i - \bar{y}_C\right)^2 + \frac{M}{m}\sum_{i \in a} N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\hat{S}_i^2$
$V_{Plano2}\left(\hat{Y}_{Plano2}\right)$	$\hat{V}_{Plano2}\left(\hat{Y}_{Plano2}\right) = \frac{1}{m(m-1)}\sum_{i \in a}\left(\frac{N_i\bar{y}_i}{p_i} - \hat{Y}_{Plano2}\right)^2$
$V_{Plano3}\left(\hat{Y}_{Plano3}\right)$	$\hat{V}_{Plano3}\left(\hat{Y}_{Plano3}\right) = \frac{1}{m}\frac{M}{M-1}\sum_{i \in a}\left(\frac{\hat{Y}_i}{p_i} - \hat{Y}_{Plano3}\right)^2 (1 - mp_i)p_i + \sum_{i \in a}\frac{1}{mp_i}N_i^2\left(\frac{1}{n_i} - \frac{1}{N_i}\right)\hat{S}_i^2$

13.8.4 Ideias básicas na escolha de unidades compostas

A seguir são apresentadas algumas ideias básicas a serem adotadas na escolha de unidades compostas em termos da definição das unidades e do plano amostral, das medidas de tamanho e informação auxiliar, da escolha dos tamanhos das UPAs e do número de unidades a serem selecionadas em cada estágio.

Definição das unidades e do plano amostral no caso de pesquisas domiciliares:

a. Usar unidades claramente definidas: mapeamento preciso, para evitar omissões e dupla contagem.

b. Boa cobertura para evitar omissão de unidades.

c. Procedimentos precisos para operações de listagem ou cadastramento de unidades, se necessário.

d. Plano amostral deve ser simples de implementar.

e. Preferir cadastros ou listas disponíveis em computador.

Em termos de medidas de tamanho e informação auxiliar:

f. Medidas de tamanho devem ser tão precisas quanto possível.

g. Informações auxiliares são necessárias para estratificar UPAs, para usar na estimação, etc.

h. UPAs menores geralmente apresentam menores custos de listagem e de deslocamento entre unidades. Porém UPAs menores são mais homogêneas e podem aumentar a variância.

Em termos da escolha dos tamanhos:

i. Geralmente se adota UPAs do maior tamanho possível, tal que uma UPA possa ser coberta por apenas um entrevistador durante a operação de coleta.

j. UPAs de tamanhos parecidos reduzem a variância.

Em termos da escolha de m e n_i:

k. Geralmente, utiliza-se uma das opções equiponderadas dos Planos 1 e 2, para manter a simplicidade.

Sob o Plano 1, precisa escolher: $f_1 = m/M$, a fração amostral do primeiro estágio; e $f_{2i} = n_i/N_i$, as frações amostrais do segundo estágio;

Sob o Plano 2, precisa escolher: $f_1 = m/M$ e $n =$ (tamanho amostral fixo no segundo estágio, que considera que $x_i = N_i$).

l. Em cada caso, há dois “parâmetros” de planejamento a especificar.

Restrições orçamentárias geralmente permitem reduzir a escolha a apenas um “parâmetro”.

Algumas vezes, para o Plano 2, determinamos primeiro n, a carga de trabalho por entrevistador em cada UPA.

Mas em geral, para tamanhos totais de amostra fixados, pode-se escolher entre:

- aumentar m e reduzir n, o que aumenta a precisão mas aumenta o custo, ou então;
- reduzir m e aumentar n, o que reduz a precisão mas também reduz o custo.

Um compromisso é necessário!

Um guia importante é considerar:

$$EPA(AC2;\ \bar{y}_R) \doteq 1 + (n-1)\rho$$

Da equação (13.45) de imediato, segue-se que:

i. se $\rho > 0 \Rightarrow [1 + (n-1)\rho] << [1 + (N-1)\rho$, que é o efeito de conglomeração na AC1. Logo, é interessante manter n pequeno, o que implica em ter mais unidades primárias e subamostras menores.

ii. se $\rho < 0 \Rightarrow [1 + (n-1)\rho] > [1 + (N-1)\rho$. Logo, a melhor alternativa é fazer $n = N$, isto é fazer AC1, tomando menos unidades primárias.

Vale registrar que se a fração de primeiro estágio cresce e, como em geral o custo da undade primária é maior que o da unidade secundária, então o fator custo não deve ser ignorado na determinação dos tamanhos da amostra.

13.8.5 Determinando tamanhos de amostra com AC2PPT

Passo 1: Determine o tamanho amostral requerido para estimar a média populacional $\bar{Y}$ por uma AAS com a margem de erro aceitável d especificada de comum acordo com o cliente:

$$n_{AAS} = \frac{Nz^2_{\alpha/2}S^2_y}{Nd^2 + z^2_{\alpha/2}S^2_y}$$

Mas note que n_{AAS} é solução da seguinte equação:

$$V_{AAS}(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)S^2_y = \left(\frac{d}{z_{\alpha/2}}\right)^2$$

Passo 2: Note que:

$$EPA(AC2PPT;\ \bar{y}_W) = \frac{V_{AC2PPT}(\bar{y}_W)}{V_{AAS}(\bar{y})} \doteq 1 + (\bar{n}-1)\rho$$

(Ver por exemplo Nascimento Silva & Moura (1986), p. 31).

Logo, segue-se que

$$V_{AAS}(\bar{y}) = V_{AC2PPT}(\bar{y}_W) / EPA(AC2PPT; \bar{y}_W)$$

Então podemos escrever:

$$\frac{V_{AC2PPT}(\bar{y}_W)}{EPA(AC2PPT; \bar{y}_W)} = \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2$$

Daí segue-se que:

$$\frac{1}{n} S_y^2 = \frac{1}{N} S_y^2 + \frac{V_{AC2PPT}(\bar{y}_W)}{EPA(AC2PPT; \bar{y}_W)} \implies$$

$$n = S_y^2 \Big/ \left(\frac{1}{N} S_y^2 + \frac{V_{AC2PPT}(\bar{y}_W)}{EPA(AC2PPT; \bar{y}_W)} \right) \implies$$

$$n = N S_y^2 \Big/ \left(S_y^2 + N \frac{V_{AC2PPT}(\bar{y}_W)}{EPA(AC2PPT; \bar{y}_W)} \right)$$

Mas queremos com o plano AC2PPT obter a mesma margem de erro admissível d, logo segue que:

$$V_{AC2PPT}(\bar{y}_W) = \left(\frac{d}{z_{\alpha/2}} \right)^2$$

Levando a equação (13.51) na equação (13.50), segue-se que:

$$n = N S_y^2 \Big/ \left(S_y^2 + \frac{N d^2}{z_{\alpha/2}^2 EPA(AC2PPT; \bar{y}_W)} \right)$$

Que leva à seguinte solução:

$$n_{AC2PPT} = \frac{EPA(AC2PPT;\ \bar{y}_W) N z^2_{\alpha/2} S^2_y}{Nd^2 + EPA(AC2PPT;\ \bar{y}_W) z^2_{\alpha/2} S^2_y} \Longrightarrow$$

$$n_{AC2PPT} \doteq n_{AAS} \times EPA(AC2PPT;\ \bar{y}_W)$$

Ou alternativamente:

$$n_{AC2PPT} \doteq n_{AAS} \times [1 + (n-1)\rho]$$

1. A PME foi encerrada em março de 2016, com a divulgação dos resultados referentes ao mês de fevereiro de 2016, tendo sido substituída, com metodologia atualizada, pela Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua, que abrange todo o País. Detalhes sobre a PME podem ser vistos em https://www.ibge.gov.br/estatisticas/sociais/trabalho/9180-pesquisa-mensal-de-emprego.html?=&t=o-que-e.↩