Capítulo 3 Estimação Baseada no Plano Amostral

3.1 Estimação de Totais

Devido a sua importância para os desenvolvimentos teóricos em vários dos capítulos subsequentes, alguns resultados básicos relativos à estimação de totais da população finita numa abordagem baseada no plano amostral são relembrados nesta seção. A referência básica usada foi a Seção 2.8 de Särndal et al. (1992). O leitor pode também consultar o capítulo 3 de Silva et al. (2020).

Consideremos o problema de estimar o vetor $\mathbf{Y} = \sum_{i \in U} \mathbf{y}_i$ de totais das P variáveis da pesquisa na população, a partir de uma amostra observada s. Naturalmente, qualquer estimador viável do total $\mathbf{Y}$ só pode depender dos valores das variáveis de pesquisa observados na amostra, contidos em $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{i_n}$, mas não dos valores dessas variáveis para os elementos não pesquisados ($i \in U - s$).

Um estimador usual baseado no plano amostral para o total $\mathbf{Y}$ é o estimador de Horvitz-Thompson (veja seção 3.7 de Silva et al. (2020)), dado por:

$$\widehat{\mathbf{Y}}_{HT} = \sum_{i \in s} \mathbf{y}_i / \pi_i = \sum_{i \in s} d_i \mathbf{y}_i \tag{3.1}$$

onde $d_i = 1/\pi_i$ é o *peso básico* da unidade i.

Na abordagem baseada no planejamento amostral, as propriedades de uma estatística ou estimador são avaliadas com respeito à sua *distribuição de aleatorização*. Denotemos por $E_p(.)$ e $V_p(.)$ os operadores de esperança e variância referentes à distribuição de probabilidades induzida pelo planejamento amostral $p(s)$, que chamaremos daqui por diante de `esperança de aleatorização` e `variância de aleatorização`.

O estimador $\hat{\mathbf{Y}}_{HT}$ é não-viciado para o total $\mathbf{Y}$ com respeito à distribuição de aleatorização, isto é:

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

Além disto, sua variância de aleatorização é dada por

$$V_p\left(\widehat{\mathbf{Y}}_{HT}\right) = \sum_{i\in U}\sum_{j\in U}\left(\frac{d_i d_j}{d_{ij}} - 1\right)\mathbf{y}_i\mathbf{y}'_j \tag{3.2}$$

Um estimador não viciado para a variância de aleatorização de $\hat{\mathbf{Y}}_{HT}$ é dado por:

$$\widehat{V_p}\left(\widehat{\mathbf{Y}}_{HT}\right) = \sum_{i\in s}\sum_{j\in a}\left(d_i d_j - d_{ij}\right)\mathbf{y}_i\mathbf{y}'_j \tag{3.3}$$

O estimador de variância em (3.3) é um estimador não-viciado da variância de aleatorização de $\hat{\mathbf{Y}}_{\pi}$, isto é

$$E_p\left[\widehat{V_p}\left(\widehat{\mathbf{Y}}_{HT}\right)\right] = V_p\left(\widehat{\mathbf{Y}}_{HT}\right) \tag{3.4}$$

desde que $\pi_{ij} > 0 \quad \forall i \neq j \in U$, como vamos supor neste livro.

Exemplo 3.1 Amostragem Aleatória Simples Sem Reposição (AAS)
Quando o plano amostral empregado num levantamento é amostragem aleatória simples sem reposição (AAS), as expressões apresentadas para o estimador de total, sua variância e estimadores desta variância simplificam bastante, porque as probabilidades de inclusão e os pesos básicos das unidades ficam iguais a

$$\pi_i = \frac{n}{N} \text{ e } d_i = \frac{N}{N} \ \forall i \in U \tag{3.5}$$

e

$$\pi_{ij} = \frac{n(n-1)}{N(N-1)} \text{ e } d_{ij} = \frac{N(N-1)}{n(n-1)} \ \forall\, i \neq j \in U \tag{3.6}$$

Essas probabilidades de inclusão e pesos básicos levam às seguintes expressões simplificadas para o caso AAS:

$$\widehat{\mathbf{Y}}_{AAS} = \frac{N}{n}\sum_{i\in s}\mathbf{y}_i = N\,\overline{\mathbf{y}} \tag{3.7}$$

onde

$$\overline{\mathbf{y}} = \frac{1}{n}\sum_{i\in s}\mathbf{y}_i \tag{3.8}$$

$$V_{AAS}\left(\widehat{\mathbf{Y}}_{AAS}\right) = N^2\left(\frac{1}{n} - \frac{1}{N}\right)\mathbf{S}_y\,, \tag{3.9}$$

onde

$$\mathbf{S}_y = \frac{1}{N-1} \sum_{i \in U} \left(\mathbf{y}_i - \overline{\mathbf{Y}} \right) \left(\mathbf{y}_i - \overline{\mathbf{Y}} \right)' \tag{3.10}$$

$$\overline{\mathbf{Y}} = \frac{1}{N} \sum_{i \in U} \mathbf{y}_i = \frac{1}{N} \mathbf{Y} \tag{3.11}$$

Sob AAS, o estimador da variância do estimador de total simplifica para:

$$\widehat{V}_{AAS} \left(\widehat{\mathbf{Y}}_{AAS} \right) = N^2 \left(\frac{1}{n} - \frac{1}{N} \right) \widehat{\mathbf{S}}_y \tag{3.12}$$

onde

$$\widehat{\mathbf{S}}_y = \frac{1}{n-1} \sum_{i \in s} \left(\mathbf{y}_i - \overline{\mathbf{y}} \right) \left(\mathbf{y}_i - \overline{\mathbf{y}} \right)' \tag{3.13}$$

Vários outros estimadores de totais estão disponíveis na literatura de amostragem, porém os que são comumente usados na prática são estimadores ponderados (lineares) da forma

$$\widehat{\mathbf{Y}}_w = \sum_{i \in s} w_i \mathbf{y}_i \tag{3.14}$$

onde w_i é um peso associado à unidade i da amostra ($i \in s$).

O estimador de Horvitz-Thompson é um caso particular de $\widehat{\mathbf{Y}}_w$ em (3.14) quando os pesos w_i são da forma

$$w_i^{HT} = d_i = 1/\pi_i \quad \forall\, i \in s.$$

Outros dois estimadores de totais comumente usados pelos praticantes de amostragem são o *estimador de razão simples* $\widehat{\mathbf{Y}}_R$ e o *estimador de regressão simples* $\widehat{\mathbf{Y}}_{REG}$, dados respectivamente por

$$\widehat{\mathbf{Y}}_R = \sum_{i \in s} w_i^R \, \mathbf{y}_i \tag{3.15}$$

com

$$w_i^R = d_i \times \frac{\sum_{i \in U} x_i}{\sum_{i \in s} d_i \, x_i} = d_i \times \frac{X}{\widehat{X}_{HT}} \tag{3.16}$$

e

$$\widehat{\mathbf{Y}}_{REG} = \sum_{i \in s} w_i^{REG} \, \mathbf{y}_i \tag{3.17}$$

onde

$$w_i^{REG} = d_i \times g_i \tag{3.18}$$

sendo

$$g_i = 1 + x_i \, (X - \widehat{X}_{HT}) / \sum_{i \in s} d_i x_i^2$$

O fator multiplicativo de ajuste de regressão g_i depende de conhecermos o total populacional $\sum_{i \in U} x_i = X$ de uma variável auxiliar x, e do estimador tipo Horvitz-Thompson para esse total dado por $\widehat{X}_{HT} = \sum_{i \in s} d_i \, x_i$.

O estimador de regressão descrito em (3.17) é um caso particular do *estimador de regressão generalizado*, obtido quando se consideram vetores de variáveis auxiliares em vez de uma única variável auxiliar x como aqui. Para uma discussão detalhada do *estimador de regressão generalizado* veja o capítulo 3 de Silva (1996), ou o excelente livro de Särndal et al. (1992). Por sua vez, o *estimador de regressão generalizado* é caso particular da família mais ampla dos *estimadores de calibração*, definidos por Deville e Särndal (1992). Mais informações sobre esta família de estimadores no capítulo 13 de Silva et al. (2020).

Para completar a descrição dos procedimentos de inferência para médias e totais baseados em estimadores ponderados do tipo razão ou regressão, é necessário identificar estimadores para as variâncias de aleatorização correspondentes. Entretanto, os estimadores de razão e regressão são viciados sob a distribuição de aleatorização para pequenas amostras. Em ambos os casos, o vício é desprezível para amostras grandes, e estão disponíveis expressões assintóticas para as respectivas variâncias de aleatorização.

Partindo destas expressões foram então construídos estimadores amostrais das variâncias dos estimadores de razão e regressão, que podem ser encontrados na excelente revisão sobre o tema contida em Särndal et al. (1992), Seção 6.6 e cap. 7. Apesar de sua importância para os praticantes de amostragem, a discussão detalhada desse problema não será incluída neste livro.

O problema da estimação das variâncias de aleatorização para estimadores como os de razão e regressão nos remete a uma questão central da teoria da amostragem. Trata-se dos métodos disponíveis para estimar variâncias de estimadores `complexos` . O caso dos estimadores de razão e regressão para totais e médias foi resolvido faz tempo, e não há muito o que discutir aqui. Entretanto, a variedade de métodos empregados para estimação de variâncias merece uma discussão em separado, pois as técnicas de ajuste consideradas neste livro para incorporar pesos e plano amostral na inferência partindo de dados de pesquisas amostrais complexas depende em grande medida da aplicação de tais técnicas.

3.2 Estimação de Variâncias - Motivação

Em Amostragem, como de resto na Estatística Clássica, a estimação de variâncias é um componente `essencial` da abordagem inferencial adotada: sem estimativas de variância, nenhuma indicação da precisão (e portanto, da qualidade) das estimativas de interesse está disponível. Nesse caso, uma tentação que assola muitos usuários incautos é esquecer que os resultados são baseados apenas em dados de uma amostra da população, e portanto sujeitos a incerteza, que não pode ser quantificada sem medidas de precisão amostral.

Em geral, a obtenção de estimativas de variâncias (alternativamente, de desvios padrões ou mesmo de coeficientes de variação) é requerida para que intervalos de confiança possam ser calculados, e outras formas de inferência realizadas. Intervalos de confiança elaborados com estimativas amostrais são geralmente baseados em aproximações assintóticas da distribuição amostral do estimador pela distribuição normal, usando resultados análogos ao TCL para populações finitas - ver Fuller (2009), tais que intervalos da forma

$$IC\left[\hat{\theta}; \widehat{V}_p\left(\hat{\theta}\right); 1-\alpha\right] = \left[\ \hat{\theta} \pm z_{\alpha/2}\sqrt{\widehat{V}_p\left(\hat{\theta}\right)}\ \right]$$

têm probabilidade de cobertura aproximada $1-\alpha$, com $z_{\alpha/2}$ sendo o quantil que deixa área de $1-\alpha/2$ à sua esquerda na distribuição Normal padrão.

Estimativas de variância podem ser úteis também para outras finalidades, tais como a detecção de problemas não antecipados, tais como observações suspeitas, celas raras em tabelas de contingência, etc.

A estimação de variâncias para os casos padrões de amostragem, isto é, quando os estimadores são lineares nas observações amostrais, não viciados, e todas as probabilidades de inclusão conjuntas são não nulas, é tratada em todos os livros de amostragem convencionais. Apesar disso, os pacotes estatísticos usuais, tais como SAS, SPSS, MINITAB e outros, por muito tempo não ofereciam rotinas prontas para estimar variâncias considerando o plano amostral, nem mesmo para estatísticas simples como estimadores de totais e médias.

Felizmente tal situação mudou, e agora já é possível contar com ferramentas no SAS (procedimentos `SURVEY`), no SPSS (módulo `COMPLEX SAMPLES`) e no STATA (funções `svy`). Mas, a nosso ver, é no pacote `survey` do sistema R que estão disponíveis as melhores ferramentas para estimação de parâmetros a partir de dados de amostras complexas.

Para alguns planos amostrais utilizados na prática, as probabilidades de inclusão conjuntas podem ser nulas (caso de amostragem sistemática) ou difíceis de calcular (caso de alguns esquemas de seleção com probabilidades desiguais). Nesses casos, as expressões fornecidas na Seção 3.1 para os estimadores das variâncias dos estimadores de totais não são mais adequadas.

Em muitos outros casos, como se verá no restante deste livro, os parâmetros de interesse são `não lineares` (diferentes de totais, médias e proporções, por exemplo). Casos comuns que consideraremos mais adiante são a estimação de razões, coeficientes de modelos de regressão etc. Nesses casos é comum que as estatísticas empregadas para estimar tais parâmetros também sejam `não lineares` .

Finalmente, alguns estimadores de variância podem, em alguns casos, produzir estimativas negativas da variância, que são inaceitáveis de um ponto de vista prático (tais como o estimador da expressão (3.3) para alguns esquemas de seleção com probabilidades desiguais e determinadas configurações peculiares da amostra).

Em todos esses casos, é requerido o emprego de técnicas especiais de estimação de variância. É de algumas dessas técnicas que tratam as seções seguintes deste capítulo. A seleção das técnicas discutidas aqui não é exaustiva, e um tratamento mais completo e aprofundado da questão pode ser encontrado no livro de Kirk M. Wolter (2007). Discutimos inicialmente a técnica de `Linearização de Taylor` , em seguida uma abordagem comumente adotada para estimar variâncias para planos amostrais estratificados e conglomerados em vários estágios, com seleção de unidades primárias com probabilidades desiguais, denominada `Método do Conglomerado Primário` (do inglês *Ultimate Cluster*). Por último, trataremos brevemente de uma técnica baseada na ideia de pseudo-replicações da amostra, denominada `Bootstrap` . A combinação dessas três idéias suporta os desenvolvimentos teóricos dos algoritmos empregados pelo pacote `survey` do sistema R para estimação de variâncias - veja T. Lumley (2006) e Thomas Lumley (2010).

3.3 Linearização de Taylor (ou Delta) para Estimar variâncias

Um problema que ocorre frequentemente é o de estimar um vetor de parâmetros $\theta = (\theta_1, \ldots, \theta_K)$ de uma população finita U, que pode ser escrito na forma:

$$\theta = \mathbf{g}(\mathbf{Y})$$

onde $\mathbf{Y} = \sum_{i \in U} \mathbf{y}_i$ é o vetor de totais de Q variáveis de pesquisa.

Poderíamos usar como estimador para o vetor de parâmetros θ o estimador $\hat{\theta}$ dado por:

$$\hat{\theta} = \mathbf{g}\left(\widehat{\mathbf{Y}}_{HT}\right) = \mathbf{g}\left(\sum_{i \in s} d_i \, \mathbf{y}_i\right)$$

No caso particular em que $\mathbf{g}(\bullet)$ é uma função linear dos totais das variáveis de pesquisa, isto é:

$$\theta = \mathbf{A}\mathbf{Y}$$

onde $\mathbf{A}$ é uma matriz de constantes de dimensão $K \times Q$, o estimador $\hat{\theta}$ de θ neste caso seria

$$\hat{\theta} = \mathbf{A}\widehat{\mathbf{Y}}_{HT}$$

Nesse caso particular, é fácil estudar as propriedades do estimador $\hat{\theta}$. Este estimador é não-viciado e tem variância de aleatorização dada por:

$$V_p\left(\hat{\theta}\right) = \mathbf{A}\left[V_p\left(\widehat{\mathbf{Y}}_{HT}\right)\right]\mathbf{A}^{'}$$

onde $V_p\left(\widehat{\mathbf{Y}}_{HT}\right)$ é dado em (3.2).

Quando $\mathbf{g}(\bullet)$ é uma função não linear, podemos usar a técnica de `Linearização de Taylor` (ou Método Delta) para obter aproximações assintóticas para a variância de $\hat{\theta} = \mathbf{g}\left(\widehat{\mathbf{Y}}_{HT}\right)$. Para maiores detalhes sobre esse método, veja por exemplo p. 172 de Särndal et al. (1992) ou p. 486 de Bishop et al. (1975).

Vamos considerar a expansão de $\mathbf{g}\left(\widehat{\mathbf{Y}}_{HT}\right)$ em torno de $\mathbf{Y}$, até o termo de primeira ordem, desprezando o resto, dada por:

$$\hat{\theta} \simeq \hat{\theta}_L = \mathbf{g}(\mathbf{Y}) + \mathbf{\Delta g}(\mathbf{Y})\left(\widehat{\mathbf{Y}}_{HT} - \mathbf{Y}\right) \tag{3.19}$$

onde $\mathbf{\Delta g}(\mathbf{Y})$ é a matriz Jacobiana $K \times Q$ cuja q-ésima coluna é $\partial \mathbf{g}(\mathbf{Y}) / \partial Y_q$, para $q = 1, \ldots, Q$.

A ideia básica do método de linearização é aproximar a variância do estimador $\hat{\theta}$ pela variância do `estimador linearizado` $\hat{\theta}_L$ dado pelo lado direito da expressão (3.19). Para obter a variância do estimador linearizado, note que $ $ é uma constante, e que

$$\begin{aligned}
\mathbf{\Delta g}(\mathbf{Y})\left(\widehat{\mathbf{Y}}_{HT} - \mathbf{Y}\right) &= \mathbf{\Delta g}(\mathbf{Y})\widehat{\mathbf{Y}}_{HT} - \mathbf{\Delta g}(\mathbf{Y})\mathbf{Y} \\
&= \textstyle\sum_{i \in s} d_i \, \mathbf{\Delta g}(\mathbf{Y})\mathbf{y}_i - \sum_{i \in U} \mathbf{\Delta g}(\mathbf{Y})\mathbf{y}_i \\
&= \textstyle\sum_{i \in s} d_i \, \mathbf{z}_i - \sum_{i \in U} \mathbf{z}_i = \widehat{\mathbf{Z}}_{HT} - \mathbf{Z}
\end{aligned}$$

onde $\mathbf{z}_i = \boldsymbol{\Delta}\mathbf{g}(\mathbf{Y})\mathbf{y}_i$.

Logo, a variância aproximada por linearização do estimador $\hat{\theta}$ pode ser obtida usando a expressão (3.2)

$$V_p\left(\hat{\theta}\right) \doteq V_p\left(\widehat{\mathbf{Z}}_{HT}\right)$$

Este resultado segue porque na expressão do lado direito o único termo que tem variância de aleatorização é $\widehat{\mathbf{Z}}_{HT}$.

Um estimador consistente de $V_p\left(\hat{\theta}\right)$ é dado por:

$$\widehat{V}_p\left(\hat{\theta}\right) = \widehat{V}_p\left(\widehat{\mathbf{Z}}_{HT}\right) \tag{3.20}$$

onde $\widehat{V}_p\left(\widehat{\mathbf{Z}}_{HT}\right)$ é dado em (3.3), onde substituímos o vetor de variáveis resposta original $\mathbf{y}_i$ pelo vetor de variáveis linearizadas $\mathbf{z}_i = \boldsymbol{\Delta}\mathbf{g}(\mathbf{Y})\mathbf{y}_i$.

Linearização de Taylor pode ser trabalhosa, porque para cada parâmetro/estimador de interesse são requeridas derivações e cálculos específicos. Felizmente, grande parte das situações de interesse prático estão hoje cobertas por pacotes estatísticos especializados na estimação de medidas descritivas e parâmetros de modelos, e suas respectivas variâncias de aleatorização empregando o método de linearização, de modo que essa desvantagem potencial tende a se diluir.

Linearização de Taylor pode não ser imediatamente possível, pois pode ocorrer que as quantidades de interesse não podem ser expressas como funções de totais ou médias populacionais (este é o caso de quantis de distribuições, por exemplo). Para estes casos será necessário recorrer a outras técnicas de estimação de variâncias, como discutido, por exemplo, em Kirk M. Wolter (2007).

3.4 Equações de Estimação

Até aqui, falamos da estimação de totais e de parâmetros que podem ser escritos como funções de totais. O caminho para obter resultados gerais referentes a muitos outros parâmetros de interesse é o que discutimos nesta seção.

Se um parâmetro populacional de interesse θ_U é uma solução única de um sistema de equações de estimação definidas como

$$\sum_{i \in U} \mathbf{u_i}(\theta) = \mathbf{0} \tag{3.20}$$

para uma função $\mathbf{u}(\bullet)$ conhecida, então é possível estimar o parâmetro θ_U usando o estimador $\hat{\theta}$ obtido resolvendo as equações de estimação amostrais:

$$\sum_{i \in s} d_i \, \mathbf{u_i}(\theta) = \mathbf{0} \tag{3.21}$$

O estimador $\hat{\theta}$ é consistente paraθ_U, e adiante mostraremos como o método de Linearização de Taylor pode ser usado para estimar a sua variância. Antes, porém, vamos usar alguns exemplos para ilustrar casos particulares relevantes de como aplicar essa ideia.

Exemplo 3.2 Estimação de médias populacionais
Para ilustrar a aplicação da abordagem de equações de estimação, considere o caso em que a função $\mathbf{u_i}(\theta) = y_i - \theta$. Nesse caso, as equações de estimação populacionais (3.20) simplificam para:

$$\sum_{i \in U} \mathbf{u_i}(\theta) = \sum_{i \in U} (y_i - \theta) = \mathbf{0}$$

Resolvendo esta equação, obtemos:

$$\theta_U = \frac{1}{N} \sum_{i \in U} y_i = \overline{Y}$$

A solução das equações de estimação amostrais fornece:

$$\hat{\theta} = \frac{\sum_{i \in s} d_i y_i}{\sum_{i \in s} d_i} = \overline{y}_{H\grave{a}jek}$$

que é o conhecido estimador de Hàjek da média populacional.

Exemplo 3.3 Estimação de razões populacionais
Considere agora o caso em que a função $\mathbf{u_i}(\theta) = y_i - \theta z_i$. Nesse caso, as equações de estimação populacionais (3.20) simplificam para:

$$\sum_{i \in U} \mathbf{u_i}(\theta) = \sum_{i \in U} (y_i - \theta z_i) = \mathbf{0}$$

Resolvendo esta equação, obtemos:

$$\theta_U = \frac{\sum_{i \in U} y_i}{\sum_{i \in U} z_i} = \frac{Y}{Z} = R$$

A solução das equações de estimação amostrais correspondentes fornece:

$$\hat{\theta} = \frac{\sum_{i \in s} d_i y_i}{\sum_{i \in s} d_i z_i} = \frac{\widehat{Y}_{HT}}{\widehat{Z}_{HT}} = \widehat{R}$$

Os exemplos apresentados ilustram que a estimação de médias e razões populacionais são casos particulares simples da abordagem mais geral de *equações de estimação*. Essa abordagem também se mostrará útil quando lidarmos com a estimação de parâmetros sob vários tipos de modelos paramétricos, que será apresentada nos capítulos seguintes deste livro. É também graças a ela que foi possível desenvolver software genérico para estimação a partir de amostras complexas, como é o caso do pacote `survey` do sistema R.

A estimação de variâncias nesse caso pode ser feita usando o método de Linearização de Taylor, empregando a estratégia de calcular variáveis linearizadas z definidas como na Seção 3.3. Esta é a estratégia adotada no pacote `survey` do sistema R.

3.5 Método do Conglomerado Primário

A ideia central do `Método do Conglomerado Primário` (do inglês `Ultimate Cluster`) para estimação de variâncias para estimadores de totais e médias em planos amostrais de múltiplos estágios, proposto por Hansen et al. (1953), é considerar apenas a variação entre informações disponíveis no nível das unidades primárias de amostragem (UPAs), isto é, dos *conglomerados primários*, e admitir que estes teriam sido selecionados com reposição da população de UPAs. Esta ideia é simples, porém bastante poderosa, porque permite acomodar uma enorme variedade de planos amostrais envolvendo estratificação, amostragem conglomerada e seleção com probabilidades desiguais (com ou sem reposição) tanto das UPAs como das demais unidades de amostragem.

Os requisitos fundamentais para permitir a aplicação deste método são que estejam disponíveis estimadores não viciados dos totais da variável de interesse para cada um dos conglomerados primários selecionados, e que pelo menos dois destes sejam selecionados em cada estrato (se a amostra for estratificada no primeiro estágio).

Embora o método tenha sido originalmente proposto para estimação de totais, pode ser aplicado também para estimar (por linearização) quantidades populacionais que possam ser representadas como funções de totais, conforme discutido na Seção 3.3. De fato, esse método fornece a base para ferramentas dos sistemas estatísticos para cálculo de variâncias considerando o plano amostral, tais como o pacote `survey` do R, as funções `svy` do STATA, o módulo `Complex Samples` do SPSS, as procs `Survey` do SAS, entre outros.

Para descrever o método, considere um plano amostral em vários estágios, no qual n_h unidades primárias de amostragem (UPAs) foram selecionadas no estrato h, com $h = 1, \ldots, H$. Denotemos por π_{hi} a probabilidade de inclusão na amostra da UPA

(conglomerado primário) i do estrato h, e por $\widehat{Y}_{hi}$ um estimador não viciado do total Y_{hi} da variável de pesquisa y na i-ésima UPA do estrato h. Então um estimador não viciado do total $Y = \sum_{h=1}^{H} \sum_{i=1}^{N_h} Y_{hi}$ da variável de pesquisa y na população é dado por

$$\widehat{Y}_{CP} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \widehat{Y}_{hi}/\pi_{hi} \tag{3.22}$$

e um estimador não viciado da variância de aleatorização correspondente por

$$\widehat{V}_p\left(\widehat{Y}_{CP}\right) = \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\frac{\widehat{Y}_{hi}}{\pi_{hi}} - \frac{\widehat{Y}_h}{n_h}\right)^2 \tag{3.23}$$

onde $\widehat{Y}_h = \sum_{i=1}^{n_h} \widehat{Y}_{hi}/\pi_{hi}$ para $h = 1, \ldots, H$. (Veja por exemplo, Shah et al. (1993), p. 4).

Embora na prática a seleção das UPAs seja geralmente feita sem reposição, o estimador do Método do Conglomerado Primário (MCP) aqui apresentado pode fornecer uma aproximação razoável da correspondente variância de aleatorização, especialmente nos casos em que as frações amostrais de UPAs são pequenas nos estratos. Isso ocorre porque planos amostrais sem reposição são em geral mais eficientes que planos com reposição de igual tamanho.

Tal aproximação é largamente utilizada pelos praticantes de amostragem para estimar variâncias de quantidades descritivas usuais tais como totais e médias (com a devida adaptação) devido à sua simplicidade, comparada com a complexidade muito maior envolvida com o emprego de estimadores de variância que tentam incorporar todas as etapas de planos amostrais conglomerados em vários estágios. Uma discussão sobre a qualidade dessa aproximação e alternativas pode ser encontrada em Särndal et al. (1992), p. 153.

3.6 Métodos de Replicação

A ideia de usar métodos indiretos ou de replicação para estimar variâncias em amostragem não é nova. Mahalanobis (1939), Mahalanobis (1944) e Deming (1956) foram os precursores e muitos desenvolvimentos importantes se seguiram. Hoje em dia várias técnicas baseadas nessa ideia são rotineiramente empregadas por praticantes de amostragem, e inclusive formam a base para pacotes especializados de estimação tais como WesVarPC (veja Westat (1996)).

A ideia básica original foi construir a amostra de tamanho n como a união de G amostras de tamanho n/G cada uma, selecionadas de forma independente e usando o mesmo plano amostral, onde G é o número de `réplicas` . Nesse caso, se θ é o parâmetro-alvo, e $\hat{\theta}_g$ é um estimador não viciado de θ baseado na g-ésima réplica $(g = 1, \ldots, G)$, segue-se que

$$\hat{\theta}_R = \frac{1}{G}\sum_{g=1}^{G}\hat{\theta}_g$$

é também um estimador não viciado de θ e

$$\widehat{V}_R\left(\hat{\theta}_R\right) = \frac{1}{G\left(G-1\right)}\sum_{g=1}^{G}\left(\hat{\theta}_g - \hat{\theta}_R\right)^2 \tag{3.24}$$

é um estimador não viciado (de replicação) da variância do estimador $\hat{\theta}_R$.

Note que desde que as réplicas sejam construídas de forma independente conforme indicado, os estimadores $\hat{\theta}_R$ e $\widehat{V}_R\left(\hat{\theta}_R\right)$ são não viciados qualquer que seja o plano amostral empregado para selecionar a amostra de cada réplica, o que faz desta uma técnica flexível e genérica. Além disso, a abordagem de replicação é bastante geral, pois os estimadores aos quais se aplica não precisam ser necessariamente expressos como funções de totais, como ocorre com a técnica de Linearização de Taylor discutida na Seção 3.3.

Apesar destas vantagens, a aplicação prática desta técnica de forma exata é restrita porque, em geral, é menos eficiente, inconveniente e mais caro selecionar G amostras (réplicas) independentes com o mesmo esquema, se comparado à seleção de uma única amostra de tamanho n diretamente. Além disto, se o número de réplicas G for pequeno, o estimador de variância pode ser instável.

Mesmo quando a amostra não foi selecionada exatamente dessa forma, a construção de réplicas a posteriori para fins de estimação de variâncias em situações complexas é também uma ideia simples de aplicar, poderosa e flexível, por acomodar uma ampla gama de planos amostrais e situações de estimação de interesse. Quando as réplicas são construídas após a pesquisa (a posteriori), mediante repartição (por sorteio) da amostra pesquisada em G grupos mutuamente exclusivos de igual tamanho, estas são chamadas de `réplicas dependentes` ou `grupos aleatórios` (do inglês *random groups*). As expressões fornecidas para o estimador de replicação e sua variância são também empregadas nesse caso como uma aproximação, mas não possuem as mesmas propriedades do caso de réplicas independentes.

Uma pesquisa importante e de grande porte em que esta ideia é aplicada é a pesquisa de preços para formar o índice de Preços ao Consumidor (do inglês *Consumer Price Index - CPI*) do US Bureau of Labour Statistics (2020), p. 46, que utiliza duas ou mais réplicas para formar a amostra de itens cujos preços são pesquisados.

É importante observar que a repartição da amostra em grupos aleatórios a posteriori precisa considerar o plano amostral empregado e pode não ser possível em algumas situações. Idealmente, tal repartição deveria ser feita respeitando estratos e alocando UPAs inteiras (isto é, com todas as respectivas unidades subordinadas). K. M. Wolter (1985), p. 31, discute algumas regras sobre como fazer para respeitar o plano amostral ao fazer a repartição da amostra a posteriori, porém recomendamos que o interessado no uso dessa técnica exerça cautela.

Além da modificação da interpretação das réplicas no caso de serem formadas a posteriori, é comum também nesse caso empregar um estimador para o parâmetro θ baseado na amostra completa (denotado $\hat{\theta}$), e um estimador de variância mais conservador que o estimador $\widehat{V}_R\left(\hat{\theta}_R\right)$ anteriormente apresentado, dado por

$$\widehat{V}_{RG}\left(\hat{\theta}\right) = \frac{1}{G\left(G-1\right)}\sum_{g=1}^{G}\left(\hat{\theta}_g - \hat{\theta}\right)^2 \tag{3.25}$$

Um exemplo de aplicação desta técnica pode ser encontrado na forma recomendada para estimação de variâncias a partir das Amostras de Uso Público do Censo Demográfico Brasileiro de 80 (veja IBGE (1985)).

Nesta seção descreveremos duas outras dessas técnicas baseadas em replicações. A primeira é o método de `jackknife`. Este método foi originalmente proposto por Quenoille (1949) e Quenoille (1956) como uma técnica para redução de vício de estimadores, num contexto da Estatística Clássica. A ideia central consiste em repartir a amostra (a posteriori, como no caso do método dos grupos aleatórios) em G grupos mutuamente exclusivos de igual tamanho n/G. Em seguida, para cada grupo formado calcular os chamados pseudo-estimadores dados por

$$\hat{\theta}_{(g)} = G\hat{\theta} - \left(G-1\right)\hat{\theta}_g$$

onde $\hat{\theta}_g$ é um estimador de θ obtido da amostra após eliminar os elementos do grupo g, empregando a mesma forma funcional adotada no cálculo do estimador $\hat{\theta}$ que considera a amostra inteira.

A estimação da variância por esse método pode então ser feita de duas maneiras alternativas, usando um dos estimadores dados por

$$\widehat{V}_{J1}\left(\hat{\theta}\right) = \frac{1}{G\left(G-1\right)}\sum_{g=1}^{G}\left(\hat{\theta}_{(g)} - \hat{\theta}_J\right)^2 \tag{3.26}$$

ou

$$\widehat{V}_{J2}\left(\hat{\theta}\right) = \frac{1}{G\left(G-1\right)}\sum_{g=1}^{G}\left(\hat{\theta}_{(g)} - \hat{\theta}\right)^2 \tag{3.27}$$

onde $\hat{\theta}_J = \frac{1}{G}\sum_{g=1}^{G}\hat{\theta}_{(g)}$ é um estimador pontual *jackknife* para θ, alternativo ao estimador da amostra inteira $\hat{\theta}$.

Observação A descrição do método *jackknife* aqui apresentada não cobre o caso de planos amostrais estratificados, que é mais complexo. Para detalhes sobre este caso, consulte K. M. Wolter (1985), pág. 174.
Observação O estimador $\widehat{V}_{J2}\left(\hat{\theta}\right)$ é mais conservador que o estimador $\widehat{V}_{J1}\left(\hat{\theta}\right)$.
Observação É comum aplicar a técnica fazendo o número de grupos igual ao tamanho da amostra, isto é, tomando $G = n$ e portanto eliminando uma observação da amostra de cada vez ao calcular os pseudo-valores. Essa regra deve ser aplicada considerando o número de UPAs quando o plano amostral é em múltiplos estágios, pois as UPAs devem sempre ser eliminadas com todas as unidades subordinadas.
Os estimadores de variância do método `jackknife` fornecem resultado idêntico aos dos estimadores usuais de variância quando aplicados para o caso de estimadores lineares nas observações amostrais. Além disso, suas propriedades são razoáveis para vários outros casos de estimadores não lineares de interesse (veja, por exemplo, Cochran (1977), p. 321 e K. M. Wolter (1985), p. 306). A situação merece maiores cuidados para o caso de quantis ou estatísticas de ordem, tais como a mediana e o máximo, pois neste caso essa técnica não funciona bem K. M. Wolter (1985), p. 163.

O pacote WesVarPC - Westat (1996) - baseia suas estimativas de variância principalmente no método `jackknife` , embora também possua uma opção para usar outro método conhecido como de replicações de meias amostras balanceadas (do inglês *balanced half-sample replication*).

O outro método de replicação que vamos considerar é uma variante do método *bootstrap* proposta por Rao et al. (1992). O método consiste dos seguintes passos:

1. Selecione amostras aleatórias simples com reposição de m_h das n_h UPAs de cada estrato $h = 1, \ldots, H$.

2. Calcule as contagens m^*_{hi} de vezes que cada UPA i aparece na amostra selecionada no estrato h; note que $\sum_i m^*_{hi} = m_h$ para todo estrato h;

3. Defina *pesos bootstrap* para as unidades da amostra selecionada em (1) usando:

$$w^*_{hik} = \left[1 - \left(\frac{m_h}{n_h - 1}\right)^{1/2} + \left(\frac{m_h}{n_h - 1}\right)^{1/2} \times \frac{n_h}{m_h} \times m^*_{hi}\right] \times w_{hik} \tag{3.28}$$

onde w_{hik} é o peso da unidade k da UPA i do estrato h. Note que quando uma UPA i não é selecionada, sua contagem m^*_{hi} será igual a zero, e o terceiro termo dentro do colchete será nulo.

4. Calcule uma estimativa $\hat{\theta}_b$ para o parâmetro de interesse usando os pesos bootstrap w^*_{hik} em lugar dos pesos originais w_{hik}.
5. Repita os passos 1) a 4) um número B grande de vezes.
6. Estime a variância do estimador $\hat{\theta}$ com:

$$\widehat{V}_B\left(\hat{\theta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b - \hat{\theta}\right)^2 \tag{3.29}$$

A Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) do IBGE passou a usar este método bootstrap para estimação da precisão dos indicadores que divulga a partir do terceiro trimestre de 2021.

Embora computacionalmente mais custoso que o método da Linearização de Taylor, o método bootstrap aqui descrito tem como vantagem a aplicação em casos onde o estimador não é função suave de totais populacionais, tais como separatrizes (quantis), algumas medidas de desigualdade e pobreza etc. Além disso, o método pode ser aplicado com qualquer software que permita implementar o algoritmo descrito, e não requer pacotes especializados. Vale mencionar, entretanto, que este método está disponível no pacote `survey` do sistema R. Sua utilização será ilustrada em capítulos posteriores.

XXX Parei aqui

3.7 Laboratório de R

Vamos utilizar dados da Pesquisa de Padrão de Vida (PPV) do IBGE para ilustrar alguns métodos de estimação de variâncias. Vamos considerar a estimação da proporção de analfabetos na faixa etária acima de 14 anos. Os dados da pesquisa encontram-se no data frame . A variável `analf2` é indicadora da condição de analfabetismo na faixa etária acima de 14 anos e a variável `faixa2` é indicadora da faixa etária acima de 14 anos. Queremos estimar a proporção de analfabetos na faixa etária acima de 14 anos na região Sudeste. Antes apresentamos o método de estimação de variância por linearização de Taylor

Vamos criar duas variáveis:

- analf - variável indicadora da condição de analfabetismo: `v04a01` ou `v04a02` igual a 2;

- faixa - variável indicadora de faixa etária entre 7 e 14 anos.

```
library(survey)
ppv_dat <- readRDS("./data/ppv.rds") # carrega dados
# cria objeto de desenho
ppv_plan<-svydesign(ids = ~nsetor, strata = ~estratof,
data = ppv_dat, nest = TRUE, weights = ~pesof)
# atualiza objeto de desenho com novas variáveis
ppv_plan<-update(ppv_plan,
  analf=(v04a01 == 2 | v04a02 == 2)*1,
  faixa=(v02a08 >= 7 & v02a08 <= 14) *1,
  analf.faixa= (analf==1 & faixa==1)*1
)
```

Como estamos interessados em estimativas relativas à Região Sudeste, vamos restringir o desenho a esse domínio:

```
ppv_se_plan <- subset(ppv_plan, regiao == 2)
```

Vamos estimar os totais das variáveis `analf.faixa` e `faixa`:

```
analf_faixa_tot_est<-svytotal(~analf.faixa+faixa ,ppv_se_plan )
Vcov.Y1.Y2<-vcov(analf_faixa_tot_est)
```

Substituindo os valores na expressão (3.20), obtemos a estimativa da variância da razão de totais das variáveis `analf.faixa` e `faixa`.

```
y1hat<-coef(analf_faixa_tot_est)[1]
y2hat<-coef(analf_faixa_tot_est)[2]
Var.raz<-(1/y2hat)*(1/y2hat)*Vcov.Y1.Y2[1,1]+2*(1/y2hat)*(-y1hat/y2hat^2)*Vcov.Y1.Y2[1,2
(-y1hat/y2hat^2)*(-y1hat/y2hat^2)*Vcov.Y1.Y2[2,2]
# estimativa do desvio-padrão
sqrt(Var.raz)
```

```
##  faixa
## 0,0118
```

Podemos calcular diretamente o desvio-padrão:

```
svyratio(~analf.faixa, ~faixa, ppv_se_plan)
```

```
## Ratio estimator: svyratio.survey.design2(~analf.faixa, ~faixa, ppv_se_plan)
## Ratios=
##             faixa
## analf.faixa 0,119
## SEs=
##              faixa
## analf.faixa 0,0118
```

A estimativa do desvio-padrão obtida por meio da função `svyratio` coincide com a obtida diretamente pelo método de linearização, e é igual a 0,012. O método default para estimar variâncias usado pela library `survey` Thomas Lumley (2017) do R é o de linearização de Taylor.

A library `survey` dispõe de métodos alternativos para a estimação de variância. Vamos utilizar os métodos de replicação de *Jackknife* e de *Bootstrap* para estimar esta variância de razão. Inicialmente, vamos converter o objeto de desenho `ppv1_se_plan` em um objeto de desenho de replicação de tipo *Jackknife*, contendo as réplicas de pesos que fornecem correspondentes réplicas de estimativas.

```
ppv_se_plan_jkn<-as.svrepdesign(ppv_se_plan,type="JKn")
svyratio(~analf.faixa, ~faixa, ppv_se_plan_jkn)
```

```
## Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv_se_plan_jkn)
## Ratios=
##             faixa
## analf.faixa 0,119
## SEs=
##        [,1]
## [1,] 0,0118
```

Para o tipo *Bootstrap*, temos:

```
ppv_se_plan_boot<-as.svrepdesign(ppv_se_plan,type="bootstrap")
svyratio(~analf.faixa, ~faixa, ppv_se_plan_boot)
```

```
## Ratio estimator: svyratio.svyrep.design(~analf.faixa, ~faixa, ppv_se_plan_boot)
## Ratios=
##             faixa
## analf.faixa 0,119
## SEs=
##       [,1]
## [1,] 0,013
```

Vamos apresentar mais detalhes sobre a obtenção dos estimadores de *Jackknife* e *Bootstrap* na library `survey` Thomas Lumley (2017). A classe do objeto `ppv_se_plan_jkn` é `svyrep.design` e ele contém as seguintes componentes:

```
class(ppv_se_plan_jkn)
```

```
## [1] "svyrep.design"
```

```
names(ppv_se_plan_jkn)
```

```
##  [1] "repweights"       "pweights"         "type"             "rho"
##  [5] "scale"            "rscales"          "call"             "combined.weights"
##  [9] "selfrep"          "mse"              "variables"        "degf"
```

A componente `repweights` é uma lista com duas componentes: `weights` e `index` . A componente `weights` é uma matriz de dimensão 276×276, onde 276 é o número de conglomerados primários do plano amostral da PPV na região Sudeste. A partir desta matriz, podemos obter 276 réplicas de pesos de desenho de Jackknife.

```
ppv_se_dat<-ppv_se_plan_jkn$variables
nrow(ppv_se_dat)
```

```
## [1] 8903
```

```
ncong<-sum(with(ppv_se_dat,tapply( nsetor,estratof, function(t) length(unique(t)))))
ncong
```

```
## [1] 276
```

O argumento `compress` da função `as.svrepdesign` permite especificar se, na saída da função, a matriz `weights` será na forma comprimida ou não. Na aplicação feita foi usado o valor default que é a forma comprimida. A forma não comprimida da matriz `weights` tem 8903 linhas e 276 colunas. A forma comprimida permite economizar memória, e pode ser facilmente convertida para a forma não comprimida, utilizando-se a componente `index` .

No método *jackknife*, cada um dos conglomerados primários é removido, e a réplica correspondente dos pesos é o produto do peso amostral original por um fator apropriado, definido da forma a seguir. Suponhamos que foi removido um conglomerado no estrato h, então os pesos do plano amostral serão multiplicados por:

- 0 para as unidades no conglomerado removido;
- $m_h/(m_h-1)$ para unidades pertencentes a outros conglomerados do estrato h;
- 1 para unidades em estratos $h' \neq h$.

Podemos obter a matriz de fatores de correção do peso amostral na forma não comprimida da seguinte maneira:

```
fact_peso_comp_mat<-ppv_se_plan_jkn$repweights[[1]]
ind_cong <-ppv_se_plan_jkn$repweights[[2]]
fat_pesos_mat<- fact_peso_comp_mat[ind_cong,]
str(fat_pesos_mat)
```

```
##  num [1:8903, 1:276] 0 0 1,06 1,06 1,06 ...
```

Podemos obter matriz de réplicas de pesos multiplicando cada coluna dessa matriz pelos pesos do plano amostra:

```
rep_pesos_mat<-weights(ppv_se_plan)*fat_pesos_mat
```

Utilizando esta matriz de réplicas de pesos, podemos obter réplicas correspondentes de estimativas da razão.

```
rep_est_raz<-numeric(ncol(rep_pesos_mat))
for (i in 1:ncol(rep_pesos_mat)){
rep_est_raz[i]<-sum(rep_pesos_mat[,i]*ppv_se_dat$analf.faixa)/sum(rep_pesos_mat[,i]*ppv_
}
```

A partir destas réplicas de estimativas da razão, finalmente estimamos a variância:

```
mean_raz<-mean( rep_est_raz[ppv_se_plan_jkn$rscales>0])
var_jack_raz<- sum((rep_est_raz-mean_raz)^2*ppv_se_plan_jkn$rscales)*ppv_se_plan_jkn$sca
round(sqrt(var_jack_raz),5)
```

```
## [1] 0,0118
```

A library `survey` Thomas Lumley (2017) fornece uma função para estimar a variância de uma função de totais a partir das réplicas de pesos:

```
var_raz_rep<-withReplicates(ppv_se_plan_jkn, function(w,ppv_se_dat) sum(w*ppv_se_dat$ana
var_raz_rep
```

```
##      theta   SE
## [1,] 0,119 0,01
```

Resultado que coincide com a estimativa obtida pela aplicação da função `svyratio` .

A vantagem de utilizar métodos de replicação é a facilidade com que estimamos a variância de qualquer característica da população, cujo estimador pontual é conhecido. Por exemplo, se quisermos estimar a variância da razão das taxas de analfabetos nas faixas etárias de 0 a 14 anos e acima de 14 anos podemos usar as mesmas réplicas de pesos:

```
withReplicates (ppv_se_plan_jkn,function(w,ppv_se_dat) with(ppv_se_dat,
(sum(w*(analf==1&faixa==1))/sum(w*(faixa==1)))/(sum(w*(analf==1&faixa==0))/sum(w*(faixa=
))
```

```
##      theta   SE
## [1,] 0,504 0,05
```

O erro padrão da razão entre razões estimada no exemplo anterior pode ser estimado por linearização de Taylor, usando-se a função `svycontrast()` da library survey:

```
# cria variáveis dummies:
ppv_se_plan <- update(ppv_se_plan,
num1 = as.numeric(analf==1 & faixa==1),
num2 = as.numeric(analf==1 & faixa==0),
den1 = as.numeric (faixa == 1),
den2 = as.numeric(faixa == 0)
)
# estima totais e matriz de covariância de estimativas de totais
comp.tot <- svytotal(~num1+num2+den1+den2, ppv_se_plan)

# estima razão de razões:
svycontrast(comp.tot, quote((num1/den1)/(num2/den2)))
```

```
##          nlcon   SE
## contrast 0,504 0,05
```