ISSN: 2317-0840

Estimação sequencial bayesiana da proporção de *loci* em equilíbrio de Hardy-Weinberg

Isabela S. Lima[1†], Carla R. G. Brighenti[1,2], Gabriel M. Yazbeck[2], Mírian Rosa[1], Edilene C. P. Azarias[1]

[1] *Universidade Federal de Lavras (UFLA).*
[2] *Universidade Federal de São João Del-Rei (UFSJ).*

Resumo: *A abordagem sequencial Bayesiana utiliza-se amostras de tamanho variável, sem a necessidade de determinar o tamanho previamente. A decisão de interromper a amostragem é baseada em um critério de parada de comparação de riscos. Essa abordagem é útil em processos que envolvem amostras destrutivas, com alto tempo e custo financeiro, ou em situações em que o tamanho amostral não é definido por uma regra pré-estabelecida. Diante disso, essa abordagem pode ser utilizada no contexto de genética de populações, para estimar a proporção de loci que estão em equilíbrio de Hardy-Weinberg (EHW), pois não há um padrão da quantidade de loci que são selecionados para a caracterização de uma população. Como existe um custo e tempo alto de operação laboratorial envolvidos, os loci são selecionados de acordo com os recursos disponíveis. O objetivo deste trabalho foi estimar a proporção de loci que estão em EHW, através da abordagem sequencial bayesiana, para caracterização da variabilidade genética do peixe dourado (Salminus brasiliensis). Cada locus foi verificado se estava em EHW, como a variável resposta é binária (está ou não em EHW) a distribuição de probabilidade associada é a binomial, então utilizou-se uma priori conjugada beta, cujos hiperparâmetros foram calculados com base em análises anteriores. Portanto, utilizando o critério de comparação dos riscos, após a avaliação do equilíbrio em 28 loci, o processo foi interrompido, e considerando uma função de perda quadrática, a estimação é dada pela média da distribuição beta a posteriori, o que resultou em uma estimativa de 50%.*

Palavras-chave: *Distribuição binomial; Genética de populações; Salminus brasiliensis.*

Bayesian sequential estimation of the proportion of *loci* in Hardy-Weinberg equilibrium

Abstract: *The Bayesian sequential approach uses samples of variable size, without the need to determine the size beforehand. The decision to stop sampling is based on a stopping criterion that compares risks. This approach is useful in processes involving destructive samples, with high time and financial costs, or in situations where the sample size is not defined by a pre-established rule. Given this, this approach can be used in the context of population genetics to estimate the proportion of loci in Hardy-Weinberg equilibrium (HWE), as there is no standard for the quantity of loci selected for population characterization. Due to the high cost and time involved in laboratory operations, loci are selected based on available resources. The objective of this study was to estimate the proportion of loci in HWE using the Bayesian sequential approach to characterize the genetic variability of the golden dorado (Salminus brasiliensis). Each locus was checked for HWE. Since the response variable is binary (in HWE or not), the associated probability distribution is binomial. A conjugate beta prior was used, whose hyperparameters were calculated based on previous analyses. Therefore, using the risk comparison criterion, after evaluating equilibrium in 28 loci, the process was terminated. Considering a quadratic loss function, the estimation is given by the mean of the posterior beta distribution, resulting in an estimate of 50%.*

Keywords: *Binomial distribution; Population genetics; Salminus brasiliensis.*

[†]Autor correspondente: isabela_lima30@hotmail.com.

Introdução

O uso de métodos matemáticos para representar as relações genéticas de uma população iniciou no século XX, com a famosa Lei de Hardy-Weinberg, que foi proposta em 1908. Uma população está em Equilíbrio de Hardy-Weinberg (EHW) quando apresenta proporção de alelos e genótipos constantes de uma geração para outra e há ocorrência de cruzamentos ao acaso, fenômeno denominado como panmixia (HARTL; CLARK, 2010).

Um parâmetro de interesse em genética de populações é a proporção de *loci* de DNA que estão em equilíbrio de Hardy-Weinberg. Nesse contexto, estimar esse parâmetro é de grande relevância, uma vez que não existe um padrão definido para a quantidade de *loci* a serem selecionados para caracterização da variabilidade genética da população (CHAGAS et al., 2015).

No entanto, existe um custo e tempo alto de operação laboratorial envolvidos para verificar o EHW em cada *loci* e, portanto, essa quantidade de *loci* é selecionada de acordo com os recursos disponíveis. Assim, possuir uma regra que dê o menor tamanho amostral possível que é necessário para a caracterização genética de uma população, sem perdas de informações, é vantajoso.

Ao verificar se cada *locus* está em equilíbrio de Hardy-Weinberg (EHW), obtém-se uma variável resposta binária (está ou não em EHW) sendo a distribuição de probabilidade associada a binomial. Além disso, no sistema de um gene com dois alelos, a proporção de genótipos em uma população pode ser descrita por uma distribuição multinomial com três categorias: homozigoto AA, heterozigoto Aa e homozigoto aa (REIS et al., 2011).

A amostragem sequencial caracteriza-se por utilizar amostras de tamanho variável, logo, o tamanho amostral não é fixado antes do experimento, ele é dado em função das observações realizadas. A inclusão de técnicas Bayesianas à amostragem sequencial permite a utilização de informações *a priori* que podem otimizar o plano de amostragem e melhorar a estimação de parâmetros (BERGER, 1985).

Assim, na abordagem sequencial Bayesiana, o procedimento é interrompido quando são obtidas informações suficientes para estimar os parâmetros desejados, de acordo com um critério de parada, que compara os valores de risco imediato e esperado a cada elemento amostral. Toma-se a decisão de interromper a amostragem quando risco imediato for menor que o risco esperado e estima-se os parâmetros de interesse.

A amostragem sequencial é útil principalmente em processos em que a amostra é destrutiva, de alto custo financeiro e alto tempo de execução, ou ainda, quando há investigações ou experimentos em que o tamanho amostral não é bem definido por uma regra (SCHILLING; NEUBAUER, 2017). Diante disso, a estimação sequencial Bayesiana pode ser utilizada no contexto de genética de populações, para estimar a proporção de *loci* que estão em equilíbrio de Hardy-Weinberg (EHW), evitando excesso de gastos financeiros e temporais.

Desse modo, o objetivo desse trabalho foi estimar a proporção de *loci* que estão em equilíbrio de Hardy-Weinberg, através da abordagem sequencial Bayesiana, para caracterização da variabilidade genética de uma população do peixe dourado (*Salminus brasiliensis*), de uma importante estação de criação ambiental (Volta Grande, MG), visando avaliar os impactos do setor hidrelétrico sobre peixes migradores de água doce, no Rio Grande, MG, Brasil.

Materiais e Métodos

A estatística do Equilibro de Hardy-Weinberg (EHW)

Para verificar o EHW, inicialmente realiza-se a análise dos perfis genotípicos em cada *locus* de DNA microssatélite, em que calcula-se a proporção dos genótipos observados, dada por:

$$p_{ij} = \frac{n_{ij}}{n} \tag{1}$$

A proporção de alelos:

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

em que, n_{ij} é a contagem observada em cada categoria $i, j = 1, ..., k$, $j \geq i$, e n a contagem total.

E as proporções esperadas dos genótipos calculadas com base nas proporções alélicas p_1 e p_2, que representam a proporção dos alelos A e a, respectivamente, na população, no caso de um gene com dois alelos. Desse modo, as proporções genotípicas esperadas são estimadas a partir da equação de Hardy-Weinberg, que é dada por uma expansão quadrática das proporções alélicas (HARTL; CLARK, 2010):

$$(p_1 + p_2)^2 = 1 \Rightarrow p_1^2 + 2p_1p_2 + p_2^2 = 1, \tag{3}$$

onde p_1^2 representa a proporção de homozigotos AA, $2p_1p_2$ a proporção de heterozigotos Aa e p_2^2 a proporção de homozigotos aa.

No caso de um sistema de um gene com k alelos, $k \geq 2$, ou seja, quando há uma diversidade genética, sendo um fenômeno muito comum observado quando há uma variedade de alelos. A distribuição de probabilidade associada tanto aos alelos quanto aos genótipos é a multinomial, e no caso dos genótipos esta possui $\frac{k(k+1)}{2}$ classes/categorias. Pois para calcular o número de classes da distribuição multinomial, considera-se a combinação da quantidade de alelos, tomados 2 a 2, somados a quantidade de alelos presentes, considerando que as espécies de maneira geral possuem um número diploide de cromossomos. Ou seja,

$$C_{k,2} + k = \frac{k!}{(k-2)!2!} + k = \frac{k(k-1)(k-2)!}{(k-2)!2!} + k = \frac{k(k-1) + 2k}{2} = \frac{k^2 + k}{2} = \frac{k(k+1)}{2}. \tag{4}$$

A equação de Hardy-Weinberg para estimar as proporções genotípicas esperadas quando existe uma diversidade genética, ou seja, quando a população possui mais de dois alelos ($k \geq 2$), é dada generalizando a expansão $(p_1 + p_2 + ... + p_k)^2 = 1$ em que p_n, com $n = 1, ..., k$, é a proporção dos alelos, que resulta em:

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

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

Considerando que $p_ip_j = p_jp_i$, portanto, $p_{ij} = 2p_ip_j$, $i, j = 1, ..., k$, se $i \neq j$ e $p_{ij} = p_i^2$, se $i = j$, em que p_{ij} são as proporções dos genótipos A_iA_j e p_i as proporções dos alelos, com $j \geq i$. Ou seja:

$$\begin{cases} p_{ij} = p_i^2, & \text{se } i = j \\ p_{ij} = 2p_ip_j, & \text{se } i \neq j \end{cases} \tag{5}$$

Pode-se escrever também essa relação de forma matricial, onde na diagonal principal tem-se as proporções esperadas dos genótipos homozigotos, e o restante são as proporções esperadas dos genótipos heterozigotos:

$$\begin{bmatrix} p_1^2 & 2p_1p_2 & \cdots & 2p_1p_j \\ 0 & p_2^2 & \cdots & 2p_2p_j \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_i^2 \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1j} \\ 0 & p_{22} & \cdots & p_{2j} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{ij} \end{bmatrix} \tag{6}$$

Realiza-se o teste de qui-quadrado para verificar se cada *locus* está ou não em equilíbrio de Hardy-Weinberg.

O teste qui-quadrado é utilizado para verificar se a frequência de um determinado evento observado em uma amostra difere significativamente da frequência esperada desse evento. Essa

quantificação é feita através da estatística de qui-quadrado, definida como (BUSSAB; MORETTIN, 2017):

$$\chi^2{}_{calculado} = \frac{\sum_{i=1}^{n} (observado_i - esperado_i)^2}{esperado_i}, \tag{7}$$

onde observado e esperado se referem as frequências observadas e esperadas em cada classe genotípica.

Após calcular o valor de $\chi^2{}_{calculado}$ (Equação **??**), este é comparado com o valor $\chi^2{}_{tabelado}$ para o número de graus de liberdade (g.l.) apropriado e nível de significância (α) desejado. Se $\chi^2{}_{calculado} \geq \chi^2{}_{tabelado}$ rejeita-se H_0, caso contrário, não se rejeita H_0. Os graus de liberdade (g.l.) são calculados por: número de classes dos dados - número de parâmetros estimados - 1 (HARTL; CLARK, 2010).

Desse modo, testa-se se o número de indivíduos em cada classe genotípica corresponde ao esperado sob a hipótese da população estar em equilíbrio de Hardy-Weinberg, ao nível de 5% de probabilidade. Cujas hipóteses são:

$$\begin{cases} H_0 : \text{A população está EHW.} \\ H_1 : \text{A população não está em EHW.} \end{cases}$$

Critério de parada na abordagem sequencial Bayesiana para a distribuição binomial

Para estimar o parâmetro p da distribuição binomial, referente à proporção, considerando o tamanho da amostra como uma variável aleatória na estimação sequencial Bayesiana, extrai-se elementos de uma amostra um por vez, e depois que cada um desses é observado, a partir do critério de parada toma-se a decisão de interromper a amostragem e estimar o parâmetro de interesse ou continuar.

Alguns conceitos estão envolvidos no procedimento de estimação sequencial Bayesiana para obtenção do critério de parada. Como função de perda, risco de Bayes, risco imediato, risco esperado e custo. Os quais são dados a seguir.

A cada decisão d e a cada possível valor do parâmetro p pode-se associar uma perda, que assume valores positivos, além de um custo $C(n)$, que indica o custo de tomar n observações.

A função de perda é definida como: $L(p, d)$. Considerou-se uma função de perda quadrática para obtenção do critério de parada, que segundo Ali (2015), é a função de perda mais utilizada em problemas de estimação. Ela é definida como:

$$L(\hat{p}, p) = (\hat{p} - p)^2. \tag{8}$$

De acordo com Berger (1985), o risco de uma regra de decisão, denotado por $R(p, d)$, é a perda esperada *a posteriori*, isto é:

$$R(p, d) = E_{posteriori}[L(p, d)]. \tag{9}$$

O risco de Bayes de um procedimento sequencial d é definido por:

$$r(\pi, d) = E^{\pi}[R(p, d)], \tag{10}$$

isto é, o risco esperado associado ao procedimento de estimação do parâmetro p dado *a priori* π, depois de n observações.

Logo, o estimador de Bayes de p com respeito à função perda é aquele com menor risco de Bayes. No caso da função de perda quadrática, o estimador de Bayes para o parâmetro p será a média de sua distribuição atualizada, ou seja, a média da sua distribuição *a posteriori* (BERGER, 1985).

Entre os métodos para desenvolver critérios de parada para o procedimento de estimação sequencial Bayesiana propostos na literatura, o método "*one-step look ahead*", pode ser considerado um dos mais úteis. A partir desse método, tem-se que o risco de Bayes de tomar uma decisão imediata é $r_0(\pi^n, n) = \underset{a \in A}{inf}\ r_0(\pi^n, a, n)$, onde A é o conjunto de ações disponíveis e $r_0(\pi^n, a, n) = E^{\pi^n}[L(p, a, n)]$ é a perda *a posteriori* esperada da ação a em n (BERGER, 1985).

Pratt et al. (1964), demonstraram que o menor risco de Bayes *a posteriori* é a variância da distribuição *a posteriori*, denotada por $var_{post}(n)$, isto é

$$r_0(\pi^n, n) = var_{post}(n). \tag{11}$$

Desse modo, o risco de Bayes *a posteriori* esperado, quando uma outra observação é feita, é a esperança desta variância, ou seja (PHAM-GIA, 1998):

$$r^1(\pi^n, n) = E[var_{post}(n)]. \tag{12}$$

Nesse sentido, para determinar o critério de parada, é necessário calcular o risco imediato, $r_0(\pi^n, n)$ acrescido do custo de n observações, e o risco esperado, $r^1(\pi^n, n)$ com o acréscimo no custo de mais uma observação (BERGER, 1985).

O valor atribuído ao custo deve possuir uma ordem de grandeza semelhante à ordem de grandeza da função de perda, o que irá garantir que a função de risco não seja exclusivamente dominada pelo custo. Ao considerar a função de perda quadrática, $L = (p - \hat{p})^2$, como a perda é o quadrado da diferença entre os valores das proporções que estão entre 0 e 1, os resultados são sempre próximos de zero e o custo também deverá ser próximo de zero (BACH, 2015).

Assim, o procedimento consiste em comparar $r_0(\pi^n, n)$ com $r^1(\pi^n, n)$, depois de avaliar a n-ésima observação. Se $r_0(\pi^n, n) > r^1(\pi^n, n)$, a amostragem continua; se $r_0(\pi^n, n) \leq r^1(\pi^n, n)$, a amostragem para.

A distribuição binomial é uma distribuição discreta de probabilidade e é definida sendo X o número de sucessos obtidos na realização de n ensaios de Bernoulli independentes. Apenas dois resultados são possíveis em cada repetição: sucesso ou fracasso. Assim, X tem distribuição binomial com parâmetros n e p, em que p é a probabilidade de sucesso em cada ensaio, se sua função de probabilidade for dada por (CASELLA; BERGER, 2002):

$$P(X = x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{13}$$

em que $x = 0, 1, \ldots, n$, logo quando a variável aleatória X tiver distribuição binomial com parâmetros n e p, escreve-se $X \sim b(n, p)$.

A média e a variância de uma variável aleatória binomial, com parâmetros n e p são dadas, respectivamente, por:

$$E(X) = np, \quad \text{Var}(X) = np(1-p). \tag{14}$$

A distribuição *a priori* conjugada da distribuição binomial é a distribuição beta, que é caracterizada por dois parâmetros a e b, dada por (CASELLA; BERGER, 2002):

$$\pi(p \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}, \tag{15}$$

em que $0 \leq p \leq 1$ e $\Gamma(t) = \int_0^\infty u^{t-1} e^{-u} du$ é a função gama.

A média e a variância da distribuição beta são dadas por:

$$E(p) = \frac{a}{a+b}, \quad \text{Var}(p) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{16}$$

A distribuição *a posteriori* também é uma beta, com parâmetros $(a + x, b + n - x)$, cuja média e variância são dadas por:

$$E(p \mid X) = \frac{a + x}{(a + x + b + n - x)}, \quad \text{Var}(p \mid X) = \frac{(a + x)(b + n - x)}{(a + b + n)^2(a + b + n + 1)}. \tag{17}$$

Com base no método "*one-step look ahead*", sabe-se que o critério de parada se resume em comparar os valores de risco imediato e esperado a cada elemento amostral, até que o risco imediato seja menor que o risco esperado de tomar uma decisão.

Desse modo, é necessário calcular para a distribuição binomial o risco imediato, que é dado pela variância *a posteriori* acrescida do custo de n observações, e o risco esperado, dado pela esperança da variância *a posteriori* e acréscimo no custo de mais uma observação, cujas expressões são, respectivamente:

$$r_0(\pi^n, n) = var_{post}(n) = \frac{(a + x)(b + n - x)}{(a + b + n)^2(a + b + n + 1)} + C(n). \tag{18}$$

$$r^1(\pi^n, n) = E[var_{posteriori}] + C(n + 1). \tag{19}$$

No entanto, os riscos são dados por uma relação de recorrência e resolver uma recorrência é encontrar uma fórmula fechada, e a função $E[var_{post}(n)]$ geralmente não está disponível na forma fechada, o que torna todo o cálculo altamente complexo. Pham-Gia (1998), apresentou a solução completa para o caso Bernoulli, e como a binomial compreende-se em n ensaios de Bernoulli, pode-se generalizar essa solução, assim como Brighenti et al. (2011) e Lima (2022) apresentaram.

Portanto, de acordo com Pham-Gia (1988), Brighenti et al. (2011) e Lima (2022), para encontrar a $E[var_{posteriori}]$, fazendo $W(n) = var_{posteriori}$, tem-se que:

$$W(n) = \frac{(a + x)(b + n - x)}{(a + b + n)^2(a + b + n + 1)}.$$

e $x = \sum_{i=1}^{n} X_i$, então a esperança da variância de tomar uma observação x_{n+1} será:

$$E[W(n)] = W(n + 1)P[sucesso] + W(n + 1)P[fracasso].$$

$$E[W(n)] = \frac{(a + x + 1)(b + n + 1 - x - 1)}{(a + b + n + 1)^2(a + b + n + 2)}p + \frac{(a + x + 0)(b + n + 1 - x - 0)}{(a + b + n + 1)^2(a + b + n + 2)}(1 - p). \tag{20}$$

Como beta é *a priori* do parâmetro p da Bernoulli, então, a probabilidade de sucesso $p = E(p) =$ média da beta *a posteriori* com n observações, ou seja, $a' = (a + x)$ e $b' = (b + n - x)$.

$$p = \frac{(a + x)}{(a + x) + (b + n - x)} = \frac{a'}{a' + b'} = \frac{(a + x)}{(a + b + n)}.$$

Então $1 - p = 1 - \left(\frac{a + x}{a + b + n}\right) = \frac{a + b + n - a - x}{a + b + n} = \frac{b + n - x}{a + b + n}$.

Substituindo p e $1 - p$ em (**??**), tem-se:

$$E[W(n)] = \frac{(a + x + 1)(b + n - x)}{(a + b + n + 1)^2(a + b + n + 2)} \frac{(a + x)}{(a + b + n)} +$$

$$+ \frac{(a + x)(b + n - x + 1)}{(a + b + n + 1)^2(a + b + n + 2)} \frac{(b + n - x)}{(a + b + n)}.$$

$$E[W(n)] = \frac{(a+x)(b+n-x)}{(a+b+n)(a+b+n+1)} \left[\frac{(a+x+1)+(b+n-x+1)}{(a+b+n+1)(a+b+n+2)}\right] \frac{(a+b+n)}{(a+b+n)}.$$

$$E[W(n)] = \frac{(a+x)(b+n-x)}{(a+b+n)(a+b+n+1)} \frac{(a+x+1+b+n-x+1)}{(a+b+n+1)(a+b+n+2)} \frac{(a+b+n)}{(a+b+n)}.$$

Substituindo o valor correspondente por $W(n)$, tem-se:

$$E[W(n)] = W(n)\frac{(a+b+n+2)}{(a+b+n+1)} \frac{(a+b+n)}{(a+b+n+2)}.$$

$$E[W(n)] = W(n)\frac{(a+b+n)}{(a+b+n+1)}.$$

$$E[var_{posteriori}] = var_{posteriori}\frac{(a+b+n)}{(a+b+n+1)}.$$

Então, o risco esperado, será dado por:

$$r^1(\pi^n, n) = E[W(n)] + C(n+1).$$

$$r^1(\pi^n, n) = \left(\frac{a+b+n}{a+b+n+1}\right) \frac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)} + C(n+1). \tag{21}$$

Após interromper a amostragem utiliza-se como estimador Bayesiano da proporção, considerando uma função de perda quadrática, a média da distribuição beta *a posteriori* com parâmetros $(a+x, b+n-x)$, dada por:

$$\mu_{post} = \hat{p}_{bayesiano} = E(p|X) = \frac{a+x}{(a+x+b+n-x)}. \tag{22}$$

Aplicação em dados reais

Verificou-se o equilibro de Hardy-Weinberg em um conjunto de dados de *locus* de DNA microssatélite de 56 peixes dourado (*Salminus brasiliensis*) (Characiformes: Bryconidae). *Salminus brasiliensis* (Cuvier, 1816), é um dos peixes mais emblemáticos da região Neotropical, uma espécie indicativa adequada para a conservação de ecossistemas de água doce e um recurso potencial para o desenvolvimento da aquicultura na América do Sul, ocorrendo em cinco países.

Assim, inicialmente realizou-se a análise dos perfis genotípicos em cada *locus* de DNA microssatélite, em 56 peixes, para determinação da proporção dos alelos e genótipos observados.

Calculou-se a proporção dos genótipos observados, conforme a Equação (**??**), a proporção de alelos (**??**). E as proporções esperadas dos genótipos calculadas conforme a equação de Hardy-Weinberg, dada em (**??**).

Em seguida, realizou-se o teste de qui-quadrado para verificar se cada *locus* está ou não em equilíbrio de Hardy-Weinberg.

Ao verificar se o *locus* está em EHW, há duas respostas possíveis: sim ou não. Logo, a distribuição de probabilidade associada é a binomial. Assim, em seguida, utilizou-se a abordagem sequencial Bayesiana para a distribuição binomial para estimar a proporção de *loci* que estão em EHW.

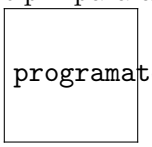
Para isso, utilizou-se uma *priori* conjugada beta, cujos hiper parâmetros foram calculados com base nos dados do histórico de análises de variabilidade genética, realizadas anteriormente, como no estudo de (CARMO et al., 2015).

Utilizou-se o programa em Delphi para realizar a estimação sequencial Bayesiana da proporção de *loci* em EHW, utilizando a distribuição binomial, para facilitar o processo, mas também poderia ter utilizado um script no *software* R (R CORE TEAM, 2023). Por fim, realizou-se um pequeno exemplo para mostrar o passo a passo do método.

Resultados e Discussão

Na Figura **??** é apresentada a interface do programa em Delphi utilizado. Neste programa, definiu-se os parâmetros iniciais, e a cada *locus* observado, marcou-se presente para quando estava em EHW, e ausente caso contrário.

Figura 1: Tela do programa em Delphi para a estimação sequencial Bayesiana.



Fonte: Autores.

Utilizou-se a informação *a priori* de um estudo anterior de Carmo et al. (2015), no qual totalizou que 50 % dos *loci* avaliados estavam em equilíbrio de Hardy-Weinberg, assim introduziu-se esse valor de proporção, com uma variação média, que representa 10,20 %. O programa calculou os valores dos hiper parâmetros da *priori* com base nessas informações, resultando em $a = 0,7255$ e $b = 0,7255$. Além disso, considerou-se um custo de 10^{-5}.

Portanto, utilizando o critério de parada de comparação dos riscos, após a avaliação do equilíbrio em 28 *loci*, o processo foi interrompido, e considerando uma função de perda quadrática, a estimação é dada pela média da distribuição beta *a posteriori*, o que resultou em uma estimativa de 50,00 %.

Na Tabela **??** é apresentado a resposta para cada um dos *loci* avaliados, com a respectiva quantidade de alelos encontrada em cada *locus*.

Tabela 1: Verificação do EHW.

	Loci	**Número de alelos**	**EHW**		***Loci***	**Número de alelos**	**EHW**
1	Sbra01	9	sim	15	Sbra18	2	sim
2	Sbra03	5	não	16	Sbra19	6	não
3	Sbra04	8	não	17	Sbra20	4	não
4	Sbra05	9	sim	18	Sbra21	11	sim
5	Sbra06	10	não	19	Sbra22	7	não
6	Sbra07	8	sim	*20*	Sbra23	6	não
7	Sbra08	7	não	21	Sbra24	8	sim
8	Sbra09	6	sim	22	Sbra25	8	não
9	Sbra10	9	sim	23	Sbra26	5	não
10	Sbra11	11	sim	24	Sbra27	9	sim
11	Sbra12	7	não	25	Sbra28	6	sim
12	Sbra14	8	não	26	Sbra29	6	não
13	Sbra15	6	sim	27	Sbra30	12	não
14	Sbra16	9	sim	28	Sbra31	2	sim

Fonte: Dos autores (2023).

Na Tabela **??** estão as estatísticas descritivas dos 28 *loci* avaliados.

Tabela 2: Estatísticas descritivas.

	Média	**Variância**	**Desvio padrão**
Número de alelos	7,2857	5,98942	2,4473
***Loci* em EHW**	0,5000	0,2593	0,5092

Fonte: Autores.

O relatório que o programa em Delphi fornece ao interromper o procedimento é apresentado na Figura **??**.

Figura 2: Relatório do programa em Delphi para Análise Sequencial Bayesiana.

Fonte: Autores.

Sendo assim, tem-se que a estimativa da proporção de *loci* em equilíbrio de Hardy-Weinberg, de uma população do peixe dourado, foi 50,00 %. Avaliando apenas 28 *loci*.

Pode-se observar na Figura **??** que o relatório apresenta além da estimativa da proporção, informações importantes, como o tamanho amostral, assim como os hiper parâmetros da distribuição *a priori* e os parâmetros da distribuição *a posteriori*.

Realizou-se um exemplo do passo a passo do método, que é realizado no programa em Delphi, para melhor compreensão.

Exemplo:

- **Passo 1** → Determinar os valores:
  1. dos hiper parâmetros (a, b) para estabelecimento da distribuição *a priori*,
  2. tamanho da amostra(n), ou seja o número da observação da sequência,
  3. soma dos ensaios de Bernoulli (x), que é a quantidade de sucessos observados,
  4. e o custo (C);

  - Os valores de n e x devem ser atualizados a cada amostragem;

- **Passo 2** → Cálculo dos parâmetros da distribuição *a posteriori*:

$$a' = a + x, \quad b' = b + n - x.$$

- **Passo 3** → Cálculo dos riscos imediato e esperado:

$$r_0 = \frac{(a+x)\,(b+n-x)}{(a+b+n)^2(a+b+n+1)} + C(n),$$
$$r^1 = \frac{(a+b+n)}{(a+b+n+1)} r_0 + C(n+1).$$

- **Passo 4** → Critério de parada:

- **Continue**: enquanto $r^1 < r_0$, atualizar n (realizar uma nova observação) e x (somar 1 quando obter sucesso ou 0 quando fracasso).

- **Pare**: quando $r^1 > r_0$, estimar o parâmetro proporção fazendo o cálculo da média da distribuição *a posteriori* beta, dada por:

$$\mu_{post} = \hat{p}_{bayesiano} = E(p|X) = \frac{a+x}{a+b+n}.$$

Assim, aplicando aos dados de *loci* de DNA, considerou-se o custo por observação igual a 10^{-5}, hiperparâmetros $a = 0,7255$ e $b = 0,7255$, e foram observados sequencialmente os *locus* de DNA um a um, verificando-se a presença (1) ou ausência (0) do equilíbrio de Hardy-Weinberg, obtendo-se o seguinte resultado:

- *locus 1* = está em equilíbrio de Hardy-Weinberg

$$x = 1, \quad a' = 1,7255, b' = 0,7255.$$
$$r_0 = \frac{(0,7255+1)(0,7255+1-1)}{(0,7255+0,7255+1)^2(0,7255+0,7255+1+1)} + 0,00001(1) = 0,0604,$$
$$r^1 = \frac{(0,7255+0,7255+1)}{(0,7255+0,7255+1+1)}0,0604 + 0,00001(2) = 0,0429.$$

Decisão: Como $r_0 > r^1$ continua-se a amostragem

- **locus 2** = não está em equilíbrio de Hardy-Weinberg

$$x = 1, \quad a' = 1,7255, \quad b' = 1,7255$$
$$r_0 = \frac{(0,7255+1)(0,7255+2-1)}{(0,7255+0,7255+2)^2(0,7255+0,7255+2+1)} + 0,00001(2) = 0,0562,$$
$$r^1 = \frac{(0,7255+0,7255+2)}{(0,7255+0,7255+2+1)}0,0604 + 0,00001(3) = 0,0436.$$

Decisão: Como $r_0 > r_l$ continua-se a amostragem.

E realiza-se esses passos sucessivamente, até que o $r_0 < r^1$. Portanto, nesse caso amostragem foi interrompida com $n = 28$ e a proporção estimada, a partir da média *a posteriori* da distribuição beta em 0,5000 .

Conclusão

Este trabalho demonstra que é possível aplicar a abordagem sequencial Bayesiana para estimar a proporção de *loci* em equilíbrio de Hardy-Weinberg, de marcadores de DNA hipervariáveis como microssatélites. Mas também essa abordagem pode ser aplicada em vários em outros procedimentos de interesse.

Agradecimentos

Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior - CAPES, pelo apoio.
CEMIG pelo projeto envolvido com o conjunto de dados utilizados neste trabalho.

Referências

ALI, S. Mixture of the inverse Rayleigh distribution: Properties and estimation in a Bayesian framework. *Applied Mathematical Modelling*, v.39, n.2, p.515-530, 2015.

BACH, D. R. A cost minimisation and bayesian inference model predicts startle reflex modulation across species. *Journal of Theoretical Biology*, v.370, p.53–60, 2015.

BERGER, J. O. *Statistical decision theory and Bayesian analysis*. 2. ed. New York: Springer Science & Business Media, 1985. 617p.

BRIGHENTI, C. R. G.; RESENDE, M.; BRIGHENTI, D. M. Estimação sequencial bayesiana aplicada à proporção de infestação de psilídeos em alecrim do campo. *Revista Brasileira de Biometria*, v.29, n.2, p.342–354, 2011.

BUSSAB, W. O.; MORETTIN, P. A. *Estatística Básica.* 9. ed. São Paulo: Editora Saraiva, 2017. 568p.

CARMO, F. M. d. S.; POLO, É. M.; SILVA, M. A. d.; YAZBECK, G. M..Optimization of heterologous microsatellites in Piracanjuba. *Pesquisa Agropecuária Brasileira*, v.50, p. 1236-1239, 2015.

CASELLA, G.; BERGER, R. L. *Statiscal Inference.* 2. ed. df: Duxbury Press, 2002. 686p.

CHAGAS, K. P. T.; SOUSA, R. F.; FAJARDO, C. G.; VIEIRA, F. A. Seleção de marcadores ISSR e diversidade genética em uma população de Elaeis guineensis. *Revista Brasileira de Ciências Agrárias*, v.10, n.1, p.147-152, 2015.

HARTL, D. L.; CLARK, A. G. *Princípios de Genética de Populações.* 4. ed. Porto Alegre: Artmed. 2010. 659p.

LIMA. I. S. *Estatística sequencial bayesiana dos parâmetros da distribuição multinomial.* Dissertação de mestrado. Universidade Federal de Lavras, Lavras, 2022.

PHAM-GIA, T. Distribution of the stopping time in Bayesian sequential sampling. *Australian & New Zealand Journal of Statistics*, v.40, n.2, p. 221–227, 1998.

PRATT, J. W.; RAIFFA, H.; SCHLAIFER, R. The foundations of decision under uncertainty: An elementary exposition. *Journal of the American statistical association*, v.59, n.306, p.353–375, 1964.

R CORE TEAM. *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2023. ISBN 3-900051-07-0, URL http://www.R-project.org/.

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*, v.41, n.5, p.834–840, 2011.

SCHILLING, E. G.; NEUBAUER, D. V.*Acceptance sampling in quality control.* London: Crc Press, 2017. 882p.