Analysis of Variance for Binary Data in Unbalanced Designs

Roberta de Souza
Departamento de Pesquisa Farmacêutica,
Grupo EMS-Sigma Pharma - SP, Brazil

Hildete Prisco Pinheiro
Departamento de Estatística,
Universidade Estadual de Campinas,
Caixa Postal 6065, 13083-970 Campinas, SP - Brazil

Cibele Queiroz da Silva Departamento de Estatística,
Universidade Federal de Minas Gerais - MG, Brazil

Sérgio Furtado dos Reis Departamento de Parasitologia,
Universidade Estadual de Campinas - SP, Brazil

Summary

In the study of genetic divergence among organisms, generally the analysis is done directly from the DNA molecule. Therefore, a possible outcome is binary (dominant or recessive phenotype). Comparison of groups of molecular data is a great interest in molecular genetics and evolutionary biology. Some work have been done on analysis of variance for genetic data (Weir, 1990; Pinheiro et al., 2000; Pinheiro et al., 2001; Pinheiro et al., 2002 and others). Weir (1990) proposed a genetic diversity measure, the *heterozygosity*, and developed an analysis of variance for binary data in a balanced design. Here, we extend the work of Weir developing an analysis of variance for binary data with the purpose of comparing groups in unbalanced designs. In order to test the null hypothesis of homogeneity among groups, the asymptotic distribution of the test statistic was found. 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.

Key words: Analysis of variance; Binary data; Bootstrap; Asymptotic distribution; Molecular data; Statistical genetics, RAPD.

1 Introduction

Random amplified polymorphic DNA molecular markers are obtained using a single random oligonucleotide primer in a polymerase chain reaction (PCR). These primers are short and therefore there is a high probability that the genome will contain several priming sites at varying distances from one another that are in an inverted orientation (Williams, 1991; Welsh et al., 1991). The main advantages of RAPD markers include suitability to probe anonymous genomes, applicability to problems where only minute amounts of DNA are available, and efficiency and low cost. The amplification profile of products are resolved on agarose gels with RAPD molecular markers behaving as dominant markers, with dominant homozygous and heterozygous represented as the phenotype band-present and the recessive homozygous represented as the phenotype band-absent (Williams, 1991; Welsh et al., 1991).

Molecular markers targeted by arbitrary primers that generate randomly amplified polymorphic DNA (RAPD) have been increasingly employed with success to quantify and describe patterns of genetic variation within populations and to partition genetic variation among populations of animals and plants (Comes and Abbott, 2000). These markers have also proven instrumental to infer patterns of population structure with important implications for evolutionary and conservation biology (Haig et al., 1994; Souza et al., 2002).

The study of species evolution can be characterized by extensions and causes of genetic variation. There are many different ways to measure genetic variation; among them one can think on the proportion of heterozygous in a population, the *heterozygosity*, since each individual carries different alleles, which represents the existence of variation. The continuous presence of different homozygous also can result in variation; for those situa-

tions the genetic diversity is an appropriate measure (Weir, 1990). Genetic differences can also be encountered by direct molecular analysis of DNA. In this case the variation can be measured by comparison of nucleotides (Pinheiro et al., 2000, Pinheiro et al., 2001).

The main interest here is the comparison of groups of different sizes, when the response variable is categorical (binary, in this particular case). In the classical analysis of variance this comparison is done when the response variable is continuous. We would like to develop an analysis of variance when the outcome variable is binary and the samples are unbalanced. For example, one of the techniques to detect genetic polymorphism, for the comparison of groups, is the class of molecular markers RAPD, where polymorphism is detected through a binary outcome (dominant or recessive phenotype).

Weir (1990) proposed the observed *heterozygosity* as a measure of diversity and a table of analysis of variance for binary data in balanced designs was developed. In our case, the groups have different sizes and we extended some of his results of the table of analysis of variance for unbalanced designs (Section 2). In Section 3 a test statistic and its asymptotic distribution are developed to assess homogeneity among groups of binary unbalanced data. The power of the test is discussed in Section 4 and the paper closes with an application of the test statistic to real data in Section 5.

2 The ANOVA Table for Binary Data

Using a similar measure of diversity, which is defined as the proportion of dominant phenotypes, an analysis of variance was developed considering that the loci are randomly sampled, as it is the case of the class of markers RAPD. By the nature of this RAPD marker, it does not make any biological sense to evaluate the contribution of the loci. Therefore, the table of analysis of variance presented here does not consider the loci effect (this effect will be incorporated in the residual).

Making an analogy to the ANOVA table for heterozygosity (Weir, 1990) with the

RAPD markers, we have,

$$X_{gik} = \begin{cases} 1, & \text{if the } i\text{-th individual of the } g\text{-th population is dominant} \\ & \text{at locus } k \text{ (band present).} \\ 0, & \text{elsewhere (band absent).} \end{cases}$$

As mentioned earlier, using RAPD markers, one can consider that the loci are randomly sampled in the individuals and populations and, since they are also randomly sampled without any guaranty that the loci are the same for all the individuals, we decided to incorporate the effect of locus in the residual term.

Table 1 shows the ANOVA table for RAPD data.

Table 1: Analysis of Variance for RAPD Data in Unbalanced Designs

Source of Variation	d.f.	Sum of Squares	E(SS)
Population	$G-1$	PSS	$E\left(\frac{PSS}{G-1}\right)$
Individuals within populations	$\sum_g (N_g - 1)$	ISS	$E\left(\frac{ISS}{N_T-G}\right)$
Residual	$(K-1)N_T$	RSS	$E\left(\frac{RSS}{(K-1)N_T}\right)$
Total	$KN_T - 1$	TSS	

The Population Sum of Squares (PSS), Individual Sum of Squares (ISS), Residual Sum of Squares (RSS) and the Total Sum of Squares (TSS) are as follows

$$\text{PSS} = \sum_{g=1}^{G} KN_g \left(\bar{X}_{g\cdot\cdot} - \bar{X}_{\cdot\cdot\cdot}\right)^2 \tag{2.1}$$

$$\text{ISS} = \sum_{g=1}^{G} \sum_{i=1}^{N_g} K \left(\bar{X}_{gi\cdot} - \bar{X}_{g\cdot\cdot}\right)^2, \tag{2.2}$$

$$\text{TSS} = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\sum_{k=1}^{K}(X_{gik} - \bar{X}_{...})^2 \quad \text{and} \quad \text{RSS} = \text{TSS} - \text{PSS} - \text{ISS} \tag{2.3}$$

3 The Test Statistic and its Asymptotic Distribution

Now, one would like to develop a test statistic to test the hypothesis of homogeneity among groups (or populations). Then, the asymptotic distribution of this test statistic will be of interest.

The outcomes one obtains from RAPD markers are like random vectors of binary data. Considering that the loci are independent and that the groups and individuals are also independent, the response variable X_{gik} follows a Bernoulli distribution, i.e.,

$$P(X_{gik} = x_{gik}) = p_{gk}^{x_{gik}}(1 - p_{gk})^{(1-x_{gik})}\mathbf{I}_{\{0,1\}}(x_{gik}), \tag{3.1}$$

where p_{gk} is the probability that an individual of population g be dominant at locus k, $i = 1, \ldots, N_g$; g $= 1, \ldots, G$; $k = 1, \ldots, K$. Therefore, $\text{E}(X_{gik}) = p_{gk}$ and $\text{Var}(X_{gik}) = p_{gk}(1 - p_{gk})$.

Note that for RAPD markers, $\bar{X}_{gi\cdot}$ represents the proportion, of dominant phenotype in individual i of group g, $\bar{X}_{g\cdot\cdot}$ represents the proportion of dominant phenotype in group g and $\bar{X}_{...}$ is the general proportion (or mean) of dominant phenotype in the whole sample:

$$\bar{X}_{gi\cdot} = \frac{X_{gi\cdot}}{K}, \qquad \bar{X}_{g\cdot\cdot} = \frac{X_{g\cdot\cdot}}{KN_g} \qquad \text{e} \qquad \bar{X}_{...} = \frac{\sum_g N_g \bar{X}_{g\cdot\cdot}}{N_T}.$$

As our interest is to test the hypothesis of homogeneity among groups, i.e., $H_0 : p_{gk} = p_k$, for all g, observing Table 1 we propose as the test statistic $F = MSP/MSI$, where

$$MSP = \frac{PSS}{G-1} \qquad \text{and} \qquad MSI = \frac{ISS}{N_T - G},$$

with PSS being the population sum of squares given in (2.1), which measures the variability among populations; ISS the sum of squares of individuals within population, which measures the variability among individuals within a group (population), and $G - 1$ and

$N_T - G$ are, respectively, the degrees of freedom for populations and individuals within population.

In order to obtain the asymptotic distribution of the statistic F, one needs to find first the asymptotic distribution of the sum of squares of the population effect. Then, by (2.1),we have that PSS is a function of the mean number of dominant phenotypes for the g-th group ($\bar{X}_{g\cdot\cdot}$) and the total number of dominant phenotypes in the g-th group can be written as

$$X_{g\cdot\cdot} = \sum_{i=1}^{N_g} \sum_{k=1}^{K} X_{gik},$$

by model (3.1),

$$\mathrm{E}(X_{g\cdot\cdot}) = N_g \sum_k p_{gk} \qquad \text{and} \qquad \mathrm{Var}(X_{g\cdot\cdot}) = N_g \sum_k p_{gk}(1-p_{gk}).$$

As X_{gik} are independent, but not identically distributed random variables, one will use the Central Limit Theorem of Liapunov for independent random variables (James, 1996).

In this case, for a given population g, $X_{g11}, \ldots, X_{g1K}, \ldots, X_{gN_g1}, \ldots, X_{gN_gK}$, $g = 1, \ldots, G$, are independent random variables, such that $\mathrm{E}(X_{gik}) = p_{gk}$, $\mathrm{Var}(X_{gik}) = p_{gk}(1-p_{gk})$, $S_n = X_{g\cdot\cdot}$ e $s_n^2 = \mathrm{Var}(X_{g\cdot\cdot})$, where $n = KN_g$. Therefore, verifying the Liapunov condition, we have:

For $\delta = 1$ and $0 < p_{gk} < 1$,

$$\frac{1}{s_n^3} \sum_{i=1}^{N_g} \sum_{k=1}^{K} \mathrm{E}|X_{gik} - \mu_k|^3 \quad = \quad \frac{\sum_k p_{gk}(1-p_{gk})(1-2p_{gk}+2p_{gk}^2)}{\sum_k p_{gk}(1-p_{gk}) \left(\sqrt{N_g \sum_k p_{gk}(1-p_{gk})}\right)}$$

Note that

$$\frac{1}{2} \leq 1 - 2p_{gk} + 2p_{gk}^2 < 1 \Rightarrow \sum_{k=1}^{K} p_{gk}(1-p_{gk})(1-2p_{gk}+2p_{gk}^2) < \sum_{k=1}^{K} p_{gk}(1-p_{gk})$$

Then,

$$\frac{\sum_k p_{gk}(1-p_{gk})(1-2p_{gk}+2p_{gk}^2)}{\sum_k p_{gk}(1-p_{gk}) \left(\sqrt{N_g \sum_k p_{gk}(1-p_{gk})}\right)} < \frac{1}{\sqrt{N_g \sum_{k=1}^{K} p_{gk}(1-p_{gk})}}$$

If $h^\star = \min_k\{p_{gk}(1-p_{gk})\}$, then $N_g \sum_k p_{gk}(1-p_{gk}) \geq KN_g h^\star$ and hence

$$\frac{1}{\sqrt{\sum_{k=1}^{K} N_g p_{gk}(1-p_{gk})}} \leq \frac{1}{\sqrt{KN_g h^\star}} \to 0 \quad \text{when} \quad KN_g \to \infty$$

Since the Liapunov condition is satisfied,

$$\bar{X}_{g\cdot\cdot} \approx \mathrm{N}\left(\frac{\sum_k p_{gk}}{K}, \frac{\sum_k p_{gk}(1-p_{gk})}{K^2 N_g}\right)$$

for K or N_g sufficiently large.

Once the asymptotic distribution of $\bar{X}_{g\cdot\cdot}$ is normal, we could write PSS as a quadratic form of normal random variables.

Note that PSS can be written as

$$\mathrm{PSS} = \mathbf{H}'\mathbf{F}\mathbf{H}, \tag{3.2}$$

where $\mathbf{H} = (\bar{X}_{1\cdot\cdot},\ \bar{X}_{2\cdot\cdot},\ \ldots,\ \bar{X}_{G\cdot\cdot})'$ and $\mathbf{F} = K\mathbf{F}^\star$, with $\mathbf{F}^\star$ being a symmetric matrix $G \times G$ with its elements given as

$$f^\star(g,g) = N_g\left(1 - \frac{N_g}{N_T}\right) \quad \text{and} \quad f^\star(g,g') = -\frac{N_g N_{g'}}{N_T}, \quad g' \neq g \tag{3.3}$$

Therefore, asymptotically

$$\mathbf{H} \approx \mathrm{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) \tag{3.4}$$

where

$$\boldsymbol{\mu}_1 = \frac{1}{K}\left(\sum_k p_{1k},\ \sum_k p_{2k},\ \ldots,\ \sum_k p_{Gk}\right)' \quad \text{and} \quad \boldsymbol{\Sigma}_1 = \frac{1}{K^2}\boldsymbol{\Sigma}_1^\star, \tag{3.5}$$

$\boldsymbol{\Sigma}_1^\star$ is a diagonal matrix $G \times G$ with diagonal elements of the form

$$\sigma^\star(g,g) = \frac{\sum_k p_{gk}(1-p_{gk})}{N_g}.$$

Under the hypothesis of homogeneity among groups, $H_0 : p_{gk} = p_k$ for all g, from (3.4), asymptotically,

$$\mathbf{H} \approx \mathrm{N}\left(\boldsymbol{\mu}_{01}, \boldsymbol{\Sigma}_{01}\right) \tag{3.6}$$

where $\boldsymbol{\mu}_{01} = \frac{\sum_k p_k}{K}\mathbf{u_G}$, with $\mathbf{u_G}$ being a column vector of 1's of dimension G and

$$\boldsymbol{\Sigma}_{01} = \frac{\sum_k p_k(1-p_k)}{K^2}\boldsymbol{\Sigma}^{\star}_{01}, \tag{3.7}$$

with $\boldsymbol{\Sigma}^{\star}_{01}$ being a diagonal matrix $G \times G$ with elements $\sigma^{\star}_0(g,g) = N_g^{-1}$, $g = 1, \ldots, G$.

As PSS is a quadratic form of random variables with asymptotic normal distribution, one uses Cochran's Theorem (Sen and Singer, 1993) to find out the distribution of PSS under H_0.

As PSS $= \mathbf{H}'\mathbf{F}\mathbf{H} = K\mathbf{H}'\mathbf{F}^{\star}\mathbf{H}$, where $\mathbf{F}^{\star}$ is given by (3.3).

From (3.6) and (3.7) we have that, for $K \to \infty$,

$$\frac{K}{\sqrt{\sum_k p_k(1-p_k)}}(\mathbf{H} - \boldsymbol{\mu}_{01}) \xrightarrow{D} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\Sigma}^{\star}_{01}\right).$$

Note that, since $\boldsymbol{\Sigma}^{\star}_{01}$ is a diagonal matrix whose elements are all positive, $\boldsymbol{\Sigma}^{\star}_{01}$ is non singular and, therefore, $\mathbf{F}^{\star}$ is a generalized inverse of $\boldsymbol{\Sigma}^{\star}_{01}$ if and only if $\mathbf{F}^{\star}\boldsymbol{\Sigma}^{\star}_{01}$ is idempotent, i.e., $\mathbf{F}^{\star}\boldsymbol{\Sigma}^{\star}_{01}\mathbf{F}^{\star} = \mathbf{F}^{\star} \Leftrightarrow \mathbf{F}^{\star}\boldsymbol{\Sigma}^{\star}_{01}\mathbf{F}^{\star}\boldsymbol{\Sigma}^{\star}_{01} = \mathbf{F}^{\star}\boldsymbol{\Sigma}^{\star}_{01}$.

Lemma 3.1 $\mathbf{F}^{\star}\boldsymbol{\Sigma}^{\star}_{01}$ *is idempotent. (Proof in the Appendix)*

Lemma 3.2 $\mathrm{Rank}(\mathbf{F}^{\star}) = G - 1$. *(Proof in the Appendix)*

Note that

$$\boldsymbol{\mu}'_{01}\mathbf{F}^{\star}\boldsymbol{\mu}_{01} = \frac{\left(\sum_k p_k\right)^2}{K^2}\mathbf{u}'_{\mathbf{G}}\mathbf{F}^{\star}\mathbf{u_G} = 0, \tag{3.8}$$

since, by (3.3) and as $\mathbf{u_G}$ is a column vector of 1's of size G, $\mathbf{u}'_{\mathbf{G}}\mathbf{F}^{\star}$ is a row vector of size G whose i-th element, $i = 1, \ldots, G$ is

$$N_i\left(1 - \frac{N_i}{N_T}\right) - \frac{N_i}{N_T}(N_T - N_i) = 0 \tag{3.9}$$

Then, by (3.8), from Lemmas 3.1 and 3.2, and using Cochran's Theorem (Sen and Singer, 1993) for $K \to \infty$

$$\frac{K^2}{\sum_k p_k(1-p_k)}(\mathbf{H} - \boldsymbol{\mu}_{01})'\mathbf{F}^{\star}(\mathbf{H} - \boldsymbol{\mu}_{01}) \xrightarrow{D} \chi^2_{G-1}$$

Note that,

$$(\mathbf{H}-\boldsymbol{\mu}_{01})'\mathbf{F}^{\star}(\mathbf{H}-\boldsymbol{\mu}_{01}) = \mathbf{H}'\mathbf{F}^{\star}\mathbf{H} - 2\boldsymbol{\mu}_{01}'\mathbf{F}^{\star}\mathbf{H} + \boldsymbol{\mu}_{01}'\mathbf{F}^{\star}\boldsymbol{\mu}_{01},$$

therefore, by (3.8) and (3.9),

$$\frac{K^2}{\sum_k p_k(1-p_k)}(\mathbf{H}-\boldsymbol{\mu}_{01})'\mathbf{F}^{\star}(\mathbf{H}-\boldsymbol{\mu}_{01}) = \frac{K}{\sum_k p_k(1-p_k)}PSS.$$

Hence, under H_0 and for K sufficiently large

$$\frac{K}{\sum_k p_k(1-p_k)}PSS \approx \chi^2_{G-1} \tag{3.10}$$

Now we obtain the asymptotic distribution of the sum of squares due to the effect of individuals within population (ISS). Then, by (2.2), one has

$$\text{ISS} = \sum_{g=1}^{G}\sum_{i=1}^{N_g} K\left(\bar{X}_{gi\cdot} - \bar{X}_{g\cdot\cdot}\right)^2,$$

where $\bar{X}_{gi\cdot}$ represents the proportion, over K loci, of dominant phenotypes in individual i of group g and $\bar{X}_{g\cdot\cdot}$ represents the average number of dominant phenotypes in group g.

To obtain the asymptotic distribution of ISS, it is necessary to obtain the asymptotic distribution of $\bar{X}_{gi\cdot}$. Note that the number of dominant phenotypes over K loci in individual i of population g is

$$X_{gi\cdot} = \sum_{k=1}^{K} X_{gik}.$$

Therefore,

$$\text{E}(X_{gi\cdot}) = \sum_k p_{gk} \qquad \text{and} \qquad \text{Var}(X_{gi\cdot}) = \sum_k p_{gk}(1-p_{gk}).$$

In this case one has that $X_{gi1}, \ldots, X_{giK}$ are independent random variables such that

$$\text{E}(X_{gik}) = p_{gk}, \;\; \text{Var}(X_{gik}) = p_{gk}(1-p_{gk}), \;\; S_K = X_{gi\cdot} \;\text{ e }\; s_K = \sqrt{\text{Var}(X_{gi\cdot})}.$$

To verify whether the condition of Liapunov's Central Limit Theorem is satisfied one takes $\delta = 1$ and $0 < p_{gk} < 1$, then

$$\frac{1}{s_K^3}\sum_{k=1}^{K}\text{E}|X_{gik}-\mu_k|^3 \;\; = \;\; \frac{1}{s_K^3}\sum_{k=1}^{K}\text{E}|X_{gik}-p_{gk}|^3$$

$$= \frac{\sum_k p_{gk}(1-p_{gk})(1-2p_{gk}+2p_{gk}^2)}{\sum_k p_{gk}(1-p_{gk})\left(\sqrt{\sum_k p_{gk}(1-p_{gk})}\right)}$$

One has that

$$\frac{\sum_k p_{gk}(1-p_{gk})(1-2p_{gk}+2p_{gk}^2)}{\sum_k p_{gk}(1-p_{gk})\left(\sqrt{\sum_k p_{gk}(1-p_{gk})}\right)} < \frac{1}{\sqrt{\sum_{k=1}^K p_{gk}(1-p_{gk})}}.$$

If $h^\star = \min_k\{p_{gk}(1-p_{gk})\}$, then $\sum_k p_{gk}(1-p_{gk}) \geq Kh^\star$ and therefore

$$\frac{1}{\sqrt{\sum_{k=1}^K p_{gk}(1-p_{gk})}} \leq \frac{1}{\sqrt{Kh^\star}} \to 0 \text{ when } K \to \infty,$$

satisfying Liapunov's condition.

Therefore one has that when $K \to \infty$,

$$\bar{X}_{gi\cdot} \xrightarrow{D} \mathrm{N}\left(\frac{\sum_k p_{gk}}{K}, \frac{\sum_k p_{gk}(1-p_{gk})}{K^2}\right)$$

As PSS, ISS can also be written as a quadratic form of normal random variables, i.e.,

$$\mathrm{ISS} = \mathbf{H_2}'\mathbf{F_2}\mathbf{H_2} \tag{3.11}$$

where $\mathbf{H_2} = (\bar{X}_{11\cdot},\ \bar{X}_{12\cdot},\ \ldots,\ \bar{X}_{1N_1\cdot},\ \bar{X}_{21\cdot},\ \ldots,\ \bar{X}_{2N_2\cdot},\ \ldots,\ \bar{X}_{GN_G\cdot})'$ and $\mathbf{F_2} = K\mathbf{F_2^\star}$, $\mathbf{F_2^\star}$ is a symmetric matrix $N_T \times N_T$ of the form

$$\mathbf{F}_2^\star = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \ldots & \mathbf{0} \\ \cdot & \ldots & & \mathbf{0} \\ \cdot & & & \\ \cdot & & & \\ \mathbf{0} & \ldots & & \mathbf{A}_G \end{pmatrix} \tag{3.12}$$

where $\mathbf{A}_g = (a_g(i,j))\,,\ \ g = 1,\ \ldots,\ G$, is a matrix $N_g \times N_g$ such that

$$a_g(i,i) = 1 - \frac{1}{N_g} \quad \text{and} \quad a_g(i,j) = -\frac{1}{N_g},\ \ i,j = 1,\ \ldots,\ N_g \tag{3.13}$$

Therefore, asymptotically

$$\mathbf{H_2} \approx \mathrm{N}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right) \tag{3.14}$$

where

$$\boldsymbol{\mu}_2 = \frac{1}{K}\left(\sum_k p_{1k}\mathbf{u}'_{\mathbf{N_1}},\ \sum_k p_{2k}\mathbf{u}'_{\mathbf{N_2}},\ \ldots,\ \sum_k p_{Gk}\mathbf{u}'_{\mathbf{N_G}}\right)' \text{ and } \boldsymbol{\Sigma}_2 = \frac{1}{K^2}\boldsymbol{\Sigma}_2^{\star}, \tag{3.15}$$

$\mathbf{u_{N_g}}$ is a column vector of 1's, of size N_g, $\boldsymbol{\Sigma}_2^{\star}$ is a block diagonal $N_T \times N_T$ whose block diagonal elements are $\boldsymbol{\Sigma}_{2g}^{\star} = \sum_k p_{gk}(1-p_{gk})\mathbf{I}_{N_g}$, with $\mathbf{I}_{N_g}$ identity matrix $N_g \times N_g$, $g = 1, \ldots, G$.

From (3.14) one has, under H_0

$$\mathbf{H_2} \sim \mathrm{N}\left(\boldsymbol{\mu}_{02}, \boldsymbol{\Sigma}_{02}\right)$$

where $\boldsymbol{\mu}_{02} = \frac{\sum_k p_k}{K}\mathbf{u_{N_T}}$ and $\boldsymbol{\Sigma}_{02}$ is a diagonal matrix $N_T \times N_T$ of the form $\frac{\sum_k p_k(1-p_k)}{K^2}\mathbf{I}_{N_T}$.

One has then, ISS $= \mathbf{H_2}'\mathbf{F_2}\mathbf{H_2} = K\mathbf{H_2}'\mathbf{F_2^{\star}}\mathbf{H_2}$.

Under H_0,

$$\frac{K}{\sqrt{\sum_k p_k(1-p_k)}}(\mathbf{H_2} - \boldsymbol{\mu}_{02}) \xrightarrow{D} \mathrm{N}(\mathbf{0}, \mathbf{I}_{N_T}), \quad \text{when } K \to \infty,$$

by Cochran's Theorem (Sen and Singer, 1993) and given that $\mathbf{I}_{N_T}$ is a non singular matrix,

$$\frac{K^2}{\sum_k p_k(1-p_k)}(\mathbf{H_2} - \boldsymbol{\mu}_{02})'\mathbf{F_2^{\star}}(\mathbf{H_2} - \boldsymbol{\mu}_{02}) \xrightarrow{D} \chi^2_{posto(\mathbf{F_2^{\star}})}$$

if and only if $\mathbf{F_2^{\star}}\mathbf{I}_{N_T} = \mathbf{F_2^{\star}}$ is idempotent.

Lemma 3.3 $\mathbf{F_2^{\star}}$ *is idempotent. (Proof in the Appendix)*

Lemma 3.4 $\mathrm{Rank}(\mathbf{F_2^{\star}}) = N_T - G$. *(Proof in the Appendix)*

One has

$$\begin{aligned}(\mathbf{H_2} - \boldsymbol{\mu}_{02})'\mathbf{F_2^{\star}}(\mathbf{H_2} - \boldsymbol{\mu}_{02}) &= \mathbf{H_2}'\mathbf{F_2^{\star}}\mathbf{H_2} - 2\boldsymbol{\mu}'_{02}\mathbf{F_2^{\star}}\mathbf{H_2} + \boldsymbol{\mu}'_{02}\mathbf{F_2^{\star}}\boldsymbol{\mu}_{02} \\ &= \frac{ISS}{K}\end{aligned} \tag{3.16}$$

since $\boldsymbol{\mu}'_{02}\mathbf{F}^{\star}_{2}\boldsymbol{\mu}_{02} = \frac{\left(\sum_k p_k\right)^2}{K^2}\mathbf{u}'_{\mathbf{N_T}}\mathbf{F}^{\star}_{2}\mathbf{u}_{\mathbf{N_T}} = 0$, where $\mathbf{u}_{\mathbf{N_T}}$ is a N_T column vector of 1's, $\mathbf{u}'_{\mathbf{N_T}}\mathbf{F}^{\star}_{2}$ is a row vector, $1 \times N_T$ of the form $(\mathbf{a}_1\ \mathbf{a}_2\ \ldots\ \mathbf{a}_G)$, where $\mathbf{a}_g$, $g = 1, \ldots, G$ is a N_g row vector such that

$$\mathbf{a}_g(i) = 1 - \frac{1}{N_g} - \left(\frac{N_g - 1}{N_g}\right) = 0.$$

Therefore, under H_0, using Cochran's Theorem (Sen and Singer, 1993), Lemmas 3.3 and 3.4 and (3.16), for K sufficiently large,

$$\frac{K^2}{\sum_k p_k(1-p_k)}(\mathbf{H_2} - \boldsymbol{\mu}_{02})'\mathbf{F}^{\star}_{\mathbf{2}}(\mathbf{H_2} - \boldsymbol{\mu}_{02}) = \frac{K}{\sum_k p_k(1-p_k)}ISS \approx \chi^2_{N_T-G} \tag{3.17}$$

Since the interest is to compare groups of arrays of binary outcomes under the null hypothesis of homogeneity among groups and using (3.10) and (3.17), one has asymptotically $(K \to \infty)$ that,

$$\frac{K}{\sum_k p_k(1-p_k)}\text{PSS} \sim \chi^2_{G-1} \quad \text{and} \quad \frac{K}{\sum_k p_k(1-p_k)}\text{ISS} \sim \chi^2_{N_T-G}$$

The null hypothesis can be tested using the statistic $F = \dfrac{\text{MSP}}{\text{ISS}}$, where

$$\text{MSP} = \frac{\text{PSS}}{G-1} \quad \text{and} \quad \text{MSI} = \frac{\text{ISS}}{N_T - G}.$$

Lemma 3.5 PSS *and* ISS *are independent. (Proof in the Appendix)*

Therefore one has,

$$F = \frac{\text{MSP}}{\text{MSI}} \approx \frac{\left(\frac{\chi^2_{G-1}}{G-1}\right)}{\left(\frac{\chi^2_{N_T-G}}{N_T-G}\right)} \approx \text{F}_{G-1,N_T-G}, \tag{3.18}$$

In other words, asymptotically F follows the *Fisher-Snedecor* distribution with parameters $G-1$ and $N_T - G$.

When the vector of outcomes has a small dimension as is the case when RAPD markers are characterized by few *loci*, one can resort to resampling methods such as the *bootstrap*.

4 The Power of the Test

A brief study of the power of the test is now undertaken. It was seen in (3.2) and (3.11) that the sum of squares due to the population effects (PSS) and to individuals within population (ISS), respectively, can be written in matrix form:

$$\text{PSS} = \mathbf{H}'\mathbf{F}\mathbf{H} \quad \text{and} \quad \text{ISS} = \mathbf{H_2}'\mathbf{F_2}\mathbf{H_2}.$$

From (3.4) and (3.14) one has, asymptotically on the number of loci K,

$$\mathbf{H} \sim \text{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right) \quad \text{and} \quad \mathbf{H_2} \sim \text{N}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right),$$

where $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$ are defined in (3.5) and (3.15).

Since $\mathbf{F}$ is symmetric it can be decomposed as: $\mathbf{F} = \mathbf{Q_1^\star}\mathbf{D_1^\star}\mathbf{Q_1^{\star\prime}}$, where $\mathbf{Q_1^\star}$ is the orthogonal matrix of eigenvectors of $\mathbf{F}$ and $\mathbf{D_1^\star}$ is the diagonal matrix of eigenvalues of $\mathbf{F}$. Therefore,

$$\mathbf{F} = \mathbf{Q_1^\star}(\mathbf{D_1^\star})^{1/2}(\mathbf{D_1^\star})^{1/2}\mathbf{Q_1^{\star\prime}} = (\mathbf{F}^{1/2})'\mathbf{F}^{1/2} \tag{4.1}$$

Since $\mathbf{F}$ is semi-definite positive (proof in the Appendix), the elements of $\mathbf{F}^{1/2} \in I\!R$.

Therefore, PSS can be written as:

$$\text{PSS} = (\mathbf{F}^{1/2}\mathbf{H})'\mathbf{F}^{1/2}\mathbf{H} = \mathbf{X_1'}\mathbf{X_1},$$

one has then that

$$\mathbf{X_1} \sim \text{N}\left(\mathbf{F}^{1/2}\boldsymbol{\mu}_1;\ \mathbf{F}^{1/2}\boldsymbol{\Sigma}_1(\mathbf{F}^{1/2})'\right).$$

Theorem 4.1 *(proof in the Appendix)*

If $\mathbf{X}$ *is a* $n \times 1$ *random vector,* $\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{V})$*, where* $\mathbf{V}$ *is a nonsingular diagonal matrix and* $\mathbf{A}$ *is a* $n \times n$ *diagonal matrix of deterministic elements, then,*

$$\mathbf{X}'\mathbf{A}\mathbf{X} \sim \sum_{i=1}^{n} \lambda_i \left(\chi_1^2(\delta_i)\right)_i,$$

where λ_i *are the eigenvalues of matrix* $\mathbf{AV}$ *and* $\delta_i = \dfrac{\mu_i^2}{2\nu_i}$*, where* μ_i^2 *is the* i*-th element of vector* $\boldsymbol{\mu}$ *and* ν_i *are the eigenvalues of* $\mathbf{V}$*.* ■

Let $\mathbf{Q}_1$ be an orthogonal matrix such that $\mathbf{Q}_1\mathbf{F}^{1/2}\mathbf{\Sigma}_1(\mathbf{F}^{1/2})'\mathbf{Q}_1' = \mathbf{\Upsilon}_1$, where $\mathbf{\Upsilon}_1$ is a diagonal matrix.

If $\mathbf{Y_1} = \mathbf{Q}_1\mathbf{X_1} \Rightarrow \mathbf{X_1} = \mathbf{Q}_1'\mathbf{Y_1}$, then, $\mathbf{Y_1} \sim N\left(\mathbf{Q}_1\mathbf{F}^{1/2}\boldsymbol{\mu}_1;\ \mathbf{\Upsilon}_1\right)$ and

$$\mathbf{X_1}'\mathbf{X_1} = \mathbf{Y_1'Y_1} \sim \sum_{i=1}^{G} \upsilon_{1i}(\chi_1^2(\delta_{1i}))_i$$

where $\delta_{1i} = \dfrac{a_{1i}^2}{2\upsilon_{1i}}$, with a_{1i} as the i-th element of vector $\dfrac{1}{2}\mathbf{Q}_1\mathbf{F}^{-1/2}\boldsymbol{\mu}_1$ e υ_{1i}, $i = 1, \ldots, G$, are the eigenvalues of $\mathbf{\Upsilon}_1$, and therefore are the elements of the diagonal matrix $\mathbf{\Upsilon}_1$ (Theorem 4.1). Note that $\mathbf{\Upsilon}_1$ is semi-definite positive because it is a covariance matrix of normally distributed random variables and, therefore, $\upsilon_{1i} \geq 0$.

Analogously, from (4.1) one can obtain $\mathbf{F_2} = (\mathbf{F_2}^{1/2})'\mathbf{F_2}^{1/2}$, and since $\mathbf{F_2}$ is semi-definite positive (proof in the Appendix), its elements $\in I\!R$.

One then has

$$\text{ISS} = ((\mathbf{F_2})^{1/2}\mathbf{H_2})'(\mathbf{F_2})^{1/2}\mathbf{H_2} = \mathbf{X_2'X_2} = \mathbf{Y_2'Q_2Q_2'Y_2} = \mathbf{Y_2'Y_2},$$

where $\mathbf{Q}_2$ is a diagonal matrix such that $\mathbf{Q}_2(\mathbf{F_2})^{1/2}\mathbf{\Sigma}_2((\mathbf{F_2})^{1/2})'\mathbf{Q}_2' = \mathbf{\Upsilon}_2$ is a diagonal matrix.

Therefore, $\mathbf{Y_2} \sim \text{N}\left(\mathbf{Q}_2(\mathbf{F_2})^{1/2}\boldsymbol{\mu}_2;\ \mathbf{\Upsilon}_2\right)$ and by Theorem 4.1,

$$\text{ISS} = \mathbf{Y_2'Y_2} \sim \sum_{i=1}^{N_T} \upsilon_{2i}(\chi_1^2(\delta_{2i}))_i$$

where $\delta_{2i} = \dfrac{a_{2i}^2}{2\upsilon_{2i}}$, with a_{2i} as the i-th element of vector $\dfrac{1}{2}\mathbf{Q}_2(\mathbf{F_2})^{1/2}\boldsymbol{\mu}_2$ and υ_{2i}, $i = 1, \ldots, N_T$, are the eigenvalues of $\mathbf{\Upsilon}_2$, and therefore are the elements of diagonal matrix $\mathbf{\Upsilon}_2$. In this case $\upsilon_{2i} \geq 0$.

Then, from (3.18), for $u \in I\!R$,

$$\Pr(F \geq u) \quad = \quad \Pr\left(\frac{\text{PSS}}{\text{ISS}} \geq \frac{G-1}{N_T - G}u\right) \tag{4.2}$$

since PSS and ISS are linear combinations of random variables following χ_1^2 distribution whose non-centrality parameters are nonnegative and whose coefficients of the linear combination are also all nonnegative, PSS/ISS is a random variable that takes values only in $I\!R_+$.

Therefore, if $N_0 = \min_{0 \leq g \leq G} N_g$, for $N_0 \to \infty$, the probability in (4.2) tends to 1 indicating that the power of the test converges to 1, i.e.,

$$\Pr\left(\frac{\text{PSS}}{\text{ISS}} \geq \frac{G-1}{N_T - G} u\right) \longrightarrow \Pr\left(\frac{\text{PSS}}{\text{ISS}} \geq 0\right) = 1.$$

5 Application

The data set represents samples from a freshwater turtle, *Hydromedusa maximiliani* that inhabits shallow rivers and creeks in mountainous regions of the Atlantic Forest of eastern Brazil (Souza *et al.*, 2000). The study area in the state of São Paulo covers approximately $2,700$ ha containing three drainages from which a total of 44 individuals were sampled, based on the natural spatial hierarchy formed by rivers and streams.

The three drainages are referred to as I, II and III and samples sizes for each drainage were 25, 8 and 11, respectively. Drainage I, which contained the larger sample, was further subdivided into three sites representing different rivers. Sample sizes for each site were 4, 12 and 9, respectively.

Under the hypothesis of homogeneity among groups we have $H_0 : p_{1k} = p_{2k} = p_{3k} = p_k$, where p_k is the probability of having a band (the dominant phenotype) at position k. Since the number of loci is not large enough in order to apply an asymptotic test, it is necessary to generate the empirical distribution for the test statistic using resampling methods. A bootstrap procedure was used as follows:

Step 1: p_k is estimated from the data, that is, it is given by $\hat{p}_k = \dfrac{x_{\cdot\cdot k}}{N_T}$, which is the proportion of observed bands at position k for the combined sample of N_T individuals, and the observed value of the statistic F (F_{obs}) is calculated.
Step 2: $N_T = 44$ random vectors of dimension $K = 10$ are generated, where each of the K elements is taken from a Bernoulli distribution, with parameter $\hat{p}_k$.
Step 3: The value of statistic F is calculated for the simulated data.
Step 4: Steps 2 and 3 are repeated 10.000.

Following the procedure described above we generated the empirical distribution of F using *MATLAB*. The $p - value$ is given as the total number of F statistics whose values

are larger than the observed value for the statistics divided by $10,000$, that is,

$$p - value = \frac{\#F's \geq F_{obs}}{10,000}.$$

Figure 1 shows the asymptotic behavior of the distribution of the statistics F, given in (3.18), to compare drainages I, II, and III.

Using the simulated data, an estimate of $F_{obs} = 0.2702$ and a $p - value = 0.7625$ were obtained, indicating that there is no difference among the drainages at 5% level.

For the three sites within drainage I we obtained $F_{obs} = 0.5251$ and a $p - value = 0.5839$. This result also shows that there is no difference in the proportions of dominant phenotypes among freshwater turtles 1, 2 and 3. The behavior of the distributions for these cases can be seen in Figures 1 and 2.

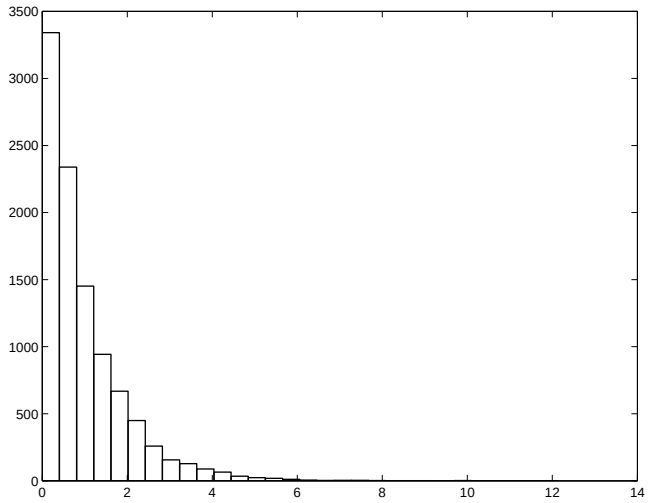

Figure 1: Empirical Distribution of F: RAPD of turtles from Drainage I, II and III.

Appendix A

Lemma 3.1 $\mathbf{F}^{\star}\mathbf{\Sigma}_{01}^{\star}$ is idempotent.

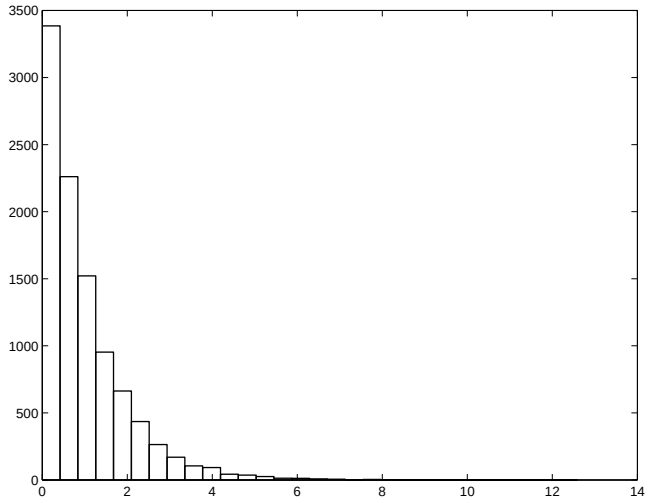

Figure 2: Empirical Distribution of F: RAPD of turtles (Sites 1, 2 and 3 from Drainage I).

Proof: By (3.3) and (3.7),

$$\mathbf{F}^{\star}\mathbf{\Sigma}^{\star}_{\mathbf{01}} = \mathbf{E_{01}},$$

where the elements of $\mathbf{E_{01}}$ are:

$$e_{01}(g,g) = 1 - \frac{N_g}{N_T} \quad e \quad e_{01}(g,g') = -\frac{N_g}{N_T}, \;\; g \neq g', \;\; g,g' = 1,...,G$$

The elements of $\mathbf{E^2_{01}}$ are

$$\begin{aligned} e^2_{01}(g,g) &= \left(1 - \frac{N_g}{N_T}\right)^2 + \frac{N_g}{N_T}(N_T - N_g) = 1 - \frac{N_g}{N_T} \\ e^2_{01}(g,g') &= -\frac{N_g}{N_T}\left(1 - \frac{N_g}{N_T} + 1 - \frac{N_{g'}}{N_T} - \frac{1}{N_T}\sum_{l \neq g,g'} N_l\right) = -\frac{N_g}{N_T} \quad \blacksquare \end{aligned}$$

Lemma 3.2 Rank$(\mathbf{F}^{\star}) = G - 1$.

Proof: Multiplying line r of matrix $\mathbf{F}^{\star}$ by the constant $\frac{1}{N_r}$, does not change the rank of $\mathbf{F}^{\star}$ (Rao, 1965), which is equivalent to pre-multiply $\mathbf{F}^{\star}$ by elementary matrices $G \times G$ known as Kronecker $\mathbf{\Delta_r}$, i.e., square diagonal matrices with nonzero elements in

the diagonal, being in this case:

$$\mathbf{\Delta_r} = (\delta_{gg'}): \quad \delta_{gg} = \begin{cases} \frac{1}{N_g} & \text{if } g = r \\ 1 & \text{if } g \neq r \end{cases}, \quad \delta_{gg'} = 0, \;\; g \neq g', \quad g,\; g' = 1, \ldots, G.$$

Premultiplying $\mathbf{F}^\star$ by G Kronecker matrices $\mathbf{\Delta_r}$, $r = 1, ..., G$ one obtains a matrix $\mathbf{E_1}$ whose elements are:

$$e_1(g,g) = 1 - \frac{N_g}{N_T}; \quad e_1(g,g') = -\frac{N_{g'}}{N_T}, \quad g \neq g', \;\; g, g' = 1, ..., G.$$

Note that $\mathbf{E_1} = \mathbf{E'_{01}}$ and therefore $\text{rank}(\mathbf{F}^\star) = \text{rank}(\mathbf{E'_{01}}) = \text{rank}(\mathbf{E_{01}})$. As the rank of an idempotent matrix is equal to its trace (Rao, 1965), $\text{rank}(\mathbf{F}^\star) = \sum_{g=1}^{G} \left(1 - \frac{N_g}{N_T}\right) = G - 1$

■

Lemma 3.3 $\mathbf{F_2^\star}$ is idempotent.

Proof: From (3.12) and (3.13) one has $\mathbf{F_2^\star} = \mathbf{A^\star}_1 + \mathbf{A^\star}_2 + \ldots + \mathbf{A^\star_G}$, given that

$$\mathbf{A^\star}_1 = \begin{pmatrix} \mathbf{A}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \ldots & \ddots & \vdots \\ \mathbf{0} & \ldots & \ldots & \mathbf{0} \end{pmatrix}, \quad \mathbf{A^\star}_2 = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \ldots & \mathbf{0} \\ \vdots & \ldots & \ddots & \vdots \\ \mathbf{0} & \ldots & \ldots & \mathbf{0} \end{pmatrix}, \ldots$$

$$\ldots, \quad \mathbf{A^\star}_G = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \vdots & \ldots & \ddots & \vdots \\ \mathbf{0} & \ldots & \ldots & \mathbf{A}_G \end{pmatrix}$$

$$\mathbf{A}_g^2 = \left(a_g^2(i,j)\right), \quad a_g^2(i,i) = 1 - \frac{1}{N_g} = a_g(i,i)$$

$$a_g^2(i,j) = -\frac{1}{N_g}\left[2\left(1 - \frac{1}{N_g}\right) - \frac{N_g - 2}{N_g}\right] = -\frac{1}{N_g} = a_g(i,j), \;\; g = 1, \ldots, G.$$

Therefore, one has
$(\mathbf{A^\star}_g)^2 = \mathbf{A^\star}_g \;\forall\; g$ and $\mathbf{A^\star}_g \mathbf{A^\star}_{g'} = \mathbf{0} \;\forall\; g \neq g' \Rightarrow (\mathbf{F_2^\star})^2 = \mathbf{F_2^\star}$ (Rao, 1965).

■

Lemma 3.4 $\text{Rank}(\mathbf{F_2^\star}) = N_T - G$.

Proof: Since $\mathbf{F_2^\star}$ is idempotent (Lema 3.3),

$$\text{Rank}(\mathbf{F_2^\star}) = \text{Trace}(\mathbf{F_2^\star}) = \sum_{g=1}^{G}\sum_{i=1}^{N_g}\left(1 - \frac{1}{N_g}\right) = N_T - G \qquad \blacksquare$$

Lemma 3.5 PSS and ISS are independent.

Proof: From (3.2) and (3.11), PSS and ISS can be written in matrix form. Note that $\mathbf{H'FH} = (\mathbf{MH_2})'\mathbf{FMH_2}$, where $\mathbf{M}$ is a matrix $G \times N_T$ whose elements are:

$$m_{ij} = \begin{cases} \dfrac{1}{N_i} & \text{if } \displaystyle\sum_{l=1}^{i-1} N_l < j \leq \sum_{l=1}^{i} N_l \\ 0 & \text{otherwise} \end{cases}$$

$i = 1 \ldots G$.

One has that $\mathbf{H_2}'\mathbf{F_2H_2}$ and $\mathbf{H_2}'\mathbf{M}'\mathbf{FMH_2}$ are independent if and only if $\mathbf{F_2\Sigma}_{02}\mathbf{M}'\mathbf{FM} = 0$ (Searle, 1971). From (3.3) and (3.12),

$$\mathbf{F_2\Sigma}_{02}\mathbf{M}'\mathbf{FM} = \frac{\sum_k p_k(1-p_k)}{N_T}\mathbf{\Phi}$$

where the elements of $\mathbf{\Phi} = (\phi_{ij})$, $i, j = 1 \ldots N_T$, are

$$\phi_{ij} = \begin{cases} \dfrac{1}{N_i}(N_T - N_i)\left[1 - \dfrac{1}{N_i} - \dfrac{N_i - 1}{N_i}\right] = 0 & \text{if } \displaystyle\sum_{l=1}^{i-1} N_l < i, j \leq \sum_{l=1}^{i} N_l \\ -\left(1 - \dfrac{1}{N_i}\right) + \dfrac{N_i - 1}{N_i} = 0 & \text{otherwise} \end{cases}$$

$\blacksquare$

Proof of Theorem 4.1: The moment generating function of a random variable Y with non central χ^2 distribution with n degrees of freedom and non centrality parameter δ, i.e., $Y \sim \chi^2_n(\delta)$ is given by:

$$M_Y(t) = (1-2t)^{-\frac{1}{2}n} e^{-\delta[1-(1-2t)^{-1}]} \qquad \text{(Searle, 1971, p49)}$$

According to Searle (1971, p57), if $\mathbf{X}$ is a random vector $n \times 1$,
$\mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{V}$ non singular, and $\mathbf{A}$ a $n \times n$ matrix of deterministic elements, then the

moment generating function of $\mathbf{X}'\mathbf{A}\mathbf{X}$ is given by:

$$M_{\mathbf{X}'\mathbf{A}\mathbf{X}}(t) = \prod_{i=1}^{n}(1-2t\lambda_i)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\mu}'\left[-\sum_{k=1}^{\infty}(2t)^k(\mathbf{A}\mathbf{V})^k\right]\mathbf{V}^{-1}\boldsymbol{\mu}\right\}.$$

With $\mathbf{A}$ and $\mathbf{V}$ diagonal matrices, $\mathbf{A}\mathbf{V}$ is a diagonal matrix whose diagonal elements are λ_i, $i = 1, \ldots, n$, Therefore $(\mathbf{A}\mathbf{V})^k$ is a diagonal matrix whose diagonal elements are λ_i^k. Then, $-\sum_{k=1}^{\infty}(2t)^k(\mathbf{A}\mathbf{V})^k$ is also a diagonal matrix with diagonal elements being

$$-\sum_{k=1}^{\infty}(2t\lambda_i)^k = 1-(1-2t\lambda_i)^{-1}, \quad \text{provided that } |t\lambda_i|<1, \quad i=1,\ldots,n.$$

Thus, as $\mathbf{V}$ is non singular, with diagonal elements being ν_i, $i = 1, \ldots, n$,

$$\boldsymbol{\mu}'\left[-\sum_{k=1}^{\infty}(2t)^k(\mathbf{A}\mathbf{V})^k\right]\mathbf{V}^{-1}\boldsymbol{\mu} = \sum_{i=1}^{n}\frac{\mu_i^2}{\nu_i}[1-(1-2t\lambda_i)^{-1}].$$

$$\begin{aligned} M_{\mathbf{X}'\mathbf{A}\mathbf{X}}(t) &= \prod_{i=1}^{n}(1-2t\lambda_i)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\mu_i^2}{\nu_i}[1-(1-2t\lambda_i)^{-1}]\right\} \\ &= \prod_{i=1}^{n} M_{Y_i}(t\lambda_i) = \prod_{i=1}^{n} M_{\lambda_i Y_i}(t), \end{aligned}$$

where $Y_i \sim \chi_1^2(\delta_i)$, with $\delta_i = \dfrac{\mu_i^2}{2\nu_i}$. ■

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

Comes, H.P., and Abbott, R.J.(2000). Random amplified polymorphic DNA (RAPD) and quantitative trait analyses across a major phylogeographical break in the Mediterranean ragwort *Senecio gallicus* Vill. (Asteraceae). Molecular Ecology 9, 61-69.

Haig, S.M., Rhymer, J.M., and Heckel, D.G. (1994). Population differentiation in randomly amplified polymorphic DNA of red-cockaded woodpeckers *Picoides borealis*. Molecular Ecology 3, 581-595.

James, B. (1996). *Introdução à Probabilidade: Um Curso em Nível Intermediário*, IMPA, Brazil.

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* (eds. P. K. Sen and C. R. Rao), Elsevier, Amsterdam, pp. 713 - 746.

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. (2002). Comparisons of Genomic Sequences using the Hamming Distance. *Research Report 72/02*. Instituto de Matemática, Estatística e Computação Científica. Universidade Estadual de Campinas.

Sen, P.K. and Singer, J.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*, Chapmann-Hall, UK.

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

Souza, F.L., Cunha, A.F., Oliveira, M.A., Pereira, G.A.G., Pinheiro, H.P. and Reis, S.F. (2002). Partitioning of molecular variation at local spatial scales in the vulnerable neotropical freshwater turtle, *Hydromedusa maximiliani* (Testudines, Chelidae): implications for the conservation of aquatic organisms in natural hierachical systems. Biological Conservation 104, 119-126.

Vucetich, L.M., Vucetich, J.A., Joshi, C.P., Waite, T.A., and Peterson, R.O. (2001). Genetic (RAPD) diversity in *Peromyscus maniculatus* populations in a naturally fragmented landscape. *Molecular Ecology* 10, 35-43.

Weir, B. (1990). *Genetic Analysis*. Sinauer.

Welsh, J., Peters, C. and Clelland, M. (1991). Polymorphisms generated by arbitrary primed PCR in the mouse: application to strain identification and genetic mapping. *Nucleic Acids Research* 20, 303-306.

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.