
**ISABELA DA SILVA LIMA**

# ESTATÍSTICA SEQUENCIAL BAYESIANA DOS PARÂMETROS DA DISTRIBUIÇÃO MULTINOMIAL

**LAVRAS – MG**

**2022**

**ISABELA DA SILVA LIMA**

# ESTATÍSTICA SEQUENCIAL BAYESIANA DOS PARÂMETROS DA DISTRIBUIÇÃO MULTINOMIAL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Prof(a). Dr(a). Carla Regina Guimarães Brighenti
Orientadora

**LAVRAS – MG**
**2022**

**Ficha catalográfica elaborada pela Coordenadoria de Processos Técnicos da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).**

Lima, Isabela da Silva

Estatística sequencial bayesiana dos parâmetros da distribuição multinomial / Isabela da Silva Lima. – Lavras : UFLA, 2022.

117 p. : il.

Dissertação (mestrado acadêmico) –Universidade Federal de Lavras, 2022.

Orientadora: Prof(a). Dr(a). Carla Regina Guimarães Brighenti.

Bibliografia.

1. Estimação Sequencial Bayesiana. 2. Critério de Parada. 3. Dirichlet. I. Brighenti, Carla Regina Guimarães. II. Título.

**ISABELA DA SILVA LIMA**

# ESTATÍSTICA SEQUENCIAL BAYESIANA DOS PARÂMETROS DA DISTRIBUIÇÃO MULTINOMIAL

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 09 de Fevereiro de 2022.

| | |
|---|---|
| Prof. Dr. Deive Ciro de Oliveira | UNIFAL |
| Prof. Dr. Peter de Matos Campos | UFSJ |
| Profa. Dra. Raquel Maria de Oliveira Pires | UFLA |

Prof(a). Dr(a). Carla Regina Guimarães Brighenti
Orientadora

**LAVRAS – MG**
**2022**

*Dedico este trabalho aos meus pais, meus maiores exemplos de vida e amor.*

## AGRADECIMENTOS

Agradeço primeiramente a Deus, pela força diária, por me permitir realizar tantos sonhos e por estar sempre presente em minha vida me guiando e iluminando meus caminhos. A São José de Cupertino, por todas as bençãos e glórias durante este percurso.

Aos meus pais, Adriana e Mérito, pelo amor incondicional, por me encorajarem a conquistar os meus sonhos e pelo incentivo que serviram de alicerce para as minhas realizações. Sou eternamente grata por todo o apoio.

Aos meus avós, padrinhos, tios, primos e todos meus familiares pelo carinho e incentivo.

Ao meu namorado Gabriel, pelo amor, companheirismo, compreensão e imenso apoio durante todos estes anos de estudo.

À minha orientadora, professora Dra. Carla Regina Guimarães Brighenti, pelos inestimáveis aprendizados transmitidos, por todo incentivo, dedicação, conselhos e confiança depositada em mim. Sou grata por todos os ensinamentos, por estar comigo desde a graduação e por ter se tornado uma grande amiga, contribuindo para o meu crescimento pessoal e profissional.

Aos professores Dr. Deive Ciro de Oliveira, Dr. Peter de Matos Campos, Dra. Raquel Maria de Oliveira Pires e Dr. Tales Jesus Fernandes por aceitaram o convite de fazer parte da banca examinadora.

Agradeço também aos professores do departamento de Estatística e Experimentação Agropecuária que contribuíram na minha formação e a todos professores que fizeram parte dessa jornada.

Aos meus amigos que o mestrado me proporcionou, em especial à Edilene e Lúcia Helena, por todos os momentos compartilhados e pela ajuda mútua durante esses anos. A todos meus amigos de Entre Rios de Minas, pelo carinho e torcida durante toda essa caminhada.

Ao Laboratório de Análise de Sementes do Departamento de Agricultura da Universidade Federal de Lavras, em especial, à professora Dra. Raquel Maria de Oliveira Pires, Elizabeth, Ana Esther e Ana Flávia, pelos dados fornecidos.

Ao programa de pós-graduação em Estatística e Experimentação Agropecuária e à Universidade Federal de Lavras (UFLA), pela oportunidade.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pelo apoio financeiro concedido.

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

*“Algumas batalhas são vencidas com espadas e lanças, outras com papel e caneta.”*
*(Tywin Lannister)*

# RESUMO

A amostragem é uma etapa importante no processo de estimação de um parâmetro, que deve ter seu custo e tempo reduzido. Assim, o uso da amostragem sequencial, que possui o tamanho de amostra variável, avalia cada elemento por vez, permite que a decisão de parar a amostragem e estimar um parâmetro seja tomada antecipadamente, sem que todos os elementos sejam avaliados como previsto no processo de inferência clássico. Além disso, pode-se incorporar a teoria da decisão bayesiana à amostragem sequencial para realizar a estimação de parâmetros, pois essa permite incluir informações *a priori* sobre o parâmetro de interesse, o que auxilia na tomada de decisão e otimiza o procedimento. O grande desafio para realizar a estimação sequencial bayesiana reside na dificuldade em estabelecer critérios de parada. Devido à dificuldade inerente ao procedimento, a maioria dos trabalhos desenvolvidos nessa área são para a distribuição binomial, e existem poucos trabalhos para a distribuição multinomial. Desse modo, o objetivo deste trabalho é definir critérios de parada para o processo de estimação sequencial bayesiana dos parâmetros da distribuição multinomial. Para validar a metodologia proposta utilizou-se um conjunto de dados reais de contagem, do teste de raios X para controle de qualidade de lotes de sementes de milho. Assim, avaliou-se a influência de duas *prioris* no critério de parada, uma uniforme e outra conjugada, com hiperparâmetros baseados em informações de referência da literatura, além do custo por observação. Os resultados obtidos pela metodologia proposta foram comparados com a abordagem frequentista e bayesiana de estimação de parâmetros, concluindo que as estimativas sequencias bayesianas tiveram um bom desempenho, e com a vantagem da redução do tamanho amostral na maioria dos lotes avaliados. Apesar da amostra ser menor, a estimação sequencial bayesiana produziu estimadores equivalentes ou tão bons quanto os frequentistas e bayesianos.



**Palavras-chave:** Estimação. Distribuição de Dirichlet. Critério de parada. Equações de programação dinâmica. Teste de raios X. Sementes de milho.

# ABSTRACT

Sampling is an important step in the process of estimating a parameter, which should have reduced cost and time. Thus, the use of sequential sampling, which has a variable sample size, evaluates each element at a time and permits the decision about when to stop sampling and estimate a parameter to be taken in advance, without having all elements evaluated as expected in the classic inference approach. In addition, bayesian decision theory can be incorporated into sequential sampling to perform parameter estimation, as this permits the inclusion of information about the parameter of interest beforehand, which helps take decisions and optimize the procedure. The greatest challenge of carrying out sequential bayesian estimation lies in the difficulty in establishing stopping criteria. Due to the inherent difficulty of the procedure, most of the works developed in this area are on the binomial distribution, and there are few works on the multinomial distribution. Therefore, the objective of this work is to define stopping criteria for the sequential bayesian estimation process of the multinomial distribution parameters. To validate the proposed methodology, a set of real count data was used, from the X-rays test for quality control of corn seed lots. Thus, the influence of two prioris on the stopping criterion was evaluated, a uniform one and another conjugate, with hyperparameters based on reference information from the literature, besides the cost per observation. The results obtained by the proposed methodology were compared with the frequentist and bayesian approach of parameter estimation, concluding that the sequential bayesian estimation had a good performance, and with the advantage of sample size reduction in most of the evaluated lots. Despite the smaller sample, the sequential bayesian estimation produced estimators that were equivalent or as good as the frequentists and bayesians.



**Keywords:** Estimation. Dirichlet distribution. Stop criterion. Dynamic programming equations. X-rays test. Corn seeds.

## LISTA DE FIGURAS


Figura 3.1 – Plano de amostragem sequencial . . . . . 24
Figura 3.2 – Curva característica da operação . . . . . 25
Figura 3.3 – Exemplo simplex . . . . . 43
Figura 3.4 – $2-$simplex $\mathbb{R}^3$ . . . . . 45
Figura 3.5 – Distribuição de Dirichlet em um 2-simplex (triângulo equilátero) para diferentes valores de $\boldsymbol{a}$. . . . . 46
Figura 3.6 – Fluxograma do procedimento de estimação sequencial bayesiana . . . . . 60
Figura 3.7 – Tipos de danos em sementes de milho *(Zea mays L.)* . . . . . 64
Figura 4.1 – Imagem radiográfica das sementes de milho, da repetição 1, do lote 100 . . . . 66
Figura 4.2 – Imagem radiográfica das classificação das sementes de milho . . . . . 67
Figura 4.3 – Interface do aplicativo em *Delphi* . . . . . 71
Figura 5.1 – Densidades das distribuições *a priori* e *a posteriori* beta das sementes sem danos 88
Figura 5.2 – Densidades das distribuições *a priori* e *a posteriori* beta das sementes sem danos 90
Figura 5.3 – Simplex da *priori* (à esquerda) e simplex da *posteriori* (à direita) . . . . . 92
Figura 5.4 – Simplex da *priori* (à esquerda) e simplex da *posteriori* (à direita) . . . . . 94
Figura 5.5 – Simplex da *priori* (à esquerda) e simplex da *posteriori* (à direita) . . . . . 96
Figura 5.6 – Simplex da *priori* (à esquerda) e simplex da *posteriori* (à direita) . . . . . 99
Figura 5.7 – Relatório do programa em *Delphi* com os resultados . . . . . 102

## LISTA DE TABELAS


Tabela 5.1 – Estimativas frequentistas das proporções de sementes sem danos e danificadas considerando a distribuição binomial . . . . . 83

Tabela 5.2 – Estimativas frequentistas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos . . . . . 84

Tabela 5.3 – Estimativas frequentistas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos 85

Tabela 5.4 – Hiperparâmetros e valores da média, variância da distribuição *a priori* uniforme 87

Tabela 5.5 – Estimativas bayesianas das proporções de sementes sem danos e danificadas considerando a distribuição binomial e *priori* uniforme . . . . . 87

Tabela 5.6 – Hiperparâmetros e valores da média, variância da distribuição *a priori* beta da literatura . . . . . 88

Tabela 5.7 – Estimativas bayesianas das proporções de sementes sem danos e danificadas considerando a distribuição binomial e *priori* da literatura . . . . . 89

Tabela 5.8 – Hiperparâmetros e valores da média, variância e covariância *a priori* com três classes de danos . . . . . 91

Tabela 5.9 – Estimativas bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos, utilizando *priori* uniforme . . . . . 91

Tabela 5.10 – Estimativas bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos, utilizando *priori* uniforme . . . . . 93

Tabela 5.11 – Hiperparâmetros e valores da média, variância e covariância *a priori* com três classes de danos . . . . . 95

Tabela 5.12 – Estimativas bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos, utilizando *priori* da literatura . . . . . 96

Tabela 5.13 – Hiperparâmetros e valores da média, variância e covariância *a priori* com três classes de danos . . . . . 97
Tabela 5.14 – Estimativas bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos, utilizando *priori* da literatura . . . . . 98
Tabela 5.15 – Valores dos hiperparâmetros, médias e variâncias das *prioris* . . . . . 100
Tabela 5.16 – Estimativas sequenciais bayesianas das proporções de sementes sem danos e danificadas considerando a distribuição binomial . . . . . 101
Tabela 5.17 – Estimativas sequenciais bayesianas considerando sementes sem danos e danificadas . . . . . 101
Tabela 5.18 – Estimativas sequenciais bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos, utilizando *priori* uniforme . . . . . 104
Tabela 5.19 – Estimativas sequenciais bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos, utilizando *priori* uniforme . . . . . 105
Tabela 5.20 – Estimativas sequenciais bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos, utilizando *priori* da literatura . . . . . 107
Tabela 5.21 – Estimativas sequenciais bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos, para *priori* da literatura . . . . . 108

## LISTA DE SÍMBOLOS

$\alpha$ Erro Tipo I

$\beta$ Erro Tipo II

$\boldsymbol{a+x}$ Vetor de parâmetros da distribuição de Dirichlet *a posteriori*

$\boldsymbol{a}$ Vetor de parâmetros da distribuição de Dirichlet *a priori*

$\pi(p)$ Distribuição *a priori*

$\pi(p/X)$ Distribuição *a posteriori*

$a+x, b+n-x$ Parâmetros da distribuição beta *a posteriori*

$A, B$ Limites do teste sequencial

$a, b$ Parâmetros da distribuição beta *a priori*

$a_0$ Soma dos componentes do vetor de parâmetros da distribuição de Dirichlet ($a_0 = a_1 + a_2 + \ldots + a_k$)

$B(x,m)$ Risco de bayes *a posteriori* esperado para a distribuição multinomial

$d$ Procedimento sequencial

$k$ Número de classes da distribuição multinomial

$L(p,d)$ Função de perda

$m, \boldsymbol{p}$ Parâmetros da distribuição multinomial, em que $\boldsymbol{p}$ é um vetor de parâmetros que indicam proporções e $m$ o tamanho amostral

$n, p$ Parâmetros da distribuição binomial, em que $p$ indica proporção e $n$ o tamanho amostral

$r(\pi,d)$ Risco de bayes

$R(p,d)$ Função de risco

$r^1(\pi^n, n)$ Risco de bayes *a priori* esperado para a distribuição binomial

$r_0(\pi^n, n)$ Risco de bayes *a posteriori* imediato para a distribuição binomial

$S(x, m)$ Risco de bayes *a posteriori* imediato para a distribuição multinomial

$x$ Números de sucessos observados

## SUMÁRIO


**1 INTRODUÇÃO** . . . . . 14
**2 OBJETIVOS** . . . . . 17
**2.1 Objetivo Geral** . . . . . 17
**2.2 Objetivos Específicos** . . . . . 17
**3 REFERENCIAL TEÓRICO** . . . . . 18
**3.1 Amostragem e Estimação** . . . . . 18
**3.2 Amostragem Sequencial** . . . . . 18
**3.2.1 Teste da Razão de Probabilidade Sequencial** . . . . . 21
**3.2.2 Estimação Sequencial** . . . . . 26
**3.3 Distribuição Binomial** . . . . . 27
**3.4 Distribuição Multinomial** . . . . . 28
**3.5 Estimação Frequentista** . . . . . 31
**3.5.1 Estimação Frequentista do parâmetro $p$ da distribuição Binomial** . . . . . 32
**3.5.2 Estimação Frequentista do parâmetro $p$ da distribuição Multinomial** . . . . . 33
**3.6 Estimação Bayesiana** . . . . . 35
**3.6.1 Estimação Bayesiana dos parâmetros da distribuição Binomial** . . . . . 38
**3.6.2 Estimação Bayesiana dos parâmetros da distribuição Multinomial** . . . . . 40
**3.6.2.1 Visualização da distribuição de Dirichlet** . . . . . 42
**3.7 Estimação Sequencial Bayesiana** . . . . . 47
**3.7.1 Estimação Sequencial Bayesiana dos parâmetros da distribuição Binomial** . . 49
**3.7.2 Estimação Sequencial Bayesiana dos parâmetros da distribuição Multinomial** 53
**3.8 Resumo das expressões** . . . . . 58
**3.9 Fluxograma do procedimento** . . . . . 60
**3.10 Aplicação a dados de contagem: Uso em teste de raios X** . . . . . 61
**3.10.1 Teste de raios X para sementes de milho** . . . . . 62
**4 METODOLOGIA** . . . . . 65

**4.1** **Critério de parada para estimação sequencial bayesiana dos parâmetros da distribuição Multinomial** . . . . . 65
**4.2** **Aplicação da teoria a dados de controle de qualidade de sementes de milho** . . 65
**4.3** **Estimação Frequentista** . . . . . 68
**4.4** **Estimação Bayesiana** . . . . . 68
**4.5** **Estimação Sequencial Bayesiana** . . . . . 70
**5** **RESULTADOS E DISCUSSÃO** . . . . . 74
**5.1** **Critério de parada para estimação sequencial bayesiana dos parâmetros da distribuição Multinomial** . . . . . 74
**5.2** **Estimação Frequentista** . . . . . 83
**5.2.1** **Estimação frequentista da proporção de sementes com danos e sem danos utilizando a distribuição Binomial** . . . . . 83
**5.2.2** **Estimação frequentista da proporção de sementes sob três classes utilizando a distribuição Multinomial** . . . . . 84
**5.3** **Estimação Bayesiana** . . . . . 86
**5.3.1** **Estimação bayesiana da proporção de sementes com danos e sem danos utilizando a distribuição Binomial** . . . . . 86
**5.3.2** **Estimação bayesiana da proporção de sementes sob três classes utilizando a distribuição Multinomial** . . . . . 90
**5.4** **Estimação Sequencial Bayesiana** . . . . . 99
**5.4.1** **Estimação sequencial bayesiana da proporção de sementes com danos e sem danos utilizando a distribuição Binomial** . . . . . 99
**5.4.2** **Estimação sequencial bayesiana da proporção de sementes sob três classes utilizando a distribuição Multinomial** . . . . . 103
**6** **CONCLUSÕES** . . . . . 109
**6.1** **Trabalhos futuros** . . . . . 110
**REFERÊNCIAS** . . . . . 111

# 1 INTRODUÇÃO

Amostragem é um procedimento estatístico base para qualquer análise de dados, já que consiste no processo de escolha da amostra representativa da população adequada para análise do todo. Ao reunir uma análise exploratória dos dados, modelos probabilísticos e amostragem, pode-se desenvolver a inferência estatística, uma área de grande relevância dentro da Estatística, que se resume em como tirar conclusões sobre parâmetros da população, com base no estudo de somente uma parte da mesma, ou seja, com base em uma amostra (MOOD; GRAYBILL; BOES, 1974).

Desta forma, as observações obtidas nas amostras podem fornecer estimativas para parâmetros populacionais, sendo a estimação de parâmetros um dos problemas básicos da inferência estatística. Dentre as estimativas, tem-se a do parâmetro $p$ sobre proporções, que é útil na identificação e mensuração de elementos portadores de uma certa característica. Como por exemplo, na mensuração de proporções de peças defeituosas e não defeituosa, de indivíduos pertencerem a uma determinada classe social, na proporção de sementes danificadas, entre outros. Desse modo, pode-se ainda observar que a estimação de proporções ambienta-se em questões de dois tipos: dicotômicas e politômicas (SILVA, 2012).

Problemas de caráter dicotômico, são aqueles que possuem duas opções de resposta, geralmente representadas pelo binômio sim/não, como por exemplo, ao mensurar a quantidade de peças defeituosas tem-se como resposta sim que corresponde à peça defeituosa, ou não, correspondente a não defeituosa. Para esse tipo de problema é suficiente utilizar a distribuição binomial para a análise dos dados, e esta possui teoria bem desenvolvida na literatura.

Já as questões politômicas são compostas de mais de duas categorias ou classes de resposta, como por exemplo, ao mensurar a raça de alunos, tem-se brancos, pardos, negros, etc. Assim, deve-se considerar a distribuição multinomial para a análise. No entanto, esta não conta com literatura tão acessível para melhor compreensão, pois é menos utilizada. Pode-se observar isto ao realizar uma simples pesquisa no *Google* Acadêmico, onde há um resultado de 1.340.000 artigos para a binomial e 430.000 para a multinomial.

A obtenção de amostras representativas da população, com possível redução dos custos, sempre foi o objetivo do processo de amostragem. Uma das formas de reduzir custos nesse pro-

cesso e consequentemente na inferência, é substituir amostras com tamanho fixo de elementos por um processo que possibilite o emprego de amostras com tamanho variável em função das observações realizadas. Isso acontece quando em cada unidade de tempo, o pesquisador realiza uma observação, e em seguida, decide parar ou continuar realizando uma nova observação. Este processo é conhecido como amostragem sequencial (WALD, 1947).

A amostragem sequencial foi desenvolvida por Wald em 1943, e na maioria dos casos proporciona um tamanho amostral menor do que seria adotado considerando uma amostra de tamanho fixo. Assim, obtendo conclusões válidas e reduzindo custos do processo (WALD, 1947).

A inferência bayesiana é uma alternativa importante em relação aos procedimentos clássicos de estimação, e esta abordagem vem sendo utilizada com sucesso em várias áreas da ciência. Ela permite a incorporação de uma informação *a priori* sobre o parâmetro que se quer estimar e toda sua teoria está baseada no teorema de Bayes (PAULINO; TURKMAN A. A.; SILVA, 2018).

A teoria de decisão bayesiana é uma forte aliada à amostragem sequencial para realizar a estimação de parâmetros, pois pode-se incluir informações relevantes ao plano de amostragem usando *prioris*, o que auxilia na tomada de decisão e otimiza o processo de estimação de parâmetros, reduzindo o tamanho amostral (BERGER, 1985).

No entanto, o grande problema para utilizar a estimação sequencial bayesiana reside na dificuldade em estabelecer critérios de parada. Pois necessita-se de uma matemática complexa pela natureza dinâmica do processo e a recursividade presente nos cálculos, já que as observações passadas irão influenciar nas decisões futuras. Além da caracterização de uma função custo por observação que não poderá dominar a regra de parada. Assim, precisa-se utilizar softwares estatísticos/matemáticos para auxilar durante o processo de tomada de decisão (FENOY, 2017).

Devido a esta dificuldade inerente ao procedimento, pode-se observar que a maioria dos trabalhos desenvolvidos nessa área são para variáveis que seguem distribuições discretas de probabilidade e no contexto univariado, mais precisamente, para a distribuição binomial, como pode ser visto em Armitage (1958), Plant e Wilson (1985), Karunamuni e Prasad (2003), entre outros. Existem poucos trabalhos realizados para distribuições multivariadas, como exemplo tem-se Jones (1976), que estabeleceu critérios de parada para a distribuição multinomial, mas considerando para isso apenas *prioris* uniformes.

O milho e a soja são as culturas de maior expressão econômica no Brasil, que apresentam desempenho crescente em produtividade, correspondendo, em conjunto, a mais de 88% da produção de grãos na safra 2020/2021 (CONAB, 2021). O aumento da produtividade agrícola está ligado não só ao desenvolvimento tecnológico nesse setor, mas também com a seleção de sementes de boa qualidade para o plantio.

Diante deste cenário, surge a necessidade de se analisar as sementes, e para isto, o teste de raios X é utilizado para avaliar a qualidade de sementes. Este teste verifica as imagens radiográficas geradas no processo, e torna-se possível a diferenciação entre sementes com boa formação e as que apresentem algum dano, seja físico, por insetos ou com variações de densidade (BRASIL, 2009).

O processo de verificação dessas imagens geralmente é realizado de forma visual, uma de cada vez, por um analista de sementes, o que é desgastante. Desta forma, o procedimento de estimação sequencial bayesiana pode ser aplicado neste contexto para estimar a proporção de sementes com diferentes tipos de danos. Portanto, estimar essas proporções utilizando um número menor de sementes requeridas por esse teste, de acordo com Brasil (2009), é vantajoso. Pois resulta em economia de recursos temporais e financeiros.

Neste sentido, este trabalho está dividido em seis capítulos. No primeiro capítulo há uma introdução ao tema estudado por meio de sua contextualização, motivação e justificativa. No capítulo dois é apresentado o objetivo geral e os específicos.

O capítulo três traz uma revisão acerca das áreas de estudo que culminam nesse trabalho. Neste capítulo é apresentado um detalhamento sobre amostragem sequencial, inferência bayesiana, distribuições multinomial e binomial, que é um caso particular da multinomial, estimação de parâmetros pelas abordagens frequentista, bayesiana e sequencial bayesiana.

No capítulo quatro é apresentada a metodologia do trabalho, descrevendo todos os passos necessários para a definição do critério de parada, e também sobre os dados advindos do teste de raios X para sementes de milho. No capítulo cinco são apresentados os resultados obtidos pela metodologia proposta, e estes foram comparados com a abordagem frequentista e bayesiana de estimação de parâmetros para discussão.

Por fim, no capítulo seis são apresentadas as conclusões sobre o trabalho, além de sugeridas possibilidades de trabalhos futuros, para a obtenção de resultados ainda melhores.

# 2 OBJETIVOS

## 2.1 Objetivo Geral

O presente trabalho tem como objetivo principal definir um critério de parada para o processo de estimação sequencial bayesiana dos parâmetros da distribuição multinomial para *priori* conjugada de Dirichlet.

## 2.2 Objetivos Específicos

- Detalhar a teoria de amostragem sequencial.
- Detalhar a teoria de estimação clássica para os parâmetros da distribuição multinomial e binomial.
- Detalhar a teoria de estimação bayesiana, considerando *priori* conjugada de Dirichlet, para os parâmetros da distribuição multinomial.
- Detalhar a teoria de estimação bayesiana, considerando *priori* conjugada Beta, para os parâmetros da distribuição binomial.
- Aplicar a teoria de estimação sequencial bayesiana à dados reais de contagem, como o do teste de raios X para controle de qualidade de lotes de sementes de milho, utilizando a distribuição multinomial, para verificar e validar a técnica.
- Avaliar a influência de duas *prioris* no critério de parada, uma uniforme e outra conjugada com hiperparâmetros baseados em informações de referência da literatura.
- Testar diferentes valores para o custo por observação.
- Comparar os resultados da metologia proposta com a abordagem frequentista e bayesiana de estimação de parâmetros.

# 3 REFERENCIAL TEÓRICO

Neste capítulo é apresentada a base teórica para a realização deste trabalho. Exibe os principais conceitos de amostragem sequencial, distribuição binomial, distribuição multinomial, inferência bayesiana, estimação de parâmetros pelas abordagens frequentista, bayesiana e sequencial bayesiana.

## 3.1 Amostragem e Estimação

Segundo Bussab e Morettin (2017), o modo de se obter uma amostra, é tão importante, e exitem muitos modos de se fazê-lo, que esses procedimentos constituem uma especialidade dentro da Estatística, conhecida como Amostragem. O processo de amostragem consiste na retirada de quantidades, mais conhecida como amostra, de um todo, população, que se deseja amostrar, de tal forma que esta amostra represente toda a população (MOOD; GRAYBILL; BOES, 1974).

Desse modo, a Amostragem está relacionada com a Inferência Estatística, no sentido em que para se realizar uma inferência depende dos processos de amostragem, uma vez que o objetivo da inferência é tirar conclusões sobre uma população, a partir do conhecimento de amostras (BUSSAB; MORETTIN, 2017).

Sendo assim, a estimação de parâmetros ou teste de hipóteses, que são os problemas básicos da inferência estatística, é subsidiado pelos processos de amostragem, pois estima-se parâmetros populacionais e testa hipóteses, a partir de amostras (CASELLA; BERGER, 2002).

Portanto, obter uma redução no tempo e custo no processo de amostragem e consequentemente na inferência, é de grande interesse, e é um dos objetivos de planos de amostragens sequenciais. Assim, sua adoção é uma alternativa viável à amostragem convencional para alcançar esses objetivos e otimizar esforços no procedimento (WALD, 1947).

## 3.2 Amostragem Sequencial

Segundo Schilling e Neubauer (2017), a técnica estatística sequencial foi criada na Segunda Guerra Mundial em processos de controle de qualidade, e os trabalhos de Dodge e Romig (1929)

e Thompson (1933) contribuíram para esta teoria estatística. Entretanto, cabe a Wald (1947) a mais importante contribuição aos estudos sequencias para reduzir a inspeção amostral necessária na indústria, e também a formalização desta técnica, em 1943, que ficou conhecida como Análise Sequencial.

A amostragem sequencial caracteriza-se por utilizar amostras de tamanho variável, dadas em função das observações realizadas, ao contrário de procedimentos convencionais, em que o tamanho amostral é fixado antes do experimento (WALD, 1947). Ela permite analisar os dados a medida que são obtidos, até que sejam suficientes para estimar os parâmetros, proporcionando, na maioria dos casos, reduções consideráveis do número total de observações que devem ser realizadas para possibilitar conclusões estatisticamente válidas (ZHUANG; BHATTACHARJEE, 2021).

Uma característica essencial de um procedimento sequencial é que o número de observações necessárias para encerrar o experimento é uma variável aleatória, pois depende do resultado das observações. Assim, pode-se calcular a esperança matemática desse número que indica o tamanho da amostra necessário, em média, para tomar uma decisão executando o teste diversas vezes em uma mesma população quando $p$ é o parâmetro verdadeiro (GOVINDARAJULU, 2004).

Lopes e Vieira (2014), apontaram a amostragem sequencial como uma possível alternativa para análises que requerem rapidez, precisão e redução de custos, ao utilizarem para atestar a pureza genética de lotes de sementes de milho. Assim, as principais vantagens deste método, é a redução do tempo e dos custos de um experimento, já que na maioria dos casos requer um tamanho amostral menor do que seria usado com a amostragem convencional.

Na literatura, pode-se observar que a amostragem sequencial quando aplicada obteve bons resultados, como por exemplo, Brack e Marshall (1990), compararam a amostragem sequencial à amostragem com tamanho fixo de amostras, e concluíram que, na primeira, ocorre uma redução no tamanho desta, entre 40 e 60%. Penteado, Oliveira e Iede (2008) testaram a técnica da amostragem sequencial para avaliação da eficiência do parasitismo de nematóide em adultos da vespa-da-madeira. Essa se mostrou como uma alternativa viável, pois resultou em redução do tamanho da amostra, dos custos da atividade e obteve precisão nas estimativas.

Ballaris et al. (2014), verificaram a eficiência de um plano de amostragem sequencial para sementes de feijão e soja, na detecção do patógeno *Sclerotinia sclerotiorum.*. Utilizaram a distri-

buição binomial para testar a presença ou ausência do patógeno, e concluíram que a amostragem sequencial foi eficiente para diminuir o tamanho amostral.

Zamba e Tsiamyrtzis (2021), utilizaram a abordagem sequencial para desenvolver uma estrutura de detecção, a ideia usou a soma das razões de verossimilhança para chegar em um critério de parada ideal. Os métodos foram aplicados com sucesso aos dados de incidência global COVID-19. Zhuang e Bhattacharjee (2021), desenvolveram um esquema de amostragem puramente sequencial que exigiu menos operações de amostragem sem perder a precisão da estimativa.

No entanto, também há algumas desvantagens em relação a amostragem sequencial, uma delas é que para utilizá-la deve existir a capacidade de amostrar em sequência. Além disso, a matemática necessária para analisar os dados para a amostragem sequencial é muito complexa. Pois existe uma relação de recorrência entre as observações e são estas que irão determinar o tamanho de amostra final. Assim, o procedimento pode requerer um auxílio computacional para simplificar o método (SCHILLING; NEUBAUER, 2017).

Um exemplo onde a amostragem sequencial apresentaria limitações a ser aplicada seria em pesquisas eleitorais, pois o processo precisa ser calculado e acompanhado a cada observação. Assim, nesse contexto seria inviável, porque o avaliador, ou ainda, entrevistador teria que ser único e deveria ter um equipamento em mãos. Ou ainda, o sistema teria que estar em nuvem, para que todos entrevistadores simultaneamente, ao realizar uma entrevista, o sistema registrasse e retornasse uma resposta imediata se seria necessário entrevistar mais um indivíduo. Desse modo, um único sistema é que acompanharia todo o procedimento, o que é praticamente impossível, pois iria ficar muito longo, já que seria necessário processar ao mesmo tempo.

Assim, o método sequencial é um pouco mais exigente ao ser conduzido, no sentido em que seria necessário o entrevistador possuir recursos computacionais no momento da pesquisa. Ou ainda, o entrevistador precisaria ter o conhecimento para avaliar quando parar o processo. Logo, a grande limitação do método sequencial, é a necessidade de ser processado imediatamente, impedindo por exemplo, que faça duas entrevistas simultâneas, pois requer do resultado de uma para posteriormente realizar outra.

O desenvolvimento de planos de amostragem sequencial são baseados no Teste Sequencial da Razão de Probabilidades de Wald (1947). São formuladas hipóteses e a decisão de aceitar,

rejeitar ou continuar amostrando são realizadas gradativamente, pois são baseadas no resultados acumulados de cada amostragem (BÁNYAI; BARABÁS, 2002).

Desse modo, ao utilizar o processo de amostragem sequencial é possível realizar inferência estatística com base em testes de hipóteses e estimação de parâmetros. Dentre os diferentes temas, aquele que tem recebido a maior atenção é o teste de hipóteses sequencial. No entanto, em algumas aplicações, a formulação de um problema como um teste de hipóteses não é adequado, assim, em algum desses casos, a estimação sequencial é mais apropriada (GOVINDARAJULU, 2004).

Portanto, o foco deste trabalho é na estimação de parâmetros utilizando uma amostragem sequencial, mas para isso é importante entender também a teoria dos testes de hipóteses.

### 3.2.1 Teste da Razão de Probabilidade Sequencial

O Teste da Razão de Probabilidade Sequencial (TRPS), desenvolvido por Wald em 1943, é uma generalização do teste da razão de máxima verossimilhança para análises sequenciais. São formuladas as hipóteses de interesses sobre os parâmetros, e o TRPS se baseia na razão de probabilidades para realizar a decisão de aceitar ou rejeitar a hipótese nula, ou ainda, de continuar a amostragem, após cada observação das amostras que são retiradas em sequência (BALLARIS et al., 2014).

Nos testes de hipóteses, ao tomar uma decisão de rejeitar ou não a hipótese nula, pode-se cometer dois tipos de erros, conhecidos como Erro Tipo I e Erro Tipo II. O Erro Tipo I é aquele cometido ao se rejeitar a hipótese nula, quando de fato ela era verdadeira, sendo este, controlado através do nível de significância do teste ($\alpha$). O Erro Tipo II é representado pela letra grega $\beta$ e corresponde ao erro cometido ao não rejeitar a hipótese nula, quando de fato ela era falsa (ANDRADE; OGLIARI, 2017).

As principais diferenças entre os testes de hipóteses convencionais e os sequenciais, é que nos convencionais fixam-se o tamanho amostral $n$ antes do experimento e o valor mínimo de significância $\alpha$, que é frequentemente adotado como 0,05 ou 0,01. Já o valor de $\beta$ não é especificado antecipadamente. No entanto, tem-se que ao fixar um tamanho de amostra, geralmente é impossí-

vel tornar ambos os tipos de probabilidades de erros arbitrariamente pequenos (BALLARIS et al., 2014).

Já no teste da razão de probabilidade sequencial os valores de erro tipo I ($\alpha$) e tipo II ($\beta$) são fixados antecipadamente, e o valor do tamanho amostral $n$ não é fixo, pois é uma função das próprias observações. Wald e Wolfowitz (1948), utilizando o Teorema de Neyman e Pearson demonstraram que os testes sequenciais de hipótese, além de permitirem o controle satisfatório das probabilidades de erro tipo I e tipo II, possuem a vantagem da redução do tempo de amostragem e de custos.

O teste sequencial da razão de probabilidades parte do pressuposto que se $X$ representa uma variável aleatória, então $f_X(x;p)$ refere-se à distribuição dessa variável. Ao descrever o teste da razão de probabilidade sequencial, para testar hipóteses sobre os parâmetros referentes a proporções, por exemplo, considerando o caso mais simples, isto é, aquele em que a proporção estudada pode assumir apenas um de dois valores possíveis $p_0$ e $p_1$, as hipóteses são (WALD, 1947):

$$\begin{cases} H_0: & p = p_0 \\ H_1: & p = p_1 \end{cases}.$$

Assim, quando tem-se a hipótese $H_0 : p = p_0$ como verdadeira, a distribuição de $X$ é dada por $f_X(x;p_0)$, e para $H_1 : p = p_1$ a distribuição de $X$ é dada por $f_X(x;p_1)$. Desse modo, a coleta de dados representa uma sequência de variáveis aleatórias $X_1, X_2, ..., X_n$ com probabilidade de obter uma amostra dada por (WALD, 1947):

$$P_{1n} = f_X(x_1;p_1) \cdot f_X(x_2;p_1) \cdot ... \cdot f_X(x_n;p_1), \text{quando } H_1 \text{ for verdadeira}$$

$$P_{0n} = f_X(x_1;p_0) \cdot f_X(x_2;p_0) \cdot ... \cdot f_X(x_n;p_0), \text{quando } H_0 \text{ for verdadeira}$$

A razão de probabilidades é definida como:

$$\frac{P_{1n}}{P_{0n}} = \prod_{i=1}^{n} \frac{f_X(x_i;p_1)}{f_X(x_i;p_0)}. \tag{3.1}$$

Para tomar uma decisão em cada etapa do experimento, este teste examina a razão das probabilidades dada por (3.1) e a compara com valores de limites representados por $A$ e $B$, estes são calculados de acordo com os valores preestabelecidos de erro tipo I $(\alpha)$ e erro tipo II $(\beta)$, pelas expressões (WALD, 1947):

$$A = \frac{1-\beta}{\alpha}; \qquad B = \frac{\beta}{1-\alpha}. \tag{3.2}$$

Desse modo, no teste da razão de probabilidade sequencial, $A$ e $B$ são encontrados, tal que $A > 1$ e $B < 1$, e em cada etapa do experimento, compara-se com a razão dada em (3.1) e chega-se a uma das três conclusões:

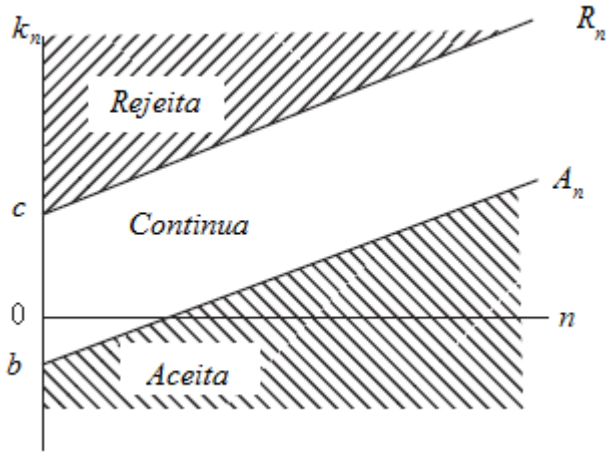
- Se $\frac{P_{1n}}{P_{0n}} \geq A$, o experimento termina com a rejeição de $H_0$ ao nível $\alpha$ de significância.
- Se $\frac{P_{1n}}{P_{0n}} \leq B$, o experimento termina com a não rejeição de $H_0$, com uma probabilidade $\beta$ de incorrer erro.
- $B < \frac{P_{1n}}{P_{0n}} < A$, continua-se o experimento e realiza-se uma nova observação.

Para fins computacionais, é interessante calcular o logaritmo das desigualdades acima, o que ajuda no cálculo da razão $P_{1n}/P_{0n}$, e alternativamente pode-se obter o plano de decisão sequencial por um procedimento gráfico, para a realização do teste sequencial da razão de probabilidade (BALLARIS et al., 2014).

O procedimento gráfico facilita a execução do teste sequencial da razão de probabilidade e consequentemente na tomada de decisão. Ele é definido em três regiões: a região de aceitação da hipótese nula, de rejeição da hipótese nula e de continuação do teste. Estas regiões são delimitadas por retas paralelas que são construídas de acordo com a distribuição da variável que está sendo levantada pela amostragem, pelos valores dos parâmetros que estão sendo testados e pelas probabilidades $\alpha$ e $\beta$ de erro (ESTEFANEL; BARBIN, 1979).

Assim, para o caso univariado, o procedimento gráfico para o plano de amostragem sequencial é organizado traçando as retas, onde o eixo das abscissas representa o número $n$ de observações realizadas, e o eixo das ordenadas representa a resposta da variável em questão, como por exemplo, no caso da distribuição binomial, pode-se ter o número de sucessos ou fracassos acumulados. Um exemplo do plano sequencial gráfico é dado pela Figura 3.1.

Figura 3.1 – Plano de amostragem sequencial




Fonte: Adaptado de Schilling e Neubauer (2017)

No plano de amostragem representado pela Figura 3.1, a região que indica a continuação da amostragem encontra-se entras as duas retas, a região de aceitação abaixo da reta $A_n$ e a região de rejeição acima da reta $R_n$.

Ao executar a amostragem os pontos $(n;k_n)$ são representados no gráfico para cada unidade amostral examinada. Se esse ponto estiver em $A_n$ ou abaixo a amostragem é terminada concluindo pela não rejeição de $H_0$. Se estiver em $R_n$ ou acima a amostragem é terminada concluindo pela rejeição de $H_0$. Enquanto $(n;k_n)$ estiver entre $A_n$ e $R_n$ novas unidades amostrais são examinadas.

Como o tamanho amostral no procedimento sequencial é uma variável aleatória, pode-se calcular a esperança matemática deste número. Desse modo, a expressão geral da esperança de $N$, que fornece o tamanho médio de amostra, segundo Wald (1947), é dada por:

$$E[N] \sim \frac{(1-\gamma)\ log\ B + \gamma\ log\ A}{E(Z)}, \tag{3.3}$$

onde $\gamma$ é a probabilidade que o teste da razão de probabilidade sequencial leve a aceitação de $H_1$. Assim, se $H = H_0$ então $\gamma = \alpha$, e se $H = H_1$, então $\gamma = 1-\beta$. E $Z$ é uma variável aleatória cuja distribuição é igual à distribuição comum das variáveis $z_i$, sendo $z_i = \dfrac{f_1(x_i)}{f_0(x_i)}$.

Ao considerar, por exemplo, uma distribuição binomial, tem-se que a expressão da esperança de $N$, de acordo com Wald (1947), é dada por:

$$E_p(N) \sim \frac{L_p \, log\frac{\beta}{1-\alpha} + (1-L_p) \, log \, \frac{1-\beta}{\alpha}}{p \, log\frac{p_1}{p_0} + (1-p) \, log \, \frac{1-p_1}{1-p_0}}, \tag{3.4}$$

onde $L_p$ é a probabilidade de aceitação e $p$ é a proporção de itens defeituosos em um lote, que são dados por:

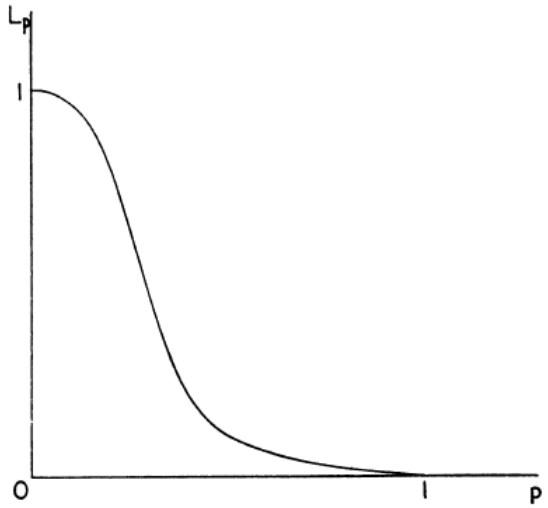
$$L_p = \frac{\left(\frac{1-\beta}{\alpha}\right)^h - 1}{\left(\frac{1-\beta}{\alpha}\right)^h - \left(\frac{\beta}{1-\alpha}\right)^h}, \tag{3.5}$$

$$p = \frac{1 - \left(\frac{1-p_1}{1-p_0}\right)^h}{\left(\frac{p_1}{p_0}\right)^h - \left(\frac{1-p_1}{1-p_0}\right)^h}, \tag{3.6}$$

em que $h$ é o número de itens avaliados.

A curva $L_p$ é chamada de curva característica de operação, em que $L_p$ será uma função de $p$ e pode ser plotada como mostrado na Figura 3.2:

Figura 3.2 – Curva característica da operação




Fonte: Wald (1947)

As curvas Características da Operação não são essenciais ao plano de amostragem sequencial, mas seu estudo permite avaliar o desempenho do mesmo. A curva Característica da Operação é a representação gráfica da Função Operatória Característica que fornece a probabilidade de terminar a amostragem não rejeitando $H_0$, ou seja, a probabilidade $L_p$ de não rejeitar $H_0$ para qualquer valor $p$ possível da proporção de uma certa característica (WALD, 1947).

Em alguns casos, a amostragem pode ser continuada indefinidamente, e assim gerar um tamanho de amostra muito grande. Para não acontecer este problema, Wald (1947) sugeriu um método de truncar o processo sequencial, colocando um limite superior ao tamanho da amostra, este limite geralmente é o tamanho de amostra dada pela amostragem convencional. Desse modo, esses esquemas são conhecidos como truncados ou ainda fechados.

Pode-se observar que Bottine (2015) desenvolveu o pacote SPRT no *software* R para realizar o Teste de Razão de Probabilidade Sequencial de Wald, entretanto apenas para variáveis com distribuição Normal, Bernoulli, exponencial ou Poisson. Através dele é possível construir o gráfico de aceitação sequencial, assim como estabelecer regiões com truncamento. Este pacote está disponível no CRAN, o endereço de pacotes para R, e pode ser facilmente instalado.

Além disso, pode-se construir planos de amostragem adequados para cada demanda e que atenda os critérios definidos pelo pesquisador, através de um programa de planilhas, assim como Schilling e Neubauer (2017), que desenvolveram um conjunto de modelos de planos de amostragem no Microsoft Excel®. Esses modelos permitem que os usuários programem e executem com mais eficiência e rapidez os planos de amostragem.

### 3.2.2 Estimação Sequencial

Segundo Berger (1985), mesmo sendo muito difícil, a maneira mais viável para realizar a estimação de parâmetros a partir de uma amostragem sequencial, é através da abordagem bayesiana, utilizando a teoria da decisão, pois ela permite a quantificação da incerteza como um auxílio na tomada de decisão, a partir de *prioris*.

Sendo assim, no procedimento de estimação a partir de uma amostragem sequencial, utilizando a abordagem bayesiana, uma função de perda está envolvida ao tomar uma decisão terminal.

Além disso, extrair uma observação de uma população envolve custos, assim, deve-se adicionar à função de perda o custo da experimentação, ou seja, a função de custo $C(n)$, que indica o custo de tomar $n$ observações, em que $n$ é uma variável aleatória (GOVINDARAJULU, 2004).

O grande desafio da estimação sequencial bayesiana é estabelecer o critério de parada e, por consequência, uma regra de decisão terminal que dará a estimativa para o parâmetro de interesse, minimizando a perda esperada e o custo, e sabe-se, que uma amostra menor custa menos, mas acarreta em uma perda esperada maior (GHOSH; MUKHOPADHYAY; SEN, 1997).

Assim, uma forma de definir critérios de parada ideais é utilizar equações de programação dinâmica, que possui o intuito de minimizar o tamanho de amostra esperado e encontrar o ponto ótimo de parada da amostragem. Basicamente, as equações de programação dinâmica funcionam por uma relação de recorrência (ALVO, 1977).

A programação dinâmica, conhecida também como otimização recursiva, é um procedimento de otimização para resolver problemas de decisão sequencial ou de múltiplos estágios relacionados. A ligação entre esta técnica e a teoria da decisão foi dada principalmente em Lindley e Barnett (1965), Freeman (1972) e Jones (1976).

Esta técnica computacional tem sido utilizada na literatura para encontrar os procedimentos sequenciais ótimos para diferentes problemas de decisão estatística. Ela é particularmente importante em processos de vários estágios, onde as decisões são tomadas sequencialmente e não são independentes, em que as decisões tomadas mais cedo afetarão nas decisões tomadas mais tarde (LINDLEY; BARNETT, 1965).

## 3.3 Distribuição Binomial

Classificar elementos amostrais segundo alguma característica está presente na maioria das áreas de conhecimento. Uma situação mais simples é aquela em que os elementos amostrais são classificados em apenas duas categorias, ou seja, a variável de interesse é dicotômica, possuindo apenas duas opções de resposta, como sim ou não, doente ou não doente, ou genericamente, sucesso ou fracasso. Desse modo, a distribuição binomial é utilizada para estimar a probabilidade de um elemento pertencer a uma das categorias (PIRES; NUNES, 2014).

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):

$$P(X = x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{3.7}$$

em que $x = 0, 1, ..., n$, 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:

$$E(X) = np, \quad Var(X) = np(1-p). \tag{3.8}$$

## 3.4 Distribuição Multinomial

Em situações mais complexas, a classificação de elementos amostrais pode ser realizada em mais de duas categorias, como por exemplo, Rossetto e Gonçalves (2015) fizeram uma classificação de acordo com a raça (branca, negra, parda, amarela), renda e escolaridade dos pais, que são variáveis categóricas, para analisar equidade nas políticas públicas de acesso na educação superior. Além disso, Georges (2019) fez uma pesquisa de opinião em que as perguntas possuíam três alternativas de resposta.

Em ambos os casos, a variável de interesse é politômica e a distribuição multinomial é utilizada para estimar a probabilidade de um elemento pertencer a mais de duas categorias, sendo esta uma distribuição discreta de probabilidade e uma generalização da distribuição binomial (MCCULLAGH; NELDER, 1989).

Desse modo, de acordo com Casella e Berger (2002), a distribuição multinomial é definida supondo um experimento cujo resultado seja um dos eventos $E_1, E_2, ..., E_k$, com probabilidade $P[E_i] = p_i$, sendo $k$ o número de classes da distribuição multinomial. Para $i = 1, 2, ..., k$, $0 \leq p_i \leq 1$

e $\sum_{i=1}^{k} p_i = 1$, e seja $X_i$ uma variável aleatória que conta o número de ocorrências de $E_i$ em $m$ repetições independentes desse experimento. Então, o vetor aleatório $(X_1, X_2, ..., X_k)$ tem distribuição chamada **multinomial**, com parâmetros $p_1, p_2, ..., p_{k-1}$ e $m$, dada por:

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

$$= \frac{m!}{x_1! x_2! \cdot ... \cdot x_k!} p_1^{x_1} p_2^{x_2} \cdot ... \cdot p_k^{x_k} = m! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!}. \tag{3.9}$$

em que cada $X_i$ é um inteiro positivo, $p_1, p_2, ..., p_k$ são proporções populacionais e $\sum_{i=1}^{k} x_i = m$. Existem $p_1, p_2, ..., p_{k-1}$ parâmetros, porque tem-se que $\sum_{i=1}^{k} p_i = 1$, portanto $p_k = 1 - \sum_{i=1}^{k-1} p_i$ Assim, escreve-se que:

$$(\boldsymbol{X}|m, \boldsymbol{p}) \sim \text{Multinomial}(m, \boldsymbol{p}).$$

Um resultado importante é que a distribuição marginal de qualquer componente $X_i$ da distribuição multinomial corresponde a uma binomial $(m, p_i)$, em que $m$ é o número de tentativas e $p_i$ é a probabilidade de obtenção de $X_i$ em cada tentativa. Para demonstrar tal fato, faz-se necessário apresentar o teorema multinomial (CASELLA; BERGER, 2002):

**Teorema multinomial:** Seja $m$ e $n$ inteiros positivos. Seja $A$ o conjunto de vetores de $x = (x_1, ..., x_k)$ de tal modo que cada $x_i$ é um inteiro não negativo e $\sum_{i=1}^{k} x_i = m$. Então para quaisquer números reais $p_1, ..., p_k$, tem-se que

$$(p_1 + ... + p_k)^m = \sum_{x \in A} \frac{m!}{x_1! \cdot ... \cdot x_k!} p_1^{x_1} \cdot ... \cdot p_k^{x_k}. \tag{3.10}$$

Sabe-se que $\sum_{x \in A} \frac{m!}{x_1! \cdot ... \cdot x_k!} p_1^{x_1} \cdot ... \cdot p_k^{x_k} = 1$, pela propriedade de variável aleatória discreta, que diz que $\sum p(x_i) = 1$. O conjunto $A$ é o conjunto dos pontos com probabilidade positiva, então a soma sobre todos esses pontos é $(p_1 + ... + p_k)^m = 1^m = 1$. Logo, este teorema diz que a soma de probabilidades de uma multinomial é 1.

Para calcular a distribuição marginal de $X_i$ de uma distribuição multinomial, considera-se a função massa de probabilidade de $X_k$, para um valor fixo de $x_k \in \{0, 1, ..., k\}$, com o objetivo de calcular a função massa de probabilidade marginal $f(x_k)$, deve-se somar todos os possíveis valores

de $(x_1,...,x_{k-1})$, ou seja, soma-se sobre todos os $(x_1,...,x_{k-1})$ de modo que os $x_i$'s sejam todos inteiros não negativos e $\sum_{i=1}^{k-1} x_i = m - x_k$, conjunto o qual será denotado por $B$. Então,

$$f(x_k) = \sum_{(x_1,...,x_{k-1})\in B} \frac{m!}{x_1! \cdot ... \cdot x_k!} p_1^{x_1} \cdot ... \cdot p_k^{x_k}$$

$$f(x_k) = \sum_{(x_1,...,x_{k-1})\in B} \frac{m!}{x_1! \cdot ... \cdot x_k!} p_1^{x_1} \cdot ... \cdot p_k^{x_k} \frac{(m-x_k)!(1-p_k)^{m-x_k}}{(m-x_k)!(1-p_k)^{m-x_k}}$$

$$f(x_k) = \frac{m!}{x_k!(m-x_k)!} p_k^{x_k} (1-p_k)^{m-x_k} \times$$

$$\times \underbrace{\sum_{(x_1,...,x_{k-1})\in B} \frac{(m-x_k)!}{x_1! \cdot ... \cdot x_{k-1}!} \left(\frac{p_1}{1-p_k}\right)^{x_1} \cdot ... \cdot \left(\frac{p_{k-1}}{1-p_k}\right)^{(x_k-1)}}_{1}$$

$$f(x_k) = \frac{m!}{x_k!(m-x_k)!} p_k^{x_k} (1-p_k)^{m-x_k}.$$

Usou-se o fato de que $x_1 + ... + x_{k-1} = m - x_k$ e $p_1 + ... + p_{k-1} = 1 - p_k$, e ainda o teorema multinomial, para concluir que o somatório anterior é igual a 1. Assim, fica demonstrado que a distribuição marginal de $X_k$ é uma binomial $(m, p_k)$. Logo, pode-se generalizar que a distribuição marginal para qualquer componente $X_i$ da distribuição multinomial corresponde a uma binomial $(m, p_i)$.

Desse modo, em função da equivalência demonstrada anteriormente, tem-se que os equacionamentos dos parâmetros de amostragem (esperança e variância) utilizados para amostragem de variáveis dicotômicas são válidos para a amostragem de variáveis politômicas. Ou seja, a esperança de $X_i$ é $mp_i$ e a sua variância é $mp_i(1-p_i)$, que são equivalentes ao caso binomial (SILVA, 2012).

Ainda, tem-se que dadas duas variáveis aleatórias multinomiais $X_i$ e $X_j$, com $i \neq j$, a covariância entre elas é dada por (OLIVEIRA; SILVA, 2018):

$$Cov(X_i, X_j) = -mp_i p_j. \tag{3.11}$$

## 3.5 Estimação Frequentista

A Inferência Estatística clássica, segundo Paulino, Turkman A. A. e Silva (2018), foi impulsionada pelos pensamentos de Karl Pearson, Ronald A. Fisher e Jerzy Neyman. Seu principal objetivo é fazer generalizações sobre uma população, com base em dados de uma amostra, em que população é o conjunto de todos os elementos ou observações possíveis de serem feitas sob condições semelhantes, e amostra é qualquer subconjunto da população.

Um dos problemas básicos da Inferência Estatística é a estimação de parâmetros, já que consiste em produzir afirmações sobre características específicas de interesse da população, a partir de informações colhidas de uma parte dessa população, ou seja, de uma amostra. Desse modo, um parâmetro nada mais é que uma medida usada para descrever uma característica da população (BUSSAB; MORETTIN, 2017).

Ainda segundo Bussab e Morettin (2017), os dados são observações de uma variável aleatória $X$ ou de $n$ variáveis aleatórias $X = (X_1, X_2, ..., X_n)$, com função de distribuição $F_X(X)$. As distribuições de probabilidade são descritas pelos seus parâmetros populacionais e na prática, frequentemente o pesquisador tem alguma ideia sobre a forma da distribuição, mas não dos valores exatos dos parâmetros que a especificam. Assim, precisa-se conhecer os parâmetros da distribuição para que ela fique completamente especificada. Então, o propósito do pesquisador é descobrir, ou seja, estimar os parâmetros da distribuição para sua posterior utilização.

No processo de estimação de parâmetros de uma população, como os dados são observações de uma variável aleatória, qualquer função desses dados é também uma variável aleatória, e essa função é chamada de estatística, sendo portanto uma característica da amostra. A estatística possui uma distribuição de probabilidade, chamada de distribuição amostral. A noção de distribuição amostral é muito importante na inferência estatística, pois permite tirar conclusões sobre o parâmetro de população correspondente com base em uma amostra aleatória (CASELLA; BERGER, 2002).

As estimativas de parâmetros podem ser obtidas pela estimação pontual ou pela estimação intervalar, na estimação por ponto, procede-se a estimação do parâmetro através de um único valor. Já na estimação por intervalo, encontra-se um limite inferior e um limite superior, os quais formam

um intervalo de valores, dentro do qual espera-se, com certo grau de confiança, que o verdadeiro valor do parâmetro esteja incluído.

Os procedimentos clássicos são realizados a partir da amostragem repetida e todos efetuados nas mesmas condições. Uma das características deste princípio reside precisamente na interpretação frequentista de probabilidade, em que, as medidas de incertezas são frequências geradas (NEIVA, 2019).

O estimador do parâmetro $p$, também denominado proporção amostral, de acordo com Bussab e Morettin (2017) é definido como:

$$\hat{p} = \frac{X}{n}. \tag{3.12}$$

sendo que, $X$ denota o número de elementos na amostra que apresentam a característica e $n$ é o tamanho da amostra coletada.

Ao observar o valor $x$ da variável aleatória $X$, obtém-se $\hat{p} = x/n$, que é denominado como estimativa para $p$. Desse modo, $\hat{p}$ dado pela equação (3.12), é uma variável aleatória, ao passo que $x/n$ é um número, ou seja, um valor da variável aleatória.

### 3.5.1 Estimação Frequentista do parâmetro $p$ da distribuição Binomial

De acordo com Bussab e Morettin (2017), o estimador de máxima verossimilhança do parâmetro $p$ de uma distribuição binomial é:

$$\hat{p}_{MV} = \frac{X}{n}. \tag{3.13}$$

Para demonstrar tal fato, pode-se observar que a função de verossimilhança nesse caso é:

$$L(p) = \binom{n}{x} p^x (1-p)^{n-x}. \tag{3.14}$$

que é a probabilidade de se obter $x$ sucessos e $n-x$ fracassos. O máximo dessa função ocorre no mesmo ponto que $l(p) = log_e L(p)$. Assim, denotando o logaritmo natural apenas por $log$, tem-se

que a função de log-verossimilhança é:

$$l(p) = log \binom{n}{x} + x\ log\ p + (n-x)\ log\ (1-p). \tag{3.15}$$

Derivando:

$$l'(p) = \frac{x}{p} - \frac{n-x}{1-p}. \tag{3.16}$$

Igualando a zero, $p$ vira uma estimativa, então tem-se:

$$l'(\hat{p}) = \frac{x}{\hat{p}} - \frac{n-x}{1-\hat{p}} = 0 \Rightarrow \tag{3.17}$$

$$\Rightarrow \frac{x}{\hat{p}} = \frac{n-x}{1-\hat{p}} \Rightarrow$$

$$\Rightarrow \hat{p}(n-x) = x(1-\hat{p}) \Rightarrow$$

$$\Rightarrow \hat{p}n - \hat{p}x = x - \hat{p}x \Rightarrow$$

$$\Rightarrow \hat{p}n - \cancel{\hat{p}x} + \cancel{\hat{p}x} = x \Rightarrow$$

$$\Rightarrow \hat{p} = \frac{x}{n}.$$

Logo, obtém-se que $\hat{p}_{MV} = \frac{x}{n}$.

Assim, o procedimento de se obter o estimador do parâmetro $p$ da distribuição binomial pelo método da máxima verossimilhança, se resume em obter a função de verossimilhança, que depende dos parâmetros desconhecidos e dos valores amostrais, e depois maximizar essa função, ou ainda, o logaritmo dela, o que pode ser mais conveniente em determinadas situações.

### 3.5.2 Estimação Frequentista do parâmetro $p$ da distribuição Multinomial

A função de verossimilhança da distribuição multinomial é dada por (MURPHY, 2006):

$$L(p) = \prod_{i=1}^{k} p_i^{x_i}. \tag{3.18}$$

O máximo dessa função ocorre no mesmo ponto em que $l(p) = log_e L(p)$. Logo, a função de log-verossimilhança é:

$$l(p) = \sum_{i=1}^{k} x_i \, log p_i. \tag{3.19}$$

Na distribuição multinomial, tem-se a restrição $\sum_{i=1}^{k} p_i = 1$, logo, para encontrar o máximo da função (3.19), para essa restrição, utiliza-se um multiplicador de Lagrange. Desse modo, a função de custo restrito torna-se:

$$\hat{p} = \sum_{i=1}^{k} x_i \, log \; p_i + \lambda \left( 1 - \sum_{i=1}^{k} p_i \right). \tag{3.20}$$

Derivando e igualando a zero, tem-se:

$$\frac{\partial \hat{l}}{\partial p_i} = \frac{x_i}{p_i} - \lambda = 0. \tag{3.21}$$

Derivando em relação a $\lambda$:

$$\frac{\partial \hat{l}}{\partial \lambda} = \left( 1 - \sum_{i=1}^{k} p_i \right) = 0. \tag{3.22}$$

Assim, de (3.21), tem-se:

$$x_i = \lambda p_i. \tag{3.23}$$

Tomando o somatório de ambos os lados, tem-se:

$$\sum_{i=1}^{k} x_i = \lambda \sum_{i=1}^{k} p_i. \tag{3.24}$$

Da distribuição multinomial, sabe-se que $\sum_{i=1}^{k} x_i = m$ e $\sum_{i=1}^{k} p_i = 1$, substituindo por isso em (3.24), tem-se:

$$m = \lambda \cdot 1 \Rightarrow m = \lambda. \tag{3.25}$$

Logo, voltando em (3.23), e usando o fato encontrado em (3.25), tem-se que o estimador de máxima verossimilhança para o parâmetro $p_i$ da distribuição multinomial, é:

$$\hat{p}_{iMV} = \frac{X_i}{m}. \tag{3.26}$$

## 3.6 Estimação Bayesiana

A metodologia bayesiana surgiu em consequência da publicação "*An essay towards solving a problem in the doctrine of changes*", atribuída ao Reverendo Thomas Bayes e comunicado à Royal Statistical Society após sua morte por Richard Price, em 1763 (TIMPANI; NASCIMENTO, 2015).

A estatística bayesiana ganhou força ao longo do tempo com a evolução dos computadores em termos de capacidade de processamento, pois auxilia nos cálculos, e despertou grande interesse principalmente em áreas aplicadas, como uma metodologia alternativa em relação aos procedimentos clássicos de estimação (REIS et al., 2009).

Depois da análise dos dados, o propósito de qualquer estatístico é fazer inferências ou predições, com certo grau de confiança, sobre o fenômeno que se está estudando, a partir dos dados que representam a variabilidade ou a incerteza na observação da característica ou fenômeno (GELMAN et al., 2013).

Na estatística clássica, ou ainda, frequentista, as inferências sobre um fenômeno ou característica que ocorre na população estudada são baseadas na avaliação dos parâmetros estimados de amostras retiradas dessa população. Assim, toda a inferência é feita a partir dos dados disponíveis, e as conclusões ocorrem a partir dos dados amostrais. Ela se baseia no princípio da repetibilidade, que consiste em considerar-se a variabilidade que seria observada caso um mesmo experimento fosse repetido, sob as mesmas condições, um grande número de vezes (PAULINO; TURKMAN A. A.; SILVA, 2018).

Diferentemente da estatística frequentista, em que somente se admite probabilidade através de medidas de frequências relativas, na análise bayesiana entende-se que a probabilidade é uma medida racional e condicional de incerteza. Nesta abordagem, o valor do parâmetro de interesse

$\theta$ é desconhecido, o qual é considerado uma variável aleatória, e o intuito é tentar reduzir esse desconhecimento sobre $\theta$ (KINAS; ANDRADE, 2020).

A intensidade da incerteza a respeito de $\theta$ pode assumir diferentes graus. Estes diferentes graus de incerteza são representados por meio de modelos probabilísticos para $\theta$, que é a principal característica da inferência bayesiana, utilizar probabilidade para quantificar essas incertezas, e esta é uma medida subjetiva, que pode variar de acordo com o pesquisador, pois a experiência e a fonte dessa informação que cada um possui são diferenciadas (EHLERS, 2011).

Assim, assume-se uma distribuição de probabilidade associada, conhecida como distribuição *a priori*, a qual é especificada antes da observação dos dados e descreve o conhecimento, ou ainda, o grau da crença do pesquisador sobre o parâmetro desconhecido, e esse conhecimento pode ser formalmente incorporado na análise (GELMAN et al., 2013).

A informação *a priori* é a hipótese que o pesquisador fixa como sendo o valor verdadeiro do parâmetro estudado. Essa *priori* pode ser extraída de fundamentos subjetivos, da experiência do pesquisador na área em questão por análises realizadas anteriormente, por considerações particulares ou ainda por informações disponíveis na literatura sobre o assunto estudado (TIMPANI; NASCIMENTO, 2015).

Especificar distribuições *a priori* quando a sua obtenção e quantificação são de natureza essencialmente subjetiva, e o intuito é transformá-la em uma informação que possa ser utilizada, muitas vezes é uma tarefa difícil. Essas dificuldades costumam ser contornadas através da adoção de uma forma distribucional conveniente denominada família das distribuições conjugadas (PAULINO; TURKMAN A. A.; SILVA, 2018).

Segundo Ehlers (2011), uma *priori* é dita conjugada se as distribuições *a priori* e *a posteriori* pertencem a mesma classe de distribuições. Tem-se que *prioris* conjugadas são os casos mais importantes ao utilizar uma abordagem com hiperparâmetros, que em geral, facilita a análise e consiste em definir uma família paramétrica de densidades.

Nesta abordagem, a distribuição *a priori* é representada por uma forma funcional, cujos parâmetros são especificados de acordo com o conhecimento que se tem sobre $\theta$. Desse modo, os parâmetros indexadores da família de distribuições *a priori* recebem o nome de hiperparâmetros

para distingui-los dos parâmetros de interesse $\theta$, e a ideia é que a atualização do conhecimento que se tem de $\theta$ envolva apenas uma mudança nos hiperparâmetros (EHLERS, 2011).

A teoria da inferência Bayesiana está fundamentada no teorema de Bayes, o qual é basicamente um resultado de uma probabilidade condicional. Para o caso em que o parâmetro $\theta$ assume valores contínuos num dado intervalo, o teorema é dado por (PAULINO; TURKMAN A. A.; SILVA, 2018):

$$\pi(\theta|X) = \frac{L(X|\theta)\pi(\theta)}{\int L(X|\theta)\pi(\theta)d\theta}. \tag{3.27}$$

e, no caso em que $\theta$ assume valores discretos, tem-se:

$$\pi(\theta|X) = \frac{L(X|\theta)\pi(\theta)}{\sum L(X|\theta)\pi(\theta)}. \tag{3.28}$$

onde:

- $\pi(\theta|X)$ é distribuição *a posteriori* do parâmetro $\theta$;
- $\pi(\theta)$ é a distribuição *a priori* de $\theta$;
- $L(X|\theta)$ é a função de verossimilhança de $\theta$, que é uma distribuição conjunta para os dados amostrais, a qual representa a informação sobre $\theta$ que foi obtida dos dados, ela é representada por:

$$L(X|\theta) = \prod_{i=1}^{n} f(x_i|\theta). \tag{3.29}$$

em que $x_1, x_2, ..., x_n$ é uma amostra aleatória de tamanho $n$ da variável aleatória $X$ com função densidade (ou de probabilidade) $f(x|\theta)$, em que $\theta \in \Theta$, sendo $\Theta$ o espaço paramétrico.

Uma forma equivalente das expressões (3.27) e (3.28), que é a forma usual do teorema de Bayes, visto que o denominador não depende do parâmetro, é dada por:

$$\pi(\theta|X) \propto L(X|\theta)\pi(\theta). \tag{3.30}$$

Ao omitir o denominador, a igualdade nas expressões (3.27) e (3.28) foram substituídas por uma proporcionalidade. Esta forma simplificada do teorema de Bayes é útil em problemas que envolvem estimação de parâmetros já que o denominador é apenas uma constante normalizadora.

Em outras situações, como seleção e comparação de modelos, este termo tem um papel crucial (EHLERS, 2011).

Assim, a partir do Teorema de Bayes, a distribuição *a priori* é combinada com a informação contida nos dados amostrais, ou seja, com a função de verossimilhança, obtendo assim a distribuição *a posteriori*. Portanto, toda a inferência relativa a um determinado parâmetro é realizada utilizando-se a distribuição *a posteriori*, pois esta contém toda a informação probabilística a respeito do parâmetro (FILHO et al., 2008).

Entretanto, para obter conclusões de modo mais prático é necessário identificar medidas que resumem a distribuição *a posteriori*, assim, segundo Paulino, Turkman A. A. e Silva (2018) ela pode ser resumida por meio da média, da moda, da mediana e do intervalo de credibilidade e/ou dos intervalos de máxima densidade *a posteriori*. Além disso, um caso simples é a estimação pontual de $\theta$, já que esta resume à distribuição marginal *a posteriori* por meio de um único valor, $\hat{\theta}$.

### 3.6.1 Estimação Bayesiana dos parâmetros da distribuição Binomial

Em estatística bayesiana, a distribuição *a priori* conjugada da distribuição binomial é a distribuição Beta, que é caracterizada por dois parâmetros $a$ e $b$. Ela é frequentemente utilizada na modelagem de variáveis aleatórias contínuas com domínio entre 0 e 1, tal como as proporções (PAULINO; TURKMAN A. A.; SILVA, 2018).

A função densidade de probabilidade da distribuição *a priori* Beta, para quaisquer hiperparâmetros $a > 0$ e $b > 0$, é dada por:

$$\pi(p|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}. \tag{3.31}$$

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:

$$E(p) = \frac{a}{a+b}, \quad Var(p) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{3.32}$$

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

$$E(p|X) = \frac{a+x}{(a+x+b+n-x)}, \quad Var(p|X) = \frac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)}. \tag{3.33}$$

Através da média e a variância para a distribuição *a priori* Beta $(a,b)$ apresentadas em (3.32), tem-se que:

$$ab = (a+b)^2(a+b+1)\sigma^2, \tag{3.34}$$

e que

$$b = \frac{a}{\mu} - a. \tag{3.35}$$

Desse modo, as estimativas de $a$ e $b$ podem ser calculadas da seguinte maneira:

Substituindo (3.35) em (3.34), encontra-se $a$:

$$\frac{a^2}{\mu} - a^2 = \left(a + \frac{a}{\mu} - a\right)^2 \left(a + \frac{a}{\mu} - a + 1\right) \sigma^2$$

$$\frac{a^2}{\mu} - a^2 = \left(\frac{a}{\mu}\right)^2 \left(\frac{a}{\mu} + 1\right) \sigma^2$$

$$a^2 \left(\frac{1}{\mu} - 1\right) = a^2 \left(\frac{1}{\mu^2}\right) \left(\frac{a}{\mu} + 1\right) \sigma^2$$

$$\left(\frac{1}{\mu} - 1\right) = \left(\frac{1}{\mu^2}\right) \left(\frac{a}{\mu} + 1\right) \sigma^2$$

$$\left(\frac{1}{\mu} - 1\right) \left(\frac{\mu^2}{\sigma^2}\right) = \left(\frac{a}{\mu} + 1\right)$$

$$\left(\frac{\mu(1-\mu)}{\sigma^2}\right) = \left(\frac{a}{\mu} + 1\right)$$

$$\left(\frac{\mu(1-\mu)}{\sigma^2}\right) = \left(\frac{a}{\mu} + 1\right)$$

$$a = \mu \left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right). \tag{3.36}$$

Substituindo (3.36) em (3.35), encontra-se $b$:

$$b = a\left(\frac{1-\mu}{\mu}\right)$$

$$b = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right)\left(\frac{1-\mu}{\mu}\right)$$

$$b = (1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right). \tag{3.37}$$

Logo, o procedimento para construir as *prioris*, inicia-se com o fornecimento de valores para a média e a variância.

### 3.6.2 Estimação Bayesiana dos parâmetros da distribuição Multinomial

A distribuição de Dirichlet é a generalização multivariada da distribuição beta, com um parâmetro vetorial não-negativo e real $\boldsymbol{a}$. É uma distribuição discreta multivariada amplamente utilizada no contexto bayesiano como a distribuição *a priori* conjugada da distribuição multinomial (PAULINO; TURKMAN A. A.; SILVA, 2018).

Seja $\boldsymbol{X} = (X_1, ..., X_k)^T$ um vetor com $k$ componentes, então ele segue uma distribuição de Dirichlet de ordem $k \geq 2$ com um vetor de parâmetros $\boldsymbol{a} = (a_1, ..., a_k)^T$, isto é (EHLERS, 2011):

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

E sua função densidade de probabilidade é dada por:

$$f_{\boldsymbol{X}}(x_1, ..., x_k) = \frac{\Gamma(a_0)}{\prod_{i=1}^{k}\Gamma(a_i)}\prod_{i=1}^{k} p_i^{a_i-1},\ 0 < p_i < 1, \tag{3.38}$$

sendo $a_0 = \sum_{i=1}^{k} a_i$, $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx$ a função gama e $\sum_{i=1}^{k} p_i = 1$. A distribuição marginal é uma beta com parâmetros $a_i$ e $(a_0 - a_i)$, para cada $i$, de onde tem-se:

$$E(X_i) = \frac{a_i}{a_0}, \quad Var(X_i) = \frac{a_i(a_0-a_i)}{a_0^2(a_0+1)}, \quad Cov(X_i, X_j) = \frac{-a_i a_j}{a_0^2(a_0+1)}. \tag{3.39}$$

Em (3.38) tem-se um vetor k-dimensional, em que os elementos são interpretados como as probabilidades de ter as proporções $(p_1, ..., p_k)$.

O modelo Multinomial-Dirichlet é dado pela combinação do modelo probabilístico multinomial com uma distribuição *a priori* de Dirichlet, generalizando assim os resultados obtidos com o modelo binomial e a distribuição *a priori* beta (LINDLEY, 1964).

Ele possui a seguinte estrutura hierárquica:

$$(X_1, ..., X_k | m, \boldsymbol{p}) \sim \text{Multinomial}(m, p_1, ..., p_k).$$

$$(p_1, ..., p_k | \boldsymbol{a}) \sim \text{Dirichlet}(a_1, ..., a_k).$$

com função de probabilidade e densidade conjunta definidas em (3.9) e (3.38), respectivamente. Ou seja, os vetores de probabilidade $\boldsymbol{p}$, parâmetros da distribuição Multinomial, seguem uma distribuição de Dirichlet com parâmetros $\boldsymbol{a}$.

Sendo assim, se a distribuição *a priori* é uma Dirichlet e a variável observada segue uma distribuição Multinomial, então a distribuição *a posteriori* será uma distribuição de Dirichlet, com outro parâmetro. A distribuição *a posteriori* é encontrada utilizando o teorema de Bayes (3.27), segundo Avetisyan e Fox (2012), como segue:

$$\pi(\boldsymbol{p}|X) = \frac{\left(\frac{m!}{x_1! \cdot ... \cdot x_k!}\right) \prod_{i=1}^{k} p_i^{x_i} \frac{\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} p_i^{a_i - 1}}{\int_{\boldsymbol{p}} \left(\frac{m!}{x_1! \cdot ... \cdot x_k!}\right) \prod_{i=1}^{k} p_i^{x_i} \frac{\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} p_i^{a_i - 1} d\boldsymbol{p}}$$

$$\pi(\boldsymbol{p}|X) \propto \left(\frac{m!}{x_1! \cdot ... \cdot x_k!}\right) \prod_{i=1}^{k} p_i^{x_i} \frac{\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} p_i^{a_i - 1}$$

$$\pi(\boldsymbol{p}|X) \propto \prod_{i=1}^{k} p_i^{x_i + a_i - 1}.$$

Portanto, *a posteriori* segue uma distribuição de Dirichlet com parâmetros:

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

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

Assim, a média, a variância e a covariância da distribuição *a posteriori* de Dirichlet são dadas respectivamente por:

$$E(\boldsymbol{p}|X_i) = \frac{x_i + a_i}{\sum_{i=1}^{k}(x_i + a_i)}, \tag{3.40}$$

$$Var(\boldsymbol{p}|X_i) = \frac{(x_i + a_i)\{\left[\sum_{i=1}^{k}(x_i + a_i)\right] - (x_i + a_i)\}}{[\sum_{i=1}^{k}(x_i + a_i)]^2\{\left[\sum_{i=1}^{k}(x_i + a_i)\right] + 1\}}, \tag{3.41}$$

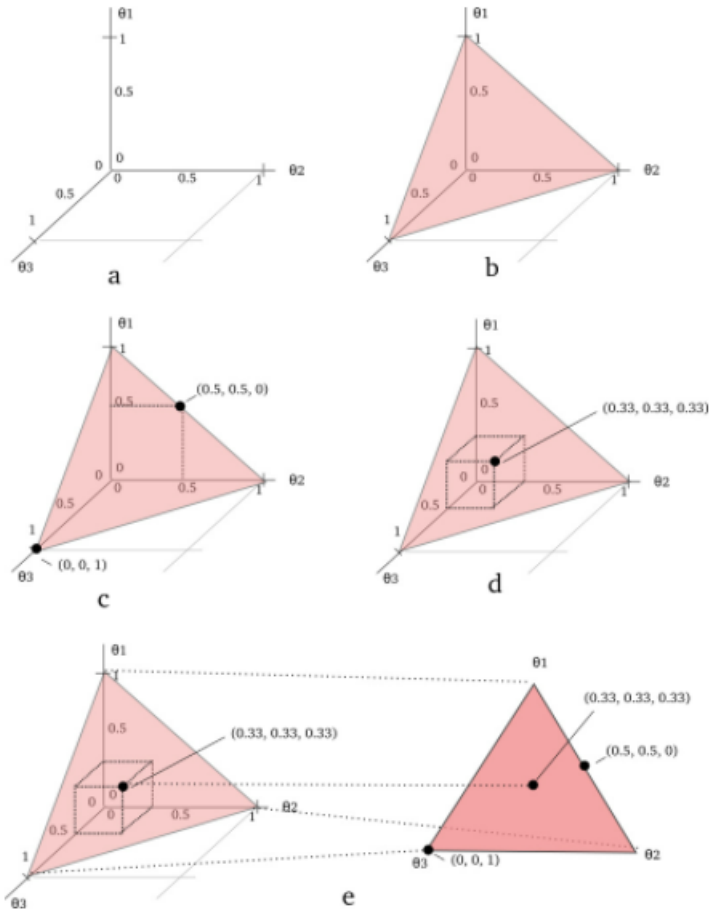
$$Cov(\boldsymbol{p}|X_i, X_j) = \frac{-(x_i + a_i)(x_j + a_j)}{[\sum_{i=1}^{k}(x_i + a_i)]^2\{\left[\sum_{i=1}^{k}(x_i + a_i)\right] + 1\}}. \tag{3.42}$$

#### 3.6.2.1 Visualização da distribuição de Dirichlet

A distribuição de Dirichlet é uma distribuição multivariada, sendo assim, ela pode ter mais de duas dimensões. Ela é muito utilizada para representar variáveis contínuas situadas em um intervalo limitado que representam composição, mais conhecidas como variáveis composicionais, cujas componentes são as proporções. Como por exemplo, a composição do solo, dado pelos teores de argila, silte e areia, a composição de um alimento, dado pela proporção de carboidrato, proteína, gordura e outros (SHIMIZU; ACHCAR; TARUMMOTO, 2014).

A distribuição de Dirichlet tem suporte sobre vetores de números reais positivos entre 0 e 1, que juntos somam 1, que é uma característica desta distribuição. Ela é definida no k$-$1-Simplex, que é uma generalização de um triângulo, em que $k$ é o número de componentes do parâmetro vetorial $\boldsymbol{a}$. Por exemplo, para $k = 3$, o suporte é um triângulo equilátero com vértices em $(1,0,0)$, $(0,1,0)$ e $(0,0,1)$ (BARNDORFF-NIELSEN; JØRGENSEN, 1991). A Figura 3.3 mostra exemplos do simplex e alguns pontos plotados:

Figura 3.3 – Exemplo simplex




Fonte: https://stats.stackexchange.com/questions/81136/the-meaning-of-representing-the-simplex-as-a-triangle-surface-in-dirichlet-distr

O parâmetro da distribuição de Dirichlet pode caracterizar a variabilidade aleatória de uma distribuição multinomial, ou seja, pode modelar a aleatoriedade de uma função de massa de probabilidade (BLEI; NG; JORDAN, 2003). Para entender melhor o que a distribuição de Dirichlet descreve, considerou-se o exemplo dado em Boggs (2019): suponha-se que são fabricados dados de 6 lados. Mas, para este exemplo, estes dados admitem como resultado de uma jogada apenas os números 1, 2 ou 3.

Assim, ao produzir um dado, o número 1 será colocado em duas das faces e será feito da mesma forma para os números 2 e 3. Se o dado produzido for um dado "justo", então as probabilidades dos três resultados serão semelhantes e iguais a $1/3$. Pode-se representar as probabilidades

para os resultados possíveis como um vetor:

$$\boldsymbol{\theta} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$$

O vetor $\boldsymbol{\theta}$ possui duas propriedades importantes:

**1)** A soma das probabilidades para cada resultado deve ser igual a 1:

$$\sum \boldsymbol{\theta} = \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2 + \boldsymbol{\theta}_3 = 1$$

**2)** Nenhuma das probabilidades pode ser negativa.

Desse modo, quando essas condições se mantêm, as probabilidades associadas aos resultados do lançamento do dado são descritas por uma distribuição multinomial.

Independentemente de produzir dados justos ou não, haverá alguma variabilidade nas características dos dados, pois os processos de fabricação são bons, mas não perfeitos. Assim, ao lançar 1000 dados, as chances teóricas de qualquer número seria $1/3$, mas não se obtém essa distribuição exata em um experimento real devido a defeitos de fabricação.

Se os dados fossem esculpidos de um bloco de madeira e à mão, haveria uma variabilidade significativa nos dados, devido às limitações da habilidade ao construí-los, variabilidade na densidade da madeira, características das ferramentas, entre outros.

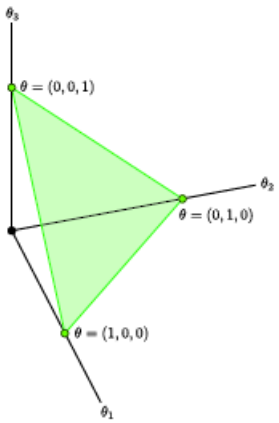
Se fosse utilizada uma impressora 3D sofisticada para fabricar os dados, espera-se que eles seriam muito mais precisos, tendo uma variabilidade significativamente menor do que os dados esculpidos à mão. Entretanto, poderá haver um pouco de oscilação no peso. Logo, cada dado terá sua própria função de massa de probabilidade.

Para caracterizar matematicamente essa variabilidade, precisa-se conhecer a densidade de probabilidade de cada valor possível de $\boldsymbol{\theta}$ para um determinado processo de fabricação. Para fazer isso, considera-se cada elemento de $\boldsymbol{\theta}$ sendo uma variável independente. Ou seja, para $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$, pode-se tratar $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ e $\boldsymbol{\theta}_3$ cada um como uma variável independente e $\boldsymbol{\theta}$ como um vetor em um espaço tridimensional (GOMES; CRIBARI-NETO; VASCONCELLOS, 2008).

Uma vez que a distribuição multinomial requer que essas três variáveis somem 1, sabe-se que os valores permitidos de $\boldsymbol{\theta}$ estão confinados a um plano. Além disso, como cada valor de $\boldsymbol{\theta}$

deve ser maior ou igual a zero, o conjunto de todos os valores permitidos de $\boldsymbol{\theta}$ está confinado a um triângulo equilátero (um 2-simplex) como mostrado na Figura 3.4:

Figura 3.4 – $2-$simplex $\mathbb{R}^3$




Fonte: Boggs (2019)

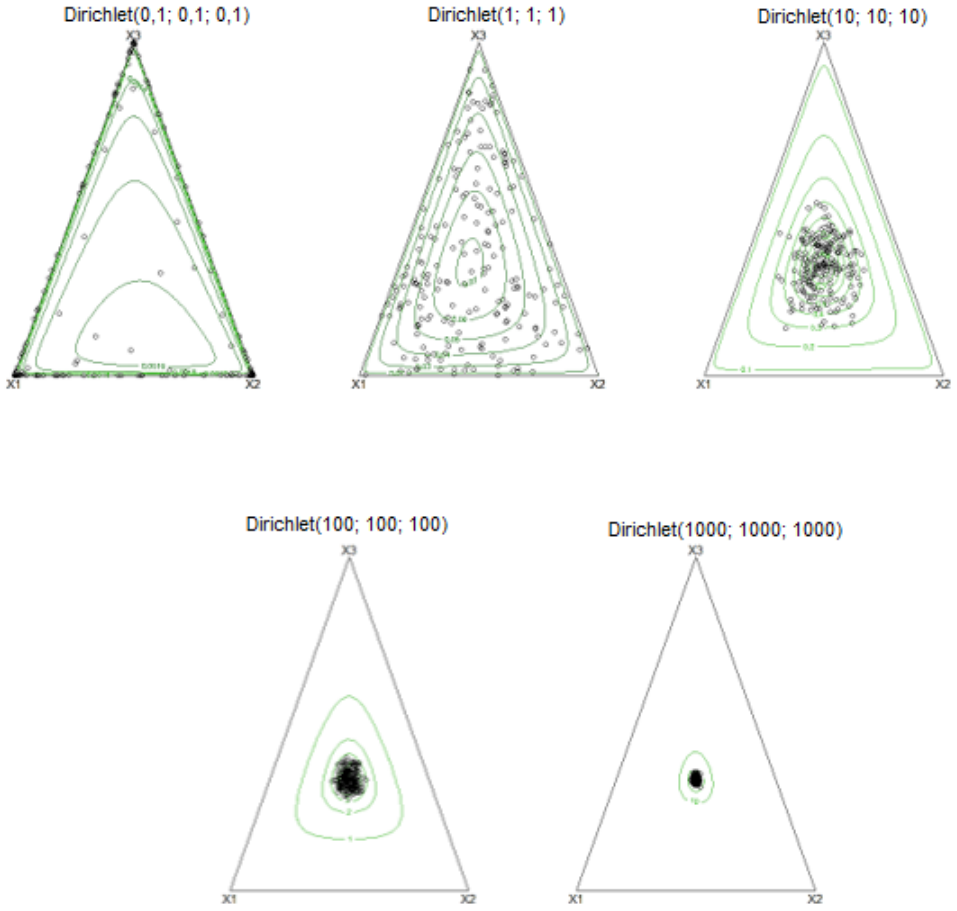
O interesse é encontrar a densidade de probabilidade em cada ponto deste triângulo. Assim, a distribuição Dirichlet auxilia neste processo, pois pode-se utilizá-la como distribuição *a priori* para a distribuição multinomial. A distribuição de Dirichlet define uma densidade de probabilidade para uma entrada de valor vetorial com as mesmas características que o parâmetro multinomial ($\boldsymbol{\theta}$) (BLEI; NG; JORDAN, 2003).

Para o caso, $n = 3$, pode-se interpretar $p(\boldsymbol{\theta}|\boldsymbol{a})$ respondendo à pergunta "qual é a densidade de probabilidade associada à distribuição multinomial $\boldsymbol{\theta}$, dado que a distribuição de Dirichlet tem parâmetro $\boldsymbol{a}$?".

Para responder a essa pergunta, o objetivo é visualizar como Dir($\boldsymbol{a}$) varia em relação ao simplex de valores permitidos de $\boldsymbol{\theta}$ para um determinado valor de $\boldsymbol{a}$. Para isso, pode-se utilizar o pacote *Compositional*, do software R (R Core Team, 2020).

Assim, considerando $n = 3$, gerou-se 200 amostras de uma distribuição de Dirichlet e construiu-se um 2-simplex, que corresponde a um triângulo equilátero, para diferentes valores de $\boldsymbol{a}$, com a função *biv.contour* do pacote *Compositional*. Onde estão representados na Figura 3.5 a seguir:

Figura 3.5 – Distribuição de Dirichlet em um 2-simplex (triângulo equilátero) para diferentes valores de $\boldsymbol{a}$.




Fonte: Elaboração própria, 2021

Pode-se observar, pela Figura 3.5, que o parâmetro $\boldsymbol{a}$ que governa as formas da distribuição, em particular, a soma $a_0 = \sum a_i$. Para valores de $a_i < 1$, a distribuição se concentra nos cantos e ao longo dos limites do simplex. Para o caso de $\boldsymbol{a} = (1, 1, 1)$ produz uma distribuição uniforme, onde todos os pontos no simplex são igualmente prováveis. Para valores de $a_i > 1$, a distribuição tende em direção ao centro do simplex.

E a medida que $\boldsymbol{a}$ aumenta, a distribuição torna-se mais concentrada em torno do centro do simplex. É importante salientar, que os pontos plotados são uma proporção entre as probabilidades resultantes, pois um ponto é a soma das três probabilidades que resultam em 1.

No contexto do exemplo de fabricação de dados, produziria-se dados consistentemente justos no limite de $\boldsymbol{a} \to \infty$. Ou ainda, para uma Dirichlet simétrica com $\boldsymbol{a > 1}$. Se o objetivo fosse

produzir dados carregados, por exemplo, com uma maior probabilidade de aparecer um 3, a distribuição de Dirichlet seria assimétrica, não central, com um valor mais alto para $\boldsymbol{a}_3$.

Desse modo, este exemplo mostra a aparência da distribuição de Dirichlet e as implicações de usá-la como distribuição *a priori* para uma função de verossimilhança multinomial.

## 3.7 Estimação Sequencial Bayesiana

Como foi visto na seção Estimação Sequencial, para realizar a estimação de parâmetros a partir de uma amostragem sequencial o mais adequado é utilizar a teoria da decisão bayesiana, já que ela permite a incorporação de informações *a priori* sobre o parâmetro de interesse, assim auxiliando na tomada de decisão.

Portanto, no procedimento de estimação sequencial bayesiana alguns conceitos estão envolvidos, como função de perda, risco de bayes, risco imediato, risco esperado, função custo, os quais serão apresentados a seguir.

Um procedimento sequencial $d$ é uma regra de decisão, que nada mais é que uma função definida no espaço dos resultados possíveis de um experimento que assume valores no espaço de possíveis ações.

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

$$L(p,d). \tag{3.43}$$

Segundo Ehlers (2011), a função de perda mais utilizada em problemas de estimação é a função de perda quadrática, definida como:

$$L(\hat{p},p) = (\hat{p}-p)^2. \tag{3.44}$$

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:

$$R(p,d) = E_{posteriori}[L(p,d)]. \tag{3.45}$$

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

$$r(\pi,d) = E^{\pi}[R(p,d)]. \tag{3.46}$$

isto é, o risco esperado associado ao procedimento de estimação do parâmetro $p$ dado *a priori* $\pi$, 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).

Segundo Berger (1985), a principal ideia envolvida no procedimento de estimação sequencial bayesiana, é que após cada observação realizada deve-se comparar o risco de Bayes *a posteriori* de tomar uma decisão imediata com o risco de Bayes *a posteriori* esperado, este será obtido se mais observações forem tomadas.

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 $r_0(\pi^n,a,n) = E^{\pi^n}[L(\theta,a,n)]$ é a perda *a posteriori* esperada da ação $a$ em $n$ (BERGER, 1985).

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

$$r_0(\pi^n,n) = var_{post}(n). \tag{3.47}$$

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):

$$r^1(\pi^n, n) = E[var_{post}(n)]. \tag{3.48}$$

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 Regra Sequencial de Bayes também pode ser conhecida como aprendizagem bayesiana, porque a distribuição *a posteriori* calculada no atual $n$ será usada para atualizar a distribuição *a priori* ainda a ser usada na $(n+1)$-ésima inspeção (BERGER, 1985).

### 3.7.1 Estimação Sequencial Bayesiana dos parâmetros da 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.

Assim, 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:

$$r_0(\pi^n, n) = var_{post}(n) = \frac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)} + C(n). \tag{3.49}$$

$$r^1(\pi^n, n) = E[var_{posteriori}] + C(n+1). \tag{3.50}$$

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, Resende e Brighenti (2011) apresentou:

Portanto, de acordo com Pham-Gia (1998) e Brighenti, Resende e Brighenti (2011), 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:

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

$$E[W(n)] = \frac{(a+x+1)(b+n+1-x-1)}{(a+b+n+1)^2(a+b+n+2)}p + \frac{(a+x+0)(b+n+1-x-0)}{(a+b+n+1)^2(a+b+n+2)}(1-p). \tag{3.51}$$

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 (3.51), 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:

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

$$r^1(\pi^n,n) = \left(\frac{a+b+n}{a+b+n+1}\right)\frac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)} + C(n+1). \tag{3.52}$$

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

$$\hat{p}_{bayesiano} = E(p|X) = \frac{a+x}{(a+x+b+n-x)}. \tag{3.53}$$

Portanto, foi possível calcular os riscos para a distribuição binomial. Pode-se observar que Jones (1974), calculou os riscos para essa mesma distribuição utilizando uma adaptação do método

"*one-step look ahead*" a partir de equações de programação dinâmica, e mostrou a proximidade entre os métodos.

Ele considerou a função de perda quadrática e custo de observação constante, no entanto, encontrou o estimador de bayes, o risco imediato e o risco esperado apenas para *priori* uniforme, assim considerando os parâmetros da distribuição beta como $a = 1$ e $b = 1$.

Logo, para verificar se as expressões encontradas por Pham-Gia (1998) e Brighenti, Resende e Brighenti (2011) são as mesmas das encontradas por Jones (1974), substituiu-se por 1 os parâmetros das expressões do estimador de bayes e do risco imediato, como pode ser visto a seguir:

O estimador de Bayes, é a média da distribuição *a posteriori* Beta, então para *priori* uniforme, tem-se:

$$E(p/X) = \frac{a+x}{a+b+n} = \frac{x+1}{n+2} \tag{3.54}$$

O risco imediato considerando *priori* uniforme, é:

$$r_0(\pi^n, n) = var_{post}(n) = \frac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)} + C(n) = \frac{(x+1)(n-x+1)}{(n+2)^2(n+3)} + C(n) \tag{3.55}$$

Desse modo, concluiu-se que as expressões (3.54) e (3.55) são idênticas às apresentadas por Jones (1974), apenas com a diferença que consideramos o custo constante aditivo e Jones (1974) considerou um custo constante multiplicativo, concluindo assim na conformidade entre os métodos, e que portanto, possuem o mesmo objetivo.

Jones (1974) comparou os métodos e concluiu que o obtido por equações de programação dinâmica resultou em pouco aumento no risco, no entanto, com a vantagem de ser mais simples de realizar os cálculos. Assim, para calcular os riscos para outras distribuições pode ser mais viável realizar essa adaptação ao método "*one-step look ahead*" utilizando equações de programação dinâmica, do que calcular $E[var_{post}(n)]$, pois para outras distribuições pode ser ainda mais complicado, assim tem-se uma alternativa para facilitar os cálculos.

### 3.7.2 Estimação Sequencial Bayesiana dos parâmetros da distribuição Multinomial

Para a distribuição multinomial com $(k+1)$ classes, de acordo com Jones (1976), a probabilidade de uma observação na i-ésima classe é $p_i$, com $i = 1,2,...,k$, e na $(k+1)$-ésima classe é $(1-\sum_{i=1}^{k} p_i)$, pois $p_{(k+1)} = (1-\sum_{i=1}^{k} p_i)$, já que $\sum_{i=1}^{k+1} p_i = 1$.

Sabe-se que a informação *a priori* sobre o parâmetro $\boldsymbol{p}$, que representa proporção, que neste caso é um vetor de parâmetros, dado por: $\boldsymbol{p} = (p_1, p_2, ..., p_k)^T$, pode ser adequadamente representada por um membro do conjugado natural Dirichlet, família de distribuições com parâmetros inteiros $a_0$, $a_i$, $i = 1,2,...,k$, com densidade proporcional a:

$$\prod_{i=1}^{k} p_i^{a_i-1} \left(1-\sum_{i=1}^{k} p_i\right)^{a_0-\sum a_i-1}.$$

com $p_i \geq 0$ e $\sum_{i=1}^{k} p_i \leq 1$.

O parâmetro $a_i$ da distribuição de Dirichlet é um parâmetro vetorial e o outro parâmetro dessa distribuição pode ser escrito como $a_0 = \sum_{i=1}^{k+1} a_i$. Logo, $a_{(k+1)} = a_0 - \sum_{i=1}^{k} a_i$, conforme a definição anterior para $k+1$ classes.

De acordo com o teorema de Bayes, após $m$ observações resultando em $x_i$ nas $i$-ésimas classes, a densidade *posteriori* de $\boldsymbol{p}$ será dada por uma Dirichlet, com parâmetros $a_0 + m$, $a_i + x_i$, onde $m$ é número total de observações, ou ainda, o tamanho amostral, e $x_i$ é o número de observações em cada umas das i-ésimas classes (JONES; MADHI, 1988).

O resultado da amostragem pode ser representado como um caminho de amostra, que começa no ponto $(\boldsymbol{a}, a_0)$, $\boldsymbol{a} = (a_1, a_2, ..., a_k)$, no espaço inteiro dimensional $(k+1)$ e é interrompido quando o limite de parada, que tem que ser determinado, é atingido. Se $a_0 = k+1$ e $a_i = 1$, então tem-se uma *priori* uniforme para $\boldsymbol{p}$, assim se uma *priori* uniforme é tomada como a origem, então qualquer outra *priori* apropriada com parâmetros inteiros fornecerá caminhos de amostra começando no ponto $(\boldsymbol{a} - \mathbf{1}, a_0 - k - 1)$, sendo $\mathbf{1}$ o vetor de linha unitária (JONES, 1976).

A distribuição uniforme é um caso especial da distribuição de Dirichlet, correspondendo ao caso em que $a_1 = a_2 = ... = a_k = 1$. A *priori* uniforme é não informativa, pois todos os possíveis valores do parâmetro de interesse são igualmente prováveis. O fato de a classe Dirichlet incluir

esses antecedentes naturais "não informativos" é uma razão para usá-la como distribuição *a priori* da distribuição multinomial (ASSIS, 2017).

Para obter os limites de parada, considerou-se a perda quadrática na estimativa de $\boldsymbol{p}$ por $\boldsymbol{d} = (d_1, d_2, ..., d_k)^T$, e esta possui a forma quadrática geral $(\boldsymbol{p}-\boldsymbol{d})^T \boldsymbol{K}(\boldsymbol{p}-\boldsymbol{d})$, onde $\boldsymbol{K}$ é uma matriz $I \times I$ simétrica positiva definida de perda constante (CHEN, 1988; JONES, 1976; OWEN, 1970). Então, segundo Ehlers (2011) utilizando uma perda quadrática, o estimador de Bayes $\boldsymbol{d}^*$ é a média da distribuição *posteriori* de $(\boldsymbol{x}, m)$, ou seja, é a média da distribuição de Dirichlet *a posteriori*, dada por:

$$d_i^* = E(\boldsymbol{p}|X_i) = \frac{a_i + x_i}{\sum(a_i + x_i)} = \frac{a_i + x_i}{a_0 + \sum x_i} = \frac{a_i + x_i}{a_0 + m}. \tag{3.56}$$

As igualdades acimas vem do fato de que $\sum_{i=1}^{k+1} a_i = a_0$, da Dirichlet e $x_i$ vem da verossimilhança, neste caso, multinomial, que representa o número de observações em cada classe, e portanto, sabe-se que $\sum_{i=1}^{k+1} x_i = m$.

No caso da utilização de uma *priori* uniforme, então os parâmetros são considerados como $a_i = 1$ e $a_0 = k+1$, assim substituindo-os na expressão (3.56) encontrada, tem-se que o estimador de Bayes usando uma *priori* uniforme é dado por (JONES, 1976):

$$d_i^* = \frac{(x_i + 1)}{(m + k + 1)}. \tag{3.57}$$

A expressão dada por $d_i^*$ também é a probabilidade marginal *posteriori* da próxima observação cair dentro da $i$-ésima classe.

O risco de Bayes ou risco imediato, ou ainda, o risco de parada de tomar uma decisão, é dado por (JONES, 1976):

$$S((x_1, ..., x_k), m) = S(\boldsymbol{x}, m) = \text{traço } \boldsymbol{K}\boldsymbol{\Sigma}. \tag{3.58}$$

onde $\boldsymbol{K}$ é uma matriz $I \times I$ simétrica positiva definida de perda constante, e $\boldsymbol{\Sigma}$ é a matriz de dispersão da distribuição *posteriori* de Dirichlet. Esta possui como diagonal principal as variâncias *a posteriori*, e nos outros componentes as covariâncias *a posteriori* da distribuição de Dirichlet.

Inicialmente, com base em Jones (1976), utilizou-se a *priori* uniforme para obter os limites de parada, como modo de facilitar os cálculos e encontrar os riscos imediato e esperado, já que esta é um caso particular da distribuição de Dirichlet. Sendo assim, considerando uma *priori* uniforme, a matriz de dispersão da distribuição *posteriori* possuirá os elementos:

$$var(p_i) = \frac{d_i^*(1-d_i^*)}{m+k+2} \quad \text{e} \quad cov(p_i, p_j) = -\frac{d_i^* d_j^*}{m+k+2}. \tag{3.59}$$

Dado que a matriz $\boldsymbol{K} = \begin{bmatrix} K_{ii} & K_{ij} & \cdots & K_{ij} \\ K_{ij} & K_{ii} & \cdots & K_{ij} \\ \vdots & \vdots & \ddots & \vdots \\ K_{ij} & K_{ij} & \cdots & K_{ii} \end{bmatrix}$ é uma matriz simétrica positiva definida de perda constante, então sabe-se que $\boldsymbol{K}$ é uma matriz quadrada e $K_{ij} = K_{ji}$.

E seja a matriz de dispersão $\boldsymbol{\Sigma} = \begin{bmatrix} \frac{d_i^*(1-d_i^*)}{m+k+2} & \frac{-d_i^* d_j^*}{m+k+2} & \cdots & \frac{-d_i^* d_j^*}{m+k+2} \\ \frac{-d_i^* d_j^*}{m+k+2} & \frac{d_i^*(1-d_i^*)}{m+k+2} & \cdots & \frac{-d_i^* d_j^*}{m+k+2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-d_i^* d_j^*}{m+k+2} & \frac{-d_i^* d_j^*}{m+k+2} & \cdots & \frac{d_i^*(1-d_i^*)}{m+k+2} \end{bmatrix}$.

Logo,

$$S(\boldsymbol{x}, m) = \text{traço } \boldsymbol{K\Sigma} =$$

$$= \frac{1}{m+k+2} \left( \sum_{i=1}^{k} K_{ii} d_i^* - \sum_{i=1}^{k} K_{ii} d_i^* d_i^* - \sum_{i,j=1}^{k} K_{ij} d_i^* d_j^* \right). \tag{3.60}$$

Para o caso em que $i = j$, tem-se que $\sum_{i=1}^{k} K_{ij} d_i^* d_j^* = \sum_{i=1}^{k} K_{ii} d_i^* d_i^*$. Logo, a Equação (3.60), se reduz à:

$$S(\boldsymbol{x}, m) = \frac{\left( \sum_{i=1}^{k} K_{ii} d_i^* - \sum_{i,j=1}^{k} K_{ij} d_i^* d_j^* \right)}{m+k+2}. \tag{3.61}$$

As demonstrações para se chegar nessas expressões serão apresentadas no capítulo dos Resultados. O resultado geral para a função perda quadrática é dado em Owen (1970) e o vetor de médias e a matriz dispersão da distribuição de Dirichlet é dado por Wilks (1962).

Com o propósito de estimar as probabilidades desconhecidas $p_i$, em que $i = 1,2,...,k$, a partir de uma amostragem sequencial, onde cada observação é feita uma após a outra até tomar uma decisão de interromper a amostragem e estimar os parâmetros, utilizou-se as equações de programação dinâmica, pois elas fornecem uma decisão ótima em cada ponto. A partir dessas equações obtém-se o critério de parada para a distribuição multinomial, como pode ser visto a seguir.

Para isso, considerou-se um ponto $(\boldsymbol{x},m) = (x_1,x_2,\cdots,x_k,m)$, e seja $c$ o custo de amostragem de uma observação e $B(\boldsymbol{x},m)$ o risco de fazer uma observação adicional a um custo $c$ prosseguindo de forma otimizada posteriormente (risco de continuação ou também risco esperado), e seja $D(\boldsymbol{x},m)$ o risco mínimo ou também conhecido como risco ideal, então as equações de programação dinâmica dando a partição em pontos de parada e continuação são (JONES, 1976; JONES; MADHI, 1988):

$$D(\boldsymbol{x},m) = min[S(\boldsymbol{x},m),B(\boldsymbol{x},m)], \tag{3.62}$$

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k}[D(\boldsymbol{x}+\boldsymbol{e}_i,m+1)d_i^*] + D(\boldsymbol{x},m+1)\left(1-\sum_{i=1}^{k}d_i^*\right), \tag{3.63}$$

para cada ponto no espaço inteiro $(k+1)$, existem $(k+1)$ transições possíveis, $(\boldsymbol{x}+\boldsymbol{e}_i,m+1)$, com probabilidade $d_i^*$, em que $\boldsymbol{e_i}$ é o vetor linha com 1 na $i$-ésima posição e zeros nas outras posições: $\boldsymbol{e_i} = (0,\cdots,1,0,0)$. Estas equações são semelhantes àquelas em Freeman (1972), Jones (1974) e Lindley (1964).

Logo, tem-se que $S(\boldsymbol{x},m)$ é o risco imediato, $B(\boldsymbol{x},m)$ é o risco esperado e $D(\boldsymbol{x},m)$ é o risco ideal, que é dado pelo mínimo entre o imediato e o esperado. O risco ideal será igual ao esperado quando a decisão for continuar a amostragem, caso contrário, quando a decisão for parar a amostragem ele será igual ao imediato.

Uma vez que $S(\boldsymbol{x},m) \to 0$ e $B(\boldsymbol{x},m) \to c$ como $m \to \infty$, haverá um grande valor de $m = N^*$, de modo que todos os pontos $(\boldsymbol{x},N^*)$ sejam pontos de parada para $\sum_{i=1}^{k} x_i \leq N^*$. As equações de

programação dinâmica (3.62) e (3.63) podem agora ser usadas sucessivamente para $m \leq N^*$, para encontrar o menor inteiro $m$ que satisfaz (JONES, 1976):

$$B(\boldsymbol{x},m) > S(\boldsymbol{x},m), \quad D(\boldsymbol{x},m+1) = S(\boldsymbol{x},m+1) \text{ para todo } \boldsymbol{x}, \tag{3.64}$$

e este $m$ será o tamanho máximo da amostra. Assim, caracterizando o procedimento de inspeção sequencial conhecido como "*one-step look ahead*", em que a inspeção termina no menor inteiro $m$ que satisfaz as duas condições em (3.64).

Portanto, para encontrar a expressão do risco esperado $B(\boldsymbol{x},m)$ que estabelece o critério de parada, para *prioris* uniformes, tem-se que:

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k} [D(\boldsymbol{x}+\boldsymbol{e}_i, m+1) d_i^*] + D(\boldsymbol{x},m+1) \left(1 - \sum_{i=1}^{k} d_i^*\right)$$

$$B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m) \left(\frac{m+k+1}{m+k+2}\right). \tag{3.65}$$

Logo, a Equação (3.65) é o risco esperado ao utilizar *prioris* uniformes.

Portanto, o critério de parada se resume em comparar os valores dos riscos imediatos e esperados para cada observação. Quando o risco imediato for menor que o esperado, ou seja $S(\boldsymbol{x},m) < B(\boldsymbol{x},m)$, a amostragem para e estima-se os parâmetros de interesse. Caso contrário, se $S(\boldsymbol{x},m) > B(\boldsymbol{x},m)$ a amostragem continua, realizando mais uma observação até que possa tomar uma decisão.

## 3.8 Resumo das expressões

Como uma forma de resumir e apresentar as principais expressões para as funções de distribuição de probabilidades, média e variância/covariância, respectivamente, construiu-se os esquemas abaixo:

**Para a distribuição binomial:**

| | **Distribuição** | | |
|---|---|---|---|
| | **Binomial** | **Priori Beta** | **Posteriori Beta** |
| **Função** | $\binom{n}{x} p^x (1-p)^{n-x}$ | $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$ | $\frac{\Gamma(a+b+n)}{\Gamma(a+x)\Gamma(b+n-x)} p^{a+x-1}(1-p)^{b+n-x-1}$ |
| **Média** | $np$ | $\frac{a}{a+b}$ | $\frac{a+x}{(a+b+n)}$ |
| **Variância** | $np(1-p)$ | $\frac{ab}{(a+b)^2(a+b+1)}$ | $\frac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)}$ |

**Para a distribuição multinomial:**

| | **Distribuição** | | |
|---|---|---|---|
| | **Multinomial** | **Priori de Dirichlet** | **Posteriori de Dirichlet** |
| **Função** | $m! \prod_{i=1}^{k} \frac{p_i^{x_i}}{x_i!}$ | $\frac{\Gamma(a_0)}{\prod_{i=1}^{k}\Gamma(a_i)} \prod_{i=1}^{k} p_i^{a_i-1}$ | $\frac{\Gamma(\sum(x_i+a_i)}{\prod_{i=1}^{k}\Gamma(x_i+a_i)} \prod_{i=1}^{k} p_i^{x_i+a_i-1}$ |
| **Média** | $mp_i$ | $\frac{a_i}{a_0}$ | $\frac{x_i+a_i}{\sum_{i=1}^{k}(x_i+a_i)}$ |
| **Variância** | $mp_i(1-p_i)$ | $\frac{a_i(a_0-a_i)}{a_0^2(a_0+1)}$ | $\frac{(x_i+a_i)[(\sum_{i=1}^{k} x_i+a_i)-(x_i+a_i)]}{[\sum_{i=1}^{k}(x_i+a_i)]^2[(\sum_{i=1}^{k} x_i+a_i)+1]}$ |
| **Covariância** | $-mp_ip_j$ | $\frac{a_ia_j}{a_0^2(a_0+1)}$ | $\frac{-(x_i+a_i)(x_j+a_j)}{[\sum_{i=1}^{k}(x_i+a_i)]^2\{[\sum_{i=1}^{k}(x_i+a_i)]+1\}}$ |

Os riscos envolvidos no processo de estimação sequencial bayesiana e o respectivo estimador da proporção encontrado para as distribuições binomial e multinomial:

| | **Binomial** |
|---|---|
| **Risco Imediato** | $\dfrac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)}+C(n)$ |
| **Risco Esperado** | $\left(\dfrac{a+b+n}{a+b+n+1}\right)\dfrac{(a+x)(b+n-x)}{(a+b+n)^2(a+b+n+1)}+C(n+1)$ |
| **Estimador proporção** | $\dfrac{a+x}{(a+b+n)}$ |

| | **Multinomial** | |
|---|---|---|
| | ***Priori* uniforme** | **Qualquer *priori*** |
| **Risco Imediato** | $\dfrac{\left\{\sum_{i=1}^{k} K_{ii}d_i^* - \sum_{i,j=1}^{k} K_{ij}d_i^* d_j^*\right\}}{m+k+2}$ | $\dfrac{\sum_{i=1}^{k} K_{ii}(x_i+a_i)\left(\sum_{i=1}^{k}(x_i+a_i)\right) - \sum_{i,j=1}^{k} K_{ij}(x_i+a_i)(x_j+a_j)}{[\sum_{i=1}^{k}(x_i+a_i)]^2\{[\sum_{i=1}^{k}(x_i+a_i)]+1\}}$ |
| **Risco Esperado** | $c+\left[\dfrac{m+k+1}{m+k+2}\right]S(\boldsymbol{x},m)$ | $c+S(\boldsymbol{x},m)\left(\dfrac{a_0+m}{a_0+m+1}\right)$ |
| **Estimador proporção** | $\dfrac{a_i+x_i}{a_0+m}$ | $\dfrac{a_i+x_i}{a_0+m}$ |

## 3.9 Fluxograma do procedimento

Construiu-se um fluxograma para resumir as etapas do procedimento de estimação sequencial bayesiana, como poder ser visto na Figura 3.6:

Figura 3.6 – Fluxograma do procedimento de estimação sequencial bayesiana


Fonte: As autoras (2021)

## 3.10 Aplicação a dados de contagem: Uso em teste de raios X

O teste de raios X é uma técnica importante para a avaliação da qualidade de sementes. Segundo Brasil (2009), o objetivo desse teste é determinar a proporção de sementes cheias, vazias, danificadas por insetos e danificadas mecanicamente, pelas características morfológicas evidenciadas na radiografia.

Esse teste cria um arquivo fotográfico das estruturas internas das sementes analisadas e se destaca por ser considerado um método rápido e preciso de controle de qualidade de sementes. O procedimento consiste na absorção de raios X em diferentes níveis pelos distintos tecidos das sementes, que é determinado pela espessura, densidade, composição e comprimento de onda da radiação ionizante (BRASIL, 2009).

As sementes são colocadas entre uma fonte de raios X de baixa energia e um filme ou papel fotossensível. Ao atravessar as sementes, um feixe de raios X cria uma imagem permanente sobre o filme ou o papel. Após o processamento desse filme ou papel, forma-se uma imagem visível, de sombras claras e escuras (BRASIL, 2009). A imagem pode apresentar áreas mais claras que representam as partes mais densas da semente, e áreas mais escuras que correspondem àquelas partes em que os raios X são absorvidos mais facilmente (ISTA, 1985).

A análise das sementes pela técnica de raios X permite verificar a ocorrência de alterações morfológicas internas por meio de raios X, e com isso permite a seleção de sementes sem danos para formação de lotes de melhor qualidade. Danos internos, independentemente da causa, afetam a viabilidade das sementes, com exceção daqueles danos de menores dimensões, distantes do eixo embrionário (CARVALHO; CARVALHO; DAVIDE, 2009).

Quando o embrião não apresenta nenhuma anormalidade, a semente germinada também não apresenta problemas de vigor. No entanto, injúrias nas estruturas internas da semente limitam sua viabilidade e podem reduzir seu vigor, produzindo plântulas fracas e susceptíveis às condições adversas. Dessa forma, a captura e o processamento de imagem radiografada permite o estabelecimento de relações entre integridade, morfologia e determinação da germinação das sementes (FERNANDES et al., 2016).

Os testes de raios X apresentam a vantagem de não alterar a viabilidade das sementes analisadas, permitindo que sejam semeadas para comparação com o teste de germinação, possibilitando o estudo da germinação em relação à imagem radiográfica. Esta metodologia de análise de sementes tem sido aprimorada para avaliação de outras espécies, visando identificar problemas não visíveis à ótica humana e sua relação com a germinação (AMARAL et al., 2011).

Neste sentido, o teste de raios X tem sido utilizado principalmente na análise da morfologia interna de sementes, possibilitando o estudo da relação entre a ocorrência de danos e o prejuízo causado à germinação, analisando a qualidade das sementes. Isso pode ser verificado em pesquisas realizadas na agricultura com sementes de soja (FORTI; CICERO; PINTO, 2010), milho (CICERO; JUNIOR, 2003), feijão (MONDO et al., 2009), tomate (BORGES et al., 2019), braquiária (JEROMINI et al., 2019), abóbora (CARVALHO et al., 2009), dentre outras espécies.

As informações gerais para a realização do teste de raios X para várias espécies estão prescritas nas Regras para Análise de Sementes (BRASIL, 2009; ISTA, 1985) e têm sido aprimoradas nos últimos anos.

### 3.10.1 Teste de raios X para sementes de milho

As empresas produtoras de sementes necessitam de um elevado e rígido controle interno de qualidade para atenderem às exigências dos produtores. Essas exigências se refletem principalmente nas culturas de maior expressão econômica no Brasil, como é o caso do milho, que apresenta desempenho crescente em produção e produtividade (CONAB, 2021).

Para Carvalho e Camargo (2003), o termo qualidade de sementes diz respeito a todos os aspectos físicos, fisiológicos, genéticos e sanitários, que, se avaliados de forma correta e integrada, permitem o conhecimento do valor real e do potencial de utilização de um lote de sementes.

A utilização das técnicas de análise de imagens, a partir do teste de raios X, para o controle de qualidade de sementes é muito eficaz, já que as sementes podem ser examinadas individualmente em imagens ampliadas e capazes de indicar, com detalhes, a área danificada, a localização e a extensão do dano (CICERO; JUNIOR, 2003).

Alguns exemplos em que o teste de raios X contribuíram para os resultados precisos, em especial para sementes de milho, cujo nome científico é *Zea mays* L., podem ser vistos em Cicero e Junior (2003), que utilizaram a referida técnica para estudar os efeitos dos danos mecânicos sobre o vigor de sementes de milho. Mondo e Cicero (2005), utilizaram para estudar os efeitos das diferentes posições da semente de milho na espiga, sobre a qualidade.

Junior e Cicero (2012) avaliaram a eficiência do teste de raios X na identificação de danos mecânicos em sementes de milho doce e determinaram sua relação com germinação e vigor. Girardin, Chavagnat e Bockstaller (1993) e Carvalho et al. (1999), ao trabalhar com sementes de milho com danos de estresse em pré-colheita, reforçaram a afirmação de que a análise de imagens é o melhor método para avaliar as características morfológicas internas da semente.

Assim, tem sido presenciada uma evolução favorável ao aperfeiçoamento de técnicas computadorizadas, mais sensíveis para a captação e mais precisas para o processamento e extração de informações úteis para a indústria de sementes, definindo-se uma amplificação de sensibilidade por vias digitais (TEIXEIRA; CICERO; NETO, 2003).

Dentre os principais tipos de danos encontrados nas sementes de milho que influenciam na sua qualidade física, pode-se destacar: danos físicos, danos causados por inseto e variação de densidade. Os danos físicos, ou ainda, conhecidos como mecânicos, é considerado um dos principais fatores que afetam a qualidade de sementes. Podem ocorrer como: quebraduras, trincas, abrasões, cortes ou pressões (BRANDÃO-JUNIOR et al., 1999). Durante a debulha, processamento e semeadura, as sementes sofrem impactos, abrasões e cortes, que além de provocarem danos imediatos e latentes, podem reduzir, sensivelmente, as qualidades física e fisiológica dos lotes (BORBA et al., 1996).

Os insetos que ocasionam danos nas sementes podem ser de campo ou de armazenamento. No caso das sementes de milho, os insetos-pragas que atacam as sementes no campo são as lagartas *Helicoverpa zea* e *Spodoptera frugiperda*, onde estas irão provocar danos nas sementes por conta do seu tipo de aparato bucal. Já as pragas de armazenamento podem ser classificadas em primárias, onde perfuram as sementes e alimentam-se do seu interior, podendo se desenvolverem no interior da semente e possibilitando a instalação de outros agentes de deterioração. Exemplo dessas pragas

são: *Sitophilus oryza*, *Sitophilus zeamais* e *Sitotroga cerearella* (MATRANGOLO; CRUZ; LÚCIA, 1997).

Os danos que causam variações de densidade nas sementes são àqueles que alteram a sua estrutura interna, ocasionando em sementes menos densas. Esses danos podem ser por deterioração de tecidos através de microorganismos, por má formação, entre outros. Isto é decorrente de condições adversas de ambiente encontradas no campo após a semente atingir o ponto de maturidade fisiológica, ou ainda por descontroles ocorridos no armazenamento. O milho é a gramínea mais sensível à variação na densidade de plantas (CRUZ et al., 2010).

Na Figura 3.7 tem-se exemplos das sementes de milhos com os principais tipos de danos.

Figura 3.7 – Tipos de danos em sementes de milho *(Zea mays L.)*


Fonte: Laboratório Central de Sementes do Departamento de Agricultura da UFLA

# 4 METODOLOGIA

No presente capítulo, toda estrutura metodológica utilizada para o desenvolvimento deste trabalho, em concordância com os objetivos propostos, é descrita nas seções a seguir:

## 4.1 Critério de parada para estimação sequencial bayesiana dos parâmetros da distribuição Multinomial

Como o grande desafio do trabalho parte do estabelecimento de um critério de parada para a estimação sequencial bayesiana dos parâmetros da distribuição multinomial, primeiramente estudou-se como realizar tal feito. Desse modo, baseou-se no artigo de Jones (1976), onde são apresentadas as expressões envolvidas no procedimento, como para o risco imediato e esperado, com base em programação dinâmica, mas que foram desenvolvidas apenas considerando *prioris* uniformes.

Assim, a primeira etapa do trabalho se resumiu em entender como Jones (1976) encontrou essas expressões, e generalizá-las para a utilização de qualquer outra *priori* conjugada de Dirichlet, a partir dos caminhos dados pelo autor.

Portanto, realizou-se a demonstração das expressões contidas em Jones (1976) para *prioris* uniformes, e a partir delas, conseguiu-se estabelecer expressões gerais, para qualquer outra *priori* conjugada de Dirichlet a ser utilizada, apresentando também suas respectivas demonstrações.

## 4.2 Aplicação da teoria a dados de controle de qualidade de sementes de milho

Depois da teoria para a estimação sequencial bayesiana consolidada, aplicou-se aos dados de sementes de milho submetidas ao teste de raios X para controle de qualidade, com o objetivo de verificar e discutir os resultados da técnica.

O teste de raios X para análise de sementes de milho foi conduzido no Laboratório de Análise de Sementes da Universidade Federal de Lavras, na cidade de Lavras- MG. As imagens radiográficas foram geradas por um aparelho Faxitron MX-20 (*Faxitron X-ray Corp. Wheeling*, IL, EUA) conectado a um computador e monitor. O equipamento foi configurado a 26 kV, e as

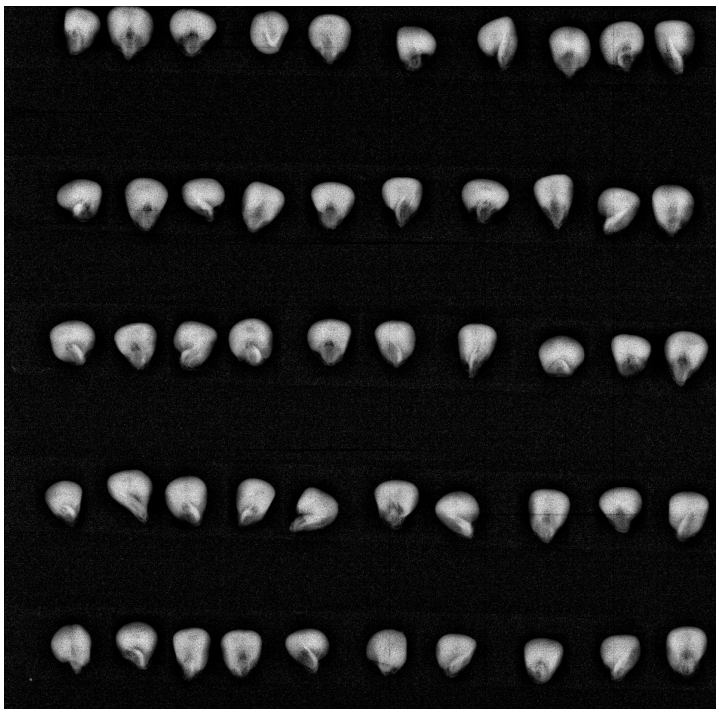
sementes foram expostas à radiação por 20 segundos. O contraste da imagem foi calibrado para 16.383 (largura) × 4849 (centro). Foram analisados 100 lotes de sementes de milho.

Foram utilizadas quatro repetições de 50 sementes para cada lote, totalizando em 200 sementes por lote, fixadas de forma ordenada em placas de acrílico (21x15 cm) com fita transparente dupla face devidamente identificadas com o número do lote, da repetição e a posição de cada semente para permitir a identificação individual em análises posteriores. Cada semente foi analisada individualmente.

O teste de raios X gerou radiografias digitais que foram analisadas visualmente para classificação quanto a presença ou não de danos, sendo considerada semente intacta aquelas sem nenhum tipo de dano, danos por insetos, danos físicos e com danos por variações de densidade.

Na Figura 4.1, tem-se um exemplo da imagem radiográfica gerada da repetição 1, contendo 50 sementes, do lote 100.

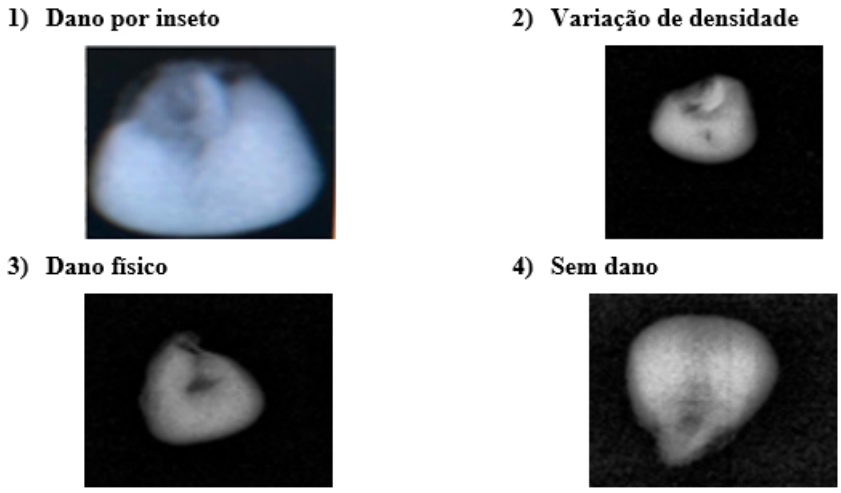
Figura 4.1 – Imagem radiográfica das sementes de milho, da repetição 1, do lote 100



Fonte: Laboratório Central de Sementes do Departamento de Agricultura da UFLA

Na Figura 4.2, tem-se um exemplo das classificações da sementes de acordo com a anatomia interna revelada pela radiografia, sendo: 1 - dano por inseto, 2 - dano por variação de densidade, 3 - dano físico e 4 - semente normal, sem nenhum tipo de dano.

Figura 4.2 – Imagem radiográfica das classificação das sementes de milho




Fonte: Laboratório Central de Sementes do Departamento de Agricultura da UFLA

No entanto, como o interesse é estimar as proporções, que são parâmetros das distribuições binomial e multinomial, organizou-se os mesmos dados de formas diferentes, do seguinte modo:

- **Etapa 1:** proporção de sementes sem danos e danificadas (2 classes). Portanto, utilizando para análise a distribuição binomial, considerando como sucesso a semente não possuir danos.
- **Etapa 2:** proporção de sementes com três classes de danos, utilizando portanto a distribuição multinomial para realizar as análises.
  - **Situação a:** Apenas com os três tipos de danos: por inseto, com variações de densidade e com danos físicos. Nesse caso, é importante salientar que os tamanhos amostrais dos lotes se diferiram entre si, pois considerou-se apenas as sementes danificadas.
  - **Situação b:** Sementes sem danos, com variações de densidade e com outros tipos de danos (danos por inseto + danos físicos).

Para a distribuição multinomial foi realizada duas organizações para testar qual teve um melhor desempenho no procedimento de estimação sequencial bayesiana.

Desse modo, estimou-se as proporções de interesse conforme as configurações realizadas com os dados, utilizando as distribuições binomial e multinomial, sob as abordagens frequentista, bayesiana e sequencial bayesiana.

## 4.3 Estimação Frequentista

Realizou-se a estimação das proporções de interesse, considerando a distribuição binomial e multinomial, a partir da abordagem frequentista utilizando o estimador de máxima verossimilhança, que é dado para a distribuição binomial e multinomial, respectivamente: $\hat{p} = \frac{X}{n}$ e $\hat{p}_i = \frac{X_i}{m}$.

Além disso, obteve-se a média e o desvio padrão das proporções dos 100 lotes. Os resultados foram apresentados em tabelas.

## 4.4 Estimação Bayesiana

Para estimar a proporção de sementes que pertencem as classes definidas para as distribuições binomial e multinomial sob a abordagem bayesiana, primeiramente precisou-se estabelecer os parâmetros das distribuições *a priori*, chamados de hiperparâmetros. Para a distribuição binomial, sabe-se que a *priori* conjugada é uma distribuição Beta com parâmetros $a$ e $b$. E para a distribuição multinomial *a priori* conjugada é uma Dirichlet, sendo uma generalização multivariada da distribuição Beta, com parâmetro $\boldsymbol{a}$.

Assim, considerou-se duas *prioris* para realizar o procedimento de estimação bayesiana para posterior discussão e comparação dos resultados. A primeira foi uma *priori* uniforme, onde tem-se que os valores dos hiperparâmetros são igualmente prováveis, logo serão iguais a 1. A outra *priori* foi advinda da literatura, construída com base nos resultados do artigo Javorski e Cicero (2017), que avaliaram os danos nas sementes de sorgo, gramínea como o milho, utilizando testes de raios X.

Os hiperparâmetros da *priori* baseada na literatura foram obtidos a partir dos valores de média e variância extraídos do artigo, e utilizando as expressões para a distribuição binomial:

$$a = \mu\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right), \quad b = (1-\mu)\left(\frac{\mu(1-\mu)}{\sigma^2} - 1\right). \tag{4.1}$$

E para a distribuição multinomial:

$$E(X_i) = \frac{a_i}{a_0}. \tag{4.2}$$

Desse modo, após observar as amostras, obteve-se as distribuições *a posteriori* Beta $(a + x, b + n - x)$ para os dados que foram organizados sob duas classes, e de Dirichlet $(a_i + x_i)$, para as organizações sob três classes, com o auxílio do software R (R Core Team, 2020).

A partir das distribuições *a posteriori* obteve-se as estimativas para as proporções de interesse, pois elas são dadas pela média da distribuição *a posteriori* obtida, calculadas conforme a expressão para a distribuição Beta:

$$E(p|X) = \frac{a + x}{a + b + n}. \tag{4.3}$$

E para a distribuição de Dirichlet:

$$E(\boldsymbol{p}|X) = \frac{x_i + a_i}{\sum_i^k (x_i + a_i)}. \tag{4.4}$$

Por fim, foram plotados os gráficos das densidades das *prioris* e *posterioris* Beta, utilizando o software R (R Core Team, 2020). E também foram plotados os simplex das *prioris* e *posterioris* de Dirichlet, utilizando o pacote *Compositional*, do software R (R Core Team, 2020), para melhor compreensão.

## 4.5 Estimação Sequencial Bayesiana

Para realizar a estimação do parâmetro proporção através do processo sequencial bayesiano, onde o tamanho de amostra é considerado como variável aleatória, as sementes de milho foram avaliadas uma a uma, até que toma-se a decisão de parar a amostragem e estimar o parâmetro, a partir de uma regra de parada definida.

Para isso, foram calculados o risco imediato e o risco esperado, assim como as médias das distribuições *a posteriori* obtidas, que são as estimativas das proporções. Além disso, foram testados diferentes valores de custo por observação: $10^{-4}$, $10^{-5}$ e $10^{-6}$, para entender o comportamento do processo de estimação sequencial bayesiana em relação aos dados analisados. Os custos foram considerados como constantes e aditivos na função de perda. Selecionou-se o melhor de acordo com os dados, de modo que este possua a mesma ordem de magnitude ou menor do que a da função de perda, e portanto, que não domine a função de risco.

Para a distribuição binomial, o procedimento foi realizado através do aplicativo em *Delphi*, desenvolvido por Resende et al. (2012). Esse aplicativo permite que seja definido a proporção média inicial, ou seja, a média da distribuição *a priori*, a variação e o custo. Como pode ser visto na Figura 4.3 a seguir:

Figura 4.3 – Interface do aplicativo em *Delphi*


Fonte: As autoras (2021)

Assim, com o intuito de utilizar duas *prioris*, a *priori* uniforme, sendo uma distribuição não informativa, e outra da literatura, sendo mais informativa, como foi realizado para a estimação bayesiana. Definiu-se a proporção inicial, a variação e o custo, com base nos valores calculados para a estimação bayesiana. No entanto, é importante salientar, que o aplicativo não permite utilizar por exemplo um valor exato para a variância e o custo, pois são definidos como baixo, médio e alto. Assim, terá uma distribuição *a priori* diferente da uniforme e da literatura.

Logo, para cada semente avaliada foi marcado como “Ausente” referente a semente que não possui dano, e “Presente” como a que possui dano, até que o risco imediato seja menor que o esperado e o programa pare, exibindo um relatório com os resultados e a estimava final para a proporção de sementes sem danos. Portanto, realizou-se esse processo para os 100 lotes, cuja estimativa final da proporção com o respectivo tamanho amostral foram apresentados em tabelas.

Para a distribuição multinomial, realizou-se todos os cálculos necessários para o procedimento de estimação sequencial bayesiana, como os riscos, médias *a posteriori*, para cada um dos 100 lotes, através de uma planilha dinâmica construída no Microsoft Excel ®.

Utilizou-se duas *prioris*, uma uniforme, onde os hiperparâmetros são iguais à um, logo são todos igualmente prováveis e assim a densidade de Dirichlet se reduz à uniforme, sendo uma *priori* não informativa. E outra *priori* cujos hiperparâmetros foram encontrados baseados na literatura, nas informações e resultados do artigo Javorski e Cicero (2017), sendo assim uma *priori* mais informativa. A expressão para encontrar os valores dos hiperparâmetros utilizada foi:

$$E(X_i) = \frac{a_i}{a_0}. \tag{4.5}$$

A cada nova observação os cálculos das estimativas *a posteriori* são atualizados utilizando os anteriores, ou seja, as estimativas da *priori*, logo, apenas para a primeira semente, tem-se uma *priori* uniforme, e portanto foram utilizadas as seguintes expressões para o cálculos dos riscos imediato e esperado, respectivamente, na primeira avaliação, baseadas em Jones (1976):

$$S(\boldsymbol{x},m) = \frac{1}{m+k+2}\left(\sum_{i=1}^{k} K_{ii}d_i^* - \sum_{i=1}^{k} K_{ii}d_i^*d_i^* - \sum_{i,j=1}^{k} K_{ij}d_i^*d_j^*\right). \tag{4.6}$$

$$B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m)\left(\frac{m+k+1}{m+k+2}\right). \tag{4.7}$$

A partir da segunda semente avaliada, as estimativas anteriores não são mais iguais à um, e portanto a *priori* para atualizar as informações não se resume em uma uniforme, e sim em uma Dirichlet, e assim utilizou-se as expressões gerais para os cálculos dos riscos imediato e esperado, respectivamente:

$$S(\boldsymbol{x},m) = \frac{\sum_{i=1}^{k} K_{ii}(x_i+a_i)\left(\sum_{i=1}^{k}(x_i+a_i)\right) - \sum_{i,j=1}^{k} K_{ij}(x_i+a_i)(x_j+a_j)}{[\sum_{i=1}^{k}(x_i+a_i)]^2\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]+1\right\}}. \tag{4.8}$$

$$B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m)\left(\frac{a_0+m}{a_0+m+1}\right). \tag{4.9}$$

Já para o caso em que a *priori* foi baseada na literatura, utilizou-se diretamente as expressões (4.8) e (4.9) para os cálculos dos riscos.

Portanto, realizou-se a estimação sequencial bayesiana utilizando as duas *prioris* obtidas, para as duas organizações de dados, uma considerando apenas os três tipos de danos, e a outra considerando as sementes sem danos, variações de densidade e outros tipos de danos, compreendendo portanto em uma distribuição multinomial com três classes para a análise em ambos os casos.

# 5 RESULTADOS E DISCUSSÃO

O presente capítulo apresenta todos os resultados obtidos deste trabalho, separados por seções, onde tem-se as demonstrações das expressões para o critério de parada da estimação sequencial bayesiana dos parâmetros da distribuição multinomial, e a aplicação aos dados de controle de qualidade de sementes de milho, obtendo a estimação dos parâmetros pelas três abordagens: frequentista, bayesiana e sequencial bayesiana.

## 5.1 Critério de parada para estimação sequencial bayesiana dos parâmetros da distribuição Multinomial

No artigo de Jones (1976), é apresentado as expressões envolvidas no procedimento de estimação sequencial bayesiana para os parâmetros da distribuição multinomial, como para o risco imediato e esperado, com base em programação dinâmica, mas que foram desenvolvidas apenas considerando *prioris* uniformes.

A demonstração das expressões contidas em Jones (1976) para *prioris* uniformes, para melhor entendê-las e generalizá-las para a utilização de qualquer outra *priori* conjugada de Dirichlet, estão a seguir:

Jones (1976), utilizou a *priori* uniforme para obter os limites de parada, como modo de facilitar os cálculos e encontrar os riscos imediato e esperado, já que esta é um caso particular da distribuição de Dirichlet, quando os parâmetros assumem o valor 1. Sendo assim, considerando uma *priori* uniforme, a matriz de dispersão da distribuição *posteriori* possuirá os elementos:

$$var(p_i) = \frac{d_i^*(1-d_i^*)}{m+k+2} \quad \text{e} \quad cov(p_i, p_j) = -\frac{d_i^* d_j^*}{m+k+2}. \tag{5.1}$$

Dado que a matriz $\boldsymbol{K} = \begin{bmatrix} K_{ii} & K_{ij} & \cdots & K_{ij} \\ K_{ij} & K_{ii} & \cdots & K_{ij} \\ \vdots & \vdots & \ddots & \vdots \\ K_{ij} & K_{ij} & \cdots & K_{ii} \end{bmatrix}$, é uma matriz simétrica positiva definida de perda constante, então sabe-se que $\boldsymbol{K}$ é uma matriz quadrada e $K_{ij} = K_{ji}$.

E seja a matriz de dispersão $\boldsymbol{\Sigma} = \begin{bmatrix} \frac{d_i^*(1-d_i^*)}{m+k+2} & \frac{-d_i^* d_j^*}{m+k+2} & \cdots & \frac{-d_i^* d_j^*}{m+k+2} \\ \frac{-d_i^* d_j^*}{m+k+2} & \frac{d_i^*(1-d_i^*)}{m+k+2} & \cdots & \frac{-d_i^* d_j^*}{m+k+2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-d_i^* d_j^*}{m+k+2} & \frac{-d_i^* d_j^*}{m+k+2} & \cdots & \frac{d_i^*(1-d_i^*)}{m+k+2} \end{bmatrix}$.

Logo,

$$S(\boldsymbol{x},m) = \text{traço } \boldsymbol{K\Sigma} = \text{traço} \begin{bmatrix} K_{ii} & K_{ij} & \cdots & K_{ij} \\ K_{ij} & K_{ii} & \cdots & K_{ij} \\ \vdots & \vdots & \ddots & \vdots \\ K_{ij} & K_{ij} & \cdots & K_{ii} \end{bmatrix} \begin{bmatrix} \frac{d_i^*(1-d_i^*)}{m+k+2} & \frac{-d_i^* d_j^*}{m+k+2} & \cdots & \frac{-d_i^* d_j^*}{m+k+2} \\ \frac{-d_i^* d_j^*}{m+k+2} & \frac{d_i^*(1-d_i^*)}{m+k+2} & \cdots & \frac{-d_i^* d_j^*}{m+k+2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{-d_i^* d_j^*}{m+k+2} & \frac{-d_i^* d_j^*}{m+k+2} & \cdots & \frac{d_i^*(1-d_i^*)}{m+k+2} \end{bmatrix} =$$

$$= \text{traço}$$

$$\begin{bmatrix} \frac{K_{ii}d_i^*(1-d_i^*)}{m+k+2} - \frac{K_{ij}d_i^*d_j^*}{m+k+2} + \cdots - \frac{K_{ij}d_i^*d_j^*}{m+k+2} & -\frac{K_{ii}d_i^*d_j^*}{m+k+2} + \frac{K_{ij}d_i^*(1-d_i^*)}{m+k+2} + \cdots - \frac{K_{ij}d_i^*d_j^*}{m+k+2} & \cdots & -\frac{K_{ii}d_i^*d_j^*}{m+k+2} - \frac{K_{ij}d_i^*d_j^*}{m+k+2} + \cdots + \frac{K_{ij}d_i^*(1-d_i^*)}{m+k+2} \\ \frac{K_{ij}d_i^*(1-d_i^*)}{m+k+2} - \frac{K_{ii}d_i^*d_j^*}{m+k+2} + \cdots - \frac{K_{ij}d_i^*d_j^*}{m+k+2} & -\frac{K_{ij}d_i^*d_j^*}{m+k+2} + \frac{K_{ii}d_i^*(1-d_i^*)}{m+k+2} + \cdots - \frac{K_{ij}d_i^*d_j^*}{m+k+2} & \cdots & -\frac{K_{ij}d_i^*d_j^*}{m+k+2} - \frac{K_{ii}d_i^*d_j^*}{m+k+2} + \cdots + \frac{K_{ij}d_i^*(1-d_i^*)}{m+k+2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{K_{ij}d_i^*(1-d_i^*)}{m+k+2} - \frac{K_{ij}d_i^*d_j^*}{m+k+2} + \cdots - \frac{K_{ii}d_i^*d_j^*}{m+k+2} & -\frac{K_{ij}d_i^*d_j^*}{m+k+2} + \frac{K_{ij}d_i^*(1-d_i^*)}{m+k+2} + \cdots - \frac{K_{ii}d_i^*d_j^*}{m+k+2} & \cdots & -\frac{K_{ij}d_i^*d_j^*}{m+k+2} - \frac{K_{ij}d_i^*d_j^*}{m+k+2} + \cdots + \frac{K_{ii}d_i^*(1-d_i^*)}{m+k+2} \end{bmatrix} =$$

$$= \frac{K_{ii}d_i^*(1-d_i^*)}{m+k+2} - \frac{K_{ij}d_i^*d_j^*}{m+k+2} + \cdots - \frac{K_{ij}d_i^*d_j^*}{m+k+2} - \frac{K_{ij}d_i^*d_j^*}{m+k+2} + \frac{K_{ii}d_i^*(1-d_i^*)}{m+k+2} + \cdots - \frac{K_{ij}d_i^*d_j^*}{m+k+2} -$$

$$- \frac{K_{ij}d_i^*d_j^*}{m+k+2} - \frac{K_{ij}d_i^*d_j^*}{m+k+2} + \cdots + \frac{K_{ii}d_i^*(1-d_i^*)}{m+k+2} =$$

$$= \frac{K_{ii}d_i^* - K_{ii}d_i^*d_i^* - K_{ij}d_i^*d_j^* + \cdots - K_{ij}d_i^*d_j^* - K_{ij}d_i^*d_j^* + K_{ii}d_i^* - K_{ii}d_i^*d_i^* + \cdots - K_{ij}d_i^*d_j^*}{m+k+2} -$$

$$\frac{-K_{ij}d_i^*d_j^* - K_{ij}d_i^*d_j^* + \cdots + K_{ii}d_i^* - K_{ii}d_i^*d_i^*}{m+k+2} =$$

$$= \frac{1}{m+k+2}\left(\sum_{i=1}^{k} K_{ii}d_i^* - \sum_{i=1}^{k} K_{ii}d_i^*d_i^* - \sum_{i,j=1}^{k} K_{ij}d_i^*d_j^*\right). \tag{5.2}$$

Para o caso em que $i = j$, tem-se que $\sum_{i=1}^{k} K_{ij}d_i^*d_j^* = \sum_{i=1}^{k} K_{ii}d_i^*d_i^*$. Logo, a Equação (5.2), se reduz à:

$$S(\boldsymbol{x},m) = \frac{\left(\sum_{i=1}^{k} K_{ii}d_i^* - \sum_{i,j=1}^{k} K_{ij}d_i^*d_j^*\right)}{m+k+2}. \tag{5.3}$$

Com o propósito de estimar as probabilidades desconhecidas $p_i$, em que $i = 1,2,...,k$, a partir de uma amostragem sequencial, utilizou-se as equações de programação dinâmica, pois elas fornecem uma decisão ótima em cada ponto. A partir dessas equações obtém-se o critério de parada para a distribuição multinomial, como pode ser visto a seguir.

Para isso, considerou-se um ponto $(\boldsymbol{x},m) = (x_1,x_2,\cdots,x_k,m)$, e seja $c$ o custo de amostragem de uma observação e $B(\boldsymbol{x},m)$ o risco de fazer uma observação adicional a um custo $c$ prosseguindo de forma otimizada posteriormente (risco de continuação ou também risco esperado), e seja $D(\boldsymbol{x},m)$ o risco mínimo ou também conhecido como risco ideal, então as equações de programação dinâmica dando a partição em pontos de parada e continuação são (JONES, 1976; JONES; MADHI, 1988):

$$D(\boldsymbol{x},m) = min[S(\boldsymbol{x},m),B(\boldsymbol{x},m)], \tag{5.4}$$

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k}[D(\boldsymbol{x}+\boldsymbol{e}_i,m+1)d_i^*] + D(\boldsymbol{x},m+1)\left(1 - \sum_{i=1}^{k} d_i^*\right), \tag{5.5}$$

para cada ponto no espaço inteiro $(k+1)$, existem $(k+1)$ transições possíveis, $(\boldsymbol{x}+\boldsymbol{e}_i,m+1)$, com probabilidade $d_i^*$, em que $\boldsymbol{e_i}$ é o vetor linha com 1 na $i$-ésima posição e zeros nas outras posições: $\boldsymbol{e_i} = (0,\cdots,1,0,0)$.

Logo, tem-se que $S(\boldsymbol{x},m)$ é o risco imediato, $B(\boldsymbol{x},m)$ é o risco esperado e $D(\boldsymbol{x},m)$ é o risco ideal, que é dado pelo mínimo entre o imediato e o esperado. O risco ideal será igual ao

esperado quando a decisão for continuar a amostragem, caso contrário, quando a decisão for parar a amostragem ele será igual ao imediato.

Uma vez que $S(\boldsymbol{x},m) \to 0$ e $B(\boldsymbol{x},m) \to c$ como $m \to \infty$, haverá um grande valor de $m = N^*$, de modo que todos os pontos $(\boldsymbol{x},N^*)$ sejam pontos de parada para $\sum_{i=1}^{k} x_i \leq N^*$. As equações de programação dinâmica (5.4) e (5.5) podem agora ser usadas sucessivamente para $m \leq N^*$, para encontrar o menor inteiro $m$ que satisfaz (JONES, 1976):

$$B(\boldsymbol{x},m) > S(\boldsymbol{x},m), \quad D(\boldsymbol{x},m+1) = S(\boldsymbol{x},m+1) \text{ para todo } \boldsymbol{x}, \tag{5.6}$$

e este $m$ será o tamanho máximo da amostra. Assim, caracterizando o procedimento de inspeção sequencial conhecido como "*one-step look ahead*", em que a inspeção termina no menor inteiro $m$ que satisfaz as duas condições em (5.6).

Portanto, para encontrar a expressão do risco esperado $B(\boldsymbol{x},m)$ que estabelece o critério de parada, para *prioris* uniformes, tem-se que:

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k} [D(\boldsymbol{x}+\boldsymbol{e}_i, m+1) d_i^*] + D(\boldsymbol{x},m+1)\left(1 - \sum_{i=1}^{k} d_i^*\right)$$

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k} \left[D(\boldsymbol{x}+\boldsymbol{e}_i, m+1)\left(\frac{x_i+1}{m+k+1}\right)\right] + D(\boldsymbol{x},m+1)\left(1 - \sum_{i=1}^{k} \frac{x_i+1}{m+k+1}\right)$$

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k} \left[D(\boldsymbol{x}+\boldsymbol{e}_i, m+1)\left(\frac{x_i+1}{m+k+1}\right)\right] + D(\boldsymbol{x},m+1)\left(\frac{m+k+1-\sum_{i=1}^{k}(x_i+1)}{m+k+1}\right)$$

Utilizando o fato de que quando $S(\boldsymbol{x},m) < B(\boldsymbol{x},m)$ tem-se a decisão de parar a análise e assim $D(\boldsymbol{x},m+1) = S(\boldsymbol{x},m+1)$, já que o interesse é encontrar $B(\boldsymbol{x},m)$ para o qual tem-se a regra de parada. Então, pode-se substituir $D(\boldsymbol{x},m+1)$ por $S(\boldsymbol{x},m+1)$ :

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k} \left[S(\boldsymbol{x}+\boldsymbol{e}_i, m+1)\left(\frac{x_i+1}{m+k+1}\right)\right] + S(\boldsymbol{x},m+1)\left(\frac{m+k+1-\sum_{i=1}^{k}(x_i+1)}{m+k+1}\right)$$

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k} \left[S(\boldsymbol{x},m)\left(\frac{x_i+1}{(m+1)+k+1}\right)\right] + S(\boldsymbol{x},m)\left(\frac{m+k+1-\sum_{i=1}^{k}(x_i+1)}{(m+1)+k+1}\right)$$

$$B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m)\left\{\left[\sum_{i=1}^{k}\left(\frac{x_i+1}{(m+1)+k+1}\right)\right] + \left(\frac{m+k+1-\sum_{i=1}^{k}(x_i+1)}{(m+1)+k+1}\right)\right\}$$

$$B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m)\left\{\left(\frac{\sum_{i=1}^{k}x_i+k}{(m+1)+k+1}\right) + \left(\frac{m+k+1-\sum_{i=1}^{k}x_i-k}{(m+1)+k+1}\right)\right\}$$

$$B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m)\left\{\left(\frac{\cancel{\sum_{i=1}^{k}x_i}+\cancel{k}+m+k+1-\cancel{\sum_{i=1}^{k}x_i}-\cancel{k}}{(m+1)+k+1}\right)\right\}$$

$$B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m)\left(\frac{m+k+1}{m+k+2}\right). \tag{5.7}$$

Logo, a Equação (5.7) é o risco esperado ao utilizar *prioris* uniformes.

Portanto, o critério de parada se resume em comparar os valores dos riscos imediatos e esperados para cada observação. E quando $S(\boldsymbol{x},m) < B(\boldsymbol{x},m)$, a amostragem para e estima-se os parâmetros de interesse. Caso contrário, se $S(\boldsymbol{x},m) > B(\boldsymbol{x},m)$ a amostragem continua, realizando mais uma observação até que possa tomar uma decisão.

A partir das demonstrações anteriores e dos caminhos dados por Jones (1976), conseguiu-se estabelecer expressões gerais, para qualquer outra *priori* a ser utilizada, cujas demonstrações estão apresentadas a seguir:

Utilizando a expressão geral para qualquer *priori* conjugada de Dirichlet, não restringindo apenas a *priori* uniforme para encontrar o risco imediato, tem-se que a matriz de dispersão da distribuição *posteriori* possuirá os elementos:

$$Var(\boldsymbol{p}|X_i) = \frac{(x_i+a_i)\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]-(x_i+a_i)\right\}}{\left[\sum_{i=1}^{k}(x_i+a_i)\right]^2\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]+1\right\}}, \tag{5.8}$$

$$Cov(\boldsymbol{p}|X_i,X_j) = \frac{-(x_i+a_i)(x_j+a_j)}{\left[\sum_{i=1}^{k}(x_i+a_i)\right]^2\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]+1\right\}}. \tag{5.9}$$

$$\text{Logo, } \boldsymbol{\Sigma} = \begin{bmatrix} \frac{(x_i+a_i)\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]-(x_i+a_i)\right\}}{\left[\sum_{i=1}^{k}(x_i+a_i)\right]^2\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]+1\right\}} & \cdots & \frac{-(x_i+a_i)(x_j+a_j)}{\left[\sum_{i=1}^{k}(x_i+a_i)\right]^2\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]+1\right\}} \\ \vdots & \ddots & \vdots \\ \frac{-(x_i+a_i)(x_j+a_j)}{\left[\sum_{i=1}^{k}(x_i+a_i)\right]^2\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]+1\right\}} & \cdots & \frac{(x_i+a_i)\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]-(x_i+a_i)\right\}}{\left[\sum_{i=1}^{k}(x_i+a_i)\right]^2\left\{\left[\sum_{i=1}^{k}(x_i+a_i)\right]+1\right\}} \end{bmatrix}.$$

Assim, tem-se que: $S(\boldsymbol{x},m) = \text{traço}\ \boldsymbol{K\Sigma} =$

$$=\text{traço}\begin{bmatrix} K_{ii} & \cdots & K_{ij} \\ \vdots & \ddots & \vdots \\ K_{ij} & \cdots & K_{ii} \end{bmatrix}\begin{bmatrix} \frac{(x_i+a_i)\{[\sum_{i=1}^k (x_i+a_i)]-(x_i+a_i)\}}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}} & \cdots & \frac{-(x_i+a_i)(x_j+\alpha_j)}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}} \\ \vdots & \ddots & \vdots \\ \frac{-(x_i+a_i)(x_j+a_j)}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}} & \cdots & \frac{(x_i+a_i)\{[\sum_{i=1}^k (x_i+a_i)]-(x_i+a_i)\}}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}} \end{bmatrix} =$$

$$= \frac{K_{ii}(x_i+a_i)\{[\sum_{i=1}^k (x_i+a_i)]-(x_i+a_i)\}}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}} + \cdots - \frac{K_{ij}(x_i+a_i)(x_j+a_j)}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}} + \cdots -$$

$$- \frac{K_{ij}(x_i+a_i)(x_j+a_j)}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}} + \cdots + \frac{K_{ii}(x_i+a_i)\{[\sum_{i=1}^k (x_i+a_i)]-(x_i+a_i)\}}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}} =$$

$$= \frac{K_{ii}\left[(x_i+a_i)\{[\sum_{i=1}^k (x_i+a_i)]-(x_i+a_i)\}\right] + \cdots - K_{ij}\left[(x_i+a_i)(x_j+a_j)\right] + \cdots -}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}}$$

$$\frac{-K_{ij}\left[(x_i+a_i)(x_j+a_j)\right] + \cdots + K_{ii}\left[(x_i+a_i)\{[\sum_{i=1}^k (x_i+a_i)]-(x_i+a_i)\}\right]}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}} =$$

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

$$= \frac{\sum_{i=1}^k K_{ii}(x_i+a_i)\left(\sum_{i=1}^k (x_i+a_i)\right) - \sum_{i=1}^k K_{ii}(x_i+a_i)(x_i+a_i) - \sum_{i,j=1}^k K_{ij}(x_i+a_i)(x_j+a_j)}{[\sum_{i=1}^k (x_i+a_i)]^2\{[\sum_{i=1}^k (x_i+a_i)]+1\}}. \tag{5.10}$$

Para o caso em que $i = j$, tem-se que $\sum_{i,j=1}^{k} K_{ij}(x_i+a_i)(x_j+a_j) = \sum_{i=1}^{k} K_{ii}(x_i+a_i)(x_i+a_i)$. Logo, a Equação (5.10), se reduz à:

$$\boxed{S(\boldsymbol{x},m) = \frac{\sum_{i=1}^{k} K_{ii}(x_i+a_i)\left(\sum_{i=1}^{k}(x_i+a_i)\right) - \sum_{i,j=1}^{k} K_{ij}(x_i+a_i)(x_j+a_j)}{[\sum_{i=1}^{k}(x_i+a_i)]^2\{[\sum_{i=1}^{k}(x_i+a_i)]+1\}}.} \tag{5.11}$$

Portanto, tem-se que a expressão (5.11) é a expressão geral para o risco imediato para qualquer *priori* conjugada de Dirichlet, e não apenas uniformes.

Para encontrar a expressão do risco esperado $B(\boldsymbol{x},m)$ que estabelece o critério de parada, para qualuer *priori* Dirichlet, tem-se que:

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k}[D(\boldsymbol{x}+\boldsymbol{e}_i,m+1)d_i^*] + D(\boldsymbol{x},m+1)\left(1-\sum_{i=1}^{k} d_i^*\right)$$

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k}\left[D(\boldsymbol{x}+\boldsymbol{e}_i,m+1)\left(\frac{a_i+x_i}{a_0+m}\right)\right] + D(\boldsymbol{x},m+1)\left(1-\sum_{i=1}^{k}\frac{a_i+x_i}{a_0+m}\right)$$

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k}\left[D(\boldsymbol{x}+\boldsymbol{e}_i,m+1)\left(\frac{a_i+x_i}{a_0+m}\right)\right] + D(\boldsymbol{x},m+1)\left(\frac{a_0+m-\sum_{i=1}^{k}a_i-\sum_{i=1}^{k}x_i}{a_0+m}\right)$$

Utilizando o fato de que quando $S(\boldsymbol{x},m) < B(\boldsymbol{x},m)$ tem-se a decisão de parar a análise e assim $D(\boldsymbol{x},m+1) = S(\boldsymbol{x},m+1)$, já que o interesse é encontrar $B(\boldsymbol{x},m)$ para o qual tem-se a regra de parada. Então, pode-se substituir $D(\boldsymbol{x},m+1)$ por $S(\boldsymbol{x},m+1)$ :

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k}\left[S(\boldsymbol{x}+\boldsymbol{e}_i,m+1)\left(\frac{a_i+x_i}{a_0+m}\right)\right] + S(\boldsymbol{x},m+1)\left(\frac{a_0+m-\sum_{i=1}^{k}a_i-\sum_{i=1}^{k}x_i}{a_0+m}\right)$$

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k}\left[S(\boldsymbol{x},m)\left(\frac{a_i+x_i}{a_0+(m+1)}\right)\right] + S(\boldsymbol{x},m)\left(\frac{a_0+m-\sum_{i=1}^{k}a_i-\sum_{i=1}^{k}x_i}{a_0+(m+1)}\right)$$

$$B(\boldsymbol{x},m) = c + \frac{S(\boldsymbol{x},m)}{a_0+(m+1)}\left(\cancel{\sum_{i=1}^{k}a_i} + \cancel{\sum_{i=1}^{k}x_i} + a_0 + m - \cancel{\sum_{i=1}^{k}a_i} - \cancel{\sum_{i=1}^{k}x_i}\right)$$

$$B(\boldsymbol{x},m) = c + \frac{S(\boldsymbol{x},m)}{a_0+(m+1)}(a_0+m).$$

Logo, a expressão geral do risco esperado para qualquer *priori* conjugada Dirichlet, e não somente uniforme, é dada por (5.12):

$$\boxed{B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m)\left(\frac{a_0+m}{a_0+m+1}\right).} \tag{5.12}$$

Como uma forma de melhor compreender o procedimento, foi realizado um exemplo, logo:

- **1º passo:** Determinar *priori* e custo:
  - *Priori* uniforme $\Rightarrow a_1 = 1$, $a_2 = 1$ e $a_3 = 1$.
  - Custo = $0,00001 = 10^{-5}$.
- **2º passo:** Realiza-se a primeira observação. (Vamos considerar lote 1, apenas sementes danificadas) $\Rightarrow x_1$ : por inseto, $x_2$ : variação de densidade e $x_3$ : físico.
  - Tem-se uma semente com dano por variação de densidade $\Rightarrow x_1 = 0, x_2 = 1, x_3 = 0$.
- **3º passo:** Calcular a estimativa da proporção (média da distribuição *a posteriori*):

$$\boxed{d_i^* = \frac{(x_i+1)}{(m+k+1)}} \Rightarrow d_1^* = d_3^* = \frac{(0+1)}{(1+2+1)} = \frac{1}{4} = 0,25.$$

$$d_2^* = \frac{(1+1)}{(1+2+1)} = \frac{2}{4} = 0,5.$$

- **4º passo:** Calcular risco imediato:

$$\boxed{S(\boldsymbol{x},m) = \frac{1}{m+k+2}\left(\sum_{i=1}^{k} K_{ii}d_i^* - \sum_{i=1}^{k} K_{ii}d_i^*d_i^* - \sum_{i,j=1}^{k} K_{ij}d_i^*d_j^*\right) =}$$

$$= \frac{10}{80} = 0,125.$$

- **5º passo:** Calcular risco esperado:

$$\boxed{B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m)\left(\frac{m+k+1}{m+k+2}\right)} = 0,10001.$$

- Agora sem a fórmula, para compreender melhor a programação dinâmica:

$$B(\boldsymbol{x},m) = c + \sum_{i=1}^{k} [D(\boldsymbol{x}+\boldsymbol{e}_i, m+1)d_i^*] + D(\boldsymbol{x}, m+1)\left(1 - \sum_{i=1}^{k} d_i^*\right).$$

  - Para a primeira observação ocorreu: $(0,1,0,1)$.
  - Para a segunda observação poderá ocorrer: $(1,1,0,2)$, $(0,2,0,2)$, $(0,1,1,2)$. Então:

$$B(\boldsymbol{x},m) = 0,00001 + [D(1,1,0,2)d_1^* + D(0,2,0,2)d_2^*] + D(0,1,1,2)\left[1 - (d_1^* + d_2^*)\right].$$

- Como o interesse é em B para o qual o procedimento para, então D=S:

$$B(\boldsymbol{x},m) = 0,00001 + [S(1,1,0,2)d_1^* + S(0,2,0,2)d_2^*] + S(0,1,1,2)\left[1 - (d_1^* + d_2^*)\right].$$

- Portanto, calcula-se os S:
  $S(1,1,0,2) = \dfrac{16}{150}$, $S(0,2,0,2) = \dfrac{14}{150}$ e $S(0,1,1,2) = \dfrac{16}{150}$.
- Logo,

$$B(\boldsymbol{x},m) = 0,00001 + \frac{16}{150}0,25 + \frac{14}{150}0,5 + \frac{16}{250}0,25 = 0,10001.$$

- **6º passo:** Comparar risco imediato com esperado:
  - Como, $0,125 > 0,10001 \Rightarrow \boxed{Continua}$
- Realiza-se uma nova observação:
  - *Priori* será a estimativa anterior $\Rightarrow a_1 = 0,25$, $a_2 = 0,5$ e $a_3 = 0,25$.
  - Semente com dano por variação novamente $\Rightarrow x_1 = 0, x_2 = 2, x_3 = 0$.
  - Calcula-se novamente a estimativa da proporção, risco imediato e esperado, até que o risco imediato seja menor que o esperado.
- Estimativa da proporção:

$$\boxed{d_i^* = \frac{a_i + x_i}{a_0 + m}}$$

- Risco imediato:

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

- Risco esperado:

$$\boxed{B(\boldsymbol{x},m) = c + S(\boldsymbol{x},m)\left(\frac{a_0+m}{a_0+m+1}\right)}$$

## 5.2 Estimação Frequentista

### 5.2.1 Estimação frequentista da proporção de sementes com danos e sem danos utilizando a distribuição Binomial

Na abordagem frequentista, a estimação da proporção de sementes com danos e sem danos é feita utilizando-se a distribuição binomial, já que existem duas classes. Desse modo, primeiramente estimou-se a proporção das sementes sem danos e das sementes danificadas de cada um dos 100 lotes, utilizando o estimador de máxima verossimilhança para o parâmetro $p$, dado por $\hat{p} = \frac{X}{n}$, como pode ser visto na Tabela 5.1:

Tabela 5.1 – Estimativas frequentistas das proporções de sementes sem danos e danificadas considerando a distribuição binomial

| **Lote** | **Sementes sem danos (%)** | **Sementes danificadas (%)** |
|---|---|---|
| **1** | 86,50 | 13,50 |
| **2** | 87,50 | 12,50 |
| **3** | 78,50 | 21,50 |
| **4** | 86,00 | 14,00 |
| ⋮ | ⋮ | ⋮ |
| **98** | 86,00 | 14,00 |
| **99** | 77,00 | 23,00 |
| **100** | 82,50 | 17,50 |
| **Média** (%) | 83,82 | 16,18 |
| **Desvio padrão** (%) | 5,18 | 5,18 |

Fonte: As autoras (2021)

Cada lote possui 200 sementes, e de acordo com a Tabela 5.1, pode-se afirmar que a média de sementes sem danos nos lotes avaliados foi de $83,82\%$, e de sementes danificadas foi de $16,18\%$, com desvio-padrão de $5,18\%$. Este é o modo pelo qual é realizado a estimação das proporções de interesse ao realizar um teste de raios X, utilizando um total de 200 sementes para o resultado de cada lote.

### 5.2.2 Estimação frequentista da proporção de sementes sob três classes utilizando a distribuição Multinomial

No teste de raios X realizado verificou-se que as sementes de milho apresentaram três tipos de danos, sendo eles por inseto, variação de densidade e danos físicos. Como há mais de duas classificações para os danos, a estimação da proporção de sementes de cada categoria é feita utilizando a distribuição multinomial. Sendo assim, para cada um dos 100 lotes, inicialmente, estimou-se a proporção de sementes que pertencem a cada uma das classes: com danos por inseto, com variação de densidade e com danos físicos, utilizando o estimador de máxima verossimilhança para $\boldsymbol{p}$, dado por: $\hat{p}_i = \frac{X_i}{m}$, sendo o total de danos considerado 100%. Deve-se ressaltar que nestes casos, os tamanhos amostrais diferem entre os lotes, pois eles são apenas contabilizados nas sementes danificadas. Os resultados estão apresentados na Tabela 5.2:

Tabela 5.2 – Estimativas frequentistas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos

| **Lote** | **m** | $\hat{\boldsymbol{p}}_1$ **:** **Danos por inseto (%)** | $\hat{\boldsymbol{p}}_2$ **:** **Variações de densidade (%)** | $\hat{\boldsymbol{p}}_3$ **:** **Danos físicos (%)** |
|---|---|---|---|---|
| **1** | 27 | 33,33 | 66,67 | 0,00 |
| **2** | 25 | 16,00 | 80,00 | 4,00 |
| **3** | 43 | 6,98 | 88,37 | 4,65 |
| **4** | 28 | 7,14 | 92,86 | 0,00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **98** | 28 | 35,71 | 39,29 | 25,00 |
| **99** | 46 | 19,57 | 65,22 | 15,22 |
| **100** | 35 | 8,57 | 68,57 | 22,86 |
| **Média (%)** | 32,36 | 17,76 | 69,61 | 12,63 |
| **Desvio padrão (%)** | 10,36 | 11,42 | 20,46 | 13,23 |

Fonte: As autoras (2021)

Resultou-se em um total de 3236 sementes danificadas nos 100 lotes, dessas, como pode ser observado pela Tabela 5.2, em média de $17,76\%$ apresentaram dano por inseto, com desvio padrão de $11,42\%$, enquanto que $69,61\%$ tiveram variações de densidade, com desvio padrão de $20,46\%$ e $12,63\%$ foram danificadas fisicamente, com desvio padrão de $13,23$.

Posteriormente realizou-se a estimação frequentista da proporção considerando outras três classes: as sementes sem danos, as sementes que possuem variações de densidade e as que possuem

outros tipos de danos (somando os danos por inseto e físicos). Cada um dos 100 lotes possui 200 sementes. Os resultados estão apresentados na Tabela 5.3:

Tabela 5.3 – Estimativas frequentistas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos

| **Lote** | $\hat{\boldsymbol{p}}_1$ **:** **Sem danos (%)** | $\hat{\boldsymbol{p}}_2$ **:** **Variações de densidade (%)** | $\hat{\boldsymbol{p}}_3$ **:** **Outros tipos de danos (%)** |
|---|---|---|---|
| **1** | 86,50 | 9,00 | 4,50 |
| **2** | 87,50 | 10,00 | 2,50 |
| **3** | 78,50 | 19,00 | 2,50 |
| **4** | 86,00 | 13,00 | 1,00 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| **98** | 86,00 | 5,50 | 8,50 |
| **99** | 77,00 | 15,00 | 8,00 |
| **100** | 82,50 | 12,00 | 5,50 |
| **Média (%)** | 83,82 | 11,38 | 4,80 |
| **Desvio padrão (%)** | 5,18 | 5,07 | 3,49 |

Fonte: As autoras (2021)

Desse modo, pode-se observar na Tabela 5.3, que a média de sementes sem danos de $83,82\%$ é a que sobressai, com desvio padrão de $5,18\%$, enquanto que as que possuem variações de densidade possuem média de $11,38\%$, com desvio padrão de $5,07\%$, e as que possuem outros tipos de danos têm média de $4,80\%$, com desvio padrão de $3,49\%$.

Os resultados obtidos estão de acordo com os dados encontrados na literatura, como pode ser visto em Bragachini et al. (1992), Borba et al. (1996), Sparks et al. (1966) e Javorski e Cicero (2017).

Segundo Bragachini et al. (1992), danos físicos acima de 20% são considerados excessivos, e pode-se observar pela Tabela 5.2 que em média, os lotes apresentaram $12,63\%$ danos físicos ao considerar apenas os danos, e $4,80\%$ ao estarem na classe de outros tipos de danos, evidenciando conformidade com a literatura, já que não possui excesso desse tipo de dano.

No trabalho realizado por Borba et al. (1996), concluíram que a germinação das sementes de milho não foi afetada logo após a debulha mecânica, mas após seis meses de armazenamento, tanto a germinação quanto o vigor foram significativamente reduzidos, quando ocorreram danos físicos na faixa de $5,6$ a $23,9\%$. Por outro lado, a germinação das sementes de milho de outra cultivar, não foi afetada quando ocorreram danos físicos, na faixa de $0,5$ a $10,3\%$, mas o vigor foi

significativamente reduzido. Desse modo, pode-se inferir que das sementes analisadas o vigor pode ser reduzido devido as taxas de danos encontradas.

O impacto dos danos por inseto nos lotes analisados não foi alto, pois para a situação a, a proporção média foi de $17,76\%$ e $4,80\%$ para a situação b. Na literatura, Sparks et al. (1966) observaram que a praga *P. rileyi* produziu danos em cerca de $2\%$ dos grãos de milho, em cinco estados dos EUA em 1962, não sendo também um número considerável.

Os danos que geraram maior impacto nos 100 lotes analisados, foram os que proporcionaram variações de densidade nas sementes, apresentando em média $69,61\%$ para a situação a, e $11,38\%$ para a situação b, assim como em Javorski e Cicero (2017). Esses danos podem ocorrer por má formação, por deterioração de tecidos através de microorganismos, entre outros. Em Javorski e Cicero (2017), o dano por deterioração de tecidos foi o maior responsável pela perda de germinação das sementes. Portanto, pode-se concluir que, dentre as sementes danificadas, grande parte das sementes analisadas possuirão a germinação afetada, já que os danos de variações de densidades foi o que apareceu com maior recorrência.

## 5.3 Estimação Bayesiana

### 5.3.1 Estimação bayesiana da proporção de sementes com danos e sem danos utilizando a distribuição Binomial

Para estimar a proporção de sementes sob duas categorias: sem danos e com danos, dos 100 lotes, através da abordagem bayesiana, utilizou-se uma *priori* conjugada Beta $(a,b)$, onde tem-se que após observar a amostra a distribuição *a posteriori* é uma Beta com parâmetros $(a+x, b+n-x)$, onde $x$ é o total de sementes sem danos e $n$ o tamanho amostal. Tem-se uma conjugada Beta já que compreende-se em uma binomial, considerando a probabilidade de sucesso a semente não apresentar danos.

Inicialmente utilizou-se uma *priori* uniforme, com $a = 1$ e $b = 1$, onde os valores dos parâmetros, chamados de hiperparâmetros, são igualmente prováveis. Na Tabela 5.4 estão apresentados os hiperparâmetros da distribuição *a priori* e a respectiva média e variância.

Tabela 5.4 – Hiperparâmetros e valores da média, variância da distribuição *a priori* uniforme

| **Hiperparâmetros** | | **Média (%)** | **Variância (%)** |
|---|---|---|---|
| a = 1 | b = 1 | 50,00 | 8,33 |

Fonte: As autoras (2021)

Na Tabela 5.5 são apresentados os resultados das estimativas das proporções de sementes sem danos e com danos, que são dadas pela média da distribuição *a posteriori*, e suas respectivas variâncias para cada lote.

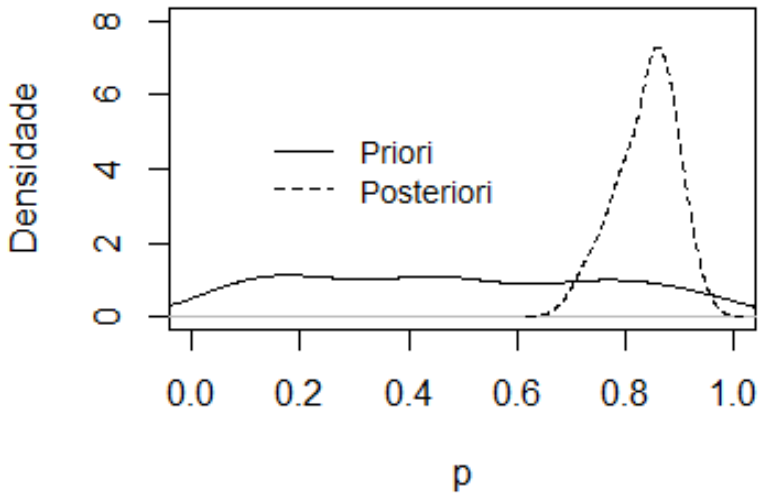
Tabela 5.5 – Estimativas bayesianas das proporções de sementes sem danos e danificadas considerando a distribuição binomial e *priori* uniforme

| **Lote** | **Sementes sem danos** | **Sementes danificadas** | |
|---|---|---|---|
| | **Média (%)** | **Média(%)** | **Variância** |
| 1 | 86,14 | 13,86 | 0,0588 |
| 2 | 87,13 | 12,87 | 0,0552 |
| 3 | 78,22 | 21,78 | 0,0839 |
| 4 | 85,64 | 14,36 | 0,0606 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 98 | 85,64 | 14,36 | 0,0606 |
| 99 | 76,73 | 23,27 | 0,0879 |
| 100 | 82,18 | 17,82 | 0,0721 |
| **Média (%)** | 83,49 | 16,51 | |
| **Desvio Padrão (%)** | 5,13 | 5,13 | |

Fonte: As autoras (2021)

É possível observar que as estimativas estão muito próximas das obtidas pela abordagem frequentista, observando a Tabela 5.1. A partir dos hiperparâmetros da *priori* e da *posteriori*, foi plotado o gráfico das duas densidades na Figura 5.1 para melhor compreensão.

Figura 5.1 – Densidades das distribuições *a priori* e *a posteriori* beta das sementes sem danos




Fonte: As autoras (2021)

A partir do gráfico apresentado na Figura 5.1, pode-se observar que as curvas evidenciam maior dispersão da *priori*, isso pode ser devido a *priori* ser uma *priori* não informativa, mas que não influenciou na *posteriori*, obtendo estimativas precisas.

Posteriormente, utilizou-se uma *priori* mais informativa, que foi construída com base nos resultados apresentados por Javorski e Cicero (2017), ao avaliarem a morfologia interna das sementes de sorgo, por meio de raios X e identificarem danos. No trabalho, eles registraram média de 90% de sementes sem danos e 10% de sementes danificadas, com um variância de $0,2283\%$.

A partir dos valores de média e variância de sementes sem danos da referência citada, encontrou-se os hiperparâmetros da distribuição *a priori* beta, de acordo com as expressões (3.36) e (3.37), que estão apresentados na Tabela 5.6:

Tabela 5.6 – Hiperparâmetros e valores da média, variância da distribuição *a priori* beta da literatura

| **Hiperparâmetros** | | **Média (%)** | **Variância (%)** |
|---|---|---|---|
| a = 34,5745 | b = 3,8416 | 90,00 | 0,22833 |

Fonte: As autoras (2021)

Na Tabela 5.7 são apresentados os resultados das estimativas das proporções de sementes sem danos e com danos, que são dadas pela média da distribuição *a posteriori*, e suas respectivas variâncias para cada lote, utilizando a *priori* construída com base no artigo.

Tabela 5.7 – Estimativas bayesianas das proporções de sementes sem danos e danificadas considerando a distribuição binomial e *priori* da literatura

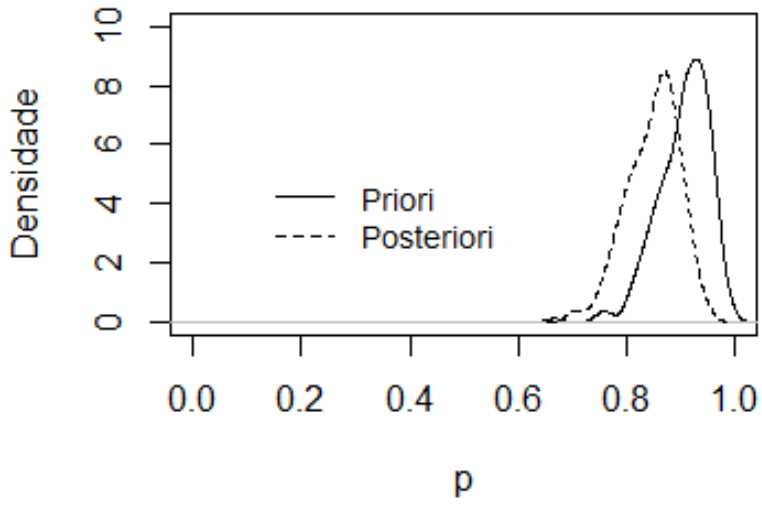
| **Lote** | **Sementes sem danos** | **Sementes danificadas** | |
|---|---|---|---|
| | **Média (%)** | **Média(%)** | **Variância** |
| 1 | 87,06 | 12,94 | 0,0470 |
| 2 | 87,90 | 12,10 | 0,0444 |
| 3 | 80,35 | 19,65 | 0,0659 |
| 4 | 86,64 | 13,36 | 0,0483 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 98 | 86,64 | 13,36 | 0,0483 |
| 99 | 79,09 | 20,91 | 0,0691 |
| 100 | 83,71 | 16,29 | 0,0570 |
| **Média (%)** | 84,82 | 15,18 | |
| **Desvio Padrão (%)** | 4,35 | 4,35 | |

Fonte: As autoras (2021)

É possível observar que as estimativas ainda estão próximas das obtidas pela abordagem frequentista, observando a Tabela 5.1. Entretanto, mais distantes das obtidas utilizando *priori* uniforme, já que são mais altas. Isso se deve ao fato de que a *priori* da literatura teve média de $90,00\%$, influenciando consideravelmente na *posteriori* a ter resultados maiores em relação à *priori* uniforme, que a média foi de 50%.

A partir dos hiperparâmetros da *priori* e da *posteriori*, foram plotados os gráficos das duas densidades na Figura 5.2:

Figura 5.2 – Densidades das distribuições *a priori* e *a posteriori* beta das sementes sem danos




Fonte: As autoras (2021)

Tem-se que as densidades da *priori* e da *posteriori* são muito próximas. Portanto, pode-se concluir que o uso de *prioris* diferentes, uma uniforme, onde os valores dos parâmetros são igualmente prováveis, e outra mais informativa, extraída da literatura, se obtêm resultados muito próximos na *posteriori*, indicando uma consistência nas análises.

Percebe-se que os valores estimados com a *priori* uniforme são sempre inferiores, mas mais próximos dos valores obtidos pela abordagem frequentista apresentados na Tabela 5.1, entretanto, a variância foi sempre maior do que da Tabela 5.7 das estimativas da *priori* da literatura.

### 5.3.2 Estimação bayesiana da proporção de sementes sob três classes utilizando a distribuição Multinomial

Para a estimação bayesiana da proporção de sementes com os três tipos de danos, por inseto, com variação de densidade e físicos, compreendendo a distribuição multinomial, utilizou-se *a priori* conjugada de Dirichlet $(a_1, a_2, a_3)$. Após observar a amostra sabe-se que a distribuição *a posteriori* é uma Dirichlet $(a_1 + x_1, a_2 + x_2, a_3 + x_3)$, onde $x_i$ é a frequência de cada classe, sendo $i = 1, 2, 3$ o número de classes, que nesse caso são 3, já que são 3 tipos de danos.

Inicialmente, considerou-se uma *priori* uniforme para a estimação das proporções, já que esta é uma caso particular da Dirichlet, logo, tem-se que os parâmetros da *priori*, denotados por hiperparâmetros, são: $a_1 = 1, a_2 = 1, a_3 = 1$, sendo todos os possíveis valores do parâmetro igualmente prováveis. Portanto, na Tabela 5.8 tem-se os valores dos hiperparâmetros de cada classe com suas respectivas médias, variâncias e covariâncias *a priori*, calculadas a partir das expressões dadas em (3.39):

Tabela 5.8 – Hiperparâmetros e valores da média, variância e covariância *a priori* com três classes de danos

| **Tipos de danos** | $a_i$ | **Média (%)** | **Variância (%)** | **Covariância (%)** |
|---|---|---|---|---|
| **Por inseto** | 1 | 33,33 | 5,56 | -2,78 |
| **Variação de densidade** | 1 | 33,33 | 5,56 | -2,78 |
| **Físicos** | 1 | 33,33 | 5,56 | -2,78 |

Fonte: As autoras (2021)

Nos 100 lotes analisados utilizou-se a frequência dos danos por inseto, variações de densidade e físicos para atualizar a informação *a priori* e obter cada uma das distribuições *a posteriori* dos lotes, conforme apresentado em (3.40). Desse modo, os resultados da estimação bayesiana para a proporção dos três tipos de danos para cada lote, que são dadas pela média da distribuição *a posteriori*, considerando a *priori* uniforme, podem ser observados na Tabela (5.9):

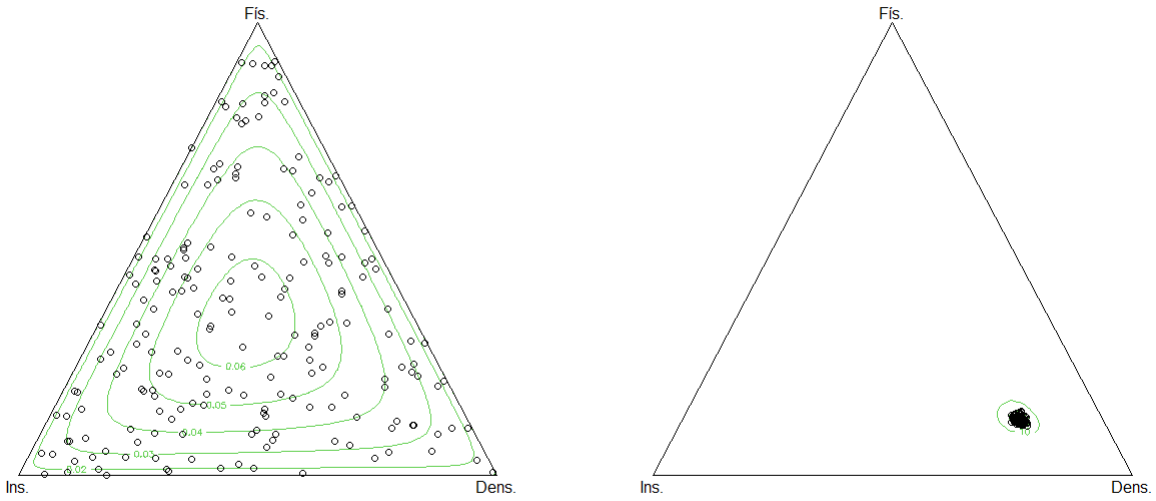
Tabela 5.9 – Estimativas bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos, utilizando *priori* uniforme

| **Tipo de dano** | $\hat{\boldsymbol{p}}_1$**: Por inseto** | | $\hat{\boldsymbol{p}}_2$**: Variações de densidade** | | $\hat{\boldsymbol{p}}_3$**: Físicos** | |
|---|---|---|---|---|---|---|
| **Lote** | **Média (%)** | **Variância (%)** | **Média (%)** | **Variância (%)** | **Média (%)** | **Variância (%)** |
| **1** | 33,33 | 0,72 | 63,33 | 0,75 | 3,33 | 0,10 |
| **2** | 17,86 | 0,51 | 75,00 | 0,65 | 7,14 | 0,23 |
| **3** | 8,70 | 0,17 | 84,78 | 0,27 | 6,52 | 0,13 |
| **4** | 9,68 | 0,27 | 87,10 | 0,35 | 3,22 | 0,10 |
| **⋮** | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **98** | 35,48 | 0,72 | 38,71 | 0,74 | 25,81 | 0,60 |
| **99** | 20,41 | 0,32 | 63,26 | 0,46 | 16,33 | 0,27 |
| **100** | 10,53 | 0,24 | 65,79 | 0,58 | 23,68 | 0,46 |
| **Média (%)** | 19,14 | | 66,32 | | 14,54 | |
| **Desvio padrão (%)** | 10,35 | | 18,67 | | 11,97 | |

Fonte: As autoras (2021)

Pode-se observar que as estimativas ficaram bem próximas das frequentistas, como pode ser visto na Tabela 5.2, mas para se ter melhor compreensão do impacto do parâmetro da *priori* e da *posteriori* é apresentado o simplex de dimensão 3, que pode ser observado na Figura 5.3:

Figura 5.3 – Simplex da *priori* (à esquerda) e simplex da *posteriori* (à direita)




Fonte: As autoras (2021)

O parâmetro da distribuição uniforme é ($a_1 = 1, a_2 = 1, a_3 = 1$), e o parâmetro da *posteriori* é ($a_1^* = 563, a_2^* = 2277, a_3^* = 399$), pois foi construída considerando os danos totais de cada classe, assim, nos 100 lotes analisados a frequência dos danos por inseto foi de 562 sementes, a de variação de densidade de 2276 sementes e de dano físico de 398.

Assim, pode-se observar que para a *priori* uniforme (simplex à esquerda) todos os pontos no simplex são igualmente prováveis, pois estão espalhados por todo o simplex. Já no simplex à direita, que representa o da *posteriori*, é possível perceber claramente que para os 100 lotes avaliados, em média, tem-se maior ocorrência de variações de densidade. Portanto, não houve influência na *posteriori* quando a *priori* é uniforme, pois é uma *priori* não informativa, e neste caso, a *posteriori* é a própria função de verossimilhança.

Para estimar a proporção de sementes utilizando a abordagem bayesiana considerando as outras três classes, assim como também foi realizado na abordagem frequentista: sementes sem danos, sementes com variações de densidade e sementes com outros tipos de danos, inicialmente utilizou-se uma *priori* uniforme como no caso anterior, onde os hiperparâmetros são: $a_1 = 1, a_2 = 1, a_3 = 1$.

Portanto, os valores dos hiperparâmetros e suas respectivas médias, variâncias e covariâncias *a priori*, que foram calculadas a partir das expressões dadas em (3.39), serão iguais aos que estão apresentados na Tabela 5.8.

Ao registrar a frequência de cada tipo de dano em cada um dos 100 lotes, atualizou-se a informação *a priori* e foram obtidas as distribuições *a posteriori* de cada lote. Logo, as estimativas através da abordagem bayesiana das proporções de cada lote, que são dadas pela média da distribuição *a posteriori*, são apresentadas na Tabela 5.10 a seguir:

Tabela 5.10 – Estimativas bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos, utilizando *priori* uniforme

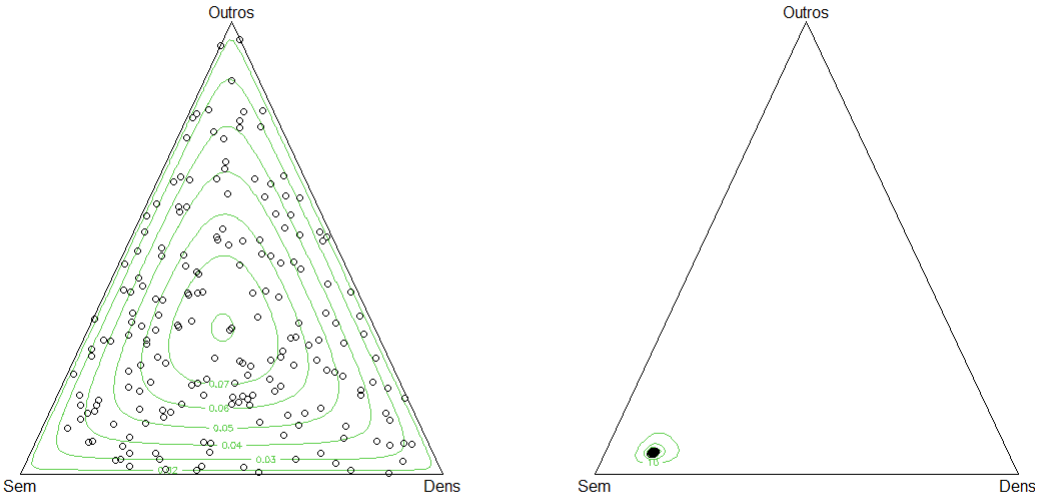
| **Classes** | $\hat{\boldsymbol{p}}_1$**: Sem danos** | | $\hat{\boldsymbol{p}}_2$**: Variações de densidade** | | $\hat{\boldsymbol{p}}_3$**: Outros tipos de danos** | |
|---|---|---|---|---|---|---|
| **Lote** | **Média (%)** | **Variância (%)** | **Média (%)** | **Variância (%)** | **Média (%)** | **Variância (%)** |
| **1** | 85,71 | 0,06 | 9,36 | 0,04 | 4,93 | 2,29 |
| **2** | 86,70 | 0,06 | 10,34 | 0,05 | 2,96 | 1,41 |
| **3** | 77,83 | 0,08 | 19,21 | 0,08 | 2,96 | 1,41 |
| **4** | 85,22 | 0,06 | 13,30 | 0,06 | 1,48 | 7,14 |
| **⋮** | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **98** | 85,22 | 0,06 | 5,91 | 0,03 | 8,87 | 3,96 |
| **99** | 76,35 | 0,09 | 15,27 | 0,06 | 8,37 | 3,76 |
| **100** | 81,77 | 0,07 | 12,32 | 0,05 | 5,91 | 2,73 |
| **Média (%)** | 83,07 | | 11,70 | | 5,22 | |
| **Desvio padrão (%)** | 5,10 | | 4,99 | | 3,44 | |

Fonte: As autoras (2021)

Pode-se observar que as estimativas ficaram bem próximas das frequentistas da Tabela 5.3. Além disso, para essa organização dos dados as estimativas ficaram melhores, pois considerando apenas os tipos de danos tem-se menos dados e portanto aconteceu uma superestimação, mas ainda sim não ficando ruim ao comparar com as estimativas frequentistas.

Para se ter melhor compreensão do impacto do parâmetro da *priori* e da *posteriori* é apresentado o simplex de dimensão 3, que pode ser observado na Figura 5.4:

Figura 5.4 – Simplex da *priori* (à esquerda) e simplex da *posteriori* (à direita)




Fonte: As autoras (2021)

O parâmetro da distribuição uniforme é ($a_1 = 1$, $a_2 = 1$, $a_3 = 1$), e o parâmetro da *posteriori* é ($a_1^* = 16765$, $a_2^* = 2277$, $a_3^* = 961$), pois foi construída considerando a frequência total de sementes sem danos, de sementes com variações de densidade e com outros tipos de danos, assim, nos 100 lotes analisados a frequência foi respectivamente: 16764, 2276 e 960.

Portanto, pode-se observar que para a *priori* uniforme (simplex à esquerda) todos os pontos no simplex são igualmente prováveis, pois estão espalhados por todo o simplex. Já no simplex à direita, que representa o da *posteriori*, é possível perceber claramente que para os 100 lotes avaliados, em média, tem-se maior ocorrência de sementes sem danos, concluindo novamente que a *priori* uniforme não influenciou na *posteriori*.

Em seguida, realizou-se toda a estimação novamente para as duas organizações de dados considerando uma distribuição multinomial, mas para uma *priori* mais informativa, extraída da literatura.

A *priori* foi construída com base no artigo de Javorski e Cicero (2017), onde foi identificado através do teste de raios X, danos por deterioração de tecidos, danos mecânicos, danos por má formação e danos por insetos, nas sementes de sorgo, que é uma outra espécie, mas ainda é uma gramínea semelhante ao milho. Assim, através desse artigo, obteve-se como média $0,55\%$ de sementes com danos por insetos, $8,45\%$ com variações de densidade e $1,00\%$ com danos físicos,

Logo, para encontrar os valores dos parâmetros da distribuição *a priori* de Dirichlet que resultarão nas médias anteriores, primeiramente realizou-se uma proporção em 100%, do seguinte modo:

| | | | | | |
|---|---|---|---|---|---|
| $10\%$ | $100\%$ | $10\%$ | $100\%$ | $10\%$ | $100\%$ |
| $0,55\%$ | $x$ | $8,45\%$ | $x$ | $1,00\%$ | $x$ |

O que resultou nas respectivas médias: $5,5\%$, $84,5\%$ e $10,0\%$. Desta forma, os parâmetros foram encontrados utilizando a expressão da média em (3.39). Na Tabela 5.11 está apresentado os valores dos parâmetros (hiperparâmetros) encontrados para a *priori* de Dirichlet, com suas respectivas médias, variâncias e covariâncias *a priori*:

Tabela 5.11 – Hiperparâmetros e valores da média, variância e covariância *a priori* com três classes de danos

| **Tipos de danos** | $a_i$ | **Média (%)** | **Variância (%)** | **Covariância (%)** |
|---|---|---|---|---|
| **Por inseto** | 0,0275 | 5,5 | 3,47 | -3,10 |
| **Variação de densidade** | 0,4225 | 84,5 | 8,73 | -0,37 |
| **Físicos** | 0,05 | 10,0 | 6,00 | -5,63 |

Fonte: As autoras (2021)

Os resultados da estimação bayesiana para a proporção dos três tipos de danos para cada lote, que são dadas pela média da distribuição *a posteriori*, considerando a *priori* construída através do artigo, podem ser observados na Tabela (5.12)

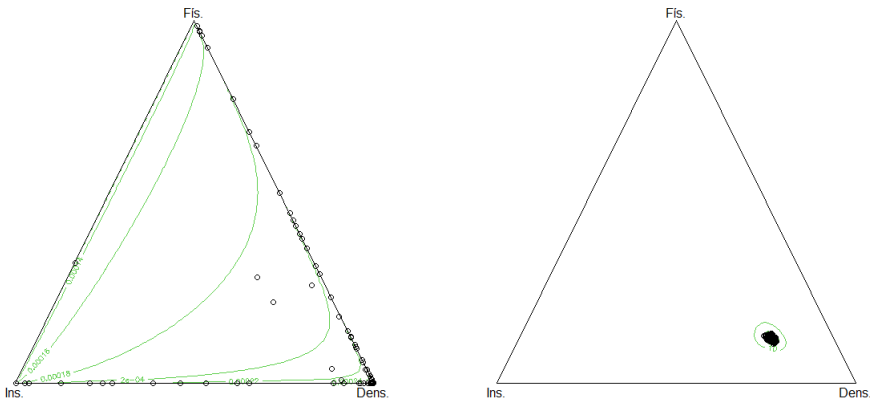
Tabela 5.12 – Estimativas bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos, utilizando *priori* da literatura

| **Tipo de dano** | $\hat{\boldsymbol{p}}_1$**: Por inseto** | | $\hat{\boldsymbol{p}}_2$**: Variações de densidade** | | $\hat{\boldsymbol{p}}_3$**: Físicos** | |
|---|---|---|---|---|---|---|
| **Lote** | **Média (%)** | **Variância (%)** | **Média (%)** | **Variância (%)** | **Média (%)** | **Variância (%)** |
| **1** | 32,83 | 0,77 | 66,99 | 0,78 | 0,18 | 0,01 |
| **2** | 15,79 | 0,50 | 80,09 | 0,60 | 4,12 | 0,15 |
| **3** | 6,96 | 0,15 | 88,33 | 0,23 | 4,71 | 0,10 |
| **4** | 7,11 | 0,22 | 92,71 | 0,23 | 0,18 | 0,01 |
| **⋮** | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **98** | 35,18 | 0,77 | 40,08 | 0,81 | 24,74 | 0,63 |
| **99** | 19,41 | 0,33 | 65,42 | 0,48 | 15,16 | 0,27 |
| **100** | 8,53 | 0,21 | 68,80 | 0,59 | 22,68 | 0,48 |
| **Média (%)** | 17,54 | | 69,88 | | 12,58 | |
| **Desvio padrão (%)** | 11,19 | | 20,09 | | 12,99 | |

Fonte: As autoras (2021)

Pode-se observar que as estimativas ficaram bem próximas das estimativas frequentistas da Tabela 5.2, ainda mais próximas de quando utilizou-se uma *priori* uniforme, como pode ser visto em 5.9, assim, concluindo que ao utilizar uma *priori* informativa acarretou em estimativas mais precisas, diferentemente do que aconteceu no caso binomial. Para se ter melhor compreensão do impacto do parâmetro da *priori* e da *posteriori* é apresentado o simplex de dimensão 3, na Figura 5.5:

Figura 5.5 – Simplex da *priori* (à esquerda) e simplex da *posteriori* (à direita)




Fonte: As autoras (2021)

O parâmetro da distribuição *a priori* foi ($a_1 = 0,0275$, $a_2 = 0,4225$, $a_3 = 0,05$), e o parâmetro da *posteriori* foi ($a_1^* = 562,0275$; $a_2^* = 2276,4225$ e $a_3^* = 398,05$), pois foi construída considerando a frequência total de sementes sem danos, de sementes com variações de densidade e com outros tipos de danos.

No simplex à esquerda os pontos estão mais próximos da categoria variações de densidade. No simplex à direita, que representa o da *posteriori*, é possível perceber claramente que para os 100 lotes avaliados, em média, tem-se maior ocorrência de sementes com variações de densidade também.

Agora, considerando outras três classes: sementes sem danos, com variações de densidade e com outros tipos de danos, construiu-se uma *priori* ainda com base no artigo de Javorski e Cicero (2017), onde obteve-se como média $90,00\%$ de sementes sem danos, $8,45\%$ com variações de densidade e $1,55\%$ com outros tipos de danos. Como ao somar resulta em 100% não foi necessário fazer a proporção das médias.

Do mesmo modo, os parâmetros foram encontrados utilizando a expressão da média em (3.39). Na Tabela 5.13 estão apresentados os valores dos parâmetros (hiperparâmetros) encontrados para a *priori* de Dirichlet, com suas respectivas médias, variâncias e covariâncias *a priori*:

Tabela 5.13 – Hiperparâmetros e valores da média, variância e covariância *a priori* com três classes de danos

| **Tipos de danos** | $a_i$ | **Média (%)** | **Variância (%)** | **Covariância (%)** |
|---|---|---|---|---|
| **Por inseto** | 0,45 | 90,0 | 6,00 | -5,07 |
| **Variação de densidade** | 0,04225 | 8,45 | 5,16 | -0,93 |
| **Físicos** | 0,00775 | 1,55 | 1,02 | -0,09 |

Fonte: As autoras (2021)

Os resultados da estimação bayesiana para a proporção das três categorias, que são dadas pela média da distribuição *a posteriori*, considerando a *priori* construída através do artigo, podem ser observados na Tabela (5.14)

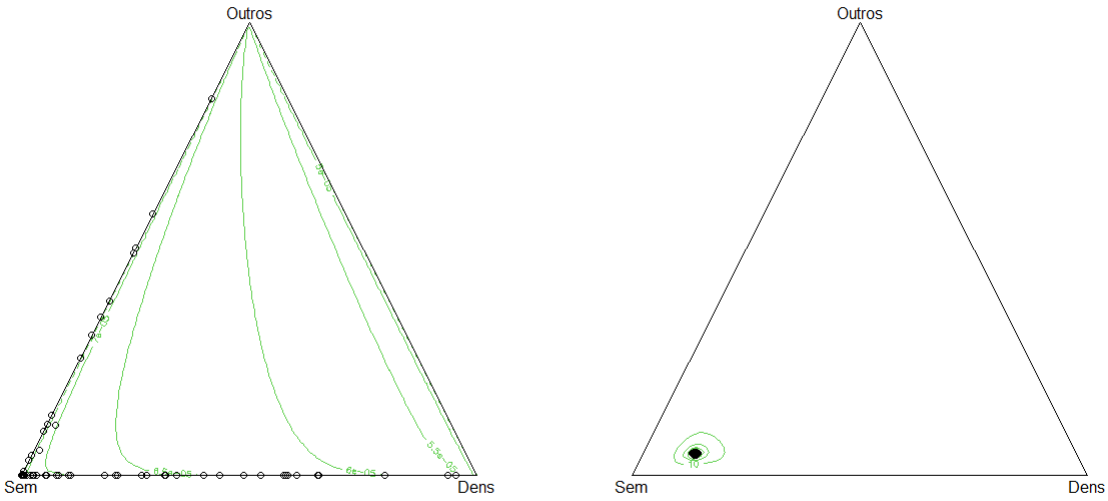
Tabela 5.14 – Estimativas bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos, utilizando *priori* da literatura

| **Classes** | $\hat{\boldsymbol{p}}_1$**: Sem danos** | | $\hat{\boldsymbol{p}}_2$**: Variações de densidade** | | $\hat{\boldsymbol{p}}_3$**: Outros tipos de danos** | |
|---|---|---|---|---|---|---|
| **Lote** | **Média (%)** | **Variância (%)** | **Média (%)** | **Variância (%)** | **Média (%)** | **Variância (%)** |
| **1** | 86,51 | 0,06 | 9,00 | 0,04 | 4,49 | 0,02 |
| **2** | 87,51 | 0,05 | 10,00 | 0,04 | 2,50 | 0,01 |
| **3** | 78,53 | 0,08 | 18,97 | 0,08 | 2,50 | 0,01 |
| **4** | 86,01 | 0,06 | 12,99 | 0,06 | 1,00 | 0,05 |
| **⋮** | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **98** | 86,01 | 0,06 | 5,51 | 0,03 | 8,48 | 0,04 |
| **99** | 77,03 | 0,09 | 14,98 | 0,06 | 7,98 | 0,04 |
| **100** | 82,52 | 0,07 | 11,99 | 0,05 | 5,49 | 0,03 |
| **Média (%)** | 83,84 | | 11,37 | | 4,79 | |
| **Desvio padrão (%)** | 5,17 | | 5,05 | | 3,48 | |

Fonte: As autoras (2021)

Pode-se observar que as estimativas ficaram bem próximas das estimativas frequentistas da Tabela 5.3, ainda mais próximas de quando utilizou-se uma *priori* uniforme, como pode ser visto na Tabela 5.10, assim, concluindo novamente que a *priori* informativa acarretou em estimativas mais precisas. Para se ter melhor compreensão do impacto do parâmetro da *priori* e da *posteriori* é apresentado o simplex de dimensão 3, na Figura 5.5:

Figura 5.6 – Simplex da *priori* (à esquerda) e simplex da *posteriori* (à direita)




Fonte: As autoras (2021)

O parâmetro da distribuição *a priori* foi ($a_1 = 0,45$, $a_2 = 0,04225$, $a_3 = 0,00775$), e o parâmetro da *posteriori* foi ($a_1^* = 16764,45$; $a_2^* = 2276,04225$ e $a_3^* = 960,00775$). No simplex à esquerda os pontos estão mais próximos da categoria sementes sem danos. No simplex à direita, que representa o da *posteriori*, é possível perceber claramente que para os 100 lotes avaliados, em média, tem-se maior ocorrência de sementes sem danos.

Portanto, todos os resultados obtidos com a estimação bayesiana estão em conformidade com a abordagem frequentista e portanto, também com os dados encontrados na literatura.

## 5.4 Estimação Sequencial Bayesiana

### 5.4.1 Estimação sequencial bayesiana da proporção de sementes com danos e sem danos utilizando a distribuição Binomial

A estimação sequencial bayesiana para a distribuição binomial da proporção de sementes sem danos e danificadas dos 100 lotes, foi realizada através de um aplicativo em *Delphi* desenvol-

vido por Resende et al. (2012). O aplicativo emite um relatório com os resultados e os parâmetros utilizados para o processo de estimação.

Portanto, para realizar a estimação sequencial bayesiana utilizou-se duas *prioris*, uma em que a média foi de 50% e variância 10%, a outra com média de 90% e variância de 5%, definidas antes da análise no programa em *Delphi*. Além disso, após testar os três valores para o custo: $10^{-4}$, $10^{-5}$ e $10^{-6}$, foi adotado o valor de $10^{-5}$, pois foi o mais adequado. Para $10^{-6}$ o custo foi muito baixo e a análise não se interrompeu, e para $10^{-4}$ o custo foi muito alto, fazendo com que o procedimento parasse após a análise de algumas poucas sementes.

Na Tabela 5.15 tem-se os valores de médias e variâncias das *prioris*, assim como seus respectivos hiperparâmetros da distribuição beta *a priori*.

Tabela 5.15 – Valores dos hiperparâmetros, médias e variâncias das *prioris*

| **Hiperparâmetros** | | **Média (%)** | **Variância (%)** |
|---|---|---|---|
| **a** | **b** | | |
| 0,75 | 0,75 | 50,00 | 10,00 |
| 0,72 | 0,08 | 90,00 | 5,00 |

Fonte: As autoras (2021)

Os hiperparâmetros utilizados não refletem exatamente uma *priori* uniforme e a que foi obtida através do artigo como na estimação bayesiana, pois o programa em Delphi não admite que entre com valores específicos de parâmetros. É possível definir a média da *priori*, mas a variância e o custo são definidos apenas como pequenos, médios ou grandes. Assim, para a média 50%, a variância escolhida foi "média", e para a média de 90% a variância escolhida foi "pequena", e ambos os custos como "pequenos". O aplicativo então para variância pequena adota o valor de 5% e para média, o valor de 10%. Para o custo pequeno o valor é de $10^{-5}$.

Os resultados dos relatórios da estimação sequencial bayesiana para a *priori* cuja média é 50% estão na Tabela 5.16, e da outra *priori* com média 90% estão na Tabela 5.17, com seus respectivos tamanhos amostrais.

Tabela 5.16 – Estimativas sequenciais bayesianas das proporções de sementes sem danos e danificadas considerando a distribuição binomial

| **Lote** | | **Sementes sem danos** | **Sementes danificadas** |
|---|---|---|---|
| | $n_{seq}$ | **Média (%)** | **Média (%)** |
| 1 | 19 | 91,46 | 8,54 |
| 2 | 23 | 84,69 | 15,31 |
| 3 | 19 | 91,46 | 8,54 |
| 4 | 21 | 87,78 | 12,22 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 98 | 21 | 87,78 | 12,22 |
| 99 | 26 | 75,45 | 24,55 |
| 100 | 24 | 81,37 | 18,63 |
| **Média (%)** | | 80,82 | 19,18 |
| **Desvio Padrão (%)** | | 7,84 | 7,84 |

Fonte: As autoras (2021)

Tabela 5.17 – Estimativas sequenciais bayesianas considerando sementes sem danos e danificadas

| **Lote** | | **Sementes sem danos** | **Sementes danificadas** |
|---|---|---|---|
| | $n_{seq}$ | **Média (%)** | **Média (%)** |
| 1 | 19 | 91,31 | 8,69 |
| 2 | 23 | 84,37 | 15,63 |
| 3 | 19 | 91,31 | 8,69 |
| 4 | 21 | 87,52 | 12,48 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 98 | 21 | 87,52 | 12,48 |
| 99 | 26 | 74,93 | 25,07 |
| 100 | 24 | 80,97 | 19,03 |
| **Média (%)** | | 80,47 | 19,53 |
| **Desvio Padrão (%)** | | 7,93 | 7,93 |

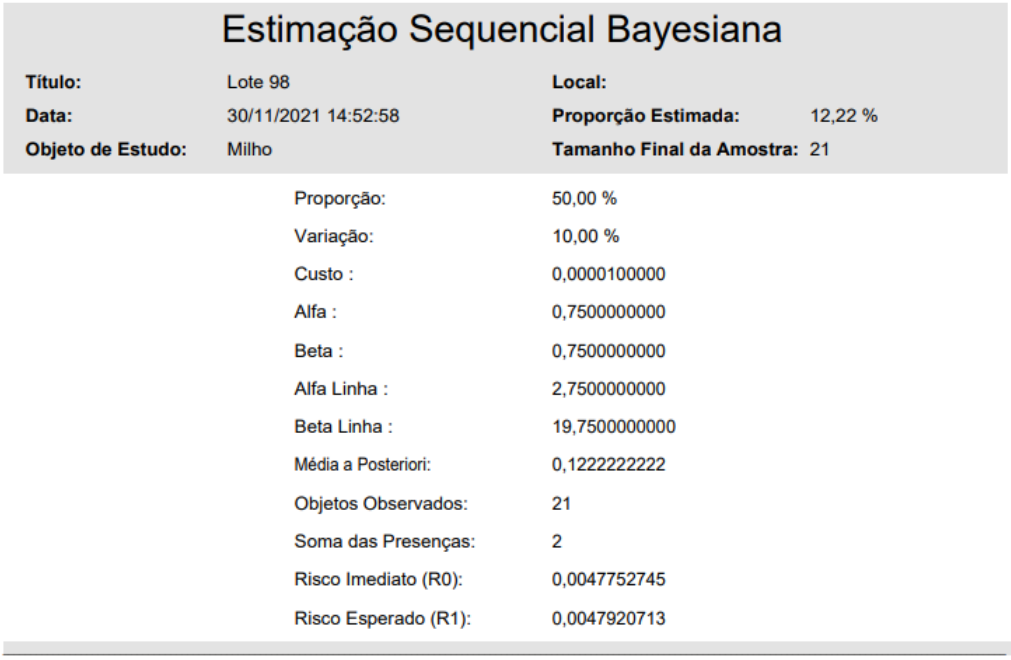
Fonte: As autoras (2021)

Portanto, pode-se observar que as estimativas também não ficaram distantes das obtidas pela abordagem frequentista. No entanto, alguns lotes, como o lote 1 e lote 3, houve uma superestimação, isso aconteceu devido ao fato de haver uma classe discrepante em relação à outra. Então, quando avaliou-se as sementes observou-se muitas sementes sem danos em sequência, assim acumulando os resultados para essa classe, fazendo com que interrompa a amostragem mais rapidamente e realize uma estimação não muito precisa.

Logo, para proporções mais extremas obtiveram estimativas mais distantes, mas para os lotes em que as proporções são intermediárias, mais distribuídas entre as classes, as estimativas sequencias bayesianas foram melhores, obtendo sucesso em todos os casos avaliados, para as duas *prioris* utilizadas. Além disso, resultou em um tamanho amostral menor em todos os lotes analisados, os quais foram iguais para as duas *prioris*.

Na Figura 5.7 tem-se um exemplo do relatório final para o lote 98 emitido pelo aplicativo em *Delphi*.

Figura 5.7 – Relatório do programa em *Delphi* com os resultados



Estimação Sequencial Bayesiana

| | | | |
|---|---|---|---|
| **Título:** | Lote 98 | **Local:** | |
| **Data:** | 30/11/2021 14:52:58 | **Proporção Estimada:** | 12,22 % |
| **Objeto de Estudo:** | Milho | **Tamanho Final da Amostra:** | 21 |

| | |
|---|---|
| Proporção: | 50,00 % |
| Variação: | 10,00 % |
| Custo : | 0,0000100000 |
| Alfa : | 0,7500000000 |
| Beta : | 0,7500000000 |
| Alfa Linha : | 2,7500000000 |
| Beta Linha : | 19,7500000000 |
| Média a Posteriori: | 0,1222222222 |
| Objetos Observados: | 21 |
| Soma das Presenças: | 2 |
| Risco Imediato (R0): | 0,0047752745 |
| Risco Esperado (R1): | 0,0047920713 |

Fonte: As autoras (2021)

Portanto, conclui-se que as estimativas estão um pouco mais distantes das obtidas pela abordagem frequentista e pela abordagem bayesiana, principalmente para proporções mais extremas, mas ainda assim não tão ruins, pois acarretou-se em um tamanho de amostra menor.

Além disso, o caso apresentado aqui, em que as estimativas ficaram mais distantes, pode ser devido a *priori* não refletir os lotes analisados, ou seja, não ser condizente com os lotes, havendo as-

sim forte influência da distribuição *a priori* que não era compatível com os resultados frequentistas. Mas que para proporções intermediárias ainda sim teve um bom desempenho.

### 5.4.2 Estimação sequencial bayesiana da proporção de sementes sob três classes utilizando a distribuição Multinomial

Para a estimação sequencial bayesiana, foi visto na Figura 3.6, que deve-se calcular o risco imediato e o risco esperado e compará-los até que o risco imediato seja menor ou igual ao esperado, interrompendo assim a amostragem e realizando a estimação. No entanto, para calcular os riscos, antes é necessário encontrar as estimativas a *posteriori*, ou seja, a média da distribuição a *posteriori*, e incluir também um custo por observação no procedimento.

Testou-se os valores $10^{-4}$, $10^{-5}$ e $10^{-6}$ para o custo por observação, entretanto o que apresentou o melhor desempenho foi o $10^{-5}$, pois para o $10^{-4}$ o custo foi muito alto e a amostragem interrompeu rapidamente para a maioria dos lotes analisados, apenas com algumas sementes avaliadas, fazendo com que o valor obtido para a estimativa da proporção não fizesse sentido com o verdadeiro valor do parâmetro. Já para o custo de $10^{-6}$ foi um custo muito baixo, e para a maioria dos lotes não houve a interrupção da amostragem, requerendo um número de sementes maior do que o disponível, e portanto obteve-se as mesmas estimativas das obtidas pela abordagem frequentista.

A seleção do custo de $10^{-5}$ está de acordo também com Bach (2015), já que este possui uma ordem de magnitude semelhante à ordem de magnitude de $(\boldsymbol{p}-\boldsymbol{d})^T\boldsymbol{K}(\boldsymbol{p}-\boldsymbol{d})$ da função de perda, pois assim assegura que a função de risco não seja dominada exclusivamente pelo custo. Como a perda é o quadrado de uma diferença entre os valores de proporção real e estimada, que estão compreendidos no intervalo [0,1], os resultados são sempre próximos de zero, e portanto, o custo deve ser próximo de zero também.

Desse modo, adotou-se como custo por observação o valor de $10^{-5}$, e realizou-se todos os cálculos necessários para cada um dos 100 lotes através de uma tabela dinâmica construída no Microsoft Excel ®. Utilizou-se uma *priori* uniforme para o início do processo de estimação sequencial bayesiana, onde todos os possíveis valores do parâmetro são igualmente prováveis, logo

os riscos foram calculados com base nas expressões dadas em (4.6) e (4.7), de acordo com Jones (1976). A partir da segunda semente avaliada, utilizou-se as estimativas anteriores como *priori* para atualizar as informações. Desta forma, tem-se *prioris* que seguem uma distribuição de Dirichlet, e os riscos foram calculados a partir das expressões (4.8) e (4.9) desenvolvidas.

Inicialmente, realizou-se a estimação considerando apenas as sementes danificadas, e portanto tem-se as seguintes classes: sementes com danos por inseto, com variações de densidade e com danos físicos, utilizando *priori* uniforme, cujos resultados se encontram na Tabela 5.18.

Tabela 5.18 – Estimativas sequenciais bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos, utilizando *priori* uniforme

| **Lotes** | $m_{seq}$ | **Tipos de danos** | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\hat{p}_1$**: Por inseto** | | $\hat{p}_2$**: Variações de densidade** | | $\hat{p}_3$**: Físicos** | |
| | | $m_1$ | $\hat{p}_{1\,seq}$ **(%)** | $m_2$ | $\hat{p}_{2\,seq}$ **(%)** | $m_3$ | $\hat{p}_{3\,seq}$ **(%)** |
| **1** | 6 | 0 | 0,00 | 6 | 100,00 | 0 | 0,00 |
| **2**$^*$ | 25 | 4 | 16,00 | 20 | 80,00 | 1 | 4,00 |
| **3**$^*$ | 43 | 3 | 6,98 | 38 | 88,37 | 2 | 4,65 |
| **4** | 6 | 0 | 0,00 | 6 | 100,00 | 0 | 0,00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **98**$^*$ | 28 | 10 | 35,71 | 11 | 39,29 | 7 | 25,00 |
| **99**$^*$ | 46 | 9 | 19,57 | 30 | 65,22 | 7 | 15,22 |
| **100**$^*$ | 35 | 3 | 8,58 | 24 | 68,57 | 8 | 22,86 |
| **Média (%)** | | | 16,01 | | 71,75 | | 12,24 |
| **Desvio Padrão (%)** | | | 12,94 | | 22,65 | | 13,50 |

Fonte: As autoras (2021)

$*$: A amostragem não foi interrompida

Nota-se que os lotes que apresentam asteriscos $*$, são aqueles em que a amostragem não foi interrompida, então utilizou-se todas as sementes disponíveis para realizar a estimação, e portanto, as estimativas nesses casos foram idênticas às estimativas frequentistas.

Assim, pode-se concluir que para essa organização de dados, considerando apenas os tipos de danos, o procedimento de estimação sequencial bayesiana foi prejudicado, pois diminuiu-se os tamanhos dos lotes ficando com poucas sementes, e então para a maioria dos lotes não houve a interrupção da amostragem, obtendo as mesmas estimativas pela abordagem frequentista e com o mesmo número de sementes avaliadas.

Já para os lotes em que houve a interrupção da amostragem esta aconteceu precocemente, fazendo com que as estimativas não ficassem precisas. Isso ocorreu porque havia uma classe muito discrepante em relação às outras, a de variações de densidade, pois a maioria das sementes possuiu dano relacionado a variação de densidade, e portanto, os resultados foram acumulados e a tendência é que análise acabe rapidamente, que foi o ocorrido, resultando em um tamanho amostral de 6 sementes analisadas.

Apesar disso, o custo de $10^{-5}$ ainda foi o melhor, pois para essa organização de dados, ao adotar o custo de $10^{-6}$ nenhum lote interrompeu a amostragem, e para $10^{-4}$ interrompeu mais rapidamente do que para $10^{-5}$, apenas com 5 sementes avaliadas. Desse modo, os resultados comprovam que o custo não é o principal fator na tomada de decisão, mas sim a variação entre as classes e a quantidade de dados disponíveis para análise, que foi o que mais prejudicou o procedimento.

Posteriormente realizou-se a estimação sequencial bayesiana considerando outras três classes: sementes sem danos, sementes que possuem variações de densidade, e as que possuem outros tipos de danos. Os resultados estão na Tabela 5.19.

Tabela 5.19 – Estimativas sequenciais bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos, utilizando *priori* uniforme

| **Lotes** | $\boldsymbol{m}_{seq}$ | **Tipos de danos** | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\boldsymbol{p}}_1$**: Sem danos** | | $\hat{\boldsymbol{p}}_2$**: Variações de densidade** | | $\hat{\boldsymbol{p}}_3$**: Outros tipos de danos** | |
| | | $\boldsymbol{m_1}$ | $\hat{\boldsymbol{p}}_{1\,seq}$ **(%)** | $\boldsymbol{m_2}$ | $\hat{\boldsymbol{p}}_{2\,seq}$ **(%)** | $\boldsymbol{m_3}$ | $\hat{\boldsymbol{p}}_{3\,seq}$ **(%)** |
| **1** | 155 | 134 | 86,45 | 13 | 8,39 | 8 | 5,16 |
| **2** | 6 | 6 | 99,98 | 0 | 0,01 | 0 | 0,01 |
| **3** | 6 | 6 | 99,98 | 0 | 0,01 | 0 | 0,01 |
| **4** | 141 | 125 | 88,65 | 15 | 10,64 | 1 | 0,71 |
| **⋮** | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **98** | 6 | 6 | 99,98 | 0 | 0,01 | 0 | 0,01 |
| **99** | 193 | 149 | 77,20 | 29 | 15,03 | 15 | 7,77 |
| **100** | 170 | 141 | 82,94 | 21 | 12,35 | 8 | 4,71 |
| **Média (%)** | | | 87,34 | | 8,70 | | 3,97 |
| **Desvio Padrão (%)** | | | 8,70 | | 6,79 | | 3,86 |

Fonte: As autoras (2021)

Assim, pode-se observar que para essa organização de dados, as estimativas pelo método sequencial bayesiano foram melhores do que para a organização considerando apenas os tipos

de danos. As estimativas estão próximas das obtidas pela abordagem frequentista e bayesiana, confirmando a adequação do método proposto. No entanto, com a vantagem de requerer um número menor de sementes em alguns lotes, diminuindo assim o tamanho amostral.

Para alguns lotes como o 2, 3 e 98, de acordo com a Tabela (5.19), a amostragem foi interrompida muito rapidamente e as estimativas não foram muito precisas, isso aconteceu também porque há uma classe muito discrepante em relação às outras, neste caso, a classe de sementes sem danos. No entanto, ao comparar esses resultados com os obtidos pela outra organização de dados, conclui-se que esses foram melhores, validando o procedimento de estimação sequencial bayesiana.

Os lotes 52, 82, 85 e 92 não interromperam a amostragem e portanto necessitou-se das 200 sementes para a estimação, possuindo assim a mesma estimativa da frequentista. O número máximo de sementes nos lotes em que a amostragem foi interrompida foi de 197, nos lotes 16 e 84.

Nos demais lotes houve uma redução considerável de sementes avaliadas para a estimação da proporção, isso indica de como o método sequencial bayesiano para a distribuição multinomial considerando as três classes reduziu o tempo de amostragem necessário para sentenciar um lote, em comparação ao método tradicional de 200 sementes.

De modo análogo ao anterior, realizou-se a estimação sequencial bayesiana primeiramente apenas para as sementes danificadas, no entanto, utilizando agora uma *priori* construída com base no artigo Javorski e Cicero (2017), com os seguintes parâmetros: $Dirichlet(0,0275; 0,4225; 0,05)$, como na Tabela 5.11. Os resultados estão apresentados na Tabela 5.20.

Tabela 5.20 – Estimativas sequenciais bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** danos por inseto, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** danos físicos, utilizando *priori* da literatura

| **Lotes** | $\boldsymbol{m}_{seq}$ | **Tipos de danos** | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\boldsymbol{p}}_1$**: Por inseto** | | $\hat{\boldsymbol{p}}_2$**: Variações de densidade** | | $\hat{\boldsymbol{p}}_3$**: Físicos** | |
| | | $\boldsymbol{m_1}$ | $\hat{\boldsymbol{p}}_{1\,seq}$ **(%)** | $\boldsymbol{m_2}$ | $\hat{\boldsymbol{p}}_{2\,seq}$ **(%)** | $\boldsymbol{m_3}$ | $\hat{\boldsymbol{p}}_{3\,seq}$ **(%)** |
| **1** | 5 | 0 | 0,00 | 5 | 100,00 | 0 | 0,00 |
| **2**$^*$ | 25 | 4 | 16,00 | 20 | 80,00 | 1 | 4,00 |
| **3**$^*$ | 43 | 3 | 6,98 | 38 | 88,37 | 2 | 4,65 |
| **4** | 5 | 0 | 0,00 | 5 | 100,00 | 0 | 0,00 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **98**$^*$ | 28 | 10 | 35,71 | 11 | 39,29 | 7 | 25,00 |
| **99**$^*$ | 46 | 9 | 19,57 | 30 | 65,22 | 7 | 15,22 |
| **100**$^*$ | 35 | 3 | 8,58 | 24 | 68,57 | 8 | 22,86 |
| **Média (%)** | | | 15,82 | | 71,94 | | 12,24 |
| **Desvio Padrão (%)** | | | 13,09 | | 22,84 | | 13,50 |

Fonte: As autoras (2021)

$*$: A amostragem não foi interrompida

Pode-se concluir que aconteceu a mesma coisa para quando utilizou-se *priori* uniforme, para a maioria dos lotes não houve a interrupção da amostragem e para os que houve, estas foram muito precoces, fazendo com que a estimativa não seja muito precisa. E o motivo pelo qual ocorreu isso foi semelhante ao anterior, a organização dos dados, possuindo uma classe discrepante e também a pouca quantidade de sementes disponíveis para a análise.

Realizou-se também a estimação sequencial bayesiana considerando outras três classes: sementes sem danos, sementes que possuem variações de densidade, e as que possuem outros tipos de danos, utilizando a *priori* também obtida através do artigo Javorski e Cicero (2017): $Dirichlet(0,45; 0,04225; 0,00775)$, assim como está apresentada na Tabela 5.13. Os resultados estão na Tabela 5.21:

Tabela 5.21 – Estimativas sequenciais bayesianas das proporções de acordo com as três classificações: $\hat{\boldsymbol{p}}_1$ **:** sementes sem danos, $\hat{\boldsymbol{p}}_2$ **:** variações de densidade, $\hat{\boldsymbol{p}}_3$ **:** outros tipos de danos, para *priori* da literatura

| **Lotes** | $\boldsymbol{m}_{seq}$ | **Tipos de danos** | | | | | |
|---|---|---|---|---|---|---|---|
| | | $\hat{\boldsymbol{p}}_1$**: Sem danos** | | $\hat{\boldsymbol{p}}_2$**: Variações de densidade** | | $\hat{\boldsymbol{p}}_3$**: Outros tipos de danos** | |
| | | $\boldsymbol{m_1}$ | $\hat{\boldsymbol{p}}_{1\,seq}$ **(%)** | $\boldsymbol{m_2}$ | $\hat{\boldsymbol{p}}_{2\,seq}$ **(%)** | $\boldsymbol{m_3}$ | $\hat{\boldsymbol{p}}_{3\,seq}$ **(%)** |
| **1** | 5 | 5 | 99,99 | 0 | 0,005 | 0 | 0,005 |
| **2** | 5 | 5 | 99,99 | 0 | 0,005 | 0 | 0,005 |
| **3** | 5 | 5 | 99,99 | 0 | 0,005 | 0 | 0,005 |
| **4** | 141 | 125 | 88,65 | 15 | 10,64 | 1 | 0,71 |
| **⋮** | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| **98** | 5 | 5 | 99,99 | 0 | 0,005 | 0 | 0,005 |
| **99** | 5 | 5 | 99,99 | 0 | 0,005 | 0 | 0,005 |
| **100** | 170 | 141 | 82,94 | 21 | 12,35 | 8 | 4,71 |
| **Média (%)** | | | 89,33 | | 7,43 | | 3,24 |
| **Desvio Padrão (%)** | | | 9,27 | | 7,09 | | 3,84 |

Fonte: As autoras (2021)

Pode-se notar que assim como aconteceu para *priori* uniforme, as estimativas também foram melhores para essa organização de dados, mas ainda, ao compará-las com as estimativas obtidas utilizando a *priori* uniforme, pode-se concluir que essas foram melhores, pois a *priori* da literatura influenciou na *posteriori* de forma à acarretar em uma superestimação. Assim, os valores das estimavas ficaram bem altos, e ainda aumentou o número de lotes para os quais a amostragem parou rapidamente, com um tamanho amostral de 5 sementes.

# 6 CONCLUSÕES

Diante do exposto, conclui-se que foi possível determinar um critério de parada para o procedimento de estimação sequencial bayesiana dos parâmetros da distribuição multinomial, para *prioris* conjugadas de Dirichlet, utilizando equações de programação dinâmica.

Além disso, pode-se observar que a aplicação para estimar a proporção de danos em sementes de milho considerando as três abordagens: frequentista, bayesiana e sequencial bayesiana, foi adequada. No entanto, no caso da abordagem bayesiana, tanto de amostragem fixa quanto sequencial, vale destacar a importância da escolha da distribuição *a priori*, pois esta influenciará nas estimativas. O uso de artigos da área de sementes de milho para encontrar os hiperparâmetros das distribuições *a priori* foram eficientes, ocasionando na maioria dos casos, estimativas mais precisas.

No procedimento de estimação sequencial bayesiana, observou-se que o modo de organizar os dados é importante, pois no qual considerou-se apenas sementes danificadas, a maioria dos lotes não interromperam a amostragem, pois as sementes foram insuficientes. Mas para o outro tipo de organização, as estimativas foram mais precisas e obteve-se a vantagem da redução do tamanho amostral na maioria dos lotes avaliados.

Ainda na estimação sequencial bayesiana, observou-se também que nos lotes em que houve interrupção da amostragem precocemente, é porque existia uma classe muito discrepante em relação às outras, e portanto, os resultados foram acumulados e a tendência é que a análise acabe rapidamente, prejudicando assim o procedimento. Além do mais, quando há classes muito discrepantes, os resultados mostraram que o custo não é o principal fator na tomada de decisão, mas sim a variação entre as classes e a quantidade de dados disponíveis para análise.

Portanto, concluiu-se que para proporções mais extremas, onde por exemplo eram muito altas (99%) ou muito baixas (1%), obteve estimativas mais distantes, mas para os lotes em que as proporções são intermediárias, mais distribuídas entre as classes, as estimativas sequencias bayesianas foram melhores, obtendo sucesso em todos os casos avaliados, para as duas *prioris* utilizadas, reduzindo o tamanho amostral.

## 6.1 Trabalhos futuros

A estimação sequencial bayesiana para a distribuição multinomial foi realizada através de planilhas no Microsoft Excel®, o que foi eficaz para o objetivo do trabalho. No entanto, como uma possibilidade para futuros trabalhos, propõe-se criar uma ferramenta de fácil manuseio, como um aplicativo interativo, em algum software como R ou Python. Permitindo aos usuários aplicarem a teoria de estimação sequencial bayesiana com facilidade, e que esta possa ser utilizada em outras áreas, resultando assim em avanços relacionados a aspectos computacionais da tomada de decisão sob incerteza.

Além disso, propõe-se também como trabalhos futuros, verificar a compatibilidade entre as abordagens através de um método de comparação, como pelo erro percentual, análise de tendências, entre outros.

# REFERÊNCIAS


ALVO, M. Bayesian sequential estimation. **The Annals of Statistics**, p. 955–968, 1977.

AMARAL, J. B. d.; MARTINS, L.; FORTI, V. A.; CÍCERO, S. M.; FILHO, J. M. Teste de raios x para avaliação do potencial fisiológico de sementes de ipê-roxo. **Revista Brasileira de Sementes**, v. 33, p. 601–607, 2011.

ANDRADE, D. F.; OGLIARI, P. **Estatística para as ciências agrárias e biológicas, com noções de experimentação**. 3. ed. Florianópolis: Editora da UFSC, 2017.

ANDRADE, R. V. d.; BORBA, C. S. Fatores que afetam a qualidade das sementes. **Tecnologia para produção de sementes de milho**, EMBRAPA - Centro Nacional de Pesquisa de Milho e Sorgo, v. 19, 1993.

ARMITAGE, P. Numerical studies in the sequential estimation of a binomial parameter. **Biometrika**, v. 45, n. 1/2, p. 1–15, 1958.

ASSIS, R. C. d. **Inferência em modelos de mistura via algoritmo EM estocástico modificado**. Dissertação (Mestrado), 2017.

AVETISYAN, M.; FOX, J.-P. The dirichlet-multinomial model for multivariate randomized response data and small samples. **Psicologica: International Journal of Methodology and Experimental Psychology**, v. 33, n. 2, p. 362–390, 2012.

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.

BALLARIS, A. de L.; MACHADO, J. da C.; CARVALHO, M. L. M. de; CAVARIANI, C. Sequential sampling of soybean and beans seeds for *sclerotinia sclerotiorum* detection (Lib.) DeBary. **Journal of Seed Science**, v. 36, n. 3, p. 295–304, 2014.

BÁNYAI, J.; BARABÁS, J. **Handbook on statistics in seed testing**. Switzerland: International Seed Testing Association Bassersdorf, 2002.

BARNDORFF-NIELSEN, O. E.; JØRGENSEN, B. Some parametric models on the simplex. **Journal of multivariate analysis**, v. 39, n. 1, p. 106–116, 1991.

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

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **the Journal of machine Learning research**, v. 3, p. 993–1022, 2003.

BOGGS, T. P. **Probabilistic Topic Modeling for Hyperspectral Image Classification**. Tese (Doutorado) — George Mason University, 2019.

BORBA, C. S.; ANDRADE, R. V. de; ANDREOLI, C.; AZEVEDO, J. T. de; OLIVEIRA, A. C. de. Ocorrência de danos mecânicos e qualidade fisiológica de sementes de milho (zea mays l.). **Embrapa Milho e Sorgo-Artigo**, Informativo Abrates, Brasília, DF., v. 6, n. 2/3, p. 57–61, 1996.

BORGES, S. R. d. S.; SILVA, P. P. d.; ARAÚJO, F. S.; SOUZA, F. F. d. J.; NASCIMENTO, W. M. Análise de imagens na avaliação de sementes de tomate durante a maturação. **Journal of Seed Science**, v. 41, n. 1, p. 22–31, 2019.

BOTTINE, S. M. Wald's sequential probability ratio test, Package: SPRT. **CRAN**, 2015. Disponível em: <https://CRAN.R-project.org/package=SPRT>.

BRACK, C. L.; MARSHALL, P. Sequential sampling and modelling for mean dominant height estimation. **Australian Forestry**, v. 53, n. 1, p. 41–46, 1990.

BRAGACHINI, M. A.; BONETTO, L. A.; BONGIOVANNI, R. G.; CASINI, C. Maiz: cosecha, secado y almacenamiento. 1992.

BRANDÃO-JUNIOR, D. S.; DINIZ, A. R.; CARVALHO, M. L. M.; VIEIRA, M. G. G. C.; OLIVEIRA, M. S.; OLIVEIRA, J. A. Avaliação de danos mecânicos e seus efeitos na qualidade fisiológica de sementes de milho. **Revista Brasileira de Sementes**, v. 21, n. 2, p. 53–58, 1999.

BRASIL. **Regras para análise de sementes**. Brasília, DF. Brasil: Ministério da Agricultura, Pecuária e Abastecimento. Secretaria de Defesa Agropecuária. MAPA/ACS, 2009.

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.

CARVALHO, L. R. d.; CARVALHO, M. L. M. d.; DAVIDE, A. C. Utilização do teste de raios x na avaliação da qualidade de sementes de espécies florestais de lauraceae. **Revista Brasileira de Sementes**, v. 31, n. 4, p. 57–66, 2009.

CARVALHO, M. L. M.; CAMARGO, R. Aspectos bioquímicos da deterioração de sementes. **Informe Abrates**, v. 13, n. 1, p. 66–88, 2003.

CARVALHO, M. L. M. d.; SILVA, C. D. d.; OLIVEIRA, L. M. d.; SILVA, D. G.; CALDEIRA, C. M. Teste de raios x na avaliação da qualidade de sementes de abóbora. **Revista Brasileira de Sementes**, v. 31, p. 221–227, 2009.

CARVALHO, M. L. M. de; AELST, A. C. van; ECK, J. W. van; HOEKSTRA, F. A. Pre-harvest stress cracks in maize (zea mays l.) kernels as characterized by visual, x-ray and low temperature scanning electron microscopical analysis: effect on kernel quality. **Seed Science Research**, v. 9, n. 3, p. 227–236, 1999.

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

CHEN, S.-Y. Restricted risk bayes estimation for the mean of the multivariate normal distribution. **Journal of Multivariate Analysis**, v. 24, n. 2, p. 207–217, 1988.

CICERO, S. M.; JUNIOR, H. L. B. Avaliação do relacionamento entre danos mecânicos e vigor, em sementes de milho, por meio da análise de imagens. **Revista Brasileira de Sementes**, v. 25, p. 29–36, 2003.

CONAB – Companhia Nacional de Abastecimento. **Série histórica das safras – Grãos por produtos.** Distrito Federal: Brasília., 2021. Disponível em: <https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras>.

CRUZ, J. C. C.; ALVARENGA, R. C.; NOVOTNY, E. H.; FILHO, I. A. P.; SANTANA, D. P.; PEREIRA, F. T. F.; HERNANI, L. C. **Cultivo do milho**. Sete Lagoas: Embrapa Milho e Sorgo, 2010. v. 1.

DODGE, H. F.; ROMIG, H. G. A method of sampling inspection. **The Bell System Technical Journal**, Nokia Bell Labs, v. 8, n. 4, p. 613–631, 1929.

EHLERS, R. S. Inferência bayesiana. **Departamento de Matemática Aplicada e Estatıstica, ICMC-USP**, v. 64, 2011.

ESTEFANEL, V.; BARBIN, D. Amostragem seqüencial baseada no teste seqüencial da razão de probabilidades e seu uso na determinação da época de controle das lagartas da soja no Estado do Rio Grande do Sul. **Revista do Centro de Ciências Rurais**, v. 9, n. 1, 1979.

FENOY, M. M. The invariant optimal sampling plan in a sequentially planned decision procedure. **Sequential Analysis**, Taylor & Francis, v. 36, n. 2, p. 194–209, 2017.

FERNANDES, J. S.; SILVA, D. F. da; SANTOS, H. O. dos; PINHO, É. V. d. R. V. Teste de raios x na avaliação da qualidade de sementes de frutos de fisális em diferentes estádios de desenvolvimento. **Revista de Ciências Agroveterinárias**, v. 15, n. 2, p. 165–168, 2016.

FILHO, S. M.; SILVA, F. F.; CARNEIRO, A. P. S.; MUNIZ, J. A. Abordagem bayesiana das curvas de crescimento de duas cultivares de feijoeiro. **Ciência Rural**, v. 38, n. 6, p. 1516–1521, 2008.

FORTI, V. A.; CICERO, S. M.; PINTO, T. L. F. Avaliação da evolução de danos por"umidade"e redução do vigor em sementes de soja, cultivar tmg113-rr, durante o armazenamento, utilizando imagens de raios x e testes de potencial fisiológico. **Revista Brasileira de Sementes**, v. 32, p. 123–133, 2010.

FREEMAN, P. R. Sequential estimation of the size of a population. **Biometrika**, v. 59, n. 1, p. 9–17, 1972.

GELMAN, A.; CARLIN, J. B.; STERN, H. S.; DUNSON, D. B.; VEHTARI, A.; RUBIN, D. B. **Bayesian data analysis**. 3. ed. London: Chapman & Hall, 2013.

GEORGES, M. R. R. Amostragem estratificada e distribuição multinomial na pesquisa de opinião: aplicação com alunos de administração. **Revista de Empreendedorismo, Negócios e Inovação**, v. 4, n. 1, p. 42–54, 2019.

GHOSH, M.; MUKHOPADHYAY, N.; SEN, P. K. **Sequential estimation**. New York: John Wiley & Sons, 1997. v. 904.

GIRARDIN, P.; CHAVAGNAT, A.; BOCKSTALLER, C. Détermination des caractéristiques des semences de maïs grâce à la radiographie aux rayons x. **Seed Science and Technology**, v. 21, n. 3, p. 545–551, 1993.

GOMES, G. S. S.; CRIBARI-NETO, F.; VASCONCELLOS, K. L. P. Uma aplicação de influência para a distribuição dirichlet. **Revista Brasileira de Estatística**, Rio de Janeiro, v. 69, n. 231, p. 7–32, 2008.

GOVINDARAJULU, Z. **Sequential statistics**. Singapura: World Scientific, 2004.

ISTA. International rules for seed testing. rules 1985. **Seed science and technology**, v. 13, n. 2, p. 299–513, 1985.

JAVORSKI, M.; CICERO, S. M. Utilização de raios x na avaliação da morfologia interna de sementes de sorgo. **Brazilian Journal of Maize and Sorghum**, v. 16, n. 2, p. 310–318, 2017.

JEROMINI, T. S.; MARTINS, C. C.; PEREIRA, F. E. C. B.; GOMES, F. G. The use of x-ray to evaluate brachiaria brizantha seeds quality during seed processing1. **Revista Ciência Agronômica**, v. 50, p. 439–446, 2019.

JONES, P. W. A note on the bayesian sequential estimation of a binomial parameter. **Biometrika**, p. 642–644, 1974.

JONES, P. W. Bayes Sequential Estimation of Multinomial Parameters. **Mathematische Operationsforschung und Statistik**, v. 7, n. 1, p. 123–127, 1976.

JONES, P. W.; MADHI, S. A. Bayesian sequential methods for choosing the best multinomial cell: some simulation results. **Statistical Papers**, v. 29, n. 1, p. 125–132, 1988.

JUNIOR, F. G. G.; CICERO, S. M. Análise de raios x para a avaliação de injúrias mecânicas em sementes de milho doce. **Revista Brasileira de Sementes**, v. 34, n. 1, p. 78–85, 2012.

KARUNAMUNI, R.; PRASAD, N. Empirical bayes sequential estimation of binomial probabilities. **Communications in Statistics-Simulation and Computation**, v. 32, n. 1, p. 61–71, 2003.

KINAS, P. G.; ANDRADE, H. A. **Introdução à análise bayesiana (com R)**. 2. ed. Porto Alegre: Consultor Editorial, 2020.

LINDLEY, D. V. The Bayesian analysis of contingency tables. **The Annals of Mathematical Statistics**, p. 1622–1643, 1964.

LINDLEY, D. V.; BARNETT, B. Sequential sampling: two decision problems with linear losses for binomial and normal random variables. **Biometrika**, v. 52, n. 3/4, p. 507–532, 1965.

LOPES, M. M.; VIEIRA, M. d. G. G. C. Amostragem sequencial e microssatelites na avaliação da qualidade genética em lotes de sementes de milho. **Bioscience Journal**, v. 30, n. 3, p. 262–271, 2014.

MATRANGOLO, W. J. R.; CRUZ, I.; LÚCIA, T. M. C. D. Insetos fitófagos presentes em estilos-estigma e espigas de milho e avaliação de dano. **Pesquisa Agropecuária Brasileira**, v. 32, n. 8, p. 773–779, 1997.

MCCULLAGH, P.; NELDER, J. A. **Generalized Linear Models**. 2. ed. London: Chapman & Hall, 1989.

MONDO, V. H. V.; CICERO, S. M. Análise de imagens na avaliação da qualidade de sementes de milho localizadas em diferentes posições na espiga. **Revista Brasileira de Sementes**, v. 27, p. 9–18, 2005.

MONDO, V. H. V.; JUNIOR, F. G. G.; PUPIM, T. L.; CICERO, S. M. Avaliação de danos mecânicos em sementes de feijão por meio da análise de imagens. **Revista Brasileira de Sementes**, v. 31, p. 27–35, 2009.

MOOD, A. M.; GRAYBILL, F. A.; BOES, D. C. **Introduction to the theory of statistics**. 3. ed. Singapore: McGraw-Hill International, 1974.

MURPHY, K. P. Binomial and multinomial distributions. **University of British Columbia, Tech. Rep**, 2006.

NEIVA, A. Sobre a interpretação frequentista de probabilidade. **Cognitio-Estudos: revista eletrônica de filosofia**, v. 16, n. 2, p. 233–244, 2019.

OLIVEIRA, J. P. D. D.; SILVA, T. T. D. Sobre as distribuições binomial e multinomial. **Revista de Matemática**, v. 5, n. 1, p. 1–28, 2018.

OWEN, R. J. The optimum design of a two-factor experiment using prior information. **The Annals of Mathematical Statistics**, v. 41, n. 6, p. 1917–1934, 1970.

PAULINO, C. D.; TURKMAN A. A., M. B.; SILVA, G. L. **Estatística bayesiana**. 2. ed. Lisboa: Fundação Calouste Gulbenkian, 2018.

PENTEADO, S. d. R. C.; OLIVEIRA, E. B. d.; IEDE, E. T. Utilização da amostragem seqüencial para avaliar a eficiência do parasitismo de *deladenus (beddingia) siricidicola (nematoda: neotylenchidae)* em adultos de *sirex noctilio (hymenoptera: siricidae)*. **Ciência Florestal**, v. 18, n. 2, p. 223–231, 2008.

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.

PIRES, M. C.; NUNES, L. S. Estimação bayesiana no modelo multinomial com erros de classificação e classficações repetidas. **Revista da Estatística da Universidade Federal de Ouro Preto**, v. 3, n. 2, p. 126–136, 2014.

PLANT, R. E.; WILSON, L. T. A bayesian method for sequential sampling and forecasting in agricultural pest management. **Biometrics**, p. 203–214, 1985.

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, 2020. Disponível em: <http://www.R-project.org/>.

REIS, R. L. d.; MUNIZ, J. A.; SILVA, F. F.; SÁFADI, T.; AQUINO, L. H. d. Abordagem bayesiana da sensitividade de modelos para o coeficiente de endogamia. **Ciência Rural**, Santa Maria, v. 39, n. 6, p. 1752–1759, 2009.

RESENDE, J. M.; RESENDE, M.; SANTOS, C. M. B. dos; BRIGHENTI, C. R. G. Software para estimação sequencial bayesiana da proporção (pp. 181). **Revista da Estatística da Universidade Federal de Ouro Preto**, v. 2, 2012.

ROSSETTO, C. B.; GONÇALVES, F. d. O. Equidade na educação superior no Brasil: uma análise multinomial das políticas públicas de acesso. **Dados**, v. 58, n. 3, p. 791–824, 2015.

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

SHIMIZU, T. K. O.; ACHCAR, J. A.; TARUMMOTO, M. Análise estatística de dados composicionais longitudinais. **Rev. Bras. Biom**, v. 32, n. 1, p. 42–58, 2014.

SILVA, A. H. L. da. Estimativa de proporções em questões politômicas. **Revista do TCU**, n. 125, p. 18–27, 2012.

SPARKS, K. J.; COX, H. C.; MCMILLIAN, W. W.; BURTON, R. L. Damage to corn the pink scavenger caterpillar and its relationship to corn earworm and rice weevil damage. **Journal of Economic Entomology**, v. 59, n. 4, p. 931–934, 1966.

TEIXEIRA, E. F.; CICERO, S. M.; NETO, D. D. Noções básicas sobre imagens digitais: captura, processamento e reconhecimento voltados para a pesquisa em tecnologia de sementes. **Informativo ABRATES**, v. 13, n. 1, p. 59–65, 2003.

THOMPSON, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. **Biometrika**, JSTOR, v. 25, n. 3/4, p. 285–294, 1933.

TIMPANI, V. D.; NASCIMENTO, T. E. C. do. Uma breve introdução à estatística bayesiana aplicada ao melhoramento genético animal. **Embrapa Amazônia Oriental-Documentos (INFOTECA-E)**, 2015.

WALD, A. **Sequential Analysis**. New York: John Willey & Sons, 1947.

WALD, A.; WOLFOWITZ, J. Optimum character of the sequential probability ratio test. **The Annals of Mathematical Statistics**, v. 19, n. 3, p. 326–339, 1948.

WILKS, S. S. **Mathematical statistics**. New York: J.Wiley, 1962.

ZAMBA, K. D.; TSIAMYRTZIS, P. Sequential detection framework for real-time biosurveillance based on shiryaev-roberts procedure with illustrations using covid-19 incidence data. **Sequential Analysis**, Taylor & Francis, p. 1–0, 2021.

ZHUANG, Y.; BHATTACHARJEE, D. Minimum risk point estimation of the size of a finite population under mark–recapture strategy. **Sequential Analysis**, Taylor & Francis, p. 1–8, 2021.