

## Analysis of variance for binary data in unbalanced designs

Roberta de Souza<sup>1</sup>, Hildete P. Pinheiro<sup>2</sup>,  
Cibele Q. da Silva<sup>3</sup> and Sérgio F. dos Reis<sup>2</sup>

<sup>1</sup> Instituto Internacional de Pesquisas Farmacêuticas

<sup>2</sup> Universidade Estadual de Campinas

<sup>3</sup> Universidade Federal de Minas Gerais

**Abstract:** 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 of great interest in molecular genetics and evolutionary biology. Some work have been done on analysis of variance for genetic data (Weir, 1996; Pinheiro et al., 2000; Pinheiro et al., 2001; Pinheiro et al., 2003 and others). Weir (1996) 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, asymptotic distribution, binary data, bootstrap, molecular data, RAPD, statistical genetics.

## 1 Introduction

The description and quantification of molecular variation and the characterization of genetic polymorphisms are essential for studies involving gene mapping, strain identification, parentage determination, population and conservation genetics, and molecular phylogenetics (Welsh and McClelland, 1990; Williams et al., 1990; Welsh et al., 1991; Frankham et al., 2002; Felsenstein, 2003). Of the several molecular markers available to assess levels of variation and genetic polymorphisms, randomly amplified polymorphic DNA (RAPD) has proven to be extremely valuable (e.g., Grosberg et al., 1996). Advantages of RAPD over other molecular markers include the suitability for work on anonymous genomes, the small amounts of DNA required for experimental assays, and efficiency and low cost (Grosberg et al., 1996).

RAPD markers consist of relatively short DNA fragments, ranging in size from 200 to 2000 nucleotide base pairs (bp) long, which can be isolated in the

laboratory by the polymerase chain reaction using primers of 10 bp in length whose sequence of nucleotides is arbitrary (Welsh et al., 1991; Grosberg et al., 1996). Pairs of such arbitrary primers anneal to priming sites in the target sequence in opposite orientations in the genome of interest. Given that the primers are short in length there is a high probability that the genome will contain several priming sites at varying distances from one another (Welsh et al., 1991; Grosberg et al., 1996). Following the accepted convention (Grosberg et al., 1996), the pair of inverted priming sites in the genome plus the intervening sequence of nucleotides is a RAPD locus, and the amplified product from a particular locus is a RAPD marker. The amplification profile of products are electrophoretically resolved by size on agarose gels and appear as bands when visualized by specific stains. The presence or absence of RAPD bands in any given individual is determined by the genotype of the individual at a given locus. RAPD alleles are expressed as dominant markers, that is, if either allele at a locus carries an amplifiable fragment the resulting phenotype is the presence of a band and the individual can be either a dominant homozygous or a heterozygous (Welsh et al., 1991; Grosberg et al., 1996). On the other hand, the absence of a band indicates that the individual is a recessive homozygous (Welsh et al., 1991; Grosberg et al., 1996). An elementary description of the molecular basis of polymorphism and experimental detection of RAPD markers can be found in Griffiths et al. (2000).

RAPD molecular markers have been increasingly employed with success to quantify and describe patterns of variation within and among populations of animals and plants (Persson et al., 1998; Vucetich et al., 2001; Comes and Abbott, 2002; Gunter et al., 2003; Verovnik et al., 2003). These markers have also proven instrumental to infer patterns of population structure with important implications for evolutionary and conservation biology (Haig et al., 1994; Frankham et al., 2002; Souza et al., 2002).

In all these studies a primary interest is to quantify genetic variation. There are many different ways to measure genetic variation; among them one can think of the proportion of heterozygotes individuals in a population, the heterozygosity, since heterozygotes individuals carry different alleles, which are responsible for the existence of variation. The continuous presence of different homozygous also can result in variation; for those situations the gene diversity is an appropriate measure (Simpson, 1949; Nei, 1972; Weir, 1996; Pinheiro and Seillier-Moiseiwitsch, 2001). 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; Pinheiro et al., 2005).

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 phenotypes).

Weir (1996) 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

For any given pair of RAPD primers, the same loci (position in the genome) are amplified in all sampled individuals. In other words, the same loci are repeatedly scored and therefore the loci can be considered as a fixed effect (Weir, 1996).

The main biological interest here is to test for heterogeneity within and between population across loci, that is, to evaluate the amount of variability among individuals sampled from different groups (e.g., geographical regions, strains or species). So, it does not make 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, 1996) with the RAPD markers, let us define,

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

with  $i = 1, \dots, N_g$ ;  $g = 1, \dots, G$ ;  $k = 1, \dots, K$ .

In Table 2 sums of squares of indicator variables are constructed to reflect the sampling structure, and the corresponding expected mean squares are written down as though the variables could be represented by a linear model of the form

$$X_{gik} = \alpha_g + \beta_{gi} + \xi_{gik} ,$$

where  $\alpha_g$  is the population or group effect;  $\beta_{gi}$  is the individual within population effect and  $\xi_{gik}$  is the residual effect. The group and individual within group effects are considered random with  $E(\alpha_g) = 0$ ,  $E(\beta_{gi}) = 0$  and  $E(\xi_{gik}) = H_k$ . The variances of the group, individual within group and residual are, respectively,  $\sigma_g^2$ ,  $\sigma_{i/g}^2$  and  $\sigma^2$ . Note that in Table 2

$$L = \frac{1}{K} \sum_k H_k^2 - \frac{1}{K(K-1)} \sum_k \sum_{k' \neq k} H_k H_{k'}.$$

Since the interest here is to test for homogeneity among groups, a natural test statistic is to take the Mean Square Population (*MSP*) divided by the Mean Square Individual (*MSI*), which will be developed in Section 3.

The Population Sum of Squares (*PSS*), Individual Sum of Squares (*ISS*), Residual

**Table 1** Analysis of variance for RAPD data in unbalanced designs

Source of Variation	d.f.	Sum of Squares	E(MS)
Population	$G - 1$	$PSS$	$\sigma^2 + K\sigma_{i/g}^2 + \frac{N_T^2 - G \sum_g N_g^2}{N_T K (G-1)} (\sum_k H_k)^2 + \frac{K}{(G-1)N_T} \sum_g \sum_{g' \neq g} N_g N_{g'} \sigma_g^2$
Individuals within populations	$N_T - G$	$ISS$	$\sigma^2 + K\sigma_{i/g}^2$
Residual	$(K - 1)N_T$	$RSS$	$\sigma^2 + L$
Total	$KN_T - 1$	$TSS$	

Sum of Squares (RSS) and the Total Sum of Squares (TSS) are as follows,

$$PSS = \sum_{g=1}^G KN_g (\bar{X}_{g..} - \bar{X}_{...})^2, \tag{2.1}$$

$$ISS = \sum_{g=1}^G \sum_{i=1}^{N_g} K (\bar{X}_{g_i.} - \bar{X}_{g..})^2, \tag{2.2}$$

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

where,  $N_T = \sum_{g=1}^G N_g$  is the total number of individuals;  $N_g$  is the number of individuals in the  $g$ -th group;  $G$  is the number of groups (populations) and  $K$  is the number of loci for each sequence (individual).

$$\bar{X}_{g_i.} = \frac{X_{g_i.}}{K}, \quad \bar{X}_{g..} = \frac{X_{g..}}{KN_g} \text{ and } \bar{X}_{...} = \frac{\sum_g \sum_i \sum_k X_{gik}}{N_T}. \tag{2.4}$$

### 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. The individuals can be considered independent since we have a random sample of individuals. The groups are assumed to be independent. As mentioned in Section 2, the positions of the loci in the genome are not known so one cannot know or assume any kind of dependence structure among loci. Therefore, the loci are assumed to be independent (Evetts and Weir, 1998). As  $X_{gik}$  is a binary variable, it follows a Bernoulli distribution, i.e.,

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

where  $p_{gk}$  is the probability that an individual of population  $g$  be dominant at locus  $k$ ,  $i = 1, \dots, N_g$ ;  $g = 1, \dots, G$ ;  $k = 1, \dots, K$ . Therefore,  $E(X_{gik}) = p_{gk}$  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  $g$ ,  $\bar{X}_{g\cdot}$  represents the proportion of dominant phenotype in group  $g$  and  $\bar{X}_{\dots}$  is the general proportion (or mean) of dominant phenotype in the whole sample. The expressions for all those terms are given in (2.4).

As our interest is to test the hypothesis of homogeneity among groups, i.e.,

$$H_0 : p_{gk} = p_k, \quad \text{for all } g = 1, \dots, G, \quad (3.2)$$

observing Table 2 we propose as the test statistic  $F = MSP/MSI$ , where

$$MSP = \frac{PSS}{G - 1} \quad \text{and} \quad MSI = \frac{ISS}{N_T - G}, \quad (3.3)$$

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 ( $\bar{X}_{g\cdot}$ ) and the total number of dominant phenotypes in the  $g$ -th group can be written as  $X_{g\cdot} = \sum_{i=1}^{N_g} \sum_{k=1}^K X_{gik}$ . By model (3.1),

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

As  $X_{gik}$  are independent, but not identically distributed random variables, one will use the Central Limit Theorem of Liapunov for independent random variables (Lehmann, 1999).

In this case, for a given population  $g$ ,  $X_{g11}, \dots, X_{g1K}, \dots, X_{gN_g1}, \dots, X_{gN_gK}$ ,  $g = 1, \dots, G$ , are independent random variables, such that  $E(X_{gik}) =$

$p_{gk}$ ,  $\text{Var}(X_{gik}) = p_{gk}(1 - p_{gk})$ , and let  $s_n = \sqrt{\text{Var}(\bar{X}_{g..})}$ , where  $n = KN_g$ . Therefore, verifying the Liapunov condition, we have:

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

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

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,

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

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

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

Since the Liapunov condition is satisfied,

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

for  $K$  or  $N_g$  sufficiently large.

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

Note that PSS can be written as

$$\text{PSS} = \mathbf{H}' \mathbf{F} \mathbf{H} , \quad (3.4)$$

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

$$f^*(g, g) = N_g \left( 1 - \frac{N_g}{N_T} \right) \quad \text{and} \quad f^*(g, g') = -\frac{N_g N_{g'}}{N_T}, \quad g' \neq g \quad (3.5)$$

Therefore, asymptotically

$$\mathbf{H} \approx N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) , \quad (3.6)$$

where

$$\boldsymbol{\mu}_1 = \frac{1}{K} \left( \sum_k p_{1k}, \sum_k p_{2k}, \dots, \sum_k p_{Gk} \right)' \quad \text{and} \quad \boldsymbol{\Sigma}_1 = \frac{1}{K^2} \boldsymbol{\Sigma}_1^* , \quad (3.7)$$

$\Sigma_1^*$  is a diagonal matrix  $G \times G$  with diagonal elements of the form

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

Under the hypothesis of homogeneity among groups,  $H_0 : p_{gk} = p_k \forall g$ , from (3.6), asymptotically,

$$\mathbf{H} \approx N(\boldsymbol{\mu}_{01}, \boldsymbol{\Sigma}_{01}) , \quad (3.8)$$

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

$$\boldsymbol{\Sigma}_{01} = \frac{\sum_k p_k(1 - p_k)}{K^2} \boldsymbol{\Sigma}_{01}^* , \quad (3.9)$$

with  $\boldsymbol{\Sigma}_{01}^*$  being a diagonal matrix  $G \times G$  with elements  $\sigma_0^*(g, g) = N_g^{-1}$ ,  $g = 1, \dots, G$ .

As  $\text{PSS} = \mathbf{H}'\mathbf{F}\mathbf{H} = K\mathbf{H}'\mathbf{F}^*\mathbf{H}$ , with elements of  $\mathbf{F}^*$  given by (3.5), PSS is a quadratic form of random variables with asymptotic normal distribution. Then, using Cochran's Theorem (Sen and Singer, 1993) we can find out the distribution of PSS under  $H_0$ .

From (3.8) and (3.9) we have that, for  $K \rightarrow \infty$ ,

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

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

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

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

Note that

$$\boldsymbol{\mu}'_{01} \mathbf{F}^* \boldsymbol{\mu}_{01} = \frac{(\sum_k p_k)^2}{K^2} \mathbf{1}'_G \mathbf{F}^* \mathbf{1}_G = 0 , \quad (3.10)$$

since, by (3.5) and as  $\mathbf{1}_G$  is a column vector of 1's of size  $G$ ,  $\mathbf{1}'_G \mathbf{F}^*$  is a row vector of dimension  $G$  whose  $i$ -th element,  $i = 1, \dots, G$  is

$$N_i \left( 1 - \frac{N_i}{N_T} \right) - \frac{N_i}{N_T} (N_T - N_i) = 0 . \quad (3.11)$$

Then, by (3.10), from Lemmas 3.1 and 3.2, and using Cochran's Theorem (Sen and Singer, 1993) for  $K \rightarrow \infty$

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

Note that,

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

therefore, by (3.10) and (3.11),

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

Hence, under  $H_0$  and for  $K$  sufficiently large

$$\frac{K}{\sum_k p_k (1 - p_k)} PSS \approx \chi^2_{G-1} . \tag{3.12}$$

It is important to point out that a large number of RAPD loci ( $K$ ) can be obtained experimentally because each polymerase chain reaction involves only a single primer, and therefore a large number of primers can be assayed in a short amount of time (Grosberg et al., 1996).

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

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

where  $\bar{X}_{gi.}$  represents the proportion, over  $K$  loci, of dominant phenotypes in individual  $i$  of group  $g$  and  $\bar{X}_{g..}$  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  $\bar{X}_{gi.}$ . Note that the number of dominant phenotypes over  $K$  loci in individual  $i$  of population  $g$  is  $X_{gi.} = \sum_{k=1}^K X_{gik}$ . Therefore,

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

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

$$E(X_{gik}) = p_{gk}, \quad \text{Var}(X_{gik}) = p_{gk}(1 - p_{gk}), \quad \text{and} \quad s_K = \sqrt{\text{Var}(X_{gi.})} .$$

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

$$\frac{1}{s_K^3} \sum_{k=1}^K E|X_{gik} - \mu_k|^3 = \frac{\sum_k p_{gk}(1 - p_{gk})(1 - 2p_{gk} + 2p_{gk}^2)}{\sum_k p_{gk}(1 - p_{gk}) (\sqrt{\sum_k p_{gk}(1 - p_{gk})})} .$$

Analogously to the Liapunov's condition verified in (3.4), we have,

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

satisfying Liapunov's condition, where  $h^* = \min_k \{p_{gk}(1-p_{gk})\}$ .

Therefore, when  $K \rightarrow \infty$ ,

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

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

$$\text{ISS} = \mathbf{H}_2' \mathbf{F}_2 \mathbf{H}_2, \quad (3.13)$$

where  $\mathbf{H}_2 = (\bar{X}_{11}, \bar{X}_{12}, \dots, \bar{X}_{1N_1}, \bar{X}_{21}, \dots, \bar{X}_{2N_2}, \dots, \bar{X}_{GN_G})'$  and  $\mathbf{F}_2 = K\mathbf{F}_2^*$ ,  $\mathbf{F}_2^*$  is a block diagonal matrix,  $N_T \times N_T$ , with diagonal blocks

$$\mathbf{A}_g = \mathbf{I}_{N_g} - N_g^{-1} \mathbf{1}_{N_g} \mathbf{1}_{N_g}' . \quad (3.14)$$

Therefore, asymptotically

$$\mathbf{H}_2 \approx N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \quad (3.15)$$

where

$$\boldsymbol{\mu}_2 = \frac{1}{K} \left( \sum_k p_{1k} \mathbf{1}'_{N_1}, \sum_k p_{2k} \mathbf{1}'_{N_2}, \dots, \sum_k p_{Gk} \mathbf{1}'_{N_G} \right)' \text{ and } \boldsymbol{\Sigma}_2 = \frac{1}{K^2} \boldsymbol{\Sigma}_2^*, \quad (3.16)$$

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

From (3.15) one has that, under  $H_0$ ,

$$\mathbf{H}_2 \sim N(\boldsymbol{\mu}_{02}, \boldsymbol{\Sigma}_{02}),$$

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

Then,

$$\frac{K}{\sqrt{\sum_k p_k(1-p_k)}} (\mathbf{H}_2 - \boldsymbol{\mu}_{02}) \xrightarrow{D} N(\mathbf{0}, \mathbf{I}_{N_T}), \text{ when } K \rightarrow \infty,$$

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

$$\frac{K^2}{\sum_k p_k(1-p_k)} (\mathbf{H}_2 - \boldsymbol{\mu}_{02})' \mathbf{F}_2^* (\mathbf{H}_2 - \boldsymbol{\mu}_{02}) \xrightarrow{D} \chi_{\text{rank}(\mathbf{F}_2^*)}^2,$$

if and only if  $\mathbf{F}_2^* \mathbf{I}_{N_T} = \mathbf{F}_2^*$  is idempotent.

**Lemma 3.3.**  $\mathbf{F}_2^*$  is idempotent. (Proof in the Appendix)

**Lemma 3.4.**  $\text{Rank}(\mathbf{F}_2^*) = N_T - G$ . (Proof in the Appendix)

One has

$$\begin{aligned} (\mathbf{H}_2 - \boldsymbol{\mu}_{02})' \mathbf{F}_2^* (\mathbf{H}_2 - \boldsymbol{\mu}_{02}) &= \mathbf{H}_2' \mathbf{F}_2^* \mathbf{H}_2 - 2\boldsymbol{\mu}'_{02} \mathbf{F}_2^* \mathbf{H}_2 + \boldsymbol{\mu}'_{02} \mathbf{F}_2^* \boldsymbol{\mu}_{02} \\ &= \frac{ISS}{K}, \end{aligned} \tag{3.17}$$

since  $\boldsymbol{\mu}'_{02} \mathbf{F}_2^* \boldsymbol{\mu}_{02} = \frac{(\sum_k p_k)^2}{K^2} \mathbf{u}'_{N_T} \mathbf{F}_2^* \mathbf{1}_{N_T} = 0$ , where  $\mathbf{1}_{N_T}$  is a  $N_T$  column vector of 1's,  $\mathbf{1}'_{N_T} \mathbf{F}_2^*$  is a row vector,  $1 \times N_T$  of the form  $(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_G)$ , where  $\mathbf{a}_g, g = 1, \dots, G$  is a row vector of dimension  $N_g$ , 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 (3.17), for  $K$  sufficiently large,

$$\frac{K^2}{\sum_k p_k(1 - p_k)} (\mathbf{H}_2 - \boldsymbol{\mu}_{02})' \mathbf{F}_2^* (\mathbf{H}_2 - \boldsymbol{\mu}_{02}) = \frac{K}{\sum_k p_k(1 - p_k)} ISS \approx \chi^2_{N_T - G}. \tag{3.18}$$

Since the interest is to compare groups of arrays of binary outcomes, under the null hypothesis of homogeneity among groups and using (3.12) and (3.18), one has asymptotically ( $K \rightarrow \infty$ ) 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, given in (3.2), can be tested using the statistic  $F = \frac{\text{MSP}}{\text{MSI}}$ , where MSP and MSI are given in (3.3).

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

Therefore one has,

$$F \approx \frac{\left( \frac{\chi^2_{G-1}}{G-1} \right)}{\left( \frac{\chi^2_{N_T - G}}{N_T - G} \right)} \approx F_{G-1, N_T - G}, \tag{3.19}$$

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*. More details about the use of resampling methods applied to RAPD markers are given in a case study developed in Section 5.

## 4 The power of the test

A brief study of the power of the test is now undertaken. It was seen in (3.4) and (3.13) 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 (3.6) and (3.15) one has, asymptotically on the number of loci  $K$ ,

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

where  $\boldsymbol{\mu}_1$ ,  $\boldsymbol{\Sigma}_1$ ,  $\boldsymbol{\mu}_2$  and  $\boldsymbol{\Sigma}_2$  are defined in (3.7) and (3.16).

Since  $\mathbf{F}$  is positive semi-definite (see proof in the Appendix) it can be decomposed as:  $\mathbf{F} = \mathbf{Q}_1^*\mathbf{D}_1^*\mathbf{Q}_1^{*\prime}$ , where  $\mathbf{Q}_1^*$  is the orthogonal matrix of eigenvectors of  $\mathbf{F}$  and  $\mathbf{D}_1^*$  is the diagonal matrix of eigenvalues of  $\mathbf{F}$ . Then,

$$\mathbf{F} = \mathbf{Q}_1^*(\mathbf{D}_1^*)^{1/2}(\mathbf{D}_1^*)^{1/2}\mathbf{Q}_1^{*\prime} = (\mathbf{F}^{1/2})'\mathbf{F}^{1/2} . \quad (4.1)$$

Therefore, PSS can be written as:

$$\text{PSS} = (\mathbf{F}^{1/2}\mathbf{H})'\mathbf{F}^{1/2}\mathbf{H} = \mathbf{X}_1'\mathbf{X}_1 .$$

Then,

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

**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,

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

where  $\lambda_i$  are the eigenvalues of matrix  $\mathbf{A}\mathbf{V}$  and  $\delta_i = \frac{\mu_i^2}{2\nu_i}$ , where  $\mu_i$  is the  $i$ -th element of vector  $\boldsymbol{\mu}$  and  $\nu_i$  are the eigenvalues of  $\mathbf{V}$ . ■

Let  $\mathbf{Q}_1$  be an orthogonal matrix such that  $\mathbf{Q}_1\mathbf{F}^{1/2}\boldsymbol{\Sigma}_1(\mathbf{F}^{1/2})'\mathbf{Q}_1' = \boldsymbol{\Upsilon}_1$ , where  $\boldsymbol{\Upsilon}_1$  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,  $\mathbf{Y}_1 \sim N\left(\mathbf{Q}_1\mathbf{F}^{1/2}\boldsymbol{\mu}_1; \boldsymbol{\Upsilon}_1\right)$  and

$$\mathbf{X}_1'\mathbf{X}_1 = \mathbf{Y}_1'\mathbf{Y}_1 \sim \sum_{i=1}^G v_{1i} \chi_1^2(\delta_{1i}) ,$$

where  $\delta_{1i} = \frac{a_{1i}^2}{2v_{1i}}$ , with  $a_{1i}$  as the  $i$ -th element of vector  $\frac{1}{2}\mathbf{Q}_1\mathbf{F}^{1/2}\boldsymbol{\mu}_1$  and  $v_{1i}$ ,  $i = 1, \dots, G$ , are the eigenvalues of  $\boldsymbol{\Upsilon}_1$ , and therefore are the elements of the

diagonal matrix  $\mathbf{\Upsilon}_1$  (Theorem 4.1). Note that  $\mathbf{\Upsilon}_1$  is positive semi-definite because it is a covariance matrix of normally distributed random variables and, therefore,  $v_{1i} \geq 0$ .

Analogously, from (4.1) and since  $\mathbf{F}_2$  is positive semi-definite (proof in the Appendix), one can obtain  $\mathbf{F}_2 = (\mathbf{F}_2^{1/2})' \mathbf{F}_2^{1/2}$ . Then,

$$\text{ISS} = ((\mathbf{F}_2)^{1/2} \mathbf{H}_2)' (\mathbf{F}_2)^{1/2} \mathbf{H}_2 = \mathbf{X}_2' \mathbf{X}_2 = \mathbf{Y}_2' \mathbf{Q}_2 \mathbf{Q}_2' \mathbf{Y}_2 = \mathbf{Y}_2' \mathbf{Y}_2 ,$$

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

Therefore,  $\mathbf{Y}_2 \sim N(\mathbf{Q}_2 (\mathbf{F}_2)^{1/2} \boldsymbol{\mu}_2; \mathbf{\Upsilon}_2)$  and by Theorem 4.1,

$$\text{ISS} = \mathbf{Y}_2' \mathbf{Y}_2 \sim \sum_{i=1}^{N_T} v_{2i} \chi_1^2(\delta_{2i}) ,$$

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

Then, from (3.19), for  $u \in \mathbb{R}$ ,

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

since PSS and ISS are linear combinations of random variables following  $\chi_1^2$  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  $\mathbb{R}_+$ .

Therefore, if  $N_0 = \min_{0 \leq g \leq G} N_g$ , for  $N_0 \rightarrow \infty$ , 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 presented in this section are derived from a study of population genetic structure using RAPD molecular markers in the freshwater turtle *Hydromedusa maximiliani*, conducted in the state of São Paulo in southeastern Brazil (Souza et al., 2002). This freshwater turtle inhabits topologically complex habitats characterized by sequences of ridges and valleys, each drained by river and stream systems. For this study, an area of approximately 2700ha containing three drainages (hereafter drainages I, II and III) was sampled based on the natural spatial hierarchy formed by rivers and streams. Within each drainage, specimens of *H.*

*maximiliani* were randomly hand-caught in the natural habitat of shallow rivers and streams. A total of 44 individuals were randomly sampled with sample sizes of 25, 8, and 11 for drainages I, II and III, respectively. Drainage I, the larger drainage sampled and which yielded the larger sample size, was further subdivided according to the spatial hierarchy of the main rivers and their tributaries, resulting in three sample sites. Sample sizes for each site were 4, 12, and 9, respectively.

From this sample 10 RAPD polymorphic loci were isolated and molecular patterns of variation across drainages were evaluated using the molecular analysis of variance (AMOVA; Excoffier et al., 1992). The AMOVA procedure is based on a simple Euclidean metric which is used as a measure of distance between individual RAPD molecular phenotypes. In our approach the measure of variation is based on theoretical framework of the classical analysis of variance and the same data derived by Souza et al. (2002) is used here to illustrate our methodology.

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 a band present (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 under the null hypothesis of homogeneity among groups, that is, it is given by  $\hat{p}_k = \frac{x_{..k}}{N_T}$ , which is the proportion of observed bands present at position  $k$  for the combined sample of  $N_T = 44$  individuals, and the observed value of the statistic  $F (F_{obs})$  is computed.

**Step 2:**  $N_T = 44$  random vectors of dimension  $K = 10$  loci are generated, where each of the  $K$  elements is drawn 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 times.

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 or equal the observed value of the test statistic divided by 10,000, that is,

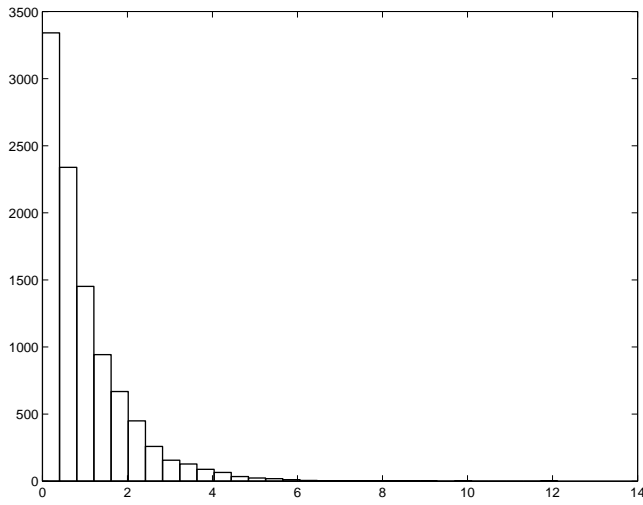
$$p - value = \frac{\#F's \geq F_{obs}}{10,000}.$$

Figure 1 shows the empirical distribution of the  $F$  statistic, given in (3.19), to compare drainages I, II, and III.

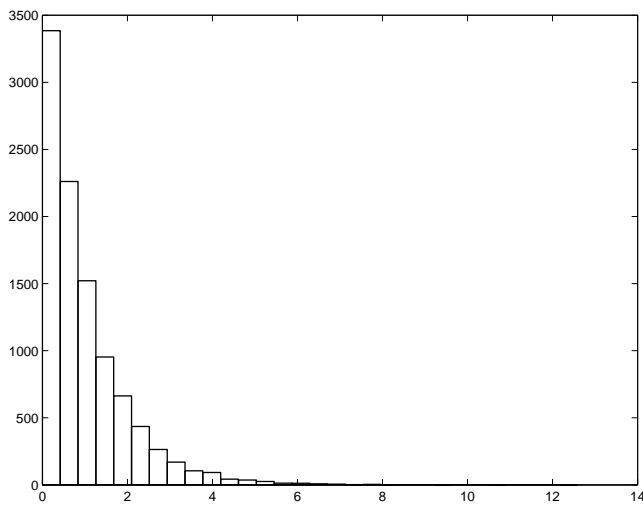
For an estimate of  $F_{obs} = 0.2702$  and using the bootstrapped (resampled) data, a  $p$  - value = 0.7625 was obtained, indicating that there is no difference among the drainages. If we use the asymptotic distribution given in (3.19), the  $p$  - value is the  $P(F_{2,41} > 0.2702) = 0.7646$ .

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 empirical distribution for this case can be seen in Figure 2.

Using the asymptotic  $p$  - value in this case, we have a  $p$  - value =  $P(F_{2,22} > 0.5251) = 0.5987$ . In both cases the asymptotic and the bootstrap  $p$  - values are



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



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

very close, which suggests that even for small data sets the asymptotic results work appropriately.

## Appendix

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

**Proof.** By (3.5) and (3.9),

$$\mathbf{F}^* \boldsymbol{\Sigma}_{01}^* = \mathbf{E}_{01},$$

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

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

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

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

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

**Proof.** Multiplying line  $r$  of matrix  $\mathbf{F}^*$  by the constant  $\frac{1}{N_r}$ , does not change the rank of  $\mathbf{F}^*$ , which is equivalent to pre-multiply  $\mathbf{F}^*$  by elementary matrices  $G \times G$  known as Kronecker (Rao, 1965, p29)  $\boldsymbol{\Delta}_r$ , i.e., square diagonal matrices with nonzero elements in the diagonal, being in this case:

$$\boldsymbol{\Delta}_r = (\delta_{gg'}) : \quad \delta_{gg} = \begin{cases} \frac{1}{N_g} & \text{if } g = r \\ 1 & \text{if } g \neq r \end{cases}, \quad \delta_{gg'} = 0, \quad g \neq g', \quad g, g' = 1, \dots, G.$$

Premultiplying  $\mathbf{F}^*$  by  $G$  Kronecker matrices  $\boldsymbol{\Delta}_r$ ,  $r = 1, \dots, G$  one obtains a matrix  $\mathbf{E}_1$  whose elements are:

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

Note that  $\mathbf{E}_1 = \mathbf{E}'_{01}$  and therefore  $\text{rank}(\mathbf{F}^*) = \text{rank}(\mathbf{E}'_{01}) = \text{rank}(\mathbf{E}_{01})$ . As the rank of an idempotent matrix is equal to its trace (Rao, 1965, p28),  $\text{rank}(\mathbf{F}^*) = \sum_{g=1}^G \left(1 - \frac{N_g}{N_T}\right) = G - 1$ .  $\blacksquare$

**Lemma 3.3**  $\mathbf{F}_2^*$  is idempotent.

**Proof.** Note that  $\mathbf{F}_2^*$  is a block diagonal matrix,  $N_T \times N_T$ , with diagonal blocks  $\mathbf{A}_g$ .

$$\begin{aligned} \mathbf{A}_g^2 &= (\mathbf{I}_{N_g} - N_g^{-1} \mathbf{1}_{N_g} \mathbf{1}'_{N_g})^2 = \mathbf{I}_{N_g}^2 - 2N_g^{-1} \mathbf{1}_{N_g} \mathbf{1}'_{N_g} + N_g N_g^{-2} \mathbf{1}_{N_g} \mathbf{1}'_{N_g} \\ &= \mathbf{I}_{N_g} - N_g^{-1} \mathbf{1}_{N_g} \mathbf{1}'_{N_g} = \mathbf{A}_g . \end{aligned}$$

**Lemma 3.4**  $\text{Rank}(\mathbf{F}_2^*) = N_T - G$ . ■

**Proof.** Since  $\mathbf{F}_2^*$  is idempotent (Lemma 3.3),

$$\text{Rank}(\mathbf{F}_2^*) = \text{Trace}(\mathbf{F}_2^*) = \sum_{g=1}^G \sum_{i=1}^{N_g} \left(1 - \frac{1}{N_g}\right) = N_T - G .$$

**Lemma 3.5** PSS and ISS are independent. ■

**Proof.** From (3.4) and (3.13), PSS and ISS can be written in matrix form. Note that  $\mathbf{H}'\mathbf{F}\mathbf{H} = (\mathbf{M}\mathbf{H}_2)'\mathbf{F}\mathbf{M}\mathbf{H}_2$ , where  $\mathbf{M}$  is a matrix  $G \times N_T$  with elements:

$$m_{ij} = \begin{cases} \frac{1}{N_i} & \text{if } \sum_{l=1}^{i-1} N_l < j \leq \sum_{l=1}^i N_l, \quad i = 1 \dots G. \\ 0 & \text{otherwise .} \end{cases}$$

Note that  $\mathbf{H}_2'\mathbf{F}_2\mathbf{H}_2$  and  $\mathbf{H}_2'\mathbf{M}'\mathbf{F}\mathbf{M}\mathbf{H}_2$  are independent if and only if  $\mathbf{F}_2 \Sigma_{02} \mathbf{M}'\mathbf{F}\mathbf{M} = \mathbf{0}$  (Searle, 1971). From (3.5) and (3.14),

$$\mathbf{F}_2 \Sigma_{02} \mathbf{M}'\mathbf{F}\mathbf{M} = \frac{\sum_k p_k (1 - p_k)}{N_T} \Phi ,$$

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

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

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

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

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

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

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

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

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

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

**Proof that  $\mathbf{F}$  is positive semi-definite.**

From (3.5) we have that  $\mathbf{F} = K\mathbf{F}^*$ , where  $\mathbf{F}^*$  is a  $G \times G$  symmetric matrix with elements

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

i.e.,  $\mathbf{F} = \frac{K}{N_T} \mathbf{F}^{**}$ , where  $\mathbf{F}^{**}$  is a symmetric matrix with elements

$$f^{**}(g, g) = N_g \sum_{j \neq g} N_j \quad \text{and} \quad f^{**}(g, g') = -N_g N_{g'} \quad g' \neq g$$

Let  $\mathbf{x} = (x_1, x_2, \dots, x_G)'$ ,  $\mathbf{x} \neq \mathbf{0}$ , be a column vector of dimension  $G$ , then

$$\mathbf{x}'\mathbf{F}\mathbf{x} = \frac{K}{N_T} \mathbf{x}'\mathbf{F}^{**}\mathbf{x} ,$$

$\mathbf{x}'\mathbf{F}^{**}$  is a row vector of dimension  $G$  with the  $i$ -th element being:

$$x_i N_i \sum_{j \neq i} N_j - N_i \sum_{j \neq i} x_j N_j .$$

Therefore,

$$\begin{aligned} \mathbf{x}'\mathbf{F}^{**}\mathbf{x} &= \sum_i x_i \left( x_i N_i \sum_{j \neq i} N_j - N_i \sum_{j \neq i} x_j N_j \right) \\ &= \sum_i \sum_{j \neq i} (N_i N_j x_i^2 - N_i N_j x_i x_j) \\ &= \sum_i \sum_{j \neq i} N_i N_j (x_i^2 - x_i x_j) = \sum_i \sum_{j > i} N_i N_j (x_i - x_j)^2 \geq 0 . \end{aligned}$$

■

**Proof that  $\mathbf{F}_2$  is positive semi-definite.**

We have that  $\mathbf{F}_2 = K\mathbf{F}_2^*$  is a  $N_T \times N_T$  symmetric matrix, with  $\mathbf{F}_2^*$  having diagonal blocks  $\mathbf{A}_g$  given in (3.14). Let  $\mathbf{x} = (x_1, x_2, \dots, x_{N_T})'$ ,  $\mathbf{x} \neq \mathbf{0}$  be a column vector of dimension  $N_T$ . From Lemma (3.3) we have that  $\mathbf{F}_2^*$  is idempotent, therefore,

$$\begin{aligned} \mathbf{x}'\mathbf{F}_2\mathbf{x} &= K\mathbf{x}'\mathbf{F}_2^*\mathbf{x} = K\mathbf{x}'\mathbf{F}_2^*\mathbf{F}_2^*\mathbf{x} \\ &= K\mathbf{x}'(\mathbf{F}_2^*)'\mathbf{F}_2^*\mathbf{x} = K(\mathbf{F}_2^*\mathbf{x})'(\mathbf{F}_2^*\mathbf{x}) \geq 0 . \end{aligned}$$

■

## Acknowledgements

We are sincerely thankful for the comments made by the anonymous reviewers. 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.

(Received April, 2003. Accepted August, 2004.)

## References

- Comes, H. P., and Abbott, R. J.(2002). 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.

- Evett, I. W. and Weir, B. S. (1998). *Interpreting DNA Evidence: Statistical Genetics for Forensic Scientists*. Sunderland: Sinauer.
- Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distance among DNA haplotypes application to human mitochondrial DNA restriction data. *Genetics*, **131**, 479-491.
- Felsenstein, J. (2003). *Inferring Phylogenies*. Sunderland: Sinauer.
- Frankham, R., Ballou, J. D. and Briscoe, D. A. (2002). *Introduction to Conservation Genetics*. Cambridge: Cambridge University Press.
- Griffiths, A. J. F., Miller, J. H., Suzuki, D. T., Lewontin, R. C. and Gelbart, W. M. (2000). *An Introduction to Genetic Analysis*. 7th ed. W. H. Freeman.
- Grosberg, R. K., Levitan, D. R. and Cameron, B. B. (1996). Characterization of genetic structure and genealogies using RAPD-PCR markers: A random primer for the novice and nervous. In *Molecular Zoology: Advances, Strategies and Protocols* (eds. J.D. Ferraris and S.R. Palumbi). New York: Wiley-Liss, pp.67-100.
- Gunter L. E., Black A. S., Ratnayeke S., Tuskan G. A. and Wullschleger S. D. (2003). Assessment of genetic similarity among 'Alamo' switchgrass seed lots using RAPD markers. *Seed Science Technology*, **31**, 681-689.
- 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.
- Lehmann, E. L. (1999). *Elements of Large-Sample Theory*. New York: Springer-Verlag.
- Nei, M. (1972). Genetic distance between populations. *American Naturalist*, **106**, 283-292.
- Persson, H. A., Lundquist, K. and Nybom, H. (1998). RAPD analysis of genetic variation within and among populations of Turk's-cap lily (*Lilium martagon* L.). *Hereditas*, **128**, 213-220.
- Pinheiro, H. P., Pinheiro, A. S. and Sen, P. K. (2005). Comparison of genomic sequences using the Hamming distance. *Journal of Statistical Planning and Inference*, **130**, 325-339.
- Pinheiro, H. P. and Seillier-Moiseiwitsch, F. (2001). Quantifying heterogeneity in the HIV genome. In *Computational and Evolutionary Analysis of HIV Molecular Sequences* (eds. A.G. Rodrigo and G.H. Learn Jr.). Boston: Kluwer Academic Publishers, pp. 91-119.

- 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., 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). Amsterdam: Elsevier, pp. 713-746.
- Rao, C. R. (1965). *Linear Statistical Inference and Its Applications*. New York: John Wiley and Sons.
- Searle, S. L. (1971). *Linear Models*. New York: John Wiley and Sons.
- Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*. London: Chapman-Hall.
- 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 hierarchical systems. *Biological Conservation*, **104**, 119-126.
- Verovnik R., Sket B., Prevorcnik S. and Trontelj P. (2003). Random amplified polymorphic DNA diversity among surface and subterranean populations of *Asellus aquaticus* (Crustacea: Isopoda). *Genetica*, **119**, 155-165.
- 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.S. (1996). *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland: Sinauer.
- Welsh, J. and McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research*, **18**, 7213-7218.
- Welsh, J., Peters, C. and McClelland, 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.

**Roberta de Souza**

Instituto Internacional de Pesquisas Farmacêuticas-IIPF  
CEP 13186-481, Hortolândia, SP, Brazil.  
E-mail: robertasouza@institutoipf.com.br

**Hildete P. Pinheiro**

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

**Cibele Q. da Silva**

Departamento de Estatística  
Universidade Federal de Minas Gerais  
MG, Brazil.  
E-mail: cibeles@est.ufmg.br

**Sérgio F. dos Reis**

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