Universidade de Brasília
IE – Departamento de Estatística

Estágio Supervisionado II

ANA LUIZA COELHO BARBOSA

ALTERNATIVAS PARA ANÁLISE DE DADOS NUTRICIONAIS

Brasília

Julho, 2014

ANA LUIZA COELHO BARBOSA

ALTERNATIVAS PARA ANÁLISE DE DADOS NUTRICIONAIS

Monografia apresentada ao Departamento de Estatística da Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: Prof. Lúcio José Vivaldi

Banca examinadora: Prof. Eduardo Freitas da Silva

Prof.ª Maria Teresa Leão Costa

Brasília

Julho, 2014

Dedico este trabalho às minhas maiores dádivas: meus pais e meu irmão.

AGRADECIMENTOS

Agradeço aos meus pais pelo amor, confiança e oportunidade que me deram de chegar até aqui.

Ao meu irmão pelo exemplo de ser humano e por ser a minha melhor companhia.

Ao meu orientador por todo o conhecimento transmitido durante a graduação e principalmente pela atenção, pela excelente orientação e pelo enorme conhecimento passado durante a realização deste trabalho.

Agradeço aos demais professores e funcionários da Universidade de Brasília pela sabedoria e pelo apoio passado.

À minha família por serem meu maior conforto e minha maior certeza.

Aos amigos que conquistei antes e durante a graduação. Amigos que entraram na minha vida por acaso e que, por razões que são sentidas, não vão deixá-la mais. Não poderia deixar de citar dois colegas de curso e, felizmente, amigos que me ajudaram de maneira imensurável durante minha graduação: Marcus Vinícius Fagundes e Roberto Lazarte. Obrigada por me fazerem acreditar no curso e por todo apoio que vocês me deram. Também quero agradecer pela amizade e companheirismo desde o início do curso de pessoas extremamente especiais para mim: Iran Barros, Fátima Lira, João Vitor Chieregatti, Bianca Agapito, Rayany de Oliveira e João Paulo Costa.

À professora Tereza Costa do departamento de nutrição da Universidade de Brasília por ter cedido os dados usados na aplicação deste trabalho.

Ao SAS Institute Brasil por possibilitar a utilização do software por meio de parceria acadêmica com o Departamento de Estatística da UnB.

À Deus por toda a luz.

RESUMO

Devido às mudanças significativas nos hábitos alimentares da população, diversos estudos utilizando dados nutricionais estão sendo desenvolvidos e aprimorados com o passar do tempo. O objetivo deste estudo é ter uma estimativa da média diária de nutrientes ingeridos pelos indivíduos de uma determinada população e, mediante as médias, estabelecer estimativas dos percentis de cada nutriente. Essa informação é útil para nutricionistas, pois, se o percentil encontrado estiver abaixo ou acima do considerado saudável para o organismo, é necessário tomar alguma medida para converter esse quadro.

A dificuldade para analisar dados nessa natureza é em decorrência das duas fontes de variação que estão naturalmente presente nas pessoas: variação entre os indivíduos e a variação entre as observações de cada indivíduo. Outro problema é que, na maioria dos casos, eles seguem uma distribuição assimétrica à direita, uma vez que poucos indivíduos ingerem muito e outros muitos indivíduos ingerem pouco de determinado nutriente.

Atualmente, o modelo de análise mais usado para este tipo de estudo é o modelo chamado S-Nusser proposto por Hoffmann et al.(2002) baseado na análise de variância do modelo aleatório, com os dados transformados pela transformação de Box e Cox(1964). O método contém um estimador tipo “shrinkage” que diminui a influência da variabilidade dentro do indivíduo.

A primeira alternativa à este método é aplicar o método de Máxima Verossimilhança Restrita aos dados transformados e utilizar as equações de Henderson (Henderson et al.(1959)) para se obter o BLUP (Best Linear Unbiased Estimator) de cada indivíduo, isto é, prover um valor predito para a média de cada indivíduo. Neste caso, mesmo encontrando o BLUP é necessário realizar a retransformação dos dados. A segunda alternativa é a utilização de um método Bootstrap para gerar as médias ajustadas de cada indivíduo sem fazer transformação dos dados.

Como exemplo de aplicação, será usada uma variável importante de um estudo nutricional, gentilmente, cedido pela Professora Teresa Macedo do departamento de

nutrição da Universidade de Brasília que utiliza uma amostra de adolescentes residentes no DF.

Os dois primeiros métodos exigem uma intensa manipulação dos dados, que são transformados e depois retransformados. Este fato pesa positivamente a favor do método Bootstrap que não usa nenhuma transformação nos dados. No entanto, o método Bootstrap desenvolvido neste trabalho é para o caso onde o número de entrevistas é o mesmo para cada indivíduo do estudo, o que é um ponto negativo visto que os outros dois métodos não tem essa suposição.

Palavras-chaves:

Dados nutricionais; S-Nusser; Máxima Verossimilhança Restrita; Bootstrap; Modelo Aleatório; BLUP

LISTA DE TABELAS

LISTA DE FIGURAS

SUMÁRIO

INTRODUÇÃO

É possível perceber que, no decorrer das últimas décadas, existiram mudanças significativas nos hábitos alimentares da população. Em decorrência deste fato e a fim de acompanhá-las, diversos estudos, utilizando dados nutricionais, estão sendo desenvolvidos. Conhecendo a estimativa da média diária de nutrientes que a população está ingerindo, é possível analisar se a população está se alimentando de forma adequada, com todos os nutrientes essenciais que o corpo necessita ou se algo deve ser mudado na alimentação da população em estudo.

Para começar a estudar esse problema, deve-se obter uma amostra da população de interesse e, a partir dela, estimar a quantidade de ingestão diária de cada nutriente de cada indivíduo. Para isso, entretanto, é necessário que haja informação suficiente de cada elemento da amostra. Dentro deste objetivo, uma estratégia a seguir é observar o indivíduo por pelo menos dois dias e, então, calcular a média para cada um.

Estimar a ingestão média diária de cada nutriente não é tão simples em decorrência das duas fontes de variação que estão naturalmente presente nas pessoas. A primeira é a variação entre os indivíduos e a segunda é a variação entre as observações de cada indivíduo. É comum uma pessoa não consumir nada de um determinado nutriente em um certo dia, porém, no dia seguinte, esta mesma pessoa pode ingerir algum alimento rico daquele nutriente. Estes fatos causam uma variação grande e instável entre as observações de um mesmo indivíduo.

Outro problema que surge ao se analisar os dados nutricionais é que, na maioria dos casos, eles seguem uma distribuição assimétrica à direita, uma vez que poucos indivíduos ingerem muito e outros muitos indivíduos ingerem pouco de determinado nutriente. Deste modo, não se deve usar um modelo assumindo a normalidade. A fim de contornar a variação entre as observações de um mesmo indivíduo e sabendo que os dados não provêm de uma distribuição normal, alternativas aos métodos usando a normalidade devem ser elaboradas.

O objetivo dos nutricionistas em um estudo como este é ter uma estimativa da média diária de nutrientes ingeridos pelos indivíduos de uma determinada população e, mediante as médias, estabelecer estimativas dos percentis de cada nutriente. Caso o percentil encontrado esteja abaixo ou acima do considerado saudável para o organismo é necessário tomar alguma medida para converter esse quadro.

Atualmente, o modelo de análise mais usado para este tipo de estudo é o modelo chamado S-Nusser proposto por Hoffmann et al.(2002) que modificou o resultado apresentado por Nusser et al.(1996). O método é baseado na análise de variância do modelo aleatório, com os dados transformados pela transformação de Box e Cox(1964). Esta análise fornece as estimativas das variâncias entre indivíduos e dentro de indivíduos necessárias para obter o valor predito da média de um indivíduo.

Basicamente, o modelo dos dados é da forma

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

sendo α_i o efeito aleatório do individuo i e y_{ij} representa um valor transformado da variável original observada em cada indivíduo. Todas as inferências feitas com esse modelo são conduzidas com a variável transformada para a normalidade e, no final, as médias são retransformadas considerando também uma forma de diminuir o viés devido à retransformação direta. O método contém um estimador tipo “shrinkage” que diminui a influência da variabilidade dentro do individuo.

Neste trabalho, será apresentado o desenvolvimento estatístico do método S-Nusser e de duas alternativas de análise dos dados. A primeira alternativa à este método é aplicar o método de Máxima Verossimilhança Restrita aos dados transformados e utilizar as equações de Henderson (Henderson et al.(1959)) para se obter o BLUP (Best Linear Unbiased Estimator) de cada indivíduo, isto é, prover um valor predito para a média de cada indivíduo. Deste caso, mesmo encontrando o BLUP é necessário realizar a retransformação dos dados. O estimador das equações de Henderson é diferente do apresentado pelo método S-Nusser, assim, sabemos que o estimador de S-Nusser não é BLUP.

A segunda alternativa é a utilização de um método Bootstrap para gerar as médias ajustadas de cada indivíduo sem fazer transformação dos dados. Será usado o desenvolvimento apresentado por Davison e Hinkley (1997) para dados hierárquicos, que são gerados pela amostragem citada, Bootstrap. Como exemplo de aplicação, será usada uma variável importante de um estudo nutricional, gentilmente, cedido pela Professora Teresa Macedo do departamento de nutrição da Universidade de Brasília.

Como o modelo aleatório é pouco conhecido, será feito, primeiramente, um desenvolvimento deste modelo fundamentado em Scheffé (1959). Em seguida, será abordado o método S-Nusser baseado em artigos relacionados. Também será feita uma descrição sobre o BLUP seguindo as linhas de Robinson (1991) e Searle et al. (2006), visto que predição de efeitos aleatórios também é pouco conhecida. Em sequência, é apresentada uma síntese do método de Máxima Verossimilhança Restrita também seguindo Searle et al. (2006) e uma apresentação prática do método Bootstrap. Um ponto importante deste trabalho é a parte computacional, bastante intensa nos três métodos. O software SAS será o aplicativo utilizado em todas as análises.

CAPÍTULO 1

Modelo Aleatório

1.1 Introdução

Na definição de Eisenhart (1947), existem três tipos de modelos: modelo de efeitos aleatórios, modelo de efeitos fixos e modelos mistos. O modelo de efeitos aleatórios é o mais importante para o entendimento deste estudo e, por isso, receberá mais ênfase a fim de detalhar e esclarecer a aplicação aqui proposta.

Para exemplificar, será tomado como base o artigo publicado no ano de 2007 pela professora Teresa H. M. Da Costa do departamento de nutrição da Universidade de Brasília. Um dos objetivos do trabalho é estimar a média diária de cada nutriente ingerida pelos adolescentes residentes no DF. Para esse tipo de análise, foi estudada a alimentação feita durante 24 horas em 2 dias consecutivos de uma amostra aleatória retirada da população de interesse. Pessoas experientes na área de nutrição utilizaram tabelas para saber quanto de cada nutriente (Vit. A, Proteínas, etc.) um indivíduo da amostra estudada ingeriu de acordo com a lista de alimentos com as respectivas quantidades fornecidas por elas.

Considere que y_{ij} seja a quantidade de um determinado nutriente ingerida pelo indivíduo i no dia j. Um primeiro modelo para $\mathrm{y_{ij}}$ é dado por

$$\mathrm{y_{ij}} = m_i + e_{ij}$$

sendo m_i a média do adolescente i e e_{ij} o erro aleatório do indivíduo i no dia j gerado pela repetição da observação no tempo. Existe uma variabilidade entre a quantidade ingerida por cada indivíduo durante os 2 dias, essa variabilidade é representada por e_{ij}, uma variável aleatória com variância σ_E^2.

Cada adolescente causa no modelo um efeito aleatório que pode ser definido por

$$\alpha_\mathrm{i} = m_i - \mu$$

onde μ é a média geral das observações. Deste modo, a maneira mais frequente de escrever o modelo é da forma

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

sendo α_i o efeito aleatório do individuo i e $m_i = \mu + \alpha_i$ é a média diária de cada nutriente ingerida por esse jovem. Observe que o modelo acima é formado apenas por componentes aleatórios (μ é considerado um efeito fixo, mas não é um parâmetro de interesse), assim, ele é chamado de modelo de efeitos aleatórios. As suposições usuais deste modelo são:

- $E(\alpha_i) = 0$ e a variância entre os indivíduos é dada por $Var(\alpha_i) = \sigma_G^2$

- $E(e_{ij}) = 0$ e a variância dentre os indivíduos é dada por $Var(e_{ij}) = \sigma_E^2$

Além disso, admite-se que os α_i's são independentes entre eles e entre os e_{ij}'s e ambos são normalmente distribuídos. Porém, é possível perceber que, em princípio, os e_{ij}'s não são independentes, visto que são dados repetidos no tempo para cada indivíduo. Entretanto, em um primeiro momento, para o entendimento do modelo aleatório, será considerado que os e_{ij} 's são independentes. A variância da observação y_{ij} é

$$\sigma_y^2 = \sigma_G^2 + \sigma_E^2$$

e é usual chamar σ_G^2 e σ_E^2 de componentes da variância. Por essa razão, o modelo de efeitos aleatórios também é chamado modelo de componentes de variância. No estudo clássico deste tópico, como o desenvolvido por Scheffé (1959), o principal foco do Modelo de Efeitos Aleatórios está em analisar os componentes de variância.

Normalmente, os estudos com dados nutricionais que acompanham as pessoas escolhidas na amostra, durante J dias, ocasionam um modelo não balanceado, visto que dificilmente todos os indivíduos da amostra participam do projeto todos os J dias pré-estabelecidos. Em uma situação ideal em que fosse possível coletar a informação em todos os J dias para cada indivíduo, diz-se que o modelo é balanceado.

Um dos caminhos para a estimação e demais inferências para este modelo é pela análise de variância. Se o modelo for balanceado, a análise de variância fornece os meios para se estimar as variâncias e realizar teste de hipótese através da estatística F, sendo possível ainda obter intervalos de confiança. Se representarmos por $\hat{\sigma}_x^2$ um dos componentes da variância, o resultado observado da esperança dos quadrados médios em uma ANOVA proverá uma estimativa para σ_x^2. Entretanto, esse procedimento não é tão simples para o caso desbalanceado. Desta forma, será realizada a análise de variância considerando o modelo balanceado.

1.2 Teste de hipótese, ANOVA e Estimativas

Como dito anteriormente, no modelo de efeitos aleatórios, deseja-se analisar os componentes da variância. Assim, as hipóteses dos teste são:

$$H_0: \sigma_G^2 = 0$$

$$H_a: \sigma_G^2 > 0$$

Deseja-se testar se a variância entre os indivíduos ou grupos é igual à zero, ou seja, testar se a média entre os indivíduos é igual à média geral.

Usualmente, utiliza-se a análise de variância para estimar as variâncias quando o modelo é balanceado. Desta forma, inicia-se com a soma de quadrado total

$$SQ_T = \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{..})^2$$

que mensura a variabilidade total dos dados. Essa soma pode ser escrita como

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}[(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})]^2$$

ou

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{..})^2 = J\sum_{i=1}^{I}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{i.})^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{i.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.})$$

No entanto, o último termo da equação acima é igual a zero, sabendo que

$$\sum_{j=1}^{J}(y_{ij} - \bar{y}_{i.}) = y_{i.} - J\bar{y}_{i.} = y_{i.} - J(y_{i.}/J) = 0$$

Desta forma, tem-se que

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{..})^2 = J\sum_{i=1}^{I}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}(y_{ij} - \bar{y}_{i.})^2$$

A equação anterior representa a diferença entre a média entre os indivíduos e a média geral mais a diferença entre a observação de cada indivíduo e a média dentre o indivíduo. Assim, a mesma equação pode ser escrita na forma

$$SQ_T = SQ_G + SS_E$$

onde SQ_G é a soma de quadrados entre os indivíduos e SS_E é a soma de quadrados dentre os indivíduos. Existem JI observações, assim SQ_T têm JI-1 graus de liberdade. Da mesma maneira, existem I indivíduos, então, SQ_G têm I-1 graus de liberdade. Finalmente, existem J observações para cada indivíduo, produzindo J-1 graus de liberdade para cada estimativa do erro experimental. Como há I indivíduos, SS_E possui I(J-1)=IJ-I graus de liberdade.

Os quadrados médios usado na análise de variância são dados por

$$QM_G = \frac{SQ_G}{I-1} \text{ e } QM_E = \frac{SQ_E}{IJ-I}.$$

Pelo modelo especificado no item anterior, temos que

$$\bar{y}_{i.} = \mu + \alpha_i + e_{i.}$$

e

$$\bar{y}_{..} = \mu + \alpha_{.} + e_{..}$$

Assim,

$$SQ_G = J\sum_{i=1}^{I}(\alpha_i + e_{i.} - \alpha_{.} - e_{..})^2$$

$$SQ_E = \sum_{i=1}^{I}\sum_{j=1}^{J}(e_{ij} - e_{i.})^2$$

A análise a seguir está baseado no desenvolvimento realizado por Scheffé (1959). Fazendo $g_i = \alpha_i + e_{i.}$, temos

$$SQ_G = J\sum_{i=1}^{I}(g_i - g_.)^2$$

sendo que a variável aleatória g_i é independente com distribuição normal com média zero e variância $\sigma_g^2 = \sigma_G^2 + \frac{1}{J}\sigma_E^2$, sabendo que $\sum_{i=1}^{I}(g_i - g_.)^2/\sigma_g^2$ é uma variável qui-quadrado com I-1 graus de liberdade. Assim,

$$SQ_G = J\sigma_g^2\chi_{I-1}^2 = (J\sigma_G^2 + \sigma_E^2)\chi_{I-1}^2$$

Sabendo que a esperança de uma qui-quadrado é igual ao número de graus de liberdade da mesma, então, a esperança do quadrado médio entre grupos é

$$E(QM_G) = E\left(\frac{SQ_G}{I-1}\right) = \frac{E(SQ_G)}{I-1} = \frac{(J\sigma_G^2 + \sigma_E^2)E(\chi_{I-1}^2)}{I-1} = J\sigma_G^2 + \sigma_E^2$$

Da mesma forma, será desenvolvido para o quadrado médio do erro. Como a variável aleatória e_{ij} é independente com distribuição normal com média zero e variância σ_e^2, então $\sum_{i=1}^{I}\sum_{j=1}^{J}(e_{ij} - e_{i.})^2/\sigma_E^2$ é uma qui-quadrado com IJ-I. Desta forma,

$$SQ_E = \sigma_E^2\chi_{IJ-I}^2$$

e

$$E(QM_E) = E\left(\frac{SQ_E}{IJ-I}\right) = \frac{E(SQ_E)}{IJ-I} = \frac{\sigma_E^2 E(\chi_{IJ-I}^2)}{IJ-I} = \sigma_E^2$$

A decisão sobre a hipótese nula que será testada será por meio da estatística F_0, dada pela razão

$$F_0 = \frac{QM_G}{QM_E} = \frac{\frac{SQ_G}{I-1}}{\frac{SQ_E}{IJ-I}} = \frac{(IJ-I)(J\sigma_G^2 + \sigma_E^2)\chi_{I-1}^2}{(I-1)\sigma_E^2\chi_{IJ-I}^2} = \left(\frac{J\sigma_G^2 + \sigma_E^2}{\sigma_E^2}\right)F_{(I-1),(IJ-I)}$$

onde $F_{(I-1),(IJ-I)}$ é a estatística F central com I-1 e IJ-I graus de liberdade. Portanto, se σ_G^2 for zero (H_0 é verdade), a hipótese nula será rejeitada se

$F_0 \geq F_{0,05;(I-1),(IJ-1)}$.

Finalmente, a tabela da análise de variância para os modelos aleatórios balanceados pode ser organizada.

Tabela 1 – Tabela da Análise de Variância para Modelos Aleatórios Balanceados

Fonte de variação	G.L.	SQ	QM	E(QM)	F_0
Grupo	I-1	$(J\sigma_G^2 + \sigma_E^2)\chi_{I-1}^2$	$SQ_G/(I-1)$	$J\sigma_G^2 + \sigma_E^2$	QM_G/QM_E
Erro	IJ-I	$\sigma_E^2\chi_{IJ-I}^2$	$SQ_E/(IJ-I)$	σ_E^2	
Total	IJ-1				

Observando a Tabela 1, fica fácil perceber que

$$QM_G = J\hat{\sigma}_G^2 + \hat{\sigma}_E^2 \quad \text{e} \quad QM_E = \hat{\sigma}_E^2$$

Assim, as estimativas para os componentes da variância podem ser encontradas. São elas

$$\hat{\sigma}_E^2 = QM_E$$

e

$$\hat{\sigma}_G^2 = \frac{QM_G - QM_E}{J}$$

Segundo Scheffé (1959), se algum QM distribuído como uma constante vezes uma variável qui-quadrado dividido pelo número de graus de liberdade,

$$QM = \frac{c\chi_{g.l.}^2}{g.l.}$$

então, $c = E(QM)$. Assim,

$$Var(QM) = \frac{2[E(QM)]^2}{g.l.}$$

Onde $Var(QM) = c^2 Var\left(\frac{\chi_{g.l.}^2}{g.l.}\right) = 2\frac{c^2}{g.l.}$. Assim

$$Var(\hat{\sigma}_E^2) = Var(QM_E) = 2\frac{\sigma_E^4}{IJ - I}$$

e

$$Var(QM_G) = \frac{2(J\hat{\sigma}_G^2 + \hat{\sigma}_E^2)}{(I-1)}$$

Agora, será desenvolvida a variância do estimador $\hat{\sigma}_G^2$. Sabendo que $\mathrm{QM_G}$ e $\mathrm{QM_E}$ são estatísticas independentes, temos que

$$Var(\hat{\sigma}_G^2) = \frac{Var(QM_G) + Var(QM_E)}{J^2} = \frac{2\left(\sigma_G^2 + \frac{\sigma_E^2}{J}\right)^2}{I-1} + \frac{2\left(\frac{\sigma_E^2}{J}\right)^2}{IJ - I}$$

O desenvolvimento do caso não balanceado é o mesmo no que tange a ANOVA e as estimativas dos componentes de variância. No entanto, não é possível realizar um teste de hipótese nem outras inferências que dependem da normalidade, visto que não se conhece a distribuição do QM_G. As estimativas dos componentes principais podem ser encontradas da mesma forma que no caso balanceado.

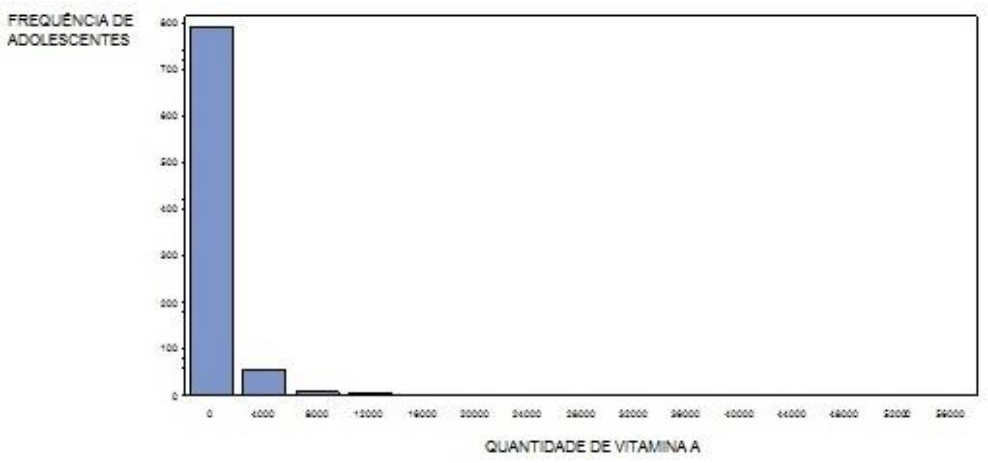
CAPÍTULO 2

Método S-Nusser

2.1 Introdução

A maioria dos métodos utilizados para estimar a média de nutrientes ingeridos pela população baseia-se no pressuposto de que as quantidades ingeridas pelos indivíduos durante os J dias seguem uma distribuição normal. Como dito na introdução deste projeto, na maioria dos casos, os dados seguem uma distribuição assimétrica à direita. Para exemplificar, a Figura 1 ilustra os resultados para a Vitamina A obtida a partir no trabalho da professora Teresa no ano de 2007 como citado no capítulo anterior.

Figura 1 – Distribuição da quantidade ingerida de Vitamina A



Nota-se que muitos indivíduos ingerem pouca vitamina A enquanto poucos ingerem uma quantidade maior da mesma, ocasionando em uma distribuição assimétrica. Para contornar essa situação, uma das alternativas é a transformação de dados, como a de Box e Cox(1964), para se obter uma distribuição normal aproximada. Nesta parte inicial do projeto, abordaremos o método de S-Nusser, chamado método simplificado de Nusser, usado para estimar a média de ingestão diária de nutrientes.

2.2 Método de S-Nusser

O método S-Nusser aparece de forma detalhada no trabalho de Hoffmann at al.(2002). Como dito na introdução deste projeto, os dados seguem uma distribuição assimétrica à direita. Assim, o primeiro passo é transformar os dados por meio da transformação de Box-Cox (1946). Esta transformação consiste em encontrar um λ tal que os dados transformados se aproximem de uma distribuição normal. A variável y_{ij} será transformada em uma variável z_{ij}, onde

$$z_{ij} = \begin{matrix} \dfrac{y^{\lambda} - 1}{\lambda} \; se \; \lambda \neq 0 \\ \log(y_{ij}) \; se \; \lambda = 0 \end{matrix}$$

Destaca-se aqui que o método S-Nusser usa outra estatística para se obter o melhor valor de lambda. A estatística mais conhecida é a fornecida pela função de verossimilhança obtida para cada valor de lambda dentro de um intervalo de -2 à 2. Aquela que produz o máximo da função é a escolhida. O método S-Nusser usa o teste de Shaphiro-Wilk para testar a hipótese nula de que os dados tem distribuição normal. O λ que produzir o maior p-valor é o escolhido.

O modelo utilizado é então dado por

$$z_{ij} = \mu + A_i + e_{ij}$$

onde A_i é o efeito aleatório do indivíduo i, de acordo com a escala dada pela transformação de Box e Cox. Os efeitos $A_i's$ são independentes entre eles e seguem uma distribuição normal com média zero e variância σ_G^2. Também como no modelo com a escala original, para prosseguir com as análises, o modelo precisa atender ao pressuposto de que os $e_{ij}'s$ são independentes e seguem uma distribuição com média zero e variância σ_E^2. Além disso, os $A_i's$ e os $e_{ij}'s$ são independentes entre eles. Com a escala transformada, é possível estimar as variâncias e a média para cada indivíduo i.

Um ponto importante é destacar que a variância da média do indivíduo i é dada por

$$Var(\bar{y}_{i.}) = \sigma_G^2 + \frac{\sigma_E^2}{J}$$

Ocorre que σ_E^2 pode ser muito grande e uma maneira de ajustar a média do indivíduo, à esta circunstância, é usar um estimador especial chamado de Estimador Shrinkage da média do indivíduo i dado por

$$\bar{z}_{iA} = \alpha\bar{z}_i + (1-\alpha)\bar{z}_{..}$$

onde $\bar{z}_i$ é a média, na escala transformada, de nutrientes ingerida pelo indivíduo i nos J dias, $\bar{z}_{..}$ é a média geral com a escala transformada e α é dado por

$$\alpha = \sqrt{\frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_E^2}{J}}}$$

Nota-se que se não existe variância entre os indivíduos, ou seja, $\sigma_G^2 = 0$, α também é igual à zero e $\bar{z}_{iA} = \bar{z}_{..}$ (média geral) . Por outro lado, se não existe variância dentre os indivíduos, ou seja, $\sigma_E^2 = 0$, α é igual a um e $\bar{z}_{iA} = \bar{z}_i$ (média do indivíduo i). Esses dois casos são extremos e não deve ocorrer, o que acontece na prática é que α fica em entre 0 e 1 e, desta forma, a média "Shrinkage" de um indivíduo fica mais próximo da média geral. Esse nome "Shrinkage" vem dessa característica do novo estimador em que a sua distância para a média geral é diminuída.

As variâncias σ_G^2 e σ_E^2 são estimadas pela análise de variância explicada anteriormente. Vale ressaltar que os tempos observados em um mesmo indivíduo foram considerados independentes, isto é aceito visto que o intervalo de uma entrevista para outra é grande o suficiente para não haver correlação entre eles.

Depois de obter $\bar{z}_{iA}$, outro passo do método é voltar os dados transformados para a escala original, isto é, obter as médias estimadas na escala original. A primeira ideia é usar a transformação inversa, ou seja, se $z_{ij} = g(y_{ij})$, então $\bar{y}_{i.} = g^{-1}(\bar{z}_{i.})$, onde $\bar{y}_{i.}$ é a medida de ingestão do indivíduo i. Entretanto, a volta direta de uma média com dados transformados, via função inversa, é viesada como foi estudado por Neyman e Scott (1960) e John e Quenouille (1977).

Embora o foco em delineamento de experimentos, Neyman e Scott (1960) fizeram um detalhado estudo sobre como ajustar os dados para corrigir o viés da comparação entre dois tratamentos. Esta volta, frequentemente, é feita sem ajustamento. Baseado nos autores citador, mas com a adaptação para o modelo aleatório, o método de S-Nusser também contém uma proposta para corrigir o viés gerado pela volta direta à escala original que é desenvolvida a seguir.

Seja z a média transformada, y a média não transformada e suponha que foi encontrado $\lambda \neq 0$, então, pela transformação de Box e Cox temos que

$$z = z_{ij} = g(y_{ij}) = \frac{{y_{ij}}^{\lambda} - 1}{\lambda}.$$

Assim,

$$y = (\lambda z + 1)^{1/\lambda} = h(z).$$

Depois da transformação, sabemos que z segue uma distribuição Normal com média μ e variância σ_Z^2. Em vez de prosseguir com a volta direta para a escala original, devemos encontrar a esperança de $h(z)$, isto é,

$$y = E(h(z)) = \int_R h(z) f(z) dz = \int_R (\lambda z + 1)^{1/\lambda} f(z) dz$$

onde $f(z)$ é a densidade da normal $z \approx N(\mu, \sigma_Z^2)$. Na prática, μ e σ_Z^2 são substituidos pelas suas respectivas estimativas.

Essa é a forma de voltar à escala original proposta pelo método de S-Nusser. O processo é o mesmo se $\lambda = 0$ e, portanto, a transformação utilizando o logaritmo.

Um outro desenvolvimento para corrigir o viés da volta direta é proposto por Dodd at. al. (2006), já relacionado com o modelo em questão. Suponha novamente que $\lambda \neq 0$ e, então,

$$z = g(y_{ij}) = \frac{{y_{ij}}^{\lambda} - 1}{\lambda}$$

e $y = (\lambda z + 1)^{1/\lambda}$ representa a volta direta. Segundo o autor, deve-se diminuir desse valor $\frac{1}{2}(1 - \lambda)(\lambda z + 1)^{\frac{1-2\lambda}{\lambda}} \sigma_E^2$, onde σ_E^2 é a variância do erro nos dados transformados.

É possível notar a semelhança até certo ponto com a forma proposta por Neyman e Scott (1960).

CAPÍTULO 3

BLUP

3.1 Best Linear Unbiased Predictor (BLUP)

No Capítulo 1 neste trabalho, foi apresentada a especificação de um modelo aleatório através do artigo da professora Teresa H. M. da Costa (2007) do departamento de Brasília que estudou a alimentação feita durante J dias consecutivos de uma amostra de adolescentes residentes no DF. O modelo para a quantidade de um determinado nutriente ingerida pelo indivíduo i é dado por

$$y_{ij} = m_i + e_{ij}$$

sendo m_i a média da quantidade de nutrientes que o adolescente i ingeriu nos J dias e e_{ij} o erro aleatório do indivíduo i no dia j.

Como dito anteriormente, cada adolescente acrescenta no modelo um efeito aleatório α_i. Assim, é possível obter a seguinte relação

$$m_i = \alpha_i + \mu$$

onde μ é a média geral das observações.

Ao se realizar um estudo desse tipo, conhecemos a informação da média ingerida por cada indivíduo durante os J dias dada por $\bar{y}_{i.}$. Para avaliar a contribuição do indivíduo i no modelo, será medida a informação de $\alpha_i + \mu$. Portanto, deve-se predizer o valor de $\alpha_i + \mu$ dado que $\bar{y}_{i.}$ é conhecido, isto é,

$$E(\alpha_i + \mu \mid \bar{y}_{i.})$$

Admite-se que α_i e e_{ij} seguem uma distribuição normal com média zero e variância σ_G^2 e σ_E^2, respectivamente, e admite-se também que os α_i's são independentes entre eles e entre os e_{ij}'s e ambos são normalmente distribuídos.

Temos que a média $\bar{y}_{i.}$ possui variância dada por $\mathrm{Var}(\bar{y}_{i.}) = \sigma_G^2 + \frac{\sigma_E^2}{J}$, visto que $\bar{y}_{i.} = \frac{\sum_{j=1}^{J} y_{ij}}{J} = \alpha_i + \mu + \frac{\sum_{j=1}^{J} e_{ij}}{J}$. Já a covariância entre a média $\bar{y}_{i.}$ e o efeito aleatório α_i é igual à $\mathrm{Cov}(\bar{y}_{i.}, \alpha_i) = \sigma_G^2$, pois $\mathrm{Cov}\left(\alpha_i + \mu + \frac{\sum_{j=1}^{J} e_{ij}}{J}, \alpha_i\right) = \mathrm{Var}(\alpha_i) + \mathrm{Cov}\left(\mu + \frac{\sum_{j=1}^{J} e_{ij}}{J}, \alpha_i\right) = \mathrm{Var}(\alpha_i) = \sigma_G^2$.

Como μ é uma constante, o valor de $\alpha_i + \mu$ que desejamos predizer pode ser escrito por

$$\mathrm{E}(\alpha_i + \mu \mid \bar{y}_{i.}) = \mu + \mathrm{E}(\alpha_i \mid \bar{y}_{i.}).$$

Usando a distribuição Normal, tem-se que

$$\begin{bmatrix} \alpha_i \\ \bar{y}_{i.} \end{bmatrix} \approx N\left\{ \begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_G^2 & \sigma_G^2 \\ \sigma_G^2 & \sigma_G^2 + \frac{\sigma_E^2}{J} \end{bmatrix} \right\}.$$

Assim, pode-se encontrar

$$\mathrm{E}(\alpha_i \mid \bar{y}_{i.}) = \frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_E^2}{J}} (\bar{y}_{i.} - \mu)$$

e

$$\mathrm{Var}(\alpha_i \mid \bar{y}_{i.}) = \sigma_G^2 - \frac{(\sigma_G^2)^2}{\sigma_G^2 + \frac{\sigma_E^2}{J}}$$

Com algumas manipulações algébricas e fazendo $\beta = \frac{\sigma_E^2}{\sigma_E^2 + J\sigma_G^2}$, temos que,

$$\mathrm{E}(\alpha_i + \mu \mid \bar{y}_{i.}) = (1 - \beta)\bar{y}_{i.} + \beta\mu$$

Embora semelhante, note que este estimado não é o mesmo que o “Shrinkage estimador” definido no capítulo anterior no Método S-Nusser. O estimador encontrado também é conhecido como Best Linear Unbiased Predictor (BLUP). Significa dizer que é o melhor estimador com o menor erro quadrático médio, linear e não viesado. As demonstrações destas propriedades estão em Searle et al. (2006). Um

estudo panorâmico sobre a predição dos efeitos aleatórios e em particular do BLUP e suas aplicações, pode ser encontrado no trabalho de Robinson (1991).

3.2 Equações de Henderson

Uma outra maneira de encontrar o Best Linear Unbiased Predictor é mediante o desenvolvimento de modelos mistos proposto por Henderson (1950) e Henderson et al. (1959). Considere o modelo na forma matricial a seguir:

$$\mathrm{y} = \mathrm{X}\beta + \mathrm{Z}\alpha + \varepsilon$$

onde y é o vetor de observações; X é a matriz de incidência dos efeitos fixos (conhecida), β é o vetor de efeitos fixos, Z é a matriz de incidência dos efeitos aleatórios, α é o vetor de efeitos aleatórios e, por fim, ε é o vetor de erros aleatórios.

Assume-se que os efeitos aleatórios e os erros seguem uma distribuição normal com média zero e matrizes de variâncias G e R, respectivamente, isto é,

$$\mathrm{Var}(\alpha) = \mathrm{G} \text{ e } Var(\varepsilon) = R$$

Temos também que $\mathrm{Cov}(\alpha, \mathrm{y}) = \mathrm{GZ}'$. Desta maneira, tem-se que

$$\mathrm{V} = \mathrm{Var}(\mathrm{y}) = \mathrm{Var}(\mathrm{Z}\alpha) + \mathrm{Var}(\varepsilon) = \mathrm{ZGZ}' + \mathrm{R}$$

e

$$\mathrm{E}(\mathrm{y}) = \mathrm{X}\beta$$

Segundo Henderson, para obter soluções para os efeitos fixos e predizer os efeitos aleatórios, deve ser feito a derivação das equações de modelos mistos pela maximização da função de densidade de probabilidade conjunta de y e α.

A função de densidade de probabilidade conjunta de y e α pode ser escrita como o produto da função de probabilidade de y dado o efeito aleatório α com a função de densidade de α. Assim, temos que

$$\mathrm{f}(\mathrm{y}, \alpha) = \mathrm{f}(\mathrm{y}|\alpha)\ \mathrm{f}(\alpha)$$

Considerando R e G fixos, a equação acima pode ser reescrita como

$$f(y,\alpha) = \frac{1}{(2\pi)^{n/2}[R]^{1/2}}\exp\{-\frac{1}{2}[(y-X\beta-Z\alpha)'(R)^{-1}(y-X\beta-Z\alpha)]\}\frac{1}{(2\pi)^{n/2}[G]^{1/2}}\exp\{-\frac{1}{2}[(\alpha-0)'(G)^{-1}(\alpha-0)]\}$$

Sabendo que os pontos de máximo de $f(y,\alpha)$ e $\log[f(y,\alpha)]$ são coincidentes, é mais fácil utilizar $L = \log[f(y,\alpha)]$ para dar continuidade à maximização. Em seguida, deve-se derivar L em relação aos parâmetros α e β e igualar a zero. Desta forma, obtemos as seguintes equações:

$$\begin{bmatrix}\frac{\partial L}{\partial \beta}\\ \frac{\partial L}{\partial \alpha}\end{bmatrix} = \begin{bmatrix} -X'R^{-1}y + X'R^{-1}X\hat{\beta} + X'R^{-1}Z\hat{\alpha} \\ -Z'R^{-1}y + Z'R^{-1}X\hat{\beta} + Z'R^{-1}Z\hat{\alpha} + G^{-1}\hat{\alpha}\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$

Isolando os termos que contêm y para o lado direito das equações, obtém-se

$$\begin{bmatrix} X'R^{-1}X\hat{\beta} + X'R^{-1}Z\hat{\alpha} \\ Z'R^{-1}X\hat{\beta} + Z'R^{-1}Z\hat{\alpha} + G^{-1}\hat{\alpha}\end{bmatrix} = \begin{bmatrix}X'R^{-1}y\\ Z'R^{-1}y\end{bmatrix}$$

Por fim,

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1}\end{bmatrix}\begin{bmatrix}\hat{\beta}\\ \hat{\alpha}\end{bmatrix} = \begin{bmatrix}X'R^{-1}y\\ Z'R^{-1}y\end{bmatrix}$$

Essas equações, chamadas equações de Henderson que permitem encontrar as soluções para $\hat{\beta}$ e $\hat{\alpha}$. Neste caso, temos

$$\begin{bmatrix}\hat{\beta}\\ \hat{\alpha}\end{bmatrix} = \begin{bmatrix}(X'V^{-1}X)^{-1}\,X'V^{-1}y \\ GZ'V^{-1}(y - X\hat{\beta})\end{bmatrix}$$

A solução encontrada para β é a mesma obtida pelos métodos de quadrados mínimos ponderados, máxima verossimilhança e máxima verossimilhança restrita. Será apresentado, no capítulo seguinte, este último método citado.

Encontrando a solução para α pelas equações de Henderson para o modelo que estamos usando, é possível verificar que (Searle at al. 2006) a solução é o valor predito de $E(\alpha|\bar{y}_{i.})$, obtido anteriormente.

CAPÍTULO 4

Método da Máxima Verossimilhança Restrita

4.1 Método da Máxima Verossimilhança Restrita (REML)

Um outro método de estimação de componentes de variâncias é chamado de Método da Máxima Verossimilhança Restrita (REML). Segundo Searle (1992), uma ideia básica do método é que as estimativas das componentes das variâncias baseiam-se nos resíduos calculados após os ajustes por quadrados mínimos ordinários. Uma propriedade importante do REML é o fato deste método levar em consideração os graus de liberdade envolvidos na estimação dos efeitos fixos para, assim, estimar as variâncias. Esta é uma das principais diferenças em relação ao Método de Máxima Verossimilhança que não considera os graus de liberdade na estimação.

O Método da Máxima Verossimilhança Restrita consiste em maximizar a parte da função de verossimilhança invariante ao parâmetro de locação, ou seja, os efeitos fixos. É preciso fazer combinações lineares nos elementos de y de forma que os efeitos fixos sejam eliminados.

De maneira geral, considere o modelo misto da forma

$$y = X\beta + Z\alpha + \varepsilon$$

onde $\varepsilon \sim N(0, R)$, $\alpha \sim N(0, G)$ e, além disso, os α_i's são independentes entre eles e entre os e_{ij}'s. Assim, temos que y também segue uma distribuição Normal com média $X\beta$ e variância V, considerando $V = ZGZ' + R$.

Precisamos encontrar uma matriz $K_{n \times (n-t)}$ de posto $(n - t)$ de forma que $K'Y = K'X\beta + K'Z\alpha + K'\varepsilon$ seja invariante à $X\beta$. Assim, $K'X\beta = 0$ para todo β e, portanto, $K'X = 0$. Desta forma, temos que $K'Y$ segue uma distribuição Normal com média zero e variância $K'VK$.

Agora, deve-se aplicar o Método da Máxima Verossimilhança em $K'Y$ para obter os estimadores das variâncias chamados de estimadores de Máxima Verossimilhança Restrita.

Segundo Searle, os resultados dos estimadores são invariantes também à escolha da matriz K'. Na maioria dos casos, as matrizes G e R, matriz de variância e covariância do erro e dos efeitos aleatórios, respectivamente, possuem estruturas conhecidas como, por exemplo, a matriz identidade vezes uma constante.

Pelo método da Máxima Verossimilhança e com a estimativa da matriz V, obtém-se a estimativa do vetor de efeitos fixos $\hat{\beta}$ chamado de estimador de verossimilhança restrita de β dado por

$$\hat{\beta} = \left(X'\hat{V}^{-1}X\right)^{-}X'\hat{V}^{-1}Y.$$

Uma propriedade importante é que, para dados balanceados, as soluções pelo REML são as mesmas que os estimadores encontrado pela análise de variância (ANOVA). Já para dados não balanceados, o método foi abordado de maneira ampla por Patterson e R. Thompson (1971).

Para que seja possível estudar este método na aplicação deste estudo com dados nutricionais definido no primeiro capítulo, considere $\mathrm{X} = 1$ e $\beta = \mu$. Lembrando que Z é a matriz de incidências dos indivíduos e α o vetor dos efeitos aleatórios, a forma matricial do modelo abordado no primeiro capítulo é dada por

$$\mathrm{y} = \mu 1 + \mathrm{Z}\alpha + \varepsilon$$

Admite-se que a média dos efeitos aleatórios α é zero e sua variância é dada por $Var(\alpha) = \mathrm{I}\sigma^2_{\mathrm{G}}$; a média dos erros é zero e a variância residual é dada por $Var(\varepsilon) = \mathrm{I}\sigma^2_{\mathrm{E}}$; os α_i's são independentes entre eles e entre os e_{ij}'s.

Conhecendo a variância de α e ε, sabe-se que a matriz de variância de y é

$$V = Z\mathrm{I}\sigma^2_{\mathrm{G}}Z' + \mathrm{I}\sigma^2_{\mathrm{E}} = ZZ'\sigma^2_{\mathrm{G}} + \mathrm{I}\sigma^2_{\mathrm{E}}$$

Desta forma, é possível aplicar o método da Máxima Verossimilhança Restrita para obter as estimativas das variâncias e o estimador de BLUP para cada indivíduo.

CAPÍTULO 5

Método Bootstrap

5.1 Introdução

Nos capítulos anteriores foi feito um estudo sobre as alternativas para se obter o valor predito da média de ingestão de um determinado nutriente para os indivíduos da amostra e, utilizando os valores preditos, estimar os respectivos percentis para os nutrientes que servirão de referência para a área de nutrição identificar o estado nutricional da população de interesse. Essas alternativas fazem uso da distribuição Normal e da transformação de dados devido ao fato das inferências com modelos aleatórios serem baseadas na distribuição Normal e porque, na maioria dos casos, as variáveis que representam dados nutricionais não tem distribuição Normal.

Os resultados obtidos pelos vários modelos de análise usando normalidade são amplamente aceitos pela comunidade científica, entretanto existem outras formas de inferências que não dependem da normalidade que podem ser usadas neste estudo. Neste capítulo, será abordado uma dessas formas: a inferência via Bootstrap.

5.2 Bootstrap

Estudos utilizando o método Bootstrap são considerados recentes devido à grande necessidade da tecnologia dos computadores para que sua implementação fosse facilitada. O avanço computacional foi essencial neste aspecto, visto que as inferências estatísticas utilizando método Bootstrap são feitas por meio de simulações (Efron e Tibshirani (1993)).

Existem diversos artigos e livros sobre o método, entretanto, as primeiras ideias sugiram no trabalho pioneiro de Bradley Efron em 1979. Alguns pontos em relação ao método Bootstrap serão discutidos neste trabalho, no entanto, os desenvolvimentos

práticos e teóricos sobre esse importante tópico são encontrados em Efron e Tibshirani (1993) e Davison Hinkley (1997), entre outros.

A maioria das aplicações de Bootstrap é feita em conjuntos de dados independentes, entretanto, dados nutricionais que estão sendo trabalhos neste estudo não são independentes, por se tratar de um modelo hierárquico com dois níveis, o dos indivíduos e do tempo, o que gera uma correlação entre os dados de um mesmo indivíduo. Dessa forma, a aplicação para este caso é diferente em alguns pontos e estudos envolvendo esse tipo de modelo foi desenvolvido por Davison e Hinkley (1997). Mais adiante será abordada a aplicação proposta por esses dois autores, mas, para conseguirmos adaptar para o estudo em questão utilizando dados nutricionais, será relembrado qual modelo está sendo utilizado neste trabalho.

O modelo mais frequente para estudos com dados nutricionais é dado por

$$\mathrm{y_{ij}} = \mu + \alpha_i + e_{ij}$$

onde y_{ij} é a quantidade de um determinado nutriente ingerida pelo indivíduo i no dia j, μ é a média geral das observações, α_i o efeito aleatório do individuo i e e_{ij} o efeito aleatório do indivíduo i no dia j. Supõe-se que α_i e e_{ij} possuem média zero e variância denotador por σ_G^2 e σ_E^2, respectivamente.

Para as alternativas de análise anteriores supomos também que os α_i's são independentes entre eles e entre os e_{ij}'s e ambos são normalmente distribuídos. No entanto, para estudar o método de Bootstrap, segundo Davison e Hinkley (1997), não são necessárias essas suposições. Segundo esses dois autores, para aplicar Bootstrap nesse tipo de dados, é preciso seguir os seguintes passos:

1) De uma amostra de tamanho n, selecionar m amostras Bootstrap com reposição de tamanho n utilizando os dados do primeiro nível (indivíduo) e mantendo os dados do segundo nível;

2) Em cada amostra, estima-se os componentes da variância por um método não-paramétrico como o de Rao (1977), por exemplo[1];

3) Em cada amostra e com as estimativas dos componentes da variância, calcula-se um estimador "Shrinkage" da média de cada indivíduo. Dois estimadores podem ser obtidos: primeiro o estimador Shrinkage usado no método de S-Nusser e segundo o estimador encontrado pelo BLUP.

Lembrando que o estimador Shirinkage da média do indivíduo i para o método de S-Nusser é dado por

$$\widetilde{\bar{y}}_{i.} = \alpha \bar{y}_{i.} + (1-\alpha)\bar{y}_{..}$$

sabendo que $\bar{y}_{i.}$ é a estimativa antiga da média do indivíduo i, $\bar{y}_{..}$ é a média geral e α é dado por

$$\alpha = \sqrt{\frac{\sigma_G^2}{\sigma_G^2 + \frac{\sigma_E^2}{J}}}$$

Já no caso do estimador BLUP para predizer o valor de $\alpha_i + \mu$ dado que $\bar{y}_{i.}$ é conhecido, temos que

$$E(\alpha_i + \mu \mid \bar{y}_{i.}) = (1 - \beta)\bar{y}_{i.} + \beta\mu$$

onde

$$\beta = \frac{\sigma_E^2}{\sigma_E^2 + J\sigma_G^2}.$$

4) Por fim, calcula-se a média das estimativas encontradas por cada amostra para cada indivíduo.

Mesmo utilizando um número m grande de amostras para o processo de Bootstrap, a frequência de cada indivíduo nas m amostras não é constante, pois a amostragem é com reposição. Entretanto, com m=10000 a frequência de cada

[1] Como a amostragem é com reposição, um indivíduo pode aparecer mais de uma vez na amostra e isto deve ser levado em consideração na estimativa dos componentes da variância em cada amostra. O PROC MIXED do SAS (2003) possui uma opção para este caso.

indivíduo ficou em torno de 62%. Um programa, encontrado no anexo, em linguagem MACRO do SAS foi construído para gerar os resultados.

CAPÍTULO 6

Aplicação

Neste trabalho foram apresentados três métodos alternativos para analisar dados nutricionais. Como exemplo, para as três alternativas, será usado os dados fornecidos pela professora Teresa Costa do departamento de nutrição da Universidade de Brasília. Para compor o banco de dados, foi formada uma amostra aleatória de 326 adolescentes residentes no DF que praticavam atividade física na época. Os jovens foram entrevistados durante 4 dias consecutivos. Em cada dia, os jovens informavam quais alimentos eles ingeriram nas 24h anteriores à entrevista. Essas informações eram passadas para pessoas experientes na área de nutrição para que elas pudessem informar a quantidade de cada nutriente que o indivíduo ingeriu de acordo com o foi dito na entrevista.

Era de se esperar que os dados coletados não fossem balanceados, visto que nem todos os 326 adolescentes da amostra participaram dos 4 dias de entrevista. No entanto, para as análises neste trabalho foi utilizado apenas 2 dias de entrevista para que fosse possível ter o mesmo número de entrevistas para todos os jovens. Desta forma, o banco de dados para esse estudo é considerado balanceado.

Seja y_{ij} a quantidade de um nutriente ingerida pelo indivídio i no dia j, o modelo mais usadao para y_{ij} é dado por

$$\mathrm{y_{ij}} = \mu + \alpha_i + e_{ij}$$

onde μ é a média geral das observações, α_i o efeito aleatório do individuo i e e_{ij} é o erro aleatório do indivídio i no dia j.

Sabendo que a maioria das quantidades de cada nutriente ingerida pelos jovens segue comportamento semelhante, foi considerado para as análises apenas um nutriente: o caroteno. As médias para os 326 indivíduos foram obtidas de acordo com as definições de cada método e utilizando o software SAS (2003). Os resultados estão representados na Tabela 2 mostrada a seguir.

Tabela 2 – Médias originais e ajustadas para o nutriente caroteno

(Continua)

Indivíduo	Original	S-Nusser	BLUP	Nusser_B	BLUP_B	Indivíduo	Original	S-Nusser	BLUP	Nusser_B	BLUP_B
1	2,7750	3,3061	2,7728	2,4919	2,5776	44	2,4400	3,0989	2,6596	2,3913	2,4071
2	1,0150	1,6392	1,7647	1,9620	1,6810	45	4,1350	3,7778	3,0215	2,9031	3,2718
3	1,5900	2,3769	2,2419	2,1342	1,9733	46	2,3350	3,0466	2,6306	2,3601	2,3540
4	10,6700	5,2434	3,7320	4,8462	6,5670	47	3,1200	3,0244	2,6182	2,5953	2,7530
5	6,3400	5,7517	3,9612	3,5698	4,4003	48	1,9950	2,2745	2,1791	2,2574	2,1806
6	6,6350	3,4761	2,8639	3,6476	4,5348	49	0,9200	1,6493	1,7717	1,9320	1,6313
7	1,8000	2,5570	2,3499	2,1975	2,0804	50	2,7100	3,3394	2,7907	2,4725	2,5448
8	4,4950	2,6265	2,3908	3,0073	3,4503	51	0,8100	1,4836	1,6548	1,8995	1,5759
9	5,8900	4,9703	3,6056	3,4299	4,1645	52	0,3250	0,8050	1,1162	1,7519	1,3271
10	5,9700	5,4257	3,8150	3,4558	4,2083	53	0,1550	0,5086	0,8303	1,7016	1,2415
11	1,5800	1,8925	1,9358	2,1306	1,9676	54	6,4050	5,3983	3,8026	3,5867	4,4288
12	0,3850	0,7178	1,0366	1,7708	1,3583	55	5,7300	5,1920	3,7084	3,3818	4,0840
13	1,7750	2,5421	2,3410	2,1906	2,0680	56	0,8300	0,9984	1,2822	1,9033	1,5845
14	0,1750	0,5729	0,8966	1,7071	1,2509	57	0,5150	1,1064	1,3699	1,8114	1,4261
15	1,9150	2,6255	2,3902	2,2331	2,1397	58	7,8500	6,5738	4,3172	4,0351	5,1843
16	3,6200	4,0408	3,1554	2,7462	3,0077	59	5,3200	3,9272	3,0980	3,2577	3,8722
17	2,1350	1,7237	1,8227	2,2996	2,2520	60	0,5900	1,2564	1,4868	1,8345	1,4651
18	2,2850	2,9428	2,5725	2,3427	2,3268	61	2,0500	2,5062	2,3197	2,2741	2,2087
19	1,9950	2,7444	2,4594	2,2560	2,1796	62	1,8950	2,0369	2,0297	2,2269	2,1293
20	3,2000	1,8303	1,8946	2,6181	2,7923	63	1,5050	2,1786	2,1195	2,1091	1,9301
21	0,6000	1,2660	1,4941	1,8349	1,4676	64	0,7600	1,3628	1,5667	1,8850	1,5508
22	0,5950	1,0192	1,2994	1,8332	1,4647	65	5,1400	4,2154	3,2426	3,2039	3,7821
23	2,1900	2,9199	2,5596	2,3150	2,2791	66	4,9000	3,8811	3,0745	3,1299	3,6577
24	3,4900	3,2792	2,7582	2,7072	2,9412	67	1,0950	1,8164	1,8853	1,9859	1,7216
25	0,7800	1,1090	1,3720	1,8905	1,5610	68	0,7150	1,4054	1,5981	1,8696	1,5261
26	12,2000	8,5411	5,1102	5,3839	7,4549	69	0,6600	1,0077	1,2899	1,8529	1,4981
27	1,4600	2,1076	2,0747	2,0958	1,9077	70	0,6850	0,9580	1,2485	1,8605	1,5109
28	0,2800	0,7657	1,0807	1,7403	1,3059	71	2,2750	2,4981	2,3148	2,3426	2,3238
29	5,2850	3,5658	2,9112	3,2436	3,8510	72	1,4200	1,9844	1,9958	2,0814	1,8854
30	0,9750	0,8755	1,1782	1,9484	1,6594	73	1,2000	1,9849	1,9961	2,0180	1,7759
31	0,6300	1,1987	1,4424	1,8463	1,4855	74	0,8200	1,2134	1,4538	1,9002	1,5790
32	0,6950	1,3894	1,5864	1,8664	1,5187	75	1,4100	1,9505	1,9738	2,0814	1,8827
33	4,5350	2,8538	2,5221	3,0193	3,4708	76	0,5750	1,2355	1,4708	1,8299	1,4571
34	5,7400	5,4350	3,8193	3,3886	4,0936	77	0,7250	1,4185	1,6077	1,8731	1,5314
35	0,2750	0,7223	1,0408	1,7361	1,3008	78	3,7800	2,5319	2,3349	2,7933	3,0880
36	0,7400	1,4150	1,6051	1,8769	1,5391	79	0,6350	1,1577	1,4105	1,8478	1,4875
37	1,8800	2,2140	2,1417	2,2213	2,1209	80	2,2550	2,1828	2,1222	2,3343	2,3119
38	2,4350	2,6296	2,3926	2,3904	2,4051	81	2,2700	2,8506	2,5203	2,3388	2,3196
39	0,5100	1,1426	1,3986	1,8112	1,4250	82	5,7250	4,1841	3,2270	3,3789	4,0780
40	0,4400	1,0399	1,3163	1,7862	1,3856	83	0,4000	0,6772	0,9986	1,7752	1,3663
41	0,3950	0,9322	1,2268	1,7744	1,3644	84	0,4650	0,9497	1,2416	1,7968	1,4012
42	1,0550	1,8239	1,8903	1,9715	1,6990	85	2,8650	1,9999	2,0059	2,5176	2,6225
43	1,6200	2,3467	2,2235	2,1430	1,9882	86	4,1400	4,3856	3,3263	2,9045	3,2743

Tabela 2 – Médias originais e ajustadas para o nutriente caroteno

(Continuação)

Indivíduo	Original	S-Nusser	BLUP	Nusser_B	BLUP_B	Indivíduo	Original	S-Nusser	BLUP	Nusser_B	BLUP_B
87	1,7300	2,4893	2,3096	2,1781	2,0459	129	1,4250	2,0893	2,0631	2,0852	1,8896
88	0,5150	0,9194	1,2159	1,8117	1,4265	130	5,4150	3,4515	2,8508	3,2840	3,9185
89	1,4650	1,3580	1,5632	2,0951	1,9083	131	0,8100	1,5372	1,6931	1,9008	1,5769
90	0,3500	0,8858	1,1871	1,7611	1,3413	132	2,5600	2,9573	2,5807	2,4259	2,4672
91	0,2950	0,7245	1,0429	1,7436	1,3129	133	1,1700	1,7832	1,8630	2,0064	1,7578
92	0,9750	1,6940	1,8024	1,9491	1,6597	134	2,2700	2,9061	2,5518	2,3395	2,3201
93	2,9000	2,6901	2,4280	2,5289	2,6407	135	7,9800	5,3320	3,7725	4,0567	5,2261
94	4,3700	3,4493	2,8496	2,9696	3,3871	136	3,5900	3,3713	2,8079	2,7359	2,9912
95	0,9400	1,6342	1,7612	1,9362	1,6402	137	2,1550	2,5951	2,3724	2,3041	2,2611
96	0,5500	1,0812	1,3497	1,8226	1,4449	138	1,1250	1,8408	1,9016	1,9942	1,7363
97	1,0800	1,7036	1,8090	1,9837	1,7161	139	1,2600	1,6396	1,7649	2,0345	1,8045
98	10,5000	2,7933	2,4876	4,7773	6,4574	140	0,7800	1,4229	1,6109	1,8894	1,5596
99	5,8700	4,4946	3,3793	3,4227	4,1525	141	0,3900	0,9205	1,2168	1,7716	1,3604
100	1,7450	2,4816	2,3050	2,1804	2,0517	142	0,6700	1,3628	1,5667	1,8565	1,5038
101	5,1350	4,7320	3,4933	3,2043	3,7828	143	0,1950	0,6125	0,9360	1,7115	1,2594
102	0,9500	1,6898	1,7995	1,9402	1,6462	144	0,6400	1,2430	1,4766	1,8488	1,4895
103	0,6150	1,2842	1,5079	1,8403	1,4761	145	2,5500	3,1055	2,6632	2,4228	2,4620
104	0,6900	1,2782	1,5034	1,8623	1,5134	146	0,9650	1,7140	1,8161	1,9460	1,6549
105	1,1150	1,8931	1,9362	1,9906	1,7306	147	0,7650	1,0196	1,2996	1,8845	1,5515
106	3,4550	3,8318	3,0493	2,6968	2,9244	148	1,1200	1,8890	1,9335	1,9959	1,7363
107	0,7750	1,3949	1,5904	1,8861	1,5553	149	0,3600	0,9027	1,2016	1,7638	1,3460
108	0,5900	0,9672	1,2562	1,8334	1,4644	150	1,7050	2,4646	2,2948	2,1694	2,0323
109	4,2350	3,3823	2,8138	2,9291	3,3190	151	1,2350	0,9936	1,2782	2,0269	1,7920
110	6,2750	5,5607	3,8759	3,5497	4,3656	152	0,9450	1,3655	1,5687	1,9397	1,6444
111	0,6600	1,3244	1,5382	1,8544	1,4992	153	1,1300	1,8202	1,8878	1,9957	1,7388
112	4,6900	4,6387	3,4488	3,0664	3,5517	154	1,6750	1,8485	1,9066	2,1611	2,0175
113	4,5550	3,8863	3,0772	3,0283	3,4842	155	1,4450	2,2362	2,1554	2,0906	1,8994
114	11,6650	6,9970	4,4942	5,1730	7,1131	156	3,3450	2,0050	2,0091	2,6617	2,8662
115	0,8750	1,1557	1,4089	1,9191	1,6092	157	4,7500	2,4218	2,2691	3,0831	3,5794
116	3,8300	3,1069	2,6640	2,8100	3,1149	158	1,1450	1,4519	1,6320	2,0002	1,7466
117	3,4400	3,8962	3,0822	2,6906	2,9159	159	2,7250	2,8825	2,5384	2,4771	2,5523
118	0,7850	1,4909	1,6601	1,8903	1,5614	160	0,5250	1,1270	1,3863	1,8140	1,4304
119	1,0550	1,8172	1,8858	1,9739	1,7012	161	1,5400	2,3109	2,2016	2,1200	1,9482
120	0,4550	0,9617	1,2516	1,7919	1,3944	162	0,0900	0,3682	0,6743	1,6810	1,2072
121	2,1950	2,6960	2,4313	2,3169	2,2818	163	1,3500	2,0723	2,0523	2,0638	1,8525
122	1,4600	2,0704	2,0511	2,0955	1,9076	164	1,8600	2,5883	2,3684	2,2165	2,1116
123	1,5150	2,2241	2,1479	2,1120	1,9352	165	3,7750	3,9914	3,1305	2,7926	3,0869
124	3,0250	3,5847	2,9211	2,5670	2,7050	166	0,5100	1,1387	1,3955	1,8101	1,4235
125	0,3000	0,8021	1,1135	1,7462	1,3163	167	1,7850	1,7367	1,8316	2,1939	2,0735
126	0,6600	1,3505	1,5576	1,8554	1,4998	168	0,7750	1,4919	1,6608	1,8895	1,5584
127	0,9300	1,3938	1,5896	1,9350	1,6366	169	1,4700	2,0132	2,0144	2,0990	1,9129
128	0,5550	1,2078	1,4495	1,8237	1,4468	170	1,2300	1,8769	1,9255	2,0260	1,7896

Tabela 2 – Médias originais e ajustadas para o nutriente caroteno

(Continuação)

Indivíduo	Original	S-Nusser	BLUP	Nusser_B	BLUP_B	Indivíduo	Original	S-Nusser	BLUP	Nusser_B	BLUP_B
171	2,9850	3,5312	2,8930	2,5544	2,6842	213	0,5350	0,8074	1,1183	1,8159	1,4349
172	4,7400	2,8453	2,5173	3,0801	3,5741	214	2,6400	3,1095	2,6654	2,4528	2,5101
173	0,7800	1,4679	1,6435	1,8911	1,5609	215	0,2650	0,7432	1,0602	1,7356	1,2984
174	2,0600	1,6800	1,7928	2,2767	2,2135	216	0,8850	1,5306	1,6885	1,9216	1,6140
175	1,5350	2,3258	2,2107	2,1192	1,9466	217	3,5750	4,0091	3,1394	2,7328	2,9863
176	0,5200	1,0642	1,3360	1,8116	1,4275	218	2,5100	2,6842	2,4245	2,4114	2,4421
177	0,9250	1,6174	1,7495	1,9354	1,6358	219	3,2850	3,6585	2,9597	2,6457	2,8377
178	1,7350	1,8303	1,8946	2,1777	2,0471	220	2,4500	3,1176	2,6699	2,3946	2,4125
179	1,5950	2,3802	2,2439	2,1376	1,9773	221	2,1950	2,6602	2,4105	2,3174	2,2822
180	0,4150	0,9057	1,2042	1,7811	1,3747	222	5,1750	4,4475	3,3565	3,2140	3,7990
181	0,7850	1,3746	1,5755	1,8915	1,5628	223	1,5950	2,2718	2,1775	2,1363	1,9762
182	1,9400	2,5673	2,3559	2,2399	2,1517	224	0,3500	0,8556	1,1609	1,7613	1,3417
183	2,0150	2,0579	2,0431	2,2630	2,1904	225	0,4050	0,8755	1,1782	1,7764	1,3683
184	3,9350	3,1375	2,6809	2,8404	3,1672	226	1,4500	1,7995	1,8740	2,0920	1,9017
185	2,0800	2,8278	2,5073	2,2816	2,2229	227	18,7250	9,8754	5,6111	7,3973	10,8470
186	1,4650	2,1878	2,1253	2,0960	1,9091	228	6,1350	5,2356	3,7284	3,5050	4,2907
187	2,5300	3,1596	2,6930	2,4189	2,4534	229	2,1400	2,7780	2,4788	2,3007	2,2541
188	4,9950	4,9628	3,6021	3,1626	3,7114	230	0,7100	1,4123	1,6031	1,8674	1,5228
189	1,5500	1,2336	1,4694	2,1211	1,9522	231	3,3650	3,6041	2,9313	2,6680	2,8770
190	0,5600	1,1820	1,4295	1,8225	1,4464	232	2,6250	2,1176	2,0811	2,4463	2,5009
191	4,7250	4,7265	3,4906	3,0794	3,5716	233	1,3900	2,0315	2,0262	2,0759	1,8729
192	2,1100	2,1184	2,0816	2,2921	2,2393	234	0,6550	1,2802	1,5049	1,8528	1,4968
193	0,5100	1,1013	1,3658	1,8080	1,4216	235	10,7400	7,6986	4,7795	4,9216	6,6784
194	0,9750	1,6632	1,7812	1,9490	1,6596	236	2,5300	3,0659	2,6413	2,4179	2,4526
195	0,2950	0,5813	0,9050	1,7446	1,3135	237	3,6750	2,5591	2,3511	2,7621	3,0353
196	0,1900	0,5968	0,9204	1,7120	1,2591	238	1,4250	2,1155	2,0798	2,0850	1,8896
197	1,1100	1,8092	1,8805	1,9901	1,7290	239	1,3350	2,0185	2,0178	2,0561	1,8423
198	0,3450	0,8270	1,1357	1,7574	1,3367	240	2,4050	2,8249	2,5056	2,3788	2,3880
199	1,6150	2,0924	2,0651	2,1423	1,9863	241	0,3300	0,8185	1,1281	1,7561	1,3319
200	0,4600	0,8480	1,1542	1,7922	1,3956	242	0,9750	1,7286	1,8260	1,9479	1,6588
201	0,3200	0,7722	1,0866	1,7517	1,3255	243	2,0350	2,6734	2,4182	2,2688	2,2006
202	1,1050	1,8348	1,8975	1,9880	1,7261	244	0,3700	0,9223	1,2183	1,7666	1,3516
203	0,8500	1,1032	1,3674	1,9113	1,5961	245	4,6450	4,4215	3,3438	3,0538	3,5287
204	0,8750	1,4707	1,6455	1,9189	1,6088	246	4,3350	2,5745	2,3602	2,9584	3,3679
205	0,8800	1,6248	1,7547	1,9213	1,6121	247	1,3950	2,1855	2,1239	2,0747	1,8732
206	0,8200	1,3922	1,5885	1,9030	1,5814	248	0,4950	0,8892	1,1900	1,8038	1,4148
207	0,8750	1,1748	1,4239	1,9190	1,6086	249	0,5650	0,9339	1,2282	1,8261	1,4512
208	0,8000	1,5286	1,6870	1,8950	1,5696	250	3,8700	4,0074	3,1386	2,8201	3,1342
209	0,7650	1,4836	1,6548	1,8842	1,5512	251	0,4050	0,7729	1,0872	1,7766	1,3683
210	2,8250	3,4445	2,8470	2,5062	2,6027	252	0,2250	0,6239	0,9471	1,7222	1,2769
211	2,3150	2,9975	2,6032	2,3529	2,3431	253	1,1850	1,2386	1,4732	2,0127	1,7673
212	1,4250	2,0984	2,0689	2,0851	1,8896	254	2,7950	2,2163	2,1431	2,4974	2,5874

Tabela 2 – Médias originais e ajustadas para o nutriente caroteno

(Conclusão)

Indivíduo	Original	S-Nusser	BLUP	Nusser_B	BLUP_B	Indivíduo	Original	S-Nusser	BLUP	Nusser_B	BLUP_B
255	2,9650	3,0996	2,6600	2,5488	2,6741	291	3,0950	3,2986	2,7688	2,5884	2,7408
256	1,9350	1,9615	1,9810	2,2395	2,1502	292	2,4350	3,1107	2,6661	2,3884	2,4037
257	1,0000	1,7121	1,8148	1,9585	1,6745	293	1,7100	2,3632	2,2335	2,1706	2,0345
258	0,9700	1,7304	1,8273	1,9476	1,6572	294	1,4850	2,2739	2,1788	2,1035	1,9203
259	6,0250	3,8150	3,0407	3,4673	4,2276	295	2,4500	2,9933	2,6008	2,3932	2,4114
260	17,5700	6,2403	4,1748	6,8596	9,9955	296	2,3150	2,2662	2,1740	2,3533	2,3432
261	3,1650	2,3984	2,2549	2,6096	2,7766	297	2,6700	3,3207	2,7807	2,4605	2,5243
262	0,2450	0,4443	0,7611	1,7275	1,2857	298	4,7750	4,8127	3,5315	3,0952	3,5982
263	5,6900	4,6669	3,4623	3,3693	4,0619	299	2,4000	2,4691	2,2975	2,3776	2,3856
264	0,8850	1,6239	1,7541	1,9208	1,6131	300	3,8200	2,2521	2,1653	2,8045	3,1074
265	0,6750	1,2429	1,4765	1,8587	1,5069	301	2,7050	2,9680	2,5866	2,4701	2,5413
266	2,8000	3,2242	2,7284	2,4992	2,5901	302	2,8100	3,4165	2,8321	2,5021	2,5954
267	3,8100	4,1747	3,2224	2,8035	3,1056	303	2,8150	3,4072	2,8271	2,5052	2,5989
268	5,1150	4,5607	3,4113	3,1972	3,7703	304	3,3500	3,8443	3,0557	2,6646	2,8702
269	1,2000	1,9352	1,9638	2,0175	1,7751	305	0,5700	1,2257	1,4633	1,8276	1,4535
270	0,3750	0,9383	1,2319	1,7681	1,3537	306	1,6050	2,3357	2,2168	2,1400	1,9819
271	26,1300	5,3898	3,7988	9,1661	13,9807	307	3,5550	3,5827	2,9201	2,7256	2,9738
272	2,4150	3,0816	2,6500	2,3834	2,3942	308	1,5100	2,2170	2,1435	2,1116	1,9334
273	3,5700	4,0033	3,1365	2,7302	2,9820	309	3,2350	2,9839	2,5956	2,6292	2,8109
274	0,6950	1,3598	1,5645	1,8656	1,5176	310	2,3900	3,0937	2,6567	2,3760	2,3815
275	1,2100	1,9699	1,9864	2,0203	1,7798	311	0,4650	0,9605	1,2506	1,7941	1,3986
276	11,6300	8,5555	5,1157	5,2089	7,1584	312	0,4000	0,9094	1,2073	1,7775	1,3684
277	0,7100	1,3514	1,5583	1,8706	1,5259	313	0,3500	0,8680	1,1716	1,7610	1,3418
278	0,3950	0,7460	1,0628	1,7745	1,3649	314	3,2500	3,7649	3,0149	2,6342	2,8193
279	6,9300	6,1405	4,1316	3,7498	4,7037	315	1,6400	2,3669	2,2358	2,1501	1,9995
280	2,5550	2,1801	2,1205	2,4256	2,4656	316	5,2900	4,2626	3,2659	3,2474	3,8560
281	5,0700	4,6314	3,4452	3,1834	3,7471	317	1,2150	1,7841	1,8636	2,0215	1,7825
282	7,6200	5,0562	3,6456	3,9481	5,0422	318	0,1650	0,5393	0,8623	1,7066	1,2483
283	1,0000	1,7613	1,8482	1,9567	1,6726	319	2,0900	2,8265	2,5065	2,2860	2,2290
284	1,0000	1,6018	1,7386	1,9557	1,6718	320	0,9950	1,6965	1,8041	1,9552	1,6702
285	2,5300	2,9075	2,5526	2,4187	2,4532	321	0,6050	1,2755	1,5013	1,8390	1,4725
286	1,0650	1,8364	1,8986	1,9758	1,7055	322	0,7050	1,3116	1,5285	1,8668	1,5215
287	1,7450	2,1502	2,1017	2,1831	2,0539	323	3,3050	3,7414	3,0028	2,6508	2,8471
288	2,0850	2,8323	2,5099	2,2854	2,2270	324	1,1500	1,8052	1,8778	2,0043	1,7511
289	0,4900	1,1146	1,3764	1,8029	1,4122	325	2,8500	3,4558	2,8530	2,5153	2,6163
290	2,0800	2,8186	2,5020	2,2833	2,2241	326	0,5600	1,1380	1,3950	1,8228	1,4469

Nota:

Original = Média dos dados originais;

S-Nusser = Média predita pelo método S-Nusser

BLUP = Média BLUP (Máxima Verossimilhança Restrita)

Nusser_B = Média predita pelo método Bootstrap usando o shrinkage de S-Nusser

BLUP_B = média predita pelo método Bootstrap usando o shrinkage BLUP

É possível observar que, para um mesmo indivíduo, os métodos fornecem médias diferentes, entretanto, espera-se que todas as médias estejam mais próximas da média geral que é 2,3644.

Nota-se, como exemplo, o 4º indivíduo da Tabela 2 que ingeriu uma média alta do nutriente caroteno comparada com a média geral. A média original deste indivíduo foi igual a 10,6700 e as demais médias ajustadas são menores que esse valor e mais próximas da média geral. Para exemplificar o caso em que a média original do indivíduo é menor que a média geral, nota-se o valor da média original do indivíduo $i = 12$. A média deste indivíduo foi igual a 0,3850 e as médias ajustadas são maiores que esse valor e mais próximas da média geral comparado com a média original.

Na Tabela 3, são fornecidos os principais percentis para as médias mostradas na Tabela 2. O nutricionista deve analisar os percentis para saber se a população está se alimentando de forma correta em relação à ingestão de um determinado nutriente. Caso o percentil encontrado esteja abaixo ou acima do considerado saudável para o organismo é necessário tomar alguma medida para converter o quadro.

Tabela 3 – Percentis para as médias originais e ajustadas para o nutriente caroteno.

Percentil	Original	S-Nusser	BLUP	Nusser_B	BLUP_B
15°	0,5500	1,0196	1,2996	1,8225	1,4449
20°	0,6600	1,1987	1,4424	1,8529	1,4981
25°	0,7650	1,3244	1,5382	1,8845	1,5512
30°	0,8750	1,4185	1,6077	1,9189	1,6086
35°	0,9750	1,6392	1,7647	1,9491	1,6597
40°	1,1300	1,7832	1,8630	1,9959	1,7388
45°	1,4100	1,8890	1,9335	2,0814	1,8827
50°	1,5375	2,0714	2,0517	2,1196	1,9474
55°	1,7450	2,2170	2,1435	2,1831	2,0539
60°	2,0800	2,3984	2,2549	2,2816	2,2229
65°	2,3150	2,6255	2,3902	2,3529	2,3431
70°	2,5550	2,8506	2,5203	2,4256	2,4656
75°	2,9000	3,0937	2,6567	2,5289	2,6407
80°	3,4550	3,3713	2,8079	2,6968	2,9244
85°	4,2350	3,7778	3,0215	2,9291	3,3190

Analisando a média ajustada utilizando o método bootstrap com o estimador do método da Máxima verossimilhança, por exemplo, a estimativa do 50º percentil para a quantidade ingerida do nutriente caroteno é de 1,9474. Um nutricionista ou qualquer especialista da área deve julgar se a quantidade é suficiente e correta para a população em estudo ou se alguma medida deve ser tomada.

Outros comentários sobre os resultados estão no capítulo final deste estudo.

CAPÍTULO FINAL

O primeiro método apresentado como alternativas para analisar dados nutricionais foi o de S-Nusser, o segundo foi o método de Máxima Verossimilhança Restrita e, por último, o método Bootstrap. Será pontuado neste capítulo as principais diferenças entre os métodos.

Foi analisado que o estimador "shrinkage" é o BLUP utilizando o método da Máxima Verossimilhança Restrita, entretanto não é o BLUP usando o método de S-Nusser. Este fato pode ser flexibilizado ao se usar o método Bootstrap. Vimos também que o método S-Nusser e o método da Máxima Verossimilhança Restrita fazem uso da suposição de normalidade dos dados, o que não ocorre. Na tentativa de deixá-los mais próximos da Normal, esses dois primeiros métodos exigem uma intensa manipulação dos dados, que são transformados e depois retransformados. Este fato pesa positivamente a favor do método Bootstrap que não usa nenhuma transformação nos dados, o que não é suficiente para declará-lo o melhor, pois o método Bootstrap, desenvolvido por Davinson e Hinkley (1997), é para o caso onde o número de entrevistas é o mesmo para cada indivíduo do estudo, o que é um ponto negativo visto que os outros dois métodos não tem essa suposição.

As médias ajustadas para um determinado indivíduo diferem de método para método provocando questionamento sobre qual usar. As diferenças pontuadas neste estudo nos dá uma direção de qual método usar. Em um trabalho futuro, também poderiam ser construídos, em cada método, intervalos de confiança para a média predita e para os percentis.

Os programas em SAS (2003) utilizados no desenvolvimento deste trabalho estão no anexo. A parte computacional para o método S-Nusser e para o método de Máxima Verossimilhança é relativamente simples comparadas com o programa utilizando para o método de Bootstrap que é bastante elaborado.

Evidentemente, outros métodos existem para serem aplicados em dados desta natureza. Como os dados são extremamente assimétricos à direita, pode-se optar, por

exemplo, pela distribuição gama e usar o modelo linear generalizado misto, que fornece o um preditor BLUP para as médias.

Este estudo visou apresentar alternativas para predizer a média de nutrientes ingerida por cada pessoa e outros tópicos relacionados ainda precisam ser pesquisados no futuro, como por exemplo, formas de se retransformar os dados, outros métodos que podem ser usados e formas de comparação dos métodos.

REFERÊNCIAS BIBLIOGRÁFICAS

BOX, G. E. P. E COX, D. R. **An analysis of transformations**. Journal of the Royal Statistical Society, Series B, 26, 211-243. 1964

CARY, NC. **SAS** Instituto SAS. 2003

DAVINSON, A. C. e HINKLEY, D. V. **Bootstrap Methods and their Applications**. Cambridge University Press. 1997

EFRON, B. **Bootstrap methods: another look at the Jackknife**. Annals of Statistics 7, 1-26. 1979

EFRON, B. e TIBSHIRANI, R **An Introduction to the Bootstrap**. New York: Chapman & Hall. 1993

EISENHART, C. **The assumptions underlying the analysis of variance.** Biometrics 3, 1-21. 1947

HENDERSON, C. R. **Estimation of genetic parameters (abstract).** Ann. Math. Statist. 21 309-310. 1950

HENDERSON, C. R. ; KEMPTHORNE,O. ; SEARLE, S.R. ; KROSIGK, C. N. **Estimation of environmental and genetic trends from records subject to culling**. Biometrics 15, 192-218. 1959

HOFFMANN, K.; BOEING, H.; DUFOUR, A.; VOLATIER, JL.; TELMAN, J.; VIRTANEN, M.; BECKER, W.; DE HENAUW, S. **Estimating the distribution of usual dietary intake by short-term measurements.** European Journal of Clinical Nutrition. 56, Suppl. 2, 553-562. 2002

JOHN, J. A. ; QUENOUILLE, M. H. **Experiments: Design and Analysis**. Charles Griffin & Company . 1977

LITTELL, R.C.; MILLIKEN, G.A.; STROUP, W.W.; WOLFINGER,R.D. **SAS System for Mixed Models**. 2006

NEYMAN, J. ; SCOTT, E. L. **Correction for bias introduced by a transformation of variables**. Ann. Math. Statist. , 31, 643-655. 1960

P.S.R.S. RAO (1977), **Theory of the MINQUE: A review**, Sankyã Ser. B. 1977

QUENOUILLE, M. H. **The Design and Analysis of Experiment**. Hafner, NY, 1953

ROBINSON, G.K. **That BLUP is a Good Thing: The Estimation of Random Effects**. Statistical Science, Vol. 6, No. 1 . 15-32. 1991

SCHEFFÉ, H. **The Analysis of Variance**. Wiley. 1959

SEARLE, S.R. **Linear Models. Wiley**. 1971

SEARLE, S.R.; CASELLA, G.; MC CULLOCH, C.E. **Variance Components**. Wiley.1992

SEARLE, S.R.; CASELLA, G.; MC CULLOCH, C.E. **Variance Components**. Wiley. 2006

SOUSA, E. F.; DA COSTA, T. H. M.; NOGUEIRA, J. A. D.; VIVALDI, L. J**. Assessment of nutrient and water intake among adolescents from sports federations in the Federal District, Brazil .** British Journal of Nutrition 2007

STEIN, C. **A two-sample test for a linear hypothesis whose power is independent of the variance**. Ann. Math. Statist. 16 243-258. 1945

ANEXOS

ANEXO A – Programa em SAS

```
/*************************************************************************************/
/* ADOLESCENTES ..  ELIENE                                                           */
/*INPUT PLANO $ 1-4 TIPO 6 IDN $ 7-11 Vit A   carotene       Vit E   Vit B12 Vit B1   */
/*  Vit B2 Vit B6      folic   Vit C   niacine sodium  potassium      calcium magnesium */
/* phosphorus copper   iron    zinc    protein fat     carbohydr;                    */
/*************************************************************************************/


LIBNAME NUTRI "H:\KINGSTON\TERESA\ELIENE"; RUN;
OPTIONS LS=120;
DATA BELEM3; SET NUTRI.ELIENE1;
IF TIPO < 3;
PROC PRINT DATA=BELEM3;  RUN;

DATA TUDO; SUJEITO="0000";RUN;

%MACRO MMMM(DO1);

    %DO L1 =&DO1 %TO &DO1 ;    /* DO PARA LER CADA VARIAVEL */

DATA TESTE1; SET BELEM3;
SUJEITO=INDIVIDUO;
 Z=N&L1;
IF Z=. THEN DELETE;
KEEP SUJEITO  TIPO  Z;
OPTIONS LS=90 NODATE NONUMBER;
TITLE LISTAGEM DOS DADOS;
/*PROC PRINT DATA=TESTE1; RUN;*/


TITLE ESTATISTICAS DESCRITIVAS;
PROC UNIVARIATE  DATA=TESTE1 normal;
VAR Z;
RUN;

TITLE H=2 'DISTRIBUIÇÃO DOS DADOS ORIGINAIS ';
PROC GCHART DATA=TESTE1;
VBAR Z /LEVELS=15 TYPE=FREQ;
RUN; QUIT;

PROC MEANS DATA=TESTE1 MEAN NOPRINT;
VAR Z;
OUTPUT OUT=MGR MEAN=MO;
RUN;
/*PROC PRINT DATA=MGR; RUN;*/
DATA TESTE1; MERGE TESTE1 MGR;
PLUS+MO;
MO=PLUS;
DROP PLUS _TYPE_ _FREQ_; RUN;
        /*PROC PRINT DATA=TESTE1;RUN; */


/*************************************************************/
/* CALCULO DO VALOR DE K PARA CADA SUJEITO                   */
```

```
/*************************************************************/


TITLE CALCULO DE  K(K=_FREQ_);
DATA VAK; SET TESTE1; C=1;
PROC SORT DATA=VAK ; BY SUJEITO;
PROC MEANS   DATA=VAK MEAN NOPRINT ; BY SUJEITO;
VAR C ;
OUTPUT OUT=VAK1 MEAN=C;
RUN;
/*proc PRINT DATA=VAK1; RUN;*/

TITLE CALCULO INTERMEDIARIO PARA VARIANCIA;
DATA VARS; SET VAK1;
K2=_FREQ_**2;
PROC MEANS DATA=VARS SUM noprint;
VAR _FREQ_ K2;
OUTPUT OUT=INTER SUM=TOTAL K2;
RUN;
/*PROC PRINT DATA=INTER; RUN;*/


%END;

%MEND  MMMM;                    /* FIM DO MACRO */
RUN;

%MMMM(2)

/*************************************************************/
/*  TRANSFORMACAO DOS DADOS                                  */
/*************************************************************/


      %LET NOME1=N2;
       %LET NOME2=N2_C;
          RUN;

 /*  DATA SET  AUXILIAR PARA ACUMULACAO */
  DATA DESCA;
 INPUT  WJ  J skewness kurtosis normal  probn;
 DATALINES;
 . 100 0 0 0 0 0
 ;
 RUN;

 /*PROC PRINT DATA=DESCA; RUN; */

 %MACRO TRANS;

 %MMMM(2)

        %DO S1 = 1 %TO 40 %BY 3;
          %DO H1 = 1 %TO 40 %BY 3;
            DATA MA1; SET TESTE1;

 S=0&S1;

H=0&H1;

WJ= (1/(H))*MO;
```

```
J=1/(S);

Y = ((Z+WJ)**J -1)/J ; RUN;

 TITLE  TESTE DE SHAPHIRO-WILKS;
 proc univariate DATA=MA1 NOPRINT; BY WJ J;
 output out=MBOX2 probn=probn normal=normal skewness=skewness
 kurtosis=kurtosis;
 var Y;
 ID S; RUN;

 DATA DESCA; SET DESCA MBOX2; RUN;

   %END;
    %END;
 %MEND TRANS;        /* FIM DO MACRO */
 RUN;
 %TRANS

 PROC PRINT DATA=DESCA; RUN;

 DATA TLOG; SET TESTE1;
 Y= LOG(Z ); WJ=0; J=0;
 run;
 /*PROC PRINT DATA=TLOG;*/
 TITLE  TESTE DE SHAPHIRO-WILKS;
 proc univariate DATA=TLOG NOPRINT;
 output out=MBOX3 probn=probn normal=normal skewness=skewness
 kurtosis=kurtosis;
 var Y;
 RUN;
 DATA MBOX4; SET MBOX3; WJ=0; J=0; RUN;
 /*PROC PRINT DATA=MBOX4; RUN;*/

 DATA BOX1; SET DESCA MBOX4;
 IF J > 99 THEN DELETE;CRITERIO=ABS(skewness);
 RUN;
 /*PROC PRINT DATA=BOX1; RUN;*/

 OPTIONS PS=200;
 PROC SORT DATA=BOX1; BY descending /* CRITERIO*/ normal ;RUN;
 PROC PRINT DATA=BOX1; RUN;


 /*****************************************************************/
 /*DADOS DA MELHOR TRANSFORMAÇÃO  USANDO A ASSIMETRIA OU TESTE DE S-W */
 /*****************************************************************/

 TITLE  DADOS DA MELHOR TRANSFORMAÇÃO ;
 DATA MELHOR; SET BOX1;
 NN +1;
 IF NN=1;
 KEEP WJ J;
 PROC PRINT DATA=MELHOR; RUN;

 DATA TESTE1; MERGE TESTE1 MELHOR;
 PLUS1+WJ;
 PLUS2+J;
 WJ=PLUS1;
 J=PLUS2;
 DROP PLUS1 PLUS2 ;
```

```
RUN;
/*PROC PRINT DATA=TESTE1;*/

 TITLE TRANSFORMACAO LOG OU OUTRA;
DATA TRR; SET TESTE1;
 IF J  = 0 THEN   Y = LOG(Z) ;
 IF J >  0 THEN  Y=((Z+WJ)**J -1)/J ;
  RUN;
PROC PRINT DATA=TRR; run;

/*****************************************/
/*****************************************/

PROC MIXED DATA= TRR;
CLASS SUJEITO;
MODEL Y=/SOLUTION;
RANDOM SUJEITO/SOLUTION;
RUN;
QUIT;

/*****************************************/
/*****************************************/

ods html;
ods graphics on;
TITLE H=2 'DISTRIBUIÇÃO  DA VARIAVEL TRANSFORMADA';
PROC GCHART DATA=TRR;
VBAR Y /TYPE=FREQ ;
RUN;
ods graphics off;
ods html close;
QUIT;

TITLE CALCULO DOS SIGMAS( SIGMAT, SIGMAE E SIGMAX=SIGMA MEDIA);
PROC GLM  DATA=TRR OUTSTAT=D3 NOPRINT;
CLASS SUJEITO;
MODEL  Y= SUJEITO /SS3;
RUN;

PROC PRINT DATA=D3; RUN;
PROC VARCOMP  DATA=TRR METHOD=TYPE1;
CLASS SUJEITO;
MODEL  Y= SUJEITO ;
RUN;

TITLE ESTIMATIVA DE SIGMAE;
DATA VAR1; SET D3; IF _SOURCE_="ERROR" ;
SIGMAE=SS/DF;
DROP _TYPE_;
PROC PRINT DATA=VAR1; RUN;


DATA VAR2; SET D3; IF _SOURCE_="SUJEITO" ;
DF_S=DF;
SQ_S=SS;
KEEP  DF SS;
PROC PRINT DATA=VAR2; RUN;

TITLE CALCULO DO SIGMAT;
DATA VAR3; MERGE VAR1 VAR2  INTER;
CONS=(1/(_FREQ_- 1))*(TOTAL - K2/TOTAL);
```

```
SIGMAT=(SS/DF - SIGMAE)/CONS;
KEEP SIGMAT SIGMAE;

PROC PRINT DATA=VAR3; RUN;
DATA VARCO; MERGE VAR3 VAK1;
SERR +SIGMAE;
SSUJEITO+SIGMAT;
SIGMAE=SERR;
SIGMAT=SSUJEITO;
DROP SERR SSUJEITO; RUN;
PROC PRINT DATA=VARCO; RUN;

TITLE CALCULO DAS MEDIAS(POR SUJEITO E GERAL);
PROC SORT DATA=TRR  ; BY SUJEITO;
PROC MEANS DATA=TRR MEAN NOPRINT; BY SUJEITO;
VAR Y;
OUTPUT OUT=MEDIAS MEAN=MS;
RUN;
PROC PRINT DATA=MEDIAS; RUN;

TITLE MEDIA GERAL DOS DADOS TRANSFORMADOS;
PROC MEANS DATA=TRR MEAN ;
VAR Y;
OUTPUT OUT=MEDIAG MEAN=MG;
RUN;
PROC PRINT DATA=MEDIAG; RUN;

TITLE CALCULO INTERMEDIARIO;
DATA JUNTO1; MERGE VARCO MEDIAG;
GERAL+ MG; DROP MG _FREQ_; RUN;

/*PROC PRINT DATA=JUNTO1; RUN; */


PROC SORT DATA=VAK1; BY SUJEITO; RUN;
PROC SORT  DATA=JUNTO1 ; BY SUJEITO;
PROC SORT  DATA=MEDIAS ; BY SUJEITO;
PROC SORT  DATA=TRR ; BY SUJEITO;
DATA JUNTO2; MERGE TRR VAK1 MEDIAS JUNTO1 ; BY SUJEITO;
K=_FREQ_; DROP _FREQ_;
SIGMAX=SIGMAT + SIGMAE/K;
V1 = (SQRT(SIGMAX-SIGMAE/K))/(SQRT(SIGMAX));
MG=GERAL; DROP GERAL;
TI = V1*(MS-MG) + MG;
alfa = sigmat/(sigmat + sigmae/k);
yma=alfa*ms + (1-alfa)*mg;

RUN;

OPTIONS PS=100;
PROC PRINT DATA=JUNTO2; run;

title verificar caracteristica do blup;
proc means data=junto2 mean noprint ; by sujeito;
var  ms ti yma;
output out=verificar mean=ms ti yma;
run;
proc print data=verificar; run;

proc means data=junto2 mean ;
var  ms ;
```

```
run;


TITLE VOLTA AO ORIGINAL(OUTRA TRANSFOMAÇÃO DIFERENTE DO LOG)  ;
DATA NORMAL; SET JUNTO2;
  V=0;
  DO YN = -9 TO 9 BY 0.005;
  FY=(J*(TI + YN)+1)**(1/J);
  PHI=FY*((1/SQRT(2*3.1416*SIGMAE))*(EXP(-(YN**2)/(2*SIGMAE))))*0.005;
  V=V+PHI ;
            END; RUN;
/*PROC PRINT DATA=NORMAL;
VAR Z WJ J FY PHI V ;
RUN; */

DATA BACK; SET NORMAL;
VOLTA= V- WJ; RUN;
/*PROC PRINT DATA=BACK;
VAR  WJ J  SUJEITO TIPO  Z TI VOLTA;
RUN;*/

PROC SORT  DATA=BACK ; BY SUJEITO; RUN;
PROC MEANS DATA=BACK MEAN NOPRINT; BY SUJEITO;
 VAR Z Volta TI;
OUTPUT OUT=FINAL4 MEAN=Z Z_C TI;
RUN;

DATA FINAL5; SET FINAL4;
  &NOME1=Z;
   &NOME2=Z_C;

   KEEP SUJEITO TIPO   &NOME1 &NOME2  ;   RUN;

PROC SORT DATA=FINAL5; BY SUJEITO; RUN;   PROC PRINT DATA=FINAL5; RUN;
TITLE VALORES CORRIGIDOS;
/*PROC PRINT DATA=FINAL5; RUN;*/


PROC SORT DATA=TUDO; BY SUJEITO; RUN;
PROC SORT DATA=FINAL5; BY SUJEITO; RUN;
DATA TUDO; MERGE TUDO FINAL5; BY SUJEITO; RUN;

PROC PRINT DATA=TUDO;

 RUN;
 QUIT;

/*****************************************************************/

DATA NUTRI.BELEM_R; SET TUDO ; RUN;

PROC PRINT DATA=NUTRI.BELEM_R; ; RUN;

/*******************************************************/

TITLE VOLTA AO ORIGINAL SE A TRANSFORMAÇÃO FOI LOG(EXATA);
DATA BLOG; SET JUNTO2;
VOLTA= EXP(TI + SIGMAE/2);
VOLTAblup= EXP(yma + SIGMAE/2);
RUN;
```

```
PROC PRINT DATA=BLOG;
VAR SUJEITO TIPO WJ  Z TI  VOLTA VOLTABLUP; ;
RUN;

PROC MEANS DATA=BLOG MEAN ;
var z;
run;
quit;

PROC MEANS DATA=BLOG MEAN NOPRINT; BY SUJEITO;
VAR Z  VOLTA VOLTABLUP TI;
OUTPUT OUT=FINAL1 MEAN=Z Z_C ZBLUP TI;
RUN;
DATA FINAL5; SET FINAL1;
 &NOME1=Z;
   &NOME2=Z_C;
   KEEP SUJEITO TIPO   &NOME1 &NOME2 ZBLUP;
   PROC PRINT DATA=FINAL5; RUN;

TITLE CAROTENO :  MEDIAS ORIGINAIS E AJUSTADAS;
PROC SORT DATA=FINAL5; BY SUJEITO; RUN;
PROC PRINT DATA=FINAL5;
VAR  SUJEITO N2   N2_C ZBLUP ;
RUN;

DATA MEDIAS_3; SET FINAL5;
N2_MEDIA=N2;
N2_NUSSER=N2_C;
N2_BLUP=ZBLUP;
DROP N2 N2_C ZBLUP;
PROC PRINT DATA=MEDIAS_3; RUN;

DATA NUTRI.MEDIAS_TODAS1; SET MEDIAS_3;
PROC PRINT DATA=NUTRI.MEDIAS_TODAS1; RUN;

LIBNAME BOOTSN "H:\KINGSTON\ANA LUISA\BOOTSTRAP"; RUN;
DATA BOOT1; SET BOOTSN.MEDIAS;
SUJEITO=INDIVIDUO;
KEEP SUJEITO N2_ N2_NUSSERB  N2_BLUPB;
PROC PRINT DATA=BOOT1; RUN;


PROC SORT DATA=NUTRI.MEDIAS_TODAS1; BY SUJEITO; RUN;
PROC SORT DATA=BOOT1; BY SUJEITO; RUN;

DATA BOOTSN.MEDIAS_TODAS; MERGE NUTRI.MEDIAS_TODAS1 BOOT1; BY SUJEITO;
PROC PRINT DATA=BOOTSN.MEDIAS_TODAS; RUN;

TITLE CALCULO DOS PERCENTIS;
PROC UNIVARIATE DATA=BOOTSN.MEDIAS_TODAS;
VAR  N2_MEDIA   N2_NUSSER   N2_BLUP   N2_NUSSERB   N2_BLUPB;
OUTPUT OUT=LIMITES25 PCTLPRE=N2_MEDIA   N2_NUSSER   N2_BLUP
N2_NUSSERB N2_BLUPB
PCTLPTS=25;
RUN;
PROC PRINT DATA=LIMITES25; RUN;

/* MACRO PARA PERCENTIL*/

DATA TUDO;
INPUT  PORC $ N2_MEDIA N2_NUSSER N2_BLUP N2_NUSSERB N2_BLUPB;
```

```
DATALINES;
0 0 0 0 0 0
;
RUN ;
%MACRO PERCENTIL;

    %DO P =15 %TO 85 %BY 5;    /* DO PARA LER CADA VARIAVEL */

 TITLE CALCULO DOS PERCENTIS;
PROC UNIVARIATE DATA=BOOTSN.MEDIAS_TODAS NOPRINT;
VAR  N2_MEDIA   N2_NUSSER   N2_BLUP   N2_NUSSERB   N2_BLUPB;
OUTPUT OUT=LIMITES&P PCTLPRE=N2_MEDIA   N2_NUSSER   N2_BLUP
N2_NUSSERB N2_BLUPB
PCTLPTS=&P; RUN;
DATA PER&P; SET
LIMITES&P;
PORC="PERC_&P";
ARRAY ORI {5}  N2_MEDIA&P N2_NUSSER&P N2_BLUP&P N2_NUSSERB&P
N2_BLUPB&P;
ARRAY PER {5} N2_MEDIA N2_NUSSER N2_BLUP N2_NUSSERB  N2_BLUPB;
DO J=1 TO 5;
PER(J)=ORI[J]; END;
DROP J N2_MEDIA&P N2_NUSSER&P N2_BLUP&P N2_NUSSERB&P  N2_BLUPB&P;RUN;
DATA TUDO; SET TUDO PER&P;
 %END;
%MEND PERCENTIL;                  /* FIM DO MACRO */
RUN;
%PERCENTIL

PROC PRINT  DATA=TUDO; RUN;

TITLE MEDIAS;
OPTIONS LS=90 NODATE NONUMBER;
DATA PERCENTI; SET TUDO;
IF PORC="0    " THEN DELETE;

PROC PRINT NOOBS DATA=PERCENTI; RUN;


/****************************************/
/* PROGRAMA PARA ESTIMAR MEDIA BOOTSTRAP*/
/****************************************/


LIBNAME NUTRI_E "H:\KINGSTON\TERESA\ELIENE"; RUN;
DATA NUTRI_E.ELIENE1; SET ELIENE; RUN;
OPTIONS LS=160;
PROC PRINT DATA=NUTRI_E.ELIENE1; RUN;

/* ESTIMATIVA BOOTSTRAP*/

TITLE SELEÇÃO DE UMA VARIAVEL;
OPTIONS LS=90 NODATE NONUMBER;
DATA ELIENE2; SET NUTRI_E.ELIENE1;
KEEP Individuo TIPO   N2 ;
IF TIPO < 3;
PROC PRINT DATA=ELIENE2; RUN;

TITLE MEDIA E FREQUENCIA DE  CADA INDIVIDUO;
PROC SORT DATA=ELIENE2 ;BY INDIVIDUO;
PROC MEANS   DATA=ELIENE2 MEAN NOPRINT;BY INDIVIDUO;
```

```
VAR N2;
OUTPUT OUT=MEDIA_IND MEAN=N2_MEDIA;
RUN;
QUIT;
PROC PRINT DATA=MEDIA_IND; RUN; /* DEFINI OS GRUPOS*/

/* PROC MEANS DATA=MEDIA_IND MEAN;
VAR _FREQ_;
OUTPUT OUT=FREQ_MEDIA  MEAN=TIPO_MED;
RUN;QUIT;
PROC PRINT DATA=FREQ_MEDIA ; RUN;
DATA FREQ1_MEDIA; SET MEDIA_IND;
IF _FREQ_ < 3;
PROC PRINT DATA=FREQ1_MEDIA ; RUN; /* TIPO =2*/


TITLE SELEÇÃO DE UM TESTE PARA O BOOTSTRAP MEDIA(PRIMEIRO NIVEL);
DATA ELIENEB1; SET MEDIA_IND;
C+1;
/*IF C < 31;*/
DROP C;
PROC PRINT DATA=ELIENEB1; RUN;

title dados nulos;
DATA MEDIA_BOOT1; SET ELIENEB1; /* dados nulos*/
  N2_MEDIA1=0 ;  N2_SHR1=0;  N2_SHR_FUL1=0;  N2_CBOOT=0; PRE1=0;
KEEP Individuo N2_MEDIA1  N2_SHR1  N2_SHR_FUL1 N2_CBOOT  PRE1;
PROC PRINT DATA=MEDIA_BOOT1; RUN;

TITLE DADOS COMPLETOS DOS INDIVIDUOS SELECIONADOS;
DATA ELIENE3; MERGE ELIENEB1 ELIENE2; BY INDIVIDUO;
IF _TYPE_=. THEN DELETE;
PROC PRINT DATA=ELIENE3; RUN;

TITLE ESTIMATIVA DAS VARIANCIAS -  ANOVA;
PROC VARCOMP DATA=ELIENE3 METHOD=TYPE1 ; CLASS INDIVIDUO;
MODEL N2=INDIVIDUO;
ODS OUTPUT ESTIMATES=VARD12;
RUN;
QUIT;
PROC PRINT DATA=VARD12; RUN;


TITLE ESTIMATIVA DAS VARIANCIAS -  MVR;
PROC MIXED DATA=ELIENE3  ;
CLASS INDIVIDUO; MODEL N2=;
RANDOM INDIVIDUO;
RUN;
QUIT;


DATA VAR1; SET VARD12;
IF VARCOMP="Var(Individuo)";
VAR_IND=ESTIMATE;
KEEP VAR_IND;
PROC PRINT DATA=VAR1; RUN;

DATA VAR2; SET VARD12;
IF  VARCOMP="Var(Error)";
VAR_ERRO=ESTIMATE;
KEEP VAR_ERRO;
PROC PRINT DATA=VAR2; RUN;
```

```
DATA VAR_12; MERGE VAR1 VAR2;
PROC PRINT DATA=VAR_12; RUN;


TITLE DATA SET PARA SOMAS ACUMULADAS;
DATA MEDIA_BOOT1; SET ELIENEB1;
N2_MEDIA1=0 ;  N2_BLUP1=0;  N2_NUSSER1=0;   PRE1=0; KEEP
Individuo N2_MEDIA1  N2_BLUP1  N2_NUSSER1  PRE1;
PROC PRINT DATA=MEDIA_BOOT1; RUN;

proc print data=ELIENEB1;RUN;
TITLE SELECAO DA AMOSTRA COM REPOSIÇÃO (PRIMEIRO NIVEL HINKLEY);
PROC SURVEYSELECT  DATA=ELIENEB1 METHOD=URS SAMPSIZE=30 OUT=BOOT1
seed = 2078
NOPRINT;
RUN;
PROC PRINT DATA=BOOT1; RUN;

TITLE RECUPERAR SEGUNDO NIVEL NA AMOSTRA SEM REPETIÇÃO DOS DADOS;
DATA TESTE1; MERGE BOOT1 ELIENE2; BY INDIVIDUO;
IF NUMBERHITS=. THEN DELETE;
PROC PRINT DATA=TESTE1 ;RUN;
TITLE RECUPERAR SEGUNDO NIVEL NA AMOSTRA COM REPETIÇÃO DOS DADOS;
DATA TESTE2; SET TESTE1;
X=NUMBERHITS;
DO J = 1 TO X;
OUTPUT;
END;
PROC PRINT DATA=TESTE2; RUN;
PROC SORT DATA=TESTE2; BY INDIVIDUO TIPO;
PROC PRINT DATA=TESTE2; RUN;

TITLE ESTIMAR VARIANCIAS COM PESOS - MIXED- RAO;
PROC MIXED DATA=TESTE2  method= MIVQUE0 ;
CLASS INDIVIDUO;
MODEL N2=;
RANDOM INDIVIDUO;
ODS SELECT COVPARMS;
ODS OUTPUT CovParms=VARDB12;
RUN;
QUIT;
PROC PRINT DATA=VARDB12; RUN;

DATA VAR1B; SET VARDB12;
IF COVPARM="Individuo";
VAR_IND=ESTIMATE;
KEEP VAR_IND;
PROC PRINT DATA=VAR1B; RUN;

DATA VAR2B; SET VARDB12;
IF COVPARM="Residual";
VAR_ERRO=ESTIMATE;
KEEP VAR_ERRO;
PROC PRINT DATA=VAR2B; RUN;

TITLE VARIANCIAS ESTIMADAS;
DATA VARB_12; MERGE VAR1B VAR2B;
PROC PRINT DATA=VARB_12; RUN;

PROC PRINT DATA=BOOT1; RUN;
```

```
TITLE CALCULO DA MEDIA GERAL BOOTSTRAP;
PROC MEANS DATA=BOOT1 MEAN NOPRINT ;
VAR N2_MEDIA;
WEIGHT NUMBERHITS;
OUTPUT OUT=GERAL1 MEAN=N_GERAL;
RUN;
QUIT;
PROC PRINT DATA=GERAL1; RUN;

DATA GERAL ; SET GERAL1;
KEEP N_GERAL;
PROC PRINT DATA=GERAL; RUN;

TITLE VARIANCIAS E MEDIA GERAL;
DATA VAR_MG; MERGE VARB_12 GERAL;
PROC PRINT DATA=VAR_MG; RUN;

PROC PRINT DATA=BOOT1; RUN;
TITLE SHRINKAGE ESTIMATOR;
DATA TUDO1; MERGE BOOT1 VAR_MG;
C1 + VAR_IND;
C2 + VAR_ERRO;
C3 + N_GERAL;
VAR_IND=C1;
VAR_ERRO=C2;
N_GERAL=C3;
DROP C1 C2 C3;
BETA=VAR_ERRO/(VAR_ERRO + _FREQ_*VAR_IND);
N2_BLUP=(1-BETA)*N2_MEDIA + BETA*N_GERAL;
BETA_FUL=SQRT(BETA);
N2_NUSSER=(1-BETA_FUL)*N2_MEDIA + BETA_FUL*N_GERAL;
PROC PRINT DATA=TUDO1;
VAR INDIVIDUO N2_MEDIA N2_BLUP N2_NUSSER BETA BETA_FUL N_GERAL;
RUN;

TITLE JUNTAR OS DOIS DATAS;
OPTIONS LS=130;
DATA MEDIA_BOOT2; MERGE  MEDIA_BOOT1 TUDO1; BY INDIVIDUO;
PRE=1;
IF VAR_IND=. THEN PRE=0;
DROP BETA BETA_FUL VAR_IND VAR_ERRO N_GERAL;
IF N2_MEDIA=. THEN  N2_MEDIA=0 ;
IF N2_BLUP=. THEN N2_BLUP=0 ;
IF N2_NUSSER=. THEN N2_NUSSER=0 ;
PROC  PRINT DATA=MEDIA_BOOT2;  RUN;

TITLE SOMAR OS RESULTADOS;
DATA SOMAB; SET MEDIA_BOOT2;
N2_MEDIA1=N2_MEDIA1 + N2_MEDIA;
N2_BLUP1=N2_BLUP1 + N2_BLUP;
N2_NUSSER1=N2_NUSSER1 + N2_NUSSER;
PRE1=PRE1 + PRE;

PROC PRINT DATA=SOMAB; RUN;


TITLE DADOS SOMADOS ATE O MOMENTO;
DATA MEDIA_BOOT2; SET SOMAB;
KEEP  Individuo N2_MEDIA1  N2_BLUP1  N2_NUSSER1   PRE1;
PROC  PRINT DATA=MEDIA_BOOT2; RUN;
```

```
DATA NUTRI_E.MEDIAB; SET MEDIA_BOOT1;
PROC PRINT DATA=NUTRI_E.MEDIAB; RUN;

PROC  PRINT DATA=MEDIA_BOOT1; RUN;

options nosource nosource2 NONOTES nostimer nomprint nosymbolgen
nomlogic;
       PROC PRINTTO NEW UNIT=20;

%MACRO GERAL1M;

options nosource nosource2 NONOTES nostimer nomprint nosymbolgen
       nomlogic; PROC PRINTTO NEW UNIT=20;

TITLE DATA SET PARA SOMAS ACUMULADAS;
DATA MEDIA_BOOT2; SET ELIENEB1;
N2_MEDIA1=0 ;  N2_BLUP1=0;  N2_NUSSER1=0;   PRE1=0;
KEEP Individuo N2_MEDIA1  N2_BLUP1  N2_NUSSER1   PRE1;
PROC PRINT DATA=MEDIA_BOOT2; RUN;


          %DO AMOSTRA = 1 %TO 10000;
   TITLE SELECAO DA AMOSTRA COM REPOSIÇÃO (PRIMEIRO NIVEL HINKLEY);
PROC SURVEYSELECT  DATA=ELIENEB1 METHOD=URS SAMPSIZE=326 OUT=BOOT1
seed =511&AMOSTRA NOPRINT ;
RUN;
PROC PRINT DATA=BOOT1; RUN;

TITLE RECUPERAR SEGUNDO NIVEL NA AMOSTRA SEM REPETIÇÃO DOS DADOS;
DATA TESTE1; MERGE BOOT1 ELIENE2; BY INDIVIDUO;
IF NUMBERHITS=. THEN DELETE;


TITLE RECUPERAR SEGUNDO NIVEL NA AMOSTRA COM REPETIÇÃO DOS DADOS;
DATA TESTE2; SET TESTE1;
X=NUMBERHITS;
DO J = 1 TO X;
OUTPUT;
END;
PROC PRINT DATA=TESTE2; RUN;
PROC SORT DATA=TESTE2; BY INDIVIDUO TIPO;
PROC PRINT DATA=TESTE2; RUN;

TITLE ESTIMAR VARIANCIAS COM PESOS - MIXED- RAO;
PROC MIXED DATA=TESTE2   ;
CLASS INDIVIDUO;
MODEL N2=;
RANDOM INDIVIDUO;
ODS SELECT COVPARMS;
ODS OUTPUT CovParms=VARDB12;

RUN;
QUIT;
RUN;

DATA VAR1B; SET VARDB12;
IF COVPARM="Individuo";
VAR_IND=ESTIMATE;
KEEP VAR_IND; RUN;
```

```
 DATA VAR2B; SET VARDB12;
 IF COVPARM="Residual";
 VAR_ERRO=ESTIMATE;
 KEEP VAR_ERRO; RUN;

 TITLE VARIANCIAS ESTIMADAS;
 DATA VARB_12; MERGE VAR1B VAR2B;
  RUN;
    PROC PRINT DATA=VARB_12; RUN;

 TITLE CALCULO DA MEDIA GERAL BOOTSTRAP;
 PROC MEANS DATA=BOOT1 MEAN NOPRINT ;
 VAR N2_MEDIA;
 WEIGHT NUMBERHITS;
 OUTPUT OUT=GERAL1 MEAN=N_GERAL;
 RUN;
 QUIT;
 RUN;

 DATA GERAL ; SET GERAL1;
 KEEP N_GERAL;
 RUN;

 TITLE VARIANCIAS E MEDIA GERAL;
 DATA VAR_MG; MERGE VARB_12 GERAL;
 RUN;
 PROC PRINT DATA=VAR_MG; RUN;

 TITLE SHRINKAGE ESTIMATOR;
DATA TUDO1; MERGE BOOT1 VAR_MG;

C1 + VAR_IND;

 C2 + VAR_ERRO;
 C3 + N_GERAL;
 VAR_IND=C1;
 VAR_ERRO=C2;
 N_GERAL=C3;
 DROP C1 C2 C3;
 BETA=VAR_ERRO/(VAR_ERRO + _FREQ_*VAR_IND);
 N2_BLUP=(1-BETA)*N2_MEDIA + BETA*N_GERAL;
 BETA_FUL=SQRT(BETA);
 N2_NUSSER=(1-BETA_FUL)*N2_MEDIA + BETA_FUL*N_GERAL;
 PROC PRINT DATA=TUDO1;
 VAR INDIVIDUO N2_MEDIA N2_BLUP N2_NUSSER BETA BETA_FUL N_GERAL;
 RUN;

 TITLE JUNTAR OS DOIS DATAS;
 OPTIONS LS=130;
 DATA MEDIA_BOOT2; MERGE  MEDIA_BOOT2 TUDO1; BY INDIVIDUO;
 PRE=1;
 IF VAR_IND=. THEN PRE=0;
 DROP BETA BETA_FUL VAR_IND VAR_ERRO ;
 IF N2_MEDIA=. THEN  N2_MEDIA=0 ;
 IF N2_BLUP=. THEN N2_BLUP=0 ;
 IF N2_NUSSER=. THEN N2_NUSSER=0 ;
 PROC  PRINT DATA=MEDIA_BOOT2;  RUN;

 TITLE SOMAR OS RESULTADOS;
 DATA SOMAB; SET MEDIA_BOOT2;
 N2_MEDIA1=N2_MEDIA1 + N2_MEDIA;
```

```
N2_BLUP1=N2_BLUP1 + N2_BLUP;
N2_NUSSER1=N2_NUSSER1 + N2_NUSSER;
PRE1=PRE1 + PRE;
PROC PRINT DATA=SOMAB; RUN;

DATA MEDIA_BOOT2; SET SOMAB;
KEEP Individuo N2_MEDIA1  N2_BLUP1  N2_NUSSER1   PRE1;
PROC PRINT DATA=MEDIA_BOOT2; RUN;
%END;

%MEND GERAL1M;
 RUN;


  %GERAL1M

options source source2 NOTES stimer ;
   PROC PRINTTO ;RUN;
PROC PRINT DATA=MEDIA_BOOT2; RUN;

TITLE MEDIAS INDIVIDUAIS BOOTSTRAP;

DATA MEDIASB_BOOT; SET MEDIA_BOOT2;

N2_NUSSERB=N2_NUSSER1/PRE1;

N2_BLUPB=N2_BLUP1/PRE1;

N2_MEDIA1T=N2_MEDIA1/PRE1;

  PROC PRINT DATA=MEDIASB_BOOT;
  RUN;

  LIBNAME BOOTSN "H:\KINGSTON\ANA LUIZA\BOOTSTRAP"; RUN;
  DATA BOOTSN.MEDIAS; SET MEDIASB_BOOT; RUN;
PROC PRINT DATA=BOOTSN.MEDIAS; RUN;
```