ANALYSIS OF VARIANCE FOR GENOMIC SEQUENCES IN UNBALANCED DESIGNS

Roberta de Souza[1], **Hildete P. Pinheiro**[2], **Cibele Q. da Silva**[3] and **Sérgio F. dos Reis**[4]

[1]Departamento de Pesquisa Farmacêutica, Grupo EMS-Sigma Pharma - SP, Brazil

[2] Departamento de Estatística, Universidade Estadual de Campinas, Caixa Postal 6065, CEP 13083-970, Campinas, SP, Brazil.
E-mail:hildete@ime.unicamp.br

[3] Departamento de Estatística, Universidade Federal de Minas Gerais - MG, Brazil. E-mail: cibeleqs@est.ufmg.br

[4] Departamento de Parasitologia, Universidade Estadual de Campinas - SP, Brazil. E-mail: sfreis@unicamp.br

Summary
In the study of genetic divergence among organisms, generally the analysis is done directly from the DNA molecule. Therefore, a possible outcome is categorical being one out of four categories (looking at the nucleotide level). Light & Margolin (1971) developed an analysis of variance for categorical data (CATANOVA) and Pinheiro et al. (2000) employed a similar measure of variation and extended the CATANOVA procedure taking into account several positions in the sequence for balanced designs. Here we consider variable number of sequences in each group, that is, the samples are unbalanced. In order to test the null hypothesis of homogeneity among groups, the asymptotic distribution of the test statistic was found and its power is evaluated. An application of the test to real data is illustrated using resampling methods such as the bootstrap to generate the empirical distribution of the test statistics.

KEYWORDS: Analysis of variance; Bootstrap; Categorical data; Asymptotic distribution; Molecular data; Statistical genetics; Unbalanced designs.

1. Introduction

The fundamental question asked in evolutionary genetics is that given a col-

lection of DNA sequences, what are the underlying forces responsible for the observed patterns of variability (Durrett, 2002). Consequently, a large effort has been devoted to develop methods for the estimation of parameters and hypothesis testing from data derived from DNA sequences (see Weir, 1990; Pinheiro et al., 2000, 2001; Pinheiro et al., 2003 and others). One hypothesis of interest is to test for homogeneity among groups of individuals sampled from a given region of the genome. Since DNA sequences are essentially categorical data in nature, i.e., if one looks at the nucleotide level, the response is one out of four categories (A, C, T, G), one for instance can use methods of categorical analysis of variance.

Based on a measure of variation for categorical data, expressed as frequencies for each category, Light & Margolin (1971) developed an analysis of variance for categorical data (CATANOVA). The properties of the variance components were investigated based on the multinomial model, which made possible the comparison of variability in the response variable within and between groups. This method can be applied to genomic sequences with only one position, but in the case of DNA sequences, one position does not provide enough information.

Pinheiro et al. (2000) employed a similar measure of variation and extended the CATANOVA procedure taking into account several positions in the sequence for balanced designs. When there is a binary response in each position, such as the case with molecular marker of the class of Random Amplified Polimorphic DNA (Williams et al., 1990), an analysis of variance for binary data in unbalanced designs was proposed by Souza et al. (2004).

Here we consider variable number of sequences in each group, that is, the samples are unbalanced. According to factors considered by Pinheiro et al. (2000), measures of diversity for unbalanced designs were obtained on Section 2. Assuming independence between sites we studied the asymptotic properties of the test statistic under the null hypothesis of homogeneity between groups. On section 6 a study of the power of the test is also presented. Finally, an application to real data is given in section 7.

2. Measures of diversity in unbalanced designs and the probability model

Let $\mathbf{X_i^g} = \left(X_{i1}^g,\ X_{i2}^g,\ \ldots X_{iK}^g\right)'$ be a random vector representing sequence i of group g, where X_{ik}^g represents the category present at site k of sequence i of group g, $i = 1, \ldots, n_g$, $k = 1, \ldots, K$ e $g = 1 \ldots, G$. Note that for the case of nucleotide sequences $X_{ik}^g \in \{\mathbf{A},\ \mathbf{C},\ \mathbf{T},\ \mathbf{G}\}$.

Using Simpson's index of diversity (Simpson, 1949) and following Pin-

heiro et al. (2000), the measures of diversity for unbalanced designs are:

$$TSI = 1 - \sum_{c=1}^{C} \left(\frac{n_{c\cdot\cdot}}{Kn}\right)^2; \quad (0.1)$$

$$WSI = 1 - \frac{1}{G}\sum_{g=1}^{G}\sum_{c=1}^{C} \left(\frac{n_{cg\cdot}}{Kn_g}\right)^2; \quad (0.2)$$

$$BSI = TSI - WSI = \sum_{c=1}^{C}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\frac{n_{cg\cdot}}{Kn_g}\right)^2 - \left(\frac{n_{c\cdot\cdot}}{Kn}\right)^2\right], \quad (0.3)$$

where TSI is the total Simpson index, which is the total variation in the pooled sample; WSI is the total variation within group and BSI is the variation between groups.

3. The probabilistic model

Denote by N_{cgk} the number of responses in category c, at site k for group g, and let p_{cgk} stand for the probability of falling into category c at position k for group g. Assuming that responses in different groups are independent, for groups g and position K, the responses $(N_{1gk}, N_{2gk}, ..., N_{Cgk})$ follow a Multinomial distribution:

$$Pr\{N_{1gk} = n_{1gk}, N_{2gk} = n_{2gk}, ..., N_{Cgk} = n_{Cgk}\} = \frac{n_g!}{\prod_{c=1}^{C} n_{cgk}!}\prod_{c=1}^{C}(p_{cgk})^{n_{cgk}}$$

where $\sum_{c=1}^{C} n_{cgk} = n_g$, $\sum_{c=1}^{C} p_{cgk} = 1$, $p_{cgk} > 0$, $c = 1, ..., C$, $\forall\ k = 1, ..., K$ and $g = 1, ..., G$. Therefore, $\mathrm{E}(n_{cgk}) = n_g p_{cgk}$, $\mathrm{Var}(n_{cgk}) = n_g p_{cgk}(1 - p_{cgk})$ and $\mathrm{Cov}(N_{c_1g_1k_1}, N_{c_2g_2k_2}) = -\delta n_g p_{c_1g_1k_1} p_{c_2g_2k_2}$,
where $\delta = I\!I(g_1 = g_2 = g \text{ and } k_1 = k_2)$.

If we assume that the positions are independent, the model is the product multinomial given by

$$\prod_{k=1}^{K}\prod_{g=1}^{G} Pr\{n_{1gk}, n_{2gk}, ..., n_{Cgk}\} = \prod_{k=1}^{K}\prod_{g=1}^{G}\frac{n_g!}{\prod_{c=1}^{C} n_{cgk}!}\prod_{c=1}^{C}(p_{cgk})^{n_{cgk}}.$$

If $\mathbf{V}_{gk} = (N_{1gk},\ \ldots,\ N_{Cgk})'$ and $\mathbf{P}_{gk} = (p_{1gk},\ \ldots\ p_{Cgk})'$, $k = 1, \ldots, K$; $g = 1, \ldots, G$, are vectors $C \times 1$ we can write

$$\mathrm{E}(\mathbf{V}_{gk}) = n_g \mathbf{P}_{gk} \quad \text{and} \quad \mathrm{Cov}(\mathbf{V}_{gk}) = n_g \mathbf{\Sigma}^{\diamond}_{gk}$$

where $\mathbf{\Sigma}^{\diamond}_{gk} = \mathbf{D}_{gk} - \mathbf{P}_{\diamond gk}\mathbf{P}'_{\diamond gk}$, with $\mathbf{D}_{gk}$ being a diagonal matrix $C \times C$ whose diagonal elements are p_{1gk}, ..., p_{Cgk}.

Now, let $\mathbf{V}_g = (\mathbf{V}_{g1},\ \mathbf{V}_{g2},\ \ldots,\ \mathbf{V}_{gK})'$ and $\mathbf{P}_g = (\mathbf{P}_{g1},\ \ldots,\ \mathbf{P}_{gK})'$ be vectors $CK \times 1$ and $\mathbf{V} = (\mathbf{V}_1,\ \mathbf{V}_2,\ \ldots,\ \mathbf{V}_G)'$ a vector $GCK \times 1$. Therefore,

$$\begin{aligned} \mathrm{Cov}(\mathbf{V}) \equiv \mathbf{\Sigma} &= \mathbf{\Sigma}_{11} \oplus \mathbf{\Sigma}_{12} \oplus ... \oplus \mathbf{\Sigma}_{1K} \oplus \mathbf{\Sigma}_{21} \oplus ... \oplus \mathbf{\Sigma}_{2K} \oplus ... \oplus \mathbf{\Sigma}_{GK} \\ &= n_1(\mathbf{\Sigma}^{\diamond}_{11} \oplus ... \oplus \mathbf{\Sigma}^{\diamond}_{1K}) \oplus n_2(\mathbf{\Sigma}^{\diamond}_{21} \oplus ... \oplus \mathbf{\Sigma}^{\diamond}_{2K}) \oplus ... \oplus \\ &\quad \oplus n_G(\mathbf{\Sigma}^{\diamond}_{G1} \oplus ... \oplus \mathbf{\Sigma}^{\diamond}_{GK}), \end{aligned} \tag{0.4}$$

4. Moments of diversity measures

Let

$$\mathbf{T} = \frac{1}{(Kn)^2}\mathbf{U}_{KG} \otimes \mathbf{I}_C \tag{0.5}$$

where $n = \sum_g n_g$, $\mathbf{U}_{KG}$ is a matrix $KG \times KG$ of $1's$, $\mathbf{I}_C$ is an identity matrix $C \times C$ and $\otimes$ is the Kronecker product (Searle, 1982). The expression (0.1) can thus be expressed in matrix form as

$$TSI = 1 - \sum_{c=1}^{C}\left(\frac{n_{c\cdot\cdot}}{KN_T}\right)^2 = 1 - \mathbf{V}'\mathbf{T}\mathbf{V}; \tag{0.6}$$

Let $\mathbf{M}$ be a diagonal matrix $G \times G$ whose diagonal elements $M_{gg} = Gn^2_{\cdot g\cdot} = G(Kn_g)^2$. Then $\mathbf{M}^{-1}$ is a diagonal matrix $G \times G$ with diagonal elements $M^{-1}_{gg} = 1/[G(Kn_g)^2]$.
We can define a matrix $\mathbf{W}$ as follows:

$$\mathbf{W} = [(\mathbf{M}^{-1} \otimes \mathbf{U}_K) \otimes \mathbf{I}_C]. \tag{0.7}$$

It is thus possible to write expression (0.2) in matrix form:

$$WSI = 1 - \frac{1}{G}\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{n_{cg\cdot}}{KN_g}\right)^2 = 1 - \mathbf{V}'\mathbf{W}\mathbf{V} \tag{0.8}$$

and

$$BSI = -\mathbf{V}'\mathbf{T}\mathbf{V} + \mathbf{V}'\mathbf{W}\mathbf{V} = \mathbf{V}'(-\mathbf{T} + \mathbf{W})\mathbf{V} = \mathbf{V}'\mathbf{B}\mathbf{V}; \tag{0.9}$$

where

$$\mathbf{B} = -\mathbf{T} + \mathbf{W} = \left(-\frac{1}{(KN_T)^2}\mathbf{U}_{KG} + \mathbf{M}^{-1} \otimes \mathbf{U}_K\right) \otimes \mathbf{I}_C. \tag{0.10}$$

Thus, according to classic results from linear models (Searle, 1971),

$$\begin{aligned}
E(TSI) &= 1 - \frac{1}{Kn_T} + \frac{1}{(Kn_T)^2}\sum_{c=1}^{C}\left[\sum_{g=1}^{G}\sum_{k=1}^{K} n_g p_{cgk}^2 - \left(\sum_g n_g p_{cg\cdot}\right)^2\right];\\
E(WSI) &= 1 - \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{n_g} + \frac{1}{GK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\left[\sum_{k=1}^{K}\frac{p_{cgk}^2}{n_g} - p_{cg\cdot}^2\right];\\
E(BSI) &= -\frac{1}{Kn_T} + \frac{1}{(Kn_T)^2}\sum_{c=1}^{C}\left[\sum_{g=1}^{G}\sum_{k=1}^{K} n_g p_{cgk}^2 - \left(\sum_g n_g p_{cg\cdot}\right)^2\right]\\
&+ \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{n_g} - \frac{1}{GK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\left[\sum_{k=1}^{K}\frac{p_{cgk}^2}{n_g} - p_{cg\cdot}^2\right].
\end{aligned}$$

Define the population variation within group g at position k

$$I_S(\mathbf{p}_{gk}) = 1 - \sum_{c=1}^{C} p_{cgk}^2\ . \tag{0.11}$$

Since our interest is to assess homogeneity between groups, the null hypothesis is $H_0 : p_{cgk} = p_{ck}$, where p_{cgk} is the probability of falling into category c at position k in group g. Under the null hypothesis, for all g, we have that $I_S(\mathbf{p}_{1k}) = I_S(\mathbf{p}_{2k}) = ... = I_S(\mathbf{p}_{Gk}) = I_S(\mathbf{p}_k)$, that is, the within-group variance at the k-th site is the same for all groups, where $\mathbf{p}_{gk} = (p_{1gk},\ p_{2gk},\ ..., p_{Cgk})'$ is a vector $C \times 1$ representing the probabilities associated with categories $c = 1, ...C$ of group g at site k.

Under the null hypothesis, the expected values of diversity measures are:

$$\mathrm{E}_0(TSI) = 1 - \frac{1}{Kn_T} + \frac{1}{K^2 n_T}\sum_{c=1}^{C}\left[\sum_{k=1}^{K} p_{ck}^2 - n_T p_{c\cdot}^2\right]; \tag{0.12}$$

$$\mathrm{E}_0(WSI) = 1 - \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{n_g} + \frac{1}{GK^2}\sum_{c=1}^{C}\left[\left(\sum_{g=1}^{G}\frac{1}{n_g}\right)\sum_{k=1}^{K} p_{ck}^2 - G p_{c\cdot}^2\right]; \tag{0.13}$$

$$\begin{aligned}
\mathrm{E}_0(BSI) &= -\frac{1}{KN_T} + \frac{1}{K^2 N_T}\sum_{c=1}^{C}\left[\sum_{k=1}^{K} p_{ck}^2 - N_T p_{c\cdot}^2\right]\\
&+ \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{n_g} - \frac{1}{GK^2}\sum_{c=1}^{C}\left[\left(\sum_{g=1}^{G}\frac{1}{n_g}\right)\sum_{k=1}^{K} p_{ck}^2 - G p_{c\cdot}^2\right]
\end{aligned} \tag{0.14}$$

Since $\mathbf{V}$ follows a multinomial distribution we can use the Central Limit Theorem and

$$\mathbf{V} \xrightarrow{D} N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{when} \quad N_0 \to \infty, \tag{0.15}$$

where $\boldsymbol{\mu}$ is the vector of expected values of $\mathbf{V}$, $\boldsymbol{\Sigma}$ is given in (0.4), and $N_0 = \min_{1 \le g \le G} n_g$.

Under H_0, for $g = 1, ..., G$,

$$\boldsymbol{\Sigma}^{\diamond}_{gk} = \boldsymbol{\Sigma}^{\diamond}_{0k} \qquad \text{and} \qquad \boldsymbol{\Sigma}^{\diamond}_{g} = \boldsymbol{\Sigma}^{\diamond}_{0} = \boldsymbol{\Sigma}^{\diamond}_{01} \oplus \boldsymbol{\Sigma}^{\diamond}_{02} ... \oplus \boldsymbol{\Sigma}^{\diamond}_{0K}; \tag{0.16}$$

where $\boldsymbol{\Sigma}^{\diamond}_{0k}$ is a matrix $C \times C$, of the form

$$\boldsymbol{\Sigma}^{\diamond}_{0k} = \mathbf{D}_k - \boldsymbol{\mu}_{\diamond k} \boldsymbol{\mu}'_{\diamond k}, \tag{0.17}$$

where $\mathbf{D}_k$ is a diagonal matrix $C \times C$ whose diagonal elements are $p_{1k}, \ \ldots, \ p_{Ck}$ and $\boldsymbol{\mu}_{\diamond k} = (p_{1k}, \ \ldots, \ p_{Ck})'$. Thus, under H_0

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 = \boldsymbol{\eta} \otimes \boldsymbol{\Sigma}^{\diamond}_0 \tag{0.18}$$

where $\boldsymbol{\eta}$ is a diagonal matrix $G \times G$ whose diagonal elements are $\eta_{gg} = n_g$ and $\boldsymbol{\Sigma}^{\diamond}_0$ is given by (0.16) . Therefore, under H_0, asymptotically,

$$\mathbf{V} \xrightarrow{D} N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0); \quad \text{where} \quad \boldsymbol{\mu}_0 = ((n_1, \ n_2, \ \ldots, \ n_G) \otimes \mathbf{P}_0)' \tag{0.19}$$

with $\mathbf{P}_0 = (p_{11}, \ \ldots, \ p_{C1}, \ p_{12}, \ \ldots, \ p_{C2}, \ \ldots, \ p_{1K}, \ \ldots, \ p_{CK})'$.

5. Asymptotic distribution of the test statistic

The interest now is to derive a statistic to test the hypothesis of homogeneity among groups. To that end we propose a test statistic in terms of Simpson's indexes and, therefore, we obtain the asymptotic distribution of the statistics $\mathbf{V'BV}$,$\mathbf{V'TV}$ and $\mathbf{V'WV}$ that are related to Simpson's indexes BSI, TSI and WSI, respectively.

From (0.9), BSI can be written as,

$$BSI = \mathbf{V'BV} = \sum_{c=1}^{C} \left[\frac{1}{G} \sum_{g=1}^{G} \left(\frac{N_{cg\cdot}}{K N_g} \right)^2 - \left(\frac{N_{c\cdot\cdot}}{K N_T} \right)^2 \right]. \tag{0.20}$$

Let $\theta_{cgk} = N_{cgk} - E_0(N_{cgk}) = N_{cgk} - n_g p_{ck}$. Then,

$$\theta_{c\cdot\cdot} = \sum_{g=1}^{G} \sum_{k=1}^{K} \theta_{cgk} = N_{c\cdot\cdot} - \sum_{g=1}^{G} n_g \sum_{k=1}^{K} p_{ck}. \tag{0.21}$$

As $N_0 = \min_{1\leq g\leq G} n_g$, under H_0,

$$\boldsymbol{\theta} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_0), \quad \text{when } N_0 \to \infty, \tag{0.22}$$

where $\boldsymbol{\theta} = (\theta_{111} \ \ldots \ \theta_{Cg1} \ \ldots \ \theta_{CGK})'$ and $\boldsymbol{\Sigma_0}$ is given by (0.18). Therefore, we can write

$$\begin{aligned} BSI &= \sum_{c=1}^{C}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\theta_{cg\cdot}+n_g p_{c\cdot}}{Kn_g}\right)^2 - \left(\frac{\theta_{c\cdot\cdot}+p_{c\cdot}n}{Kn}\right)^2\right] \\ &= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta}; \end{aligned} \tag{0.23}$$

where $\mathbf{A} = (a_1\mathbf{A}^\star \ a_2\mathbf{A}^\star \ \ldots \ a_G\mathbf{A}^\star) = \mathbf{a}\otimes\mathbf{A}^\star$ is a vector $1\times CGK$, $\mathbf{a} = (a_g)$ is a vector $1\times G$ being $a_g = \frac{1}{GN_g} - \frac{1}{N_T}$, e $\mathbf{A}^\star$ is a vector $1\times CK$ of the form

$$\mathbf{A}^\star = \frac{2}{GK^2}(p_{1\cdot},\ \ldots,\ p_{C\cdot},\ p_{1\cdot},\ \ldots,\ p_{C\cdot},\ \ldots,\ p_{1\cdot},\ \ldots,\ p_{C\cdot}) \tag{0.24}$$

As we have already seen, $\boldsymbol{\theta}$ has normal asymptotic distribution, then

$$\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} \xrightarrow{D} \sum_{i=1}^{CGK}\lambda_i\left(\chi_1^2\right)_i; \quad \text{when } N_0\to\infty \ \text{(Searle, 1971)} \tag{0.25}$$

where $\left(\chi_1^2\right)_i' s$ are independent random variables chi-square distributed with com 1 degree of freedom and $\{\lambda_i,\ i=1,\ ...\ CGK\}$ is a set of eigenvalues of

$$\mathbf{B}\boldsymbol{\Sigma}_0 = \mathbf{B}\left(\boldsymbol{\eta}\otimes\boldsymbol{\Sigma}_0^\diamond\right) = \left[\left(-\frac{1}{(KN_T)^2}\mathbf{U}_{KG} + \left(\mathbf{M}^{-1}\otimes\mathbf{U}_K\right)\right)\otimes\mathbf{I}_C\right](\boldsymbol{\eta}\otimes\boldsymbol{\Sigma}_0^\diamond)$$

by (0.10) and (0.18). Note also that from (0.22),

$$\mathbf{A}\boldsymbol{\theta} \xrightarrow{D} N\left(0, \mathbf{A}\boldsymbol{\Sigma}_0^\diamond\mathbf{A}'\right); \tag{0.26}$$

with

$$\begin{aligned} \mathbf{A}\boldsymbol{\Sigma}_0^\diamond\mathbf{A}' &= \frac{4}{(GK^2)^2}\sum_{g=1}^{G}\left[\left(\frac{1}{n_g}-\frac{1}{N_T}\right)^2\sum_{c=1}^{C}\left(\sum_{k=1}^{K}p_{c\cdot}^2\left[n_g p_{cgk}(1-p_{cgk})\right] + \right.\right. \\ &+ \left.\left.\sum_{c'\neq c=1}^{C}\sum_{k=1}^{K} -n_g p_{c\cdot}p_{c'\cdot}p_{cgk}p_{c'gk}\right)\right]. \end{aligned}$$

Then, we can say that, under H_0,

$$BSI = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta} \xrightarrow{D} \sum_{i=1}^{CGK} \lambda_i \left(\chi_1^2\right)_i + N\left(0, \mathbf{A}\boldsymbol{\Sigma}_0^{\diamond}\mathbf{A}'\right); \tag{0.27}$$

that is, BSI is the sum of a linear combination of random variables χ_1^2 distributed and the other normally distributed. By Lemma 5.1 we have that $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ and $\mathbf{A}\boldsymbol{\theta}$ are not independent and therefore the distribution of $\mathbf{V}'\mathbf{B}\mathbf{V}$ is not a convolution of these random variables.

Lemma 5.1 $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ *and* $\mathbf{A}\boldsymbol{\theta}$ *are not independent. (Proof in A.1)* ■

As an alternative to obtain the distribution of $\mathbf{V}'\mathbf{B}\mathbf{V}$, let us define $\mathbf{R}$ as a matrix $CGK \times CGK$ such that $\mathbf{R}\boldsymbol{\Sigma}\mathbf{R}' = \mathbf{I}_{CGK}$ and $\mathbf{Y} = \mathbf{R}\mathbf{V} \Rightarrow \mathbf{V} = \mathbf{R}^{-1}\mathbf{Y}$. Therefore, as $N_0 \to \infty$,

$$\mathbf{Y} \xrightarrow{D} N(\mathbf{R}\boldsymbol{\mu}, \mathbf{I}_{CGK}) \quad \text{and} \quad \mathbf{V}'\mathbf{B}\mathbf{V} = \mathbf{Y}'(\mathbf{R}^{-1})'\mathbf{B}\mathbf{R}^{-1}\mathbf{Y} \equiv \mathbf{Y}'\mathbf{C}\mathbf{Y},$$

where $\mathbf{C} = (\mathbf{R}^{-1})'\mathbf{B}\mathbf{R}^{-1}$.

Let $\mathbf{P}$ be an orthogonal matrix such that $\mathbf{P}\mathbf{C}\mathbf{P}'$ is a diagonal matrix and $\mathbf{Y}^\star = \mathbf{P}\mathbf{Y} \Rightarrow \mathbf{Y} = \mathbf{P}^{-1}\mathbf{Y}^\star = \mathbf{P}'\mathbf{Y}^\star$. Therefore,

$$\mathbf{Y}^\star \xrightarrow{D} N(\mathbf{P}\mathbf{R}\boldsymbol{\mu}, \mathbf{I}_{CGK});$$

$\mathbf{V}'\mathbf{B}\mathbf{V} = \mathbf{Y}'\mathbf{C}\mathbf{Y} = (\mathbf{Y}^\star)'\mathbf{P}\mathbf{C}\mathbf{P}'\mathbf{Y}^\star = (\mathbf{Y}^\star)'\mathbf{C}^\star\mathbf{Y}^\star$,
with $\mathbf{C}^\star \equiv \mathbf{P}(\mathbf{R}^{-1})'\mathbf{B}\mathbf{R}^{-1}\mathbf{P}'$. Therefore,

$$BSI = (\mathbf{Y}^\star)'\mathbf{C}^\star\mathbf{Y}^\star \xrightarrow{D} \sum_{i=1}^{CGK} c_i^\star \left(\chi_1^2(\delta_{1i})\right); \tag{5.28}$$

where $\delta_{1i} \equiv \dfrac{(\mu_i^\star)^2}{2}$, being $\mu_i^\star$ the i-th element of the vector $\boldsymbol{\mu}^\star = \mathbf{P}\mathbf{R}\boldsymbol{\mu}$ and $c_i^\star$'s the elements of the diagonal of $\mathbf{C}^\star$, $c_i^\star \in \mathbb{R}$.

Reminding that, $TSI = 1 - \mathbf{V}'\mathbf{T}\mathbf{V}$, under H_0, $\mathbf{V} \xrightarrow{D} N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Since $\mathbf{T}\boldsymbol{\Sigma}_0$ is not idempotent, the distribution of $\mathbf{V}'\mathbf{T}\mathbf{V}$ is not $\chi^2_{(rank(\mathbf{T}), \boldsymbol{\mu}_0'\mathbf{T}\boldsymbol{\mu}_0)}$. Nevertheless,

$$\mathbf{V}'\mathbf{T}\mathbf{V} \quad = \quad \boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta} + \mathbf{A_2}\boldsymbol{\theta} + \delta; \tag{5.29}$$

where $\mathbf{A_2}' = (\mathbf{A_2^\star}\ \mathbf{A_2^\star}\ ...\ \mathbf{A_2^\star})' = (\mathbf{1}_G \otimes \mathbf{A_2^\star})'$ is a vector $CGK \times 1$, $\mathbf{1}_G$ is a row vector of 1's of size G and $\mathbf{A_2^\star}$ is a vector $1 \times CK$ of the form

$$\mathbf{A_2^\star} = \frac{2}{K^2 N_T}\left(p_{1\cdot},\ \ldots,\ p_{C\cdot},\ p_{1\cdot},\ \ldots,\ p_{C\cdot},\ \ldots,\ p_{1\cdot},\ \ldots,\ p_{C\cdot}\right).$$

According to Lemma 5.2, $\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ and $\mathbf{A_2}\boldsymbol{\theta}$ are not independent and in this case $\mathbf{V}'\mathbf{T}\mathbf{V}$ will not be the convolution of these variables.

Lemma 5.2 $\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ *e* $\mathbf{A_2}\boldsymbol{\theta}$ *are not independent. (Proof in A.2)* ■

Under H_0,

$$\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta} \xrightarrow{D} \sum_{i=1}^{CGK} \lambda_{2i} \left(\chi_1^2\right)_i \quad \textit{and} \quad \mathbf{A_2}\boldsymbol{\theta} \xrightarrow{D} N(0, \mathbf{A}_2\boldsymbol{\Sigma}_0\mathbf{A}_2'), \quad \text{as } N_0 \to \infty.$$

Therefore, asymptotically, we can say that $\mathbf{V}'\mathbf{T}\mathbf{V}$ is the sum of a linear combination of χ_1^2 random variables and a normally distributed random variable.

$$\mathbf{V}'\mathbf{T}\mathbf{V} \xrightarrow{D} \sum_{i=1}^{CGK} \lambda_{2i} \left(\chi_1^2\right)_i + N(0, \mathbf{A}_2\boldsymbol{\Sigma}_0\mathbf{A}_2') + \delta;$$

where $\lambda_{2i}, i = 1, ..., CGK$ is the set of eigenvalues of

$$\mathbf{T}\boldsymbol{\Sigma}_0 = \frac{1}{(Kn)^2}\mathbf{T}^{\diamond}\boldsymbol{\Sigma}_0 = \frac{1}{(Kn)^2}\mathbf{T}^{\diamond}\boldsymbol{\eta} \otimes \boldsymbol{\Sigma}_0^{\diamond}.$$

Alternatively, we have that the distribution of $\mathbf{V}'\mathbf{T}\mathbf{V}$ is analogous to that obtained in (5.28):

$$\mathbf{V}'\mathbf{T}\mathbf{V} = (\mathbf{Y_2^\star})'\mathbf{C_2^\star}\mathbf{Y_2^\star} \xrightarrow{D} \sum_{i=1}^{CGK} c_{2i}^\star \left(\chi_1^2(\delta_{2i})\right), \quad \text{with } \delta_{2i} \equiv \frac{(\mu_{2i}^\star)^2}{2}; \quad (5.30)$$

where $c_{2i}^\star$'s , $c_{2i}^\star \in \mathbb{R}$, are the elements of the diagonal matrix

$$\mathbf{C_2^\star} \equiv \mathbf{P_2}(\mathbf{R_2}^{-1})'\mathbf{T}\mathbf{R_2}^{-1}\mathbf{P_2},$$

where $\mathbf{P_2}$ is an orthogonal matrix and $\mu_{2i}^\star$ the i-th element of the vector $\boldsymbol{\mu}_2^\star = \mathbf{P_2}\mathbf{R_2}\boldsymbol{\mu}$, with $\mathbf{R_2}$ being a matrix $CGK \times CGK$ such that $\mathbf{R_2}\boldsymbol{\Sigma}\mathbf{R_2}' = \mathbf{I}_{CGK}$. Note that

$$\mathbf{Y_2^\star} \xrightarrow{D} N(\boldsymbol{\mu}_2^\star, \mathbf{I}_{CGK});$$

In the case of the within-group Simpson's index we have $WSI = 1 - \mathbf{V}'\mathbf{W}\mathbf{V}$ and under H_0,

$$\mathbf{V}'\mathbf{W}\mathbf{V} \;=\; \boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta} + \mathbf{A_3}\boldsymbol{\theta} + \delta; \quad (5.31)$$

where $\mathbf{A_3}' = (\mathbf{A_3^\star}\ \mathbf{A_3^\star}\ ...\ \mathbf{A_3^\star})' = (\mathbf{n_3} \otimes \mathbf{A_3^\star})'$ is a vector $CGK \times 1$, $\mathbf{n_3} = \left(\frac{1}{n_1}\ \ldots\ \frac{1}{n_g}\right)$ is a row vector of size G and $\mathbf{A_3^\star}$ is a vector $1 \times CK$ of the form

$$\mathbf{A_3^\star} = \frac{2}{K^2 G}\left(p_{1.},\ \ldots,\ p_{C.},\ p_{1.},\ \ldots,\ p_{C.},\ \ldots,\ p_{1.},\ \ldots,\ p_{C.}\right).$$

We also showed (Lemma 5.3) that $\mathbf{V'WV}$ is not a convolution of $\boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta}$ and $\mathbf{A_3}\boldsymbol{\theta}$.

Lemma 5.3 *$\boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta}$ and $\mathbf{A_3}\boldsymbol{\theta}$ are not independent. (Proof in A.3)* ■

Therefore,

$$\mathbf{V'WV} \xrightarrow{D} \sum_{i=1}^{CGK} \lambda_{3i}\left(\chi_1^2\right)_i + N(0, \mathbf{A}_3\boldsymbol{\Sigma}_0\mathbf{A}_3') + \delta;$$

where $\lambda_{3i}, i = 1, ..., CGK$ is the set of eigenvalues of

$$\mathbf{W}\boldsymbol{\Sigma}_0 = [(\mathbf{M}^{-1} \otimes \mathbf{U}_K) \otimes \mathbf{I}_C]\boldsymbol{\Sigma}_0.$$

In other words, $\mathbf{V'WV}$ is the sum of a linear combination of χ_1^2 random variables, a normally distributed random variable and a constant.

As in (5.28) we obtained that

$$\mathbf{V'WV} = (\mathbf{Y_3^\star})'\mathbf{C_3^\star}\mathbf{Y_3^\star} \xrightarrow{D} \sum_{i=1}^{CGK} c_{3i}^\star\left(\chi_1^2(\delta_{3i})\right) \quad \text{with } \delta_{3i} \equiv \frac{(\mu_{3i}^\star)^2}{2} \tag{5.32}$$

where $c_{3i}^\star$'s, $c_{3i}^\star \in \mathbb{R}$, are the elements of the diagonal matrix $\mathbf{C_3^\star} \equiv \mathbf{P_3}(\mathbf{R_3}^{-1})'\mathbf{T}\mathbf{R_3}^{-1}\mathbf{P_3'}$, where $\mathbf{P_3}$ is an orthogonal matrix and $\mu_{3i}^\star$ the i-th element of the vector $\boldsymbol{\mu}_3^\star = \mathbf{P_3}\mathbf{R_3}\boldsymbol{\mu}$.

Note that

$$\mathbf{Y_3^\star} \xrightarrow{D} N(\boldsymbol{\mu}_3^\star, \mathbf{I}_{CGK});$$

where $\mathbf{R_3}$ is a matrix $CGK \times CGK$ such that $\mathbf{R_3}\boldsymbol{\Sigma}\mathbf{R_3'} = \mathbf{I}_{CGK}$.

To test for homogeneity between groups we need to obtain the asymptotic distribution of a statistic that is a function of BSI/WSI. We thus need to study the order of convergence of some statistics and their moments.

Let $N_0 = \min_{1 \le g \le G} n_g$. We have from (0.14), (0.12) and (0.13), respectively,

$$\begin{aligned}
\theta_1 &\equiv E_0(BSI) \\
&= -\frac{1}{KN_T} + \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{n_g} + \frac{1}{K^2N_T}\sum_{c=1}^{C}\sum_{k=1}^{K}p_{ck}^2 - \frac{1}{GK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\frac{1}{n_g}p_{ck}^2 \\
\theta_2 &\equiv E_0(TSI) = 1 + \frac{1}{KN_T}\left(\sum_{c=1}^{C}\sum_{k=1}^{K}\frac{p_{ck}^2}{K} - 1\right) - \sum_{c=1}^{C}\frac{p_{c.}^2}{K^2} \\
&= 1 + O(N_0^{-1}) - \sum_{c=1}^{C}\frac{p_{c.}^2}{K^2} \\
\theta_3 &\equiv E_0(WSI) = 1 - \frac{1}{GK}\sum_{g=1}^{G}\frac{1}{n_g} + \frac{1}{GK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\frac{1}{n_g}p_{ck}^2 - \frac{p_{c.}^2}{K^2} \\
&= 1 + O(N_0^{-1}) - \sum_{c=1}^{C}\frac{p_{c.}^2}{K^2}
\end{aligned}$$

Defining now,

$$T_1 \equiv BSI - \theta_1; \quad T_2 \equiv TSI - \theta_2 \quad \text{and} \quad T_3 \equiv WSI - \theta_3.$$

If $n_g = O(N_0)\ \forall g, g = 1, ..., G$ we have

1. $\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta} \sim \sum_{i=1}^{CGK}\lambda_{2i}\left(\chi_1^2\right)_i = O_p(N_0^{-1})$, since $\{\lambda_{2i}, i = 1, ..., CGK\}$ is the set of eigenvalues of $\mathbf{T}\boldsymbol{\Sigma}_0 = O(N_0^{-1})$;

2. $\mathbf{A_2}\boldsymbol{\theta} = O_p(N_0^{-1/2})$, since $\mathbf{A_2}\boldsymbol{\theta} \sim N(0, \mathbf{A}_2\boldsymbol{\Sigma}_0\mathbf{A}_2')$, $\mathbf{A_2} = O(N_0^{-1})$ and $\boldsymbol{\Sigma}_0 = O(N_0)$;

3. $\delta = \frac{1}{K^2}\sum_{c=1}^{C}p_{c.}^2 = O(1)$.

Analogously we have,

1. $\boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta} = \sum_{i=1}^{CGK}\lambda_{3i}\left(\chi_1^2\right)_i = O_p(N_0^{-1})$;

2. $\mathbf{A_3}\boldsymbol{\theta} = O_p(N_0^{-1/2})$.

Therefore, $T_2 = TSI - \theta_2 = O_p(N_0^{-1/2})$, $T_3 = WSI - \theta_3 = O_p(N_0^{-1/2})$ and $BSI = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta} = O_p(N_0^{-1}) + O_p(N_0^{-1/2}) = O_p(N_0^{-1/2})$. Then, to test the hypothesis of homogeneity among groups we propose the following statistic:

$$F_1 \equiv N_0^{1/2}\left(\frac{BSI}{WSI}\right); \tag{5.33}$$

And we can write F_1 as

$$\begin{aligned} F_1 &= N_0^{1/2}\left(\frac{BSI}{T_3+\theta_3}\right) = N_0^{1/2}\frac{BSI}{\theta_3}\left(1 - \frac{T_3}{\theta_3+T_3}\right) \\ &= N_0^{1/2}\frac{BSI}{\theta_3} + O(N_0^{-1/2}) = O(1); \end{aligned}$$

since $N_0^{1/2}BSI = O_p(1)$, $\theta_3 + T_3 = O(1) + O_p(N_0^{-1/2}) = O_p(1)$ and

$$\theta_3 = 1 - \sum_{c=1}^{C}\frac{p_{c.}^2}{K^2} + O(N_0^{-1}) = \theta_3^0 + O(N_0^{-1}).$$

Therefore, by (5.28), we have that, for $n_g = O(N_0)$ and $N_0 \to \infty$, F_1 is expressed as a linear combination of non-central chi-square random variables.

$$F_1 = N_0^{1/2}\frac{BSI}{\theta_3^0} \sim \sum_{i=1}^{CGK}\frac{N_0^{1/2}}{\theta_3^0}c_i^\star\left(\chi_1^2(\delta_i)\right), \tag{5.34}$$

where $c_i^\star$ and δ_i are obtained according to (5.28).

When the sample sizes $n_g s$ are small we can call upon resampling methods such as the bootstrap and obtain the empirical distribution of F_1.

6. Power of the test

We will consider now the following alternative hypotheses in order to evaluate the power of the test:

$$H_1 : p_{cgk} = \frac{1}{\sqrt{n_g}}\gamma_{cgk} + p_{ck} \quad \text{for all } g = 1, \ldots, G.$$

These are called the *Pitman alternative hypothesis* (Pinheiro et al., 2000). Therefore, it is also possible to assess the behavior of the power of the test for alternatives that get closer to the null hypothesis as N_0 increases.

We are interested in the case where $\gamma_{cgk} \neq 0$. Under the alternative hypothesis we have:

$$\begin{aligned}
\theta_{cgk} &= N_{cgk} - E(N_{cgk}) = N_{cgk} - n_g\left(\frac{\gamma_{cgk}}{\sqrt{n_g}} + p_{ck}\right), \\
\theta_{cg\cdot} &= N_{cg\cdot} - n_g\left(\frac{\gamma_{cg\cdot}}{\sqrt{n_g}} + p_{c\cdot}\right) \text{ and} \\
\theta_{c\cdot\cdot} &= N_{c\cdot\cdot} - \sum_g n_g\left(\frac{\gamma_{cg\cdot}}{\sqrt{n_g}} + p_{c\cdot}\right).
\end{aligned}$$

Now,

$$\begin{aligned}
BSI &= \sum_{c=1}^{C}\left[\frac{1}{G}\sum_{g=1}^{G}\left(\frac{\theta_{cg\cdot} + n_g p_{c\cdot}}{Kn_g}\right)^2 - \left(\frac{\theta_{c\cdot\cdot} + p_{c\cdot}n}{Kn}\right)^2\right] \\
&+ \sum_{c=1}^{C}\sum_{g=1}^{G}\left\{\frac{1}{G}\left[2\frac{\gamma_{cg\cdot}}{\sqrt{n_g}}\frac{n_g(\theta_{cg\cdot} + n_g p_{c\cdot})}{(Kn_g)^2}\right] - \left[2\frac{\sum_g n_g\gamma_{cg\cdot}}{\sqrt{n_g}}\frac{(\theta_{c\cdot\cdot} + np_{c\cdot})}{(Kn)^2}\right]\right. \\
&+ \left.\left(\frac{n_g\gamma_{cg\cdot}}{Kn_g\sqrt{n_g}}\right)^2 - \left(\frac{\sum_g n_g\gamma_{cg\cdot}}{Kn\sqrt{n_g}}\right)^2\right\}.
\end{aligned}$$

From (0.23) we have

$$\begin{aligned}
BSI &= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}\boldsymbol{\theta} + \sum_c\sum_g\left[\frac{2\sqrt{n_g}\gamma_{cg\cdot}}{K^2}\left(\frac{\theta_{cg\cdot}}{GN_g^2} - \frac{\theta_{c\cdot\cdot}}{N_T^2}\right)\right] \\
&+ \sum_c\sum_g\left[\frac{2p_{c\cdot}\sqrt{n_g}\gamma_{cg\cdot}}{K^2}\left(\frac{1}{GN_g} - \frac{1}{N_T}\right)\right] \\
&+ \sum_c\sum_g\frac{1}{GK^2}\left[\left(\frac{\gamma_{cg\cdot}}{\sqrt{n_g}}\right)^2 - \left(\frac{\sum_g\sqrt{n_g}\gamma_{cg\cdot}}{N_T}\right)^2\right] \\
&= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + (\mathbf{A} + \mathbf{A_4})\boldsymbol{\theta} + \delta^{\star\star},
\end{aligned}$$

where $\mathbf{A}$ is a vector $1 \times CGK$ defined in (0.23) and (0.24), and $\mathbf{A_4} = (\mathbf{A}_4^\star,\ \mathbf{A}_4^\star,\ \ldots,\ \mathbf{A}_4^\star)$, a vector $1 \times CGK$, being $\mathbf{A}_4^\star = (\mathbf{A}_{41}^\star,\ \mathbf{A}_{42}^\star,\ \ldots,\ \mathbf{A}_{4G}^\star)$, $1 \times CG$, with $\mathbf{A}_{4g}^\star = (\mathbf{a}_{41g}^\star,\ \mathbf{a}_{42g}^\star,\ \ldots,\ \mathbf{a}_{4Cg}^\star)$, $1 \times C$,

$$\mathbf{a}_{4cg}^\star = \frac{2\sqrt{n_g}}{K^2}\left(\frac{1}{GN_g^2} - \frac{1}{N_T}\right)\gamma_{cg} - \sum_{g' \neq g}\frac{2\sqrt{N_{g'}}}{(KN_T)^2}\gamma_{cg'}.$$

Since $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ and $\mathbf{A}\boldsymbol{\theta}$ are not independent (Lema 5.1), $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ e $(\mathbf{A}+\mathbf{A_4})\boldsymbol{\theta}$ are also not independent. Therefore, the distribution of $\mathbf{V}'\mathbf{B}\mathbf{V}$ is not a convolution of a linear combination of χ^2 random variables and a random variable normally distributed.

If $\mathbf{A}^{\star\star} = \mathbf{A} + \mathbf{A_4}$, then

$$\mathbf{V}'\mathbf{B}\mathbf{V} = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + \mathbf{A}^{\star\star}\boldsymbol{\theta} + \delta^{\star\star} = \mathbf{X}'\mathbf{X} - \frac{1}{4}(\mathbf{A}^{\star\star})'\mathbf{B}^{-1}\mathbf{A}^{\star\star} + \delta^{\star\star},$$

where $\mathbf{X} = \mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A}^{\star\star}$.

Suppose that $\mathbf{B}$ is semi-definite positive, then its elements $\in \mathbb{R}$, as is the case with balanced samples (Pinheiro *et al.*, 2000). In this case, if $\boldsymbol{\Gamma} = \mathbf{B}^{1/2}\boldsymbol{\Sigma}(\mathbf{B}^{1/2})'$ and $\boldsymbol{\mu}^{\star\star} = \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A}^{\star\star}$, then

$$\mathbf{X} \sim N(\boldsymbol{\mu}^{\star\star}; \boldsymbol{\Gamma}).$$

Let $\mathbf{P}_4$ be an orthogonal matrix such that $\mathbf{P}_4\boldsymbol{\Gamma}(\mathbf{P}_4)' = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix. If $\mathbf{Y} = \mathbf{P}_4\mathbf{X} \Rightarrow \mathbf{X} = \mathbf{P}_4'\mathbf{X}$, then,

$$\mathbf{Y} \sim N(\mathbf{P}_4\boldsymbol{\mu}^{\star\star}; \boldsymbol{\Lambda}) \quad \text{and} \quad \mathbf{X}'\mathbf{X} = \mathbf{Y}'\mathbf{P}_4\mathbf{P}_4'\mathbf{Y} \sim \sum_{i=1}^{CGK} \lambda_i(\chi_1^2(\delta_i))_i \quad (5.35)$$

where λ_i are the eigenvalues of $\boldsymbol{\Lambda}$, in this case the elements of the diagonal matrix $\boldsymbol{\Lambda}$. Note that $\boldsymbol{\Lambda}$ is semi-definite positive and therefore $\lambda_i \geq 0$. $\delta_i = \dfrac{a_i^2}{2\lambda_i}$, being a_i o i-th element of vector $\frac{1}{2}\mathbf{P}_4\mathbf{B}^{-1/2}\mathbf{A}^{\star\star}$, which is a linear combination of the γ_{cgk}'s.

If the constant $c = -\dfrac{1}{4}(\mathbf{A}^{\star\star})'\mathbf{B}^{-1}\mathbf{A}^{\star\star} + \delta^{\star\star}$, then,

$$\Pr(F_1 \geq u) = \Pr(\sqrt{N_0}\mathbf{X}'\mathbf{X} \geq \theta_3^o u - \sqrt{N_0}c). \quad (5.36)$$

However,

$$c = -\frac{1}{4}(\mathbf{A}^{\star\star})'\mathbf{B}^{-1}\mathbf{A}^{\star\star} + \delta^{\star\star} = O(N_0^{-1/2}),$$

since $n_g = O(N_0)$. Therefore, $\sqrt{N_0}c = O(1)$, with the increase of the non-centrality parameter δ_i the distribution of the χ^2 random variable tends to the right and, for $N_0 \to \infty$, the probability in (5.36) tends to 1, indicating that the power of the test converges to 1.

In cases where **B** is not semi-definite positive, more studies are needed on the power of the test. Nevertheless, in these cases the power of the test can be evaluated numerically.

7. Application

The data presented in this section are derived from a study of population genetic structure using sequences of the mitochondrial DNA genome of the freshwater turtle *Hydromedusa maximiliani*, conducted in the state of São Paulo in southeastern Brazil (Souza et al., 2003). This freshwater turtle inhabits topologically complex habitats characterized by sequences of ridges and valleys, each drained by river and stream systems. For this study, an area of approximately 2700ha containing three drainages (hereafter drainages I, II and III) was sampled based on the natural spatial hierarchy formed by rivers and streams. Within each drainage, specimens of *H. maximiliani* were randomly hand-caught in the natural habitat of shallow rivers and streams. The data set consists of 48 sequences of the mitochondrial DNA genome of freshwater turtles of the species *H. maximiliani* collected from watersheds I, II and III, having sample sizes of 30, 7 and 11 sequences, respectively (Souza et al., 2003). Drainage I, the larger drainage sampled and which yielded the larger sample size, was further subdivided according to the spatial hierarchy of the main rivers and their tributaries, resulting in three sample sites. Sample sizes for each site were 4, 12, and 9, respectively. From each individual in these samples, a 1,400 bp fragment of the mitochondrial region encompassing cytochrome b, 12S, and Thr-proline genes, as well as the D loop region was obtained. Sequences were obtained from these fragments with a 377 Automated DNA sequencer. Details of the molecular procedures can be found in Souza et al. (2003). These sequences were obtained from two different regions of the mitochondrial genome, the cytochrome b gene with 262 sites and the control region with 413 sites.

Since the sequences were taken from two different regions of the mitochondrial genome the analysis will be carried out separately. Under H_0 : $p_{cgk} = p_{ck}$, where p_{ck} is the proportions of sequences in site k showing category c, where $c \in \{\mathbf{A},\ \mathbf{T},\ \mathbf{C},\ \mathbf{G}\}$. Initially, we compare groups corresponding to the three watersheds sampled for the cytochrome b and then we compare sequences from the control region. These comparisons were also performed for the three partitions within watershed I. Since the sample sizes are small, we call upon resampling techniques to generate the empirical distribution of F_1 under H_0. The procedure is as follows:

Step 1: Estimate p_{ck} from the data, i.e., $\hat{p}_{ck} = \dfrac{n_{c1k} + n_{c2k} + n_{c3k}}{n}$, which is

the observed proportion of sequences in the pooled sample that in position k falls in category c, and compute the observed value of the test statistic F_1 (F_{1obs}).
Step 2: Generate $n = 48$ sequences with $K = 262$ (for cytochrome b gene) and $K = 413$ (for the control region) positions each from a Multinomial distribution $(48; \hat{p}_{1k}, \hat{p}_{2k}, \hat{p}_{3k}, \hat{p}_{4k})$.
Step 3: Compute the value of the test statistic F_1 from the generated data.
Step 4: Repeat steps 2 and 3 $10,000$ times.

Table 1 shows the observed values of F_1 and the respective $p - values$ corresponding to the cytochrome b gene and control region sequences of the mitochondrial DNA sampled from the freshwater turtle populations in watersheds I, II and III.

The histograms of Figure 1 show the behavior of the distribution of the test statistic F_1. Figure 2 shows the behavior of the distribution of F_1 for subpopulations of freshwater turtles within watershed I. The observed values of F_1 and the respective p-values from cytochrome b gene and control region for the comparisons of partitions 1, 2 and 3 within watershed I are on Table 2.

For the significance level of 5%, we reject the hypothesis of homogeneity among groups if the $p - value \leq 0.05$. If the observed value of F_1 is negative then the $p - value = 2P(F_1 \leq F_{1obs})$, otherwise, $p - value = 2P(F_1 \geq F_{1obs})$. Among the three watersheds we find strong evidence to reject the hypothesis of homogeneity among groups, that is, analyzing the sequences for the control region there is statistical evidence for genetic variation among the three populations from watersheds I, II and III. Comparing sequences for the control region of individuals from watershed I, we also obtained strong evidence for the rejection of the null hypothesis of homogeneity among groups. Therefore, at the level of DNA sequences it was possible to observe genetic variation among individuals of watershed I.

A pairwise comparison of partitions indicated which turtle samples differed between partitions 1, 2 and 3, within watershed I (see Table 2). Pairwise comparisons were adjusted by the Bonferroni correction and at the significance level of 5% the hypothesis of homogeneity between groups is rejected if the $p - value < (0.05/3) \approx 0.017$.

Therefore, there is strong evidence for genetic diversity between watersheds I and II (Table 1). Diversity is also evident between partitions 1 and 2 and between 1 and 3 (Table 2). The empirical distribution of F_1 for the pairwise comparisons have symmetric distributions around zero and the case of pairwise comparisons between partitions within watershed I is shown on

Figure 3.

Acknowledgements

This research was funded in part by Fundação de Amparo à Pesquisa do Estado de São Paulo (00/00805-9), Fundo de Apoio ao Ensino e Pesquisa (0023/00) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

References

Durrett, R. (2002). *Probability Models for DNA Sequence Evolution*. New York: Springer Verlag.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics* **19,** 293-325.

Light, R. J. and Margolin, B. H. (1971). An Analysis of Variance for Categorical Data. *Journal of the American Statistical Association* **66,** 534-544.

Pinheiro, H. P., Seillier-Moiseiwitsch, F., Sen, P. K. and Eron, J. (2000). Genomic sequence analysis and quasi-multivariate CATANOVA. In *Handbook of Statistics, Volume 18: Bioenvironmental and Public Health Statistics*, P. K. Sen and C. R. Rao (eds), 713-746, Amsterdam: Elsevier.

Pinheiro, H. P., Seillier-Moiseiwitsch, F. and Sen, P. K. (2001). *Analysis of variance for Hamming distances applied to unbalanced designs*. Research Report 30/01. Instituto de Matemática, Estatística e Computação Científica. Universidade Estadual de Campinas.

Pinheiro, H. P., Pinheiro, A. S. and Sen, P. K. (2003). Comparisons of Genomic Sequences using the Hamming Distance. *Journal of Statistical Planning and Inference (in press)*.

Searle, S. L. (1971). *Linear Models*. New York: John Wiley & Sons.

Searle, S. L. (1982). *Matrix Algebra Useful for Statistics*. New York: John Wiley & Sons.

Simpson, E. H. (1949). The measurement of diversity. *Nature* **163,** 688.

Souza, F. L., A. F. Cunha, M. A. Oliveira, G. A. G. Pereira and S. F. dos Reis. 2003. Preliminary phylogeographic analysis of the neotropical freshwater turtle *Hydromedusa maximiliani* (Chelidae). *Journal of Herpetology* **37,** 199-205.

Souza, R., Pinheiro, H. P., Silva, C. Q. and Reis, S. F. (2004). *Analysis of variance for binary data in unbalanced designs. Brazilian Journal of Probability and Statistics* (submitted).

Weir, B. S. (1996). *Genetic Data Analysis 2: Methods for Discrete Population Genetic Data*. Sunderland: Sinauer.

Williams, J.K.G., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* 18, 6531-6535.

Appendix

A.1 Proof of Lemma 5.1

$\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ and $\mathbf{A}\boldsymbol{\theta}$ would be independent if and only if $\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{B} = 0$ (Searle, 1971).

$$\begin{aligned}
\mathbf{A}\boldsymbol{\Sigma}_0\mathbf{B} &= -\frac{1}{(Kn)^2}\mathbf{A}(\boldsymbol{\eta}\otimes\boldsymbol{\Sigma}_0^{\diamond})(\mathbf{U}_{KG}\otimes\mathbf{I}_C) + \mathbf{A}(\boldsymbol{\eta}\otimes\boldsymbol{\Sigma}_0^{\diamond})(\mathbf{M}^{-1}\otimes\mathbf{U}_K\otimes\mathbf{I}_C) \\
&= -\frac{1}{(Kn)^2}(\mathbf{a}\otimes\mathbf{A}^{\star})\left[\boldsymbol{\eta}\mathbf{U}_G\otimes\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)\right] \\
&+ (\mathbf{a}\otimes\mathbf{A}^{\star})\left[\boldsymbol{\eta}\mathbf{M}^{-1}\otimes\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)\right] \\
&= \left[\mathbf{a}\boldsymbol{\eta}\left(-\frac{1}{(Kn)^2}\mathbf{U}_G + \mathbf{M}^{-1}\right)\right]\otimes[\mathbf{A}^{\star}\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C)]
\end{aligned}$$

Let $\mathbf{a}^{\star} = (p_{1\cdot}\ldots p_{C\cdot})$. Remember that $\boldsymbol{\Sigma}_0^{\diamond} = \boldsymbol{\Sigma}_{01}^{\diamond}\oplus\boldsymbol{\Sigma}_{02}^{\diamond}\oplus\ldots\oplus\boldsymbol{\Sigma}_{0K}^{\diamond}$.

$$\mathbf{A}^{\star}\boldsymbol{\Sigma}_0^{\diamond}(\mathbf{U}_K\otimes\mathbf{I}_C) = \frac{2}{GK^2}\left(\mathbf{a}^{\star}\Sigma_{01}^{\diamond}\ \mathbf{a}^{\star}\Sigma_{02}^{\diamond}\ \ldots\ \mathbf{a}^{\star}\Sigma_{0K}^{\diamond}\right)(\mathbf{U}_K\otimes\mathbf{I}_C).$$

For each k, of (0.17) we have,

$$\mathbf{a}^{\star}\Sigma_{0k}^{\diamond} = \left(p_{1k}\left(p_{1\cdot} - \sum_{c=1}^{C}p_{c\cdot}p_{ck}\right)\ p_{2k}\left(p_{2\cdot} - \sum_{c=1}^{C}p_{c\cdot}p_{ck}\right)\ \ldots\ p_{Ck}\left(p_{C\cdot} - \sum_{c=1}^{C}p_{c\cdot}p_{ck}\right)\right).$$

The first element of the vector $\mathbf{A}^\star\mathbf{\Sigma}_0^\diamond(\mathbf{U}_K \otimes \mathbf{I}_C)$ is

$$\frac{2}{GK^2}\left(p_{11}\left(p_{1\cdot} - \sum_c p_{c\cdot}p_{c1}\right) + p_{12}\left(p_{1\cdot} - \sum_c p_{c\cdot}p_{c2}\right) + \ldots + p_{1K}\left(p_{1\cdot} - \sum_c p_{c\cdot}p_{cK}\right)\right)$$
$$= \frac{2}{GK^2}\left(\sum_k p_{1k}\left(p_{1\cdot} - \sum_c p_{c\cdot}p_{ck}\right)\right) = \frac{2}{K^2}\left(p_{1\cdot} - \sum_c p_{c\cdot}\sum_k p_{ck}\right) \neq 0.$$

As $\mathbf{a}\boldsymbol{\eta}\left(-\frac{1}{(KN_T)^2}\mathbf{U}_G + \mathbf{M}^{-1}\right) \neq 0 \Rightarrow \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ e $\mathbf{A}\boldsymbol{\theta}$ are not independent. ■

A.2 Proof of Lemma 5.2

$\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ and $\mathbf{A_2}\boldsymbol{\theta}$ would be independent if and only if $\mathbf{A_2}\mathbf{\Sigma}_0\mathbf{T} = 0$ (Searle, 1971).

$$\begin{aligned}\mathbf{A_2}\mathbf{\Sigma}_0\mathbf{T} &= \frac{1}{(Kn)^2}\mathbf{A_2}(\boldsymbol{\eta}\otimes\mathbf{\Sigma}_0^\diamond)(\mathbf{U}_G\otimes\mathbf{U}_K\otimes\mathbf{I}_C)\\ &= \frac{1}{(Kn)^2}(\mathbf{1}_G\otimes\mathbf{A_2^\star})[\boldsymbol{\eta}\mathbf{U}_G\otimes\mathbf{\Sigma}_0^\diamond(\mathbf{U}_K\otimes\mathbf{I}_C)]\\ &= \frac{1}{(Kn)^2}(\mathbf{1}_G\boldsymbol{\eta}\mathbf{U}_G\otimes\mathbf{A_2^\star}\mathbf{\Sigma}_0^\diamond(\mathbf{U}_K\otimes\mathbf{I}_C)).\end{aligned}$$

From Lemma 5.1 we have

$$\mathbf{A_2^\star}\mathbf{\Sigma}_0^\diamond(\mathbf{U}_K\otimes\mathbf{I}_C)) = \frac{2}{K^2n}\left(\mathbf{a}^\star\Sigma_{01}^\diamond\ \mathbf{a}^\star\Sigma_{02}^\diamond\ \ldots\ \mathbf{a}^\star\Sigma_{0K}^\diamond\right)(\mathbf{U}_K\otimes\mathbf{I}_C) \neq 0.$$

As $\mathbf{1}_G\boldsymbol{\eta}\mathbf{U}_G \neq 0 \Rightarrow \boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ and $\mathbf{A_2}\boldsymbol{\theta}$ are not independent. ■

A.3 Proof of Lemma 5.3

$\boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta}$ and $\mathbf{A_3}\boldsymbol{\theta}$ would be independent if and only if $\mathbf{A_3}\mathbf{\Sigma}_0\mathbf{W} = 0$ (Searle, 1971).

$$\begin{aligned}\mathbf{A_3}\mathbf{\Sigma}_0\mathbf{W} &= \mathbf{A_3}(\boldsymbol{\eta}\otimes\mathbf{\Sigma}_0^\diamond)[(\mathbf{M}^{-1}\otimes\mathbf{U}_K)\otimes\mathbf{I}_C]\\ &= (\mathbf{n_3}\otimes\mathbf{A_3^\star})[\boldsymbol{\eta}\mathbf{M}^{-1}\otimes\mathbf{\Sigma}_0^\diamond(\mathbf{U}_K\otimes\mathbf{I}_C)]\\ &= \mathbf{n_3}\boldsymbol{\eta}\mathbf{M}^{-1}\otimes\mathbf{A_3^\star}\mathbf{\Sigma}_0^\diamond(\mathbf{U}_K\otimes\mathbf{I}_C).\end{aligned}$$

From Lemma 5.1

$$\mathbf{A_3^\star}\mathbf{\Sigma}_0^\diamond(\mathbf{U}_K\otimes\mathbf{I}_C)) = \frac{2}{K^2G}\left(\mathbf{a}^\star\Sigma_{01}^\diamond\ \mathbf{a}^\star\Sigma_{02}^\diamond\ \ldots\ \mathbf{a}^\star\Sigma_{0K}^\diamond\right)(\mathbf{U}_K\otimes\mathbf{I}_C) \neq 0.$$

As $\mathbf{n_3}\boldsymbol{\eta}\mathbf{M}^{-1} \neq 0 \Rightarrow \boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ e $\mathbf{A_3}\boldsymbol{\theta}$ are not independent. ■

Table 1: Observed Values of F_1 and p-values

Sequence	Watersheds	F_{1obs}	$p-value$
Cytochrome b gene	I,II, III	0.0000	0.4923
Control region	I, II, III	0.0003	0.0025
Control region	I, II	0.0004	0.0070
	I, III	0.0001	0.1955
	II, III	0.0001	0.0270

Table 2: Observed Values of F_1 and p-values: Watershed I

Sequence	Partitions	F_{1obs}	$p-value$
Cytochrome b gene	1,2,3	0.0000	0.8112
Control region	1,2,3	-0.0003	0.0020
Control region	1,2	-0.0005	< 0.0002
	1,3	-0.0020	0.0136
	2,3	-0.0010	0.1330

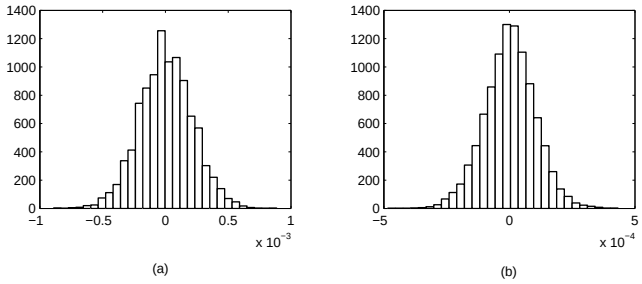

Figure 1: Empirical Distribution of F_1: DNA sequences of freshwater turtle populations. (a) Cytochrome b gene. (b) Control region.

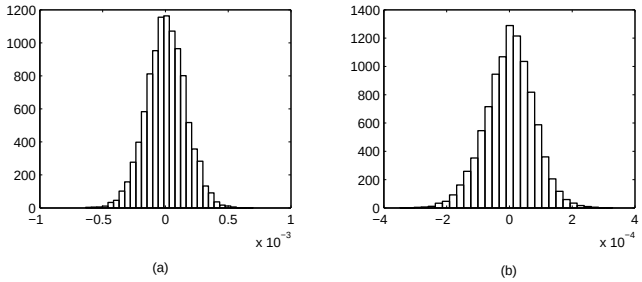

Figure 2: Empirical Distribution of F_1: DNA Sequences of Turtles from Watershed I. (a) cytochrome b gene. (b) Control region.

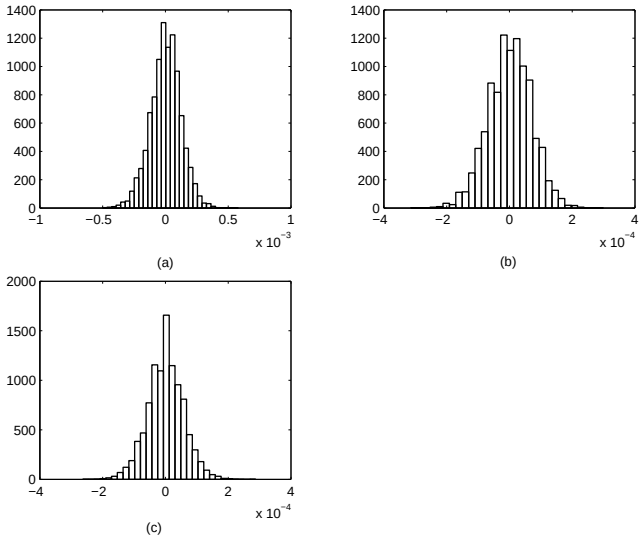

Figure 3: Empirical distribution of F_1: DNA Sequences for the Control Region from Watershed I. (a) Partitions 1 and 2. (b) Partitions 1 and 3. (c) Partitions 2 and 3.