Modelo Dirichlet-Multinomial para respostas categorizadas

Damaris de Sá Motta Regina [1], Tuany de Paula Castro [1], Julio Singer [1]

Claudia R. Furquim de Andrade [2], Talita Fortunato Tavares [2]

[1] Instituto de Matemática e Estatística- IME/USP
[2] Faculdade de Medicina da Universidade de São Paulo - FM/USP

Resumo: Utilizamos um modelo Dirichlet-Multinomial para comparar vetores de probabilidades de ensaios multinomiais realizados por unidades amostrais sob diferentes tratamentos. Indicamos como os resultados podem ser implementados no pacote R e ilustramos a metodologia por meio de um exemplo na área de Fonoaudiologia.

Palavras-chave: análise de dados categorizados, modelo Dirichlet-Multinomial.

1. INTRODUÇÃO

Em muitas situações, há interesse em comparar unidades amostrais sob diferentes tratamentos quanto aos vetores de probabilidades de ensaios multinomiais. Por exemplo, em um estudo realizado na Faculdade de Medicina da Universidade de São Paulo em 2012 [Fortunato-Tavares et al. (2012)], em cada uma de 60 crianças, 15 com desenvolvimento típico de linguagem (DTL), 15 com distúrbio específico de linguagem (DEL), 15 com desordens do espectro autista (DEA) e 15 com síndrome de Down (SD), foram aplicados 26 testes e cada uma de suas respostas foi avaliada como correta (C), incorreta com resposta hierárquica (H), incorreta com troca de posição (P) ou incorreta com resposta reversa (R). A finalidade do estudo era comparar os vetores de probabilidades de categorias de respostas entre as crianças de diferentes grupos.

2. METODOLOGIA

Seja o vetor de respostas $\boldsymbol{n}_{(ij)} = (n_{(ij)1}, n_{(ij)2}, n_{(ij)3}, n_{(ij)4})$, em que $n_{(ij)k}$ representa o número de questões classificadas na categoria k, para o questionário da i-ésima criança do grupo j e seja $\boldsymbol{\pi}_{(ij)} = (\pi_{(ij)1}, \pi_{(ij)2}, \pi_{(ij)3}, \pi_{(ij)4})$ um vetor de parâmetros, tais que $\pi_{(ij)k} > 0$ e

$\sum_{k=1}^{4} \pi_{(ij)k} = 1$. $i = 1, ..., 15$; $j = 1, 2, 3, 4$ e $k = 1, 2, 3, 4$, com:

$$j = \begin{cases} 1, & \text{se grupo DTL} \\ 2, & \text{se grupo DEL} \\ 3, & \text{se grupo DEA} \\ 4, & \text{se grupo SD} \end{cases} \qquad k = \begin{cases} 1, & \text{se resposta correta} \\ 2, & \text{se resposta incorreta hierárquica} \\ 3, & \text{se resposta incorreta com troca de posição} \\ 4, & \text{se resposta incorreta reversa} \end{cases} \tag{1}$$

Podemos admitir que, para cada criança, os dados obedecem a uma distribuição Multinomial com parâmetros n = 26 e vetor de probabilidades $\boldsymbol{\pi}_{(ij)}$ em que o elemento $\pi_{(ij)k}$ corresponde à probabilidade de uma questão selecionada ao acaso do questionário aplicado à i-ésima criança do grupo j ter resposta classificada na categoria k.

Caso o interesse de comparação entre os grupos fosse restrito às unidades amostrais, isto é, considerando fixo o efeito de crianças, as comparações dos vetores de probabilidades poderiam ser feitas por meio de um ajuste de um modelo produto de Multinomiais [Paulino e Singer (2006)]. Entretanto, como o objetivo é expandir as conclusões obtidas para a população (conceitual) de crianças com as características indicadas acima, pode-se levar em conta a aleatoriedade da amostra por meio de um modelo hierárquico. Com essa finalidade, consideramos um modelo em que:

$$\begin{gathered} \boldsymbol{n}_{(ij)} | \boldsymbol{\pi}_{(ij)} \sim Multinomial(n, \boldsymbol{\pi}_{(ij)}), \text{ independentes.} \\ \boldsymbol{\pi}_{(ij)} | (\nu_j, \boldsymbol{\alpha}_j) \sim Dirichlet(\nu_j, \boldsymbol{\alpha}_j), \text{ independentes,} \end{gathered} \tag{2}$$

ou seja, $\boldsymbol{\pi}_{(ij)}$ tem função densidade de probabilidade:

$$f(\boldsymbol{\pi}_{(ij)} | (\nu_j, \boldsymbol{\alpha}_j)) = \Gamma(\nu_j) \prod_{k=1}^{4} \frac{(\pi_{(ij)k})^{\nu_j \alpha_{jk} - 1}}{\Gamma(\nu_j \alpha_{jk})}, \tag{3}$$

com $\boldsymbol{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, \alpha_{j3}, \alpha_{j4})$, j = 1, 2, 3, 4, tais que $\alpha_{jk} > 0$ e $\sum_{k=1}^{4} \alpha_{jk} = 1$.

Sob essa parametrização da distribuição Dirichlet, ν_j é o parâmetro de precisão para o grupo j e $\alpha_{jk} = \mathbb{E}[\pi_{(ij)k}]$.

Essencialmente, o modelo implica que os dados de cada criança, dado o correspondente vetor de probabilidades de categorias de resposta, segue uma distribuição Multinomial e esse vetor de probabilidades segue uma distribuição Dirichlet com mesmos parâmetros para

crianças de mesmo grupo. Dessa forma, o vetor de respostas $\boldsymbol{n}_{(ij)}$ segue uma distribuição Dirichlet-Multinomial denotada por:

$$\boldsymbol{n}_{(ij)} \sim Dirichlet - Multinomial(n, \nu_j, \boldsymbol{\alpha}_j) \tag{4}$$

cuja função densidade de probabilidade é:

$$f(\boldsymbol{n}_{(ij)}|n, \nu_j, \boldsymbol{\alpha}_j) = \begin{pmatrix} n \\ \boldsymbol{n}_{(ij)} \end{pmatrix} \frac{\Gamma(\nu_j)}{\Gamma(n+\nu_j)} \prod_{k=1}^{4} \frac{\Gamma(n_{(ij)k} + \nu_j \alpha_{jk})}{\Gamma(\nu_j \alpha_{jk})} \tag{5}$$

Este modelo é uma generalização do modelo Beta-Binomial [Agresti (2002)], o qual seria utilizado no caso de apenas duas categorias de resposta.

A correspondente função de verossimilhança é:

$$L(\nu_j, \boldsymbol{\alpha}_j|n, \boldsymbol{n}_{(ij)}) \propto \prod_{j=1}^{4} \prod_{i=1}^{15} \left(\frac{\Gamma(\nu_j)}{\Gamma(n+\nu_j)} \prod_{k=1}^{4} \frac{\Gamma(n_{(ij)k} + \nu_j \alpha_{jk})}{\Gamma(\nu_j \alpha_{jk})} \right) \tag{6}$$

Os estimadores de máxima verossimilhança para ν_j e α_{jk} podem ser obtidos resolvendo $U(\nu_j, \alpha_{jk}) = 0$ por métodos numéricos [Minka (2012)], em que:

$$U(\nu_j, \alpha_{jk}) = \begin{pmatrix} \frac{d}{d\nu_j} \log L(\nu_j, \boldsymbol{\alpha}_j|n, \boldsymbol{n}_{(ij)}) \\ \frac{d}{d\alpha_{jk}} \log L(\nu_j, \boldsymbol{\alpha}_j|n, \boldsymbol{n}_{(ij)}) \end{pmatrix}, \text{com:} \tag{7}$$

$$\frac{d}{d\nu_j} \log L(\nu_j, \boldsymbol{\alpha}_j|n, \boldsymbol{n}_{(ij)}) = \sum_{i=1}^{15} \left(\psi(\nu_j) - \psi(n+\nu_j) + \sum_{k=1}^{4} \alpha_{jk} \left(\psi(n_{(ij)k} + \nu_j \alpha_{jk}) - \psi(\nu_j \alpha_{jk}) \right) \right) \tag{8}$$

$$\frac{d}{d\alpha_{jk}} \log L(\nu_j, \boldsymbol{\alpha}_j|n, \boldsymbol{n}_{(ij)}) = \sum_{i=1}^{15} \nu_j \left(\psi(n_{(ij)k} + \nu_j \alpha_{jk}) - \psi(\nu_j \alpha_{jk}) \right) \tag{9}$$

e $\psi(x)$ é a função digama dada por $\frac{d}{dx} \log \Gamma(x)$.

A comparação dos vetores esperados de probabilidades foi efetuada por meio de testes de Wald.

A implementação computacional do método de máxima verossimilhança foi realizada no *software* R por meio da função vglm do pacote VGAM. Os comandos para sua aplicação no exemplo considerado podem ser encontrados em www.ime.usp.br/∼jmsinger.

3. RESULTADOS

Na comparação das crianças com DTL, DEL, DEA e SD quanto às frequências esperadas de respostas nas categorias C, H, P e R, conjuntamente, obteve-se p $<$ 0,001, indicando que os grupos não são homogêneos com relação ao vetor esperado de probabilidades, o que pode ser observado na Tabela 3.1.

Tabela 3.1. Comparação dos grupos quanto às frequências esperadas de respostas conjuntamente.

Resposta	**g.l.**	χ^2	**Valor p**
C, H, P, R	9	417,53	$<$ 0,001

Com a finalidade de identificar as diferenças entre os grupos para cada categoria de resposta, construímos os testes apresentados na Tabela 3.2.

Tabela 3.2. Comparação dos grupos quanto às frequências esperadas de respostas para cada categoria.

Resposta	**g.l.**	χ^2	**Valor p**
C	3	352,28	$<$ 0,001
H	3	97,44	$<$ 0,001
P	3	118,00	$<$ 0,001
R	3	854,19	$<$ 0,001

Em todas as categorias de resposta, houve evidências para rejeitar a hipótese de igualdade entre as frequências esperadas para os diferentes grupos. Intervalos de confiança para a diferença entre os grupos DEL, DEA e SD em relação ao DTL quanto aos vetores de probabilidades esperadas $(\boldsymbol{\alpha_j} - \boldsymbol{\alpha_1}, j = 2, 3, 4)$ estão apresentados na Figura 1.

Frequências relativas esperadas para cada grupo, sob o modelo Dirichlet-Multinomial, são apresentadas na Tabela 3.3.

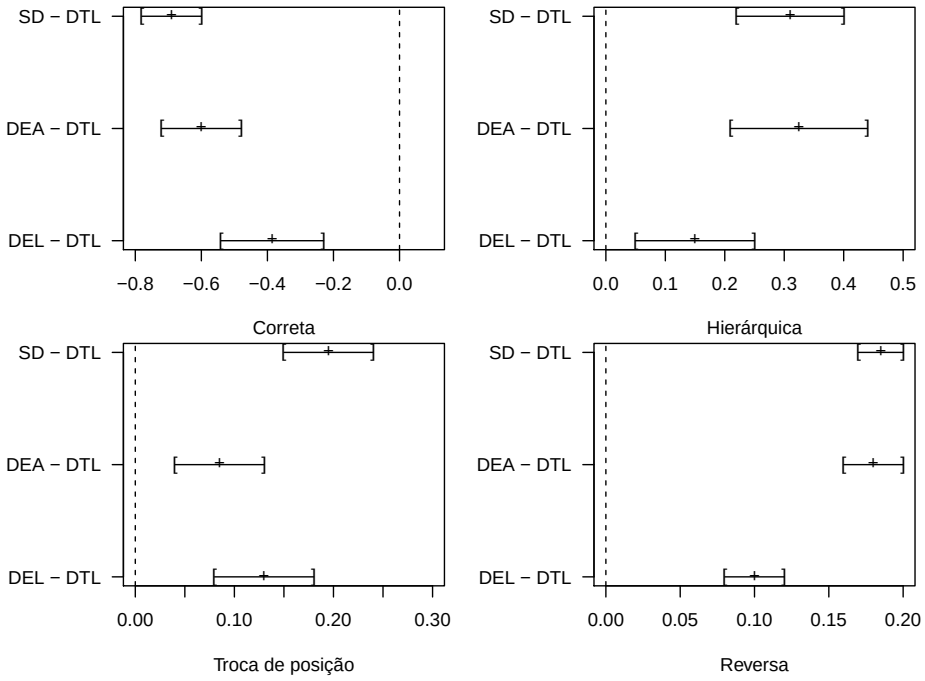

Figura 1: Intervalos de confiança para a diferença de probabilidade esperada para cada categoria de resposta

Tabela 3.3. Frequências de respostas esperadas sob o modelo.

Grupo	Certa(C)	Hierárquica (H)	Com troca de preposição (P)	Reversa (R)	Total
DTL	95%	2%	2%	1%	100%
DEL	56%	17%	15%	12%	100%
DEA	35%	35%	11%	19%	100%
SD	26%	33%	21%	20%	100%

4. DISCUSSÃO

O modelo Dirichlet-Multinomial proposto é adequado para comparar vetores de probabilidades de ensaios multinomiais realizados por unidades amostrais sob diferentes tratamentos, entretanto, é importante lembrar que os resultados são aproximados, uma vez que os estimadores de máxima verossimilhança para os parâmetros seguem assintoticamente distribuição Normal. Dessa forma, tanto os intervalos de confiança quanto os testes de Wald são assintóticos e, portanto, mais confiáveis conforme maior o tamanho da amostra.

Uma abordagem bayesiana também poderia utilizada para este fim, quando se tem algum

conhecimento a priori do comportamento dos parâmetros da distribuição Dirichlet.

REFERÊNCIAS

Agresti, A. (2002). *Categorical Data Analysis.* 2 ed. New Jersey: John Wiley & Sons

Fortunato-Tavares, T., Andrade, C. R. F., Befi-Lopes, D. M., Hestvik, A., Epstein, B., Tornyova, L., e Schwartz, R. G. (in press). *Syntactic structural assignment in Brazilian Portuguese-speaking children with specific language impairment.* Journal of Speech, Language, Hearing Research.

Minka, T.P. (2012). *Estimating a Dirichlet Distribution.* http://research.microsoft.com/en-us/um/people/minka/papers/dirichlet/minka-dirichlet.pdf

Paulino, C.D. e Singer, J.M. (2006). *Análise de dados categorizados.* São Paulo: Blücher