Multivariate CATANOVA and Applications to DNA Sequences in Categorical Data

By Hildete Prisco Pinheiro [*]
State University of Campinas - Brazil
Françoise Seillier-Moiseiwitsch

Pranab Kumar Sen
University of North Carolina at Chapel Hill

Abstract

Simpson (1949) proposed a measure of diversity for categorical data. On the basis of a similar measure of variation, Light & Margolin (1971) developed an analysis of variance (CATANOVA), for one-way tables, suitable for categorical variables. This framework can be used to compare the variability of the response variable at a single position between and within groups. We extend these to deal with variation at a number of sites because, in the context of interest (sequences from the human immunodeficiency virus), a single position yields little information. Components of variation are derived based on the fact that the sum of squares of deviations from the mean can be expressed as a function of the squares of the pairwise differences for all possible pairs. We assume independence among positions and develop a test statistic for the null hypothesis of homogeneity among groups.

1. Introduction The motivation here is to develop a multivariate analysis of variance for categorical data when the response variable is not ordered. The focus here is the comparison of sets of sequences. For example, we are interested in comparing DNA sequences from the human immunodeficiency virus (HIV) from different geographical areas to see whether the variability is similar in each. Similarly, when we study several individuals and obtain a set of sequences from each individual at different time points, our interest lies in estimating the variability between and within individuals.

[*] This research was funded in part by CAPES (Brazilian Institution), the National Science Foundation, the American Foundation for AIDS Research and the National Institutes of Health.

Key words and phrases: Multivariate CATANOVA, Analysis of Variance, Simpson Index of Diversity

Simpson (1949) proposed a measure of diversity for categorical data in terms of frequencies for each category. On the basis of a similar measure of variation, Light & Margolin (1971) developed an analysis of variance (CATANOVA), for one-way tables, suitable for categorical variables. This framework can be used to compare the variability of the response variable at a single position between and within groups. We consider a number of sites because, in the context of interest (the analysis of HIV-1 sequences), a single position yields little information. The basic motivation and discussions about the model assumptions are in Section 2. Components of variation are derived from the fact that the sum of squares of deviations from the mean can be expressed as a function of the squares of the pairwise differences for all possible pairs (Section 3). The sequences are not considered on an individual basis but only as contributing to the overall variability in the distribution of the categorical response. We partition the measures of diversity according to the factors considered (Section 4) assuming independence among positions (Section 5). We develop a test statistic for the null hypothesis of homogeneity among groups (Sections 6 and 7) and assess its power (Section 8). Simulations are performed to evaluate the relevance to the asymptotic results when sample sizes are moderate (Section 9). A brief data analysis follows (Section 10).

2. Basic Motivation

Light & Margolin (1971) developed an analysis of variance for categorical data (CATANOVA) for two-dimensional contingency tables (i.e., one-way tables). They investigated the properties of the components of variation under a common multinomial model. Anderson & Landis (1980) extended the CATANOVA procedure to multidimensional contingency tables involving several factors and a categorical response variable. For the case of comparisons of DNA sequences, we would have a data set which can be summarized in Table 1.

To understand Anderson & Landis approach, we could say that, making an analogy to the analysis of variance in experimental design, the groups play the role of blocks and the positions are treated as a factor. Since we usually have a large number of positions (at least 35 in the aminoacid level and even bigger in the nucleotide level), we would have a factor with at least 35 levels, which may be a problem using this approach. Also, the main interest is the difference among groups not necessarily among the positions.

The interest is in assessing the homogeneity among groups: the null hypothesis is that $p_{cgk} = p_{ck}$ where p_{cgk} is the population probability of belonging to category c in group g at position k. One could argue that this is the classical Pearson's χ^2 test, but note that the classical Pearson's χ^2 statistic for Table 1 is

$$\chi^2_P = \sum_{g=1}^{G}\sum_{c=1}^{C}\sum_{k=1}^{K} \frac{G\left(n_{cgk} - \dfrac{n_{c\cdot k}}{G}\right)^2}{N n_{c\cdot k}}$$

with $K(G-1)(C-1)$ degrees of freedom. The limiting χ^2-distribution is a close approximation only when the cell frequencies n_{cgk}'s are all large (at least 5). In analyzing amino-acid sequences, we know that these conditions are not met. The distribution at a single position usually exhibits a few polymorphisms with very low frequencies. Just as for Fisher's exact test, the exact null distribution is difficult to implement for small values of N, when G or K is not small. Moreover, if the number of degrees of freedom of the χ^2-statistic is large but the noncentrality parameter is not proportionally so, the resulting test is likely to have less power than some tests directed towards specific alternatives. Note that the number of degrees of freedom here is usually large: for instance, comparing two groups of sequences 100 nucleotide long yields 300 degrees of freedom. For these reasons we need to use another approach to assess homogeneity among groups.

Note that the two approaches discussed above (Anderson & Landis and Pearsons χ^2) all assume independence among groups, sequences and positions, which leads to a product multinomial model.

One can argue that genetic types of individuals in a population are not generally independent because they may be related due to their shared ancestry. In the specific case of HIV, assuming that the individuals are epidemiologically independent may not be such a strong assumption, because of the rapid evolution of HIV (Hahn et al., 1985; 1986; Coffin, 1986; Seillier-Moiseiwitsch et al., 1994; Mansky & Temin, 1995).

Another issue is the fact that we do not have any information about the underlying structure that reviews any possible ordering of the categories or the positions. On the other hand, these positions should be stochastically interrelated. For this reason the classical logistic model may not work out here without some further information that relate to possible underlying structure for the categories and positions.

An important point is the problem of known dependence among positions in DNA sequences. If we assume independence among groups and individuals (sequences), but nonindependence among positions, we no longer have a product multinomial model and any model we choose will have too many parameters to be estimated (taking into account all the possible correlations between positions), requiring a very large data set. In the binary case, we may use the Bahadur representation model (Bahadur, 1961) or the model suggested by Liang, Zeger & Qaqish (1992) with only pairwise dependence, but we still have $K + \binom{K}{2}$ parameters in the model. Ideally the ratio between the number of sequences (individuals) and the number of parameters in the models to should be at least 5. Therefore, the number of sequences should be at least 5 times $\frac{K(K+1))}{2}$ and in practice this is almost impossible. Since we do not have any information about the structure of these sequences, we may reduce this model assuming that the correlation between positions is the same for all positions. Under these circustances we may have feasible models with $K+1$ parameters.

When we have polychotomous response the model proposed by Liang, Zeger & Qaqish (1992) have too many parameters (for C categories, $KC + C^2\binom{K}{2}$) and we will need to reduce the number of paramenters by, for example, assuming equal correlation structure between all the positions. These situations we will pursue in a near future.

As a first step, we would like to test the difference between and within the groups, using a measure of diversity for categorical response assuming independence among positions. Note that, if K is large and the dependence between the positions is not small, the distribution may degenerate to lower dimension ones. This means that, given some of the positions, the others may be conditionally redundant, but we do not know this information. So, we shall assume that the degree of dependence becomes small and small as K becomes larger. This also allows us to use an exchangeable model for large K with small intraclass dependence pattern. Our proposed measure here is averaging over K. Therefore, a large K may have a smoothing effect. Hence, this assumption of small dependence and independence would not be so divergent.

3. Variation in Categorical Data

For categorical data, the mean is an ill-defined concept. Therefore, measures of variation, such as the variance, which are meaningful for continuous variables, no longer apply. Gini (1912) found an alternative way of characteryzing variation and developed a measure of variation for categorical data.

Let $X_1, X_2, \ldots, X_N$ denote measurements of N independent experimental units. The variance of X may be expressed as $\mathrm{E}\phi(X_1, X_2)$, where $\phi(a,b) = \frac{1}{2}(a-b)^2$ (Hoeffding, 1948). In a similar fashion, the sum of squares is

$$SS \quad = \quad \sum_{i=1}^{N}(X_i - \bar{X})^2 = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}(X_i - X_j)^2 = \frac{1}{N}\sum_{1\leq i \leq N} d_{ij}^2 \tag{2.1}$$

where $\bar{X} = \sum_{i=1}^{N} X_i/N$ and $d_{ij} = X_i - X_j$.

In the present context, each X_i falls into one of C possible categories. Define $d(X_i, X_j) = d_{ij}$ as

$$d_{ij} = \begin{cases} 1 \text{ if } X_i \text{ and } X_j \text{ name different categories} \\ 0 \text{ if } X_i \text{ and } X_j \text{ name the same category.} \end{cases} \tag{2.2}$$

Definition 1
The variation for categorical responses $X_1, \ldots, X_N$ is

$$\frac{1}{2N}\sum_{j=1}^{N}\sum_{i=1}^{N} d_{ij}^2 = \frac{1}{2N}\sum_{j=1}^{N}\sum_{i=1}^{N} d_{ij} \tag{2.3}$$

where d_{ij} is defined in (2.2). ■

As each response assumes one and only one of the C possible categories, the data is summarized by the vector $\boldsymbol{\Phi} = (n_1, \ldots, n_C)$ where n_i is the number of responses in the ith category ($i =$

$1, \ldots, C$), so that $\sum_{i=1}^{C} n_i = N$. Then the variation in the responses is defined as

$$D_N \equiv \frac{1}{2N} \sum_{i \neq j} n_i n_j = \frac{N}{2} \left\{ 1 - \sum_{i=1}^{C} \left(\frac{n_i}{N} \right)^2 \right\} \quad . \tag{2.4}$$

If $p_1, \ldots, p_C$ stand for the probabilities of X belonging to these C categories, the Simpson Index of ecological diversity (Simpson, 1949) is defined as

$$I_S(\mathbf{p}) \equiv 1 - \mathbf{p}'\mathbf{p} = 1 - \sum_{i=1}^{C} p_i^2 \tag{2.5}$$

and its corresponding sample counterpart is

$$\hat{I}_S(\mathbf{p}) = 1 - \hat{\mathbf{p}}'\hat{\mathbf{p}} = 1 - \sum_{i=1}^{C} \hat{p}_i^2, \tag{2.6}$$

where $\hat{p}_i = n_i/N$, $i = 1, \ldots, C$ relate to the sample proportion. Therefore, we have

$$D_N = \frac{N}{2} \hat{I}_S(\mathbf{p})$$

The definitions (2.4) and (2.5) are motivated by two properties.

1. The variation of N categorical responses is minimized if and only if they all belong to the same category, i.e., $p_i = 1, \ \forall \ i = 1, \ldots, C$.

2. The variation of N responses is maximized when the responses are distributed among the available categories as evenly as possible, i.e., $p_i = 1/C, \ \forall \ i = 1, \ldots, C$.

4. Partitioning the Measures of Diversity

Let $\mathbf{X}_i^g = (X_{i1}^g, X_{i2}^g, \ldots, X_{iK}^g)'$ be a random vector representing sequence i of group g. Suppose $i = 1, \ldots, N$, $k = 1, \ldots, K$ and $g = 1, \ldots, G$. So, X_{ik}^g represents position k of sequence i of group g. X_{ik}^g is a categorical variable assuming C (unordered) categories. For instance, if comparisons are made at the nucleotide level, $x_{ik}^g \in \{\mathbf{A}, \mathbf{C}, \mathbf{T}, \mathbf{G}\}$ and there are 4 categories.

First, assume there is only one position for each sequence. We summarize the data in Table 2.

Now d_{ij} is defined as

$$d_{ij} = \begin{cases} 1 \text{ if } X_i^g \neq X_j^g \\ 0 \text{ if } X_i^g \neq X_j^g \ . \end{cases}$$

The total number of responses is

$$NG = \sum_{g=1}^{G} n_{\cdot g} = \sum_{c=1}^{C} n_{c\cdot} = \sum_{c=1}^{C}\sum_{g=1}^{G} n_{cg}$$

where n_{cg} is the number of responses in category c for group g and $N = n_{\cdot g} = \sum_{c=1}^{C} n_{cg}$ is the number of responses for group g, which here is simply the number of sequences in each group. The Total Simpson Index (TSI) is

$$TSI = 1 - \sum_{c=1}^{C} \left(\frac{n_{c\cdot}}{NG}\right)^2 \tag{3.1}$$

The dispersion within group g (i.e., within $\{x_1^g, x_2^g, \ldots, x_N^g\}$) is

$$1 - \sum_{c=1}^{C} \left(\frac{n_{cg}}{n_{\cdot g}}\right)^2 \tag{3.2}$$

Therefore, the within-group Simpson Index (WSI) is found by averaging (3.2) over all g's:

$$WSI = \frac{1}{G}\sum_{g=1}^{G}\left\{1 - \sum_{c=1}^{C}\left(\frac{n_{cg}}{n_{\cdot g}}\right)^2\right\} = 1 - G\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{n_{cg}}{NG}\right)^2 \tag{3.3}$$

The between-group Simpson Index (BSI) is

$$BSI = TSI - WSI = G\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{n_{cg}}{NG}\right)^2 - \sum_{c=1}^{C}\left(\frac{n_{c\cdot}}{NG}\right)^2 \tag{3.4}$$

Now, assume there are K positions along each sequence. We have $\mathbf{X}_i^g = (X_{i1}^g, X_{i2}^g, \ldots, X_{iK}^g)'$ and $\mathbf{X}^g = (\mathbf{X}_1^g, \mathbf{X}_2^g, \ldots, \mathbf{X}_N^g)'$. The data are summarized in Table 1.

The total number of responses is

$$NGK = \sum_{g=1}^{G} n_{\cdot g\cdot} = \sum_{c=1}^{C} n_{c\cdot\cdot} = \sum_{k=1}^{K} n_{\cdot\cdot k} = \sum_{c=1}^{C}\sum_{g=1}^{G}\sum_{k=1}^{K} n_{cgk}$$

The variation within the gth group at the kth position is

$$1 - \sum_{c=1}^{C}\left(\frac{n_{cgk}}{n_{\cdot gk}}\right)^2 = 1 - \sum_{c=1}^{C}\left(\frac{n_{cgk}}{N}\right)^2,$$

since $n_{\cdot gk} = N$. The variation within the gth group is

$$1 - \sum_{c=1}^{C}\left(\frac{n_{cg.}}{n_{\cdot g\cdot}}\right)^2 = 1 - \sum_{c=1}^{C}\left(\frac{n_{cg\cdot}}{NK}\right)^2$$

since $n_{\cdot g \cdot} = NK$. The measures of dispersions are

$$WSI = \frac{1}{G}\sum_{g=1}^{G}\left\{1-\sum_{c=1}^{C}\left(\frac{n_{cg\cdot}}{NK}\right)^2\right\} = 1 - G\sum_{c=1}^{C}\left(\frac{n_{cg\cdot}}{NGK}\right)^2 \tag{3.5}$$

$$TSI = 1-\sum_{c=1}^{C}\left(\frac{n_{c\cdot\cdot}}{NGK}\right)^2 \quad , \tag{3.6}$$

$$\text{and} \quad BSI = TSI - WSI = G\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{n_{cg\cdot}}{NGK}\right)^2 - \sum_{c=1}^{C}\left(\frac{n_{c\cdot\cdot}}{NGK}\right)^2 \quad . \tag{3.7}$$

5. The Probabilistic Model

Assuming that responses in different groups are independent, for each group and each position, the responses $(n_{1gk}, n_{2gk}, \ldots, n_{Cgk})$ follow a multinomial distribution:

$$\Pr\{n_{1gk}, n_{2gk}, \ldots, n_{Cgk}\} = \binom{N}{n_{1gk}\ldots n_{Cgk}}\prod_{c=1}^{C}(p_{cgk})^{n_{cgk}},$$

where $\sum_{c=1}^{C} p_{cgk} = 1$, $p_{cgk} > 0$, $c = 1, \ldots, C$, $k = 1, \ldots, K$ and $g = 1, \ldots, G$

$$\mathrm{E}(n_{cgk}) = Np_{cgk} \qquad \mathrm{Var}(n_{cgk}) = Np_{cgk}(1-p_{cgk})$$

and $\mathrm{Cov}(n_{c_1g_1k_1}, n_{c_2g_2k_2}) = -\delta N p_{c_1g_1k_1} p_{c_2g_2k_2}$ where

$$\delta = \begin{cases} 1 & \text{if } g_1 = g_2 \text{ and } k_1 = k_2 \\ 0 & \text{otherwise} \end{cases}$$

n_{cgk} denotes number of responses in category c at position k for group g and
p_{cgk} the probability of being at category c at position k for group g.

If we assume that the positions are independent, the model is

$$\prod_{g=1}^{G}\prod_{k=1}^{K}\Pr\{(n_{1gk}, n_{2gk}, \ldots, n_{Cgk})\} = \prod_{g=1}^{G}\prod_{k=1}^{K}\binom{N}{n_{1gk}\ldots n_{Cgk}}\prod_{c=1}^{C}(p_{cgk})^{n_{cgk}}$$

Then $\mathbf{V}_g \equiv (n_{1g1} \ldots n_{Cg1}\, n_{1g2} \ldots n_{Cg2} \ldots n_{1gK} \ldots n_{CgK})'$ is a $CK \times 1$ vector

and $\mathbf{V} \equiv (\mathbf{V}_1\, \mathbf{V}_2 \ldots \mathbf{V}_G)'$ is a $GCK \times 1$ vector.

$$\mathrm{E}(\mathbf{V}) \equiv \boldsymbol{\mu} \equiv N\boldsymbol{\mu}_\diamond = N(p_{111} \ldots p_{C11} \ldots p_{1GK} \ldots p_{CGK})' \tag{4.1}$$

Let $\oplus$ denote the direct-sum operation. Then

$$\begin{aligned}\mathrm{Cov}(\mathbf{V}) &\equiv \boldsymbol{\Sigma} \equiv N\boldsymbol{\Sigma}^{\diamond} \\ &= N(\boldsymbol{\Sigma}_{11} \oplus \boldsymbol{\Sigma}_{12} \oplus \cdots \oplus \boldsymbol{\Sigma}_{1K} \oplus \boldsymbol{\Sigma}_{21} \oplus \cdots \oplus \boldsymbol{\Sigma}_{2K} \oplus \cdots \oplus \boldsymbol{\Sigma}_{GK}) \qquad (4.2)\end{aligned}$$

where $\boldsymbol{\Sigma}_{gk}$ is a $C \times C$ matrix of the form

$$\boldsymbol{\Sigma}_{gk} = \mathbf{D}_{gk} - \boldsymbol{\mu}_{\diamond gk}\boldsymbol{\mu}'_{\diamond gk} \tag{4.3}$$

with $\mathbf{D}_{gk}$ being a $C \times C$ diagonal matrix with elements $p_{1gk},\ \ldots,\ p_{Cgk}$ and $\boldsymbol{\mu}_{\diamond gk} = (p_{1gk}\ \ldots\ p_{Cgk})'$.

6. Moments of Diversity Measures

Let $\otimes$ denote the Kronecker product and

$$\mathbf{T} = \frac{1}{(NGK)^2}(\mathbf{U}_{KG} \otimes \mathbf{I}_C) = \frac{1}{(NGK)^2}\mathbf{T}^{\diamond} \tag{5.1}$$

where $\mathbf{U}_{KG}$ is a $KG \times KG$ matrix of 1's, $\mathbf{I}_C$ is the $C \times C$ identity matrix and $\mathbf{T}^{\diamond} \equiv (NGK)^2\mathbf{T}$ is a $CKG \times CKG$ matrix, having $KG \times KG$ partitions with each partition being $C \times C$ identity matrix. Let $\mathbf{M}$ be a $G \times G$ diagonal matrix with diagonal elements $Gn^2_{\cdot g \cdot}(= G(NK)^2$ here), i.e., $\mathbf{M} = G(NK)^2\mathbf{I}_G$. Then $\mathbf{M}^{-1} = \frac{1}{G(NK)^2}\mathbf{I}_G$.

$$\mathbf{W} \equiv \left[(\mathbf{M}^{-1} \otimes \mathbf{U}_K) \otimes \mathbf{I}_C\right] \equiv \frac{G}{(NGK)^2}[(\mathbf{I}_G \otimes \mathbf{U}_K) \otimes \mathbf{I}_C] \equiv \frac{1}{G(NK)^2}\mathbf{W}^{\diamond} \tag{5.2}$$

Then

$$TSI = 1 - \mathbf{V}'\mathbf{T}\mathbf{V} \tag{5.3}$$

$$WSI = 1 - \mathbf{V}'\mathbf{W}\mathbf{V} \tag{5.4}$$

Therefore,

$$BSI = TSI - WSI = \mathbf{V}'(-\mathbf{T} + \mathbf{W})\mathbf{V} = \mathbf{V}'\mathbf{B}\mathbf{V} \tag{5.5}$$

where

$$\begin{aligned}\mathbf{B} &= -\mathbf{T} + \mathbf{W} = \frac{-1}{(NGK)^2}(\mathbf{U}_{KG} \otimes \mathbf{I}_C) + \frac{G}{(NGK)^2}[(\mathbf{I}_G \otimes \mathbf{U}_K) \otimes \mathbf{I}_C] \\ &= \frac{G}{(NGK)^2}\left[\left((\mathbf{I}_G \otimes \mathbf{U}_K) - \frac{1}{G}\mathbf{U}_{KG}\right) \otimes \mathbf{I}_C\right] \equiv \frac{1}{G(NK)^2}\mathbf{B}^{\diamond} \qquad (5.6)\end{aligned}$$

Since $E(\mathbf{V}'\mathbf{T}\mathbf{V}) = \mathrm{trace}(\mathbf{T}\boldsymbol{\Sigma}) + \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu}$,

$$\begin{aligned}E(TSI) &= 1 - \mathrm{trace}(\mathbf{T}\boldsymbol{\Sigma}) - \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu} \\ &= 1 - \frac{1}{NGK} + \frac{1}{N(GK)^2}\sum_{c=1}^{C}\left[\sum_{g=1}^{G}\sum_{k=1}^{K}p^2_{cgk} - Np^2_{c\cdot\cdot}\right]\end{aligned}$$

$$\begin{aligned} E(WSI) &= 1-\text{trace}(\mathbf{W}\boldsymbol{\Sigma})-\boldsymbol{\mu}'\mathbf{W}\boldsymbol{\mu} \\ &= 1-\frac{1}{NK}+\frac{1}{NGK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\left[\sum_{k=1}^{K}p_{cgk}^2-Np_{cg\cdot}^2\right] \end{aligned}$$

$$\begin{aligned} \text{and} \quad E(BSI) &= \frac{G-1}{NGK}+\frac{1}{N(GK)^2}\sum_{c=1}^{C}\left[\sum_{g=1}^{G}\sum_{k=1}^{K}p_{cgk}^2-Np_{c\cdot\cdot}^2\right] \\ &- \frac{1}{NGK^2}\sum_{c=1}^{C}\sum_{g=1}^{G}\left[\sum_{k=1}^{K}p_{cgk}^2-Np_{cg\cdot}^2\right] \end{aligned}$$

Define the population variation within the gth group at the kth position as

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

$H_0 : p_{cgk}=p_{ck}$ for all g implies that

$$I_S(\mathbf{p}_{1k})=I_S(\mathbf{p}_{2k})=\cdots=I_S(\mathbf{p}_{Gk})=I_S(\mathbf{p}_k)\ \ ,$$

i.e., within-group variation at the kth position is the same over all the groups and this implies that

(5.8) $$\|\ \mathbf{p}_{1k}\ \|=\|\ \mathbf{p}_{2k}\ \|=\ \cdots\ \|\ \mathbf{p}_{Gk}\ \|$$

where $\mathbf{p}_{gk}=(p_{1gk}\ p_{2gk}\ \ \ldots\ p_{Cgk})'$ is a $C\times 1$ vector representing the probabilities of belonging to categories $c=1,\ldots,C$ in group g and position k.

If one is interested in the hypothesis stated in (5.8), the hypothesis of homogeneity among the groups ($p_{cgk}=p_{ck}$) is not necessarily true. Here, we consider $H_0 : p_{cgk}=p_{ck}$.

(5.9) $$\mathrm{E}_0(TSI) = 1-\frac{1}{NGK}+\frac{1}{NGK^2}\sum_{c=1}^{C}\left[\sum_{k=1}^{K}p_{ck}^2-NGp_{c\cdot}^2\right]$$

(5.10) $$\mathrm{E}_0(WSI) = 1-\frac{1}{NK}+\frac{1}{NK^2}\sum_{c=1}^{C}\left[\sum_{k=1}^{K}p_{ck}^2-Np_{c\cdot}^2\right]$$

(5.11) $$\mathrm{E}_0(BSI) = \frac{G-1}{NGK}\left[1-\frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K}p_{ck}^2\right]$$

Since $\mathbf{V}$ follows a multinomial distribution, from (4.1) and (4.2), asymptotically,

(5.12) $$\frac{\mathbf{V}}{\sqrt{N}}\xrightarrow{d}\mathrm{N}(\sqrt{N}\boldsymbol{\mu}_{\diamond},\boldsymbol{\Sigma}^{\diamond})$$

where $\boldsymbol{\Sigma}^{\diamond} = \boldsymbol{\Sigma}_1 \oplus \boldsymbol{\Sigma}_2 \oplus \cdots \oplus \boldsymbol{\Sigma}_G$, $\boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_{g1} \oplus \boldsymbol{\Sigma}_{g2} \oplus \cdots \oplus \boldsymbol{\Sigma}_{gK}$, $g = 1, \ldots, G$ and $\boldsymbol{\Sigma}_{gk}$ is given by (4.3). Under H_o, for any $g = 1, \ldots, G$,

$$\boldsymbol{\Sigma}_{gk} = \boldsymbol{\Sigma}_{0k} \quad \text{and} \quad \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}_0^{\star} = \boldsymbol{\Sigma}_{01} \oplus \boldsymbol{\Sigma}_{02} \oplus \cdots \oplus \boldsymbol{\Sigma}_{0K} \tag{5.13}$$

where $\boldsymbol{\Sigma}_{0k}$ is the $C \times C$ matrix

$$\boldsymbol{\Sigma}_{0k} = \mathbf{D}_k - \boldsymbol{\mu}_{\diamond k}\boldsymbol{\mu}_{\diamond k}' \tag{5.14}$$

with $\mathbf{D}_k$ being a $C \times C$ diagonal matrix with elements $p_{1k}, \ \ldots, \ p_{Ck}$ and $\boldsymbol{\mu}_{\diamond k} = (p_{1k} \ \ldots \ p_{Ck})'$.

Therefore, under H_0,

$$\boldsymbol{\Sigma} = N\boldsymbol{\Sigma}^{\diamond} = N\boldsymbol{\Sigma}_0^{\diamond} = N(\mathbf{I}_G \otimes \boldsymbol{\Sigma}_0^{\star}) \tag{5.15}$$

Now,

$$\operatorname{Cov}(BSI, WSI) = \operatorname{Cov}(BSI, TSI) - \operatorname{Var}(BSI) \tag{5.16}$$

$$\operatorname{Cov}(TSI, WSI) = \operatorname{Var}(TSI) - \operatorname{Cov}(TSI, BSI) \tag{5.17}$$

7. The Test Statistic

Note that BSI can be written as

$$BSI = \mathbf{V}'\mathbf{B}\mathbf{V} = \sum_{g=1}^{G}\sum_{c=1}^{C}\left[\frac{n_{cg\cdot}}{NK\sqrt{G}} - \frac{n_{c\cdot\cdot}NK\sqrt{G}}{(NGK)^2}\right]^2 \tag{6.1}$$

Let $\theta_{cgk} = n_{cgk} - Np_{ck} = n_{cgk} - \mathrm{E}_0(n_{cgk})$. Then

$$\theta_{c\cdot\cdot} = \sum_{g=1}^{G}\sum_{k=1}^{K}\theta_{cgk} = n_{c\cdot\cdot} - NG\sum_{k=1}^{K}p_{ck}$$

Also, under H_0, $\boldsymbol{\theta} = (\theta_{111} \ \ldots \ \theta_{Cg1} \ \ldots \ \theta_{CGK})'$ is asymptotically

$$\frac{\boldsymbol{\theta}}{\sqrt{N}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_0^{\diamond}) \tag{6.2}$$

where $\boldsymbol{\Sigma}_0^{\diamond}$ is as in (5.15). So,

$$BSI = \left[\frac{\theta_{cg\cdot}}{NK\sqrt{G}} - \frac{\theta_{c\cdot\cdot}}{(NGK)^2}NK\sqrt{G}\right]^2 = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} = \frac{1}{G(NK)^2}\boldsymbol{\theta}'\mathbf{B}^{\diamond}\boldsymbol{\theta}$$

Hence, under H_0, BSI is asymptotically

$$BSI \sim \sum_{i=1}^{CGK}\lambda_i \left(\chi_1^2\right)_i, \tag{6.3}$$

where $\left(\chi_1^2\right)_i$'s are independent χ^2-random variables with 1 degree of freedom and $\{\lambda_i, i = 1, \ldots, CGK\}$ is the set of characteristic roots of

$$N\mathbf{B}\boldsymbol{\Sigma}_0^\diamond = \frac{1}{NGK^2}\mathbf{B}^\diamond\boldsymbol{\Sigma}_0^\diamond \quad = \quad \frac{1}{NGK^2}\left[\left((\mathbf{I}_G \otimes \mathbf{U}_K) - \frac{1}{G}\mathbf{U}_{KG}\right) \otimes \mathbf{I}_C\right](\mathbf{I}_G \otimes \boldsymbol{\Sigma}_0^\star)$$

by (5.6) and (5.15).

Looking at $\boldsymbol{\Sigma}_{0k}$, it is easy to see that its rank is at most $(C-1)$, because of the restriction that $\sum_{c=1}^{C} p_{ck} = 1$. In fact, the rank of each $\boldsymbol{\Sigma}_{0k}$ is $(C-1)$, since any of its $(C-1)$ columns are linearly independent. Therefore, the rank of $\boldsymbol{\Sigma}_0^\star$ is $K(C-1)$. Further, $\frac{1}{NGK^2}\mathbf{B}^\diamond\boldsymbol{\Sigma}_0^\diamond$ is a $G \times G$ partition matrix with elements the $\boldsymbol{\Sigma}_{0k}$'s matrices premultiplied by some constants ($G-1$ or -1). In order to get the characteristic roots of $\frac{1}{NGK^2}\mathbf{B}^\diamond\boldsymbol{\Sigma}_0^\diamond$ we need to solve the equation

$$\left|\frac{1}{NGK^2}\mathbf{B}^\diamond\boldsymbol{\Sigma}_0^\diamond - \lambda\mathbf{I}_{CKG}\right| = 0 \tag{6.4}$$

Since $\{p_{ck},\ c = 1, \ldots, C\}$ is unknown and need to be estimated, so are $\{\lambda_{ik},\ i = 1, \ldots, C-1\}$. Determining the characteristic roots of a multinomial covariance matrix is not straightforward. Roy et al. (1960) studied this problem without actually presenting the closed-form expression for the roots. The characteristic equation for each k is

$$\left\{1 - \sum_{c=1}^{C}\left(\frac{p_{ck}^2}{p_{ck} - \lambda}\right)\right\}\prod_{c=1}^{C}(p_{ck} - \lambda) = 0 \tag{6.5}$$

It is easy to see that $\lambda = 0$ is a root, but identifying the other roots must proceed numerically.

Now,

$$TSI \quad = \quad 1 - \mathbf{V}'\mathbf{T}\mathbf{V}$$

Since $N\mathbf{T}\boldsymbol{\Sigma}_0^\diamond$ is not idempotent, the distribution of $\mathbf{V}'\mathbf{T}\mathbf{V}$ is not $\chi^2_{(\text{rank}(T), \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu})}$. Under H_0, however,

$$\begin{aligned}\mathbf{V}'\mathbf{T}\mathbf{V} \quad &= \quad \frac{1}{(NGK)^2}\sum_{c=1}^{C} n_{c\cdot\cdot}^2 \;=\; \frac{1}{(NGK)^2}\sum_{c=1}^{C}[\theta_{c\cdot\cdot} + NG\sum_{k=1}^{K} p_{ck}]^2 \\ &= \quad \boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta} + \frac{1}{K^2}\sum_{c=1}^{C} p_{c\cdot}^2 + \mathbf{A}'\boldsymbol{\theta} \end{aligned} \tag{6.6}$$

where $\mathbf{A} = (\mathbf{A}^\star\ \mathbf{A}^\star\ \ldots\ \mathbf{A}^\star)'$ is a $CGK \times 1$ vector and $\mathbf{A}^\star$ is a $1 \times CK$ vector of the form

$$\mathbf{A}^\star = \frac{2}{NGK^2}(p_{1\cdot}\ \ldots\ p_{C\cdot}\ p_{1\cdot}\ \ldots\ p_{C\cdot}\ \ldots\ p_{1\cdot}\ \ldots\ p_{C\cdot})$$

Lemma 1
$\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ and $\mathbf{A}'\boldsymbol{\theta}$ are not independent.

Proof:
$\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta} = \dfrac{1}{(NGK)^2}\boldsymbol{\theta}'\mathbf{T}^\diamond\boldsymbol{\theta}$ and $\mathbf{A}'\boldsymbol{\theta}$ are independent if and only if

$$\mathbf{A}'N\boldsymbol{\Sigma}_0^\diamond\frac{1}{(NGK)^2}\mathbf{T}^\diamond = \mathbf{0} \quad \text{(Searle, 1971).}$$

$$\begin{aligned}
&\frac{1}{N(GK)^2}\mathbf{A}'\boldsymbol{\Sigma}_0^\diamond\mathbf{T}^\diamond \\
&\quad = \frac{1}{N(GK)^2}(\mathbf{A}^\star\ \mathbf{A}^\star\ \ldots\ \mathbf{A}^\star)(\mathbf{I}_G\otimes\boldsymbol{\Sigma}_0^\star)(\mathbf{U}_{KG}\otimes\mathbf{I}_C) \\
&\quad = \frac{G}{N(GK)^2}(\mathbf{A}^\star\boldsymbol{\Sigma}_0^\star(\mathbf{U}_K\otimes\mathbf{I}_C)\ \mathbf{A}^\star\boldsymbol{\Sigma}_0^\star(\mathbf{U}_K\otimes\mathbf{I}_C)\ \ldots\ \mathbf{A}^\star\boldsymbol{\Sigma}_0^\star(\mathbf{U}_K\otimes\mathbf{I}_C))
\end{aligned}$$

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

$$\mathbf{A}^\star\boldsymbol{\Sigma}_0^\star(\mathbf{U}_K\otimes\mathbf{I}_C) \quad = \quad \frac{2}{NGK^2}(\mathbf{a}'\boldsymbol{\Sigma}_{01}\ \mathbf{a}'\boldsymbol{\Sigma}_{02}\ \ldots\ \mathbf{a}'\boldsymbol{\Sigma}_{0K})(\mathbf{U}_K\otimes\mathbf{I}_C)$$

For each k, from (5.14)

$$\mathbf{a}'\boldsymbol{\Sigma}_{0k} = [p_{1k}(p_{1\cdot} - \sum_{c=1}^{C} p_{c\cdot}p_{ck})\ p_{2k}(p_{2\cdot} - \sum_{c=1}^{C} p_{c\cdot}p_{ck})\ \ldots\ p_{Ck}(p_{C\cdot} - \sum_{c=1}^{C} p_{c\cdot}p_{ck})]$$

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

$$\frac{2}{NGK^2}\left(p_{1\cdot}^2 - \sum_{c=1}^{C} p_{c\cdot}\sum_{k=1}^{K} p_{1k}p_{ck}\right) \neq 0$$

Hence, $\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta}$ and $\mathbf{A}'\boldsymbol{\theta}$ are not independent. ■

Now,

$$\boldsymbol{\theta}'\mathbf{T}\boldsymbol{\theta} \sim \sum_{i=1}^{KCG}\lambda_i\left(\chi_1^2\right)_i \quad , \qquad \mathbf{A}'\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{0}, N\mathbf{A}'\boldsymbol{\Sigma}_0^\diamond\mathbf{A})$$

and

$$\mathbf{V}'\mathbf{T}\mathbf{V} \sim \sum_{i=1}^{KCG}\lambda_i\left(\chi_1^2\right)_i + \mathrm{N}(\mathbf{0}, N\mathbf{A}'\boldsymbol{\Sigma}_0^\diamond\mathbf{A}) + \delta_1$$

where $\{\lambda_i,\ i=1,\ldots,CGK\}$ is the set of characteristic roots of $N\mathbf{T}\boldsymbol{\Sigma}_0^\diamond = \dfrac{1}{N(GK)^2}\mathbf{T}^\diamond\boldsymbol{\Sigma}_0^\diamond$ and

$$(6.7) \qquad \delta_1 = \frac{1}{K^2}\sum_{c=1}^{C} p_{c\cdot}^2 = \boldsymbol{\mu}'\mathbf{T}\boldsymbol{\mu} \;\text{ under } H_0$$

As for

$$WSI \;=\; 1-\mathbf{V}'\mathbf{W}\mathbf{V}\ ,$$

under H_0

$$\begin{aligned}
\mathbf{V}'\mathbf{W}\mathbf{V} &= \frac{1}{G(NK)^2}\sum_{g=1}^{G}\sum_{c=1}^{C} n_{cg\cdot}^2 \qquad \text{from (3.5)}\\
&= \frac{1}{G(NK)^2}\sum_{g=1}^{G}\sum_{c=1}^{C}\theta_{cg\cdot}^2 + \frac{2}{NGK^2}\sum_{c=1}^{C}\theta_{c\cdot\cdot}p_{c\cdot} + \frac{1}{GK^2}\sum_{g=1}^{G}\sum_{c=1}^{C}p_{c\cdot}^2\\
(6.8) \qquad &= \boldsymbol{\theta}'\mathbf{W}\boldsymbol{\theta} + \mathbf{A}'\boldsymbol{\theta} + \delta_1 \qquad \text{from (6.7)}
\end{aligned}$$

Again,

$$\mathbf{V}'\mathbf{W}\mathbf{V} = \frac{1}{G(NK)^2}\mathbf{V}'\mathbf{W}^\diamond\mathbf{V} \sim \sum_{i=1}^{CGK}\lambda_i\left(\chi_1^2\right)_i + \mathrm{N}(\mathbf{0}, N\mathbf{A}'\boldsymbol{\Sigma}_0^\diamond\mathbf{A}) + \delta_1$$

where $\{\lambda_i,\ i=1,\ldots,CGK\}$ is the set of characteristic roots of $N\mathbf{W}\boldsymbol{\Sigma}_0^\diamond = \dfrac{1}{NGK^2}\mathbf{W}^\diamond\boldsymbol{\Sigma}_0^\diamond$. Let

$$\begin{aligned}
\theta_1 &\equiv \mathrm{E}_0(BSI) = \frac{G-1}{NGK}\left[1-\frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K}p_{ck}^2\right]\\
\theta_2 &\equiv \mathrm{E}_0(TSI) = 1-\frac{1}{NGK}+\frac{1}{NGK^2}\sum_{c=1}^{C}\left[\sum_{k=1}^{K}p_{ck}^2 - NG\,p_{c\cdot}^2\right]\\
\theta_3 &\equiv \mathrm{E}_0(WSI) = 1-\frac{1}{NK}+\frac{1}{NK^2}\sum_{c=1}^{C}\left[\sum_{k=1}^{K}p_{ck}^2 - N\,p_{c\cdot}^2\right]
\end{aligned}$$

from (5.11), (5.9) and (5.10). The variances are calculated using the following theorem.

Theorem 1 (Searle, 1971)
When $\mathbf{X}$ is $\mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, the rth cumulant of $\mathbf{X}'\mathbf{A}\mathbf{X}$ is

$$K_r(\mathbf{X}'\mathbf{A}\mathbf{X}) = 2^{r-1}(r-1)![\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma})^r + r\boldsymbol{\mu}'\mathbf{A}(\boldsymbol{\Sigma}\mathbf{A})^{r-1}\boldsymbol{\mu}]$$

■

Since $\dfrac{\mathbf{V}}{\sqrt{N}} \sim \mathrm{N}(\sqrt{N}\boldsymbol{\mu}_{\diamond}, \boldsymbol{\Sigma}^{\diamond})$ and $\dfrac{\boldsymbol{\theta}}{\sqrt{N}} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\diamond})$,

$$\begin{aligned}
&(6.9) & \mathrm{Var}(BSI) &= \mathrm{Var}(\mathbf{V}'\mathbf{B}\mathbf{V}) = \mathrm{Var}(\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}) = 2\,\mathrm{trace}(\mathbf{B}N\boldsymbol{\Sigma}^{\diamond})^2 \\
&(6.10) & \mathrm{Var}(TSI) &= \mathrm{Var}(\mathbf{V}'\mathbf{T}\mathbf{V}) = 2\,\mathrm{trace}(\mathbf{T}N\boldsymbol{\Sigma}^{\diamond})^2 + 4N\boldsymbol{\mu}_{\diamond}'\mathbf{T}N\boldsymbol{\Sigma}^{\diamond}\mathbf{T}N\boldsymbol{\mu}_{\diamond} \\
&(6.11) & \mathrm{Var}(WSI) &= \mathrm{Var}(\mathbf{V}'\mathbf{W}\mathbf{V}) = 2\,\mathrm{trace}(\mathbf{W}N\boldsymbol{\Sigma}^{\diamond})^2 + 4N\boldsymbol{\mu}_{\diamond}'\mathbf{W}N\boldsymbol{\Sigma}^{\diamond}\mathbf{W}N\boldsymbol{\mu}_{\diamond}
\end{aligned}$$

and under H_0

$$\begin{aligned}
&(6.12) & \mathrm{Var}_0(BSI) &= 2\,\mathrm{trace}\left(\frac{1}{NGK^2}\mathbf{B}^{\diamond}\boldsymbol{\Sigma}_0^{\diamond}\right)^2 = \frac{2}{(NGK^2)^2}\mathrm{trace}(\mathbf{B}^{\diamond}\boldsymbol{\Sigma}_0^{\diamond})^2 \\
&(6.13) & \mathrm{Var}_0(TSI) &= \frac{2}{N^2(GK)^4}\mathrm{trace}(\mathbf{T}^{\diamond}\boldsymbol{\Sigma}_0^{\diamond})^2 + \frac{4}{N(GK)^4}\boldsymbol{\mu}_{\diamond}'\mathbf{T}^{\diamond}\boldsymbol{\Sigma}_0^{\diamond}\mathbf{T}^{\diamond}\boldsymbol{\mu}_{\diamond} \\
&(6.14) & \mathrm{Var}_0(WSI) &= \frac{2}{(NG)^2K^4}\mathrm{trace}(\mathbf{W}^{\diamond}\boldsymbol{\Sigma}_0^{\diamond})^2 + \frac{4}{NG^2K^4}\boldsymbol{\mu}_{\diamond}'\mathbf{W}^{\diamond}\boldsymbol{\Sigma}_0^{\diamond}\mathbf{W}^{\diamond}\boldsymbol{\mu}_{\diamond}
\end{aligned}$$

Let

$$T_{N,1} \equiv BSI - \theta_1, \quad T_{N,2} \equiv TSI - \theta_2, \quad T_{N,3} \equiv WSI - \theta_3$$

Note that

(i) $\displaystyle\sum_{i=1}^{KCG} \lambda_i \left(\chi_G^2\right)_i = O_p(N^{-1})$

since $\{\lambda_i : i = 1, \ldots, KC\}$ is the set of characteristic roots of $\frac{1}{N(GK)^2}(\mathbf{U}_K \otimes \mathbf{I}_C)\boldsymbol{\Sigma}_0^{\star}$ and $\boldsymbol{\Sigma}_0^{\star} = O(1)$.

(ii) $\mathbf{A}'\boldsymbol{\theta} = O_p(N^{-1/2})$ since $\mathbf{A}'\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{0}, N\mathbf{A}'\boldsymbol{\Sigma}_0^{\diamond}\mathbf{A})$ and $\mathbf{A} = O(N^{-1})$

(iii) $\delta_1 = \dfrac{1}{K^2}\displaystyle\sum_{c=1}^{C} p_{c\cdot}^2 = O(1)$.

Then,

$$\begin{aligned}
T_{N,2} &= 1 - \mathbf{V}'\mathbf{T}\mathbf{V} - \theta_2 \\
&= 1 - \left(O_p(N^{-1}) + O_p(N^{-1/2}) + \frac{1}{K^2}\sum_{c=1}^{C} p_{c\cdot}^2\right) \\
&\quad - \left(1 + O(N^{-1}) - \frac{1}{K^2}\sum_{c=1}^{C} p_{c\cdot}^2\right) \\
&= O_p(N^{-1/2})
\end{aligned}$$

Similarly,

$$T_{N,3} = 1 - \mathbf{V}'\mathbf{W}\mathbf{V} - \theta_3 = O_p(N^{-1/2})$$

and

$$BSI = \mathbf{V}'\mathbf{B}\mathbf{V} = \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} = O_p(N^{-1})$$

Then

$$\begin{aligned} F_1 &\equiv N\left(\frac{BSI}{WSI}\right) = N\left(\frac{BSI}{T_{N,3}+\theta_3}\right) = N\left(\frac{BSI}{\theta_3}\right)\left[1+\frac{T_{N,3}}{\theta_3}\right]^{-1} \\ &= N\left(\frac{BSI}{\theta_3}\right) + O_p(N^{-1/2}) = N\left(\frac{BSI}{\theta_3^\circ}\right) + O_p(N^{-1/2}) \end{aligned}$$

since $T_{N,3} = O_p(N^{-1/2})$, $\dfrac{N(BSI)\,T_{N,3}}{\theta_3^2} = O_p(N^{-1/2})$, $N(BSI) = O_p(1)$ and

$$\theta_3 = 1 - \frac{1}{K^2}\sum_{c=1}^{C} p_{c\cdot}^2 + O(N^{-1}) = \theta_3^\circ + O(N^{-1})$$

By (6.3), asymptotically

$$F_1 = N\frac{BSI}{\theta_3^\circ} \sim \frac{1}{\theta_3^\circ}\sum_{i=1}^{CGK}\lambda_i\left(\chi_1^2\right)_i \tag{6.15}$$

where $\{\lambda_i : 1, \ldots, KGC\}$ is the set of characteristic roots of $\dfrac{1}{NGK^2}\mathbf{B}^\diamond\boldsymbol{\Sigma}_0^\star$. Under H_0, asymptotically

$$\mathrm{E}_0(F_1) \;=\; \frac{N\theta_1}{\theta_3^\circ} = \frac{(G-1)}{GK\theta_3^\circ}\left[1 - \frac{1}{K}\sum_{c=1}^{C}\sum_{k=1}^{K} p_{ck}^2\right]$$

$$\mathrm{Var}_0(F_1) = N^2\left[\frac{\mathrm{Var}_0(BSI)}{(\theta_3^\circ)^2}\right] = \frac{2\mathrm{trace}(\mathbf{B}^\diamond\Sigma_0^\diamond)^2}{(GK^2\theta_3^\circ)^2}$$

Since p_{ck}'s are unknown, one can only get estimates for the λ_i's, i.e., the characteristic roots of $\boldsymbol{\Sigma}_0^\diamond$. To derive the distribution of F_1, estimate $\{p_{ck}\}$, then get the characteristic roots of $\boldsymbol{\Sigma}_0^\diamond$ based on those estimates.

Alternatively, since the terms on the R.H.S. of (6.15) are i.i.d. χ_1^2's, if $\dfrac{\max(\lambda_i)}{\sqrt{\sum_{i=1}^{CGK}\lambda_i^2}} \to 0$, we can apply the C.L.T. for CGK large,

$$K^2\left(F_1 - N\frac{\theta_1}{\theta_3^\circ}\right) \sim N(0,\sigma^2) \tag{6.16}$$

where $\sigma^2 = \dfrac{2\,\mathrm{trace}(\mathbf{B}^\diamond \Sigma_0^\diamond)^2}{(G\theta_3^\circ)^2}$.

Using a similar approach, Light & Margolin (1971) developed an analysis of variance for categorical data (CATANOVA) and these two appraches are equivalent. Extending the CATANOVA approach for several positions, the sum of squares are

$$WSS = \frac{NGK}{2} - \frac{1}{2NK}\sum_{g=1}^{G}\sum_{c=1}^{C} n_{cg\cdot}^2 = \frac{NGK}{2} WSI \quad , \tag{6.17}$$

$$TSS = \frac{NGK}{2} - \frac{1}{2NGK}\sum_{c=1}^{C} n_{c\cdot\cdot}^2 = \frac{NGK}{2} TSI \quad , \tag{6.18}$$

$$\text{and} \quad BSS = \frac{1}{2NGK}\left(G\sum_{g=1}^{G}\sum_{c=1}^{C} n_{cg\cdot}^2 - \sum_{c=1}^{C} n_{c\cdot\cdot}^2 \right) = \frac{NGK}{2} BSI \quad . \tag{6.19}$$

In this setup a test statistic is

$$F_1^\star = \frac{BSS/(G-1)}{WSS/(NGK-G)} = \frac{BSI/(G-1)}{WSI/(NGK-1)} = \frac{(NGK-G)}{N(G-1)} F_1$$

8. Power of the Test

Let us now consider an alternative hypothesis, i.e.,

$$p_{cgk} = \frac{1}{\sqrt{N}}\gamma_{cgk} + p_{ck} \quad .$$

Thus, $\gamma_{cgk} = 0$ yields the null hypothesis $H_0 : p_{cgk} = p_{ck}$. The interest here is in the case where $\gamma_{cgk} \neq 0$. Then,

$$\theta_{cgk} = n_{cgk} - N\left(\frac{1}{\sqrt{N}}\gamma_{cgk} + p_{ck}\right)$$

$$\theta_{cg\cdot} = n_{cg\cdot} - \sqrt{N}\gamma_{cg\cdot} - Np_{c\cdot} \quad , \quad \theta_{c\cdot\cdot} = n_{c\cdot\cdot} - \sqrt{N}\gamma_{c\cdot\cdot} - NGp_{c\cdot}$$

and

$$\begin{aligned} BSI \quad &= \quad \mathbf{V}'\mathbf{B}\mathbf{V} = \sum_{g=1}^{G}\sum_{c=1}^{C}\left[\frac{n_{cg\cdot}}{NG\sqrt{G}} - \frac{n_{c\cdot\cdot}NK\sqrt{G}}{(NGK)^2}\right]^2 \\ &= \quad \sum_{g=1}^{G}\sum_{c=1}^{C}\left[\frac{\theta_{cg\cdot} + Np_{c\cdot} + \sqrt{N}\gamma_{cg\cdot}}{NK\sqrt{G}} - \frac{(\theta_{c\cdot\cdot} + NGp_{c\cdot} + \sqrt{N}\gamma_{c\cdot\cdot})NK\sqrt{G}}{(NGK)^2}\right]^2 \end{aligned}$$

$$
\begin{aligned}
&= \sum_{g=1}^{G}\sum_{c=1}^{C}\left[\frac{\theta_{cg\cdot}}{NK\sqrt{G}} - \frac{\theta_{c\cdot\cdot}\sqrt{G}}{NKG^2} + \frac{\gamma_{cg\cdot}}{K\sqrt{NG}} - \frac{\gamma_{c\cdot\cdot}\sqrt{G}}{KG^2\sqrt{N}}\right]^2 \\
&= \sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{\theta_{cg\cdot}}{NK\sqrt{G}} - \frac{\theta_{c\cdot\cdot}NK\sqrt{G}}{(NGK)^2}\right)^2 \\
&\quad + 2\sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{\theta_{cg\cdot}}{NK\sqrt{G}} - \frac{\theta_{c\cdot\cdot}NK\sqrt{G}}{(NGK)^2}\right)\left(\frac{\gamma_{cg\cdot}}{K\sqrt{NG}} - \frac{\gamma_{c\cdot\cdot}\sqrt{G}}{KG^2\sqrt{N}}\right) \\
&\quad + \sum_{g=1}^{G}\sum_{c=1}^{C}\left(\frac{\gamma_{cg\cdot}}{K\sqrt{NG}} - \frac{\gamma_{c\cdot\cdot}\sqrt{G}}{KG^2\sqrt{N}}\right)^2 \\
&= \boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} + (\mathbf{A}_1 - \mathbf{A}_2)'\boldsymbol{\theta} + N^2(\mathbf{A}_1 - \mathbf{A}_2)'(\mathbf{A}_1 - \mathbf{A}_2)
\end{aligned}
$$

where $\mathbf{B}$ is as in (5.6), $\boldsymbol{\theta} = (\theta_{111}\ \ldots\ \theta_{Cg1}\ \ldots\ \theta_{CGK})'$, $\mathbf{A}_1$ and $\mathbf{A}_2$ are $CGK \times 1$ vectors of the form

$$
\begin{aligned}
&\mathbf{A}_1 \equiv \frac{2}{K^2GN^{3/2}}(\mathbf{A}^{\star}_{11}\ \ldots\ \mathbf{A}^{\star}_{1G}) \quad \text{with} \\
&\quad \mathbf{A}^{\star}_{1g} = (\gamma_{1g\cdot}\ \ldots\ \gamma_{Cg\cdot}\ \gamma_{1g\cdot}\ \ldots\ \gamma_{Cg\cdot}\ \ldots\ \gamma_{1g\cdot}\ \ldots\ \gamma_{Cg\cdot}) \ \text{ for each } g = 1, \ldots, G \\
&\mathbf{A}_2 \equiv \frac{2}{(KG)^2N^{3/2}}(\mathbf{A}^{\star}_{2}\ \ldots\ \mathbf{A}^{\star}_{2}) \quad \text{with} \\
&\quad \mathbf{A}^{\star}_{2} = (\gamma_{1\cdot\cdot}\ \ldots\ \gamma_{C\cdot\cdot}\ \gamma_{1\cdot\cdot}\ \ldots\ \gamma_{C\cdot\cdot}\ \ldots\ \gamma_{1\cdot\cdot}\ \ldots\ \gamma_{C\cdot\cdot})
\end{aligned}
$$

Recall that $\dfrac{\boldsymbol{\theta}}{\sqrt{N}} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}^{\diamond})$ and $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta} \sim \sum_{i=1}^{CGK} \lambda_i \left(\chi^2_1\right)_i$, with λ_i's being the characteristic roots of $\dfrac{1}{NGK^2}\mathbf{B}^{\diamond}\boldsymbol{\Sigma}^{\diamond}$, where

(7.1) $$\boldsymbol{\Sigma}^{\diamond} = (\boldsymbol{\Sigma}_{11} \oplus \boldsymbol{\Sigma}_{12} \oplus \cdots \oplus \boldsymbol{\Sigma}_{1K} \oplus \boldsymbol{\Sigma}_{21} \oplus \cdots \oplus \boldsymbol{\Sigma}_{GK})$$

and $\boldsymbol{\Sigma}_{gk}$ is as in (4.3). Now, let $\mathbf{A}^{\diamond}_1 = N^{3/2}\mathbf{A}_1$ and $\mathbf{A}^{\diamond}_2 = N^{3/2}\mathbf{A}_2$. Then

$$N(\mathbf{A}_1 - \mathbf{A}_2)'\boldsymbol{\theta} \sim \mathrm{N}(\mathbf{0}, (\mathbf{A}^{\diamond}_1 - \mathbf{A}^{\diamond}_2)'\boldsymbol{\Sigma}^{\diamond}(\mathbf{A}^{\diamond}_1 - \mathbf{A}^{\diamond}_2))$$

Lemma 2
$\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ and $\mathbf{A}_1'\boldsymbol{\theta}$ are not independent.

Proof:

$\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ and $\mathbf{A}_1'\boldsymbol{\theta}$ are independent if and only if $\mathbf{A}_1'N\boldsymbol{\Sigma}^{\diamond}\mathbf{B} = \mathbf{0}$.

$$
\begin{aligned}
\mathbf{A}_1'N\boldsymbol{\Sigma}^{\diamond}\mathbf{B} &= \frac{2}{K^2GN^{3/2}}(\mathbf{A}^{\star}_{11}\ \ldots\ \mathbf{A}^{\star}_{1G})\frac{1}{NGK^2}\boldsymbol{\Sigma}^{\diamond}\left[\left((\mathbf{I}_G \otimes \mathbf{U}_K) - \frac{1}{G}\mathbf{U}_{KG}\right) \otimes \mathbf{I}_C\right] \\
&= \frac{2}{K^3G^2N^{5/2}}(\mathbf{A}^{\star}_{11}\boldsymbol{\Sigma}_1\ \ldots\ \mathbf{A}^{\star}_{1G}\boldsymbol{\Sigma}_G)\left[\left((\mathbf{I}_G \otimes \mathbf{U}_K) - \frac{1}{G}\mathbf{U}_{KG}\right) \otimes \mathbf{I}_C\right]
\end{aligned}
$$

Let $\mathbf{a}^{\star}_{1g} = (\gamma_{1g\cdot}\ \dots\ \gamma_{Cg\cdot})$ be a $1 \times C$ vector and $\mathbf{A}^{\star}_{1g} = (\mathbf{a}^{\star}_{1g}\ \dots\ \mathbf{a}^{\star}_{1g})$. $\mathbf{A}^{\star}_{1g}\boldsymbol{\Sigma}_g$ can be written as

$$\mathbf{A}^{\star}_{1g}\boldsymbol{\Sigma}_g \quad = \quad (\mathbf{a}_{1g}\boldsymbol{\Sigma}_{g1}\ \mathbf{a}_{1g}\boldsymbol{\Sigma}_{g2}\ \dots\ \mathbf{a}_{1g}\boldsymbol{\Sigma}_{gK}) \quad \text{for each } g = 1, \dots, G$$

Now,

$$\mathbf{a}_{1g}\boldsymbol{\Sigma}_{1k} = \left[p_{1gk}(\gamma_{1g\cdot} - \sum_{c=1}^{C} \gamma_{cg\cdot}\, p_{cgk})\ \dots\ p_{Cgk}(\gamma_{Cg\cdot} - \sum_{c=1}^{C} \gamma_{cg\cdot}\, p_{cgk}) \right]$$

for each $g = 1, \dots, G$ and $k = 1, \dots, K$. The first element of $\mathbf{A}_1' N \boldsymbol{\Sigma}^{\diamond}\mathbf{B}$ is

$$\frac{2}{K^3 G^2 N^{5/2}} \left(\sum_{k=1}^{K} p_{11k}(\gamma_{11\cdot} - \sum_{c=1}^{C} \gamma_{c1\cdot}\, p_{c1k}) - \frac{1}{G} \sum_{g=1}^{G} \sum_{k=1}^{K} p_{1gk}(\gamma_{1g\cdot} - \sum_{c=1}^{C} \gamma_{cg\cdot}\, p_{cgk}) \right) \neq 0$$

■

Since $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ and $\mathbf{A}_1'\boldsymbol{\theta}$ are not independent, $\boldsymbol{\theta}'\mathbf{B}\boldsymbol{\theta}$ and $(\mathbf{A}_1 - \mathbf{A}_2)'\boldsymbol{\theta}$ are also not independent. So, the distribution of $\mathbf{V}'\mathbf{B}\mathbf{V}$ is not the convolution of a linear combination of χ^2-random variables and a normal distribution.

Let $\mathbf{A} = \mathbf{A}_1 - \mathbf{A}_2$, then

$$\mathbf{V}'\mathbf{B}\mathbf{V} = (\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A})'(\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A}) + \frac{1}{4}\mathbf{A}'(4N^2\mathbf{I} - \mathbf{B}^{-1})\mathbf{A}$$

and let $\mathbf{X} = (\mathbf{B}^{1/2}\boldsymbol{\theta} + \frac{1}{2}\mathbf{B}^{-1/2}\mathbf{A})$, $\boldsymbol{\Gamma} = \dfrac{1}{GNK^2}(\mathbf{B}^{\diamond})^{1/2}\boldsymbol{\Sigma}^{\diamond}[(\mathbf{B}^{\diamond})^{1/2}]'$ and $\boldsymbol{\mu}_B = \frac{1}{2}\mathbf{B}^{1/2}\mathbf{A}$. Then,

$$\sqrt{N}\mathbf{X} \sim \mathrm{N}\left(\sqrt{N}\boldsymbol{\mu}_B,\, N\boldsymbol{\Gamma}\right)$$

Let $\mathbf{P}$ be an orthogonal matrix (i.e., $\mathbf{P}'\mathbf{P} = \mathbf{I}$) such that $\mathbf{P}\boldsymbol{\Gamma}\mathbf{P}' = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix, and

$$\mathbf{Y} = \mathbf{P}\mathbf{X} \Rightarrow \mathbf{X} = \mathbf{P}'\mathbf{Y}$$

Then,

$$\mathbf{Y} \sim \mathrm{N}(\mathbf{P}\boldsymbol{\mu}_B, \boldsymbol{\Lambda})$$

and

$$\mathbf{X}'\mathbf{X} = \mathbf{Y}'\mathbf{P}\mathbf{P}'\mathbf{Y} = \mathbf{Y}'\mathbf{Y} \sim \sum_{i=1}^{G} \lambda_i \left(\chi^2_1(\delta_i)\right)_i \tag{7.2}$$

where λ_i's are the diagonal elements of $\boldsymbol{\Lambda}$, $\delta_i = \dfrac{a_i^2}{\lambda_i}$, a_i is the ith row of the vector $\frac{1}{2}\mathbf{P}\mathbf{B}^{-1/2}\mathbf{A}$, which is a linear combination of the γ_{cgk}'s.

Let c be the constant $\frac{1}{4}\mathbf{A}'(4N^2\mathbf{I} - \mathbf{B}^{-1})\mathbf{A}$. Then,

$$\Pr(F_1 \geq u) = \Pr\left(\frac{N(\mathbf{X}'\mathbf{X} + c)}{\theta_3^\circ} \geq u\right) = \Pr\bigl(N\mathbf{X}'\mathbf{X} \geq \theta_3^\circ u - Nc\bigr) \tag{7.3}$$

Note that $Nc = O(1)$ since $\mathbf{B}^{-1} = O(N^2)$ and $\mathbf{A}'\mathbf{A} = O(N^{-3})$. As the noncentrality parameter δ_i increases, the distribution of each of the noncentral χ^2-random variables shifts to the right, therefore the probability in (7.3) goes to 1 and the power of the test converges to 1.

9. Simulations

In order to look at the suitability of the asymptotic distributions, we generated N sequences, with K positions each, in G groups and computed the test statistic $F_1^\star$ and the standardized $F_1^\star$ (i.e, $K\left(F_1^\star - \frac{a_1}{a_2^\star}\right)/\sigma_\star$). We performed 500 simulations (i.e., the above procedure was repeated 500 times). In Table 3 the data were generated using the same probability across positions, i.e., $p_{ck} = p_c$, while in Table 4, different probabilities were used across positions. The tables show the number below and above some quantiles of the standard normal distribution.

10. Data Analysis

The data set consists of two groups (subtype B and not B) with 46 sequences each. The nucleotide sequences are all from independent individuals. There are therefore four categories. After aligning the sequences and discarding the positions with no change, we end up with 155 positions.

Looking at the simulations results we see that our data set is not large enough for the asymptotic results to apply. So, we call upon resampling techniques, such as the bootstrap. Here is a summary of the procedure:

1. Estimate p_{ck} from the data, i.e., $\hat{p}_{ck} = \frac{n_{c1k}+n_{c2k}}{2N}$ and compute the statistic F_1.
2. Generate $N = 46$ sequences, with $K = 155$ positions each, in each of $G = 2$ groups, using $\hat{p}_{ck}$.
3. Recompute the test statistic F_1 from the generated data and store it.
4. Repeat steps 2 and 3 1,000 times.

The p-value is then $\dfrac{\#F_1's \geq F_1obs}{1000}$.

The results are

$$WSI = 0.7005 \qquad TSI = 0.7007 \qquad BSI = 0.0002 \quad \text{and} \quad F_1obs = 0.011$$

The percentiles of the bootstrap distribution are given in Table 5 and the observed p-value is less than 1/1001. This means that relative to the within-clade variation, there is significant variability between the two clades.

References

[1] R J Anderson and J R Landis. CATANOVA for multidimensional contingency tables: Nominal-scale response. *Communications in Statistics - Theory and Methods*, 1980.

[2] R J Anderson and J R Landis. CATANOVA for multidimensional contingency tables: Ordinal-scale response. *Communications in Statistics - Theory and Methods*, 11(3):257–270, 1982.

[3] C W Gini. Variabilita e Mutabilita. *Studi Economico-Giuridici della R. Universita di Cagliari*, 3(2):3–159, 1912.

[4] F A Graybill. *An Introduction to Linear Statistical Models.* McGraw-Hill, New York, 1961.

[5] W Hoeffding. A class of statistics with asymptotically normal distribution. *Ann. of Math. Stat.*, 19:293–325, 1948.

[6] R J Light and B H Margolin. An analysis of variance for categorical data. *Journal of the American Statistical Association*, 66:534–544, 1971.

[7] R J Light and B H Margolin. An analysis of variance for categorical data II: Small sample comparisons with chi square and other competitors. *Journal of the American Statistical Association*, 69:755–764, 1974.

[8] H P Pinheiro. *Modelling Variability in the HIV Genome.* PhD thesis, University of North Carolina, December 1997. Mimeo Series No. 2186T.

[9] C R Rao. *Linear Statistical Inference and Its Applications.* John Wiley & Sons, second edition, 1973.

[10] S N Roy, B G Greenberg, and A E Sarhan. Evaluation of determinants, characteristic equations, and their roots for a class of patterned matrices. *Journal of the Royal Statistical Society, Ser.* **B**, 22(2):348–359, 1960.

[11] S R Searle. *Linear Models.* John Wiley & Sons, 1971.

[12] S R Searle. *Matrix Algebra Useful for Statistics.* John Wiley & Sons, 1982.

[13] E H Simpson. The Measurement of Diversity. *Nature*, 163:688, 1949.

Table 1: Contingency Table (K positions)

Group	Position	1	2	...	C	Total
1	1	n_{111}	n_{211}	...	n_{C11}	$n_{\cdot 11} = N$
1	2	n_{112}	n_{212}	...	n_{C12}	$n_{\cdot 12} = N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
1	K	n_{11K}	n_{21K}	...	n_{C1K}	$n_{\cdot 1K} = N$
Total		$n_{11\cdot}$	$n_{21\cdot}$	...	$n_{C1\cdot}$	$n_{\cdot 1\cdot} = NK$
2	1	n_{121}	n_{221}	...	n_{C21}	$n_{\cdot 21} = N$
2	2	n_{122}	n_{222}	...	n_{C22}	$n_{\cdot 22} = N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
2	K	n_{12K}	n_{22K}	...	n_{C2K}	$n_{\cdot 2K} = N$
Total		$n_{12\cdot}$	$n_{22\cdot}$	...	$n_{C2\cdot}$	$n_{\cdot 2\cdot} = NK$
...	...	...	...	...	...	...
G	1	n_{1G1}	n_{2G1}	...	n_{CG1}	$n_{\cdot G1} = N$
G	2	n_{1G2}	n_{2G2}	...	n_{CG2}	$n_{\cdot G2} = N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$
G	K	n_{1GK}	n_{2GK}	...	n_{CGK}	$n_{\cdot GK} = N$
Total		$n_{1G\cdot}$	$n_{2G\cdot}$	...	$n_{CG\cdot}$	$n_{\cdot G\cdot} = NK$
TOTAL		$n_{1\cdot\cdot}$	$n_{2\cdot\cdot}$	...	$n_{C\cdot\cdot}$	$n_{\cdot\cdot\cdot} = NGK$

Table 2: Summary of the Data (one position)

	Group				
Sequence	1	2	3	...	G
1	x_1^1	x_1^2	x_1^3	...	x_1^G
2	x_2^1	x_2^2	x_2^3	...	x_2^G
3	x_3^1	x_3^2	x_3^3	...	x_3^G
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
N	x_N^1	x_N^2	x_N^3	...	x_N^G

Table 3: Results of Simulations for Diversity Measures ($p_{ck} = p_c$)

			Percentiles of the Std. Normal Dist.							
N	K	G	1	2.5	5	10	90	95	97.5	99
20	10	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	52	$36^{\star}$	$27^{\star\star}$	$16^{\star\star}$
20	10	5	$0^{\star}$	$1^{\star\star}$	$2^{\star\star}$	$20^{\star\star}$	48	30	$21^{\star}$	$16^{\star\star}$
20	50	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	59	$49^{\star\star}$	$38^{\star\star}$	$27^{\star\star}$
50	10	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	$67^{\star}$	$43^{\star\star}$	$28^{\star\star}$	$19^{\star\star}$
50	10	5	$0^{\star}$	$0^{\star\star}$	$4^{\star\star}$	$33^{\star}$	51	34	$23^{\star\star}$	$13^{\star\star}$
50	50	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	59	$41^{\star\star}$	$32^{\star\star}$	$27^{\star\star}$
100	10	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	52	$38^{\star}$	$30^{\star\star}$	$24^{\star\star}$
100	10	5	$0^{\star}$	$1^{\star\star}$	$5^{\star\star}$	$26^{\star\star}$	63	$38^{\star}$	$26^{\star\star}$	$11^{\star}$
100	10	10	$0^{\star}$	$4^{\star}$	$11^{\star}$	40	42	30	14	$10^{\star}$
100	50	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	49	$36^{\star}$	$21^{\star}$	$16^{\star\star}$
100	50	5	$0^{\star}$	$0^{\star\star}$	$3^{\star\star}$	$23^{\star\star}$	56	32	18	$12^{\star\star}$
200	10	5	$0^{\star}$	$0^{\star\star}$	$9^{\star\star}$	$30^{\star}$	48	32	$24^{\star\star}$	$18^{\star\star}$
200	10	10	1	$3^{\star}$	$14^{\star}$	40	52	31	17	7

$\star$ between 2SD and 3SD or -2SD and -3SD.

$\star\star$ greater than 3SD or smaller than -3SD.

Table 4: Results of Simulations for Diversity Measures ($p_{ck} \neq p_c$)

			Percentiles of the Std. Normal Dist.							
N	K	G	1	2.5	5	10	90	95	97.5	99
20	10	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	59	$36^{\star}$	$23^{\star\star}$	$16^{\star\star}$
20	10	5	$0^{\star}$	$0^{\star\star}$	$4^{\star\star}$	$24^{\star\star}$	54	$37^{\star}$	$22^{\star}$	$13^{\star\star}$
20	50	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	63	$39^{\star}$	$29^{\star\star}$	$20^{\star\star}$
50	10	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	60	$36^{\star}$	$28^{\star\star}$	$19^{\star\star}$
50	10	5	$0^{\star}$	$1^{\star\star}$	$6^{\star\star}$	43	51	33	17	10
50	50	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	53	$37^{\star}$	$24^{\star\star}$	$14^{\star\star}$
100	10	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	53	$42^{\star\star}$	$31^{\star\star}$	$21^{\star\star}$
100	10	5	$0^{\star}$	$1^{\star\star}$	$10^{\star\star}$	$27^{\star\star}$	41	28	19	$12^{\star\star}$
100	10	10	1	3	$13^{\star}$	49	53	30	18	7
100	50	2	$0^{\star}$	$0^{\star\star}$	$0^{\star\star}$	$0^{\star\star}$	46	28	$22^{\star}$	$17^{\star\star}$
100	50	5	$0^{\star}$	$1^{\star\star}$	$9^{\star\star}$	$28^{\star\star}$	57	$35^{\star}$	$22^{\star}$	$10^{\star}$
200	10	5	$0^{\star}$	$1^{\star\star}$	$8^{\star\star}$	$34^{\star}$	54	30	18	$12^{\star\star}$
200	10	10	1	8	17	40	54	30	17	9

$\star$ between 2SD and 3SD or -2SD and -3SD.

$\star\star$ greater than 3SD or smaller than -3SD.

Table 5: Percentiles of the Bootstrap Dist. for Diversity Measures

90%	95%	97.5%	99%	99.5%	99.9%
0.0015	0.0020	0.0023	0.0027	0.0030	0.0040