Capítulo 11 Amostragem Inversa

11.1 Introdução

Na amostragem probabilística usualmente são selecionadas n unidades, que deverão compor a amostra, de uma população ou cadastro de seleção composto de N unidades populacionais, $n < N$.

Essas unidades selecionadas para a amostra serão identificadas e as variáveis de interesse serão medidas utilizando-se a ferramenta adequada de acordo com as característicasda pesquisa em questão. No caso de pesquisas do tipo socioeconômicas, frequentemente são utilizados questionários que a ser respondidos por algum representante da unidade selecionada, que pode ser um domicílio, um estabelecimento industrial, uma escola, etc. Em casos como um processo de controle de qualidade por amostragem poderão ser utilizados instrumentos de medida para conferência das dimensões de uma peça que está sendo fabricada em série ou mesmo a realização de um ensaio específico para testar alguma característica desejável. Muitas vezes esses ensaios podem ser destrutivos e, portanto, este é um exemplo clássico onde a utilização de uma amostra se impõe pela impossibilidade de se testar todas as peças que estiverem sendo fabricadas.

Um dos problemas que podem ocorrer numa investigação por amostragem é a impossibilidade de se entrevistar ou medir todas as n unidades selecionadas para a amostra. Isso pode ocorrer por vários motivos como a dificuldade de se localizar a unidade no campo; a mudança das características da unidade excluindo a mesma do escopo da pesquisa, como, por exemplo, um endereço onde havia um domicílio selecionado que no momento da pesquisa passou a ser uma unidade comercial; a seleção de um domicílio de uso ocasional, como uma casa de veraneio; ausência da pessoa que deveria ser o informante; a simples recusa do informante responder total ou parcialmente a pesquisa. Todas essas questões implicam em uma redução no tamanho da amostra e aumento da imprecisão das estimativas geradas.

Muitas das causas que levam a essa redução da amostra talvez possam ser prevenidas e evitadas por uma atualização mais criteriosa na preparação do cadastro de seleção da amostra, porém é sabido que um cadastro é um ente dinâmico e as mudanças que podem

ocorrer no mesmo estão fora do controle do amostrista. Alterações podem ocorrer no momento da realização da pesquisa.

Existem estratégias para tentar contornar esse problema como, por exemplo, o superdimensionamento da amostra. Principalmente em pesquisas que são repetidas no tempo ou pesquisas com características semelhantes a outras já realizadas, pode-se ter uma estimativa das taxas de perda de unidades amostrais que podem ser úteis para orientar o dimensionamento de uma amostra robusta o suficiente para que, mesmo sofrendo perdas na coleta, possa estimar os parâmetros de interesse com um nível adequado de precisão. Se as perdas forem menores do que o esperado, as estimativas poderão ser até mais precisas o que é até positivo do ponto de vista estatístico.

Outro problema desse tipo de abordagem é quando se está pesquisando eventos de ocorrência rara na população e no cadastro disponível para a seleção da amostra não existe nenhuma indicação sobre essa ocorência nas unidades populacionais. São exemplos desse tipo de pesquisa os estudos sobre determinadas doenças; problemas que podem atingir somente indivíduos de determinado sexo e/ou numa faixa etária restrita; atividades econômicas restritas a um número pequeno de estabelecimentos, etc. Nestes casos pode-se realizar uma etapa prévia de atualização cadastral, comunmente chamada de ***screening*** ou **varredura**, procurando identificar na população aquelas unidades que tem, pelo menos idicativos, das características desejadas. Esse processo, porém, pode ser muito caro inviabilizando sua realização. Em pesquisas que utilizam amostragem de conglomerados o custo dessa opreação pode ser reduzido pela atualização apenas nos conglomerados selecionados para a amostra. Um exemplo desse caso pode ser visto na Pesquisa Nacional por Amostra de Domicílios Contínua - PNAD Contínua do IBGE, em que são atualizados antes da coleta apenas os setores cencitários selecionados como Unidades Primárias de Amostragem - UPA, (IBGE 2014).

As estratégias exemplificadas acima são úteis para tentar minimizar as perdas de unidades amostrais, mas não eliminam o problema da redução no tamanho final da amostra.

Uma outra forma de se tentar garantir que o número de unidades amostrais coletadas seja efetivamente igual ao valor de n desejado é a chamada Amostragem Inversa.

11.2 Amostragem Inversa

A *Amostragem Inversa* foi proposta por (Haldane 1945) como uma forma de enfrentar o problema de amostragem para estimar parâmetros relativos a eventos raros. Em termos gerais, consiste em pesquisar n unidades da população até encontrar m que contenham as

caracterísitcas desejadas. Mesmo assim podem ocorrer casos em que a busca se estenda por toda a população (ou estrato ou conglomerado onde se faça uma subamostragem) sem que se encontre as m unidades de interesse.

Suponha que p seja a proporção de unidades da população que posuam determinada característica e $q = 1 - p$ a proproção das unidades que não possuam a mesma caracterrística. Se forem obaservadas n unidades da população, selecionadas através de um sorteio aleatório, de maneira que nas $n - 1$ primeiras unidades observadas forem encontradas $m - 1$ unidades com a característica desejada e na observação n a unidade tenha a característica, perfazendo o total de m unidades com essa característica, pode-se definir um estimador não viciado para a proporção p (Cochran 1977) como:

$$\hat{p} = \frac{m-1}{n-1} \tag{11.1}$$

A variância do estimador da proporção pode ser aproximada por:

$$\widehat{V}(\hat{p}) = \frac{m(n-m)}{n^2(n-1)} \tag{11.2}$$

O esquema de amostragem inversa pode ser usado no caso geral, onde não se deseja estimar a proporção das unidades que tenha determinada característica mas investigar uma amostra das unidades da população que pertençam a uma determinada população alvo. Por exemplo, pode-se estar interessado em aplicar um questionário às mulheres com idade de 15 até 49 anos, porém não há disponibilidade de um cadastro de tais mulheres, mas apenas do cadastro dos domicílios da área da pesquisa, não havendo indicação de quais deles tem mulheres na faixa etária desejada.

Uma possíbilidade é selecionar sequencialmente os domicílios, com um esquema de seleção aleatória, e aplicar os questionários quando existirem mulheres na faixa etária de interesse e apenas registrar algumas variáveis de controle caso não existam mulheres na faixa de interesse no domicílio selecionado. Assim após visitar n domicílios serão obtidos os questionários relativos aos m domicílios com mulheres de 15 a 49 anos.

11.2.1 Amostragem Inversa Simples

Seja uma população, composta de N unidades, onde se deseja investigar uma amostra dentre as unidades que fazem parte de uma determinada subpopulação, ou seja, de um subconjunto das unidades populacionais que possuam uma determinada particularidade.

Essa população pode ser descrita como a união de dois subconjuntos, onde um deles será composto de $M = pN$ unidades com a particularidade que interessa e outro de $N - M = (1 - p)N$ unidades que não possuam tal particularidade, onde N é total de unidades, conhecido, da população e p é a proporção, desconhecida, de unidades no primeiro subconjunto.

Pode-se definir tais subpopulações como:

$$C \cup \overline{C} = U, \quad C \cap \overline{C} = \emptyset \quad \text{e} \quad C \neq \emptyset,$$

onde C é a subpopulação formada pelas unidades com a particularidade de interesse e $\overline{C}$ a subpopulação complementar.

Um exemplo prático seria o de selecionar uma amostra de domicílios numa localidade para aplicar um questionário a ser respondido apenas por mulheres na faixa etária de 15 até 45 anos. Neste caso poder-se-ia ter acesso a um cadastro de todos os domicílios da localidade, porém sem a indicação de quais desses domicílios têm moradoras na faixa etária de interesse. Os domicílios com mulheres de 15 até 45 anos formariam a subpopulação C e os demais domicílios seriam a subpopulação $\overline{C}$.

No caso de uma Amostra Inversa Simples - AIS sem reposição serão selecionadas através de um mecanismo de AAS, uma amostra de de n unidades até que m dessas unidades sejam originárias da subpopulação C. Dessa maneira, o tamanho da amostra n será uma variável aleatória hipergeométrica negativa, com distribuição de probabilidades dada por:

$$P(n = k) = \frac{M - m + 1}{N - k + 1} \frac{\binom{M}{m-1}\binom{N-M}{k-m}}{\binom{N}{k-1}} \tag{11.3}$$

com valor esperado e variância dados por:

$$E(n) = \frac{(N + 1)m}{M + 1} \tag{11.4}$$

e

$$V(n) = \frac{(N + 1)(M - m + 1)(N - M)}{(M + 1)^2(M + 2)} \tag{11.5}$$

onde $1 \leq m \leq M$ e $m \leq k \leq N$

11.2.1.1 Estimativa da média e total na AIS

Existem várias propostas para se estimar a média (ou total) populacional utilizando a amostragem inversa simples, como o estimador de Horvitz-Thompson apresentado pelos autores em (Horvitz e Thompson 1952) ou o estimador proposto por Des Raj (Raj 1956), ambos descritos em (Jordao 2012). Aqui será apresentado o estimador proposto por (Murthy 1957), que é dado por:

$$\widehat{Y}_M = \sum_{i=1}^{v} y_i \frac{P(s|i)}{P(s)} \tag{11.6}$$

onde v é o número de unidades distintas na amostra, que no caso de amostragem sem reposição será igual a n, $P(s|i)$ é a probabilidade condicional de selecionar a amostra s dado que a unidade i foi selecionada e $P(s)$ é a probabilidade de seleção da amostra s .

Para estimar a média populacional, basta dividir o estimador do total pelo tamanho, N, da população, ou seja:

$$\bar{y}_M = \frac{1}{N} \sum_{i=1}^{v} y_i \frac{P(s|i)}{P(s)} \tag{11.7}$$

No caso da amostragem inversa podemos reescrever os estimadores acima da forma sugerida por (Salehi 2001).

Seja a função indicadora I_i que assume o valor 1 se a unidade i da população foi selecionada para a amostra e 0 caso contrário. No caso da amostragem inversa sabe-se que a última unidade selecionada obrigatoria mente pertence ao subconjunto C da população, como definido acima. Portanto as demais unidades da amostra podem ser alocadas de $(n-1)!$ maneiras, de forma que a amostra s pode ser obtida de $m(n-1)!$ configurações distintas. Para o conjunto das unidades pertencentes à subpopulação C o evento $\{I_i = 1, s\}$ pode ocorres de $(m-1)(n-2)!$ maneiras, enquanto que para as unidades pertencentes à subpopulação $\overline{C}$ esse mesmo evento pode ocorrer de $m(n-2)!$ formas. Sabe-se, ainda, que a seleção das unidades da amostra é feita com equiprobabilidade, onde $P(i) = 1/N$, para $i = 1, 2, \ldots, N$. Portanto:

$$\frac{P(s|i)}{P(s)} = \begin{cases} [N(m-1)] \,/\, [(n-1)m] & \text{se } i \in C \\ N/(n-1) & \text{se } i \in \overline{C} \end{cases} \tag{11.8}$$

Dessa maneira pode-se escrever a equação (11.6) como:

$$\widehat{Y}_M = \frac{N}{n-1} \left(\sum_{i \in C} \frac{m-1}{m} y_i + \sum_{i \in \overline{C}} y_i \right) \tag{11.9}$$

Para obter o estimador da média basta dividir a expressão acima por N:

$$\overline{y}_M = \frac{1}{n-1}\left(\sum_{i\in C}\frac{m-1}{m}y_i + \sum_{i\in\overline{C}} y_i\right) \tag{11.10}$$

11.2.2 Amostragem inversa com probabilidades desiguais com reposição

Suponha que uma população finita consiste em N unidades com valores de interesse associados $y_1, \ldots, y_N$. A probabilidade de seleção inicial da $i-ésima$ unidade é denotada por z_i. O parâmetro a ser estimado é o total da população, dado por $Y = \sum_{i=1}^{N} y_i$. Suponha, como já foi visto anteriormente, que a população é formada por duas subpopulações onde C é a subpopulação formada pelas unidades com uma particularidade de interesse e $\overline{C}$ a subpopulação complementar. Em amostragem inversa de probabilidades desiguais, selecionamos unidades, uma de cada vez, com probabilidades desiguais com reposição até que tenhamos obtido um dado número m de unidades da classe C na amostra. O tamanho da amostra, n, é uma variável aleatória.

A amostra final s pode ser dividida em duas partes: uma parte é o conjunto S_C das m unidades da amostra vindas da classe C e $S_{\overline{C}}$ é o conjunto das $n-m$ unidades da amostra vindas de $\overline{C}$. Sejam k e g os números de unidades distintas em S_C e $S_{\overline{C}}$, e que são indexados por $i = 1, \ldots, k$ e $i = k+1, \ldots, v$ respectivamente. Seja é o número de vezes que a unidade i aparece na amostra. Com uma amostra ordenada, S^*, a probabilidade de obter essa amostra ordenada é $P(S^*) = \prod_{i=1}^{v}(z_i)^{r_i}$, onde a última unidade amostrada pertence ao conjunto S_C.No caso de uma amostra não-ordenada,s, com a última unidade amostrada pertencente a S_C, após alocar a i-ésima unidade de k unidades amostradas em S_C, o resto das unidades amostradas podem ser ordenadas em $\binom{n-1}{r_1, r_i-1, \ldots, v}$ maneiras. Portanto, a amostra s pode ser construída em $\sum_{i=1}^{k}\binom{n-1}{r_1, r_i-1, \ldots, v}$ maneiras. A probabilidade de obter uma particular amostra s é $P(s) = \sum_{i=1}^{k}\binom{n-1}{r_1, r_i-1, \ldots, v}\prod_{i=1}^{v}(z_i)^{r_i}$.

Assim, um estimador não viciado para o total populacional, Y, pode ser obtido por:

$$\widehat{Y} = \hat{p}\overline{y}_C + (1-\hat{p})\overline{y}_{\overline{C}}$$

com sua variância estimada por:

$$\widehat{V}(\widehat{Y}) = (\overline{y}_C - \overline{y}_{\overline{C}})^2\widehat{V}(\hat{p})\frac{\hat{\sigma}_C^2}{m}\left[(m-1)\widehat{V}(\hat{p}) - \hat{p}^*\hat{p}^2\right] + \frac{\hat{\sigma}_{\overline{C}}^2(n-m-1)}{(n-1)(n-2)}$$

onde, $\hat{p}$ é dado pela equação (11.1), $\overline{y}_C = \frac{1}{m}\sum_{i \in s_C}\frac{y_i}{z_i}$ e $\overline{y}_{\overline{C}} = \frac{1}{n-m}\sum_{i \in s_{\overline{C}}}\frac{y_i}{z_i}$

Os resultados acima podem ser demonstrados utilizando o estimador do total populacional, Y, definido por (Murthy 1957).

(Tille 2016) desenvolve a teoria de amostra inversa com probabilidades desiguais a partir de um problema sugerido por uma pesquisa do Statístics Canada, onde é selecionada uma amostra de empresas (ou unidades locais) em estratos representados por regiões econômicas e em cada uma destas empresas deve ser selecionada uma amostra das ocupações que fazem parte de uma lista de ocupações para as quais se deseja obter informações. Não se conhece *à priori* a lista de ocupações de cada empresa e, portanto, deve-se lançar mão da amostragem inversa selecionando as ocupações dentro de cada empresa, pertencentes à lista de ocupações de interesse, até se obter m ocupações distintas. Essas ocupações podem ser selecionadas com probabilidades proporcionais a sua participação (ou prevalência) na lista de ocupações interesse.