**Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”**

**Análise de agrupamento de semeadoras manuais quanto à
distribuição do número de sementes**

Patricia Peres Araripe

Dissertação apresentada para obtenção do título
de Mestra em Ciências. Área de concentração:
Estatística e Experimentação Agronômica

**Piracicaba
2016**

Patricia Peres Araripe
Licenciada em Matemática

**Análise de agrupamento de semeadoras manuais quanto à distribuição do
número de sementes**

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientador:
Prof. Dr. **SILVIO SANDOVAL ZOCCHI**

Dissertação apresentada para obtenção do título de Mestra
em Ciências. Área de concentração: Estatística e Experi-
mentação Agronômica

**Piracicaba
2016**

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA - DIBD/ESALQ/USP

Araripe, Patricia Peres

Análise de agrupamento de semeadoras manuais quanto à distribuição do número de sementes / Patricia Peres Araripe. - - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2016.
98 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura "Luiz de Queiroz".

1. Semeadora manual 2. Teste da razão de verossimilhanças 3. Fator de Bayes
4. Análise de agrupamentos I. Título

CDD 631.33
A662a

"Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor"

DEDICATÓRIA

A minha família, noivo e amigos

por acreditarem na minha capacidade de realizar este trabalho.

Dedico.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus por ser minha fortaleza, por toda fé e perseverança que me fazem vencer os obstáculos da minha caminhada.

Aos meus pais, Raimundo e Ligia, por todos os ensinamentos de vida, pelo incentivo em buscar e não desistir dos meus sonhos, por todo o amor dedicado e que se reflete na pessoa em que me tornei. As minhas irmãs Priscila e Tatiana por toda amizade, companheirismo e apoio que são essenciais na minha vida.

Agradeço, em especial, ao meu noivo Mateus por estar sempre ao meu lado, nos momentos de felicidade e de tristeza, por todo cuidado e carinho que sempre teve comigo, por todas as palavras de incentivo que foram muitas, enfim, poderia descrever tudo de maravilhoso que ele fez e representa, mas resumo isso em uma única palavra: amor.

As minhas amigas de longa data Adriana, Juliana e Tatiane por tantos momentos maravilhosos que compartilhamos e por mostrarem que não existe distância quando a amizade é verdadeira.

Agradeço a Valiana, “Cabeça”, por ser minha dupla nesse mestrado, pela ajuda e força nas dificuldades encontradas. Além disso, agradeço ao Erasnilson por ser essa pessoa do bem, sempre disposto a ajudar as pessoas.

As minhas companheiras de apartamento Cristiane e Sâmala por todo incentivo, pelas tantas conversas que serviram de motivação nessa trajetória e por todo o carinho que me deram e que é totalmente recíproco.

Ao meu orientador, Prof. Dr. Silvio Sandoval Zocchi, por toda a orientação, paciência ao sentar inúmeras vezes comigo, seus ensinamentos e por acreditar na realização desta pesquisa.

Agradeço aos docentes, funcionários e alunos do programa por todo o conhecimento que passaram, pela atenção e por demonstrarem sempre disposição em ajudar.

Por fim, a CNPQ pela concessão da bolsa ao longo desse mestrado.

SUMÁRIO

RESUMO	9
ABSTRACT	11
LISTA DE FIGURAS	13
LISTA DE TABELAS	15
LISTA DE ABREVIATURAS	17
1 INTRODUÇÃO	19
2 REVISÃO BIBLIOGRÁFICA	21
2.1 Semeadora manual	21
2.2 Tabelas de contingência	24
2.2.1 Caselas vazias e pequenas frequências em tabelas de contingência	25
2.3 Inferência clássica	26
2.3.1 Método da máxima verossimilhança	26
2.3.2 Método do mínimo-qui-quadrado	27
2.3.3 Teste da razão de verossimilhanças	28
2.3.4 Ajuste do nível descritivo (valor-p) para inferência múltipla	29
2.3.4.1 Bonferroni	30
2.3.4.2 Sidák	30
2.4 Inferência bayesiana	31
2.4.1 Distribuições a priori	32
2.4.1.1 Prioris não-informativas	33
2.4.1.2 Prioris conjugadas	34
2.4.2 Distribuição preditiva a posteriori	34
2.4.3 Fator de Bayes	34
2.5 Análise de agrupamentos	35
2.5.1 Métodos hierárquicos	38
2.5.1.1 Métodos aglomerativos	39
2.5.1.2 Métodos divisivos	41
2.5.2 Métodos não-hierárquicos	41
2.5.3 Determinação do número de grupos	42
2.5.4 Coeficiente de correlação cofenética como critério para avaliar métodos de agrupamento	43

3 MATERIAL E MÉTODOS	45
3.1 Descrição dos dados	45
3.2 Métodos	46
4 RESULTADOS E DISCUSSÃO	59
4.1 Comparações das semeadoras duas a duas por meio dos métodos descritivos . . .	60
4.2 Inferência Clássica	61
4.2.1 Semeadoras sem zeros amostrais	62
4.2.2 Todas as semeadoras	63
4.2.2.1 Agrupar as caselas	63
4.2.2.2 Adicionar uma constante as caselas	68
4.3 Inferência bayesiana	71
4.3.1 Fator de Bayes	72
4.4 Análise de Agrupamentos	75
4.4.1 Análise de agrupamentos na abordagem clássica	76
4.4.1.1 Análise de agrupamentos para semeadoras sem zeros amostrais	76
4.4.1.2 Análise de agrupamentos para todas as semeadoras agrupando-se as caselas . .	77
4.4.1.3 Análise de agrupamentos para todas as semeadoras com adição de constantes nas caselas	79
4.4.2 Análise de agrupamentos na abordagem bayesiana	82
5 CONCLUSÕES	83
5.1 Trabalhos futuros	83
REFERÊNCIAS	85
APÊNDICE	91

RESUMO

Análise de agrupamento de semeadoras manuais quanto à distribuição do número de sementes

A semeadora manual é uma ferramenta que, ainda nos dias de hoje, exerce um papel importante em diversos países do mundo que praticam a agricultura familiar e de conservação. Sua utilização é de grande importância devido a minimização do distúrbio do solo, exigências de trabalho no campo, maior produtividade sustentável entre outros fatores. De modo a avaliar e/ou comparar as semeadoras manuais existentes no mercado, diversos trabalhos têm sido realizados, porém considerando somente medidas de posição e dispersão. Neste trabalho é utilizada, como alternativa, uma metodologia para a comparação dos desempenhos das semeadoras manuais. Neste caso, estimou-se as probabilidades associadas a cada categoria de resposta e testou-se a hipótese de que essas probabilidades não variam para as semeadoras quando comparadas duas a duas, utilizando o teste da razão das verossimilhanças e o fator de Bayes nos paradigmas clássico e bayesiano, respectivamente. Por fim, as semeadoras foram agrupadas considerando, como medida de distância, a medida de divergência J-divergência na análise de agrupamento. Como ilustração da metodologia apresentada, são considerados os dados para a comparação de quinze semeadoras manuais de diferentes fabricantes analisados por Molin, Menegatti e Gimenez (2001) em que as semeadoras foram reguladas para depositarem exatamente duas sementes por golpe. Inicialmente, na abordagem clássica, foram comparadas as semeadoras que não possuíam valores nulos nas categorias de resposta, sendo as semeadoras 3, 8 e 14 as que apresentaram melhores comportamentos. Posteriormente, todas as semeadoras foram comparadas duas a duas, agrupando-se as categorias e adicionando as constantes 0,5 ou 1 à cada categoria de resposta. Ao agrupar categorias foi difícil a tomada de conclusões pelo teste da razão de verossimilhanças, evidenciando somente o fato da semeadora 15 ser diferente das demais. Adicionando 0,5 ou 1 à cada categoria não obteve-se, aparentemente, a formação de grupos distintos, como a semeadora 1 pelo teste diferiu das demais e apresentou maior frequência no depósito de duas sementes, o exigido pelo experimento agrônômico, foi a recomendada neste trabalho. Na abordagem bayesiana, utilizou-se o fator de Bayes para comparar as semeadoras duas a duas, no entanto as conclusões foram semelhantes às obtidas na abordagem clássica. Finalmente, na análise de agrupamento foi possível uma melhor visualização dos grupos de semeadoras semelhantes entre si em ambas as abordagens, reafirmando os resultados obtidos anteriormente.

Palavras-chave: Semeadora manual; Teste da razão de verossimilhanças; Fator de Bayes; Análise de agrupamentos

ABSTRACT

Cluster analysis of manual planters according to the distribution of the number of seeds

The manual planter is a tool that today still has an important role in several countries around the world, which practices family and conservation agriculture. The use of it has importance due to minimizing soil disturbance, labor requirements in the field, most sustainable productivity and other factors. In order to analyze and/or compare the commercial manual planters, several studies have been conducted, but considering only position and dispersion measures. This work presents an alternatively method for comparing the performance of manual planters. In this case, the probabilities associated with each category of response has estimated and the hypothesis that these probabilities not vary for planters when compared in pairs evaluated using the likelihood ratio test and Bayes factor in the classical and bayesian paradigms, respectively. Finally, the planters were grouped considering as a measure of distance, the divergence measure J-divergence in the cluster analysis. As an illustration of this methodology, the data from fifteen manual planters adjusted to deposit exactly two seeds per hit of different manufacturers analyzed by Molin, Menegatti and Gimenez (2001) were considered. Initially, in the classical approach, the planters without zero values in response categories were compared and the planters 3, 8 and 14 presents the better behavior. After, all the planters were compared in pairs, grouping categories and adding the constants 0,5 or 1 for each response category. Grouping categories was difficult making conclusions by the likelihood ratio test, only highlighting the fact that the planter 15 is different from others. Adding 0,5 or 1 for each category, apparently not obtained the formation of different groups, such as planter 1 which by the test differed from the others and presented more frequently the deposit of two seeds, required by agronomic experiment and recommended in this work. In the Bayesian approach, the Bayes factor was used to compare the planters in pairs, but the findings were similar to those obtained in the classical approach. Finally, the cluster analysis allowed a better idea of similar planters groups with each other in the both approaches, confirming the results obtained previously.

Keywords: Manual planter; Likelihood ratio test; Bayes factor; Cluster analysis

LISTA DE FIGURAS

Figura 1 - Semeadora manual	22
Figura 2 - Exemplo de um dendrograma para a análise de agrupamento de 5 indivíduos	39
Figura 3 - Distribuições de frequências dos números de sementes depositas por acio- namento (golpe) para 15 semeadoras diferentes, de um total de 150 acio- namentos por semeadora	59
Figura 4 - Distribuição observada das quinze semeadoras em um espaço Simplex . .	65
Figura 5 - Associação entre as quinze semeadoras em um espaço Simplex pela união das que possuem comportamentos semelhantes quanto à distribuição do número de sementes depositadas por golpe	67
Figura 6 - Dendrograma para semeadoras sem zeros amostrais, obtido utilizando os métodos: (a) da distância média; (b) do vizinho mais distante e (c) do vizinho mais próximo	76
Figura 7 - Dendrograma para semeadoras com caselas agrupadas, obtido utilizando os métodos: (a) da distância média; (b) do vizinho mais distante e (c) do vizinho mais próximo	78
Figura 8 - Dendrograma para semeadoras manuais com adição de 0,5 nas caselas, obtido utilizando os métodos: (a) da distância média, (b) do vizinho mais distante e (c) do vizinho mais próximo	80
Figura 9 - Dendrograma para semeadoras manuais com adição de 1 nas caselas, ob- tido utilizando os métodos: (a) da distância média, (b) do vizinho mais distante e (c) do vizinho mais próximo	81

LISTA DE TABELAS

Tabela 1 - Tabela de contingência genérica com S subpopulações e R categorias de resposta	24
Tabela 2 - Interpretação do fator de Bayes de acordo com Kass e Raftery (1995) . . .	35
Tabela 3 - Frequências dos números de sementes por golpe de semeadoras manuais de quinze diferentes fabricantes, reguladas para caírem 2 sementes por golpe	45
Tabela 4 - Tabela de contingência $S \times R$	46
Tabela 5 - Frequências absolutas para as semeadoras i e i^*	50
Tabela 6 - Médias, medianas, modas, desvios padrões e amplitudes interquartílicas (AIQ) dos números de sementes por golpe para as 15 semeadoras. (Considerou-se, aqui, que os valores de números de sementes por golpe maiores ou iguais a 5 são exatamente iguais a 5)	60
Tabela 7 - Estimativas de máxima verossimilhança das probabilidades associadas às categorias das semeadoras manuais	62
Tabela 8 - Comparações entre as semeadoras duas a duas e seus respectivos valores-p ajustados pelo método de Sidák	62
Tabela 9 - Frequências dos números de sementes por golpe considerando caselas agrupadas para as 15 semeadoras	64
Tabela 10 - Comparações entre as semeadoras duas a duas com categorias agrupadas e respectivos p-valores ajustados pelo método de Sidák. Em negrito são ressaltados os valores menores ou iguais ao nível de significância $\alpha = 5\%$.	66
Tabela 11 - Comparações entre as semeadoras duas a duas e respectivos valores-p ajustados pelo método de Sidák, adicionando a constante 0,5 às caselas. Em negrito são ressaltados os valores menores ou iguais ao nível de significância $\alpha = 5\%$	69
Tabela 12 - Comparações entre as semeadoras duas a duas e respectivos valores-p ajustados pelo método de Sidák, adicionando a constante 1 às caselas. Em negrito são ressaltados os valores menores ou iguais ao nível de significância $\alpha = 5\%$	70
Tabela 13 - Médias a posteriori para os parâmetros utilizando a priori não-informativa de Jeffreys, dadas por $(n_{ij} + 0,5)/153$	71

Tabela 14 -Médias a posteriori para os parâmetros utilizando a priori uniforme, dadas por $(n_{ij} + 1)/156$	72
Tabela 15 -Valores de $2 \log B_{10}$ considerando a priori não-informativa de Jeffreys. Em negrito são ressaltados os valores maiores que 6	73
Tabela 16 -Valores de $2 \log B_{10}$ considerando a priori uniforme. Em negrito são res- saltados os valores maiores que 6	74
Tabela 17 -Valores de corte e de coeficientes de correlação cofenética conforme o mé- todo de agrupamento utilizado para semeadoras sem zeros amostrais . . .	77
Tabela 18 -Valores de corte e de coeficientes de correlação cofenética conforme o mé- todo de agrupamento utilizado para semeadoras com caselas agrupadas . .	78
Tabela 19 -Valores de corte e de coeficientes de correlação cofenética conforme o mé- todo de agrupamento utilizado para semeadoras com adição da constante 0,5 nas caselas	80
Tabela 20 -Valores de corte e coeficientes de correlação cofenética conforme o método de agrupamento utilizado para semeadoras com adição da constante 1 nas caselas	81

LISTA DE ABREVIATURAS

ESALQ – Escola Superior de Agricultura “Luiz de Queiroz”

EMV – Estimador de Máxima Verossimilhança

FAO – Food and Agriculture Organization of the United Nations

FWE – Familywise Error Rate

IBGE – Instituto Brasileiro de Geografia e Estatística

MCMC – Monte Carlo via cadeias de Markov

TRV – Teste da Razão de Verossimilhanças

1 INTRODUÇÃO

Segundo a FAO (2010), a agricultura de conservação é uma gestão agroecossistema para obter maiores produtividades sustentáveis, aumento dos lucros e da segurança alimentar enquanto melhora o ambiente. Há, no mundo, diversos países que empregam os princípios dessa prática sendo a semeadora manual uma das ferramentas disponíveis para os agricultores minimizarem o distúrbio do solo e diminuírem as exigências de trabalho no campo. Assim, a semeadora manual, também conhecida como “matraca”, desempenha papel importante na agricultura familiar e de conservação.

A maioria da produção agrícola nos países da África, Ásia e América Latina é baseada na agricultura de subsistência praticada por agricultores de baixa renda cujo maquinário para a semeadura é principalmente manual e de tração animal (JOHANSEN et al., 2012).

No Brasil, segundo Wall (2007), estima-se que 200 mil hectares sejam cultivados por pequenos agricultores que exercem a agricultura de conservação, especialmente nos estados do Paraná, Santa Catarina e Rio Grande do Sul. Além disso, conforme o Censo Agropecuário do IBGE (2006), a agricultura familiar, mesmo com pequena disponibilidade de área cultivável, é responsável pela produção de 70% do feijão, 46% do milho, 34% do arroz, dentre outros alimentos.

Para a exploração dessas áreas agrícolas familiares, principalmente aquelas associadas a solos declivosos ou com alta pedregosidade a utilização da semeadora manual é ideal, pois máquinas e implementos de grande porte ou de tração animal para o preparo do solo não são facilmente disponíveis e recomendados.

A semeadora manual é utilizada por agricultores há mais de um século e ainda mantém suas características básicas estruturais e de funcionamento. Além disso, por haver uma carência de equipamentos específicos para o plantio direto as semeadoras vêm sendo construídas ou adaptadas para esse fim, porém com capacidade operacional geralmente baixa variando conforme a cultura e o operador. Em resumo, apesar de sua importância social e econômica, essa ferramenta sofreu poucas melhorias ao longo dos anos e poucos dados são disponíveis a respeito de suas características de desempenho (MOLIN; MENEGATTI; GIMENEZ, 2001). De modo a comparar o desempenho de semeadoras manuais produzidas por diversos fabricantes, entretanto, há alguns trabalhos realizados no

mundo, como os de Ukatu (2001), Wall (2007), Aikins, Bart-Plange e Opoku-Baffour (2010), Johansen et al. (2012), Molin, Menegatti e Gimenez (2001) e Anjos e Drumond (2004). Para essas comparações, no entanto, consideram somente medidas de posição e dispersão, como média e variância, mas duas semeadoras com comportamentos totalmente distintos podem ter os mesmos valores observados das medidas consideradas. Consequentemente, a utilização dessas medidas pode acarretar, na análise de agrupamento, a presença de duas ou mais semeadoras com comportamentos totalmente distintos num mesmo grupo.

Desta forma, desenvolveu-se como metodologia, a análise de agrupamento das semeadoras manuais tomando-se como distância entre duas semeadoras a medida J-divergência e testando-se a hipótese de que duas semeadoras possuem iguais comportamentos quanto ao número de sementes depositadas por golpe por meio da estatística do teste da razão de verossimilhanças e do fator de Bayes nas abordagens clássica e bayesiana, respectivamente.

2 REVISÃO BIBLIOGRÁFICA

No item 2.1, é feita a descrição da semeadora manual como ferramenta utilizada no trabalho de semeadura no Brasil e em diversos países. No item 2.2, são descritas as tabelas de contingências e particularidades a respeito de suas caselas. A inferência clássica, por sua vez, é descrita no item 2.3, sendo apresentados dois possíveis métodos de estimação de parâmetros, um método de seleção de modelos e ajustes do valor descritivo para a inferência múltipla. No item 2.4, é apresentado o paradigma bayesiano, tipos de distribuições a priori, a obtenção da distribuição preditiva a posteriori e, enfim, um critério de seleção de modelos bayesiano. Finalmente, no item 2.5, é feita a descrição da análise de agrupamentos, possíveis métodos de agrupamentos, a determinação do número de grupos e um critério para avaliação dos métodos.

2.1 Semeadora manual

As primeiras semeadoras manuais surgiram na década de 1850, originalmente desenvolvidas para áreas montanhosas e pedregosas, pois anteriormente os agricultores realizavam a semeadura com um pedaço de pau fazendo um buraco no solo onde colocavam as sementes (VOELKER, 2009).

Segundo Mialhe (2012), a semeadura foi uma das primeiras operações agrícolas a serem mecanizadas após a invenção e o desenvolvimento de inúmeras ferramentas manuais que ainda hoje são utilizadas em pequenas áreas de cultivo. Essa operação é uma das etapas mais importantes para se obter um bom rendimento da lavoura, podendo ser realizada com uma semeadora manual, o que possibilita precisão e rapidez na distribuição das sementes (SILVA et al., 2003).

A semeadora manual, como ilustra a Figura 1, é uma máquina agrícola que tem por objetivo promover a dosagem de sementes e de depositá-las em linhas de plantio sob a superfície do solo (COELHO, 1996). Como é transportada manualmente, o operador pode escolher o local onde vai colocar a semente, desviando-a de pedras e outros impedimentos.

Figura 1 – Semeadora manual
Fonte: FAO (2005)

No cenário mundial atual, diante da necessidade de criação de formas de desenvolvimento sustentável e geração de renda para a população mais pobre, a prática da agricultura de conservação e o incentivo à agricultura familiar são alternativas praticadas por pequenos e médios produtores rurais (ROMEIRO FILHO, 2012). Nesses casos, trabalhos como os de semeadura, plantio, tratos culturais, dentre outros, são realizados com o auxílio de ferramentas agrícolas simples como enxadas, ancinhos e semeadoras manuais.

Na África, a agricultura de conservação está sendo adotada por diversos países como Camarões, Gana, Quênia, Tanzânia, entre outros. Em Gana, segundo Aikins, Bart-Plange e Opoku-Baffour (2010), o milho é o alimento básico mais importante produzido e consumido, sendo semeado principalmente de forma manual com enxadas, catanas e sachos. Entretanto, essa operação tem alta exigência de mão de obra e é um trabalho extremamente árduo, o que gera a necessidade de melhores ferramentas de semeadura do milho, que permitam aos agricultores plantarem em menor tempo aumentando, assim, a produtividade e diminuindo o trabalho penoso. Nesse aspecto, as semeadoras manuais são consideradas ferramentas promissoras tanto para reduzir potencialmente o trabalho de semeadura, quanto para a aplicação de fertilizantes (UKATU, 2001).

No Brasil, de acordo com Molin, Menegatti e Gimenez (2001), as semeadoras manuais também vêm sendo utilizadas para a semeadura direta, além de alguns modelos realizarem ambas as operações de semeadura e adubação. Convém ressaltar que o interesse por construir ou adaptar semeadoras manuais para semeadura direta vem da falta de equipamentos de baixo custo no mercado (ALMEIDA, 1993). Nesse aspecto, o plantio

direto da mamona com uma semeadora manual adaptada pela Embrapa Semi-Árido no sudoeste baiano contribuirá para a preservação do meio ambiente, principalmente nas áreas em fase de degradação. Isto se deve a ação mecânica sobre o solo ser bem menor utilizando a semeadora do que com o uso de implementos com tração mecânica ou animal (ANJOS; DRUMOND, 2004).

Apesar dessas vantagens, as semeadoras manuais falham em aspectos ergonômicos, capacidade operacional em campo e precisão na dosagem, resultando em baixa produtividade (MOLIN; D'AGOSTIN, 1996).

Conforme Molin, Menegatti e Gimenez (2001), a qualidade construtiva e acabamentos das semeadoras, em particular dos mecanismos dosadores, são essenciais para a obtenção da regularidade no depósito de sementes no solo e, dessa forma, obter-se um bom desenvolvimento de uma lavoura bem como sua produção.

No entanto, apesar de alguns trabalhos estarem sendo realizados para avaliar e comparar o desempenho das semeadoras manuais, estes os fazem utilizando somente métodos estatísticos descritivos básicos. Molin, Menegatti e Gimenez (2001), por exemplo, avaliaram o desempenho de semeadoras manuais utilizadas por agricultores brasileiros e encontraram grande irregularidade na vazão do mecanismo dosador das diferentes máquinas, sendo poucas as que apresentaram comportamento satisfatório. Chegaram às conclusões, entretanto, comparando-se as médias e as variâncias dos números de sementes caídas por golpe, ou acionamento, das semeadoras.

Utilizando como medida a média de sementes de mamona por cova, Anjos e Drumond (2004), por sua vez, avaliaram o desempenho operacional da semeadora manual adaptada com quatro níveis de sementes (25%, 50%, 75% e 100%) para a semeadura direta de mamona e obtiveram, segundo os autores, bom desempenho na distribuição de sementes.

No início de 2000, muitas unidades de semeadoras manuais foram exportadas para a África. Desde então, várias tentativas foram feitas para produzi-las localmente, onde novas máquinas estão agora sendo desenvolvidas e tendo seus desempenhos avaliados por meio de experimentos (JOHANSEN et al., 2012).

Em um desses experimentos, Aikins, Bart-Plange e Opoku-Baffour (2010) avaliaram o desempenho de 30 semeadoras manuais de um mesmo modelo, produzidas por um fabricante local em Gana. Por meio da média e do intervalo de sementes depositadas pelas semeadoras para cinco variedades de milho, observaram que havia alta variabilidade

no depósito das sementes para cada variedade, considerando a exigência do experimento agrônomo de duas ou três sementes por golpe.

Na maioria dos testes de uniformidade de semeadura, as semeadoras são reguladas para depositarem um certo número fixo de sementes por golpe, sendo acionadas um certo número de vezes e anotadas, para cada semeadora, as frequências de depósito de 0, 1, 2, 3, ... sementes por cova. Os resultados, desta forma, podem ser agrupados numa tabela de contingência, descrita na seção a seguir.

2.2 Tabelas de contingência

Pesquisadores em diversas áreas deparam-se com experimentos em que a resposta de interesse é categórica, ou seja, representam categorias de informação.

Segundo Paulino e Singer (2006), dados categóricos são dados discretos relativos a uma variável definida por meio de um número finito de categorias. A variável resposta é categórica, já a variável explicativa do experimento pode ser tanto categórica quanto contínua.

Os dados categóricos (ou discretos) podem ser dispostos na forma de uma tabela de contingência $S \times R$ em que as S linhas correspondem aos níveis da variável explicativa (ou combinações dos níveis de mais de uma variável explicativa) e as R colunas correspondem às categorias da variável resposta (ou combinações dos níveis de mais de uma variável resposta). De uma forma genérica, uma tabela de contingência bidimensional pode ser representada como a Tabela 1, em que as frequências nas caselas são representadas por n_{ij} , $i = 1, 2, \dots, S$ e $j = 1, 2, \dots, R$. Além disso, n_{i+} , n_{+j} e n são denominados, respectivamente, de total marginal linha, total marginal coluna e tamanho da amostra. Esses totais podem ser considerados como fixos ou aleatórios.

Tabela 1 – Tabela de contingência genérica com S subpopulações e R categorias de resposta

Subpopulação (i)	Categorias de resposta (j)					Total
	1	2	3	...	R	
1	n_{11}	n_{12}	n_{13}	...	n_{1R}	n_{1+}
2	n_{21}	n_{22}	n_{23}	...	n_{2R}	n_{2+}
...	...	...	...	...	...	...
S	n_{S1}	n_{S2}	n_{S3}	...	n_{SR}	n_{S+}
Total	n_{+1}	n_{+2}	n_{+3}	...	n_{+R}	n

2.2.1 Caselas vazias e pequenas frequências em tabelas de contingência

Ao trabalhar com amostras em tabelas de contingência podem-se obter caselas com pequenas frequências ou dois tipos de caselas vazias, tabelas conhecidas como dispersas. O primeiro tipo de casela vazia é conhecido como zero amostral (ou aleatório) e o segundo é chamado de zero estrutural (ou fixo).

Segundo Agresti (2002), tabelas dispersas ocorrem quando o tamanho da amostra n é pequeno ou quando a amostra é grande, mas contém muitas categorias.

Os zeros amostrais correspondem às caselas com frequências nulas, $n_{ij} = 0$, registados na amostra selecionada, mas sendo possível de se obter observações na casela, ou seja, $n_{ij} > 0$, quando o tamanho da amostra é suficientemente grande. Por outro lado, uma casela vazia em que a observação é impossível de ser obtida, independentemente do tamanho da amostra, é chamada de zero estrutural e, assim, não contribui no processo de estimação ou no ajuste de modelos (AGRESTI, 2002).

Caselas vazias e pequenas frequências podem causar problemas com relação à estimação de parâmetros, ao teste da razão de verossimilhanças, à complexidade computacional e às distribuições assintóticas (SUBBIAH; SRINIVASAN, 2008).

Conforme Kraus (2012), uma variedade de propostas têm sido sugeridas para solucionar esses problemas, dentre elas:

- (i) agrupar (ou combinar) as categorias das variáveis explicativas da tabela de contingência;
- (ii) adicionar uma pequena constante às caselas;
- (iii) considerar apenas caselas com frequências observadas ou esperadas que excedam determinados valores;
- (iv) eliminar as caselas que contenham zeros.

No entanto, as alternativas (i), (iii) e (iv) levam à perda de informação e resultados com conclusões eventualmente inapropriadas. Uma outra alternativa é usar os testes exatos ou bootstrap (paramétrico) para se obter a distribuição da estatística de interesse para amostras pequenas. À medida que as dimensões da tabela crescem, entretanto, os testes exatos se tornam inviáveis e é possível aproximar esta distribuição pelos métodos de simulação Monte Carlo (FINKLER, 2010).

Alternativamente, problemas de inferência devido à presença de zeros amostrais ou pequenas frequências de amostragem podem ser tratados no âmbito da análise bayesiana. Além disso, a incorporação de uma informação a priori pode ser essencial em aplicações de algumas áreas (KATERI, 2014).

2.3 Inferência clássica

Seja θ o parâmetro de interesse. A inferência estatística consiste em estimar um valor para o parâmetro desconhecido (ou para uma função do parâmetro desconhecido) baseando-se em um conjunto de valores observados $\mathbf{x} = (x_1, x_2, \dots, x_n)$ de uma variável aleatória X com função de densidade (ou probabilidade) dada por $f_X(x|\theta)$. Uma maneira de se realizar a estimação na abordagem clássica é encontrando um estimador pontual para θ por meio dos métodos de estimação.

Segundo Mood, Graybill e Boes (1974), existem vários métodos para a obtenção de estimadores pontuais. Dentre os métodos encontrados serão descritos o da máxima verossimilhança e o método do mínimo-qui-quadrado.

2.3.1 Método da máxima verossimilhança

Segundo Casella e Berger (2002), o método da máxima verossimilhança é, sem dúvida, o mais popular para se encontrar estimadores. Neste, os estimadores são obtidos a partir da maximização da função de verossimilhança.

Sejam $x_1, x_2, \dots, x_n$ os valores observados de uma amostra aleatória $X_1, X_2, \dots, X_n$ de tamanho n da variável aleatória X com função de densidade (ou de probabilidade) $f_X(x|\theta)$ e θ um escalar pertencente ao espaço uniparamétrico $\Theta \subseteq \mathbb{R}$. Então a função de verossimilhança correspondente à amostra aleatória é dada por

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n f_X(x_i|\theta).$$

Seja $\hat{\theta} = \vartheta(x_1, x_2, \dots, x_n)$, isto é, uma função de $x_1, x_2, \dots, x_n$. Se $\hat{\theta} \in \Theta$ é o valor que maximiza a função de verossimilhança $L(\theta|\mathbf{x})$, então $\hat{\Phi} = \vartheta(X_1, X_2, \dots, X_n)$ é o estimador de máxima verossimilhança (EMV) de θ . O valor $\hat{\theta}$ é a estimativa de máxima verossimilhança de θ e também maximiza o logaritmo da função de verossimilhança denotado

por

$$l(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x}) = \sum_{i=1}^n \log f_X(x_i|\theta)$$

sendo mais simples de se trabalhar algebricamente em muitas situações (e muitas vezes computacionalmente). Assim, de modo a obter a estimativa de máxima verossimilhança $\hat{\theta}$ de θ , deve-se solucionar a equação dada por

$$\frac{dl(\theta|\mathbf{x})}{d\theta} = 0. \quad (1)$$

A solução de (1) fornece um possível candidato a ponto de máximo. Para comprovar que $\hat{\theta}$ corresponde ao ponto de máximo deve-se calcular a segunda derivada da eq. (1) e esta ser negativa, ou seja,

$$\frac{d^2l(\theta|\mathbf{x})}{d\theta^2} \Big|_{\theta=\hat{\theta}} < 0.$$

Para o caso multiparamétrico, isto é, em que $\boldsymbol{\theta}$ é um vetor de parâmetros de dimensão k , $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_k)$, a estimativa de máxima verossimilhança $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ de $\boldsymbol{\theta}$, é obtida por meio da solução do sistema de equações

$$\frac{\partial l(\boldsymbol{\theta}|\mathbf{x})}{\partial \theta_i} = 0, \quad i = 1, 2, \dots, k.$$

Segundo Casella e Berger (2002), a estimativa de máxima verossimilhança de θ é o valor para o qual a amostra observada é mais provável. Deve-se sempre fazer a distinção de que um estimador é uma função da amostra aleatória $X_1, X_2, \dots, X_n$, enquanto que uma estimativa é o valor observado de um estimador quando uma amostra é efetivamente obtida ($X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$).

2.3.2 Método do mínimo-qui-quadrado

Mood, Graybill e Boes (1974) descrevem este método conforme a seguir. Sejam $X_1, X_2, \dots, X_n$ uma amostra aleatória de uma função densidade (ou probabilidade) $f_X(x|\theta)$ e $P_1, P_2, \dots, P_k$ uma partição do intervalo de X . Seja, ainda, $p_j(\theta)$ $j = 1, 2, \dots, k$, a probabilidade de uma observação pertencer à partição (ou célula) P_j . Por exemplo, se $f_X(x|\theta)$ é a função densidade de uma variável aleatória contínua, então $p_j(\theta) = \int_{P_j} f_X(\theta|x) dx$.

Note que $\sum_{j=1}^k p_j(\theta) = 1$.

Seja N_j , $j = 1, 2, \dots, k$, a variável aleatória que denota o número de observações na amostra aleatória que pertencem à célula P_j e $\sum_{j=1}^k N_j = n$ o tamanho da amostra. A seguinte soma é formada

$$\chi^2 = \sum_{j=1}^k \frac{[n_j - np_j(\theta)]^2}{np_j(\theta)}, \quad (2)$$

em que n_j é o valor observado de N_j . Neste caso, o numerador do j -ésimo termo na eq. (2) é o quadrado da diferença entre o valor observado e o valor esperado das observações na célula P_j .

A estimativa de mínimo-qui-quadrado para θ é o valor de $\hat{\theta}$ que minimiza χ^2 . É esse θ , entre todos os possíveis valores de θ , que faz com que o número esperado da observação na célula P_j seja o “mais próximo” do número observado.

O estimador de mínimo-qui-quadrado depende, entretanto, da partição $P_1, P_2, \dots, P_k$ selecionada. Além disso, muitas vezes é difícil identificar qual o valor de $\hat{\theta}$ que minimiza χ^2 , por isso o denominador de (2) é, às vezes, alterado para n_j (se $n_j = 0$, a unidade é utilizada) formando a equação

$$\chi_m^2 = \sum_{j=1}^k \frac{[n_j - np_j(\theta)]^2}{n_j},$$

chamada *mínimo-qui-quadrado modificado*. A estimativa $\hat{\theta}$ do mínimo-qui-quadrado modificado de θ é, então, o valor de θ que minimiza o χ_m^2 modificado.

2.3.3 Teste da razão de verossimilhanças

Um método usual para a comparação de modelos aninhados ajustados pela máxima verossimilhança é o teste da razão de verossimilhança (LEHMANN, 1986).

Segundo West, Welch e Galecki (2007), dois modelos são aninhados quando um dos modelos é um caso especial do outro. Se um modelo mais simples (ou reduzido) é aninhado a um modelo mais completo (ou geral), então todos os parâmetros do modelo reduzido devem estar contidos no modelo completo.

Uma maneira de se decidir entre um modelo completo e um modelo reduzido

é testando a hipótese nula de que os dados são extraídos do modelo reduzido (LEWIS; BUTLER; GILBERT, 2011).

São consideradas as hipóteses H_0 : o modelo reduzido é adequado versus H_1 : o modelo reduzido não é adequado. A estatística do teste da razão de verossimilhanças é dada por

$$TRV = 2 \log \left[\frac{L_2}{L_1} \right] = 2 [\log (L_2) - \log (L_1)],$$

em que L_1 representa a função de verossimilhança para o modelo reduzido, isto é, com menor número de parâmetros e L_2 representa a função de verossimilhança para o modelo completo.

Sejam k_1 e k_2 , respectivamente, o número de parâmetros dos modelos reduzido e completo, em que $k_1 < k_2$. Então a estatística TRV do teste da razão de verossimilhanças segue, assintoticamente, uma distribuição χ^2 com $k_2 - k_1$ graus de liberdade (WEST; WELCH; GALECKI, 2007).

Conforme Lewis, Butler e Gilbert (2011), é usual pressupor que a distribuição χ^2 também permanece, aproximadamente, válida para uma amostra finita desde que o tamanho seja moderadamente grande.

Assim, se o nível descritivo, ou valor-p, associado ao teste é menor do que o limite pré-definido, então rejeita-se H_0 e o modelo completo se ajusta melhor aos os dados do que o modelo reduzido. Caso contrário, não se rejeita H_0 , sendo o modelo reduzido tão bom quanto o modelo completo (POSADA; BUCKLEY, 2004).

O teste da razão de verossimilhança deve ser utilizado, assim, para comparar dois modelos de cada vez.

2.3.4 Ajuste do nível descritivo (valor-p) para inferência múltipla

Testes múltiplos referem-se a testar mais de uma hipótese simultaneamente, este é um subcampo da inferência múltipla ou inferência simultânea. As hipóteses são consideradas como uma família, ou seja, como um conjunto em que os testes de significância e os erros serão controlados conjuntamente (SHAFFER, 1995).

Ao testar uma única hipótese, a probabilidade de se cometer o erro do tipo I, ou seja, de rejeitar a hipótese nula quando ela é verdadeira, é usualmente controlada por um nível de significância α fixado. Quando muitas hipóteses são testadas e cada teste tem um nível significância especificado, a probabilidade de se rejeitar indevidamente pelo menos

uma das hipóteses nulas da família, chamada de *Familywise Error Rate* (FWE), é elevada quanto maior for o número de comparações (WESTFALL; WOLFINGER, 1997).

Segundo Shaffer (1995), vários métodos têm sido propostos para controlar a FWE. Para simplificar a apresentação dos métodos de comparações múltiplas, um conceito que permite a generalização para o contexto múltiplo a partir de um teste de única hipótese é o valor-p ajustado. Por definição, o valor-p ajustado para qualquer hipótese é igual ao menor FWE em que a hipótese seria rejeitada. Por essa razão, os valores-p ajustados são facilmente interpretados como evidências contra as hipóteses nulas correspondentes, quando todos os testes são considerados como uma família (WESTFALL et al., 1999).

Segundo Wright (1992), um valor-p ajustado pode ser comparado diretamente com qualquer nível de significância α estabelecido. Assim, se o valor-p for menor ou igual a α , a hipótese nula será rejeitada. Para a maioria dos métodos, essa regra de decisão controla a FWE.

Os métodos de ajuste *single-step* necessitam de apenas uma etapa para encontrar o valor crítico adequado para todos os testes. São calculados sem referência para os outros testes de hipóteses em consideração. Os ajustes de Bonferroni e Sidák são funções simples dos valores-p e calculados computacionalmente de maneira rápida, mas ambos podem ser conservadores (WESTFALL et al., 1999).

2.3.4.1 Bonferroni

O procedimento de Bonferroni, de acordo com Westfall et al. (1999), rejeita qualquer hipótese nula, H_{0k} , cujo p-valor correspondente, p_k , é menor ou igual a α/m , em que $k = 1, 2, \dots, m$ e m é número de testes.

Isto é equivalente a rejeitar qualquer hipótese nula, H_{0k} , para o qual o valor-p ajustado

$$\tilde{p}_k = \begin{cases} mp_k, & \text{se } mp_k \leq 1; \\ 1, & \text{se } mp_k > 1, \end{cases}$$

é menor ou igual a α . O método de Bonferroni é conservador, mas sempre controla a FWE.

2.3.4.2 Sidák

Para o método de Sidák, segundo Westfall et al. (1999), rejeita-se uma hipótese individual, H_{0k} , se $p_k \leq 1 - (1 - \alpha)^{1/m}$, em que $k = 1, 2, \dots, m$ e m é número de testes.

Isto é equivalente a rejeitar qualquer hipótese nula, H_{0k} , para o qual o valor-p ajustado

$$\tilde{p}_k = 1 - (1 - p_k)^m,$$

é menor ou igual a α . O ajuste de Sidák é uma técnica ligeiramente menos conservadora que a de Bonferroni.

Conforme Westfall et al. (1999), há ainda os métodos de ajuste *stepwise*, que ordenam as hipóteses em step-up (menos significativa para o mais significativa) ou na forma step-down (mais significativa para o menos significativa), em seguida determinam sequencialmente a aceitação ou rejeição das hipóteses nulas. A descrição desses métodos não será feita no presente trabalho.

2.4 Inferência bayesiana

A abordagem clássica dominou a teoria estatística e prática durante a maior parte do século passado, mas as últimas décadas depararam-se com a prática frequente da estatística bayesiana motivada, principalmente, pela disponibilidade de novas técnicas computacionais (GELMAN et al., 2013).

Considere uma quantidade de interesse desconhecida θ pertencente a um espaço uniparamétrico Θ , a informação que temos a respeito do parâmetro θ é resumida por meio de uma distribuição a priori $\pi(\theta)$. Pode-se, então, atualizar esta informação obtendo uma amostra de uma quantidade aleatória relacionada com θ e utilizando o teorema de Bayes. Seja o vetor $\mathbf{x} = (x_1, x_2, \dots, x_n)$ de observações de uma amostra aleatória $X_1, X_2, \dots, X_n$ de tamanho n da variável aleatória X , ou seja, em que X_i , $i = 1, 2, \dots, n$, são independentes e identicamente distribuídos conforme a distribuição de X . Por meio do teorema de Bayes, temos que a distribuição a posteriori de θ é obtida por

$$\pi(\theta|\mathbf{x}) = \frac{L(\theta|\mathbf{x})\pi(\theta)}{\pi(\mathbf{x})} = \frac{L(\theta|\mathbf{x})\pi(\theta)}{\int_{\Theta} L(\theta|\mathbf{x})\pi(\theta) d\theta}, \quad (3)$$

em que $L(\theta|\mathbf{x}) = \prod_{i=1}^n f_X(x_i|\theta)$ é a função de verossimilhança.

Na eq. (3), $\pi(\theta|\mathbf{x})$ é a distribuição a posteriori de θ considerando a amostra aleatória da variável aleatória contínua X , que pode, entretanto, ser discreta.

O denominador em (3) define a verossimilhança marginal dos dados e como

independe do parâmetro θ , esta pode ser considerada como uma constante normalizadora e, assim, a eq. (3) pode ser representada na forma

$$\pi(\theta|\mathbf{x}) \propto L(\theta|\mathbf{x})\pi(\theta),$$

ou, em palavras, tem-se que a distribuição a posteriori é proporcional ao produto da verossimilhança pela distribuição a priori.

O símbolo de proporcionalidade pode ser justificado da seguinte maneira: ao multiplicar a função de verossimilhança por uma constante não alteramos a informação relativa ao parâmetro θ , logo a distribuição a posteriori não será alterada (LEANDRO, 2001).

Conforme Paulino, Turkman e Murteira (2003), toda inferência relativa a um determinado parâmetro é feita utilizando-se a distribuição a posteriori de θ , que pode ser resumida por meio da média, moda, mediana e do intervalo de credibilidade. Entretanto, em muitas situações, a distribuição a posteriori é analiticamente impossível de ser resolvida. Nesses casos, são utilizados métodos de simulação Monte Carlo via cadeias de Markov (MCMC), dentre os quais estão o Amostrador de Gibbs e o algoritmo de Metropolis-Hastings (ver PAULINO; TURKMAN; MURTEIRA, 2003).

2.4.1 Distribuições a priori

A utilização da informação a priori na inferência bayesiana requer a especificação de uma distribuição a priori para o parâmetro de interesse θ . Esta distribuição representa o conhecimento que se tem sobre θ antes da obtenção dos dados (EHLERS, 2011).

Segundo Leandro (2001), esta distribuição pode ser utilizada para representar conhecimento prévio (por exemplo, com base nos resultados ou estimativas dos parâmetros em análises anteriores) ou ignorância (quando pouco ou nada se sabe a respeito do parâmetro).

Serão abordadas algumas formas de especificação da distribuição a priori.

2.4.1.1 Prioris não-informativas

Quando a informação a respeito do parâmetro é inexistente ou vaga é necessário especificar uma priori que não influenciará a distribuição a posteriori e, assim, “deixar que os dados falem por si mesmos”. Essas distribuições são frequentemente chamadas de distribuições a priori não-informativas ou vagas (NTZOUFRAS, 2009).

Nestes casos, pode-se considerar uma distribuição a priori em que todos os possíveis valores de θ são igualmente prováveis, ou seja, uma distribuição a priori uniforme dada por

$$\pi(\theta) \propto k,$$

em que k é uma constante. No entanto, se o intervalo de variação de θ for ilimitado, então esta distribuição a priori é imprópria (não soma ou integra um).

Além disso, segundo Kaplan e Depaoli (2013), a distribuição a priori uniforme não é invariante a transformações simples. Uma transformação de uma priori uniforme pode resultar em uma distribuição a priori que não seja uniforme e, assim, favorecer alguns valores em detrimento de outros.

Ao abordar o problema de invariância associado à distribuição uniforme, Jeffreys (1961) propôs uma classe de prioris não-informativas invariantes a transformações um a um , embora essa classe seja geralmente imprópria (EHLERS, 2011).

A distribuição a priori não-informativa de Jeffreys é dada por

$$\pi(\theta) \propto [I(\theta)]^{1/2},$$

em que $I(\theta)$ é a informação esperada de Fisher para θ obtida por

$$I(\theta) = E \left\{ \left[\frac{d(\log f_X(x|\theta))}{d\theta} \right]^2 \right\}.$$

Se θ for um vetor de um espaço multiparamétrico, então $I(\theta)$ é a matriz de informação de Fisher e, assim, $\pi(\theta) \propto |\det I(\theta)|^{1/2}$.

Deve-se observar que a distribuição a priori pode ser imprópria para se obter uma distribuição a posteriori própria. No entanto, é necessário sempre certificar de que a posteriori é própria antes de fazer qualquer inferência, uma vez que esta é uma função de probabilidade ou função densidade de probabilidade.

2.4.1.2 Prioris conjugadas

De acordo com Kaplan e Depaoli (2013), uma distribuição a priori conjugada é aquela que quando combinada com a função de verossimilhança resulta em uma posteriori que pertença a mesma família de distribuições da priori. Assim, a atualização do conhecimento que se tem sobre θ envolve apenas uma mudança nos hiperparâmetros já que as distribuições permanecem as mesmas (EHLERS, 2011).

Conjugação é formalmente definida como segue (GELMAN et al., 2013). Se $\mathfrak{S}$ é uma classe de distribuições amostrais $f_X(x|\theta)$ e $\wp$ é uma classe de distribuições a priori $\pi(\theta)$ para θ , então a classe $\wp$ é conjugada a $\mathfrak{S}$ se $\forall f_X(\cdot|\theta) \in \mathfrak{S}$ e $\pi(\cdot) \in \wp$ tem-se $\pi(\theta|x) \in \wp$.

O maior interesse, conforme Gelman et al. (2013), está nas famílias de prioris conjugadas naturais, que são obtidas ao considerar a classe $\wp$ como sendo o conjunto de todas as densidades que possuem a mesma forma funcional (núcleo) que a verossimilhança.

2.4.2 Distribuição preditiva a posteriori

Após observar os dados $\mathbf{x} = (x_1, x_2, \dots, x_n)$ pode-se ter o interesse em prever os valores de “futuras” observações, $\mathbf{x}^{pred} = (x_{n+1}, \dots, x_m)$, também relacionados com θ . A distribuição de $\mathbf{x}^{pred}$ é chamada de *distribuição preditiva a posteriori*, pois é a distribuição de dados não observados condicionada ao conjunto observado $\mathbf{x}$ e é definida como

$$\pi(\mathbf{x}^{pred}|\mathbf{x}) = \int_{\Theta} \pi(\mathbf{x}^{pred}, \theta|\mathbf{x}) d\theta = \int_{\Theta} \pi(\mathbf{x}^{pred}|\theta, \mathbf{x}) \pi(\theta|\mathbf{x}) d\theta.$$

Ao assumir independência condicional entre $\mathbf{x}$ e $\mathbf{x}^{pred}$ dado θ , a distribuição preditiva pode ser representada por

$$\pi(\mathbf{x}^{pred}|\mathbf{x}) = \int_{\Theta} \pi(\mathbf{x}^{pred}|\theta) \pi(\theta|\mathbf{x}) d\theta.$$

2.4.3 Fator de Bayes

De acordo com Kass e Raftery (1995), o fator de Bayes é um critério bayesiano para a comparação de modelos. Ao considerar a comparação de modelos como um problema de teste de hipóteses em que o interesse consiste na avaliação da hipótese nula H_0 (que

corresponde a um modelo M_0) contra a hipótese alternativa H_1 (que corresponde a um modelo M_1), então o fator de Bayes avalia a evidência contra H_0 (NTZOUFRAS, 2009). Este é obtido por meio da razão das verossimilhanças marginais de dois modelos e é descrito como

$$B_{10} = \frac{\pi(\mathbf{x}|H_1)}{\pi(\mathbf{x}|H_0)} = \frac{\int_{\Theta_1} L(\theta_1|\mathbf{x}, H_1) \pi(\theta_1|H_1) d\theta_1}{\int_{\Theta_0} L(\theta_0|\mathbf{x}, H_0) \pi(\theta_0|H_0) d\theta_0}, \quad (4)$$

em que $L(\theta_1|\mathbf{x}, H_1)$ é a função de verossimilhança para o parâmetro sob modelo de H_1 e $\pi(\theta_1|H_1)$ é a distribuição a priori para o parâmetro do modelo sob a hipótese H_1 . Analogamente tem-se que $L(\theta_0|\mathbf{x}, H_0)$ é a função de verossimilhança para o modelo de H_0 e $\pi(\theta_0|H_0)$ é a distribuição a priori para o parâmetro do modelo sob a hipótese H_0 .

Como interpretação para os valores de B_{10} , Kass e Raftery (1995) apresentam uma tabela com os graus de evidência contra H_0 de acordo com o valor de B_{10} ou de $2 \log B_{10}$, reproduzida na Tabela 2.

Tabela 2 – Interpretação do fator de Bayes de acordo com Kass e Raftery (1995)

$2 \log B_{10}$	B_{10}	Evidência contra H_0
0 – 2	1 – 3	Desprezível
2 – 6	3 – 20	Positiva
6 – 10	20 – 150	Forte
> 10	> 150	Muito forte

Convém ressaltar que o trabalho de Kass e Raftery (1995), utilizado como referência para a obtenção da Tabela 2, não menciona a inclusão dos pontos extremos nos intervalos e, também, de valores menores do que zero (ao adotar $2 \log B_{10}$) ou menores do que um (ao adotar B_{10}).

Em situações envolvendo modelos mais simples, as verossimilhanças marginais da eq. (4) podem ser calculadas analiticamente. No entanto, em geral, essas integrais não são fáceis de serem resolvidas sendo, assim, necessária a utilização de métodos MCMC para obter soluções aproximadas (BORGATTO, 2004).

2.5 Análise de agrupamentos

A análise de agrupamentos tem por finalidade obter grupos homogêneos por um esquema que possibilite reunir objetos (ou indivíduos) em um determinado número de grupos, de tal forma que exista homogeneidade dentro dos grupos e heterogeneidade entre

os mesmos (JOHNSON; WICHERN, 2007).

Desse modo, conforme Everitt (1974, apud BUSSAB; MIAZAKI; ANDRADE, 1990), dada uma amostra de n objetos (ou indivíduos), cada um deles medido segundo p variáveis, procura-se um esquema de classificação que agrupe os objetos em g grupos. Usualmente pretende-se agrupar objetos semelhantes segundo suas características (variáveis), mas nada impede que o interesse seja o de agrupar variáveis segundo os valores obtidos pelos objetos.

As metodologias existentes são determinadas, basicamente, por uma medida de proximidade e um algoritmo. As medidas de proximidade são quantidades comparativas entre as observações e são definidas com base na natureza das variáveis que caracterizam os objetos em análise e, em geral, os algoritmos descrevem como o método de agrupamento deve ser realizado (MATOS, 2007).

A medida de proximidade pode ser definida como medida de similaridade ou medida de dissimilaridade. Na medida de similaridade quanto maior o valor observado, mais semelhantes serão os objetos, enquanto que na medida de dissimilaridade quanto maior o valor observado, menor a semelhança. De um modo geral, é possível construir uma medida de dissimilaridade a partir de uma similaridade e vice-versa (BUSSAB; MIAZAKI; ANDRADE, 1990).

Quando objetos são agrupados é, normalmente, utilizada alguma medida de distância. Por outro lado, as variáveis são geralmente agrupadas com base em coeficientes de correlação ou medidas similares de associação (JOHNSON; WICHERN, 2007).

Convém ressaltar que uma medida de distância deve satisfazer pelo menos às condições de não-negatividade, simetria e reflexividade (JOUSSELME; MAUPIN, 2012)

De acordo com Cha (2007), a escolha das medidas de distância depende do tipo de medição ou representação dos objetos. A função densidade de probabilidade (ou função de probabilidade) é uma das representações mais populares, sendo medidas de distâncias/divergências utilizadas para quantificar a diferença entre pares de distribuições de probabilidade.

Seja X uma variável aleatória discreta e sejam p e q duas distribuições de probabilidade de X .

Segundo Sriperumbudur et al. (2009), uma das famílias amplamente estudada e bem compreendida de distâncias/divergências entre medidas de probabilidade é a chamada

divergência f de Csiszár, que é definida como

$$D_f(p, q) = \sum_{x \in X} p(x) f\left(\frac{q(x)}{p(x)}\right), \quad (5)$$

em que $f : (0, \infty) \rightarrow \mathbb{R}$ é uma função convexa, tal que $f(1) = 0$. Algumas medidas de divergência populares obtidas escolhendo adequadamente f citadas na literatura são (KUMAR; CHHINA, 2005; CHA, 2007):

1) **Kullback-Leibler**, em que $f(u) = -\log_2(u)$, definida por

$$KL(p, q) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{q(x)},$$

em que é utilizada a função logarítmica de base 2 a menos que indicado outra base. Esta medida, entretanto, não é simétrica. Para obter uma medida simétrica, pode-se definir:

$$J(p, q) = KL(p, q) + KL(q, p) = \sum_{x \in X} (p(x) - q(x)) \log_2 \frac{p(x)}{q(x)},$$

conhecida como invariante de Jeffreys ou também chamada *J-divergência*.

2) **Distância variacional**, em que $f(u) = |u - 1|$, definida por

$$V(p, q) = \sum_{x \in X} |p(x) - q(x)|.$$

3) **Divergência- χ^2** , em que $f(u) = (u - 1)^2$, definida por

$$\chi^2(p, q) = \sum_{x \in X} \frac{(p(x) - q(x))^2}{q(x)}.$$

Esta medida, entretanto, não é simétrica, porém, de acordo com Cha (2007), várias versões simétricas de χ^2 têm sido exploradas. Uma forma simétrica é a medida de discriminação triangular dada por:

$$\Delta(p, q) = \sum_{x \in X} \frac{(p(x) - q(x))^2}{p(x) + q(x)}.$$

4) **Distância de Hellinger**, em que $f(u) = \frac{1}{2}(\sqrt{u} - 1)^2$, definida por

$$H^2(p, q) = \frac{1}{2} \sum_{x \in X} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2,$$

sendo que por vezes o fator $\frac{1}{2}$ não é utilizado na equação acima (CASTELLÓ et al., 2008).

As medidas Kullback-Leibler, J-divergência, divergência- χ^2 e discriminação triangular podem enfrentar problemas em suas expressões matemáticas, por convenção $0/0$ e $0 \log 0$ são tratados como 0. Para os casos $a/0$, $a > 0$, ou $\log 0$, o zero é substituído por um valor muito pequeno (CHA, 2007).

Seja, agora, X uma variável aleatória contínua e p e q duas funções densidades de probabilidade de X . Então a medida de divergência f de Csiszár entre essas distribuições de probabilidade é definida por

$$D_f(p, q) = \int_X p(x) f\left(\frac{q(x)}{p(x)}\right) dx.$$

Muitos métodos de agrupamento utilizam como ponto de partida uma matriz de proximidade que é uma matriz simétrica $n \times n$ contendo as medidas de dissimilaridades (distâncias) ou similaridades entre os n objetos (EVERITT et al., 2011).

Após a obtenção da matriz de dissimilaridades (similaridades), deve-se utilizar, então, algum método que auxilie na formação dos grupos similares. De um modo geral, na análise de agrupamento, os métodos podem ser classificados em dois grupos principais: métodos hierárquicos e métodos não-hierárquicos, que se aplicam em diversas áreas do conhecimento como Biologia, Medicina, Astronomia, Arqueologia, dentre outras. (EVERITT et al., 2011)

2.5.1 Métodos hierárquicos

Nos métodos hierárquicos os objetos (ou indivíduos) são classificados em grupos em diferentes etapas, de modo hierárquico, produzindo uma árvore de classificação (dendrograma) (BUSSAB; MIAZAKI; ANDRADE, 1990).

O dendrograma, conforme apresenta a Figura 2, é um gráfico bidimensional que ilustra todo o procedimento de agrupamento. Os “nós” do dendrograma representam

os agrupamentos e os comprimentos das hastes (alturas) representam as distâncias em que objetos são unidos (EVERITT et al., 2011).

Figura 2 – Exemplo de um dendrograma para a análise de agrupamento de 5 indivíduos

Segundo Frei (2006), este tipo de representação facilita a visualização da formação dos grupos e, conseqüentemente, auxilia na interpretação dos mesmos. Além disso, é a representação padrão nos softwares estatísticos.

Ao utilizar um método para determinar um ponto de corte na escala das distâncias do dendrograma, este corte determinará o número de grupos existentes nesse nível, os indivíduos que os formam e a qualidade do agrupamento (FERREIRA, 2011).

Os métodos de agrupamentos hierárquicos podem ser utilizados para agrupar objetos (ou indivíduos) como, também, para agrupar variáveis. Quando o dendrograma construído é das variáveis, a similaridade entre duas variáveis aponta forte correlação entre as mesmas. Os dendrogramas de objetos (ou indivíduos), entretanto, são mais comuns do que os de variáveis (MOITA NETO, 2004).

Os métodos hierárquicos subdividem-se em dois tipos: *métodos aglomerativos* e *métodos divisivos*. Os métodos aglomerativos realizam uma série sucessiva de uniões dos n objetos em grupos e os métodos divisivos consideram os n objetos como um único grupo e por sucessivas divisões formam grupos menores. De acordo com Bussab, Miazaki e Andrade (1990), os métodos aglomerativos são mais populares do que os divisivos.

2.5.1.1 Métodos aglomerativos

Um algoritmo geral para os agrupamentos hierárquicos aglomerativos com n objetos é dado, conforme Johnson e Wichern (2007), da seguinte maneira:

1. Inicia-se com n grupos, cada grupo contém um único elemento (objeto), e com a

matriz de distâncias (ou semelhanças) simétrica $\mathbf{D} = (d_{ij})$ de dimensão $n \times n$;

2. Identificar na matriz de distâncias os dois grupos que possuem a menor distância (mais similares). Seja a distância entre os objetos “mais semelhantes” U e V representado por d_{UV} .
3. Unir os grupos U e V e denominá-lo por (UV) . Atualizar as entradas na matriz de distâncias $\mathbf{D}$ por
 - 3.a) Eliminar as linhas e as colunas correspondentes aos grupos U e V ;
 - 3.b) Adicionar uma nova linha e uma nova coluna com as distâncias entre o grupo (UV) e os grupos restantes utilizando um dos métodos aglomerativos;
4. Repetir os passos 2 e 3 num total de $(n - 1)$ vezes até todos os objetos pertençam a um único grupo.

Conforme Ferreira (2011), alguns dos métodos aglomerativos mais utilizados são dados por:

- (i) Vizinho mais próximo (*Single linkage*). Neste caso, para o passo 3 do algoritmo geral, as distâncias entre o grupo (UV) e qualquer outro grupo W são calculadas por

$$d_{(UV)W} = \text{mín} \{d_{UW}, d_{VW}\},$$

em que as quantidades d_{UW} e d_{VW} são as distâncias entre os grupos U e W e entre os grupos V e W , respectivamente.

- (ii) Vizinho mais distante (*Complete linkage*). Neste caso, para o passo 3 do algoritmo geral, as distâncias entre o grupo (UV) e qualquer outro grupo W são calculadas por

$$d_{(UV)W} = \text{máx} \{d_{UW}, d_{VW}\},$$

em que as quantidades d_{UV} e d_{VW} foram definidos no item (i).

- (iii) Distância média (*Average linkage*). Neste caso, a distância entre dois grupos é dada pela média das distâncias entre todos os pares de objetos, no qual cada membro do

par pertence a um grupo distinto. Para o passo 3 do algoritmo geral, as distâncias entre (UV) e qualquer outro grupo W são determinadas por

$$d_{(UV)W} = \frac{\sum_{i \in (UV)} \sum_{j \in W} d_{ij}}{N_{(UV)} \times N_W}$$

em que d_{ij} é a distância entre o objeto i do grupo (UV) e o objeto j do grupo W e $N_{(UV)}$ e N_W são os números de objetos nos grupos (UV) e W , respectivamente.

- (iv) Método centróide (*Centroid method*). Neste caso, no passo 3, é necessário voltar aos dados originais para calcular a nova matriz de distâncias, o que exige um tempo computacional maior do que nos outros métodos (para maiores detalhes ver FREI, 2006). Ao contrário dos três métodos apresentados anteriormente, o método centróide não pode ser utilizado em situações nas quais se dispõe apenas da matriz de distâncias entre os n elementos amostrais (MINGOTI, 2005).
- (v) Método de Ward. Os métodos descritos anteriormente diferem somente em termos da relação de distância especificada na etapa 3. Por contraste, segundo Lattin, Carroll e Green (2011), o método Ward adota uma estratégia diferente na etapa 2. Em lugar de juntar os dois grupos mais similares, este método busca agrupar os dois grupos cuja união dá origem à menor soma de quadrados dentro do agrupamento, isto é, a variância mínima dentro do grupo (maiores detalhes em FERREIRA, 2011).

2.5.1.2 Métodos divisivos

Os métodos divisivos procedem de maneira inversa aos métodos aglomerativos. Os métodos iniciam com um único grupo contendo todos os n objetos e por divisões sucessivas vão sendo obtidos 2, 3, ... grupos até que, finalmente, todos os grupos contenham apenas um objeto. Muitos livros de análise de agrupamentos, no entanto, dão pouca atenção a este método que não tem sido considerado em grande parte dos “softwares” estatísticos (FREI, 2006).

2.5.2 Métodos não-hierárquicos

Segundo Johnson e Wichern (2007), métodos não-hierárquicos são desenvolvidos para agrupar objetos, ao invés de variáveis, em k grupos especificados antes do processo

de agrupamento. Uma maneira de alocar os n objetos em k grupos é a forma aleatória. O processo é efetuado diretamente na matriz de dados.

O método não-hierárquico mais conhecido recebe o nome de *método k -médias* ou *k -means*. O procedimento consiste em (FREI, 2006):

1. Separar os n objetos em k grupos, de forma aleatória;
2. Calcular os centróides (médias) de cada grupo;
3. Percorrer o conjunto de objetos, associando-os ao agrupamento cujo centróide está mais próximo (utilizam-se as distâncias com ou sem a padronização das variáveis); recalculando o centróide do agrupamento que recebeu o novo objeto e do agrupamento que perdeu o objeto;
4. Repetir o item 3 até que nenhuma reassociação tenha lugar.

Como uma matriz de distâncias (ou similaridades) não tem que ser determinada e os dados não precisam ser armazenados durante a execução no computador, os métodos não-hierárquicos podem ser aplicados a conjuntos de dados muito maiores do que nos métodos hierárquicos (JOHNSON; WICHERN, 2007).

2.5.3 Determinação do número de grupos

É frequente o caso em que o pesquisador não está interessado na hierarquia completa, mas apenas em uma ou duas partições obtidas a partir dela e isso envolve a decisão sobre o número de grupos (EVERITT et al., 2011).

Uma questão importante é de como se deve proceder para escolher o número final de grupos que define a partição do conjunto de dados analisado, ou de outra maneira, em qual passo o algoritmo de agrupamento deve ser interrompido. No entanto, não existe uma resposta exata para esta pergunta (MINGOTI, 2005).

Existem algumas estatísticas que se aplicam aos métodos de agrupamento hierárquicos que auxiliam na determinação do número de grupos. Essencialmente, procura-se por um grande “salto” entre os valores obtidos pela estatística utilizada. Entretanto, estas estatísticas são basicamente heurísticas, ou seja, suas distribuições são desconhecidas (ver SUBHASH, 1996 para mais detalhes).

Mojena (1977), entretanto, propôs um método baseado nas distâncias presentes no dendrograma para determinar o ponto de corte e definir o número de grupos a ser considerado. A proposta é selecionar o número de grupos no passo j que primeiro satisfazer a inequação

$$\alpha_j > \bar{\alpha} + k s_\alpha,$$

ou, de forma equivalente,

$$\frac{\alpha_j - \bar{\alpha}}{s_\alpha} > k$$

em que α_j é o valor da distância do nível de fusão correspondente ao passo j ($j = 1, 2, \dots, n$), $\bar{\alpha}$ e s_α são a média e o desvio padrão dos valores de α , respectivamente, e k é uma constante.

O valores que Mojena (1977) sugere para k estão no intervalo 2,75 e 3,5 para os melhores resultados, embora Milligan e Cooper (1985) sugiram utilizar $k = 1,25$ baseados em resultados de simulação.

2.5.4 Coeficiente de correlação cofenética como critério para avaliar métodos de agrupamento

Posteriormente a utilização dos métodos de agrupamentos hierárquicos, deve-se obter o coeficiente de correlação cofenética entre a matriz de distâncias D original e a matriz cofenética C (matriz de distâncias entre os objetos obtida a partir do dendrograma). Este coeficiente quantifica a proximidade entre as duas matrizes e é determinado por

$$r_{cof} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (c_{ij} - \bar{c}) (d_{ij} - \bar{d})}{\sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (c_{ij} - \bar{c})^2} \sqrt{\sum_{i=1}^{n-1} \sum_{j=i+1}^n (d_{ij} - \bar{d})^2}},$$

em que: c_{ij} é o valor de dissimilaridade entre os indivíduos i e j obtidos a partir da matriz cofenética; d_{ij} é o valor de dissimilaridade entre os indivíduos i e j obtidos a partir da matriz distâncias; n é a dimensão da matriz;

$$\bar{c} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=1}^n c_{ij} \quad \text{e} \quad \bar{d} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=1}^n d_{ij}.$$

Como interpretação do coeficiente de correlação cofenética, quanto mais próximo de um , melhor será a representação produzida pelo dendrograma e, assim, melhor

será a qualidade do agrupamento. Por outro lado, quanto mais próximo de zero, pior será a representação. Pode-se dizer em análise de agrupamentos que um valor em torno de 0,8 já pode ser considerado um bom ajuste (BUSSAB; MIAZAKI; ANDRADE, 1990).

3 MATERIAL E MÉTODOS

3.1 Descrição dos dados

Serão considerados, como ilustração da metodologia apresentada, os dados analisados por Molin, Menegatti e Gimenez (2001) relativos a um experimento de comparação de semeadoras conduzido no Departamento de Engenharia Rural da ESALQ/USP em 1998.

Neste experimento, quinze semeadoras manuais de diferentes fabricantes foram reguladas de modo a que caíssem exatamente duas sementes de cada vez (ou por golpe). Essas máquinas foram, então, utilizadas 150 vezes e ao final foram computados os números de sementes que caíram por golpe da semeadora e suas frequências, apresentadas na Tabela 3.

Tabela 3 – Frequências dos números de sementes por golpe de semeadoras manuais de quinze diferentes fabricantes, reguladas para caírem 2 sementes por golpe

Semeadora	Número de sementes por golpe						Total
	0	1	2	3	4	≥ 5	
1	0	19	103	18	3	7	150
2	2	26	70	48	4	0	150
3	14	21	82	27	5	1	150
4	7	19	82	42	0	0	150
5	33	32	34	42	9	0	150
6	42	32	43	22	9	2	150
7	20	26	68	31	5	0	150
8	13	28	47	50	8	4	150
9	35	26	30	30	21	8	150
10	6	41	77	18	8	0	150
11	34	41	40	27	4	4	150
12	6	38	70	35	1	0	150
13	33	28	29	38	18	4	150
14	14	28	62	41	4	1	150
15	15	59	73	3	0	0	150
Total	274	464	910	472	99	31	2250

3.2 Métodos

Considere os resultados de um experimento para a comparação de S semeadoras quanto à distribuição do número de sementes por golpe cujos resultados são agrupados em R categorias (Exemplo: 0, 1, 2, 3, 4, ≥ 5 , formando $R = 6$ categorias) e apresentados em uma tabela de contingência $S \times R$ como ilustra a Tabela 4.

Nessa tabela n_{ij} , $i = 1, 2, \dots, S$ e $j = 1, 2, \dots, R$, denota a frequência observada na casela correspondente, $n_{i+} = \sum_{j=1}^R n_{ij}$ é o total da i -ésima linha, $n_{\bullet j} = \sum_{i=1}^S n_{ij}$ é o total da j -ésima coluna e $n = \sum_{i=1}^S \sum_{j=1}^R n_{ij}$ é o total ou tamanho da amostra. Esses totais, ao serem considerados no delineamento amostral como fixos ou aleatórios, indicam o modelo probabilístico a ser considerado (GIOLO, 2004).

Tabela 4 – Tabela de contingência $S \times R$

Semeadora (i)	Categorias de resposta (j)					Total
	1	2	3	...	R	
1	n_{11}	n_{12}	n_{13}	...	n_{1R}	n_{1+}
2	n_{21}	n_{22}	n_{23}	...	n_{2R}	n_{2+}
3	n_{31}	n_{32}	n_{33}	...	n_{3R}	n_{3+}
...	...	...	...	...	...	...
S	n_{S1}	n_{S2}	n_{S3}	...	n_{SR}	n_{S+}
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	$n_{\bullet 3}$	...	$n_{\bullet R}$	n

O modelo probabilístico é usualmente assumido para explicar a ocorrência dos dados obtidos (de acordo com algum esquema amostral), na base do qual é traçada a inferência de interesse (PAULINO; SINGER, 2006).

Considere, que no caso em questão, n_{1+} , n_{2+} , ..., n_{S+} sejam fixos, isto é, estabelecidos antes da execução do experimento e seja a variável aleatória N_{ij} o número de golpes com a semeadora i cujo número caído de sementes pertence à categoria j . Nesse caso, o vetor aleatório $\mathbf{X}_i = (N_{i1}, N_{i2}, \dots, N_{iR})$ tem distribuição multinomial com n_{i+} fixo e parâmetros $\theta_{(i)j} \geq 0$, ou seja,

$$f_{\mathbf{X}_i}(\mathbf{x}_i | \boldsymbol{\theta}_i) = n_{i+}! \prod_{j=1}^R \frac{\theta_{(i)j}^{n_{ij}}}{n_{ij}!},$$

em que $\mathbf{x}_i = (n_{i1}, n_{i2}, \dots, n_{iR})$ e $\boldsymbol{\theta}_i = (\theta_{(i)1}, \theta_{(i)2}, \dots, \theta_{(i)R})$, sendo $\theta_{(i)j}$ a probabilidade de uma sementeira i depositar o número de sementes correspondente à categoria j . Neste caso em questão, o espaço multiparamétrico $\Theta_i = \left\{ \boldsymbol{\theta}_i : \theta_{(i)j} \geq 0, j = 1, 2, \dots, R \text{ e } \sum_{j=1}^R \theta_{(i)j} = 1 \right\}$ é o simplex $(R - 1)$ -dimensional, $\forall i \in \{1, 2, \dots, S\}$. Note que o último parâmetro pode ser obtido pela diferença dada por

$$\theta_{(i)R} = 1 - \sum_{j=1}^{R-1} \theta_{(i)j}.$$

Como os resultados das S sementeiras são independentes, o modelo probabilístico associado à Tabela 4, será o modelo produto de multinomiais independentes e descrito pela função de probabilidade

$$f_{\mathbf{X}^*}(\mathbf{x}^* | \boldsymbol{\theta}^*) = \prod_{i=1}^S \left(n_{i+}! \prod_{j=1}^R \frac{\theta_{(i)j}^{n_{ij}}}{n_{ij}!} \right),$$

em que $\mathbf{x}^* = (n_{11}, \dots, n_{1R}, \dots, n_{S1}, \dots, n_{SR})$ é o vetor de observações do vetor aleatório $\mathbf{X}^* = (N_{11}, \dots, N_{1R}, \dots, N_{S1}, \dots, N_{SR})$ e $\boldsymbol{\theta}^* = (\theta_{11}, \dots, \theta_{1R}, \dots, \theta_{S1}, \dots, \theta_{SR})$.

Consequentemente, a função de verossimilhança correspondente é dada por

$$L(\boldsymbol{\theta}^* | \mathbf{x}^*) = \prod_{i=1}^S \left(n_{i+}! \prod_{j=1}^R \frac{\theta_{(i)j}^{n_{ij}}}{n_{ij}!} \right). \quad (6)$$

Logo, a estimativa de máxima verossimilhança de $\boldsymbol{\theta}^*$ é aquela que maximiza a função de verossimilhança apresentada na eq. (6), o que é equivalente a maximizar o logaritmo da função de verossimilhança

$$l(\boldsymbol{\theta}^* | \mathbf{x}^*) = \sum_{i=1}^S \log n_{i+}! - \left(\sum_{i=1}^S \sum_{j=1}^R \log n_{ij}! \right) + \left(\sum_{i=1}^S \sum_{j=1}^R n_{ij} \log \theta_{(i)j} \right).$$

Segundo Bishop, Fienberg e Holland (2007), pode-se introduzir o multiplicador de Lagrange λ para maximizar a função $l(\boldsymbol{\theta}_i | \mathbf{x}_i)$ sujeita à restrição $\sum_{j=1}^R \theta_{(i)j} - 1 = 0$ com respeito a $\boldsymbol{\theta}_i$ e λ e, assim, encontrar os pontos críticos da função. Logo, para $i, i = 1, 2, \dots, S$, tem-se

$$F(\boldsymbol{\theta}_i, \lambda) = \log n_{i+}! - \left(\sum_{j=1}^R \log n_{ij}! \right) + \sum_{j=1}^R n_{ij} \log \theta_{(i)j} + \lambda \left(\sum_{j=1}^R \theta_{(i)j} - 1 \right). \quad (7)$$

Tomando-se as derivadas parciais em relação a $\theta_{(i)j}$ e igualando-as a zero, obtém-se

$$\frac{\partial F(\boldsymbol{\theta}_i, \lambda)}{\partial \theta_{(i)j}} = \frac{n_{ij}}{\theta_{(i)j}} + \lambda = 0, \quad j = 1, 2, \dots, R$$

e conseqüentemente, a solução

$$\hat{\theta}_{(i)j} = -\frac{n_{ij}}{\lambda}. \quad (8)$$

Tomando-se a derivada parcial da eq. (7) em relação a λ e igualando-se a zero, tem-se

$$\frac{\partial F(\boldsymbol{\theta}_i, \lambda)}{\partial \lambda} = \sum_{j=1}^R \theta_{(i)j} - 1 = 0 \quad (9)$$

Substituindo $\theta_{(i)j}$ por $\hat{\theta}_{(i)j}$ dado pela eq. (8) na eq. (9), tem-se

$$\begin{aligned} \sum_{j=1}^R \left(-\frac{n_{ij}}{\lambda} \right) - 1 &= 0 \\ -\frac{1}{\lambda} \times \left(\sum_{j=1}^R n_{ij} \right) &= 1 \\ -\frac{n_{i+}}{\lambda} &= 1 \end{aligned}$$

e, assim, obtém-se a seguinte solução

$$\hat{\lambda} = -n_{i+}. \quad (10)$$

Portanto, substituindo-se $\hat{\lambda}$ dado pela eq. (10) na eq. (8) encontra-se a estimativa de máxima verossimilhança de $\theta_{(i)j}$ referente a cada sementeira i , $i = 1, 2, \dots, S$, pertencente à categoria j , $j = 1, 2, \dots, R$, dada por

$$\hat{\theta}_{(i)j} = \frac{n_{ij}}{n_{i+}}.$$

Finalmente, tomando-se a segunda derivada parcial de eq. (7) e substituindo

os valores, $\hat{\theta}_{(i)j}$, comprova-se que estes são pontos de máximo, pois

$$\frac{\partial^2 F(\boldsymbol{\theta}_i, \lambda)}{\partial \theta_{(i)j}^2} \Big|_{\theta_{(i)j} = \hat{\theta}_{(i)j}} = -\frac{n_{ij}}{\theta_{(i)j}^2} < 0, \quad j = 1, 2, \dots, R$$

em que $n_{ij} \geq 0$, $\forall i = 1, 2, \dots, S$ e $\forall j = 1, 2, \dots, R$.

Considere que haja o interesse inicial de verificar se as distribuições dos números de sementes por golpe sejam iguais para as diferentes semeadoras. Pode-se, assim, estabelecer as seguintes hipóteses:

$$\left\{ \begin{array}{l} H_0 : \quad \theta_{(1)1} = \theta_{(2)1} = \dots = \theta_{(S)1} = \theta_{(\bullet)1}; \\ \quad \quad \theta_{(1)2} = \theta_{(2)2} = \dots = \theta_{(S)2} = \theta_{(\bullet)2}; \\ \quad \quad \quad \dots \\ \quad \quad \theta_{(1)R} = \theta_{(2)R} = \dots = \theta_{(S)R} = \theta_{(\bullet)R}; \\ H_1 : \quad \text{Pelo menos 2 semeadoras diferem entre si.} \end{array} \right.$$

No caso em questão, o modelo reduzido, com menor número de parâmetros, é representado pela função de verossimilhança $L_1(\theta_{(\bullet)1}, \dots, \theta_{(\bullet)R} | n_{\bullet 1}, \dots, n_{\bullet R})$ e o modelo completo, pela função de verossimilhança $L_2(\boldsymbol{\theta}^* | \mathbf{x}^*)$. Assim, a estatística do teste da razão de verossimilhanças é dada por

$$TRV = 2 \log \left(\frac{L_2(\boldsymbol{\theta}^* | \mathbf{x}^*)}{L_1(\theta_{(\bullet)1}, \dots, \theta_{(\bullet)R} | n_{\bullet 1}, \dots, n_{\bullet R})} \right) = 2 [\log(L_2) - \log(L_1)].$$

Essa estatística da razão de verossimilhanças, sob a hipótese nula, tem distribuição assintótica χ^2 , com número de graus de liberdade obtido pela diferença entre o número de parâmetros do modelo completo e o número de parâmetros do modelo reduzido.

Considere, agora, que haja o interesse de verificar se as distribuições dos números de sementes por golpe de duas semeadoras específicas sejam iguais. Considere, assim, a Tabela 5 correspondente aos dados observados respectivamente para as semeadoras i e i^* , $\forall i \neq i^*$, $i = 1, 2, \dots, S$.

Tabela 5 – Frequências absolutas para as semeadoras i e i^*

Semeadora	Categorias de resposta (j)					Total
	1	2	3	...	R	
i	n_{i1}	n_{i2}	n_{i3}	...	n_{iR}	n_{i+}
i^*	n_{i^*1}	n_{i^*2}	n_{i^*3}	...	n_{i^*R}	n_{i^*+}
Total	n_{+1}	n_{+2}	n_{+3}	...	n_{+R}	n

Neste caso, as hipóteses de interesse são

$$\left\{ \begin{array}{l} H_0 : \quad \theta_{(i)1} = \theta_{(i^*)1} = \theta_{(+)1}; \\ \quad \quad \theta_{(i)2} = \theta_{(i^*)2} = \theta_{(+)2}; \\ \quad \quad \quad \dots \\ \quad \quad \theta_{(i)R} = \theta_{(i^*)R} = \theta_{(+)R}. \\ H_1 : \quad \text{Semeadoras diferem entre si.} \end{array} \right. \quad (11)$$

A função de verossimilhança para o modelo completo é, portanto,

$$L_2(\theta_{(i)1}, \dots, \theta_{(i^*)R} | n_{i1}, \dots, n_{i^*R}) = \frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \prod_{j=1}^R \theta_{(i)j}^{n_{ij}} \times \frac{n_{i^*+}!}{\prod_{j=1}^R n_{i^*j}!} \prod_{j=1}^R \theta_{(i^*)j}^{n_{i^*j}}$$

e, para o modelo reduzido (sob a hipótese H_0) é

$$L_1(\theta_{(+)1}, \dots, \theta_{(+)R} | n_{+1}, \dots, n_{+R}) = \frac{n_{+}!}{\prod_{j=1}^R n_{+j}!} \prod_{j=1}^R \theta_{(+)j}^{n_{+j}} \times \frac{n_{+}!}{\prod_{j=1}^R n_{+j}!} \prod_{j=1}^R \theta_{(+)j}^{n_{+j}}.$$

Logo, pode-se escrever a razão das verossimilhanças como se segue

$$\begin{aligned} \frac{L_2(\theta_{(i)1}, \dots, \theta_{(i^*)R} | n_{i1}, \dots, n_{i^*R})}{L_1(\theta_{(+)1}, \dots, \theta_{(+)R} | n_{+1}, \dots, n_{+R})} &= \frac{\frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \prod_{j=1}^R \theta_{(i)j}^{n_{ij}} \times \frac{n_{i^*+}!}{\prod_{j=1}^R n_{i^*j}!} \prod_{j=1}^R \theta_{(i^*)j}^{n_{i^*j}}}{\frac{n_{+}!}{\prod_{j=1}^R n_{+j}!} \prod_{j=1}^R \theta_{(+)j}^{n_{+j}} \times \frac{n_{+}!}{\prod_{j=1}^R n_{+j}!} \prod_{j=1}^R \theta_{(+)j}^{n_{+j}}} \\ &= \frac{\prod_{j=1}^R \theta_{(i)j}^{n_{ij}} \times \prod_{j=1}^R \theta_{(i^*)j}^{n_{i^*j}}}{\prod_{j=1}^R \theta_{(+)j}^{n_{+j}} \times \prod_{j=1}^R \theta_{(+)j}^{n_{+j}}}, \end{aligned} \quad (12)$$

e, substituindo as estimativas obtidas utilizando-se o método da máxima verossimilhança dadas por $\hat{\theta}_{(i)j} = \frac{n_{ij}}{n_{i+}}$, $\hat{\theta}_{(i^*)j} = \frac{n_{i^*j}}{n_{i^*+}}$ e $\hat{\theta}_{(+)j} = \frac{n_{+j}}{n}$ na eq. (12), obtém-se

$$\begin{aligned} \frac{L_2(\theta_{(i)1}, \dots, \theta_{(i^*)R} | n_{i1}, \dots, n_{i^*R})}{L_1(\theta_{(+)1}, \dots, \theta_{(+)R} | n_{+1}, \dots, n_{+R})} &= \frac{\prod_{j=1}^R \binom{n_{ij}}{n_{i+}}^{n_{ij}} \times \prod_{j=1}^R \binom{n_{i^*j}}{n_{i^*+}}^{n_{i^*j}}}{\prod_{j=1}^R \binom{n_{+j}}{n}^{n_{+j}} \times \prod_{j=1}^R \binom{n_{+j}}{n}^{n_{+j}}} \\ &= \frac{\frac{1}{\prod_{j=1}^R (n_{i+})^{n_{ij}} \times \prod_{j=1}^R (n_{i^*+})^{n_{i^*j}}} \times \prod_{j=1}^R (n_{ij})^{n_{ij}} \times \prod_{j=1}^R (n_{i^*j})^{n_{i^*j}}}{\frac{1}{\prod_{j=1}^R (n)^{n_{+j}} \times \prod_{j=1}^R (n)^{n_{+j}}} \times \prod_{j=1}^R (n_{+j})^{n_{+j}} \times \prod_{j=1}^R (n_{+j})^{n_{+j}}}. \end{aligned}$$

Além disso, como os totais das linhas são $n_{i+} = \sum_{j=1}^R n_{ij}$ e $n_{i^*+} = \sum_{j=1}^R n_{i^*j}$,

tem-se que

$$\begin{aligned} \frac{L_2(\theta_{(i)1}, \dots, \theta_{(i^*)R} | n_{i1}, \dots, n_{i^*R})}{L_1(\theta_{(+)1}, \dots, \theta_{(+)R} | n_{+1}, \dots, n_{+R})} &= \frac{\frac{1}{(n_{i+})^{\sum_{j=1}^R n_{ij}} \times (n_{i^*+})^{\sum_{j=1}^R n_{i^*j}}} \times \prod_{j=1}^R (n_{ij})^{n_{ij}} \times \prod_{j=1}^R (n_{i^*j})^{n_{i^*j}}}{\frac{1}{(n)^{\sum_{j=1}^R n_{+j}} \times (n)^{\sum_{j=1}^R n_{+j}}} \times \prod_{j=1}^R (n_{+j})^{n_{+j}} \times \prod_{j=1}^R (n_{+j})^{n_{+j}}} \\ &= \frac{\frac{1}{(n_{i+})^{n_{i+}} \times (n_{i^*+})^{n_{i^*+}}} \times \prod_{j=1}^R (n_{ij})^{n_{ij}} \times \prod_{j=1}^R (n_{i^*j})^{n_{i^*j}}}{\frac{1}{(n)^{n_{+}} \times (n)^{n_{+}}} \times \prod_{j=1}^R (n_{+j})^{n_{+j}} \times \prod_{j=1}^R (n_{+j})^{n_{+j}}}. \end{aligned}$$

Considerando, ainda, que o tamanho da amostra é $n = n_{i+} + n_{i^*+}$ e, particularmente, o caso em que os totais marginais das linhas são fixos e iguais, ou seja, $n_{i+} = n_{i^*+}$, obtém-se

$$\begin{aligned} \frac{L_2(\theta_{(i)1}, \dots, \theta_{(i^*)R} | n_{i1}, \dots, n_{i^*R})}{L_1(\theta_{(+)1}, \dots, \theta_{(+)R} | n_{+1}, \dots, n_{+R})} &= \frac{\frac{1}{(n_{i+})^{n_{i+} + n_{i^*+}}} \times \prod_{j=1}^R (n_{ij})^{n_{ij}} \times \prod_{j=1}^R (n_{i^*j})^{n_{i^*j}}}{\frac{1}{(n)^{n_{i+} + n_{i^*+}}} \times \prod_{j=1}^R n_{+j}^{n_{ij} + n_{i^*j}}} \\ &= \frac{\frac{1}{(n_{i+})^n} \times \prod_{j=1}^R (n_{ij})^{n_{ij}} \times \prod_{j=1}^R (n_{i^*j})^{n_{i^*j}}}{\frac{1}{(n)^n} \times \prod_{j=1}^R n_{+j}^{n_{ij} + n_{i^*j}}}. \end{aligned}$$

Portanto, como $n_{+j} = n_{ij} + n_{i^*j}$, $\forall j = 1, 2, \dots, R$, a razão das verossimilhan-

ças é dada por

$$\frac{L_2(\theta_{(i)1}, \dots, \theta_{(i^*)R} | n_{i1}, \dots, n_{i^*R})}{L_1(\theta_{(+)1}, \dots, \theta_{(+)R} | n_{+1}, \dots, n_{+R})} = \left(\frac{n}{n_{i+}}\right)^n \times \frac{\prod_{j=1}^R (n_{ij})^{n_{ij}} \times \prod_{j=1}^R (n_{i^*j})^{n_{i^*j}}}{\prod_{j=1}^R n_{+j}^{n_{+j}}}. \quad (13)$$

Com base no resultado obtido em (13), o valor da estatística do teste da razão de verossimilhanças para uma amostra pode ser expresso por

$$\begin{aligned} TRV &= 2 \log \left(\frac{L_2(\theta_{(i)1}, \dots, \theta_{(i^*)R} | n_{i1}, \dots, n_{i^*R})}{L_1(\theta_{(+)1}, \dots, \theta_{(+)R} | n_{+1}, \dots, n_{+R})} \right) \\ &= 2 \log \left[\left(\frac{n}{n_{i+}}\right)^n \times \frac{\prod_{j=1}^R (n_{ij})^{n_{ij}} \times \prod_{j=1}^R (n_{i^*j})^{n_{i^*j}}}{\prod_{j=1}^R (n_{+j})^{n_{+j}}} \right], \end{aligned}$$

ou seja,

$$TRV = 2 \left[n \log n - n \log n_{i+} + \sum_{j=1}^R n_{ij} \log n_{ij} + \sum_{j=1}^R n_{i^*j} \log n_{i^*j} - \left(\sum_{j=1}^R n_{+j} \log n_{+j} \right) \right].$$

O número de parâmetros do modelo completo é $k_2 = (R-1) + (R-1) = 2 \times (R-1)$ e o número de parâmetros do modelo reduzido é $k_1 = (R-1)$, logo a estatística do teste segue uma distribuição χ^2 com $k_2 - k_1 = (R-1)$ graus de liberdade.

Na situação apresentada, há o interesse em comparar todas as sementeiras duas a duas. Consequentemente, o nível descritivo ou valor-p associado ao teste deve ser ajustado. Será utilizado, deste modo, o método *single-step* de Sidák em que o valor-p ajustado é dado por

$$\tilde{p}_k = 1 - (1 - p_k)^m, \quad k = 1, 2, \dots, m,$$

sendo

$$m = \binom{S}{2} = \frac{S!}{2!(S-2)!}$$

o número total de comparações das S sementeiras tomadas duas a duas. Assim, dado um nível de significância α , rejeita-se a hipótese nula H_0 para cada comparação se $\tilde{p}_k \leq \alpha$.

Uma maneira alternativa de se realizar a inferência do vetor de parâmetros

desconhecidos $\boldsymbol{\theta}_i = (\theta_{(i)1}, \theta_{(i)2}, \dots, \theta_{(i)R})$, para cada semeadora i , é por meio da inferência bayesiana.

Suponha que a distribuição a priori de $\boldsymbol{\theta}_i$ seja uma distribuição Dirichlet com parâmetros $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iR})$, $\alpha_{ij} > 0$, expressa por

$$\pi(\boldsymbol{\theta}_i) = \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij}-1}, \quad (14)$$

em que $B(\boldsymbol{\alpha}_i) = \frac{\prod_{j=1}^R \Gamma(\alpha_{ij})}{\Gamma\left(\sum_{j=1}^R \alpha_{ij}\right)}$ é a função beta com parâmetros $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iR})$.

A função de verossimilhança para cada semeadora i , assim, dada por

$$L(\boldsymbol{\theta}_i | \mathbf{x}_i) = \frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \prod_{j=1}^R \theta_{(i)j}^{n_{ij}}, \quad (15)$$

em que $\mathbf{x}_i = (n_{i1}, n_{i2}, \dots, n_{iR})$. Logo, por meio do teorema de Bayes, a eq. (14) é combinada com a eq. (15), produzindo a distribuição a posteriori de $\boldsymbol{\theta}_i$ dada por

$$\begin{aligned} \pi(\boldsymbol{\theta}_i | \mathbf{x}_i) &= \frac{\frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \prod_{j=1}^R \theta_{(i)j}^{n_{ij}} \times \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij}-1}}{\int_{\Theta_i} \frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \prod_{j=1}^R \theta_{(i)j}^{n_{ij}} \times \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij}-1} d\boldsymbol{\theta}_i} \\ &= \frac{\frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \prod_{j=1}^R \theta_{(i)j}^{n_{ij}} \times \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij}-1}}{\frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \times \frac{1}{B(\boldsymbol{\alpha}_i)} \int_{\Theta_i} \prod_{j=1}^R \theta_{(i)j}^{n_{ij}} \times \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij}-1} d\boldsymbol{\theta}_i} \\ &= \frac{\prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij}+n_{ij}-1}}{\int_{\Theta_i} \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij}+n_{ij}-1} d\boldsymbol{\theta}_i}. \end{aligned}$$

No entanto, como

$$\int_{\Theta_i} \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij}+n_{ij}-1} d\boldsymbol{\theta}_i = B(\boldsymbol{\alpha}_i + \mathbf{x}_i),$$

em que $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}_i + \mathbf{x}_i = (\alpha_{i1} + n_{i1}, \alpha_{i2} + n_{i2}, \dots, \alpha_{iR} + n_{iR})$ e $B(\boldsymbol{\alpha}_i + \mathbf{x}_i) = \frac{\prod_{j=1}^R \Gamma(\alpha_{ij} + n_{ij})}{\Gamma\left(n_{i+} + \sum_{j=1}^R \alpha_{ij}\right)}$,

tem-se que

$$\pi(\boldsymbol{\theta}_i | \mathbf{x}_i) = \frac{1}{B(\boldsymbol{\alpha}_i + \mathbf{x}_i)} \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij} + n_{ij} - 1},$$

ou seja, a distribuição a posteriori de $\boldsymbol{\theta}_i$ corresponde a uma distribuição Dirichlet cujos os parâmetros são dados por $\boldsymbol{\alpha}^* = (\alpha_{i1} + n_{i1}, \alpha_{i2} + n_{i2}, \dots, \alpha_{iR} + n_{iR})$.

Diante disso, percebe-se que a distribuição a posteriori pertence a mesma família de distribuições que a da priori. Portanto, tem-se que a família Dirichlet é a conjugada natural da distribuição multinomial. Nesse caso, a verossimilhança marginal do conjunto de dados, $\mathbf{x}_i = (n_{i1}, n_{i2}, \dots, n_{iR})$, expressa por

$$\begin{aligned} \pi(\mathbf{x}_i) &= \int_{\Theta_i} \frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \prod_{j=1}^R \theta_{(i)j}^{n_{ij}} \times \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij} - 1} d\boldsymbol{\theta}_i \\ &= \frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \times \frac{1}{B(\boldsymbol{\alpha}_i)} \int_{\Theta_i} \prod_{j=1}^R \theta_{(i)j}^{\alpha_{ij} + n_{ij} - 1} d\boldsymbol{\theta}_i \\ &= \frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \times \frac{1}{B(\boldsymbol{\alpha}_i)} \times B(\boldsymbol{\alpha}_i + \mathbf{x}_i), \end{aligned}$$

é denominada como distribuição Multinomial-Dirichlet.

Como $\boldsymbol{\theta}_i = (\theta_{(i)1}, \theta_{(i)2}, \dots, \theta_{(i)R})$ segue uma distribuição Dirichlet com parâmetros $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iR})$ denotada, aqui, por $\boldsymbol{\theta}_i \sim \text{Dir}(\boldsymbol{\alpha}_i)$, tem-se a média a priori dada por $E(\theta_{(i)j}) = \frac{\alpha_{ij}}{\sum_{j=1}^R \alpha_{ij}}$. A distribuição a posteriori de $\boldsymbol{\theta}_i$ corresponde a uma distribuição Dirichlet com parâmetros $\boldsymbol{\alpha}^* = (\alpha_{i1} + n_{i1}, \alpha_{i2} + n_{i2}, \dots, \alpha_{iR} + n_{iR})$, ou seja, $\boldsymbol{\theta}_i | \mathbf{x}_i \sim \text{Dir}(\boldsymbol{\alpha}^*)$, logo tem-se que a média a posteriori dada por

$$E(\theta_{(i)j} | \mathbf{x}_i) = \frac{\alpha_{ij} + n_{ij}}{\sum_{j=1}^R \alpha_{ij} + \sum_{j=1}^R n_{ij}} = \frac{\alpha_{ij} + n_{ij}}{\sum_{j=1}^R \alpha_{ij} + n_{i+}},$$

é uma estimativa de Bayes para θ .

Considere as hipóteses de interesse dadas por (11). Para realizar a comparação dos dois modelos sob cada hipótese na metodologia Bayesiana será utilizado o fator de Bayes.

A verossimilhança marginal dos dados, $\mathbf{x}' = (n_{i1}, \dots, n_{iR}, n_{i*1}, \dots, n_{i*R})$, dada a hipótese H_1 (referente ao modelo completo) é dada por

$$\pi(\mathbf{x}'|H_1) = \frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \times \frac{1}{B(\boldsymbol{\alpha}_i)} \times B(\boldsymbol{\alpha}_i + \mathbf{x}_i) \times \frac{n_{i*+}!}{\prod_{j=1}^R n_{i*j}!} \times \frac{1}{B(\boldsymbol{\alpha}_{i*})} \times B(\boldsymbol{\alpha}_{i*} + \mathbf{x}_{i*}),$$

em que $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \dots, \alpha_{iR})$, $\mathbf{x}_i = (n_{i1}, n_{i2}, \dots, n_{iR})$, $\boldsymbol{\alpha}_{i*} = (\alpha_{i*1}, \alpha_{i*2}, \dots, \alpha_{i*R})$ e $\mathbf{x}_{i*} = (n_{i*1}, n_{i*2}, \dots, n_{i*R})$. Por outro lado, tem-se a verossimilhança marginal dos dados, $\mathbf{x}' = (n_{i1}, \dots, n_{iR}, n_{i*1}, \dots, n_{i*R})$, dada a hipótese H_0 (referente ao modelo reduzido) por

$$\pi(\mathbf{x}'|H_0) = \frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \times \frac{n_{i*+}!}{\prod_{j=1}^R n_{i*j}!} \times \frac{1}{B(\boldsymbol{\alpha}_+)} \times B(\boldsymbol{\alpha}_+ + \mathbf{x}_+),$$

em que $\boldsymbol{\alpha}_+ = (\alpha_{+1}, \alpha_{+2}, \dots, \alpha_{+R})$ e $\mathbf{x}_+ = (n_{+1}, n_{+2}, \dots, n_{+R})$.

Logo, o fator de Bayes pode ser expresso na forma

$$\begin{aligned} B_{10} &= \frac{\frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \times \frac{1}{B(\boldsymbol{\alpha}_i)} \times B(\boldsymbol{\alpha}_i + \mathbf{x}_i) \times \frac{n_{i*+}!}{\prod_{j=1}^R n_{i*j}!} \times \frac{1}{B(\boldsymbol{\alpha}_{i*})} \times B(\boldsymbol{\alpha}_{i*} + \mathbf{x}_{i*})}{\frac{n_{i+}!}{\prod_{j=1}^R n_{ij}!} \times \frac{n_{i*+}!}{\prod_{j=1}^R n_{i*j}!} \times \frac{1}{B(\boldsymbol{\alpha}_+)} \times B(\boldsymbol{\alpha}_+ + \mathbf{x}_+)} \\ &= \frac{B(\boldsymbol{\alpha}_+)}{B(\boldsymbol{\alpha}_+ + \mathbf{x}_+)} \times \frac{B(\boldsymbol{\alpha}_i + \mathbf{x}_i)}{B(\boldsymbol{\alpha}_i)} \times \frac{B(\boldsymbol{\alpha}_{i*} + \mathbf{x}_{i*})}{B(\boldsymbol{\alpha}_{i*})}. \end{aligned}$$

Aplicando o logaritmo natural em ambos os lados do fator de Bayes, obtém-se

$$\begin{aligned} \log B_{10} &= \log \left[\frac{B(\boldsymbol{\alpha}_+)}{B(\boldsymbol{\alpha}_+ + \mathbf{x}_+)} \times \frac{B(\boldsymbol{\alpha}_i + \mathbf{x}_i)}{B(\boldsymbol{\alpha}_i)} \times \frac{B(\boldsymbol{\alpha}_{i*} + \mathbf{x}_{i*})}{B(\boldsymbol{\alpha}_{i*})} \right] \\ &= \log \frac{B(\boldsymbol{\alpha}_+)}{B(\boldsymbol{\alpha}_+ + \mathbf{x}_+)} + \log \frac{B(\boldsymbol{\alpha}_i + \mathbf{x}_i)}{B(\boldsymbol{\alpha}_i)} + \log \frac{B(\boldsymbol{\alpha}_{i*} + \mathbf{x}_{i*})}{B(\boldsymbol{\alpha}_{i*})} \\ &= \log B(\boldsymbol{\alpha}_+) - \log B(\boldsymbol{\alpha}_+ + \mathbf{x}_+) + \log B(\boldsymbol{\alpha}_i + \mathbf{x}_i) - \log B(\boldsymbol{\alpha}_i) + \log B(\boldsymbol{\alpha}_{i*} + \mathbf{x}_{i*}) - \\ &\quad \log B(\boldsymbol{\alpha}_{i*}). \end{aligned}$$

Portanto,

$$\log B_{10} = \log B(\boldsymbol{\alpha}_+) + \log B(\boldsymbol{\alpha}_i + \mathbf{x}_i) + \log B(\boldsymbol{\alpha}_{i*} + \mathbf{x}_{i*}) - [\log B(\boldsymbol{\alpha}_+ + \mathbf{x}_+) + \log B(\boldsymbol{\alpha}_i) + \log B(\boldsymbol{\alpha}_{i*})].$$

Para se obter a rejeição ou não-rejeição da hipótese nula, H_0 , deve-se multiplicar por dois o logaritmo natural do fator de Bayes, ou seja, $2 \log B_{10}$ e utilizar como referência a Tabela 2, que apresenta uma interpretação do fator de Bayes de acordo com Kass e Raftery (1995). Segundo Bender e Lange (2001), os métodos de Bayes não necessitam de ajustes dos valores-p. Além disso, a perspectiva bayesiana é similar a da frequentista que faz inferências em uma base por comparação (BERRY; HOCHBERG, 1999).

Finalmente, será feita uma análise de agrupamentos das S sementeiras com relação a distribuição do número de sementes. Esta tem por finalidade a formação de grupos similares por meio dos métodos hierárquicos aglomerativos, tomando como base a matriz de distâncias entre as sementeiras, de tal forma que as sementeiras dentro dos grupos sejam semelhantes, mas os grupos formados sejam distintos entre si.

Sabendo-se que cada sementeira i é representada por um vetor aleatório com distribuição multinomial, pode-se utilizar medidas de divergência como ferramenta para avaliar se duas sementeiras quaisquer estão próximas e se podem fazer parte ou não de um mesmo grupo.

Embora as medidas de divergência não obedeçam a necessariamente todas as propriedades de uma métrica, estas são interpretadas como distâncias entre duas distribuições de probabilidade.

Existem diferentes medidas de divergências para distribuições de probabilidade e cada uma delas pode produzir um determinado tipo de agrupamento. Para que se possa prosseguir com o agrupamento das sementeiras será utilizada a *J-divergência*, dada por

$$J_{\phi}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i^*}) = \sum_{j=1}^R (\theta_{(i)j} - \theta_{(i^*)j}) \log \left(\frac{\theta_{(i)j}}{\theta_{(i^*)j}} \right),$$

em que $\boldsymbol{\theta}_i = (\theta_{(i)1}, \theta_{(i)2}, \dots, \theta_{(i)R})$ e $\boldsymbol{\theta}_{i^*} = (\theta_{(i^*)1}, \theta_{(i^*)2}, \dots, \theta_{(i^*)R})$.

Uma vez escolhida a divergência, serão calculadas as distâncias entre as sementeiras i e i^* , $\forall i, i^* \in \{1, 2, \dots, S\}$, e armazenadas em uma matriz de dissimilaridades $\mathbf{D}_{n \times n}$, que é simétrica contendo zeros na diagonal principal. A escolha da medida *J-divergência* se deve ao fato de ser a medida simétrica da divergência de Kullback-Leibler que, segundo Le e Ho (2005), é a mais popular das várias medidas de dissimilaridade entre distribuições de probabilidade.

Posteriormente, serão utilizados os métodos de distância média, vizinho mais

distante e vizinho mais próximo para a obtenção dos dendrogramas, uma vez que, não necessitam voltar aos valores originais dos objetos para atualizar a matriz de distâncias. Além disso, segundo Milligan, 1980 apud Lattin, Carroll e Green (2011), o método centróide é robusto para as discrepâncias, mas pode ser superado pelo método da distância média. Por fim, será determinado o número de grupos que define a partição do conjunto de dados com a constante $k = 1,25$, conforme sugerido por Milligan e Cooper (1985) e obtido o coeficiente de correlação cofenética para avaliar os métodos utilizados.

4 RESULTADOS E DISCUSSÃO

A Figura 3 apresenta as distribuições de frequências dos números de sementes por golpe para as semeadoras 1 a 15 apresentadas na Tabela 3.

Figura 3 – Distribuições de frequências dos números de sementes depositas por acionamento (golpe) para 15 semeadoras diferentes, de um total de 150 acionamentos por semeadora

Como primeiro critério para a comparação do desempenho quanto à distribuição do número de sementes das semeadoras manuais poderia se pensar em fazê-lo por meio das frequências ocorridas pelo depósito de exatamente duas sementes por golpe. Nesse caso, as semeadoras 3 e 4 seriam consideradas semelhantes mesmo apresentando distribuições diferentes. Chega-se a mesma conclusão para as semeadoras 2 e 12. Entretanto, em 14 vezes a semeadora 3 não depositou nenhuma semente, o dobro do observado para a semeadora 4, evidenciando o pior desempenho da semeadora 3.

Na Figura 3, é possível observar a variação na distribuição de sementes por meio do mecanismo dosador de cada semeadora. Com base nas frequências obtidas de cada semeadora, Tabela 3, serão feitas análises estatísticas conforme apresentadas a seguir, as rotinas computacionais utilizadas para obtenção dos resultados nas análises se encontram no Apêndice.

4.1 Comparações das semeadoras duas a duas por meio dos métodos descritivos

A Tabela 6 apresenta medidas de posição e dispersão dos números de sementes por golpe das quinze semeadoras consideradas.

Tabela 6 – Médias, medianas, modas, desvios padrões e amplitudes interquartílicas (AIQ) dos números de sementes por golpe para as 15 semeadoras. (Considerou-se, aqui, que os valores de números de sementes por golpe maiores ou iguais a 5 são exatamente iguais a 5)

Semeadora	Média	Mediana	Moda	Desvio padrão	AIQ
1	2,1733	2	2	0,8494	0
2	2,1733	2	2	0,7921	1
3	1,9400	2	2	0,9429	0
4	2,0600	2	2	0,7705	1
5	1,7467	2	3	1,2488	2
6	1,5333	2	2	1,2779	2
7	1,8333	2	2	1,0128	1
8	2,1600	2	3	1,1415	2
9	2,0000	2	0	1,5370	2
10	1,8733	2	2	0,8693	1
11	1,5867	1,5	1	1,2434	1
12	1,9133	2	2	0,8187	1
13	1,9467	2	3	1,4321	2
14	1,9733	2	2	1,0030	2
15	1,4267	2	2	0,6985	1

Analisando-se as estatísticas apresentadas na Tabela 6 tem-se que as sementeiras 1 e 2 apresentam as mesmas médias. Porém, observa-se na Figura 3 que ambas apresentam distribuições completamente diferentes. Analogamente, tem-se que as sementeiras 4 e 9 apresentam, aproximadamente, médias similares, porém suas distribuições são completamente distintas conforme Figura 3. O uso da mediana ou moda é ainda pior devido ao fato de pouco variarem, levando a um menor poder discriminativo.

Como medida de dispersão frequentemente utilizada, o desvio padrão, também não foi capaz de discriminar diferenças entre as sementeiras 1 e 10, assim como, 1 e 12, que apresentam aproximadamente os mesmos desvios padrões.

A amplitude interquartil também não identificou diferenças entre, por exemplo, as sementeiras 8 e 9 e as sementeiras 13 e 14, conforme mostra a Tabela 6. Além disso, as amplitudes interquartílicas para as sementeiras 1 e 3 são nulas, indicando não haver variabilidade entre os números de sementes depositados por golpe, entretanto, apresentam irregularidades nas distribuições das sementes.

Ao considerar esta análise, o pesquisador poderia agrupar, por exemplo, as sementeiras 1 e 2 por considerá-las semelhantes e recomendar a utilização de ambas. Portanto, as medidas de posição e dispersão utilizadas para avaliar e comparar o desempenho das máquinas duas a duas não mostraram-se suficientes para detectar diferenças entre as sementeiras quanto ao número de sementes depositadas por golpe.

4.2 Inferência Clássica

As estimativas de máxima verossimilhança obtidas para as categorias de cada sementeira são dadas na Tabela 7.

Note que, na Tabela 7, observaram-se estimativas de probabilidades iguais a zero. Entretanto, convém observar que obter estimativas de probabilidades iguais a zero não significa que a probabilidade de cair, por exemplo, 4 sementes por golpe para a sementeira 4 seja zero caso o tamanho da amostra seja aumentado.

Tabela 7 – Estimativas de máxima verossimilhança das probabilidades associadas às categorias das semeadoras manuais

Semeadora	Número de sementes por golpe						Total
	0	1	2	3	4	≥ 5	
1	0,0000	0,1267	0,6867	0,1200	0,0200	0,0467	1
2	0,0133	0,1733	0,4667	0,3200	0,0267	0,0000	1
3	0,0933	0,1400	0,5467	0,1800	0,0333	0,0067	1
4	0,0467	0,1267	0,5467	0,2800	0,0000	0,0000	1
5	0,2200	0,2133	0,2267	0,2800	0,0600	0,0000	1
6	0,2800	0,2133	0,2867	0,1467	0,0600	0,0133	1
7	0,1333	0,1733	0,4533	0,2067	0,0333	0,0000	1
8	0,0867	0,1867	0,3133	0,3333	0,0533	0,0267	1
9	0,2333	0,1733	0,2000	0,2000	0,1400	0,0533	1
10	0,0400	0,2733	0,5133	0,1200	0,0533	0,0000	1
11	0,2267	0,2733	0,2667	0,1800	0,0267	0,0267	1
12	0,0400	0,2533	0,4667	0,2333	0,0067	0,0000	1
13	0,2200	0,1867	0,1933	0,2533	0,1200	0,0267	1
14	0,0933	0,1867	0,4133	0,2733	0,0267	0,0067	1
15	0,1000	0,3933	0,4867	0,0200	0,0000	0,0000	1

Primeiramente, na seção 4.2.1, serão consideradas somente as semeadoras que não apresentam zeros amostrais nas categorias de resposta, para os quais o teste da razão de verossimilhanças pode ser obtido analiticamente.

4.2.1 Semeadoras sem zeros amostrais

A Tabela 8 apresenta os valores-p ajustados pelo método de Sidák para as semeadoras 3, 6, 8, 9, 11 e 13 quando comparadas duas a duas.

Tabela 8 – Comparações entre as semeadoras duas a duas e seus respectivos valores-p ajustados pelo método de Sidák

	6	8	9	11	13	14
3	0,0005	0,0510	0,0000	0,0003	0,0000	1,0000
6		0,0016	0,5191	1,0000	0,7199	0,0038
8			0,0033	0,0365	0,0877	1,0000
9				0,0831	1,0000	0,0000
11					0,2747	0,0432
13						0,0004

A partir da mesma conclui-se, considerando o nível de significância $\alpha = 5\%$, que a semeadora 3 não difere estatisticamente das semeadoras 8 e 14 quanto à distribuição do número de sementes e a semeadora 6 não difere estatisticamente das semeadoras 9, 11 e 13. Note que as semeadoras 6, 9, 11 e 13 apresentam frequências bastante elevadas de nenhuma semente por golpe (> 30), podendo acarretar falhas de semeadura e, consequentemente, prejuízos extremamente sérios. Apresentam, ainda, distribuições próximas da uniforme para os números de sementes por golpe de 0 a 3. Esse grupo de semeadoras tem, portanto, comportamento muito ruim. Por outro lado, as semeadoras 3, 8 e 14 apresentam distribuições mais concentradas ao redor da mediana, 2, sendo, por esse motivo, mais adequadas para a semeadura. No entanto, apesar da existência dos dois grupos, a semeadora 8 não difere estatisticamente da 13, não devendo por este motivo ser utilizada. De forma geral, as semeadoras recomendadas dentre as seis analisadas aqui, são a 3 e a 14.

4.2.2 Todas as semeadoras

Quando a tabela de contingência apresenta zeros amostrais, são necessárias algumas medidas para realização da análise dos dados, uma vez que, zeros amostrais causam problemas com relação às estimativas dos parâmetros de interesse, nas distribuições assintóticas das estatísticas dos testes e nos cálculos necessários para a análise. As metodologias utilizadas para comparar semeadoras com zeros amostrais foram: agrupar caselas ou adicionar uma constante em cada casela, cujos resultados são apresentados nas seções 4.2.2.1 e 4.2.2.2.

4.2.2.1 Agrupar as caselas

De modo a todas as categorias de resposta terem frequências não nulas para todas as semeadoras, foram agrupadas as caselas referentes às categorias 1 e 2 e às categorias 4, 5 e 6. Logo, a tabela de contingência resultante compreende três categorias dadas por: menos que duas sementes, exatamente duas sementes e mais que duas sementes, apresentadas na Tabela 9.

Tabela 9 – Frequências dos números de sementes por golpe considerando caselas agrupadas para as 15 semeadoras

Semeadora	Número de sementes			Total
	< 2	2	> 2	
1	19	103	28	150
2	28	70	52	150
3	35	82	33	150
4	26	82	42	150
5	65	34	51	150
6	74	43	33	150
7	46	68	36	150
8	41	47	62	150
9	61	30	59	150
10	47	77	26	150
11	75	40	35	150
12	44	70	36	150
13	61	29	60	150
14	42	62	46	150
15	74	73	3	150
Total	738	910	602	2250

Com base nos dados da Tabela 9, dada a existência de apenas 3 categorias, construiu-se a Figura 4, a partir da qual pode-se dizer que, quanto mais para baixo e à esquerda o ponto se localiza, pior é a semeadora. As semeadoras 5, 6, 9, 11 e 13, assim, aparentemente pertencem a um grupo de semeadoras bastante ruins que possuem frequências baixas (≤ 50) com relação ao número de duas sementes e frequências altas (≥ 50) de depósitos de nenhuma ou uma semente por golpe.

Note que, a semeadora 15, devido ao ponto que se localiza no gráfico, possui frequências altas com relação ao número de duas sementes (≤ 70), porém é considerada ruim devido a frequências altas (≥ 70) de depósitos de nenhuma ou uma semente por golpe.

Figura 4 – Distribuição observada das quinze semeadoras em um espaço Simplex

Os resultados obtidos por meio do teste da razão de verossimilhanças para a comparação das semeadoras duas a duas com categorias agrupadas são apresentados na Tabela 10.

Tabela 10 – Comparações entre as sementeiras duas a duas com categorias agrupadas e respectivos p-valores ajustados pelo método de Sidák. Em negrito são ressaltados os valores menores ou iguais ao nível de significância $\alpha = 5\%$

[illegible]

Por meio dos resultados apresentados na Tabela 10 pode-se, então, associá-los ao gráfico apresentado na Figura 4, unindo os pontos relativos às semeadoras não diferentes estatisticamente com um nível de significância $\alpha = 0,05$, produzindo a Figura 5. Por meio desta, entretanto, as conclusões ficam difíceis de serem tiradas, evidenciando somente o fato da semeadora 15 ser diferente das demais.

Figura 5 – Associação entre as quinze semeadoras em um espaço Simplex pela união das que possuem comportamentos semelhantes quanto à distribuição do número de sementes depositadas por golpe

Ao agrupar as categorias de resposta é possível verificar, pelo teste da razão de verossimilhanças, que muitos pares de semeadoras não diferem estatisticamente entre si, ou seja, apresentam comportamentos semelhantes quanto às distribuições dos números de sementes, esse fato ocorre devido à perda de informação ao agrupar as categorias tornando a análise conservadora.

O agrupamento das categorias, entretanto, é severamente criticado por Everitt (1992), pois uma quantidade considerável de informação pode ser perdida pela combinação

das categorias o que pode afetar o interesse e a utilidade do estudo. Além disso, a aleatoriedade da amostra pode ser afetada, além de que o modo em que as categorias são agrupadas pode ter um efeito importante sobre as inferências.

4.2.2.2 Adicionar uma constante as caselas

Segundo Agresti (2002), adicionar uma pequena constante em todas as caselas é uma maneira de se obter estimativas finitas e assegurar a convergência dos algoritmos de ajuste.

Na literatura são sugeridas pequenas constantes para tabelas de contingência contendo zeros amostrais. De acordo com Subbiah e Srinivasan (2008), a adição da constante 0,5 é, geralmente, utilizada na estimativa do logaritmo da razão das verossimilhanças. No entanto, uma análise de sensibilidade adicionando-se constantes de tamanhos variados, com intuito de avaliar o efeito na estimativa do parâmetro e nas estatísticas de adequação de ajuste, também é sugerida na literatura para situações que contenham zeros amostrais.

Desse modo, no presente trabalho foram consideradas as constantes 0,5 e 1 para se obter as estimativas do teste da razão de verossimilhanças.

A Tabela 11 apresenta os resultados do teste da razão das verossimilhanças para a comparação das sementeiras duas a duas ao se adicionar 0,5 às frequências observadas, apresentadas na Tabela 3. Analogamente, tem-se na Tabela 12, os resultados quando se adicionado o valor 1 à cada casela da Tabela 3. Apesar de todos os valores-p apresentados na Tabela 11 serem menores ou iguais aos da Tabela 12, quando se utiliza o nível de significância $\alpha = 5\%$, as conclusões são exatamente as mesmas. O fato de os valores-p serem menores mostra que o aumento do valor da constante adicionada, de 0,5 para 1, fez com que os testes fossem mais conservadores. Note que, segundo Nandram, Bhatta e Bhadra (2013), adicionar 0,5 em cada casela da tabela de contingência tende a tornar as caselas dispersas igualmente prováveis e, assim, tornar a estatística χ^2 muito conservadora, efeito muito grave quando há muitas caselas.

Tabela 11 – Comparações entre as sementeiras duas a duas e respectivos valores-p ajustados pelo método de Sidák, adicionando a constante 0,5 às caselas. Em negrito são ressaltados os valores menores ou iguais ao nível de significância $\alpha = 5\%$

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,0014	0,0118	0,0021	0,0000	0,0000	0,0000	0,0000	0,0000	0,0049	0,0000	0,0005	0,0000	0,0000	0,0000
2		0,3369	1,0000	0,0000	0,0000	0,1313	0,2794	0,0000	0,0931	0,0000	1,0000	0,0000	0,9961	0,0000
3			0,9960	0,0001	0,0011	1,0000	0,1410	0,0000	0,9404	0,0009	0,9229	0,0000	1,0000	0,0000
4				0,0000	0,0000	0,6407	0,0095	0,0000	0,0069	0,0000	1,0000	0,0000	0,9965	0,0000
5					0,9999	0,2733	0,5818	0,4733	0,0000	0,9992	0,0000	1,0000	0,2171	0,0000
6						0,2543	0,0040	0,9187	0,0000	1,0000	0,0000	0,9821	0,0099	0,0000
7							0,7927	0,0000	0,5760	0,2417	0,9491	0,0010	1,0000	0,0000
8								0,0089	0,0009	0,0953	0,1120	0,2255	1,0000	0,0000
9									0,0000	0,2426	0,0000	1,0000	0,0000	0,0000
10										0,0000	0,9927	0,0000	0,2679	0,0061
11											0,0001	0,6396	0,1230	0,0000
12												0,0000	1,0000	0,0000
13													0,0011	0,0000
14														0,0000

Tabela 12 – Comparações entre as sementeiras duas a duas e respectivos valores-p ajustados pelo método de Sidák, adicionando a constante 1 às caselas. Em negrito são ressaltados os valores menores ou iguais ao nível de significância $\alpha = 5\%$

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0,0032	0,0326	0,0084	0,0000	0,0000	0,0000	0,0000	0,0000	0,0149	0,0000	0,0017	0,0000	0,0000	0,0000
2		0,4549	1,0000	0,0000	0,0000	0,1901	0,4596	0,0000	0,1171	0,0000	1,0000	0,0000	0,9996	0,0000
3			0,9999	0,0002	0,0014	1,0000	0,1698	0,0000	0,9691	0,0012	0,9698	0,0000	1,0000	0,0000
4				0,0000	0,0000	0,8135	0,0253	0,0000	0,0139	0,0000	1,0000	0,0000	0,9998	0,0000
5					1,0000	0,3015	0,7335	0,6909	0,0000	1,0000	0,0000	1,0000	0,2678	0,0000
6						0,3258	0,0052	0,9563	0,0000	1,0000	0,0000	0,9894	0,0128	0,0000
7							0,9012	0,0001	0,6470	0,3411	0,9799	0,0018	1,0000	0,0000
8								0,0117	0,0015	0,1136	0,2087	0,2625	1,0000	0,0000
9									0,0000	0,3116	0,0000	1,0000	0,0000	0,0000
10										0,0000	0,9985	0,0000	0,3343	0,0151
11											0,0003	0,7188	0,1528	0,0000
12												0,0000	1,0000	0,0000
13													0,0018	0,0000
14														0,0000

Em resumo, tem-se que, ao nível de significância $\alpha = 5\%$, a semeadora 1 difere estatisticamente das demais, bem como a semeadora 15. Entre as restantes não há aparente formação de grupos distintos. No caso em questão, como a semeadora 1 difere das demais e apresenta posição mais alta no gráfico da Figura 4, ou seja, com maior frequência de valores iguais a 2 sementes por golpe, seria a recomendada aqui.

Comparando-se, por outro lado, os resultados apresentados nas Tabelas 11 e 12 com os da Tabela 10, observa-se que há uma quantidade muito maior de comparações consideradas significativas ao nível de significância $\alpha = 5\%$. Esse fato se deve provavelmente à perda de informação ao se agruparem categorias.

4.3 Inferência bayesiana

Na Tabela 13, são apresentadas as médias a posteriori para os parâmetros referentes a cada categoria de resposta para as quinze semeadoras manuais utilizando-se a distribuição a priori Dirichlet com parâmetros iguais a 0,5, chamada a priori não-informativa de Jeffreys.

Tabela 13 – Médias a posteriori para os parâmetros utilizando a priori não-informativa de Jeffreys, dadas por $(n_{ij} + 0,5)/153$

Semeadora	Número de sementes por golpe						Total
	0	1	2	3	4	≥ 5	
1	0,0033	0,1275	0,6765	0,1209	0,0229	0,0490	1
2	0,0163	0,1732	0,4608	0,3170	0,0294	0,0033	1
3	0,0948	0,1405	0,5392	0,1797	0,0359	0,0098	1
4	0,0490	0,1275	0,5392	0,2778	0,0033	0,0033	1
5	0,2190	0,2124	0,2255	0,2778	0,0621	0,0033	1
6	0,2778	0,2124	0,2843	0,1471	0,0621	0,0163	1
7	0,1340	0,1732	0,4477	0,2059	0,0359	0,0033	1
8	0,0882	0,1863	0,3105	0,3301	0,0556	0,0294	1
9	0,2320	0,1732	0,1993	0,1993	0,1405	0,0556	1
10	0,0425	0,2712	0,5065	0,1209	0,0556	0,0033	1
11	0,2255	0,2712	0,2647	0,1797	0,0294	0,0294	1
12	0,0425	0,2516	0,4608	0,2320	0,0098	0,0033	1
13	0,2190	0,1863	0,1928	0,2516	0,1209	0,0294	1
14	0,0948	0,1863	0,4085	0,2712	0,0294	0,0098	1
15	0,1013	0,3889	0,4804	0,0229	0,0033	0,0033	1

Assim como na Tabela 13, são apresentados na Tabela 14, as médias a posteriori para os parâmetros referentes a cada categoria de resposta para as quinze semeadoras

manuais utilizando-se, entretanto, a distribuição a priori uniforme no intervalo $[0, 1]$.

Tabela 14 – Médias a posteriori para os parâmetros utilizando a priori uniforme, dadas por $(n_{ij} + 1)/156$

Semeadora	Número de sementes por golpe						Total
	0	1	2	3	4	≥ 5	
1	0,0064	0,1282	0,6667	0,1218	0,0256	0,0513	1
2	0,0192	0,1731	0,4551	0,3141	0,0321	0,0064	1
3	0,0962	0,1410	0,5321	0,1795	0,0385	0,0128	1
4	0,0513	0,1282	0,5321	0,2756	0,0064	0,0064	1
5	0,2179	0,2115	0,2244	0,2756	0,0641	0,0064	1
6	0,2756	0,2115	0,2821	0,1474	0,0641	0,0192	1
7	0,1346	0,1731	0,4423	0,2051	0,0385	0,0064	1
8	0,0897	0,1859	0,3077	0,3269	0,0577	0,0321	1
9	0,2308	0,1731	0,1987	0,1987	0,1410	0,0577	1
10	0,0449	0,2692	0,5000	0,1218	0,0577	0,0064	1
11	0,2244	0,2692	0,2628	0,1795	0,0321	0,0321	1
12	0,0449	0,2500	0,4551	0,2308	0,0128	0,0064	1
13	0,2179	0,1859	0,1923	0,2500	0,1218	0,0321	1
14	0,0962	0,1859	0,4038	0,2692	0,0321	0,0128	1
15	0,1026	0,3846	0,4744	0,0256	0,0064	0,0064	1

Pode-se observar que quando comparadas as Tabelas 13 e 14, as médias a posteriori obtidas diferem pouco entre si, bem como diferem pouco das estimativas obtidas por máxima verossimilhança apresentadas na Tabela 7. Devido à informação que se tem a priori na metodologia bayesiana, obtém-se médias a posteriori não nulas para as caselas que continham zeros amostrais, diferentemente da abordagem clássica.

Note que, ao considerar o mesmo valor para cada parâmetro da distribuição a priori Dirichlet e quanto maior for o mesmo, tem-se que a distribuição do número de sementes por golpe para cada semeadora será mais próxima da distribuição uniforme.

4.3.1 Fator de Bayes

Por meio das verossimilhanças marginais obtidas em função dos dados amostrais e das distribuições a priori, foram calculados os fatores de Bayes para comparar as semeadoras duas a duas. Os resultados podem ser visualizados considerando-se a priori não-informativa de Jeffreys e a priori uniforme, respectivamente, nas Tabelas 15 e 16.

Tabela 15 – Valores de $2 \log B_{10}$ considerando a priori não-informativa de Jeffreys. Em negrito são ressaltados os valores maiores que 6

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	14,820	10,514	15,796	88,839	74,738	30,359	41,332	91,575	13,035	68,198	17,982	92,787	27,087	53,151
2		0,347	-8,453	30,197	47,789	2,208	1,874	61,436	2,608	38,919	-10,443	48,403	-6,096	58,494
3			-5,270	17,836	13,261	-14,110	2,193	33,810	-4,769	13,888	-4,202	28,813	-11,305	30,387
4				35,301	45,714	-1,828	11,158	69,425	10,098	37,403	-9,852	57,802	-5,509	45,986
5					-7,398	-0,451	-1,011	0,307	30,050	-6,253	23,076	-9,193	1,155	63,047
6						1,076	10,435	-4,397	25,808	-12,799	31,787	-5,870	8,433	41,490
7							-2,504	22,296	-2,593	1,655	-5,342	14,816	-14,374	28,452
8								8,786	14,974	3,072	4,376	0,883	-11,782	64,796
9									48,943	1,023	56,519	-15,305	23,260	86,372
10										23,571	-6,551	43,712	0,646	10,951
11											19,801	-2,263	2,597	34,408
12												45,035	-9,658	22,330
13													13,763	84,332
14														38,879

Tabela 16 – Valores de $2 \log B_{10}$ considerando a priori uniforme. Em negrito são ressaltados os valores maiores que 6

	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	10,830	7,669	9,923	84,659	72,209	26,296	39,506	89,768	9,120	65,971	13,364	90,688	24,265	46,100
2		-1,347	-12,086	28,750	46,113	0,656	0,540	59,804	59,804	37,121	-12,775	46,892	-7,634	53,421
3			-8,213	17,694	14,064	-14,445	3,041	34,858	-5,601	14,592	-5,890	29,828	-10,993	26,508
4				32,454	42,693	-4,591	8,325	66,237	6,769	34,514	-12,845	54,739	-8,246	41,668
5					-6,733	-0,217	-0,654	1,123	29,594	-5,931	21,617	-8,302	1,177	59,333
6						1,220	12,162	-1,935	25,303	-10,968	30,170	-3,544	9,319	37,939
7							-2,551	22,435	-3,242	1,553	-6,823	15,027	-14,607	24,810
8								11,291	14,440	4,871	2,904	3,218	-10,777	60,828
9									48,480	3,426	54,759	-12,052	24,396	82,426
10										22,773	-8,401	43,297	-0,140	6,931
11											18,247	-0,053	3,468	30,926
12												43,397	-11,135	18,288
13													14,903	80,413
14														35,095

De acordo com a Tabela 2, para a interpretação do fator de Bayes, conclui-se que para valores maiores que seis a evidência contra a hipótese nula é forte. Rejeita-se, assim, a hipótese de que as duas semeadoras apresentam o mesmo comportamento quanto a distribuição do número de sementes por golpe. No caso em questão, comparando-se os valores de $2 \log B_{10}$ apresentados nas Tabelas 15 e 16 com o valor adotado de referência 6, chegam-se às mesmas conclusões que as obtidas a partir das análises das Tabelas 11 e 12, ou seja, não há aparente formação de grupos distintos. A semeadora 1 difere das demais, bem como a semeadora 15, mas como apresenta maior frequência de valores iguais a 2 sementes por golpe é a recomendada.

4.4 Análise de Agrupamentos

Do exposto até aqui nas seções 4.2 e 4.3, ficam evidentes as dificuldades para se tomarem decisões, a partir das comparações das semeaduras duas a duas. Isso ocorre a não ser nos casos em que há poucas semeadoras a serem comparadas e/ou em que há grupos de semeadoras semelhantes, grupos esses com comportamentos distintos. Nesse sentido, a análise de agrupamentos é a metodologia mais adequada para a visualização dos grupos distintos de semeadoras semelhantes entre si.

No caso, cada semeadora representa uma amostra finita de observações independentes e identicamente distribuídas de uma variável aleatória multinomial. Logo, como cada semeadora é representada por uma distribuição de probabilidade, foi utilizada a medida *J-divergência* como medida de dissimilaridade para a comparação de duas distribuições de probabilidade, distinguindo-se diferenças entre as semeadoras.

Como métodos de agrupamentos, utilizaram-se os aglomerativos hierárquicos: distância média, vizinho mais distante e vizinho mais próximo, pois estes baseiam-se somente na matriz de distâncias. Assim, as semeadoras foram classificadas em grupos em diferentes etapas de forma hierárquica produzindo um dendrograma representando uma árvore de classificação.

Para determinar o ponto de corte na escala de distâncias, obtendo o número de grupos que otimiza a qualidade de ajuste do agrupamento aos dados, foi utilizado o método proposto por Mojena (1977) com a constante sugerida por Milligan e Cooper (1985) de $k = 1,25$. Foram realizadas as análises de agrupamentos para ambas as abordagens: clássica e bayesiana.

4.4.1 Análise de agrupamentos na abordagem clássica

A análise de agrupamentos na abordagem clássica foi realizada inicialmente considerando-se as distribuições de probabilidade somente para as sementeiras sem zeros amostrais e, posteriormente, incluindo as sementeiras com zeros amostrais agrupando caselas ou adicionando constantes.

4.4.1.1 Análise de agrupamentos para sementeiras sem zeros amostrais

Os dendrogramas foram obtidos para cada método aglomerativo, como mostra a Figura 6.

Figura 6 – Dendrograma para sementeiras sem zeros amostrais, obtido utilizando os métodos: (a) da distância média; (b) do vizinho mais distante e (c) do vizinho mais próximo

Os valores de corte (linhas em azul tracejadas) e os coeficientes de correlação cofenética obtidos para cada um dos métodos utilizados para o agrupamento das sementeiras são apresentados na Tabela 17.

Tabela 17 – Valores de corte e de coeficientes de correlação cofenética conforme o método de agrupamento utilizado para sementeiras sem zeros amostrais

Método de agrupamento	Valor de corte	Coeficiente de correlação cofenética
Distância média	0,5122	0,7877
Vizinho mais distante	0,8272	0,7852
Vizinho mais próximo	0,3150	0,7724

Como se pode observar na Figura 6, todos os métodos considerados levaram à divisão em dois grupos: um com as sementeiras 3, 8 e 14 e outro com as sementeiras 6, 9, 11 e 13. O método de agrupamento da distância média, entretanto, levou o maior coeficiente de correlação cofenética, 0,7877 (ver Tabela 17), valor aproximadamente igual a 0,8, indicando bom ajuste, segundo Bussab, Miazaki e Andrade (1990). Logo, pode-se dizer que a representação da matriz de dissimilaridade com a matriz cofenética de cada método aglomerativo estão compatíveis entre si e que cada representação pelo dendrograma é válida. Esses resultados estão de acordo com os obtidos utilizando as comparações das sementeiras duas a duas, apresentado na subseção 4.2.1. As conclusões, portanto, seriam as mesmas já discutidas, ou seja, dá-se preferência pelas sementeiras 3, 8 ou 14.

4.4.1.2 Análise de agrupamentos para todas as sementeiras agrupando-se as caselas

Neste caso, procedeu-se a análise dos dados da Tabela 9, cujos resultados são apresentados nos dendrogramas da Figura 7. Observa-se, entretanto, que o método do vizinho mais distante, diferentemente dos demais, levou a formação de 3 grupos distintos. Por outro lado, o maior valor de correlação cofenética, 0,7964 (ver Tabela 18), foi obtido do uso do método da distância média, valor este aproximadamente igual a 0,8, indicando bom ajuste. Nesta análise, a sementeira 15 foi considerada diferente das demais, conclusão igual à obtida pelo método de comparações das sementeiras duas a duas apresentado na subseção 4.2.2.1.

Figura 7 – Dendrograma para semeadoras com caselas agrupadas, obtido utilizando os métodos: (a) da distância média; (b) do vizinho mais distante e (c) do vizinho mais próximo

Tabela 18 – Valores de corte e de coeficientes de correlação cofenética conforme o método de agrupamento utilizado para semeadoras com caselas agrupadas

Método de agrupamento	Valor de corte	Coeficiente de correlação cofenética
Distância média	0,6668	0,7964
Vizinho mais distante	1,1749	0,7067
Vizinho mais próximo	0,2811	0,7568

Note que, agrupar as caselas pode deixar as frequências mais próximas dentro de cada categoria. Com essa perda de informação, as semeadoras tendem a não diferir entre si, tornando mais difícil a formação de grupos, ou seja, tornando o método excessivamente

conservador.

4.4.1.3 Análise de agrupamentos para todas as semeadoras com adição de constantes nas caselas

Após adicionar uma constante igual a 0,5 em cada casela da tabela de contingência, Tabela 3, obtiveram-se as estimativas das probabilidades para o cálculo da matriz de dissimilaridades utilizada pelos métodos aglomerativos, exatamente iguais às médias das distribuições a posteriori utilizando a priori não-informativa de Jeffrey, apresentadas na Tabela 13. Analogamente, adicionar a constante 1 às caselas produz estimativas das probabilidades exatamente iguais às médias das distribuições a posteriori utilizando a priori uniforme, apresentadas na Tabela 14

A Figura 8, apresenta os dendrogramas referentes aos diferentes métodos aglomerativos, nos quais percebe-se que o agrupamento das semeadoras varia com os métodos. Entretanto, pelo fato do método do vizinho mais próximo obter o maior coeficiente de correlação cofenética, 0,6995 (ver Tabela 19), dentre os três métodos, adotou-se o agrupamento: (i) semeadora 1; (ii) semeadora 15 e (iii) demais semeadoras. Mais uma vez, as conclusões obtidas pela análise de agrupamentos foram iguais às obtidas pela comparação das semeadoras duas a duas, apresentado na subseção 4.2.2.2. Permanece, assim, a recomendação da semeadora 1 conforme discutido na subseção citada.

Figura 8 – Dendrograma para semeadoras manuais com adição de 0,5 nas caselas, obtido utilizando os métodos: (a) da distância média, (b) do vizinho mais distante e (c) do vizinho mais próximo

Tabela 19 – Valores de corte e de coeficientes de correlação cofenética conforme o método de agrupamento utilizado para semeadoras com adição da constante 0,5 nas caselas

Método de agrupamento	Valor de corte	Coeficiente de correlação cofenética
Distância média	0,8898	0,6752
Vizinho mais distante	1,5145	0,6043
Vizinho mais próximo	0,4279	0,6995

As conclusões foram exatamente as mesmas quando se considera a adição da constante 1 a todas as caselas da Tabela 3, cujos resultados são apresentados na Figura 9

e Tabela 20.

Figura 9 – Dendrograma para semeadoras manuais com adição de 1 nas caselas, obtido utilizando os métodos: (a) da distância média, (b) do vizinho mais distante e (c) do vizinho mais próximo

Tabela 20 – Valores de corte e coeficientes de correlação cofenética conforme o método de agrupamento utilizado para semeadoras com adição da constante 1 nas caselas

Método de agrupamento	Valor de corte	Coeficiente de correlação cofenética
Distância média	0,8029	0,6911
Vizinho mais distante	1,3997	0,6418
Vizinho mais próximo	0,3704	0,6973

4.4.2 Análise de agrupamentos na abordagem bayesiana

Na abordagem bayesiana, foi calculada a média a posteriori dos parâmetros para cada casela da tabela de contingência, Tabela 3, utilizando-se a distribuição a priori Dirichlet não-informativa de Jeffreys e a Uniforme. Com base nesses valores, foi construída a matriz de dissimilaridades entre as sementeiras e utilizado os métodos de agrupamentos aglomerativos.

Entretanto, dado o fato de levarem a exatamente as mesmas matrizes de dissimilaridades obtidas quando se adicionam as constantes 0,5 ou 1 às caselas da Tabela 3, respectivamente, os resultados são exatamente iguais aos obtidos na seção anterior.

5 CONCLUSÕES

Quando há poucas semeadoras, a metodologia proposta pode ser utilizada para identificar grupos distintos de semeadoras manuais quanto à distribuição do número de sementes comparando-as duas a duas e, assim, optar por aquelas que apresentam melhores comportamentos. É importante notar que quando se utilizou a média como medida de discriminação, as semeadoras 3 e 13 apresentaram, aproximadamente, médias similares não diferindo entre si quanto ao número de sementes depositadas por golpe, no entanto, por meio da metodologia utilizada foi possível verificar diferença significativa entre as mesmas com o teste da razão das verossimilhanças. Porém, não considerar todas as semeadoras geralmente não vai responder as perguntas de interesse dos pesquisadores.

Agrupar caselas na tabela de contingência pode levar a perda de informação em relação ao número de sementes depositado por golpe de modo que os testes e a análise de agrupamento se tornem mais conservadores, sendo difícil a identificação de diferentes grupos de semeadoras. Somente a semeadora 15 diferiu das demais, entretanto, não seria recomendada aqui devido as altas frequências de menos que duas sementes depositadas por golpe.

Adicionar constantes 0,5 ou 1 em todas as caselas da tabela de contingência obtém-se na abordagem clássica estimativas de máxima verossimilhança semelhantes às estimativas das médias a posteriori dos parâmetros, na bayesiana, com priori de Jeffreys ou uniforme. Logo, comparar as semeadoras duas a duas dificulta a tomada de conclusões a menos que haja grupos de semeadoras bem distintos. Assim, a semeadora recomendada por meio da metodologia apresentada é a primeira, pois além de diferir das demais foi aquela que mais depositou duas sementes por golpe, o exigido pelo experimento agrônomo. Além disso, a análise de agrupamento para ambas as abordagens provoca necessariamente as mesmas matrizes de distâncias e, portanto, os mesmos agrupamentos quando utilizados distintos métodos hierárquicos aglomerativos. Os resultados obtidos reafirmam as diferenças entre as semeadoras e as conclusões citadas acima.

5.1 Trabalhos futuros

Como ideias para trabalhos futuros, pretende-se:

- (i) considerar que a variável número de sementes por golpe segue outras distribuições

paramétricas como, por exemplo, a de Poisson e comparar as semeadoras utilizando os métodos sugeridos neste trabalho.

- (ii) utilizar modelos de classes latentes por meio de variáveis latentes (que não podem ser observadas diretamente) utilizadas para explicar a associação existente entre as variáveis observadas e, assim, identificar e caracterizar grupos de casos similares.
- (iii) fazer a estimação da metodologia apresentada para casos em que há mais de uma repetição, isto é, mais de uma semeadora por fabricante.
- (iv) repetir o experimento para algumas semeadoras, porém não perdendo a informação de que os números de sementes por golpe foram obtidos sequencialmente, isto é, formando uma série temporal.
- (v) incluir técnicas de análise de componentes principais e representação por meio de biplots para o auxílio na escolha das semeadoras ideais, uma vez que é uma técnica bastante útil que pode indicar a existência de agrupamentos entre observações.

REFERÊNCIAS

- AGRESTI, A. **Categorical data analysis**. 2. ed. Hoboken: John Wiley & Sons, 2002. 710 p.
- AIKINS, S.H.M.; BART-PLANGE, A.; OPOKU-BAFFOUR, S. Performance evaluation of jab planters for maize planting and inorganic fertilizer application. **ARPJ Journal of Agricultural and Biological Science**, Islamabad, v. 5, n. 1, p. 29–33, Jan. 2010.
- ALMEIDA, R.A. Adaptação da “matraca” ao plantio direto em pequenas propriedades. In: ENCONTRO LATINO AMERICANO SOBRE PLANTIO DIRETO NA PEQUENA PROPRIEDADE, 1993, Ponta Grossa. **Anais...** Ponta Grossa: IAPAR, 1993. p. 251–257.
- ANJOS, J.B.; DRUMOND, M.A. Adaptação de semeadora manual para plantio direto de mamona. In: CONGRESSO BRASILEIRO DE MAMONA-ENERGIA E SUSTENTABILIDADE, 2004, Campina Grande. **Anais...** Campina Grande: Embrapa Algodão, 2004. p. 4.
- BENDER, R.; LANGE, S. Adjusting for multiple testing - when and how?. **Journal of clinical epidemiology**, Amsterdam, v. 54, n. 4, p. 343–349, 2001.
- BERRY, D.A.; HOCHBERG, Y. Bayesian perspectives on multiple comparisons. **Journal of Statistical Planning and Inference**, Amsterdam, v. 82, n. 1, p. 215–227, 1999.
- BISHOP, Y.M.; FIENBERG, S.E.; HOLLAND, P.W. **Discrete multivariate analysis: theory and practice**. New York: Springer Science, 2007. 559 p.
- BORGATTO, A.F. **Modelos para Proporções com Superdispersão e Excesso de Zeros - Um Procedimento Bayesiano**. 2004. 104 p. Tese (Doutorado em Estatística e Experimentação Agronômica) — Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 2004.
- BUSSAB, W.O.; MIAZAKI, E.S.; ANDRADE, D.F. Introdução à análise de agrupamentos. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 9., 1990, São Paulo. **Anais...** São Paulo: ABE, 1990. p. 105.
- CASELLA, G.; BERGER, R.L. **Statistical inference**. 2. ed. Pacific Grove: Duxbury, 2002. 660 p.
- CASTELLÓ, P.; SBERT, M.; CHOVER, M.; FEIXAS, M. Viewpoint-based simplification using f-divergences. **Information Sciences**, London, v. 178, n. 11, p. 2375–2388, June 2008.
- CHA, S.H. Comprehensive survey on distance/similarity measures between probability density functions. **International Journal of Mathematical Models and Methods in Applied Sciences**, Singapore, v. 1, n. 4, p. 300–307, 2007.
- COELHO, J. L. D. Ensaio e certificação das máquinas para a semeadura. In: MIALHE, L. G. (Ed.). **Máquinas Agrícolas: ensaios e certificação**. Piracicaba: FEALQ, 1996. cap. 11, p. 551–570.
- EHLERS, R.S. **Inferência Bayesiana**. São Carlos: Departamento de Matemática Aplicada e Estatística, ICMC/USP, 2011. 106 p.

EVERITT, B.S. **The analysis of contingency tables**. 2. ed. London: Chapman & Hall/CRC, 1992. 164 p.

EVERITT, B.S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster Analysis**. 5. ed. Chichester: John Wiley & Sons, 2011. 330 p.

FAO. **Farming for the Future in Southern Africa**: An introduction to conservation agriculture. Johannesburg, 2010. 4 p.

FERREIRA, D.F. **Estatística multivariada**. 2. ed. Lavras: UFLA, 2011. 676 p.

FINKLER, A. **Goodness of fit statistics for sparse contingency tables**. 2010. Disponível em <<http://arxiv.org/pdf/1006.2620v1.pdf>>. Acesso em: 25 mar. 2015.

FREI, F. **Introdução à análise de agrupamentos**: teoria e prática. São Paulo: UNESP, 2006. 112 p.

GELMAN, A.; CARLIN, J.B.; STERN, H.S.; DUNSON, D.B.; VEHTARI, A.; RUBIN, D.B. **Bayesian data analysis**. 3. ed. Boca Raton: Chapman & Hall/CRC, 2013. 675 p.

GIOLO, S.R. **Análise de dados discretos**. Curitiba: Departamento de Estatística, UFPR, 2004. 90 p.

IBGE. **Censo Agropecuário 2006**: Agricultura Familiar. Primeiros Resultados. Brasil, Grandes Regiões e Unidades da Federação. 2006. 267 p.

JOHANSEN, C.; HAQUE, M.E.; BELL, R.W.; THIERFELDER, C.; ESDAILE, R.J. Conservation agriculture for small holder rainfed farming: opportunities and constraints of new mechanized seeding systems. **Field Crops Research**, Amsterdam, v. 132, p. 18–32, June 2012.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. 6. ed. Upper Saddle River: Pearson Prentice Hall, 2007. 773 p.

JOUSSELME, A.-L.; MAUPIN, P. Distances in evidence theory: Comprehensive survey and generalizations. **International Journal of Approximate Reasoning**, Paris, v. 53, n. 2, p. 118–145, 2012.

KAPLAN, D.; DEPAOLI, S. Bayesian statistical methods. In: LITTLE, T.D. (Ed.). **The Oxford Handbook of Quantitative Methods**. Oxford: Oxford University Press, 2013. cap. 20, p. 407–437.

KASS, R.E.; RAFTERY, A.E. Bayes factors. **Journal of the American Statistical Association**, Alexandria, v. 90, n. 430, p. 773–795, Mar. 1995.

KATERI, M. **Contingency Table Analysis**: methods and implementation using R. New York: Springer, 2014. 304 p.

KRAUS, K. **On the Measurement of Model Fit for Sparse Categorical Data**. 2012. 24 p. Tese (Doctor of Philosophy) — Faculty of Social Sciences, Uppsala University, Uppsala, 2012.

- KUMAR, P.; CHHINA, S. A symmetric information divergence measure of the csiszár's f-divergence class and its bounds. **Computers & Mathematics with Applications**, Amsterdam, v. 49, n. 4, p. 575–588, Feb. 2005.
- LATTIN, J.; CARROLL, J.D.; GREEN, P.E. **Análise de dados multivariados**. São Paulo: Cengage Learning, 2011. 475 p.
- LE, S.Q.; HO, T.B. An association-based dissimilarity measure for categorical data. **Pattern Recognition Letters**, Haarlem, v. 26, n. 16, p. 2549–2557, 2005.
- LEANDRO, R.A. **Introdução à estatística Bayesiana**. Piracicaba: USP/ESALQ, 2001. 51 p.
- LEHMANN, E.L. **Testing statistical hypotheses**. 2. ed. New York: Wiley, 1986. 600 p.
- LEWIS, F.; BUTLER, A.; GILBERT, L. A unified approach to model selection using the likelihood ratio test. **Methods in Ecology and Evolution**, London, v. 2, n. 2, p. 155–162, Apr. 2011.
- MATOS, R.A. **Comparação de Metodologias de Análise de Agrupamentos na Presença de Variáveis Categóricas e Contínuas**. 2007. 156 p. Dissertação (Mestrado em Estatística) — Universidade Federal Minas Gerais, Belo Horizonte, 2007.
- MIALHE, L.G. **Máquinas agrícolas para plantio**. Campinas: Millennium, 2012. 623 p.
- MILLIGAN, G.W.; COOPER, M.C. An examination of procedures for determining the number of clusters in a data set. **Psychometrika**, New York, v. 50, n. 2, p. 159–179, June 1985.
- MINGOTI, S.A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: UFMG, 2005. 295 p.
- MOITA NETO, J.M. **Estatística multivariada - Uma visão didática-metodológica**. 2004. Disponível em: <http://criticanarede.com/cien_estatistica.html>. Acesso em: 15 jan. 2015.
- MOJENA, R. Hierarchical grouping methods and stopping rules: an evaluation. **The Computer Journal**, Oxford, v. 20, n. 4, p. 359–363, 1977.
- MOLIN, J.P.; D'AGOSTIN, V. Development of a rolling punch planter for stony soil conditions. **Agricultural Mechanization in Asia, Africa and Latin America**, Tokyo, v. 27, n. 3, p. 17–19, 1996.
- MOLIN, J.P.; MENEGATTI, L.A.A.; GIMENEZ, L.M. Avaliação do desempenho de semeadoras manuais. **Revista Brasileira de Engenharia Agrícola e Ambiental**, Campina Grande, v. 5, n. 2, p. 339–343, jul. 2001.
- MOOD, A.M.; GRAYBILL, F.A.; BOES, D.C. **Introduction to the Theory of Statistics**. 3. ed. New York: McGraw-Hill, 1974. 564 p.
- NANDRAM, B.; BHATTA, D.; BHADRA, D. A likelihood ratio test of quasi-independence for sparse two-way contingency tables. **Journal of Statistical Computation and Simulation**, Philadelphia, v. 85, n. 2, p. 284–304, 2013.

NTZOUFRAS, I. **Bayesian modeling using WinBUGS**. Hoboken: John Wiley & Sons, 2009. 492 p.

PAULINO, C.D.M.; SINGER, J.M. **Análise de dados categorizados**. São Paulo: Edgard Blucher, 2006. 629 p.

PAULINO, C.D.M.; TURKMAN, M.A.A.; MURTEIRA, B. **Estatística bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003. 446 p.

POSADA, D.; BUCKLEY, T.R. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. **Systematic biology**, Oxford, v. 53, n. 5, p. 793–808, June 2004.

ROMEIRO FILHO, E. Uma abordagem centrada no usuário para o projeto de máquinas agrícolas de tração animal. **Gestão & Produção**, São Carlos, v. 19, n. 1, p. 93–102, 2012.

SHAFFER, J.P. Multiple hypothesis testing. **Annual review of psychology**, Palo Alto, v. 46, n. 1, p. 561–584, Jan. 1995.

SILVA, O.R.R.F.; GONDIM, T.M. de S.; VASCONCELOS, R. de A.; PEREIRA, J.R.; BELTRÃO, N.E. de M. Avaliação de diferentes formas de plantio da semente de algodão lintada, deslintada e deslintada e tratada. In: CONGRESSO BRASILEIRO DE ALGODÃO, 4., 2003, Goiânia. **Anais...** Campina Grande: Embrapa Algodão, 2003. p. 251–257.

SRIPERUMBUDUR, B.K.; FUKUMIZU, K.; GRETTON, A.; SCHÖLKOPF, B.; LANCKRIET, G.R.G. **On Integral Probability Metrics, ϕ -Divergences and Binary Classification**. 2009. Disponível em <<http://arxiv.org/pdf/0901.2698.pdf>>. Acesso em: 5 abr. 2015.

SUBBIAH, M.; SRINIVASAN, M.R. Classification of 2×2 sparse data sets with zero cells. **Statistics & Probability Letters**, Amsterdã, v. 78, n. 18, p. 3212–3215, Dec. 2008.

UKATU, A.C. A multi-seed jab planter. **International journal of tropical agriculture**, New Delhi, v. 19, n. 1-4, p. 131–140, 2001.

VOELKER, D. **Stepping Forward**: Hand-held corn planters. 2009. Disponível em: <<http://www.farmcollector.com/equipment/hand-held-corn-planters.aspx>>. Acesso em: 9 nov. 2014.

WALL, P.C. Tailoring conservation agriculture to the needs of small farmers in developing countries: an analysis of issues. **Journal of Crop Improvement**, Philadelphia, v. 19, n. 1-2, p. 137–155, 2007.

WEST, B.T.; WELCH, K.B.; GALECKI, A.T. **Linear mixed models: a practical guide using statistical software**. Boca Raton: Chapman & Hall/CRC Press, 2007. 353 p.

WESTFALL, P.H.; TOBIAS, R.D.; ROM, D.; WOLFINGER, R.D.; HOCHBERG, Y. **Multiple comparisons and multiple tests using the SAS system**. Cary: SAS Institute, 1999. 397 p.

WESTFALL, P.H.; WOLFINGER, R.D. Multiple tests with discrete distributions. **The American Statistician**, Philadelphia, v. 51, n. 1, p. 3–8, 1997.

WRIGHT, S.P. Adjusted p-values for simultaneous inference. **Biometrics**, Arlington, v. 48, n. 4, p. 1005–1013, Dec. 1992.

APÊNDICE

APÊNDICE A - Script do *software* R

```
rm(list=ls(all=TRUE))
sem = read.table("dados1.txt", header=T); sem      #leitura dos dados
#####
# Gráficos das distribuições de frequências
par(mfrow=c(1,5))
plot(0:5,sem[1,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 1")
plot(0:5,sem[2,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 2")
plot(0:5,sem[3,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 3")
plot(0:5,sem[4,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 4")
plot(0:5,sem[5,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 5")
plot(0:5,sem[6,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 6")
plot(0:5,sem[7,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 7")
plot(0:5,sem[8,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 8")
plot(0:5,sem[9,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 9")
plot(0:5,sem[10,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 10")
plot(0:5,sem[11,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 11")
plot(0:5,sem[12,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 12")
plot(0:5,sem[13,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 13")
plot(0:5,sem[14,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 14")
plot(0:5,sem[15,],type="h",lwd=2,ylim=c(0,110),xlab="nº sementes", ylab="Frequência",main
="Semeadora 15")
#####
#medidas de posição e dispersão
f = function(sem,num){
media=mediana=moda=desvio=ampl_inter=vector()

for (i in 1:dim(sem)[1]){
freq=rep(num,sem[i,])
tab_freq=table(freq)
media[i] = mean(freq)
mediana[i]= median(freq)
moda[i]=names(tab_freq)[tab_freq == max(tab_freq)]
desvio[i] = sd(freq)
ampl_inter[i] = IQR(freq)
}
Resultado = cbind(media,mediana,moda=as.numeric(moda),desvio,ampl_inter)
return("Result"=Resultado)
}
num=c(0:5) #entrar com o vetor contendo os números de sementes por golpe
Medidas=f(sem,num)
```

```

round(Medidas,4)
#####
#Estimativas por máxima verossimilhança
g = function(sem){
max_veross=matrix(numeric(0),dim(sem)[1],dim(sem)[2])
for(i in 1:dim(sem)[1]){
for(j in 1:dim(sem)[2]){
max_veross[i,j]=sem[i,j]/sum(sem[i,])
}
}
return("Result"= max_veross)
}

(g(sem))
round(g(sem),4)
#####

#Comparações das sementeiras duas a duas utilizando a estatística do teste da razão das
#verossimilhanças e do valor-p ajustado

h = function(Sem){
S = dim(Sem)[1] # número de colunas (ou linhas) da matriz de comparações
ncombina=choose(dim(Sem)[1],2)
Semeadora=matrix(numeric(0),ncombina,2)
C = combn(1:S,2) #combinações de 1 até S duas a duas
Comp = t(C) #matriz que recebe todas as possíveis comparações
Semeadora[,1:2]=as.numeric(row.names(Sem))[Comp[,1:2]]
TRV = Pvalue.fit = vector()
k=0
for(i in 1:(S-1)) {
for (j in i:(S-1)) {
k=k+1
total.linha = sum(Sem[i,])
total = sum(Sem[c(i,j+1),])# tamanho da amostra
y=Sem[i,]
y1=Sem[j+1,]
y2=apply(Sem[c(i,j+1),],2,sum)

int=int1=int2=0

for(l in 1:length(y)){ if(y[l]==0){cont=0}else{cont= y[l]*log(y[l])}}
int=int+cont
if(y1[l]==0){cont1=0}else{cont1= y1[l]*log(y1[l])}
int1=int1+cont1
if(y2[l]==0){cont2=0}else{cont2= y2[l]*log(y2[l])}
int2=int2+cont2
}

TRV[k] = as(2*(total*log(total)-total*log(total.linha)+int+int1-int2),"numeric")

Pvalue.fit[k] = 1-(1-(pchisq(TRV[k],df=5,lower.tail=FALSE)))^ncombina
}
}
Resultado = cbind(Semeadora,TRV,Pvalue.fit)
return("Result"=Resultado)
}

Sem = sem[c(3,6,8,9,11,13,14),] # sementeiras sem zeros amostrais

```

```

Sem = sem+0.5;   #adicionando a constante 0,5
Sem = sem+1;    #adicionando a constante 0,5
Sem = read.table("dados2.txt", header=T); sem2   #leitura dos dados com caselas agrupadas
(h(Sem))
round(h(Sem),4)
#####
#Gráfico simplex
require(compositions)
agrup = read.table("dadosnovo.txt", header=T)
dados_comp<- rcomp(agrup) #Declarar o conjunto de dados a ser composicional
plot.rcomp(dados_comp,id=T,idlabs= c("1","2","3","4","5","6","7","8","9","10",
"11","12","13","14","15"),idcol=4,pch=20,axes=T) # plota o diagrama triangular
ellipses(mean(dados_comp),var(dados_comp),r=2,col="red") # região preditiva simplex 2-sigma
agrup[-15,]
lines(rcomp(agrup[-15,]))
#####
#Cáculo da média a posteriori dos parâmetros

t = function(sem,prior){
media_post=matrix(numeric(0),dim(sem)[1],dim(sem)[2])
for(i in 1:dim(sem)[1]){
for(j in 1:dim(sem)[2]){
media_post[i,j]=(sem[i,j]+prior[j])/(sum(sem[i,])+sum(prior))
}
}
return("Result"= media_post)
}

prior=rep(0.5,6) #priori Jeffreys
prior=rep(1,6) #priori uniforme
(t(sem,prior))
round(t(sem,prior),4)
#####
#Cáculo do Fator de Bayes

g=function(sem_l,prior){
t1=sum(prior)
t2=sum(sem_l)
t3=prior+sem_l
resp=lgamma(t1)-lgamma(t1+t2)+sum(lgamma(t3))-sum(lgamma(prior))
}

f=function(sem,prior){
S = dim(sem)[1]
C = combn(1:S,2)
Comp = t(C)
FB = vector()
k = 0
for (i in 1:(S-1)){
for (j in i:(S-1)) {
k = k+1
freq = apply(sem[c(i,j+1),],2,sum)
FB[k] = 2*(g(sem[i,],prior)+g(sem[j+1,],prior)-g(freq,prior))
}
}
Resultado = cbind(Comp,FB)
return("Result"= Resultado)
}

```



```

}

prior=rep(0.5,6) #priori Jeffreys
prior=rep(1,6) #priori uniforme
fator_bayes = f(sem,prior)
round(fator_bayes,4)
#####
#Cálculo da divergência de Kullback-Leibler
f=function(Sem){
  S = dim(Sem)[1] # número de sementeiras
  KL = matrix(numeric(0),S,S)
  for(i in 1:S){
    for(j in 1:S){
      Prob1=Sem[i,]/sum(Sem[i,]) # Prob para cada categoria da sementeira i
      Prob2=Sem[j,]/sum(Sem[j,]) # Prob para cada categoria da sementeira j

      KL[i,j]=sum(Prob1*log2(Prob1/Prob2)) #divergência de Kullback-Leibler
    }
  }

  return(list("MD"= KL))#Retornar as matrizes de distancias
}

#Análise de agrupamento para as sementeiras sem zeros amostrais
Sem=sem[c(3,6,8,9,11,13,14),]
Dist=f(Sem); Dist #matriz de distâncias

d = Dist$MD+t(Dist$MD) #matriz simétrica
m = as.dist(d, diag = FALSE, upper = FALSE)
round(m,2)
#Método da distância média
hc = hclust(m, method="average")
k = 1.25
corte <- mean(hc$height) + k*sd(hc$height); corte

## Esboçando o gráfico com a determinação do ponto de corte e exibindo
plot(hc,labels = c("3","6","8","9","11","13","14"),hang = -1,cex = 0.8, ylab = "Distâncias",
xlab = "Sementeiras", main="Dendrograma", sub="Método do vizinho mais próximo")
abline(h = corte, v = NULL, col = 4, lty = 2)
G <- rect.hclust(hc, h =corte, which = c(1:2), border = 1:6); G

## Calculando o coeficiente de correlação cofenética
cof<- cophenetic(hc)
cor(m, cof)

#Método do vizinho mais longe
hc1 = hclust(m, method="complete")
k = 1.25
corte <- mean(hc1$height) + k*sd(hc1$height); corte

## Esboçando o gráfico com a determinação do ponto de corte e exibindo
plot(hc1,labels = c("3","6","8","9","11","13","14"),hang = -1,cex = 0.8, ylab = "Distâncias",
xlab = "Sementeiras", main="Dendrograma", sub="Método do vizinho mais distante")
abline(h = corte, v = NULL, col = 4, lty = 2)
G <- rect.hclust(hc1, h =corte, which = c(1:2), border = 1:6); G

## Calculando o coeficiente de correlação cofenética

```

```

cof<- cophenetic(hc1)
cor(m, cof)

#Método do vizinho mais proximo
hc2 = hclust(m, method="single")
k = 1.25
corte <- mean(hc2$height) + k*sd(hc2$height); corte

## Esboçando o gráfico com a determinação do ponto de corte e exibindo
plot(hc2,labels = c("3","6","8","9","11","13","14"),hang = -1,cex = 0.8, ylab = "Distâncias",
xlab = "Semeadoras",main="Dendrograma", sub="Método do vizinho mais próximo")
abline(h = corte, v = NULL, col = 4, lty = 2)
G <- rect.hclust(hc2, h =corte, which = c(1:2), border = 1:6); G

## Calculando o coeficiente de correlação cofenética
cof<- cophenetic(hc2)
cor(m, cof)
#####
#Análise de agrupamento para todas as sementeiras
Sem = read.table("dados2.txt", header = T) #semeadoras com caselas agrupadas
Sem= sem+1 #adicionando a constante 1
Sem= sem+0.5 #adicionando a constante 0.5
Dist=f(Sem); Dist # mostra as matrizes obtidas

# Análise de Agrupamentos usando métodos hierárquicos
d = Dist$MD+t(Dist$MD)
m = as.dist(d, diag = FALSE, upper = FALSE)
round(m,2)

#Método da distância média
hc = hclust(m, method="average")
k = 1.25
corte <- mean(hc$height) + k*sd(hc$height); corte

## Esboçando o gráfico com a determinação do ponto de corte
plot(hc,labels = c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15"),
hang = -1, cex = 0.8, ylab= "Distâncias", xlab="Semeadoras", main="Dendrograma",
sub="Método da distância média")
abline(h = corte, v = NULL, col = 4, lty = 2)
G <- rect.hclust(hc, h =corte, which = c(1:2), border = 1:6); G

## Calculando o coeficiente de correlação cofenética
cof<- cophenetic(hc)
cor(m, cof)

#Método do vizinho mais longe
hc1 = hclust(m, method="complete")
k = 1.25
corte <- mean(hc1$height) + k*sd(hc1$height); corte

## Esboçando o gráfico com a determinação do ponto de corte
plot(hc1,labels = c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15"),
hang = -1, cex = 0.8, ylab = "Distâncias",xlab="Semeadoras", main="Dendrograma",
sub="Método do vizinho mais distante")
abline(h = corte, v = NULL, col = 4, lty = 2)
G <- rect.hclust(hc1, h =corte, which = c(1:3), border = 1:6); G

## Calculando o coeficiente de correlação cofenética

```

```
cof<- cophenetic(hc1)
cor(m, cof)

#Método do vizinho mais proximo
hc2 = hclust(m, method="single")
k = 1.25
corte <- mean(hc2$height) + k*sd(hc2$height); corte

## Esboçando o gráfico com a determinação do ponto de corte
plot(hc2,labels = c("1","2","3","4","5","6","7","8","9","10","11","12","13","14","15"),
hang = -1, cex = 0.8, ylab = "Distâncias",xlab="Semeadoras", main="Dendrograma",
sub="Método do vizinho mais próximo")
abline(h = corte, v = NULL, col = 4, lty = 2)
G <- rect.hclust(hc2, h =corte, which = c(1:2), border = 1:6); G

## Calculando o coeficiente de correlação cofenética
cof<- cophenetic(hc2)
cor(m, cof)
```