4
Dados amostrais complexos e a estimação dos coeficientes de escalonabilidade

Este capítulo inicialmente discute a análise estatística adequada para dados gerados por pesquisas amostrais complexas. É apresentada a importância da análise dos dados amostrais complexos e posteriormente, o desenvolvimento da estimação pontual e da variância dos estimadores de quantidades populacionais de interesse.

No final do capítulo é abordada a estimação pontual e da variância dos estimadores propostos para os coeficientes de escalonabilidade H_{ij}, H_i e H da TRIN no contexto da amostragem complexa de populações finitas.

4.1 Dados amostrais complexos

Muitas técnicas estatísticas tradicionais são formuladas para modelar e analisar dados que são realizações de variáveis (vetores) aleatórias independentes e identicamente distribuídas (Scott e Holt, 1982; Skinner, 1986). Entretanto, dados desse tipo são raros na prática (Chambers e Skinner, 2003). Essas técnicas não levam em consideração os planos amostrais complexos que freqüentemente são empregados para a obtenção dos dados.

Tem crescido o uso de dados oriundos de amostras complexas, isto é, de amostras envolvendo múltiplos estágios de seleção, sorteio sistemático, amostragem com probabilidade proporcional a uma medida de tamanho, estratificação, etc. Logo, é necessário poder incorporar estas características do plano amostral no uso descritivo ou analítico dos dados (Heeringa, 2010).

Existem sérias conseqüências nas estimativas pontuais de quantidades populacionais de interesse e na qualidade (precisão) dessas estimativas, caso seja feita a análise estatística de dados amostrais complexos como se fossem observações amostrais independentes e identicamente distribuídas. Dessa forma, de modo geral, as estimativas pontuais são viciadas e com qualidade comprometida (Heeringa, 2010).

Para a medição estatística de variáveis latentes com base em dados que serão coletados de pesquisas amostrais complexas, é imprescindível a incorporação das seguintes características: estratificação, conglomeração em vários estágios e ponderação desigual na estimação pontual e também da variância dos estimadores dos coeficientes H_{ij}, H_i e H.

4.2 Estimação pontual

Algumas vezes dados oriundos de amostras complexas são usados para fins puramente descritivos (estimação de totais, proporções, razões, dentre outros). Outras vezes, porém, sua utilização é feita para fins analíticos (formulação, seleção, ajuste e interpretação de modelos) onde o analista busca estabelecer a natureza de relações ou associações entre variáveis, no sentido de extrair conclusões aplicáveis também para populações distintas.

Certos cuidados precisam ser tomados para a correta utilização dos dados de pesquisas amostrais complexas. A particularidade destes dados é proveniente de características de planos amostrais de pesquisas de populações finitas que envolvem: probabilidades distintas de seleção, estratificação e conglomeração das unidades, ajustes para compensar não resposta e outros ajustes (Heeringa, 2010) .

A atribuição de um peso amostral distinto para cada unidade amostral selecionada segundo um plano amostral complexo tem um papel indispensável na construção dos estimadores pontuais para usos descritivo e analítico.

No contexto das populações finitas torna-se necessário definir alguma notação básica bem como certos parâmetros populacionais de interesse que serão usados no desenvolvimento das seções seguintes.

Seja $U = \{1, 2, ..., k, ..., N\}$ uma população finita com N elementos.

Considere x e y variáveis de pesquisa definidas sobre os elementos de U.

Sejam x_k e y_k os valores das variáveis de interesse x e y, respectivamente, assumidos pelo *k-ésimo* elemento da população U.

Para cada unidade da população U associa-se um valor para cada variável de interesse x e y, e desta forma, os vetores correspondentes são denotados por $\mathbf{x} = (x_1, x_2, ..., x_N)$ e $\mathbf{y} = (y_1, y_2, ..., y_N)$.

Considerando estas variáveis de pesquisa, os seguintes parâmetros populacionais podem ser definidos:

i. Total populacional da variável x é definido por $X = \sum_{k \in U} x_k$;

ii. Média populacional da variável x é dada por $\overline{X} = \frac{\sum_{k \in U} x_k}{N}$;

iii. Quando a variável z (indicadora) assume dois valores: $z_k = 1$ ou $z_k = 0$, $k \in U$; a proporção de elementos na população U que possuem a característica de interesse $z_k = 1$ é definida por $P = \frac{\sum_{k \in U} z_k}{N}$.

iv. Razão entre os totais populacionais das variáveis y e x: $R = \frac{\sum_{k \in U} y_k}{\sum_{k \in U} x_k}$.

De modo geral, cada parâmetro populacional θ exemplificado acima pode ser representado por uma dada função f de valores de uma ou mais variáveis de pesquisa nas unidades da população U. Em outras palavras, $\theta = f(x_1, x_2, ..., x_n)$.

Como foi feita a caracterização das variáveis de pesquisa assim como de alguns parâmetros populacionais na população U torna-se necessário a caracterização também numa *amostra probabilística* dessa população que posteriormente será usada para estimar os parâmetros populacionais de interesse.

Seja $s \subset U$ uma amostra selecionada da população U sob um *plano amostral probabilístico A*. Em notação matemática, $s = \{1, 2, ..., k\}$, tal que $k \in U$.

Para cada unidade da amostra s está associado um valor da variável x e, desse modo, o vetor correspondente é representado por $\mathbf{x}_s = (x_1, x_2, ..., x_k)$, ou ainda, $\mathbf{x}_s = (x_k, k \in s)$.

Portanto, podemos estabelecer os estimadores correspondentes aos parâmetros populacionais: X, $\overline{X}$, P e R:

i. Total (*Horwitz-Thompson*) da variável x: $\hat{X}_{HT} = \sum_{k \in s} \frac{1}{\pi_k} x_k$, com π_k é a probabilidade de inclusão da unidade k na amostra s.

ii. Total da variável y: $\hat{Y} = \sum_{k \in s} w_k y_k$.

iii. Média da variável x: $\bar{x} = \frac{\sum_{k \in s} w_k x_k}{\sum_{k \in s} w_k}$.

iv. Proporção: $\hat{P} = \frac{\sum_{k \in s} w_k z_k}{\sum_{k \in s} w_k}$.

v. Razão entre os totais populacionais das variáveis y e x: $\hat{R} = \frac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k x_k}$.

Convém destacar que w_k é o peso da unidade k na amostra s e pode ser definido de modo diferente do que foi ilustrado acima no *estimador de Horwitz-Thompson* para o total. Isto pode ocorrer em virtude da correção de não-resposta, por exemplo. Em conseqüência deste fato, os estimadores de total e de variância não são únicos.

De modo geral, cada estimador do parâmetro populacional θ exemplificado acima pode ser representado por uma dada função f de valores de uma ou mais variáveis de pesquisa observadas nas unidades da amostra s, além dos pesos amostrais. Em outras palavras, $\hat{\theta} = f\left(x_1, x_2, ..., x_k ; w_1, w_2, ..., w_k\right),\ k \in s$.

4.3 Estimação da variância

A estimação da variância em pesquisas amostrais complexas torna-se imprescindível, pois permite avaliar a qualidade (*precisão*) das estimativas pontuais dos parâmetros de interesse, além da construção do arcabouço da inferência estatística.

O tipo do estimador (linear ou não linear) e as características dos planos amostrais, a saber: estratificação, conglomeração em vários estágios, métodos distintos de seleção das *unidades primárias de amostragem* (*PPT* ou seleção sistemática), ponderação diferenciada para as unidades selecionadas, interferem diretamente no processo de estimação da variância.

Entre os métodos mais usuais para a estimação da variância convém citar o *método do Conglomerado Primário* (Hansen, Hurvitz e Madow, 1953), os métodos de replicação (reamostragem): *Jackknife* e *Bootstrap* (Shao e Tu, 1995) e *Linearização em série de Taylor* (Casella e Berger, 2002).

Quando o estimador $\hat{\theta}$ é não linear, geralmente não existe uma expressão matemática exata para sua variância $V_A(\hat{\theta})$. No entanto, se este estimador $\hat{\theta}$ puder ser escrito como uma função não linear de estimadores de totais populacionais torna-se possível a construção da expressão analítica da variância aproximada $V_{A,L}(\hat{\theta})$, através da *Linearização em série de Taylor*.

De modo geral, em planos amostrais conglomerados, na estimação da variância de estimadores não lineares (exceto para percentis de uma distribuição), é comum a combinação da *Linearização em série de Taylor* com o *método do Conglomerado Primário*.

4.3.1 Método do Conglomerado Primário

O termo *Conglomerado Primário* proposto por Hansen, Hurvitz e Madow (1953) é utilizado para representar o conjunto de unidades incluídas na amostra contidas em uma mesma *unidade primária de amostragem* (*UPA*) (Lila, 2004).

Através deste método, em planos amostrais complexos com vários estágios de seleção, na estimação da variância de um estimador de total populacional, é considerada apenas a variação entre as estimativas de total no nível das unidades primárias de amostragem, sob a hipótese de que os conglomerados primários são selecionados com reposição da população de *UPA*s.

Convém ressaltar que o *método do Conglomerado Primário* combinado com a *Linearização em série de Taylor* permite a estimação da variância de estimadores não lineares, desde que estes possam ser escritos como função de estimadores não viciados de totais populacionais.

Na prática, mesmo que a seleção das unidades primárias de amostragem seja feita sem reposição é comum utilizar o estimador da variância do estimador de total obtido pelo *método do Conglomerado Primário quando* a *fração amostral* do primeiro estágio é inferior a 5%. Esta é uma prática conservadora na estimação da variância diante de planos amostrais complexos com vários estágios. Desta maneira, obtêm-se estimativas para uma cota superior de variância sob o plano amostral adotado pelas pesquisas amostrais complexas (Lila, 2004). Quando a *fração amostral* (f_a) (definida como a razão entre os tamanhos da amostra *s* e da população *U*), no primeiro estágio é superior a 5%, são sugeridas algumas

correções, dentre elas a *correção de população finita* ($1 - f_a$) na estimativa da variância.

Este método de aproximação empregado no processo de estimação de variâncias é largamente utilizado por praticantes de amostragem devido a sua simplicidade e praticidade, quando comparado com a complexidade envolvida no cálculo das diversas componentes de variância devidas a cada estágio de amostragem num plano complexo (Lila, 2004).

A popularidade do *método do Conglomerado Primário* está relacionada ao fato desta técnica fornecer a base metodológica de vários pacotes estatísticos especializados para estimação de variâncias, tais como SUDAAN, STATA, SPSS, SAS, R, entre outros.

A seguir é apresentada a expressão do estimador de total populacional e de sua variância baseados no *método do Conglomerado Primário* (Lila, 2004).

Considere um plano amostral conglomerado, em múltiplos estágios, no qual no 1° estágio existe estratificação com $n_h \geq 2$ *unidades primárias* selecionadas do estrato h, h=1,..., L. Suponha que no 1° estágio as *unidades primárias* sejam selecionadas com reposição e com probabilidade proporcional a uma medida de tamanho (*PPT*) em cada estrato h.

Seja p_{hl} a probabilidade de inclusão da *UPA l* do estrato h num sorteio, com $\sum_h p_{hl} = 1$. Seja $\hat{Y}_{hl}$ um estimador não viciado do total Y_{hl} da variável de pesquisa y na *l-ésima UPA* do estrato h, h=1,..., L.

Um estimador não viciado do total populacional Y, conforme Hansen et al (1953) é dado por:

$$\hat{Y}_{CP} = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{l=1}^{n_h} \frac{\hat{Y}_{hl}}{p_{hl}} \tag{4.1}$$

Um estimador não viciado da variância do estimador $\hat{Y}_{CP}$, de acordo com Hansen et al (1953) é definido por:

$$\hat{V}_{CP}\left(\hat{Y}_{CP}\right) = \sum_{h=1}^{L} \frac{1}{n_h\left(n_h - 1\right)} \sum_{l=1}^{n_h} \left(\frac{\hat{Y}_{hl}}{p_{hl}} - \hat{Y}_h \right)^2 \tag{4.2}$$

onde

$$\hat{Y}_h = \frac{1}{n_h} \sum_{l=1}^{n_h} \frac{\hat{Y}_{hl}}{p_{hl}}$$

4.3.2 Linearização de Taylor

É um método alicerçado na expansão em *série de Taylor* com o intuito de fornecer estimadores linearizados (aproximados) de variância para estimadores não lineares $\hat{\theta}$.

É importante destacar que o estimador $\hat{\theta}$ é um estimador de um parâmetro populacional θ que pode ser definido por uma função não linear de totais populacionais $\hat{X}$ e $\hat{Y}$.

Para que uma dada função *f* seja aproximada localmente no ponto (x_0, y_0), através de um polinômio $\mathbf{P}_n$ de ordem *n* (*série de Taylor*) é necessário que as derivadas parciais de ordem *n* da função *f* no ponto (x_0, y_0) existam (Casella e Berger, 2002).

Desta forma, o polinômio $\mathbf{P}_n$ é definido pela seguinte expressão:

$$\mathbf{P}_n(x,y) = f(x_0,y_0) + \frac{\partial f(x,y)}{\partial x}\Big|_{(x_0,y_0)}(x - x_0) + \frac{\partial f(x,y)}{\partial y}\Big|_{(x_0,y_0)}(y - y_0) + D^n(x_0,y_0), \quad (4.3)$$

onde

$$D^n(x_0, y_0) = \frac{1}{n!} f_x^n(x_0, y_0)(x - x_0)^n + \frac{1}{n!} f_y^n(x_0, y_0)(y - y_0)^n,$$

$$f_x^n(x_0, y_0) = \frac{\partial f^n(x,y)}{\partial x}\Big|_{(x_0,y_0)},$$

$$f_y^n(x_0, y_0) = \frac{\partial f^n(x,y)}{\partial y}\Big|_{(x_0,y_0)}.$$

Assim, o polinômio $\mathbf{P}_n$ tem em comum com a função *f,* além do valor no ponto (x_0, y_0), também o valor das derivadas parciais de ordem *n* da função *f* nesse ponto.

Portanto, um polinômio de aproximação para a função *f* no ponto (x_0, y_0) é dada por um *polinômio de Taylor* $\mathbf{P}_1(.)$ de primeiro grau, obtido de (4.3) :

$$\mathbf{P}_1(x,y) \cong f(x_0,y_0) + \frac{\partial f(x,y)}{\partial x}\Big|_{(x_0,y_0)}(x - x_0) + \frac{\partial f(x,y)}{\partial y}\Big|_{(x_0,y_0)}(y - y_0) \quad (4.4)$$

Em particular, o resíduo ou a qualidade da aproximação (em valor absoluto) do polinômio $\mathbf{P}_1(.)$ no ponto (x_0, y_0) é definido por $R(x_0, y_0) = | f(x_0, y_0) - \mathbf{P}_1(x_0, y_0)|$. A qualidade da aproximação no ponto (x_0, y_0) melhora à medida que

são incluídas na expressão (4.4) os termos com as derivadas parciais de ordem $n>1$ da função f nesse ponto.

Quando um estimador não linear $\hat{\theta}$ pode ser escrito por uma dada função de estimadores de totais populacionais $\hat{X}$ e $\hat{Y}$, ou seja, $\hat{\theta} = f(\hat{X}, \hat{Y})$ e, além disso, existem as derivadas parciais de primeira ordem de f em $\boldsymbol{p}=\left(E(\hat{X}),E(\hat{Y})\right)$, é possível obter a expressão matemática de um estimador linearizado $\hat{\theta}_L$ mediante um *polinômio de Taylor* de primeiro grau para a função f no ponto $\boldsymbol{p}$, conforme (4.4):

$$\hat{\theta}_L = f(\mathbf{p}) + f_{\hat{X}}(\mathbf{p})\left(\hat{X} - E\left(\hat{X}\right)\right) + f_{\hat{Y}}(\mathbf{p})\left(\hat{Y} - E\left(\hat{Y}\right)\right) \tag{4.5}$$

A variância do estimador $\hat{\theta}$ ($V_A(\hat{\theta})$), considerando um *plano amostral probabilístico A*, é aproximada pela variância do estimador $\hat{\theta}_L$, como veremos a seguir.

Pela definição de variância, temos:

$$V_A\left(\hat{\theta}_L\right) = E_A\left(\hat{\theta}_L - E_A\left(\hat{\theta}_L\right)\right)^2 \tag{4.6}$$

Mediante propriedades de variáveis aleatórias aplicadas na expressão (4.6), o estimador de variância do estimador linearizado $\hat{\theta}_L$ é escrito como combinação linear de variância e covariâncias de estimadores de totais $\hat{X}$ e $\hat{Y}$ e sua expressão é dada por:

$$V_A\left(\hat{\theta}_L\right) = a_1^2 V_A(\hat{X}) + a_2^2 V_A(\hat{Y}) + 2\,a_1\,a_2\,Cov_A(\hat{X},\hat{Y}) \tag{4.7}$$

onde

$$a_1 = \frac{\partial f}{\partial \hat{X}}\left(E(\hat{X}),E(\hat{Y})\right) \text{ e } \quad a_2 = \frac{\partial f}{\partial \hat{Y}}\left(E(\hat{X}),E(\hat{Y})\right)$$

A expressão (4.7) pode ser escrita em notação matricial:

$$V_A(\hat{\theta}_L) = [a_1\ a_2]\begin{pmatrix} V_A(\hat{X}) & Cov_A(\hat{X},\hat{Y}) \\ Cov_A(\hat{X},\hat{Y}) & V_A(\hat{Y}) \end{pmatrix}[a_1\ a_2]^T$$

A notação usada para a variância linearizada do estimador $\hat{\theta}$, considerando um *plano amostral probabilístico A*, de acordo com a expressão (4.7) é $V_{A,L}(\hat{\theta})$.

Para amostras suficientemente grandes, $V_A(\hat{\theta}) \cong V_A(\hat{\theta}_L)$.

Para a estimação da variância linearizada ($V_{A,L}(\hat{\theta})$), sob um *plano amostral A*, será necessário a estimação das variâncias e das covariâncias, além de cada

derivada parcial. O estimador da variância linearizada é denotado como $v_{A,L}(\hat{\theta}) = \hat{V}_{A,L}(\hat{\theta})$.

Convém destacar que uma das vantagens da aplicação do método de *Linearização de Taylor* é quando as derivadas parciais são conhecidas. Assim, a linearização (aproximação) quase sempre fornece uma estimativa de variância que pode ser aplicada em todos os tipos de planos amostrais.

Obter expressões analíticas para as derivadas parciais pode ser muito trabalhoso (Skinner, Holt e Smith, 1989), a presença de viés em decorrência do número pequeno de *UPA*s na amostra (Korn e Graubard, 1995) e a aplicação do método a funções complexas são algumas desvantagens da *Linearização de Taylor* como método de estimação de variâncias.

4.3.3 Método Jackknife

O método de *Linearização de Taylor* exige o cálculo de derivadas parciais da função que define o estimador $\hat{\theta}$. Por outro lado, os procedimentos de replicação (*Jackknife*, *Bootstrap*, *Balanced Repeated Replication* (BRR), entre outros) empregam uma abordagem mais simples para a estimação de variâncias de estimadores não lineares e não exigem cálculos de derivadas parciais (Shao e Tu, 1995).

Os métodos de replicação trocam formulações teóricas complicadas exigidas na aplicação dos métodos tradicionais pela inferência baseada em selecionar repetidas réplicas a partir da amostra original (Shao e Tu, 1995).

O *método Jackknife* é um dos procedimentos computacionalmente intensivos de replicação mais utilizados atualmente para estimar variância, em particular, em amostras complexas com estratificação e múltiplos estágios, com grande número de observações (Lee, 1973; Jones, 1974; Kish e Frankel, 1974; Krewski e Rao, 1981; Shao e Tu, 1995).

Quenouille (1949) introduziu a idéia original deste método para estimar o viés de um estimador $\hat{\theta}$. Esta consistia em excluir uma observação a cada vez dos dados originais e recalcular o estimador $\hat{\theta}$ baseado nos dados remanescentes. Mais tarde, Tukey (1958) denominou este procedimento como método de

Jackknife e descobriu que este poderia ser usado para a construção de estimadores de variância do estimador $\hat{\theta}$ (Efron e Tibshirani, 1993; Shao e Tu, 1995).

Para a estimação da variância sob planos amostrais complexos com vários estágios de seleção, Shao e Tu (1995, p.238) e Yee et al (1999) sugerem uma versão do *método de Jackknife* denominado de *Delete - 1 Jackknife* que está disponível no pacote estatístico *R* desde a versão 2.13 na library *survey* sob a função *svrepdesign*.

Para ilustrar este procedimento considere uma amostra s selecionada segundo um *plano amostral A* com conglomeração em vários estágios e com n_h *UPA*s sorteadas em cada estrato h, onde $h = 1, 2,.., L$.

Para a construção da réplica r, a partir da amostra original s, uma unidade primária l (*UPA*) é excluída do estrato $h=h^*$, e nos demais estratos todas as *unidades primárias de amostragem* selecionadas são mantidas, com $h= 1, 2, 3,...,$ $h^* -1$, $h^* +1,..., L$ (Figura 3.1).

Assim, após cada *UPA* ser excluída em cada estrato h, será gerado o total de $\sum_{h=1}^{L} n_h$ réplicas da amostra original s (Shao e Tu, 1995).

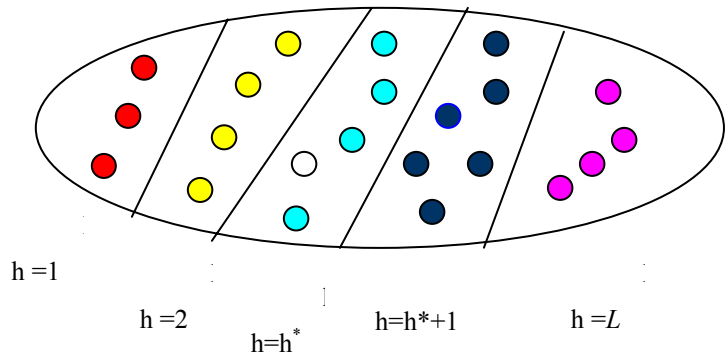

Figura 4.1: Procedimento *Jackknife* para construção da réplica r da amostra s

Finalmente, com o total de $\sum_{h=1}^{L} n_h$ estimativas do parâmetro de interesse θ, é possível obter a variância dessas estimativas que constitui uma estimativa da variância do estimador $\hat{\theta}$ obtida pelo método *Delete-1 Jackknife*:

$$v_{jackk}(\hat{\theta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{l=1}^{n_h} (\hat{\theta}_{(hl)} - \hat{\theta}_h)^2 \qquad (4.8)$$

onde

n_h : total de UPAs no estrato *h*, com *h=1,...,L*;

$\hat{\theta}_h = \frac{1}{n_h}\sum_{l=1}^{n_h}\hat{\theta}_{(hl)}$ e $\hat{\theta}_{(hl)}$: estimador de θ excluindo os elementos da unidade *l* no estrato *h*.

Convém ressaltar que no processo de estimação da variância do estimador $\hat{\theta}$ pelo método *Delete-1 Jackknife,* os pesos amostrais devem ser recalculados (reponderados) para cada unidade em cada réplica *r* da amostra *s*, onde $r \in \left\{1, 2, ..., \sum_{h=1}^{L} n_h\right\}$.

O motivo para a reponderação dos pesos consiste em manter a ″soma dos pesos em cada estrato *h* na réplica *r* ″igual à ″soma dos pesos em cada estrato *h* na amostra *s* original″. Assim, com a exclusão da unidade *l* (*UPA*) no estrato *h*=h*, o peso final no primeiro estágio (w^*_{ht}) para cada *UPA* na réplica *r* é calculado pelo produto do peso amostral original do primeiro estágio (w_{ht}) por um fator de correção adequado (Yee et al., 1999), como veremos a seguir:

$$w^*_{ht} = \begin{cases} 0\,; t = l \;\; e \;\; h = h^* \\ w_{ht} \cdot \frac{n_h}{n_h - 1}\,; t \neq l \;\; e \;\; h = h^* \\ w_{ht}\,; h = 2, 3, \ldots, L \end{cases}$$

Como existe mais de um estágio de seleção na amostra *s,* a correção atribuída ao peso no primeiro estágio é extendida aos pesos nos demais estágios.

4.3.3.1 Consistência do estimador de Jackknife da variância

Se o estimador $\hat{\theta}$ é definido por uma dada função *g* diferenciável de estimadores de totais populacionais então $v_{jackk}(\hat{\theta})$ é estimador consistente da variância $V_A(\hat{\theta})$ (Shao e Tu, § 6.4.2, p. 260, 1995). Além disso, a consistência do estimador $v_{jackk}(\hat{\theta})$ exige a continuidade da derivada de *g* (Shao e Tu, 1995).

Enfim, a qualidade da estimativa de variância obtida pelo estimador $v_{jackk}(\hat{\theta})$ depende da ″suavidade (*smoothness*) do estimador $\hat{\theta}$″, que pode ser

caracterizada pela diferenciabilidade da função *g* (Shao e Tu, 1995; Efron e Tibshirani, 1993).

4.3.3.2 As relações entre Jackknife e a Linearização de Taylor

O método *Delete-1 Jackknife* é sugerido para a estimação da variância de estimadores $\hat{\theta}$ que são funções diferenciáveis da classe C^1 de estimadores de totais populacionais. Além disso, os resultados assintóticos (Shao e Tu, 1995) são equivalentes aos resultados obtidos pelo *método do Conglomerado Primário* combinado com a *Linearização de Taylor*.

Em particular, se o estimador $\hat{\theta}$ é um estimador de total, média ou de proporção então $v_{jackk}(\hat{\theta})$ coincide com $v_{A,L}(\hat{\theta})$ (Shao e Tu, p. 240, 1995).

Na prática, é recomendado para o analista avaliar as estimativas da variância do estimador $\hat{\theta}$ obtidas pelos métodos: *Conglomerado Primário* e *Delete-1 Jackknife* para posteriormente escolher a mais adequada de acordo com o tamanho da amostra, a precisão desejada, o tipo de *plano amostral* e a presença de não resposta (Yee et al, 1999).

4.4 Estimadores dos coeficientes de escalonabilidade

Nesta seção são apresentados os estimadores pontuais para os coeficientes H_{ij}, H_i e H e seus respectivos estimadores da variância, no contexto da amostragem complexa de populações finitas.

4.4.1. Estimadores pontuais para H_{ij}, H_i e H

A atribuição do *peso amostral* para cada unidade que compõe uma amostra selecionada segundo um *plano amostral* complexo, tem papel fundamental também na construção dos estimadores pontuais para os coeficientes de escalonabilidade.

4.4.1.1 Estimador do coeficiente de escalonabilidade - $\hat{H}_{ij\,w}$

No contexto dos modelos da TRIN, é de interesse estimar parâmetros referentes a um conjunto de *J* itens rotulados como $\{1, 2, ..., i, ..., J\}$ levando em

consideração a ordenação não decrescente de dificuldade. Isto significa que os itens estão ordenados em termos da sua *popularidade* (proporção de acertos) da seguinte forma: da menor a maior popularidade, ou seja, do mais difícil ao mais fácil. Em notação matemática: $P_1 \leq P_2 \leq ... \leq P_J$.

Em particular, a notação $i < j$ denota que o item i é mais difícil que o item j, ou seja, $P_i < P_j$.

Sejam os itens i e j tais que: $i = 1, 2, ..., J$, $j = 2,3, ..., J$.

As variáveis dicotômicas $x(i)$ e $x(i,j)$ assumem o valor unitário quando o elemento $k \in U$ acertar a resposta do item i ($x_k(i)$=1) e acertar simultaneamente os itens i e j ($x_k(i,j)$=1), respectivamente.

Considerando as variáveis de interesse $x(i)$ e $x(i,j)$, podemos estabelecer os seguintes parâmetros populacionais:

1. O total de unidades na população U que acertam o item i:

$$X_i = \sum_{k \in U} x_k(i) \tag{4.9}$$

2. O total de unidades na população U que acertam simultaneamente os itens i e j:

$$X_{ij} = \sum_{k \in U} x_k(i,j) \tag{4.10}$$

3. A proporção de acertos no item i e a proporção de acertos simultâneos nos itens i e j podem ser definidas, respectivamente:

$$P_i = \frac{X_i}{N} \text{ e } P_{ij} = \frac{X_{ij}}{N} . \tag{4.11}$$

No contexto da amostragem de populações finitas, a construção dos estimadores da proporção de acertos do item i e da proporção de acertos simultâneos nos itens i e j é baseada na *Modelagem de Superpopulação* como será visto a seguir.

Considere os *Modelos de Superpopulação*:

Modelo 1: $x_1(i), x_2(i), x_3(i), .., x_N(i)$ observações de variáveis aleatórias IID com distribuição Ber (P_i). (4.12)

Modelo 2: $x_1(i,j), x_2(i,j), x_3(i,j), ..., x_N(i,j)$ observações de variáveis aleatórias IID com distribuição Ber (P_{ij}).

Supondo que todas as unidades responderam aos itens e aplicando o método de *Máxima Verossimilhança* (MV) aos valores observados "populacionais" para

estimar os parâmetros populacionais P_i e P_{ij}, respectivamente. Desta forma, os estimadores podem ser escritos como:

$$P_{i_U} = \frac{X_i}{N} \text{ e } P_{ij_U} = \frac{X_{ij}}{N}. \tag{4.13}$$

Sob os modelos especificados acima, estes estimadores de Máxima Verossimilhança são não viciados para os parâmetros populacionais P_i e P_{ij}, respectivamente.

Mas, o objetivo é a estimação destes "pseudo-parâmetros" (4.13) com base numa amostra complexa. Desta forma, considere $s \subset U$ uma amostra selecionada da população U sob um *plano amostral A*.

Os totais populacionais X_i e X_{ij} podem ser estimados com base em somas ponderadas das observações da amostra s. Desta forma, decorrem das expressões (4.9), (4.10) e (4.13) que:

$$\hat{P}_{i_w} = \frac{\hat{X}_i}{\hat{N}} = \frac{\sum_{k \in s} w_k x_k(i)}{\sum_{k \in s} w_k}. \tag{4.14}$$

$$\hat{P}_{ij_W} = \frac{\hat{X}_{ij}}{\hat{N}} = \frac{\sum_{k \in U} w_k \; x_k(i,j)}{\sum_{k \in s} w_k}. \tag{4.15}$$

onde $i = 1, 2, ..., J$, $j = 2,3, ..., J$ e w_k é o peso da unidade k na amostra s.

Assim, os estimadores pontuais $\hat{P}_{i_w}$ e $\hat{P}_{ij_w}$ são os estimadores de *Máxima Pseudo Verossimilhança* (Skinner, Holt e Smith, 1989) das proporções de acertos do item i e de acertos simultâneos aos itens i e j, respectivamente, levando em consideração o *peso amostral* estabelecido no *plano amostral A*.

Como foi estabelecido no Capítulo 3, o coeficiente de escalonabilidade H_{ij} pode ser escrito como uma função de proporções populacionais. Deste modo, empregando as expressões (4.14) e (4.15), é proposto o seguinte estimador pontual deste *coeficiente* sob o *plano amostral A*:

$$\hat{H}_{ij_W} = \frac{\hat{P}_{ij_W} - \hat{P}_{i_W}\hat{P}_{j_W}}{\hat{P}_{i\mathrm{w}} - \hat{P}_{i_W}\hat{P}_{j_W}}, \; \hat{P}_{i_W} < \hat{P}_{j_W}. \tag{4.16}$$

Além disto, o estimador $\hat{H}_{ij_W}$ pode ser definido por uma dada função f de estimadores de totais populacionais, sob o *plano amostral A*:

$$\hat{H}_{ij_W} = f(\hat{X}_{ij}, \hat{X}_i, \hat{X}_j, \hat{N}) = \frac{\hat{N}\hat{X}_{ij} - \hat{X}_i.\hat{X}_j}{\hat{N}\hat{X}_i - \hat{X}_i.\hat{X}_j},\ i < j \qquad (4.17)$$

3.4.1.2 Estimador do coeficiente de escalonabilidade - $\hat{H}_{i_w}$

Sob o *plano amostral A*, o estimador pontual $\hat{H}_{i_w}$ pode ser escrito por uma função *g* de estimadores de proporções de acordo com (4.14) e (4.15):

$$\hat{H}_{i_W} = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{J} \left(\hat{P}_{ij_W} - \hat{P}_{i_W}\hat{P}_{j_W}\right)}{\sum_{\substack{j=1 \\ j<i}}^{J} \hat{P}_{j_W}(1-\hat{P}_{i_W}) + \sum_{\substack{j=1 \\ i<j}}^{J} \hat{P}_{j_W}(1-\hat{P}_{i_W})}. \qquad (4.18)$$

Para $i = 1, 2, ..., J$, $j = 2,3, ..., J$, o estimador $\hat{H}_{i_w}$ pode ser definido também por uma dada função *g* de estimadores de totais populacionais sob o *plano amostral A*:

$$\hat{H}_{i_W} = g\left(\hat{X}_1, \hat{X}_2, ..., \hat{X}_i, ..., \hat{X}_J, \hat{X}_{i1}, ..., \hat{X}_{iJ}, \hat{N}\right) = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{J} \left(N\hat{X}_{ij} - \hat{X}_i\hat{X}_j\right)}{\sum_{\substack{j=1 \\ j\neq i \\ j<i}}^{J} \left(N\hat{X}_j - \hat{X}_i\hat{X}_j\right) + \sum_{\substack{j=1 \\ j\neq i \\ i<j}}^{J} \left(N\hat{X}_j - \hat{X}_i\hat{X}_j\right)}. \qquad (4.19)$$

3.4.1.3 Estimador do coeficiente de escalonabilidade - $\hat{H}_w$

Sob o *plano amostral A*, empregando as expressões (4.14) e (4.15), o estimador $\hat{H}_w$ do *coeficiente de escalonabilidade H* pode ser definido como uma função de estimadores de proporções:

$$\hat{H}_W = \frac{\sum_{i=1}^{J-1}\sum_{j=i+1}^{J} (\hat{P}_{ij_W} - \hat{P}_{i_W}\hat{P}_{j_W})}{\sum_{i=1}^{J-1}\sum_{j=i+1}^{J} \hat{P}_{i_W}(1-\hat{P}_{j_W})} \qquad (4.20)$$

O estimador $\hat{H}_w$ pode ser definido por uma dada função *u* de estimadores de totais populacionais sob o *plano amostral A* conforme expressões (4.14) e (4.15):

$$\hat{H}_W = u\left(\hat{X}_1,..,\hat{X}_J,\hat{X}_{i1},...,\hat{X}_{iJ},...,\hat{X}_{J-1,J},\hat{N}\right) = \frac{\sum_{\substack{j=1 \\ j\neq i}}^{J}\left(\hat{N}\hat{X}_{ij} - \hat{X}_i\hat{X}_j\right)}{\sum_{i=1}^{J-1}\sum_{\substack{j=1 \\ j<i}}^{J}\left(\hat{N}\hat{X}_j - \hat{X}_i\hat{X}_j\right) + \sum_{i=1}^{J-1}\sum_{\substack{j=1 \\ i<j}}^{J}\left(\hat{N}\hat{X}_j - \hat{X}_i\hat{X}_j\right)} \quad (4.21)$$

Convém ressaltar que em cada parcela do somatório no denominador do estimador $\hat{H}_w$ em (4.20), deve ser avaliada a relação de ordem das *popularidades* entre cada par de itens (i, j).

4.4.2. Estimadores da variância dos estimadores de H_{ij}, H_i e H

A estimação da variância dos estimadores: $\hat{H}_{ij_W}$, $\hat{H}_{i_W}$ e $\hat{H}_w$ será aqui desenvolvida pelos métodos *Linearização de Taylor* e *Delete - 1 Jackknife* de acordo Särndal et al (1992) e Shao e Tu (1995), respectivamente.

A escolha destes dois métodos é justificada em decorrência da impossibilidade de se obter expressões exatas para a variância dos estimadores: $\hat{H}_{ij_W}$, $\hat{H}_{i_W}$ e $\hat{H}_w$, pois estes são estimadores não lineares, embora possam ser escritos como funções de estimadores de totais.

4.4.2.1 Variância do estimador $\hat{H}_{ij_W}$

4.4.2.1.1. Linearização de Taylor

Considere a expressão (4.17) para o estimador $\hat{H}_{ij_W}$:

$$\hat{H}_{ij_W} = f(\hat{X}_{ij},\hat{X}_i,\hat{X}_j,\hat{N}) = \frac{\hat{N}\hat{X}_{ij} - \hat{X}_i.\hat{X}_j}{\hat{N}\hat{X}_i - \hat{X}_i.\hat{X}_j},\ i<j$$

Como a quantidade de interesse H_{ij} é estimada por uma função não linear *f* dos estimadores de totais $\hat{X}_{ij},\hat{X}_i,\hat{X}_j$ e $\hat{N}$, não existe uma expressão algébrica

exata para a variância do estimador $\hat{H}_{ij_W}$ (Särndal et al, 1992). No entanto, podemos obter uma aproximação da variância do estimador $\hat{H}_{ij_W}$ através da *Linearização em série de Taylor* de 1ª ordem da função *f* em torno do vetor $\mathbf{a} = (X_{ij}, X_i, X_j, N)$, conforme a expressão a seguir:

$$\hat{H}_{ij_W} - H_{ij} = f(\hat{X}_{ij}, \hat{X}_i, \hat{X}_j, \hat{N}) - f(X_{ij}, X_i, X_j, N)$$

$$\frac{\partial f(\mathbf{a})}{\partial \hat{X}_{ij}}(\hat{X}_{ij} - X_{ij}) + \frac{\partial f(\mathbf{a})}{\partial \hat{X}_i}(\hat{X}_i - X_i) + \frac{\partial f(\mathbf{a})}{\partial \hat{X}_j}(\hat{X}_j - X_j) + \frac{\partial f(\mathbf{a})}{\partial \hat{N}}(\hat{N} - N).$$

Seja $\Delta f(\boldsymbol{a})$ o vetor de derivadas parciais da função *f* em relação aos estimadores $\hat{X}_{ij}, \hat{X}_i, \hat{X}_j$ e $\hat{N}$ aplicadas no vetor $\boldsymbol{a}$:

$$\Delta f(\mathbf{a}) = \left[\frac{\partial f(\mathbf{a})}{\partial \hat{X}_{ij}}, \frac{\partial f(\mathbf{a})}{\partial \hat{X}_i}, \frac{\partial f(\mathbf{a})}{\partial \hat{X}_j}, \frac{\partial f(\mathbf{a})}{\partial \hat{N}} \right] \tag{4.22}$$

Seja $V_A(\hat{\mathbf{a}})$ a matriz de variâncias e covariâncias do vetor de estimadores de totais $\hat{\mathbf{a}}$, sob o *plano amostral A*:

$$V_A(\hat{\mathbf{a}}) = \begin{pmatrix} V_A(\hat{X}_{ij}) & Cov_A(\hat{X}_{ij}, \hat{X}_i) & Cov_A(\hat{X}_{ij}, \hat{X}_j) & Cov_A(\hat{X}_{ij}, \hat{N}) \\ Cov_A(\hat{X}_{ij}, \hat{X}_i) & V_A(\hat{X}_i) & Cov_A(\hat{X}_i, \hat{X}_j) & Cov_A(\hat{X}_i, \hat{N}) \\ Cov_A(\hat{X}_j, \hat{X}_{ij}) & Cov_A(\hat{X}_j, \hat{X}_i) & V_A(\hat{X}_j) & Cov_A(\hat{X}_j, \hat{N}) \\ Cov_A(\hat{N}, \hat{X}_{ij}) & Cov_A(\hat{N}, \hat{X}_i) & Cov_A(\hat{N}, \hat{X}_j) & V_A(\hat{N}) \end{pmatrix}$$

Assim, a expressão para a variância do estimador $\hat{H}_{ij_W}$ sob o *plano amostral A*, aproximada pela *Linearização de Taylor,* é dada por:

$$V_{A,L}(\hat{H}_{ij_W}) = \Delta f(\mathbf{a}) V_A(\hat{\mathbf{a}}) (\Delta f(\mathbf{a}))^T \tag{4.23}$$

Para amostras suficientemente grandes, sob o *plano amostral A*, tem-se:

$$V_A(\hat{H}_{ij_W}) = V_{A,L}(\hat{H}_{ij_W})$$

De acordo com a expressão (4.23) podemos estabelecer o estimador da variância do estimador $\hat{H}_{ij_W}$ sob o *plano amostral A*:

$$v_{A,L}(\hat{H}_{ij_W}) = \Delta f(\hat{\mathbf{a}}) \hat{V}_A(\hat{\mathbf{a}}) (\Delta f(\hat{\mathbf{a}}))^T \tag{4.24}$$

Convém ressaltar que na fórmula da variância linearizada do estimador $\hat{H}_{ij_W}$ existem as derivadas parciais da função f em termos de $\hat{X}_{ij}, \hat{X}_i, \hat{X}_j$ e $\hat{N}$ aplicadas no vetor **a**. A expressão matemática de cada derivada parcial no vetor $\Delta f(\mathbf{a})$, conforme (4.22), é dada por:

$$\frac{\partial f(\mathbf{a})}{\partial \hat{X}_{ij}} = \frac{N}{(NX_i - X_i X_j)}; \quad \frac{\partial f(\mathbf{a})}{\partial \hat{X}_i} = \frac{NX_{ij}X_j - N^2 X_{ij}}{(NX_i - X_i X_j)^2}$$

$$\frac{\partial f(\mathbf{a})}{\partial \hat{X}_j} = \frac{NX_{ij}X_i - N(X_i)^2}{(NX_i - X_i X_j)^2}; \quad \frac{\partial f(\mathbf{a})}{\partial \hat{N}} = \frac{(X_i)^2 X_j - X_{ij}X_i X_j}{(NX_i - X_i X_j)^2}$$

Para a análise do desempenho dos estimadores de variância dos estimadores dos coeficientes de escalonabilidade (ver Capítulos 7 e 8) será necessário o cálculo das variâncias populacionais linearizadas dos estimadores: $\hat{H}_{ij_w}$, $\hat{H}_{i_w}$ e $\hat{H}_w$ sob os planos amostrais *AC1S*, *AC1C* e *AASC*. Desta forma, tornam-se apropriadas definir, para estes planos amostrais, as expressões matemáticas da variância e covariância entre os dois estimadores de totais populacionais $\hat{X}$ e $\hat{Y}$.

No caso do *plano amostral AC1S*, onde *m unidades primárias de amostragem* são selecionadas por *AAS* sem reposição da população com *M* conglomerados, a variância e covariância entre os dois estimadores de totais populacionais $\hat{X}$ e $\hat{Y}$ são definidas conforme Bolfarine e Bussab (2005):

a. $$V_{AC1S}(\hat{X}) = M^2\left(\frac{1}{m} - \frac{1}{M}\right)S^2_{e_X}, \tag{4.25}$$

onde $S^2_{e_X} = \frac{1}{M-1}\sum_{c=1}^{M}(X_c - \overline{X})^2$ é a variância entre os totais dos conglomerados.

b. $$Cov_{AC1S}(\hat{X}, \hat{Y}) = M^2\left(\frac{1}{m} - \frac{1}{M}\right) S^2_{e_{XY}}, \tag{4.26}$$

onde $S^2_{e_{XY}} = \frac{1}{M-1}\sum_{c=1}^{M}(X_c - \overline{X})(Y_c - \overline{Y})$, $\overline{Y} = \frac{1}{M}\sum_{c=1}^{M} Y_c$ e $\overline{X} = \frac{1}{M}\sum_{c=1}^{M} X_c$.

Além disso, Y_c e X_c denotam os totais populacionais no conglomerado *c*.

Assim, a variância linearizada do estimador $\hat{H}_{ij_W}$, sob o *plano amostral AC1S*, definida através das expressões: (4.22), (4.23), (4.25) e (4.26); é denotada por $V_{AC1S,L}(\hat{H}_{ij_w})$.

No caso do *plano amostral AC1C*, onde *m* conglomerados são escolhidos por *AAS* com reposição da população com *M* conglomerados, a variância e

covariância entre os dois estimadores de totais populacionais $\hat{X}$ e $\hat{Y}$ são dadas por Bolfarine e Bussab (2005):

a. $V_{AC1C}(\hat{X}) = M^2 \frac{S^2_{eX}}{m}$ (4.27)

b. $Cov_{AC1C}(\hat{X},\hat{Y}) = M^2 \frac{S^2_{eXY}}{m}$ (4.28)

Logo, a variância linearizada do estimador $\hat{H}_{ij_W}$, sob o *plano amostral AC1C*, definida matematicamente pelas as expressões: (4.22), (4.23), (4.27) e (4.28); é denotada por $V_{AC1C\,,\,L}(\hat{H}_{ij_W})$.

Convém ressaltar o seguinte resultado para a razão $\frac{V_{AC1S,L}(\hat{H}_{ij_W})}{V_{AC1C,L}(\hat{H}_{ij_W})} = 1 - f_a$, onde f_a denota a *fração amostral* no primeiro estágio.

Para o *plano amostral AASC*, as *n* observações amostrais são selecionadas diretamente e com reposição da população com *N* unidades. A variância e covariância entre os dois estimadores de totais $\hat{X}$ e $\hat{Y}$, são estabelecidas conforme Bolfarine e Bussab (2005):

a. $V_{AASC}(\hat{X}) = N^2 \frac{S^2_X}{n}$, (4.29)

onde $S^2_X = \frac{1}{N-1}\sum_{l=1}^{N}(X_l - \bar{\bar{X}})^2$

b. $Cov_{AASC}(\hat{X},\hat{Y}) = N^2 \frac{1}{n} S^2_{XY}$ (4.30)

onde $S^2_{XY} = \frac{1}{N-1}\sum_{l=1}^{N}(X_l - \bar{\bar{X}})(Y_l - \bar{\bar{Y}})$, $\bar{\bar{Y}} = \frac{1}{N}\sum_{l=1}^{N} Y_l$ e $\bar{\bar{X}} = \frac{1}{N}\sum_{l=1}^{N} X_l$.

Assim, sob a hipótese de que as observações são independentes e identicamente distribuídas, a expressão para a variância linearizada $V_{AASC,L}(\hat{H}_{ij_W})$ é definida conforme (4.22), (4.23), (4.29) e (4.30) e dada por:

$$V_{AASC,L}(\hat{H}_{ij_W}) = \Delta f(\mathbf{a})\; V_{AASC}(\hat{\mathbf{a}})\; (\Delta f(\mathbf{a}))^T \quad (4.31)$$

Nesta tese, para cada *plano amostral A* adotado no Capítulo 5, são propostos estimadores de variância dos estimadores: $\hat{H}_{ij_W}$, $\hat{H}_{i_W}$ e $\hat{H}_W$ de acordo com os métodos: *Linearização de Taylor* e *Delete - 1 Jackknife*, conforme descrição a seguir.

Na estimação da variância do estimador $\hat{\theta} = \hat{H}_{ij_w}$ apresentamos os seguintes estimadores:

a) Estimador da variância do estimador $\hat{H}_{ij_w}$, sob o *plano amostral A*, obtido pela *Linearização de Taylor*, com a hipótese de que as unidades do primeiro estágio são selecionadas sem reposição: $v_2 = v_{A,L2}(\hat{\theta})$. (4.32)

b) Estimador da variância do estimador $\hat{H}_{ij_w}$, sob o *plano amostral A*, obtido pela *Linearização de Taylor*, supondo que as unidades primárias de amostragem são selecionadas com reposição: $v_1 = v_{A,L1}(\hat{\theta})$. Este estimador é obtido pelo *método do Conglomerado Primário*. (4.33)

c) Estimador da variância do estimador $\hat{H}_{ij_w}$, sob o *plano amostral A*, obtido pelo método *Delete -1 Jackknife* conforme expressão (4.8): $v_{jackk} = v_{jackk}(\hat{\theta})$. (4.34)

Cabe destacar que os estimadores de variância definidos acima são válidos quando $\hat{\theta} = \hat{H}_{i_w}$ ou $\hat{\theta} = \hat{H}_{w}$.

Além disso, é importante considerar na análise comparativa do desempenho dos estimadores de variância dos estimadores $\hat{H}_{ij_w}$, $\hat{H}_{i_w}$ e $\hat{H}_{w}$, o *estimador ingênuo* da variância ($v_0 = v_0(\hat{\theta})$). Este estimador não leva em consideração as características do *plano amostral A* na estimação da variância. (4.35)

4.4.2.2 Variância do estimador $\hat{H}_{i_w}$

Como a quantidade de interesse H_i é estimada por uma função não linear *g* dos estimadores de totais: $\hat{X}_1, \hat{X}_2, ..., \hat{X}_i, ..., \hat{X}_J, \hat{X}_{i1}, ..., \hat{X}_{iJ}$ e $\hat{N}$; não existe uma expressão algébrica exata para a variância de $\hat{H}_{i_w}$ (Särndal et al, 1992). No entanto, podemos obter uma aproximação da variância do estimador $\hat{H}_{i_w}$ através da *Linearização em série de Taylor* de 1ª ordem da função *g* em torno do vetor $\dot{\mathbf{b}} = (X_1, X_2, ..., X_i, ..., X_J, X_{i1}, ..., X_{iJ}, N)$, conforme a expressão a seguir:

$$\hat{H}_{i_w} - H_i =$$
$$= g\left(\hat{X}_1,\hat{X}_2,...,\hat{X}_i,...,\hat{X}_J,\hat{X}_{i1},...,\hat{X}_{iJ},\hat{N}\right) - g\left(\hat{X}_1,\hat{X}_2,...,\hat{X}_i,...,\hat{X}_J,\hat{X}_{i1},...,\hat{X}_{iJ},\hat{N}\right) \cong$$

$$\cong \frac{\partial g(\dot{\mathbf{b}})}{\partial \hat{X}_1}(\hat{X}_1 - X_1) + ... + \frac{\partial g(\dot{\mathbf{b}})}{\partial \hat{X}_J}(\hat{X}_J - X_J) + \frac{\partial g(\dot{\mathbf{b}})}{\partial \hat{X}_{i1}}(\hat{X}_{i1} - X_{i1}) + ... + \frac{\partial g(\dot{\mathbf{b}})}{\partial \hat{N}}(\hat{N} - N).$$

Seja $\Delta g(\dot{\mathbf{b}})$ o vetor de derivadas parciais da função g em relação aos estimadores $\hat{X}_1,\hat{X}_2,...,\hat{X}_i,...,\hat{X}_J,\hat{X}_{i1},...,\hat{X}_{iJ}$ e $\hat{N}$ aplicada no vetor $\dot{\mathbf{b}}$:

$$\Delta g(\dot{\mathbf{b}}) = \left[\frac{\partial g(\dot{b})}{\partial \hat{X}_1},..., \frac{\partial g(\dot{b})}{\partial \hat{X}_i},..., \frac{\partial g(\dot{b})}{\partial \hat{X}_J}, \frac{\partial g(\dot{b})}{\partial \hat{X}_{i1}},..., \frac{\partial g(\dot{b})}{\partial \hat{X}_{iJ}}, \frac{\partial g(\dot{b})}{\partial \hat{N}} \right] \quad (4.36)$$

A matriz de variâncias e covariâncias do vetor de estimadores de totais $\hat{\dot{\mathbf{b}}}$, sob o *plano amostral A* denotada por $V_A(\hat{\dot{\mathbf{b}}})$ é dada por:

$$V_A(\hat{\dot{\mathbf{b}}}) = \begin{pmatrix} V_A(\hat{X}_i) & Cov_A(\hat{X}_i,\hat{X}_{i+1}) & ... & Cov_A(\hat{X}_i,\hat{X}_J) & Cov_A(\hat{X}_i,\hat{X}_{ii+1}) & ... & Cov_A(\hat{X}_i,\hat{X}_{iJ}) & Cov_A(\hat{X}_i,\hat{N}) \\ & V_A(\hat{X}_{i+1}) & & & & & & \\ & & ... & & & & & \\ & & & V_A(\hat{X}_J) & & & & \\ & & & & V_A(\hat{X}_{ii+1}) & & & \\ & & ... & & & ... & & \\ & & & & & & V_A(\hat{X}_{iJ}) & \\ & & & & & & & V_A(\hat{N}) \end{pmatrix}$$

Assim, a expressão para a variância do estimador $\hat{H}_{i_w}$ sob o *plano amostral A*, aproximada pela *Linearização de Taylor* é dada por:

$$V_{A,L}(\hat{H}_{i_w}) = \Delta g(\dot{\mathbf{b}}) V_A(\hat{\dot{\mathbf{b}}}) (\Delta g(\dot{\mathbf{b}}))^T \quad (4.37)$$

Para amostras suficientemente grandes, sob o *plano amostral A*, tem-se:

$$V_A(\hat{H}_{i_W}) = V_{A,L}(\hat{H}_{i_W})$$

De acordo com a expressão (4.37) podemos estabelecer o estimador da variância do estimador $\hat{H}_{i_w}$ sob o *plano amostral A*:

$$v_{A,L}(\hat{H}_{iw}) = \Delta g(\hat{\dot{\mathbf{b}}})\hat{V}_A(\hat{\dot{\mathbf{b}}})\,(\Delta g(\hat{\dot{\mathbf{b}}}))^T \tag{4.38}$$

Convém ressaltar que na fórmula da variância linearizada do estimador $\hat{H}_{i_w}$, segundo a expressão (4.37), existem as derivadas parciais da função *g* em termos de $\hat{X}_1,\hat{X}_2,...,\hat{X}_i,...,\hat{X}_J,\hat{X}_{i1},...,\hat{X}_{iJ}$ e $\hat{N}$ aplicadas no vetor **b**. A expressão matemática de cada derivada parcial no vetor $\Delta g(\dot{\mathbf{b}})$, conforme (4.36), é dada por:

$$\frac{\partial g(\dot{\mathbf{b}})}{\partial \hat{X}_{ij}} = \frac{1}{N}\frac{1}{\sum_{i<j} P_i + \sum_{j<i} P_j - \sum_{j\neq i} P_i P_j} \tag{4.39}$$

$$\frac{\partial g(\dot{\mathbf{b}})}{\partial \hat{X}_i} = -\frac{1}{N}\frac{\sum_{j\neq i} P_j}{\sum_{i<j} P_i + \sum_{j<i} P_j - \sum_{j\neq i} P_i P_j} + \frac{1}{N}\left(\sum_{j\neq i} P_{ij} - \sum_{j\neq i} P_i P_j\right)\frac{-\sum_{i<j} 1 + 0 + \sum_{j\neq i} P_j}{\left(\sum_{i<j} P_i + \sum_{j<i} P_j - \sum_{j\neq i} P_i P_j\right)^2}$$

(4.40)

$$\frac{\partial g(\dot{\mathbf{b}})}{\partial \hat{X}_j} = -\frac{1}{N}\frac{P_i}{\left(\sum_{i<j} P_i + \sum_{j<i} P_j - \sum_{j\neq i} P_i P_j\right)} + \frac{1}{N}\left(\sum_{j\neq i} P_{ij} - \sum_{j\neq i} P_i P_j\right)\frac{\left(-1_{j<i} + P_i\right)}{\left(\sum_{i<j} P_i + \sum_{j<i} P_j - \sum_{j\neq i} P_i P_j\right)^2}$$

(4.41)

$$\frac{\partial g(\dot{\mathbf{b}})}{\partial \hat{N}} = -\frac{1}{N}\frac{\left(\sum_{j\neq i} P_{ij} - 2\sum_{j\neq i} P_i P_j\right)}{den} + \frac{1}{N}\left(\sum_{j\neq i} P_{ij} - \sum_{j\neq i} P_i P_j\right)\frac{\sum_{i<j} P_i + \sum_{j<i} P_j - 2\sum_{j\neq i} P_i P_j}{\left(\sum_{i<j} P_i + \sum_{j<i} P_j - \sum_{j\neq i} P_i P_j\right)^2} \tag{4.42}$$

4.4.2.3 Variância do estimador $\hat{H}_w$

Como a quantidade de interesse H é estimada por uma função não linear *u* dos estimadores de totais: $\hat{X}_1,\hat{X}_2,...,\hat{X}_i,...,\hat{X}_J,\hat{X}_{12},...,\hat{X}_{1J},...,\hat{X}_{ij},...,\hat{X}_{iJ},...,\hat{N}$, não existe uma expressão algébrica exata para a variância do estimador $\hat{H}_w$ (Särndal et al, 1992). No entanto, podemos obter uma aproximação da variância do estimador $\hat{H}_w$ através da *Linearização em série de Taylor* de 1ª ordem da função *u* em torno do vetor $\dot{\mathbf{c}} = (X_1,X_2,...,X_i,...,X_J,X_{12},...,X_{1J},...,X_{ij},...,X_{iJ},...,N)$, conforme expressão a seguir:

$$\hat{H}_w - H =$$
$$= u(\hat{X}_1,...,\hat{X}_J,\hat{X}_{12},...,\hat{X}_{1J},...,\hat{X}_{ij},...,\hat{X}_{iJ},...,\hat{N}) - u(X_1,...,X_J,X_{12},...,X_{1J},...,X_{ij},...,X_{iJ},...,N) \cong$$

$$\cong \frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{X}_1}(\hat{X}_1 - X_1) + ... + \frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{X}_J}(\hat{X}_J - X_J) + ... + \frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{X}_{ij}}(\hat{X}_{ij} - X_{ij}) + ... + \frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{N}}(\hat{N} - N).$$

Seja $\Delta u(\dot{\mathbf{c}})$ o vetor de derivadas parciais da função u em relação aos estimadores: $\hat{X}_1,\hat{X}_2,...,\hat{X}_i,...,\hat{X}_J,\hat{X}_{12},...,\hat{X}_{1J},...,\hat{X}_{ij},...,\hat{X}_{iJ},...,\hat{N}$ aplicada no vetor $\dot{\mathbf{c}}$:

$$\Delta u(\dot{\mathbf{c}}) = \left[\frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{X}_1},..., \frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{X}_J},...., \frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{X}_{ij}},..., \frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{N}} \right] \tag{4.43}$$

A seguir, é apresentada a matriz de variâncias e covariâncias do vetor de estimadores de totais $\hat{\mathbf{c}}$, sob o *plano amostral A*, com notação $V_A(\hat{\mathbf{c}})$:

$$V_A(\hat{\mathbf{c}}) = \begin{pmatrix} V_A(\hat{X}_1) & Cov_A(\hat{X}_1,\hat{X}_2) & ... & Cov_A(\hat{X}_1,\hat{X}_i) & Cov_A(\hat{X}_1,\hat{X}_J) & ... & Cov_A(\hat{X}_1,\hat{X}_{ij}) & Cov_A(\hat{X}_i,\hat{N}) \\ & V_A(\hat{X}_{i+1}) & & & & & & \\ & & ... & & & & & \\ & & & V_A(\hat{X}_J) & & & & \\ & & & & V_A(\hat{X}_{ii+1}) & & & \\ & & ... & & & ... & & \\ & & & & & & V_A(\hat{X}_{ij}) & \\ & & & & & & & V_A(\hat{N}) \end{pmatrix}$$

Assim, a expressão para a variância do estimador $\hat{H}_w$ sob o *plano amostral A*, aproximada pela *Linearização de Taylor,* é dada por:

$$V_{A,L}(\hat{H}_w) = \Delta u(\mathbf{c}) V_A(\hat{\mathbf{c}}) (\Delta u(\dot{\mathbf{c}}))^T \tag{4.44}$$

Para amostras suficientemente grandes, sob o *plano amostral A*, tem-se:

$$V_A(\hat{H}_w) = V_{A,L}(\hat{H}_w)$$

De acordo com a expressão (4.41) podemos estabelecer o estimador da variância do estimador $\hat{H}_w$ sob o *plano amostral A*:

$$v_{A,L}(\hat{H}_w) = \Delta u(\hat{\mathbf{c}}) V_A(\hat{\mathbf{c}}) (\Delta u(\hat{\mathbf{c}}))^T \tag{4.45}$$

Convém ressaltar que na fórmula da variância linearizada do estimador $\hat{H}_w$, segundo a expressão (4.40), existem as derivadas parciais da função *u* em termos dos estimadores: $\hat{X}_1,\hat{X}_2,...,\hat{X}_i,...,\hat{X}_J,\hat{X}_{12},...,\hat{X}_{1J},...,\hat{X}_{ij},...,\hat{X}_{iJ},...,\hat{N}$ aplicadas no vetor $\dot{\mathbf{c}}$. A expressão matemática de cada derivada parcial no vetor $\Delta u(\dot{\mathbf{c}})$, conforme (4.40), é dada por:

$$\frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{X}_{ij}} = \frac{1}{N}\frac{1}{denH} \tag{4.46}$$

$$\frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{X}_{i}} = -\frac{1}{N}\frac{\sum_{j\neq i} P_j}{denH} + \frac{1}{N}\frac{numH}{denH^2}.\left(-\sum_{i<j} 1 + \sum_{j\neq i} P_j\right) \tag{4.47}$$

$$\frac{\partial u(\dot{\mathbf{c}})}{\partial \hat{N}} = -\frac{1}{N}\frac{\left(\sum_{i=1}^{J-1}\sum_{j\neq i}^{J} P_{ij}\right) - 2\sum_{i=1}^{J-1}\sum_{j\neq i}^{J} P_i P_j}{denH} + \frac{1}{N}\frac{numH}{denH^2}\left(\sum_{i<j} P_i + \sum_{j<i} P_j - 2\sum_{j\neq i} P_i P_j\right) \tag{4.48}$$

onde

$$numH = \sum_{i}\sum_{j\neq i}^{J} P_{ij} - \sum_{i}\sum_{j\neq i}^{J} P_i P_j \text{ e } denH = \sum_{i;i<j}^{J-1} P_i - \sum_{i<j}^{J-1} (P_i P_j) + \sum_{j,j<i} P_j - \sum_{j<i}^{J} (P_i P_j).$$