Brazilian Journal of Biometrics, **42**, 171–181 (2024)
doi:10.28951/bjb.v42i2.695

BRAZILIAN JOURNAL OF BIOMETRICS
ISSN:2764-5290

ARTICLE

Sequential Bayesian approach for genetic diversity analysis of the piracanjuba fish (*Brycon orbignyanus*)

Isabela da Silva Lima,*[,1] Carla Regina Guimarães Brighenti,*[,2] and Gabriel de Menezes Yazbeck[2]

[1]Postgraduate Program in Statistics and Agricultural Experimentation, Institute of Exact and Technological Sciences, Federal University of Lavras, Lavras, Brazil
[2]Department of Animal Science, Federal University of São João del-Rei, São João del-Rei, Brazil
*Corresponding author. Email: isabela_lima30@hotmail.com; carlabrighenti@ufsj.edu.br

(Received: September 30, 2023; Revised: December 06, 2023; Accepted: January 15, 2024; Published: April 15, 2024)

Abstract

In the sequential Bayesian approach, the sample size is not fixed before the experiment; it is determined based on the observations made. The procedure concludes when there is enough information to estimate the parameters, according to a stopping criterion. A parameter of interest in population genetics is the proportion obtained from the allele frequency at one or more loci to verify Hardy-Weinberg equilibrium (HWE). The objective of this study was to assess the occurrence of HWE in a population of piracanjuba fish (Brycon orbignyanus) by estimating the allele proportion and expected genotype proportions using a sequential Bayesian approach. Additionally, a comparison was made with frequentist and Bayesian approaches. Initially, genotypic profiles were analyzed at a microsatellite DNA locus, Bh6, in 49 fish, to determine the frequency of observed alleles and genotypes at the UFSJ Genetic Resources Laboratory. Seven allele classes were observed; thus, under the assumption of sampling independence, the likelihood is multinomial. The estimation of allele and genotype proportions was then carried out using frequentist, Bayesian, and sequential Bayesian approaches. A uniform prior and a cost of 10^{-3} were considered. The estimates from the three approaches were compared, and it was concluded that the sequential approach proved effective, utilizing only 55.1% of the available data, thereby reducing the sample size and optimizing the procedure. Using a chi-square test at a 5% probability level, it was concluded that the studied sample is in Hardy-Weinberg equilibrium (p-value: 0.9800245).

Keywords: Multinomial; Dirichlet; Sampling; Population Genetics; Aquaculture.

1. Introduction

The sequential sampling is characterized by using samples of variable size; therefore, the sample size is not fixed before the experiment but is determined based on the observations made. The

incorporation of Bayesian techniques into sequential sampling allows the use of *a priori* information that can optimize the sampling plan and improve parameter estimation (Wald, 1947).

Thus, in the Bayesian sequential approach, the procedure is stopped when sufficient information is obtained to estimate the desired parameters, according to a stopping criterion that compares the immediate and expected risks at each sample element. The decision to stop sampling is made when the immediate risk is lower than the expected risk, and the parameters of interest are estimated (Berger, 1985).

This approach can be applied in various areas with the aim of reducing the required sample size for making decisions about parameters. Therefore, it is particularly useful in processes where sampling is destructive and incurs high financial and/or execution time costs (Schilling & Neubauer, 2017).

In the context of population genetics, the parameters of interest are the proportions obtained from the frequencies of alleles and genotypes at one or more loci, to check for Hardy-Weinberg equilibrium (HWE). According to Hartl & Clark (2010), a population is in HWE when the genotype proportions are distributed as expected based on the occurrence of random mating, a phenomenon known as panmixia, for a given distribution of allele proportions (genetic diversity). In general, this equilibrium occurs when allele and genotype proportions remain constant.

However, the process of genetic population description is conducted meticulously, requiring the visual analysis of one element at a time at each DNA locus. This procedure, besides consuming considerable time, demands significant resources, making it an exhaustive task. Therefore, the application of the Bayesian sequential approach is relevant in the field of population genetics, as it may result in a reduction of the required sample size to verify Hardy-Weinberg equilibrium, optimizing the procedure.

In a two-allele gene system, A and a, the proportion of these alleles follows a binomial distribution. Meanwhile, the proportion of genotypes in a population can be described by a multinomial distribution with three categories: homozygote AA, heterozygote Aa, and homozygote aa. However, alleles generally have more than two categories, exhibiting genetic diversity; thus, the allele proportion is also determined by a multinomial distribution (Rehman *et al.,* 2020).

Thus, the objective of this study was to verify the occurrence of Hardy-Weinberg equilibrium in a population of piracanjuba fish (*Brycon orbignyanus*), an endangered species, by estimating allele proportions and the expected proportions of genotypes. This was done using a Bayesian sequential approach for the multinomial distribution and comparing it with frequentist and Bayesian approaches.

2. Matherials and Methods

2.1 Multinomial distribution

The multinomial distribution is a discrete probability distribution and a generalization of the binomial distribution for polytomous response variables, used to estimate the probability of an element belonging to more than two categories.

According to Casella & Berger (2002), the multinomial distribution is defined by assuming an experiment whose result is one of the events $E_1, E_2, ..., E_k$, with probabilities $P[E_i] = p_i$, where k is the number of classes of the multinomial distribution. For $i = 1, 2, ..., k$, $0 \leq p_i \leq 1$, and $\sum_{i=1}^{k} p_i = 1$. Let X_i be a random variable that counts the number of occurrences of E_i in n independent repetitions of this experiment. Then, the random vector $(X_1, X_2, ..., X_k)$ has a distribution called **multinomial**, with parameters $p_1, p_2, ..., p_{k-1}$, and n, given by:

$$f_{X_1,X_2,...,X_k}(x_1, x_2, ..., x_k) = P[X_1 = x_1, X_2 = x_2, ..., X_k = x_k] =$$

$$= \frac{n!}{x_1!x_2! \cdot \ldots \cdot x_k!} p_1^{x_1} p_2^{x_2} \cdot \ldots \cdot p_k^{x_k} = n! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!}. \quad (1)$$

where each X_i is a positive integer, $p_1, p_2, ..., p_k$ are population proportions, and $\sum_{i=1}^{k} x_i = n$. There are $p_1, p_2, ..., p_{k-1}$ parameters because $\sum_{i=1}^{k} p_i = 1$, thus $p_k = 1 - \sum_{i=1}^{k-1} p_i$.

2.1.1 The Bayesian estimation of parameters for the multinomial distribution

The Dirichlet distribution is the multivariate generalization of the beta distribution, with a non-negative real vector parameter $\boldsymbol{a}$. It is a widely used multivariate discrete distribution in the Bayesian context as the conjugate prior distribution for the multinomial distribution (Paulino *et al.,* 2018).

Let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a vector with k components; then, it follows a Dirichlet distribution of order $k \geq 2$ with a parameter vector $\boldsymbol{a} = (a_1, ..., a_k)^T$, i.e., (Paulino *et al.,* 2018):

$$(\boldsymbol{X}|\boldsymbol{a}) \sim Dirichlet(\boldsymbol{a}).$$

Its probability density function is given by:

$$f_X(x_1, ..., x_k) = \frac{\Gamma(a_0)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} p_i^{a_i - 1}, \ 0 < p_i < 1, \quad (2)$$

where $a_0 = \sum_{i=1}^{k} a_i$, $\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} dx$ is the gamma function, and $\sum_{i=1}^{k} p_i = 1$. The marginal distribution is a beta distribution with parameters a_i and $(a_0 - a_i)$ for each i, from which:

$$E(X_i) = \frac{a_i}{a_0}, \qquad Var(X_i) = \frac{a_i(a_0 - a_i)}{a_0^2(a_0 + 1)}, \qquad Cov(X_i, X_j) = \frac{-a_i a_j}{a_0^2(a_0 + 1)}. \quad (3)$$

If the prior distribution is Dirichlet and the observed variable follows a multinomial distribution, then the posterior distribution will be another Dirichlet distribution with different parameters. Therefore, the posterior distribution follows a Dirichlet distribution with parameters:

$$\boldsymbol{p}|X \sim \text{Dirichlet}(a_1^* = x_1 + a_1, ..., a_k^* = x_k + a_k).$$

where $X = (x_1, ..., x_k, a_1, ..., a_k)^T$.

Thus, the mean, variance, and covariance of the Dirichlet posterior distribution are given by:

$$E(\boldsymbol{p}|X_i) = \frac{x_i + a_i}{\sum_{i=1}^{k}(x_i + a_i)}, \quad (4)$$

$$Var(\boldsymbol{p}|X_i) = \frac{(x_i + a_i)\left\{\left[\sum_{i=1}^{k}(x_i + a_i)\right] - (x_i + a_i)\right\}}{\left[\sum_{i=1}^{k}(x_i + a_i)\right]^2 \left\{\left[\sum_{i=1}^{k}(x_i + a_i)\right] + 1\right\}}, \quad (5)$$

$$Cov(\boldsymbol{p}|X_i, X_j) = \frac{-(x_i + a_i)(x_j + a_j)}{\left[\sum_{i=1}^{k}(x_i + a_i)\right]^2 \left\{\left[\sum_{i=1}^{k}(x_i + a_i)\right] + 1\right\}}. \quad (6)$$

2.1.2 The Bayesian sequential estimation of parameters for the multinomial distribution

According to Berger (1985), the main idea behind the Bayesian sequential estimation procedure is that when making each observation one at a time, one should compare the *a posteriori* Bayesian risk of making an immediate decision with the expected *a posteriori* Bayesian risk, which will be obtained if more observations are taken.

Moreover, the Bayesian sequential rule can also be known as Bayesian learning because the *a posteriori* distribution calculated at the current n will be used to update the *a priori* distribution yet to be used in the $(n+1)$-th evaluation (Berger, 1985).

In this regard, to determine the stopping criterion, it is necessary to calculate these risks for the multinomial distribution. Lima (2022) established these risks for Dirichlet conjugate priors and detailed the derivations; for more details, refer to the mentioned work.

To obtain the stopping criterion, a quadratic loss function was considered for the parameter estimate $\boldsymbol{p} = (p_1, p_2, ..., p_k)^T$ by $\hat{\boldsymbol{p}} = (\hat{p}_1, \hat{p}_2, ..., \hat{p}_k)^T$, and it takes the general quadratic form $(\boldsymbol{p} - \hat{\boldsymbol{p}})^T \boldsymbol{K} (\boldsymbol{p} - \hat{\boldsymbol{p}})$, where $\boldsymbol{K}$ is a positive definite symmetric $I \times I$ matrix of constant loss (Chen, 1988; Jones, 1976; Owen, 1970).

Thus, using a quadratic loss function, the Bayesian estimator $\hat{\boldsymbol{p}}$ is the mean of the *posteriori* distribution of $(\boldsymbol{x}, n)$, i.e., the mean of the *posteriori* Dirichlet distribution, given by (4).

The immediate risk, or the stopping risk of making a decision, is given by:

$$S\left((x_1, ..., x_k), n\right) = S(\boldsymbol{x}, n) = \text{trace } \boldsymbol{K}\boldsymbol{\Sigma}. \tag{7}$$

where $\boldsymbol{\Sigma}$ is the dispersion matrix of the *posteriori* Dirichlet distribution, having the *posteriori* variances on its main diagonal and the *posteriori* covariances of the Dirichlet distribution in the other components.

This results in:

$$S(\boldsymbol{x}, n) = \frac{\sum_{i=1}^{k} K_{ii}(x_i + a_i)\left(\sum_{i=1}^{k}(x_i + a_i)\right) - \sum_{i=1}^{k}\sum_{j=1}^{k} K_{ij}(x_i + a_i)(x_j + a_j)}{\left[\sum_{i=1}^{k}(x_i + a_i)\right]^2 \left\{\left[\sum_{i=1}^{k}(x_i + a_i)\right] + 1\right\}}. \tag{8}$$

Dynamic programming equations were used to find the expected risk, resulting in:

$$B(\boldsymbol{x}, n) = c + S(\boldsymbol{x}, n)\left(\frac{a_0 + n}{a_0 + n + 1}\right). \tag{9}$$

where c is the cost of sampling one observation.

Therefore, the stopping criterion boils down to comparing the values of immediate and expected risks for each observation, given by the expressions found in (8) and (9). When the immediate risk is less than the expected one, i.e., $S(\boldsymbol{x}, m) < B(\boldsymbol{x}, m)$, the sampling stops, and the parameters of interest are estimated. Otherwise, if $S(\boldsymbol{x}, m) > B(\boldsymbol{x}, m)$, the sampling continues, taking another observation until a decision can be made.

2.2 Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium is one of the main topics studied in population genetics. In 1908, the English mathematician Godfrey Harold Hardy and the German physician Wilhelm Weinberg independently and almost simultaneously arrived at the Hardy-Weinberg Equilibrium Law.

To check the HWE, one must calculate the expected proportion of each genotype based on the allelic proportions p_1 and p_2, representing the proportion of alleles A and a, respectively, in the population. Thus, the expected genotypic proportions are estimated from the Hardy-Weinberg equation, which is given by a quadratic expansion of allelic proportions (Hartl & Clark, 2010):

$$(p_1 + p_2)^2 = 1 \Rightarrow p_1^2 + 2p_1p_2 + p_2^2 = 1, \tag{10}$$

where p_1^2 represents the proportion of homozygotes AA, $2p_1p_2$ the proportion of heterozygotes Aa, and p_2^2 the proportion of homozygotes aa.

In the case of a single-gene system with k alleles, $k \geq 2$, the probability distribution associated with genotypes is a multinomial distribution, and it will have $\frac{k(k+1)}{2}$ classes because the number of categories/classes of the multinomial distribution of genotypes is given by the combination of the number of alleles taken 2 by 2, added to the number of alleles, considering that species generally have a diploid number of chromosomes. i.e.,

$$C_{k,2} + k = \frac{k!}{(k-2)!2!} + k = \frac{k(k-1)(k-2)!}{(k-2)!2!} + k = \frac{k(k-1) + 2k}{2} = \frac{k^2 + k}{2} = \frac{k(k+1)}{2}. \tag{11}$$

The Hardy-Weinberg equation to estimate the expected genotypic proportions when there is genetic diversity, i.e., when the population has more than two alleles ($k \geq 2$), is given by generalizing the expansion $(p_1 + p_2 + ... + p_k)^2 = 1$ where p_n, with $n = 1, ..., k$, is the proportion of alleles, resulting in:

$$p_1p_1 + p_1p_2 + ... + p_1p_k + p_2p_1 + p_2p_2 + ... + p_2p_k + ... + p_kp_1 + p_kp_2 + ... + p_kp_k = 1$$

$$p_1^2 + 2p_1p_2 + ... + 2p_1p_k + p_2^2 + ... + 2p_2p_k + ... + p_k^2 = 1$$

Therefore, $p_{ij} = 2p_ip_j$, $i, j = 1, ..., k$, if $i \neq j$, and $p_{ij} = p_i^2$, if $i = j$, where p_{ij} are the proportions of genotypes A_iA_j, with $j \geq i$. This relationship can be written in matrix form, where on the main diagonal, the expected proportions of homozygous genotypes are, and the rest are the expected proportions of heterozygous genotypes:

$$\begin{bmatrix} p_1^2 & 2p_1p_2 & \cdots & 2p_1p_j \\ 0 & p_2^2 & \cdots & 2p_2p_j \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_i^2 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} \\ 0 & p_{22} & \cdots & p_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{ij} \end{bmatrix}. \tag{12}$$

2.3 Application

All the theory of Hardy-Weinberg equilibrium was applied to a dataset of a microsatellite DNA locus, Bh6, from 49 fish of the species Brycon orbignyanus, commonly known as piracanjuba. These fish are from the Rio das Mortes basin, São João del-Rei - MG municipality, and are used as breeders in fish farming, being a threatened species.

Initially, the analysis of genotypic profiles of the microsatellite DNA locus Bh6 from the 49 fish was performed using polyacrylamide gel electrophoresis to determine the proportions of observed alleles and genotypes at the Laboratory of Genetic Resources of the Department of Animal Science at the Federal University of São João del-Rei, São João del-Rei, Minas Gerais (LARGE-DEZOO-UFSJ).

The allele and genotype proportions were estimated using three approaches: frequentist, Bayesian, and sequential Bayesian. The sample size for the frequentist and Bayesian approaches was 49 fish.

For the frequentist approach, the proportion of observed genotypes was calculated as follows:

$$p_{ij} = \frac{n_{ij}}{n}. \tag{13}$$

The allele proportion:

$$p_i = \frac{n_i + n_j}{2n_{ij}}, \tag{14}$$

where n_{ij} is the observed count in each category $i, j = 1, ..., k, j \geq i$, n is the total count, n_i is the sum of counts in each category in row i, and n_j is the sum of counts in each category in column j, in a contingency table with L rows and C columns.

And the expected genotype proportions calculated according to the Hardy-Weinberg equation, using the allele proportions:

$$\begin{cases} p_{ij} = p_i^2, & \text{if i=j} \\ p_{ij} = 2p_i p_j, & \text{if i}\neq\text{j} \end{cases}, \tag{15}$$

where p_{ij} are the proportions of genotypes A_iA_j, and p_i are the proportions of alleles, $i, j = 1, ..., k$ with $j \geq i$.

For the Bayesian and sequential Bayesian approaches, a Dirichlet conjugate prior was adopted with hyperparameters equal to one, thus having as a particular case of the Dirichlet, a uniform prior distribution, where all values are equally probable, being a non-informative *priori*.

In the Bayesian approach, allele proportion estimates were calculated by the mean of the Dirichlet posterior distribution given in (4). The expected genotype proportions were then calculated using (15).

For the sequential Bayesian approach, a cost of 10^{-3} was considered, which is constant and additive in the loss function. This value is associated with the precision of the p values. According to Bach (2015), the cost value should have a similar order of magnitude as the loss function. This ensures that the risk function is not exclusively dominated by the cost. By considering the quadratic loss function, as the loss is the square of a difference between real and estimated proportion values, which are in the interval $[0, 1]$, the results are always close to zero, and therefore, the cost should also be close to zero. Thus, the chosen cost value is in line with Bach (2015) and does not dominate the stopping criterion.

In the sequential Bayesian approach, the procedure stops when the immediate risk is lower than expected, given by the expressions (8) and (9), obtained through the quadratic loss function. Thus, the estimates of observed genotypes were calculated by the mean of the Dirichlet *a posteriori* distribution, given by (4). With the sample size at which the procedure stopped, allele proportions were calculated using (14), and from these, expected genotype proportions were calculated using (15).

The sampling order of the sequential procedure followed the order of the spreadsheet analyses. Thus, the first fish analyzed from the spreadsheet, the second, and so on. Just for clarification, as this aspect can affect early or late stopping in relation to the obtained result.

A chi-square test was performed to check whether the locus is in Hardy-Weinberg equilibrium or not.

The chi-square test is used to verify if the frequency of a certain event observed in a sample differs significantly from the expected frequency of that event. This quantification is done through the chi-square statistic, defined as (Bussab & Morettin, 2017):

$$\chi^2_{calculated} = \frac{\sum_{i=1}^{n} (observed_i - expected_i)^2}{expected_i}, \tag{16}$$

where observed and expected refer to the observed and expected frequencies in each genotypic class.

After calculating the value of $\chi^2_{calculated}$ (Equation 16), it is compared with the $\chi^2_{tabulated}$ value for the appropriate degrees of freedom (d.f.) and desired significance level (α). If $\chi^2_{calculated} \geq \chi^2_{tabulated}$, H_0 is rejected; otherwise, H_0 is not rejected. The degrees of freedom (d.f.) are calculated as the number of data classes - number of estimated parameters - 1 (Hartl & Clark, 2010).

Thus, it was tested whether the number of individuals in each genotypic class corresponds to the expected under the hypothesis that the population is in Hardy-Weinberg equilibrium at a 5% probability level. The hypotheses are:

$$\begin{cases} H_0 : \text{The population is in HWE.} \\ H_1 : \text{The population is not in HWE.} \end{cases}$$

The estimates from the three approaches were compared by calculating the percentage error, the correlation between the estimates, and the confidence interval for the differences in estimated proportions between two approaches.

The Percentage Error (PE) is given by the expression:

$$PE = \frac{(|p_{i1} - p_{i2}|) \times 100\%}{p_{i2}}, \tag{17}$$

where PE is the percentage error in the estimation of the proportion parameter, p_{i1} is the proportion parameter estimated by approach 1, p_{i2} is the proportion parameter estimated by approach 2 to be compared.

The confidence interval for the difference in proportions is given by:

$$IC_{1-\alpha}(p) : (\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}}. \tag{18}$$

3. Results and Discussion

In the utilized dataset, seven allele classes were observed, denoted as $k = 1, ..., 7$. Thus, assuming sample independence, the likelihood follows a multinomial distribution. The multinomial distribution for genotypes has 28 categories, calculated by the expression 11.

The estimation of observed genotype proportions and subsequently the estimation of allele proportions and expected genotype proportions were carried out using the frequentist, Bayesian, and sequential Bayesian approaches. In the sequential Bayesian approach, the procedure was interrupted, and the estimate was obtained with a sample size of 27 fish.

In Figure 1 and Table 1, the results of the allele proportion estimates by the three approaches are presented:

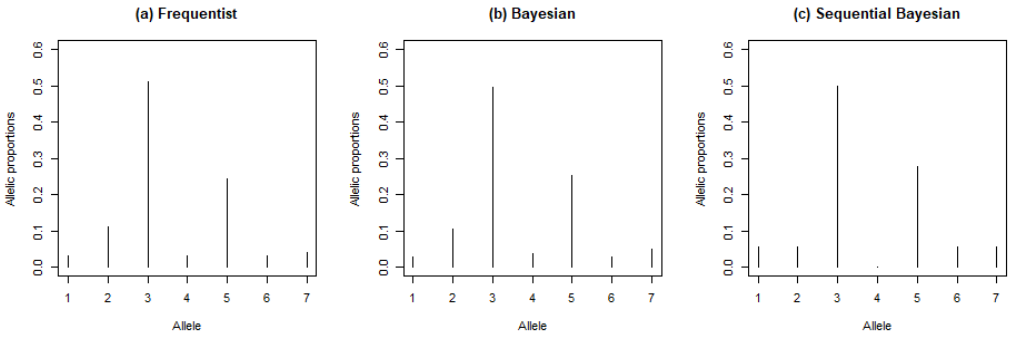

Figure 1. Allele proportions estimated by the three approaches.

Table 1. Estimates of allele proportions under the three approaches: frequentist (f.), Bayesian (b.), and sequential Bayesian (s. b.)

Parameter	$\hat{p}_i$ f.	$\hat{p}_i$ b.	$\hat{p}_i$ s. b.
p_1	0,031	0,029	0,056
p_2	0,112	0,104	0,056
p_3	0,510	0,497	0,500
p_4	0,031	0,038	0,000
p_5	0,245	0,254	0,278
p_6	0,031	0,029	0,056
p_7	0,041	0,049	0,056
n	49	49	27

Source: Authors (2023).

As the expected genotype proportions are calculated based on allele proportions, given by the Hardy-Weinberg equation (15), only the allele estimate comparisons were made.

To do this, the Percentage Error (%) between the approaches was calculated pairwise, and also the confidence interval of the differences in proportions, with a 95% confidence level. The results are shown in Table 2.

In the PE between estimates by the Bayesian and frequentist approaches, the first approach was considered Bayesian, and the second one was considered frequentist. In the PE between estimates by the sequential Bayesian and frequentist approaches, the first approach was considered sequential Bayesian, and the second one was considered frequentist. Finally, in the PE between estimates by the sequential Bayesian and Bayesian approaches, the first approach was sequential Bayesian, and the second one was Bayesian.

Table 2. Percentage Errors (PE) % and Confidence Intervals (CI) for the differences in proportions among the three approaches

Parameter	PE (f. x b.) %	PE (f. x s. b.) %	PE (b. x s. b.) %	$CI_{95\%}$ (f. - b.)	$CI_{95\%}$ (f. - s. b.)	$CI_{95\%}$ (b. - s. b.)
p_1	5,267	81,481	91,571	[-0,066; 0,069]	[-0,124; 0,074]	[-0,125; 0,072]
p_2	7,345	50,505	46,581	[-0,115; 0,131]	[-0,067; 0,180]	[-0,073; 0,170]
p_3	2,588	2,000	0,604	[-0,185; 0,211]	[-0,225; 0,245]	[-0,239; 0,232]
p_4	24,133	100,000	100,000	[-0,079; 0,065]	[-0,018; 0,079]	[-0,016; 0,092]
p_5	3,717	13,426	9,361	[-0,180; 0,162]	[-0,240; 0,175]	[-0,232; 0,185]
p_6	5,267	81,481	91,571	[-0,066; 0,069]	[-0,124; 0,074]	[-0,125; 0,072]
p_7	20,050	36,111	13,379	[-0,090; 0,074]	[-0,117; 0,088]	[-0,112; 0,099]

Source: Authors (2023).

It can be observed from Table 2 that the lowest percentage errors were between estimates from the frequentist and Bayesian approaches. However, when the parameter is larger, the sequential Bayesian approach is quite effective, as observed in the case of p_3 and p_5.

Moreover, in Table 2, all confidence intervals for the differences in proportion estimates contain zero, meaning that the difference in the proportion may be zero, indicating non-significance. This highlights that the sequential approach was efficient in estimates and utilized only 55.1% of the available data, optimizing the procedure.

Another point to be highlighted is that, with the Bayesian sequential approach, the sample size was 27 fish, of which only the presence of 6 alleles was detected. This resulted in the estimate of p_4 being equal to zero. However, the non-estimation of allele 4 is not considered a problem in population genetics, as it is a rare allele, and therefore, its influence on Hardy-Weinberg equilibrium is low. This fact can be confirmed by the low occurrence of allele 4 in the total sample of 49 individuals, where it was observed at a low frequency (3%), being present in the last individuals evaluated in the total sample (43 and 49).

Since allele 4 is a rare allele, even if it is present in the population, its frequency will be low. Therefore, the frequency of its combinations, i.e., genotypic frequencies, will also be low. This results in a low impact on the calculation of Hardy-Weinberg equilibrium, which depends only on the adherence between the observed and expected. Thus, if the observed is rare, then the expected will also be rare, resulting in little influence on the equilibrium, since estimates of rare alleles, when present, will be close to zero.

It is also important to note that, in this study, 49 fish were observed for analysis by traditional (frequentist) methods initially, of which 7 alleles were identified. But by taking a larger number than 49 fish, it may be possible to identify more rare alleles. Thus, even using other approaches with a fixed sample size, it is not possible to identify all alleles present in the population. Alleles that are considered rare, with low frequency, will be less identified. Therefore, the number of alleles cannot be pre-fixed, as it always depends on the analyzed sample size.

Another point to emphasize is that with the sequential Bayesian approach, the sample size was 27 fish, and with these 27 fish, the presence of only 6 alleles was detected, resulting in the estimate of p_4 being zero. It should be noted that even using another approach, the same thing could happen, for example, when analyzing more fish, and the presence of more alleles is detected. Therefore, the number of alleles cannot be predetermined; this quantity will always depend on the analyzed sample size.

Additionally, the correlation between the estimates was calculated, resulting in 0.9991 between the frequentist and Bayesian approaches, 0.9826 between the frequentist and sequential Bayesian approaches, and 0.9852 between the Bayesian and sequential Bayesian approaches. All values are close to 1, indicating a strong correlation.

Histograms of the posterior distributions of alleles from the sequential Bayesian approach were constructed using the R software (R Core Team, 2023), presented in Figure 2.

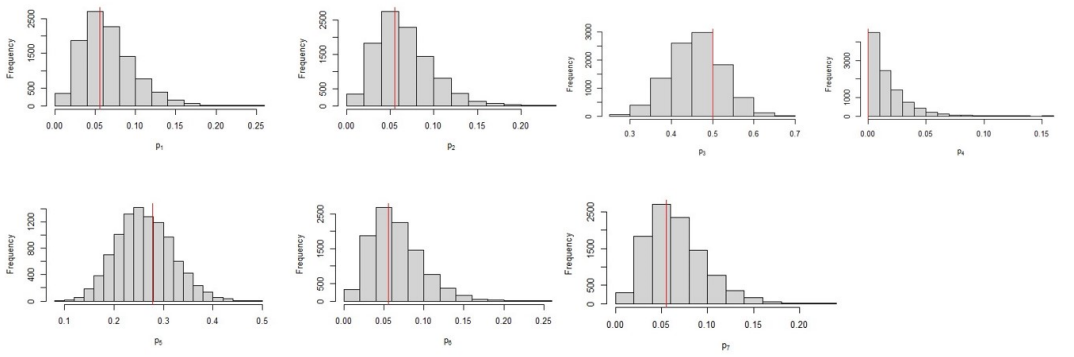

Figure 2. Histograms of the posterior distributions of alleles from the sequential Bayesian approach.

A chi-square test was conducted at a 5% probability level, with degrees of freedom = 28 – 6 – 1 = 21. It was concluded that the studied sample is in Hardy-Weinberg equilibrium, as the obtained p-value was 0.9800245, which is greater than 0.05, leading to the non-rejection of the null hypothesis.

It can be observed that the use of Bayesian methods with Dirichlet conjugate priors to assess Hardy-Weinberg equilibrium has been yielding good results. Reis *et al.* (2008) described a Bayesian method to study Hardy-Weinberg equilibrium through the inbreeding coefficient. In this work, the authors analyzed various models and concluded that the best model is the one using Dirichlet priors.

Moreover, Reis *et al.* (2011) concluded that the Bayesian methodology proved to be efficient in studying the Hardy-Weinberg model, being evaluated and confirmed by simulation studies, presenting estimates very close to the real value.

Furthermore, Cunha Filho *et al.* (2020) noticed that Bayesian analysis obtained results relatively closer to reality to verify the Hardy-Weinberg equilibrium hypothesis and has the advantage of being applicable to samples of any size. The Bayesian methods presented were efficient in testing Hardy-Weinberg equilibrium. Its application may serve as a subsidy for the researcher's decision-making to be as close as possible to reality.

It can be observed that there are works in the literature with the Bayesian approach, but with the sequential Bayesian approach in this area, it is a novelty. The results obtained with this approach were of great utility, as it optimized the procedure.

4. Conclusions

This study demonstrates that it was possible to apply the sequential Bayesian approach to estimate allele and genotype proportions, for verifying the occurrence of Hardy-Weinberg equilibrium, in a population of piracanjuba fish (*Brycon orbignyanus*).

There was efficiency in the sequential Bayesian approach, reducing the sample size and utilizing only 55.1% of the available data for analysis. Additionally, it is noteworthy that the use of this approach in this field is a novelty. Moreover, it can be applied in various other areas for different procedures of interest, aiming to reduce time and/or cost.

Acknowledgments

This work was supported by the Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior - CAPES. We also thank CEMIG for the project related to the data used in this study.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Bach, D. R. A cost minimisation and Bayesian inference model predicts startle reflex modulation across species. *Journal of Theoretical Biology* **370,** 53–60 (2015).
2. Berger, J. O. *Statistical decision theory and Bayesian analysis* 2nd ed. (Springer Science & Business Media, New York, 1985).
3. Bussab, W. O. & Morettin, P. A. *Estatística Básica* 9th ed. (Editora Saraiva, São Paulo, 2017).
4. Casella, G. & Berger, R. L. *Statiscal Inference* 2nd ed. (Duxbury Press, df, 2002).
5. Chen, S.-Y. Restricted risk Bayes estimation for the mean of the multivariate normal distribution. *Journal of Multivariate Analysis* **24,** 207–217 (1988).
6. Cunha Filho, M., de Oliveira, E. C. A., Piscoya, V. C., Moreira, G. R., Cunha, A. L. X. & Araújo Filho, R. N. Statistical analysis with a bayesian approach to the Hardy-Weinberg equilibrium. *Brazilian Journal of Biometrics* **38,** 69–78 (2020).
7. Hartl, D. L. & Clark, A. G. *Princípios de Genética de Populações-4* (Artmed Editora, 2010).
8. Jones, P. W. Bayes Sequential Estimation of Multinomial Parameters. *Mathematische Operationsforschung und Statistik* **7,** 123–127 (1976).
9. Lima, I. d. S. *Estatística sequencial bayesiana dos parâmetros da distribuição multinomial* MA thesis (Universidade Federal de Lavras, Lavras, 2022).
10. Owen, R. J. The optimum design of a two-factor experiment using prior information. *The Annals of Mathematical Statistics* **41,** 1917–1934 (1970).
11. Paulino, C. D., Turkman, A. A., Murteira, B. & Silva, G. L. *Estatística bayesiana* 2nd ed. (Fundação Calouste Gulbenkian, Lisboa, 2018).
12. R Core Team. R: A language and environment for statistical computing. Version ISBN 3-900051-07-0. *R Foundation for Statistical Computing.* http://www.R-project.org/ (2023).
13. Rehman, A.-u., Iqbal, J., Shakeel, A., Qamar, Z. u. & Rana, P. Hardy-Weinberg equilibrium study of six morphogenetic characters in a population of Punjab, Pakistan. *All Life* **13,** 213–222 (2020).
14. Reis, R. L. d., Muniz, J. A., Silva, F. F., Sáfadi, T. & Aquino, L. H. d. Bayesian inference in genetic analysis of diploid populations: inbreeding coefficient and outcrossing rate estimation. *Ciência Rural* **38,** 1258–1265 (2008).
15. Reis, R. L. d., Muniz, J. A., Silva, F. F., Sáfadi, T. & Aquino, L. H. d. Comparação bayesiana de modelos com uma aplicação para o equilíbrio de Hardy-Weinberg usando o coeficiente de desequilíbrio. *Ciência Rural* **41,** 834–840 (2011).
16. Schilling, E. G. & Neubauer, D. V. *Acceptance sampling in quality control* (Crc Press, London, 2017).
17. Wald, A. *Sequential Analysis* (John Willey & Sons, New York, 1947).