XXX Congresso de Iniciação Científica UNICAMP

O uso de U-estatísticas em testes de homogeneidade entre grupos: aplicações em genética e educação.

Palavras-chave: U-estatística; Homogeneidade; Sequências genômicas; Desempenho Acadêmico.

Aluno: Bruno Martinez de Farias [IMECC],
Orientadora: Hildete Prisco Pinheiro [IMECC]

Julho de 2022

1 Descrição

O presente projeto de pesquisa teve como objeto de estudo, a teoria de U-estatísticas que embasa a construção de uma medida de diversidade entre grupos, proposta por Pinheiro et al. (2005) como ferramenta para testar a homogeneidade entre grupos, sendo esse o principal objetivo nessa pesquisa.

Dados grupos $g, g' \in \{1, \ldots, G\}$, temos $\mathcal{H}_{gg}$ como medida de diversidade populacional média dentre grupos e $\mathcal{H}_{gg'}$ como medida de diversisdade populacional entre os grupos. Essas medidas podem ser estimadas respectivamete por

$$\overline{D}_{gg} = \binom{n_g}{2}^{-1} \sum_{1 \le i < j \le n_g} D_{ij}^{g} \tag{1}$$

e

$$\overline{D}_{gg'} = \frac{1}{n_g n_{g'}} \sum_{i=1}^{n_g} \sum_{j=1}^{n_{g'}} D_{ij}^{gg'}, \tag{2}$$

em que n_g é o tamanho da amostra do g-ésimo grupo.

Dado que, $\overline{D}_{gg}$ é uma U-estatística de grau 2 e $\overline{D}_{gg'}$ é uma U-estatística bivariada de grau $(1,1)$, a diversidade geral total pode ser expressa como uma combinação linear de U-estatísticas. Dessa fora, poemos escrever $D_n^{(0)} = D_n(W) + D_n(B)$, em que $D_n(W)$ é a medida de diversidade dentre grupos e $D_n(B)$ é a medida de diversidade entre grupos e a equação (3) é a nossa estatística de teste dada por

$$D_n(B) = \frac{1}{n(n-1)} \sum_{g<g'} n_g n_{g'} (2\overline{D}_{gg'} - \overline{D}_{gg} - \overline{D}_{g'g'}), \tag{3}$$

com n sendo o tamanho total da amostra, i.e, $n = \sum_g n_g$. A hipótese nula de homogeneidade entre grupos é dada por $H_0 : 2\mathcal{H}_{gg'} - \mathcal{H}_{gg} - \mathcal{H}_{g'g'} = 0$. Sob a hipótese nula, $\mathbb{E}(D_n(B)) = 0$ e $\mathbb{E}(D_n^{(0)}) = D_n(W)$.

Em Pinheiro et al. (2009, 2011, 2012) foi mostrado que $D_n(B)$ possuí distribuição assintóticamente normal sob H_0 para amostras suficientemete grandes. Entretanto, a variância da estatística de teste é muito complexa de ser estimada. Como alternativa, foi utilizada a técnica de reamostragem por bootstrap para obter a distribuição empirica da estatistica de teste ep-valores.

No contexto da biologia, a medida de diversidade utilizada é uma extensão do Índice de Gini-Simpson comumente utilizada como medida de variação genética. Essa medida será estimada pela distância de Hamming dada por

$$D_{ij} = \frac{1}{S}\sum_{l}^{S} \mathbb{I}(X_{i_l} \neq X_{j_l}), \tag{4}$$

em que D_{ij} é a proporção de diferenças entre as sequencias X_i e X_j e $\mathbb{I}(.)$ é a função indicadora que indica quando existe diferença no l-ésimo sítio. Para a aplicação nesse contexto compararemos trechos das sequências do RNA de diferentes cepas/variantes do virús causador da Covid-19 obtidas no repositório GISAID.

No contexto da educação, a medida de diversidade é baseada na diferença quadrática da função de ganho relativo apresentada em Maia et al. (2016) definida por

$$D_{ij}^{(g)} = (X_i^{(g)} - X_j^{(g)})^2, \tag{5}$$

em que $X_i = \frac{Cr_i - Pv_i}{n_e}$, para o i-ésimo estudante temos Cr_i é a posição do estudante baseado no Cr, Pv_i é a posição do estudante baseada na nota do vestibular e n_e é o numero de estudantes no mesmo curso e ano de ingresso. Neste caso nosso interesse é comparar o desempenho acadêmico de estudantes de graduação da Unicamp utilizando dados disponibilizados pela DAC e pela COMVEST.

2 Aplicações e resultados

As sequências genômicas utilizadas nessa pesquisa são compostas por amostras aleatórias de 500 sequências de quatro grupos, três variantes $\boldsymbol{\alpha}$,$\boldsymbol{\delta}$ e $\boldsymbol{\gamma}$ e um quarto grupo genérico, obtidas até setembro de 2020. Foram consideradas sequências completas referentes a *Glicoproteína Spike* e cada uma das 2000 sequências tem mesmo comprimento de 3822 nucleotídeos.

Para a amostra da variante $\boldsymbol{\alpha}$, identificamos 217 sítios segregantes, que significa que nesses sítios ao menos uma mudança ocorrou referente ao nucleotídeo da sequência de referência (sequência **WIV04** P. Zhou et al.). Analogamente, para a amostra composta pela variante $\boldsymbol{\delta}$, identificamos 261 sítios segregantes. Para a amostra da variante $\boldsymbol{\gamma}$, são 220 sítios segregantes. Apesar do número aproximado de sítios segregantes entre as amostras, as posições dos nucleotídeos com diferenças observadas são diferentes em cada variante. Na Figura 1 é possivel observar ao longo dos 3822 sítios onde ocorreram mudanças e onde ocorrom com maior frequência.

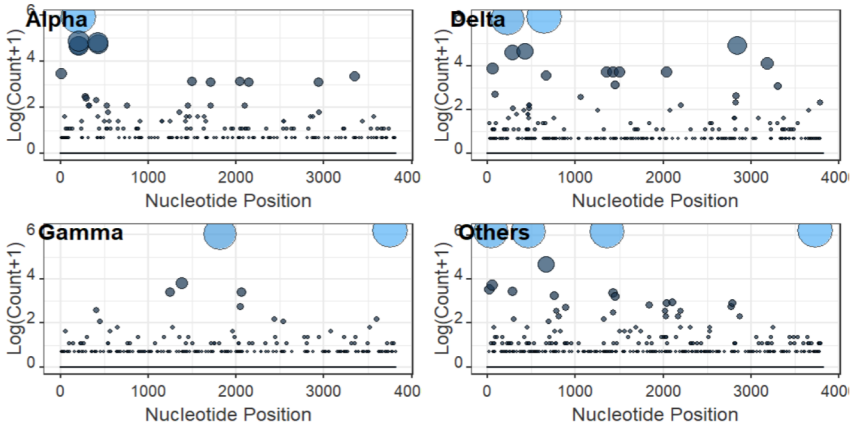

Figure 1: Frequência de nucleotídeos alterados por sítio transformado por da log(contagem +1).

Para testar a homogeneidade destes quato grupos, temos $D_n(B)$ como estatística de teste para teste dado por $H_0 : 2\mathcal{H}_{gg'} = \mathcal{H}_{gg} + \mathcal{H}_{g'g'}, \forall g \neq g'$ vs. $H_1 : 2\mathcal{H}_{gg'} > \mathcal{H}_{gg} + \mathcal{H}_{g'g'}$.

Considerando os grupos em que $\boldsymbol{g = 1 \rightarrow \alpha}$, $\boldsymbol{g = 2 \rightarrow \delta}$, $\boldsymbol{g = 3 \rightarrow \gamma}$ and $\boldsymbol{g = 4 \rightarrow}$ **outras variantes**. Temos $D_n(B)$ calculado em duas situações: i) Homogeneidade sobre os quatro grupos, com $D_n(B)_4$ a estatística de teste observada e ii) Homogeneidade entre duas a duas dos quatro grupos, com $D_n(B)_{ij}$ $(i < j \in \{1, 2, 3, 4\})$ como estatística de teste observada. Para todas essas situações o resultado está apresentado na Tabela 1.

Dn(B)	**Value**	***P*-value**
$\boldsymbol{D_n(B)_4}$	$1.95 \cdot 10^{-3}$	< 0.0001
$\boldsymbol{D_n(B)_{12}}$	$-9.11 \cdot 10^{-5}$	1.0
$\boldsymbol{D_n(B)_{13}}$	$2.11 \cdot 10^{-3}$	< 0.0001
$\boldsymbol{D_n(B)_{14}}$	$7.64 \cdot 10^{-4}$	< 0.0001
$\boldsymbol{D_n(B)_{23}}$	$1.89 \cdot 10^{-3}$	< 0.0001
$\boldsymbol{D_n(B)_{24}}$	$5.11 \cdot 10^{-4}$	< 0.0001
$\boldsymbol{D_n(B)_{34}}$	$1.38 \cdot 10^{-3}$	< 0.0001

Table 1: Test statistics and p-values for homogeneity tests.

A distribuição empírica da estatística de teste obtida por 1000 reamostragens por bootstrap estão apresentadas na Figura 2.

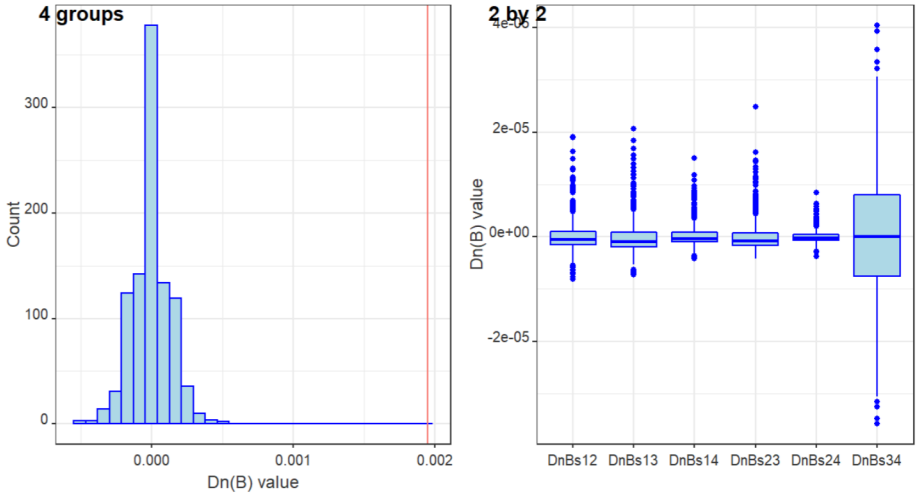

Figure 2: The empirical distributions of test statistic by bootstrap.

Em todos os casos, exceto para a comparação entre as variantes $\boldsymbol{\alpha}$ e $\boldsymbol{\gamma}$, os testes mostraram resultados significativos, i.e., há evidência estatística de que a diversidade entre os grupos é maior que a diversidade dentre as variantes. No caso da comparação entre as variantes $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, o valor observdo $D_n(B)$ é negativo, que significa que a diversidade média dentre os grupos é maior que a diversidade entre os grupos. Portanto, não há evidência estatística para rejeitar a hipótese nula de homogeneidade entre essas duas variantes.

No contexto da educação, as informações utilizadas na pesquisa são compostas por amostras aleatórias de aproximadamente 11000 estudantes, ingressantes na UNICAMP entre os anos 2000 e 2005 de todos os cursos que foram agrupados em cinco grandes áreas *Exatas*, *Humanas*, *Biológicas*, *Engenharias* e *Artes*. Para cada estudante no banco foi calculado o ganho relativo (Eq. 5).

Os testas de homogeneidade foram aplicados para comparar os sexos e tipos de escola onde cursaram o ensino médio dentro de cada grande área. Para isso novamente temos $D_n(B)$ como estatística de teste para teste dado por $H_0 : 2\mathcal{H}_{gg'} = \mathcal{H}_{gg} + \mathcal{H}_{g'g'}, \forall g \neq g'$ vs. $H_1 : 2\mathcal{H}_{gg'} > \mathcal{H}_{gg} + \mathcal{H}_{g'g'}$.

Novamente temos $D_n(B)$ calculado em duas situações: i) Homogeneidade entre tipo de escola que o estudande cursou o ensino médio dentro das grandes áreas, com a estatística de teste observada para cada área e ii) Homogeneidade entre sexos dos estudantes dentro das grandes áreas, com a estatística de teste observada também para cada área. Para todas essas situações o resultado está apresentado na Tabela 2.

	Escola Pública x Escola Privada		Sexo Masculino x Sexo Feminino	
Área	DnB	P-valor	DnB	P-valor
Exatas	$4.23 \cdot 10^{-4}$	0.022	$1.42 \cdot 13^{-3}$	< 0.0001
Humanas	$1.66 \cdot 10^{-3}$	< 0.0001	$5.15 \cdot 13^{-3}$	< 0.0001
Biológicas	$-9.45 \cdot 10^{-5}$	0.706	$7.88 \cdot 13^{-3}$	< 0.0001
Engenharias	$1.99 \cdot 10^{-4}$	< 0.0001	$1.14 \cdot 13^{-3}$	< 0.0001
Artes	$-9.39 \cdot 10^{-4}$	0.407	$3.26 \cdot 13^{-3}$	< 0.0001

Table 2: Frequência de previsões por barramento com erro absoluto menor que a previsão por carga global em 2020.

A distribuição empírica da estatística de teste obtida por 1000 reamostragens por bootstrap estão apresentadas na Figura 3.

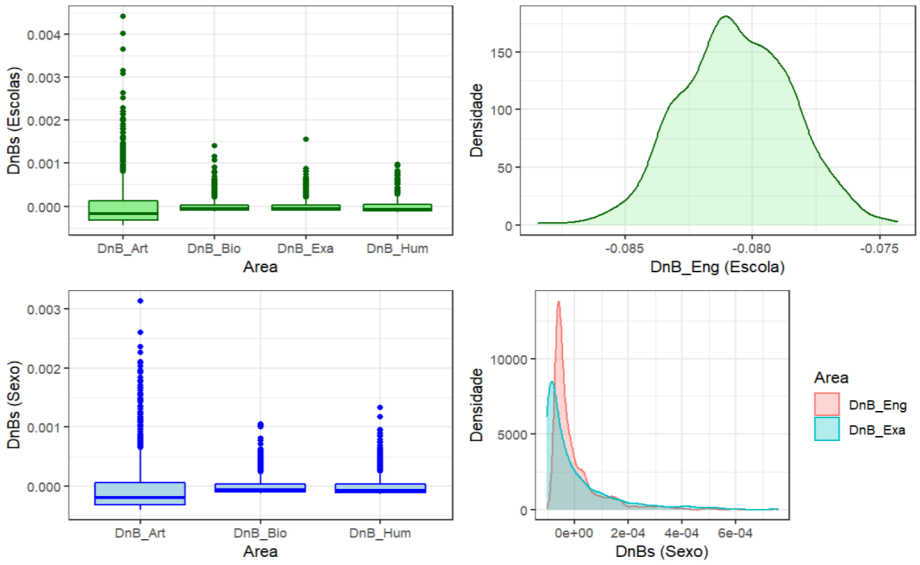

Figure 3: Distribuições empíricas da estatística de teste por bootstrap.

Entre as comparações envolvendo o sexo dos estudantes, os testes mostraram resultados significativos em todas as áreas, i.e., há evidência estatística de que a diversidade entre os os sexos é maior que a diversidade dentre os sexos. Nas comparações entre o tipode de escola, o valor observdo $D_n(B)$ é negativo apenas nas áreas de *Biológicas* e *Artes*, que significa que a diversidade média dentre os grupos é maior que a diversidade entre os grupos. Portanto, não há evidência estatística para rejeitar a hipótese nula de homogeneidade entre tipo de escolas nessas áreas.

3 Discussão

Utilizimanos nessa pesquisa um teste de homogeneidade baseado em uma classe de U-estatísticas ponderadas, chamadas Quasi U-estatísticas como ferramenta do estudo da variabilidade em sequências genômicas, que podem ter grande impacto na area de saúde pública, bem como na variabilidade presente no desempenho acadêmico de estudantes universitários que igualmente pode ter impacto na área da educação.

Com respeito a comparação sobre as variantes do SARS-COV2, este estudo mostra que a homogeneidade baseada em medidas de diversidade é capaz de capturar a variabilidade presente em sequências de nucleotídeos. Essa variabilidade em nosso contexto é caracterizada por mutações que ocorrem na sequência

referência traduzida no ínicio da pandemia ao longo do passar do tempo. Cabe citar que novas mutações contínuam aparecendo.

Os resultados são consistentes com a informação observada no conjunto de dados. Para o caso das variantes $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, os resultados podem ser explicados pelo fato de que essas duas variantes sofrem algumas mutações em comum mensionadas em Gupta et. al.(2021). Em complemento a esse fato, é possível observar um comportamento similar observado entre essas variantes na Figura 1, enquanto que nas outras as mudanças ocorrem em posições diferentes.

Em relação ao desempenho acadêmico de estudantes, este estudo mostra algumas situações onde a homogeneidade baseada em medidas de diversidade é capaz de capturar a variabilidade na performance de estudantes que adentraram na graduação. Em situações onde comparamos por exemplo estudantes de engenharias exatas, há claramente evidências de que o ganho relativo associado a estudantes de escolas públicas e privadas é significativamente diferentes, enquanto que nas áreas de Artes e Biológicas esse ganho relativo não foge da hipótese de homogeneidade.

References

[1] D. Gupta, et al. Structural and functional insights into the spike protein mutations of emerging SARS-CoV-2 variants. Cellular and Molecular Life Sciences 78.24 (2021): 7967-7989.

[2] H. P. Pinheiro, A. Pinheiro, and P. K. Sen, Comparison of genomic sequences using the hamming distance. Journal of Statistical Planning and Inference, vol. 130, no. 1-2, pp. 325-339, 2005.

[3] A. Pinheiro, P. K. Sen, and H. P. Pinheiro, Decomposability of high-dimensional diversity measures: Quasi U-statistics, martingales and nonstandard asymptotics. Journal of Multivariate Analysis, vol. 100, no. 8, pp. 1645-1656, 2009.

[4] A. Pinheiro, P. K. Sen, and H. Pinheiro, A class of asymptotically normal degenerate quasi U-statistics: Quasi U-statistics, martingales and nonstandard asymptotics. *Annals of the Institute of Statistical Mathematics*, vol. 63, no. 6, pp. 1165-1182, 2011.

[5] A. Pinheiro, H. P. Pinheiro, and P. K. Sen, The use of hamming distance in bioinformatics. *Handbook of Statistics*. Elsevier, 2012, vol. 28, pp. 129-162.

[6] P. Zhou et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579.7798 (2020): 270-273.

[7] R. P. Maia, H. P. Pinheiro, and A. Pinheiro, Academic performance of students from entrance to graduation via quasi u-statistics: a study at a brazilian research university, Journal of Applied Statistics, vol. 43, no. 1, pp. 7286, 2016.

[8] re3data.org: GISAID; editing status 2021-08-24; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R3Q59F last accessed: 2022-07-18