CLASSE DE DISTRIBUIÇÕES SÉRIE DE POTÊNCIAS INFLACIONADAS COM APLICAÇÕES

Deise Deolindo Silva

CLASSE DE DISTRIBUIÇÕES SÉRIE DE POTÊNCIAS INFLACIONADAS COM APLICAÇÕES

Deise Deolindo Silva

Orientador: Prof. Dr. Josemar Rodrigues

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos - DEs/UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

UFSCar - São Carlos

Junho/2009

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S586cd

Silva, Deise Deolindo.
Classe de distribuições série de potências inflacionadas com aplicações / Deise Deolindo Silva. -- São Carlos : UFSCar, 2009.
70 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2009.

1. Estatística matemática. 2. Distribuição (Probabilidades). 3. Séries de potências. 4. Distribuição de Poisson. 5. Distribuição binomial negativa. I. Título.

CDD: 519.5 (20ª)

UNIVERSIDADE FEDERAL DE SÃO CARLOS
Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Estatística
Via Washington Luís, Km 235 - C.P.676 - CGC 45358058/0001-40
FONE: (016) 260-8292/260-8241 - FAX: (016) 260-8243
13565-905 - SÃO CARLOS-SP-BRASIL

ATA DO EXAME DE DISSERTAÇÃO DE MESTRADO DA CANDIDATA:
Deise Deolindo Silva

Aos seis dias do mês de abril do ano de dois mil e nove, às quatorze horas, na Sala de Reuniões do Departamento de Estatística, reuniu-se a Comissão Examinadora nas formas e termos do Artigo 25° do Regimento Interno do Programa de Pós-Graduação em Estatística da UFSCar, composta pelos membros: Prof. Dr. Josemar Rodrigues (DEs-UFSCar, Orientador), Profa. Dra. Vera Lucia Damasceno Tomazella (DEs-UFSCar) e Prof. Dr. Luis Gustavo Esteves (IME-USP), para Exame de Dissertação de Mestrado da candidata Deise Deolindo Silva, sob o título "Classe de distribuições série de potências inflacionadas com aplicações". A sessão foi aberta pelo Prof. Dr. Josemar Rodrigues (Presidente), iniciando-se pela apresentação da dissertação. Em seguida, foi feita a argüição da candidata pelos membros da Comissão Examinadora. A Comissão Examinadora considerou o tema relevante para Estatística e julgou a exposição feita pela candidata clara e objetiva. A candidata respondeu satisfatoriamente as questões formuladas. Pelo apresentado acima, a comissão atribuiu as seguintes avaliações: Prof. Dr. Josemar Rodrigues, nível A; Profa. Dra. Vera Lucia Damasceno Tomazella, nível A; e Prof. Dr. Luis Gustavo Esteves, nível A. De acordo com o parágrafo 5° do Artigo 25°, a candidata foi considerada **aprovada**. Encerrada a sessão secreta, o Presidente informou o resultado da defesa. Nada mais havendo a tratar, eu, Maria Isabel Rinaldo Pessôa de Araujo, Secretária deste Programa, lavrei a presente ata, que assino juntamente com os membros da Banca Examinadora.

Prof. Dr. Josemar Rodrigues

Profa. Dra. Vera Lucia Damasceno Tomazella

Maria Isabel R. P. Araujo

Prof. Dr. Luis Gustavo Esteves

Dedico este trabalho

a Deus autor de todo o conhecimento

e aos meus pais

Luiz Deolindo Apolinário

e Josefa da Silva Deolindo.

Eis o princípio da sabedoria: adiquire a sabedoria.

Adquire a inteligência em troca de tudo o que possuis.

Provérbios 4,7.

AGRADECIMENTOS

Agradeço,

primeiramente a Deus por dedicar a mim um amor supremo, por ter me concedido o dom da vida, do amor, da sabedoria e do discernimento;

a Jesus por ser meu melhor amigo, ouvir minhas orações nas horas de aflição e meus agradecimentos nas horas de alegria;

ao Espírito Santo por me reavivar sempre nas horas de desânimo e, nesses momentos, me conceder o dom da fortaleza e por me guiar e proteger, semanalmente, nas rodovias;

à Maria mãe de Deus por ser minha mãe espiritual, ser meu espelho de mulher e, com sua história de vida, me ensinou a ouvir e a silenciar nos momentos mais difíceis;

ao meu pai - Luiz Deolindo Apolinário - pelo seu amor, por sua sabedoria em aconselhar-me, pela dedicação, apoio, incentivo e, principalmente, por ser um exemplo de ser humano e pai;

à minha mãe - Josefa da Silva Deolindo - pelo amor incondicional, pelas palavras de incentivo, pelo carinho, por não me deixar desanimar diante dos obstáculos e por toda a educação que concedeu a mim;

à minha irmã Débora por ser minha incentivadora, minha melhor amiga, ser minha defensora e por contagiar todos com sua alegria;

ao meu irmão Jorge Luiz pela paciência em me ajudar a demonstrar as fórmulas matemáticas, por ser um irmão muito especial, por alegrar todos em sua volta e pelo amor que nos une;

em especial os meus tios: Luis Carlos e Marlene, Inês e José, Aparecida e Ednaldo, Lucieide e Fernando, Maria, Dulce e Paulo;

de forma muito especial meus primos, pela infância feliz que tivemos, por sermos cúmplices em muitos acontecimentos de nossas vidas e por incentivar meu crescimento pessoal e profissional: Márcia Rosana, Álvaro, Ricardo, Márcia Deolindo,

José Antonio, Thais, Bruna, Milena, Marieli, Vanessa, Milena Spada, Carlos, Zilda, Ilma, Patrícia, Gustavo, Lara, Junior, Rodrigo, Ana Paula, Gabrielly, Gisele, Jacqueline, Rafael, Leandro, Viviane e Luis Gustavo;

aos meus amigos: Josana, Luis Carlos, Ana Paula, Luana, Cleomar, Lorena, Lucilene e Amanda pelo apoio e incentivo;

aos amigos de curso: Camila Bertini, Camila Pedrozo, Caroline, Yoshi, Priscila, Guilherme, Rubiane, Leandro e Janaina pelo companheirismo e amizade;

às amigas de república: Camila por ser minha companheira de viagem, por sua amizade, carinho e pelas nossas risadas e choros. Janaina por sua amizade, animação, dedicação, pelos momentos felizes que passamos juntas. Joice pela dedicação, companheirismo e amizade. Alessandra pelo incentivo, apoio, dedicação e amizade;

aos professores da pós-graduação em Estatística da UFSCAR: Vera, Carlos Diniz, Milan, Adriano, Aparecida, Silvia;

de forma muito especial ao prof. Dr. Josemar Rodrigues por sua orientação, paciência, por entender meus problemas de horário e, principalmente, pelas ideias durante todo este trabalho;

aos funcionários do Departamento de Estatística da UFSCAR pela atenção dispensada;

à Faculdade de Birigui - UNIESP e à Faculdade de Saúde de São Paulo - FASSP pelo incentivo ao meu crescimento profissional;

aos colegas de trabalho: Eldir, Mônica, Rute, Marlene, Leonardo, Ronaldo, Paula, Renata, Juliana, Beth, Wladimir, Sandra, Ulisses;

aos meus alunos pelo carinho, incentivo e, por apesar de ensinar, aprender muito com eles;

enfim, a todos que rogaram a Deus pedindo o meu sucesso durante a trajetória desse mestrado.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Este trabalho tem como tema central a classe de distribuições série de potências inflacionadas, em que o intuito é estudar suas principais propriedades e a aplicabilidade no contexto bayesiano. Esta classe de modelos engloba as distribuições de Poisson, binomial e binomial negativa simples e as generalizadas e, por isso é muito aplicada na modelagem de dados discretos com valores excessivos. Como caso particular propôs-se explorar *a distribuição de Poisson zero inflacionada* (ZIP), em que o objetivo principal foi verificar a eficácia de sua modelagem quando comparada à distribuição de Poisson. A mesma metodologia foi considerada para a distribuição binomial negativa inflacionada, mas comparando-a com as distribuições de Poisson, binomial negativa e ZIP. Como critérios formais para seleção de modelos foram considerados o fator de Bayes e o teste de significância completamente bayesiano.

Palavras-chave: Classe de Distribuição Série de Potências Inflacionadas, Distribuição Poisson Zero Inflacionada, Distribuição Binomial Negativa Zero Inflacionada, Seleção de Modelos.

Abstract

This work has as central theme the Inflated Modified Power Series Distributions, where the objective is to study its main properties and the applicability in the bayesian context. This class of models includes the generalized Poisson, binomial and negative binomial distributions. These probability distributions are very helpful to models discrete data with inflated values. As particular case the - *zero inflated Poisson models* (ZIP) is studied, where the main purpose was to verify the effectiveness of it when compared to the Poisson distribution. The same methodology was considered for the negative binomial inflated distribution, but comparing it with the Poisson, negative binomial and ZIP distributions. The Bayes factor and full bayesian significance test were considered for selecting models.

keyword: Inflated Modified Power Series Distributions, Zero Inflated Poisson Models, Zero Inflated Negative Binomial Models, Selection of Models.

Capítulo 1

Introdução

O bom ajuste dos dados depende dos modelos probabilísticos atribuídos a eles. Nas aplicações envolvendo dados reais sobre contagens, geralmente atribui-se modelos discretos que são amplamente desenvolvidos na literatura. No entanto, é comum encontrar uma grande quantidade zeros nos conjuntos de dados. Esses zeros excessivos dificultam a elaboração de uma análise estatística precisa para o problema, pois os modelos usuais desenvolvidos para a modelagem de dados discretos não ajustam bem tal situação.

Diante disso, é relevante pesquisar como se formam esses zeros nos conjuntos de dados. Martin et al. (2005), ressaltaram que o valor zero pode acontecer de quatro maneiras diferentes. Duas delas, podem ser definidas como zeros *verdadeiros* e duas como *aleatórios ou falsos*. Os zeros verdadeiros podem surgir da baixa freqüência de ocorrência do evento. Por exemplo, se o interesse for estudar a incidência de determinada doença em um local específico, neste caso, o excesso de zeros significa que a doença recaiu a poucos indivíduos do local.

Outra situação considerada como zero verdadeiro é quando realmente o local não havia nenhum indivíduo presente. Podemos citar as aplicações que envolvem controle de qualidade que utilizam processo de fabricação moderno e, por isso, esperam zero defeito (estado perfeito). Os zeros aleatórios ou falsos podem ser resultado de erros de amostragem ou um vício visual, ou seja, o indivíduo existe, ocupa o local, mas não estava presente durante a realização

da pesquisa ou o elemento ocupa o local, está presente, mas o pesquisador não o encontra. Esse tipo de zero ocorre geralmente em estudos ecológicos, principalmente de vida selvagem ou aquática. Por exemplo, a ausência de animais em uma determinada área pode ter sido resultado de um erro humano ou de um erro visual.

Os zeros ocorridos em um conjunto de dados podem ter sido resultado de um zero verdadeiro, de um erro humano, ou ser um zero de amostragem. Infelizmente, a distinção desses tipos de zeros é, na maioria das situações, uma tarefa impossível de ser realizada. Existe ainda, uma situação onde há a incerteza sobre a sua origem nas observações, neste caso, um procedimento usual é utilizar distribuições truncadas.

A produção dos zeros excessivos nas amostras podem ser classificadas de duas formas. Na primeira, o excesso de zeros é resultado da superdispersão, ou seja, a variância dos dados é maior que a assumida pelo modelo (Paula, 2004). Na segunda forma o excesso de zeros é formado por subpopulações distintas que podem estar relacionadas à alguma intervenção natural ou truncamento nos dados. Saito (2005) mencionou um estudo econômico realizado por Aitchison e Brown (1957) sobre despesas do lar. Em certos lares, alguns itens considerados supérfluos não fazem parte da compra. Logo, os dados apresentam uma grande quantidade de observações zero. Neste caso, a amostra pode ser separada em dois grupos, os lares que gastam com os itens e os lares que não gastam (este grupo formam uma subpopulação que produz somente observações zero).

Nas aplicações ecológicas em que o objetivo é modelar a distribuição das plantas e/ou organismos é sugerido como adequado o modelo de Poisson, mas geralmente os descendentes das plantas e/ou organismos tendem a se concentrarem próximos dos pais que em outros lugares, formando aglomerações. Assim a variância do número de plantas e/ou organismos é maior que a esperada pelo modelo de Poisson, causando superdispersão. Existem muitas técnicas para modelar esse tipo de dados baseando-se numa única distribuição (Martin et al., 2005).

Uma metodologia eficaz na modelagem dados de contagem com zeros

excessivos são as misturas de modelos, através das distribuições zero inflacionadas.

Como as distribuições zero inflacionadas requerem a definição de uma distribuição discreta, foi objeto de estudo deste trabalho a classe de distribuições série de potências. Esta classe engloba tanto as distribuições de Poisson, binomial e binomial negativa simples como as generalizadas e podem ser utilizadas em uma variedade de aplicações e seus resultados mostram eficiência. Esta classe de distribuições também pode ser estendida para os modelos inflacionados, a qual é denominada classe de distribuições série de potências inflacionadas.

Como mencionado, para o desenvolvimento da classe de distribuições série de potências inflacionadas se faz necessário ter conhecimentos sobre misturas de modelos, pois esta abordagem considera uma distribuição degenerada no ponto zero e uma distribuição discreta que se adequaria aos dados caso não existisse zeros excessivos. No entanto, em alguns conjuntos de dados o valor em excesso não é zero. Então, esta classe de distribuições série de potências inflacionadas pode ser ampliada para solucionar qualquer valor excessivo $y = s$. Neste trabalho a classe de distribuições série de potências inflacionadas foi objeto de estudo e, para estas situações ela é uma boa alternativa, conseguindo modelar bem os dados.

Rodrigues (2003) apresentou a abordagem bayesiana para distribuições zero inflacionadas utilizando um procedimento baseado em dados ampliados. O objetivo, neste caso, foi tornar a distribuição a *posteriori* $\pi(\theta, \omega|D)$ conhecida facilitando o tratamento computacional.

1.1 Objetivos

- Estudar a origem de zeros excessivos em conjuntos de dados e modelá-los através da classe de distribuições série de potências inflacionadas.

- Estudar a classe de distribuições série de potências inflacionadas, suas principais propriedades, os momentos e as relações com a classe de distribuições série de potências.

- Considerar a abordagem bayesiana para a classe de distribuições série de potências inflacinadas. Utilizar misturas de modelos para explicar a quantidade de zeros excessivos. Além disso, considerar o procedimento de dados ampliados para esta classe de distribuições, obtendo uma forma geral.

- Para verficar a eficácia deste procedimento foi proposto estudar aplicações envolvendo casos particulares da classe de distribuições série de potências inflacionadas.

- Uma delas foi a *distribuição de Poisson zero inflacionada* (ZIP), em que o objetivo principal foi verificar a eficácia de sua modelagem quando comparada à distribuição de Poisson. E a *distribuição binomial negativa zero inflacionada* (ZIBN) comparando-a com outras distribuições.

- Como critério formal para seleção de modelos utilizou-se o fator de Bayes. Este é uma medida de evidência usual neste contexto e problemático em algumas situações, principalmente quando atribui-se distribuições a *priori* vagas para alguns parâmetros e, além disso, considera uma escala arbitrária para a interpretação de seus valores.

- Como alternativa ao fator de Bayes utilizou-se o teste de signifância completamente bayesiano proposto por Pereira e Stern (2008). Um dos benefícios em utilizá-lo é porque não apresenta problemas quando são atribuídas distribuições a *priori* impróprias, como acontece no fator de Bayes. Exige somente o conhecimento da distribuição a posteriori para calcular a medida de evidência *Ev*, sem qualquer complicação relacionada com a dimensionalidade do parâmetro nem com a do espaço amostral, esta característica evita a necessidade de eliminar o parâmetro *nuisance*. Além disso, as necessidades computacionais de *Ev* não se baseiam em nenhum método assintótico, sua obtenção está baseada na otimização e na integração numérica.

- Obter as distribuições a *posteriori* dos parâmetros modelo de Poisson zero inflacionado e de Poisson utilizando o algoritimo Gibbs através do *Software* R.

- Para calcular o fator de Bayes e o teste de significância completamente bayesiano propôs-se utilizar o *Software* R e o *Winbugs*, respectivamente.

Este trabalho está organizado da seguinte forma.

No capítulo 2 é apresentado um resumo sobre as principais fontes dos zeros excessivos em contagens, como se dá a formação dos zeros inflacionados e exemplos de aplicações que envolvem valores excessivos diferentes de zero. No capítulo 3 é apresentada a classe de distribuições série de potências inflacionadas, que necessita da descrição da classe de distribuições série de potências e de algumas propriedades. Além disso, é explorada a aplicabilidade no cenário bayesiano. No Capítulo 4 são descritos os modelos para dados inflacionados de zeros, em especial, é explorado dois casos particulares: a *distribuição de Poisson zero inflacionada* e a *distribuição binomial negativa zero inflacionada*. No Capitulo 5 apresentam-se aplicações envolvendo estes modelos. Em seguida estão as considerações finais.

Capítulo 2

Número Excessivo de Valores em Contagens

Em muitas aplicações envolvendo dados reais sobre contagens são atribuídos os modelos discretos que são largamente desenvolvidos na literatura. Podemos citar as distribuições de Poisson, binomial e binomial negativa.

Geralmente, os conjuntos de dados contêm um número excessivo de zeros que não são descritos pelo modelo assumido. Esses zeros podem ser classificados de quatro modos, dois podem ser definidos como zeros *verdadeiros* e dois como *falsos* (aleatórios).

No primeiro caso, os zeros verdadeiros surgem de uma baixa freqüência de ocorrência ou realmente o local não havia nenhum indivíduo presente. No segundo caso, o indivíduo existe, ocupa o local, mas não estava presente durante a pesquisa ou o elemento ocupa o local, está presente, mas o pesquisador não o encontra. Uma possível solução, para explicar zeros verdadeiros ou falsos, é utilizar as distribuições zero inflacionadas. Estas distribuições exigem conhecimentos sobre misturas de modelos, que neste caso, considera uma distribuição degenerada no ponto zero e uma distribuição que se adequaria aos dados caso não existisse zeros excessivos.

Quando os zeros inflacionados são resultados de excesso de zeros verdadeiros e falsos, não há nenhuma discussão formal na literatura de como modelar

tais conjuntos de dados, justamente porque é difícil distinguir a origem desses zeros. Quando há a incerteza sobre a sua origem nas observações, um procedimento usual, é utilizar distribuições truncadas.

Em alguns conjuntos de dados o valor inflacionado não é o valor zero, mas esta metodologia pode ser aplicada para solucionar esse número excessivo de valores diferentes de zero. Além disso, a classe de distribuições série de potências inflacionadas pode ser aplicada em aplicações de diversas áreas, por exemplo, ecologia, atuária, controle de qualidade dentre outras.

Este capítulo está organizado como segue. Na seção 1 esta descrito sobre o número excessivo de zeros em dados, sobre a possível fonte dos zeros inflacionados, superdispersão e subpopulações distintas. Na seção 2 é apresentado um resumo sobre alguns estudos que envolveram um número excessivo de valores $y = s$ diferentes de zero.

2.1 Número Excessivo de Zeros

Em muitos estudos envolvendo a análise de dados discretos é comum existir uma grande quantidade de zero nos dados. Esse excesso dificulta a elaboração de uma análise estatística para o problema, pois os modelos usuais não conseguem modelar a presença excessiva de zeros. Por exemplo, ao considerarmos uma linha de produção na qual se aplica controle de qualidade, a contagem de defeitos de um produto apresenta-se cada vez menor, ou seja, há um grande número de zeros, isto devido à modernização dos processos de fabricação. Neste caso, esses zeros correspondem a zeros determinísticos.

Nas aplicações ecológicas os zeros podem ocorrer devido à espécie ser totalmente ausente na área amostrada ou quando a espécie está presente mas não foi observada pelo pesquisador, neste caso, os zeros são aleatórios (falsos). Nessas aplicações o problema de zeros aleatórios ocorre frequentemente devido a erros humanos ou vícios no método de amostragem. (Martin et al., 2005). Portanto, podemos dizer que em uma contagem os zeros podem ser obtidos por erro humano, por amostragem ou ser resultado de um zero verdadeiro. Na grande maioria das

situações práticas, infelizmente, é impossível separar os zeros de amostragem do zero verdadeiro.

2.1.1 Fonte dos Zeros Inflacionados

Segundo Martin et al. (2005), o valor zero acontece de quatro modos. Dois podem ser definidos como verdadeiros zeros e dois como falsos zeros. O primeiro tipo de zero verdadeiro *surge de uma baixa freqüência de ocorrência* ou *ser o resultado de um efeito que conduz para locais que não têm nenhum indivíduo presente*.

Na área ecológica uma espécie pode estar ausente por causa de processos demográficos, competição, ou qualidade de hábitat pobre. Estes zeros são verdadeiros, pois são resultados dos reais efeitos ecológicos. Secundariamente, um zero simplesmente pode acontecer por acaso, porque a espécie não satura seu hábitat inteiro (por exemplo, por causa de extinções locais), este é um exemplo para a segunda situação dos zeros verdadeiros.

O primeiro tipo de falso zero é causado por: *apesar de existir, ocupar o local, não estava presente na hora da pesquisa*. A segunda maneira: *acontece quando ocupa o local e está presente na hora de amostrar, mas o pesquisador não o encontra*. Por exemplo, no caso onde uma espécie pode estar temporariamente ausente do local de estudo, se o objetivo é quantificar onde a espécie está de forma rápida, sua ausência não constituíria um falso zero (isto é, a espécie realmente não estava lá quando inspecionado). Porém, se o interesse fosse nas áreas que estavam sendo utilizadas pela espécie considerando um grande período de tempo, então sua ausência constituíria um falso zero.

Apesar das categorias definidas acima, um número grande de zeros podem surgir nos dados de outro modo, quando observações são obtidas fora do alcance do elemento pesquisado.

A Tabela 2.1 apresenta um resumo sobre os tipos de zeros e sua respectiva definição.

Fonte: Martin et al. (2005).

TABELA 2.1: Tipos de zeros em dados

Tipo de zero	**Definição**
Zero Verdadeiro	• Baixa freqüência de ocorrência; • Realmente não existe nenhum indivíduo presente.
Zero Aleatório	• Apesar de existir, ocupar o local, não estava presente na hora da pesquisa; • Acontece quando ocupa o local e está presente na hora de amostrar, mas o pesquisador não o encontra.

Em um conjunto de dados podem ocorrer diferentes tipos de zeros: as informações podem não conter nenhum zero inflacionado, ou ter zeros inflacionados devido a zeros verdadeiros, devido a aleatórios; ou ainda, devido a excesso de ambos os casos, verdadeiros e aleatórios. Para finalizar, existe a incerteza relativa sobre a fonte dos zeros inflacionados.

Na ausência de zeros inflacionados as distribuições usuais podem modelar os dados, essas distribuições são amplamente desenvolvidas no cenário estatístico, podemos citar como exemplo: a binomial, a binomial negativa ou Poisson.

Quando esses zeros são verdadeiros é recomendada a utilização de misturas de modelos. É considerada uma média ponderada de duas distribuições, ou seja, atribui uma distribuição degenerada no ponto zero e uma distribuição discreta adequada aos outros dados. Se os falsos zeros estão presentes nos dados, a modelagem também pode ser feita através de misturas de modelos. Em muitas aplicações as distribuições zero inflacionadas são soluções possíveis. Na literatura não houve nenhuma discussão formal de como modelar conjuntos de dados que contenham excesso de zeros verdadeiros e falsos, justamente porque e difícil distingui-los. Neste caso, misturas de modelos também são recomendadas para a modelagem.

Em alguns casos não é possível determinar a origem de zero nas observações. Um procedimento usual é utilizar uma distribuição truncada, ou seja, são eliminados zeros da contagem e só as ocorrências são modeladas. Por exemplo, Baum e Myers (2004) *apud* Martin et al. (2005) em seus estudos observaram que

era impossível determinar se a ausência de capturas de tubarões era resultado de zeros verdadeiros (não existia nenhum tubarão) ou o fracasso do pescador em capturar os tubarões. Eles lidaram com esta incerteza usando um modelo binomial negativo truncado.

A Tabela 2.2 apresenta alguns cenários sobre a origem dos zeros nos conjuntos de dados e a modelagem recomendada para solucionar o excesso de zeros nas contagens.

TABELA 2.2: Cenários sobre a origem de zero nos dados e a modelagem recomendada

Zero Inflacionado	**Modelagem Apropriada**
Sem inflação de zeros	Modelos Discretos Simples
Zeros Verdadeiros	Mistura modelos, distribuições zero inflacionadas (ZIBN, ZIB ou ZIP)
Zeros Aleatórios	Mistura modelos, distribuições zero inflacionadas (ZIBN, ZIB ou ZIP)
Combinação de Ambos	Mistura de duas ou mais distribuições
Incerteza da Origem	Distribuições Truncadas

Fonte: Martin et al. (2005).

2.1.2 Origem de Zeros nos Dados

Como já ressaltado, uma observação zero pode ser resultado de um zero verdadeiro ou de um zero aleatório. Sob a ótica estatística a produção desse excesso de zeros pode ser classificada de duas formas: *superdispersão ou subpopulações distintas.*

Conforme Hinde e Demétrio (1998) *apud* Paula (2004) superdispersão é um fenômeno comum que ocorre na modelagem de dados discretos e cuja ocorrência é caracterizada quando a variância observada excede aquela assumida pelo modelo. Segundo Echavarría (2004), dado uma suposição de distribuição para os dados temos superdispersão se a variância observada dos dados é maior que a variância suposta para modelo.

Quando trabalhamos com dados de contagem, geralmente atribuímos o modelo de Poisson, sabe-se que para esta distribuição a média e a variância são iguais. Na prática é encontrada aplicações que não obedecem esta restrição sobre a igualdade entre a variância e a média. Logo, os conjuntos de dados observados tendem a ser superdispersos, ou seja, significa que a variância nos dados excede a variância assumida pela distribuição de Poisson, isto é, a superdispersão ocorre quando é esperada tal distribuição para a resposta, porém a variância é maior do que a resposta média.

A superdispersão pode ser causada por heterogeneidade ou por excesso de zeros na contagem. No primeiro caso, alguma covariável pode ter sido não quantificada e, devido a isto, ser fonte da heterogeneidade apresentada nos dados e portanto, pode ser a possível responsável pela superdispersão observada, ou as unidades amostrais podem causar a heterogeneidade devido a variabilidades inter-unidades experimentais. No segundo caso, o excesso de zeros também conduz a superdispersão (Paula, 2004).

Na área ecológica, o número de organismos ou plantas de um quadrante é geralmente modelado com a distribuição de Poisson. Entretanto, muitas vezes, os descendentes de plantas ou organismos tendem a se concentrarem mais próximos de seus pais do que em outros locais. Assim a variância do número de plantas ou organismos é maior do que a esperada pela distribuição de Poisson, caracterizando superdispersão (Saito, 2005).

Neste caso, os zeros têm origem de uma única fonte de variação. Existem muitas técnicas para modelar dados baseando-se numa única distribuição. Das três distribuições discretas apresentadas a distribuição binomial negativa é quem ajusta melhor a superdispersão apresentada pelos dados, geralmente ela é muito aplicada.

A seguir é apresentado graficamente na Figura 2.1 a relação entre a esperança e a variância das distribuições discretas binomial, binomial negativa e Poisson. Estudando essas distribuições podemos identificar algumas propriedades sobre a dispersão admitidas por elas.

Os zeros podem ser produzidos por subpopulações distintas, ou seja,

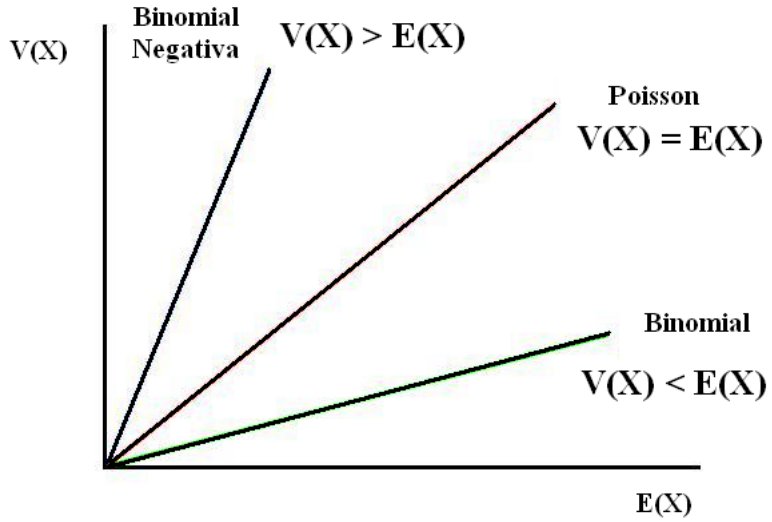

FIGURA 2.1: Relação entre a esperança e variância das distribuições discretas.

pode estar relacionado a alguma intervenção natural ou truncamento dos dados. Considerando o número de peças defeituosas encontradas na linha de produção e que a indústria possui um processo de controle de qualidade. É esperado o menor número possível de peças com defeitos, portanto, temos um agrupamento de zeros nos dados, ou em termos industriais, estamos diante de um estado perfeito. Então, podemos considerar uma subpopulação que produz somente contagens zero.

Neste caso, pequenas mudanças no meio fazem com que o processo de fabricação movimente-se aleatoriamente entre um estado perfeito e um estado imperfeito (os defeitos são possíveis, mas evitáveis), consequentemente, os zeros têm origem de duas fontes diferentes, portanto, é apropriado considerarmos uma mistura de distribuições para modelar os dados.

2.2 Número Excessivo de valores s

Existem conjuntos de dados em que os valores excessivos não são o valor zero. Para modelar esses dados, a metodologia utilizada para zeros excessivos também pode ser considerada.

Existem na literatura algumas aplicações envolvendo tais acontecimentos. Por exemplo, Saito (2005) apresentou uma aplicação sobre o número de visitas ao dentista de cidadãos suecos e observou que além do valor zero ser excessivo, o valor 1 também era. Neste estudo os zeros em excessos podem surgir porque o indivíduo pode apresentar dentes saudáveis, ter medo de dentista, não poder comparecer, ou simplesmente por livre-arbítrio. Quanto ao valor 1 ser excessivo uma explicação razoável é que em certos locais tem-se o costume de ir ao dentista como uma forma de prevenção.

Para a modelagem deste conjunto de dados foi considerado a distribuição ZOIP (Zeros and Ones Inflated Poisson), na qual o modelo formado pela média ponderada de distribuições (mistura de modelos), com esta metodologia conseguiu um bom ajuste para o conjunto de dados.

Murat e Szynal (1998) estudaram sobre o número de pétalas de flores e observaram que o valor oito possuía uma frequência relativamente maior que o esperado pelo modelo assumido, logo atribuíram a esses dados a distribuição de Poisson inflacionada neste ponto ($s = 8$).

Por este motivo este trabalho utilizou misturas de modelos para ajustar os zeros em excesso. Além disso, explorou uma classe de distribuições que abrange os modelos discretos usuais e os generalizados: *a classe de distribuições série de potências.*

Esta classe de distribuições pode ser estendida para valores inflacionados e é denominada *classe de distribuições série de potências inflacionadas.* Esta última, pode ser utilizada em todas as situações até agora mencionadas. No próximo capítulo estas classes de distribuições serão exploradas.

Capítulo 3

Classe de Distribuições Série de Potências Inflacionadas

No contexto de modelagem de dados discretos uma classe geral foi desenvolvida - *distribuições série de potências* (PSD). Esta classe engloba tanto as distribuições de Poisson, binomial e binomial negativa simples como as generalizadas e pode ser considerada para uma diversidade de aplicações obtendo bons resultados, como ressaltaram Gupta et al. (1995).

Na análise de dados discretos existem frequentemente valores inflacionados, como por exemplo, o valor zero é observado com uma frequência significativamente maior que o admitido pelo modelo assumido; consequentemente, a classe de distribuições série de potências pode ser estendida para distribuições inflacionadas e, denominamo-a de *classe de distribuições série de potências inflacionadas* (IPSD). A distribuição de Poisson inflacionada é um caso particular desta classe de distribuições.

Para exemplificação considere a seguinte aplicação sobre a produção de duas máquinas: a máquina I produz itens perfeitos e a máquina II produz defeitos de acordo com o modelo de Poisson. Ao observar a produção final sem defeitos não é possível identificar se o produto é oriundo da máquina I ou II, neste caso, o valor zero torna-se inflacionado, ou seja, é resultado de uma subpopulação que produz contagens zero. Uma modelagem adequada seria a distribuição de Poisson

zero inflacionada.

O cápitulo está organizado como segue. Na seção 1 apresenta-se a descrição da classe de distribuições série de potências e suas propriedades. Na seção 2 apresenta-se a descrição da classe de distribuições série de potências inflacionadas e algumas propriedades e para esta classe foram descritas rescpectivamente na seção 3 e 4 a abordagem bayesiana e a matriz de informação de Fisher.

3.1 A Classe de Distribuições Série de Potências

A modelagem de dados discretos pode ser feita através da classe geral das *distribuições série de potências*. Esta classe engloba modelos comumente utilizados na literatura. A forma geral da função de probabilidade é especificada a seguir:

$$P[X = x] = \frac{a(x)[g(\theta)]^x}{f(\theta)}, x = 0, 1, 2, \ldots, \tag{3.1}$$

em que $a(x) > 0$ e é independente de θ, $\theta > 0$, $f(\theta) = \sum_y a(y)[g(\theta)]^y < \infty$, $g(\theta)$ é positiva, finita, diferenciável e inversível.

Como já foi ressaltado, esta classe de modelos inclui tanto as distribuições de Poisson, binomial e binomial negativa simples como as generalizadas. Ilustrativamente é apresentado na Tabela 3.1 alguns casos especiais da *classe de distribuições série de potências*. (ver Murat e Szynal, 1998).

TABELA 3.1: Casos especiais da classe de distribuições série de potências

Modelos Ordinários	**Modelo**	$\mathbf{a}(x)$	$g(\theta)$	$\mathbf{f}(\theta)$
Poisson	$\frac{1}{x!}\theta^x e^{-\theta}$	$\frac{1}{x!}$	θ	e^{θ}
Binomial	$\binom{n}{x}\theta^x(1-\theta)^{(n-x)}$	$\binom{n}{x}$	$\theta(1-\theta)^{-1}$	$(1-\theta)^{-n}$
Binomial Negativa	$\binom{n-1}{x-1}\theta^x(1-\theta)^{(n-x)}$	$\binom{n-1}{x-1}$	$\theta(1-\theta)^{-1}$	$(1-\theta)^{-n}$
Modelos Generalizados	**Modelo**	$\mathbf{a}(x)$	$g(\theta)$	$\mathbf{f}(\theta)$
Poisson	$\frac{(1+x\alpha)^{x-1}}{x!}\left(\theta e^{-\alpha\theta}\right)^x e^{-\theta}$	$\frac{(1+x\alpha)^{x-1}}{x!}$	$\theta e^{-\alpha\theta}$	e^{θ}
Binomial Negativa	$\frac{N\Gamma(N+\beta x)}{N!\Gamma(N+\beta x-x+1)}\left(\theta(1-\theta)^{\beta-1}\right)^x(1-\theta)^N$	$\frac{n\Gamma(N+\beta x)}{N!\Gamma(N+\beta x-x+1)}$	$\theta(1-\theta)^{\beta-1}$	$(1-\theta)^{-N}$
Logarítmica	$\frac{N\Gamma(\beta x)}{x\Gamma(x)\Gamma(\beta x-x+1)}\frac{\left(\theta(1-\theta)^{\beta-1}\right)^x}{-ln(1-\theta)}$	$\frac{N\Gamma(\beta x)}{x\Gamma(x)\Gamma(\beta x-x+1)}$	$\theta(1-\theta)^{\beta-1}$	$-ln(1-\theta)$
Perda em Jogos	$\frac{\alpha}{2x-\alpha}\binom{2x-\alpha}{x}\frac{(\theta(1-\theta))^x}{\theta^{\alpha}}$	$\frac{\alpha}{2x-\alpha}\binom{2x-\alpha}{x}$	$\theta(1-\theta)$	θ^{α}
Período em Filas M/M/1	$\frac{\alpha}{2x-\alpha}\binom{2x-\alpha}{x}\left[\frac{\theta}{(1-\theta)^2}\right]^x\left(\frac{1+\theta}{\theta}\right)^{\alpha}$	$\frac{\alpha}{2x-\alpha}\binom{2x-\alpha}{x}$	$\frac{\theta}{(1-\theta)^2}$	$\left(\frac{1+\theta}{\theta}\right)^{\alpha}$

A distribuição de Poisson generalizada é adequada para modelar dados

com superdispersão e excesso de zeros. Considere a distribuição de Poisson Generalizada dada por

$$P\left(X=x\right)=\frac{\left(1+x\alpha\right)^{x-1}}{x!}\left(\theta e^{-\alpha\theta}\right)^{x}e^{-\theta}$$

Em que, $\alpha \geq 0$, e α conhecido, $\theta > 0$ e $|\alpha\theta| < 1$. Se $\alpha = 0$ o modelo se reduz ao modelo de Poisson simples. (Borgatto, 2004).

A distribuição binomial negativa generalizada também é adequada para a modelagem de dados com superdispersão e excesso de zeros. Para esta distribuição as restrições são, $0 < \theta < 1$, $|\beta\theta| < 1$, $\beta = 0$ ou $\beta \geq 1$ e β deve ser conhecido.

Se $\beta = 0$ temos que o modelo binonial negativo generalizado se reduz ao modelo binomial. Se $\beta = 1$ ele se reduz ao modelo de binomial negativo.

Na distribuição logarítmica generalizada a exigência é $0 < \theta < 1$, $|\beta\theta| < 1$ com β conhecido. Se $\beta = 1$ temos que o modelo se reduz à distribuição logarítmica de Fisher. A distribuição para perda em jogos faz a exigência de $0 < \theta < 1/2$ com $\alpha \geq 1$ e conhecido.

3.1.1 Propriedades da Classe de Distribuições Série de Potências

Esta classe de distribuições apresentam propriedades interessantes. A seguir será calculada a esperança matemática desta classe de distribuições. Para isto considere,

$$\begin{aligned}
E\left(X\right) &= \sum_{x=0}^{\infty} x\frac{a\left(x\right)\left[g\left(\theta\right)\right]^{x}}{f\left(\theta\right)} \qquad (3.2)\\
&= 0+\frac{a\left(1\right)g\left(\theta\right)}{f\left(\theta\right)}+\frac{2a\left(2\right)\left[g\left(\theta\right)\right]^{2}}{f\left(\theta\right)}+\frac{3a\left(3\right)\left[g\left(\theta\right)\right]^{3}}{f\left(\theta\right)}+\cdots\\
&= \frac{g\left(\theta\right)}{f\left(\theta\right)}\left[a\left(1\right)+2a\left(2\right)g\left(\theta\right)+3a\left(3\right)\left[g\left(\theta\right)\right]^{2}+\cdots\right]\\
&= \frac{g\left(\theta\right)}{f\left(\theta\right)}\sum_{x=1}^{\infty} xa\left(x\right)g\left(\theta\right)^{x-1}
\end{aligned}$$

Mas,

$$\begin{aligned} f(\theta) &= \sum_{x=0}^{\infty} a(x)\,[g(\theta)]^x \\ &= a(0) + a(1)\,[g(\theta)] + a(2)\,[g(\theta)]^2 + a(3)\,[g(\theta)]^3 + \cdots, \end{aligned}$$

o que implica

$$\begin{aligned} f'(\theta) &= 0 + a(1)\,g'(\theta) + 2a(2)\,[g(\theta)]^1\,g'(\theta) + 3a(3)\,[g(\theta)]^2\,g'(\theta) + \cdots \\ &= \sum_{x=1}^{\infty} xa(x)\,[g(\theta)]^{x-1}\,g'(\theta). \end{aligned}$$

Logo,

$$\sum_{x=1}^{\infty} xa(x)\,[g(\theta)]^{x-1} = \frac{f'(\theta)}{g'(\theta)}.$$

Portanto, a esperança é

$$E(X) = \frac{g(\theta)}{f(\theta)}\frac{f'(\theta)}{g'(\theta)} = \mu_1(\theta). \tag{3.3}$$

Para facilitar a obtenção da variância a forma exponencial do modelo será adotada

$$q(x,\theta) = [\exp\{xc(\theta) + d(\theta) + S(x)\}]\,I_{0,1,2,\ldots}(x). \tag{3.4}$$

Logo, o modelo descrito em (3.1) pode ser reescrito como

$$q(x,\theta) = \exp\{x\log(g(\theta)) + \log(a(x)) - \log(f(\theta))\}, \tag{3.5}$$

em que, $c(\theta) = \log(g(\theta))$, $d(\theta) = -\log(f(\theta))$ e $S(x) = \log(a(x))$.

Reparametrizando a expressão dada em (3.5) com a introdução do parâmetro na forma natural, $\eta = \log(g(\theta))$, temos

$$q(x,\eta) = \exp\{x\eta + \log(a(x)) - \log d_0(\eta)\}, \tag{3.6}$$

sendo que, $d_0(\eta) = f(g^{-1}(e^{\eta}))$.

Quando um modelo pode ser escrito na forma exponencial a esperança e a variância podem ser obtidas através de

$$E(X) = -d_0'(\eta) \text{ e } V(X) = -d_0''(\eta).$$

Então,

$$E(X) = -d_0'(\eta) = \frac{1}{f(\theta)} \frac{df(\theta)}{d\theta} \frac{d\theta}{d\eta} = \frac{f'(\theta)}{f(\theta)} \frac{g(\theta)}{g'(\theta)} = \mu_1(\theta),$$

pois,

$$\eta = \log(g(\theta)) \Rightarrow \eta' = \frac{1}{g(\theta)} g'(\theta) \Rightarrow (\eta')^{-1} = \frac{g(\theta)}{g'(\theta)}.$$

A variância será

$$V(X) = -d_0''(\eta) = \frac{d\mu_1(\theta)}{d\theta} \frac{d\theta}{d\eta} = \frac{g(\theta)}{g'(\theta)} \frac{d\mu_1(\theta)}{d\theta}. \tag{3.7}$$

Da expressão (3.7) o momento de segunda ordem pode ser obtido por

$$E\left(X^2\right) = V(X) + (E(X))^2 = \frac{g(\theta)}{g'(\theta)} \frac{d\mu_1(\theta)}{d(\theta)} + \mu_1^2(\theta) = \mu_2(\theta). \tag{3.8}$$

Na Tabela 3.2 é apresentada alguns casos especias da *classe de distribuições série de potências* e sua respectiva média.

TABELA 3.2: Média das distribuições série de potências

Modelos Ordinários	**Modelo**	$\mu_1(\theta) = \frac{g(\theta)}{f(\theta)} \frac{f'(\theta)}{g'(\theta)}$
Poisson	$\frac{1}{x!}\theta^x e^{-\theta}$	θ
Binomial	$\binom{n}{x} \theta^x (1-\theta)^{(n-x)}$	$n\theta$
Binomial Negativa	$\binom{n-1}{x-1} \theta^x (1-\theta)^{(n-x)}$	$\frac{n\theta}{1-\theta}$
Modelos Generalizados	**Modelo**	$E(X) = \frac{g(\theta)}{f(\theta)} \frac{f'(\theta)}{g'(\theta)}$
Poisson	$\frac{(1+x\alpha)^{x-1}}{x!} \left(\theta e^{-\alpha\theta}\right)^x e^{-\theta}$	$\frac{\theta}{1-\alpha\theta}$
Binomial Negativa	$\frac{N\Gamma(N+\beta x)}{N!\Gamma(N+\beta x-x+1)} \left(\theta(1-\theta)^{\beta-1}\right)^x (1-\theta)^N$	$\frac{N\theta}{1-\beta\theta}$
Logarítmica	$\frac{N\Gamma(\beta x)}{x\Gamma(x)\Gamma(\beta x-x+1)} \frac{\left(\theta(1-\theta)^{\beta-1}\right)^x}{-ln(1-\theta)}$	$\frac{\theta}{(\beta\theta-1)ln(1-\theta)}$
Perda em Jogos	$\frac{\alpha}{2x-\alpha} \binom{2x-\alpha}{x} \frac{(\theta(1-\theta))^x}{\theta^\alpha}$	$\frac{(1-\theta)\alpha}{(1-2\theta)}$
Período em Filas M/M/1	$\frac{\alpha}{2x-\alpha} \binom{2x-\alpha}{x} \left[\frac{\theta}{(1-\theta)^2}\right]^x \left(\frac{1+\theta}{\theta}\right)^\alpha$	$\frac{\alpha}{\theta-1}$

A esperança matemática desta classe de distribuições é útil para determinar a esperança da classe de distribuições série de potências inflacionadas.

Pois, como o modelo desta classe de distribuições é formado por misturas de distribuições, para os valores não inflacionados, a esperança continua sendo igual a da classe de distribuições série de potências, ponderada por um peso.

A seguir será descrita a classe de distribuições série de potências inflacionadas que é considerada uma generalização desta classe de distribuições e as propriedades descritas até o momento também se fazem necessárias.

3.2 A Classe de Distribuições Série de Potências Inflacionadas

Na análise de dados discretos inflacionados é usual considerar misturas de distribuições. Murat e Szynal (1998) descreveram a distribuição série de potências para dados inflacionados utilizando esta metodologia. O modelo é especificado da seguinte forma

$$P\left[Y=y|\,\Theta=(\theta,\omega)\right]=\begin{cases}\omega+(1-\omega)\,\frac{a(s)[g(\theta)]^{s}}{f(\theta)}, & \text{se } y=s\\ (1-\omega)\,\frac{a(y)[g(\theta)]^{y}}{f(\theta)}, & \text{se } y\neq s,\end{cases} \tag{3.9}$$

com $0 \leq \omega \leq 1$, $y \in N$, onde N representa os números naturais, $f(\theta) = \sum_y a(y)\left[g(\theta)\right]^y < \infty$, $g(\theta)$ é positiva, finita, diferenciável e inversível e $a(y)$ é não negativa e independe de θ.

3.2.1 Propriedades da Classe de Distribuições Série de Potências Inflacionadas

É importante determinar algumas propriedades desta classe de modelos, podemos citar algumas delas:

A média da classe de distribuições série de potências inflacionadas depende da média da distribuição série de potências, pois

$$
\begin{aligned}
E(Y) &= \sum_{y=0}^{\infty} yP[Y=y] \qquad (3.10)\\
&= s\left[\omega + (1-\omega)\frac{a(s)[g(\theta)]^s}{f(\theta)}\right] + (1-\omega)\sum_{y=0,y\neq s}^{\infty} y\frac{a(y)[g(\theta)]^y}{f(\theta)}\\
&= s\omega + s\left[(1-\omega)\frac{a(s)[g(\theta)]^s}{f(\theta)}\right] + (1-\omega)\sum_{y=0,y\neq s}^{\infty} y\frac{a(y)[g(\theta)]^y}{f(\theta)}\\
&= s\omega + (1-\omega)\sum_{y=0}^{\infty} y\frac{a(y)[g(\theta)]^y}{f(\theta)}\\
&= s\omega + (1-\omega)\mu_1(\theta)\\
&= s\omega + (1-\omega)E(X)
\end{aligned}
$$

Fazendo $\mu_1(\theta) = E(X)$ e $\nu_1(\theta) = E(Y)$ a equação (3.10) se reduz a

$$\nu_1(\theta) = s\omega + (1-\omega)\mu_1(\theta).$$

Para obter o segundo momento, então $\mu_2(\theta)$ e $\nu_2(\theta)$ será:

$$
\begin{aligned}
E\left(Y^2\right) &= \sum_{y=0}^{\infty} y^2 P[Y=y] \qquad (3.11)\\
&= s^2\left[\omega + (1-\omega)\frac{a(s)[g(\theta)]^s}{f(\theta)}\right] + (1-\omega)\sum_{y=0,y\neq s}^{\infty} y^2\frac{a(y)[g(\theta)]^y}{f(\theta)}\\
&= s^2\omega + (1-\omega)\sum_{y=0}^{\infty} y^2\frac{a(y)[g(\theta)]^y}{f(\theta)}\\
&= s^2\omega + (1-\omega)\mu_2(\theta)\\
&= s^2\omega + (1-\omega)E\left(X^2\right)
\end{aligned}
$$

Determinando o r-ésimo momento

$$
\begin{aligned}
E(Y^r) &= \sum_{y=0}^{\infty} y^r P[Y=y] \qquad (3.12)\\
&= s^r\left[\omega + (1-\omega)\frac{a(s)[g(\theta)]^s}{f(\theta)}\right] + (1-\omega)\sum_{y=0,y\neq s}^{\infty} y^r\frac{a(y)[g(\theta)]^y}{f(\theta)}\\
&= s^r\omega + (1-\omega)\sum_{y=0}^{\infty} y^r\frac{a(y)[g(\theta)]^y}{f(\theta)}\\
&= s^r\omega + (1-\omega)\mu_r(\theta)\\
&= s^r\omega + (1-\omega)E(X^r)
\end{aligned}
$$

Portanto, se $\mu_r(\theta)$ e $\nu_r(\theta)$ denotam o r-ésimo momento da distribuição série de potências e da distribuição série de potências inflacionada, respectivamente, então

$$\nu_r(\theta) = s^r\omega + (1-\omega)\,\mu_r(\theta), \quad r = 1, 2, 3, \ldots. \tag{3.13}$$

Ou podemos denotar $M_X(k) = E\left(X^k\right)$ representando o *k-ésimo* momento de X e $M_Y(k) = E\left(Y^k\right)$ denotando o *k-ésimo* momento de Y, logo (3.13) pode ser escrita como

$$M_Y(k) = s^k\omega + (1-\omega)\,M_X(k), \quad k = 1, 2, 3, \ldots. \tag{3.14}$$

Para determinar a variância, considere

$$\begin{aligned}
V(Y) &= E\left(Y^2\right) - (E(Y))^2 \qquad (3.15)\\
&= s^2\omega + (1-\omega)\,\mu_2(\theta) - [s\omega + (1-\omega)\,\mu_1(\theta)]^2\\
&= s^2\omega + (1-\omega)\left[V(X) + \mu_1^2(\theta)\right] - \left[(s\omega)^2 + 2s\omega(1-\omega)\,\mu_1(\theta) + (1-\omega)^2\,\mu_1^2(\theta)\right]\\
&= s^2\omega + (1-\omega)\,V(X) + (1-\omega)\,\mu_1^2(\theta) - (s\omega)^2 - 2s\omega(1-\omega)\,\mu_1(\theta) - (1-\omega)^2\,\mu_1^2(\theta)\\
&= (1-\omega)\,s^2\omega + (1-\omega)\,V(X) - 2s\omega(1-\omega)\,\mu_1(\theta) + (1-\omega)\,\mu_1^2(\theta)\,[1-(1-\omega)]\\
&= (1-\omega)\,s^2\omega + (1-\omega)\,V(X) - 2s\omega(1-\omega)\,\mu_1(\theta) + \omega(1-\omega)\,\mu_1^2(\theta)\\
&= (1-\omega)\left[s^2\omega + V(X) + \omega\mu_1(\theta)\,[-2s + \mu_1(\theta)]\right]
\end{aligned}$$

Fazendo $s = 0$ temos que (3.13) se reduz à:

$$\nu_r(\theta) = (1-\omega)\,\mu_r(\theta), \quad r = 1, 2, 3, \ldots. \tag{3.16}$$

Se considerarmos em (3.16) $r = 2$, temos

$$\nu_2(\theta) = (1-\omega)\,\mu_2(\theta) = (1-\omega)\left[V(X) + \mu_1^2(\theta)\right].$$

Assim, a variância de Y, com $s = 0$, será um caso particular de (3.15)

$$\begin{aligned}
V(Y) &= E\left(Y^2\right) - (E(Y))^2 \qquad (3.17)\\
&= (1-\omega)\left[V(X) + \mu_1^2(\theta)\right] - [(1-\omega)\,\mu_1(\theta)]^2\\
&= (1-\omega)\,V(X) + (1-\omega)\,\mu_1^2(\theta) - \left(1 - 2\omega + \omega^2\right)\mu_1^2(\theta)\\
&= (1-\omega)\,V(X) + \omega\mu_1^2(\theta) - \omega^2\mu_1^2(\theta)\\
&= (1-\omega)\,V(X) + (1-\omega)\,\omega\mu_1^2(\theta)\\
&= (1-\omega)\left[V(X) + \omega\mu_1^2(\theta)\right]\\
&= (1-\omega)\left[\frac{g(\theta)}{g'(\theta)}\frac{d\mu_1(\theta)}{d\theta} + \omega\mu_1^2(\theta)\right].
\end{aligned}$$

Além disso, podemos verificar que $V(Y) > V(X)$ se, e somente se, $\omega < 1 - \gamma^2$, em que γ é o coeficiente de variação de X. Com efeito,

$$\begin{aligned} V(Y) &= (1-\omega)\left[V(X) + \omega\mu_1^2(\theta)\right] \\ &= V(X) - \omega V(X) + \omega\mu_1^2(\theta) - \omega^2\mu_1^2(\theta). \end{aligned}$$

Logo, $V(Y) > V(X)$ se, $-\omega V(X) + \omega\mu_1^2(\theta) - \omega^2\mu_1^2(\theta) > 0$, ou seja,

$$\begin{aligned} -\omega V(X) + \omega\mu_1^2(\theta) - \omega^2\mu_1^2(\theta) &> 0 \\ \omega\left[-V(X) + \mu_1^2(\theta) - \omega\mu_1^2(\theta)\right] &> 0 \\ -V(X) + \mu_1^2(\theta) - \omega\mu_1^2(\theta) &> 0 \\ -\omega\mu_1^2(\theta) &> -\mu_1^2(\theta) + V(X) \\ \omega\mu_1^2(\theta) &< \mu_1^2(\theta) - V(X) \\ \omega &< 1 - \frac{V(X)}{\mu_1^2(\theta)} \\ \omega &< 1 - \gamma^2. \end{aligned}$$

A obtenção destas propriedades mostraram a relação entre a classe de distribuições série de potências e a classe de distribuições série de potências inflacionadas. Além disso, foram descritas as propriedades para o caso geral, facilitando a obtenção delas quando o valor excessivo for diferente de zero.

3.2.2 Função de Verossimilhança para a Classe de Distribuições Série de Potências Inflacionadas

No contexto de distribuições para dados inflacionados de zeros, as misturas de modelos conseguem incorporar o excesso de zeros apresentado pelos dados. (ver Rodrigues, 2003). Para isto considere

$$\Pr[Y = y|\Theta = \theta] = \begin{cases} I_{\{s\}}(y), y = s \\ p(y|\theta), y \neq s, \end{cases} \tag{3.18}$$

em que a função indicadora $I_{\{s\}}(y)$ é uma distribuição que está degenerada em zero e $p(y|\theta)$ é uma função de probabilidade que se adequa aos dados caso não

existisse a inflação do ponto s. Para construir a mistura de distribuições considere os seguintes pesos.

Mistura de Distribuições

Y	s	y
Peso	ω	$(1-\omega)$

em que $0 \leq \omega \leq 1$, levando à expressão

$$P\left[Y=y|\Theta=(\theta,\omega)\right]=\omega I_{\{s\}}\left(y\right)+\left(1-\omega\right)p\left(y|\theta\right),\ \ y=0,1,2,.... \tag{3.19}$$

Neste trabalho será considerado o ponto de inflação em $s=0$. Logo o modelo (3.19) se reduz a

$$P\left[Y=y|\Theta=(\theta,\omega)\right]=\omega I_{\{0\}}\left(y\right)+\left(1-\omega\right)p\left(y|\theta\right),\ \ y=0,1,2,..., \tag{3.20}$$

sendo que, $p\left(y|\theta\right)$ é uma distribuição da classe de distribuições série de potências, com parâmetro θ, que teoricamente se ajustaria bem aos dados caso não houvesse a presença excessiva de zeros. O parâmetro ω é a proporção de zeros que excede o que seria predito através de $p\left(y|\theta\right)$. Portanto, pode-se dizer que esta equação é uma mistura de uma distribuição degenerada no ponto zero e uma outra distribuição com função de probabilidades dada por $p\left(y|\theta\right)$.

Segundo o modelo (3.9) o modelo (3.20) pode ser escrito como

$$P\left[Y=y|\Theta=(\theta,\omega)\right]=\begin{cases}\omega+\left(1-\omega\right)\frac{a(0)}{f(\theta)}, & \text{se } y=0\\ \left(1-\omega\right)\frac{a(y)[g(\theta)]^{y}}{f(\theta)}, & \text{se } y\neq 0.\end{cases} \tag{3.21}$$

Suponha que $\mathbf{Y}=(Y_1,...,Y_n)$ seja um vetor de n variáveis aleatórias com uma distribuição *série de potências zero inflacionadas*. Seja A o conjunto dos valores y_i iguais a zero, ou seja, $A=\{y_i : y_i=0\}$ e m representa o total de zeros, isto é, $m=n(A)$. Assim, a função de verossimilhança é

$$\begin{aligned}L\left(\theta,\omega\right) &= \prod_{i=1}^{n}p\left(y_i|\theta,\omega\right)\\ &= \left[\omega+\left(1-\omega\right)p\left(0|\theta\right)\right]^{m}\left(1-\omega\right)^{n-m}\prod_{y_i\notin A}p\left(y_i|\theta\right)\\ &\propto \left[\omega+\left(1-\omega\right)\frac{a(0)}{f(\theta)}\right]^{m}\left(1-\omega\right)^{n-m}\frac{g(\theta)^{\sum_{y_i\notin A}y_i}}{\prod\limits_{y_i\notin A}f\left(\theta\right)}.\end{aligned}$$

Em que, os elementos do conjunto A podem ser resultado de zeros verdadeiros ou aleatórios.

3.3 Abordagem Bayesiana

Nesta seção são apresentados alguns conceitos básicos sobre a abordagem bayesiana. Esta é adequada quando existe a possibilidade de incorporar informações adicionais aos dados. Para isto considere $X = (X_1, X_2, \ldots, X_n)$ uma amostra aleatória de tamanho n gerada de uma distribuição de probabilidades $f(x|\theta)$, onde x é o vetor de amostras observadas e θ é o vetor de parâmetros de interesse de dimensão k. (ver Saito, 2005). O procedimento de inferência bayesiana é baseado no teorema de Bayes, dado por

$$\pi(\theta|x) = \frac{L(\theta; x)\,\pi(\theta)}{\int L(\theta; x)\,\pi(\theta)}$$

em que, $L(\theta; x)$ é a função de verossimilhança para a amostra, $\pi(\theta)$ a distribuição a *priori* de θ e $\pi(\theta|x)$ é a distribuição a *posteriori* de θ dado x.

Na aplicação de inferência bayesiana, quando as distribuições a *priori* e a verossimilhança são conjugadas e as componentes de variância são supostas conhecidas, resultados analíticos são obtidos. Porém, nem sempre isto é possível e métodos de aproximação númerica são necessários para resolver as integrais envolvidas na obtenção das distribuições a *posteriori* de interesse. Esses métodos podem ser subdivididos em métodos numéricos analíticos e baseados em amostragem.

Os métodos analíticos mais utilizados são: *a aproximação pela normal, os métodos de quadratura e o método de Laplace.* Esses métodos são considerados mais precisos, mas apresentam a desvantagem de serem baseados em resultados assintóticos e na suposição de normalidade. Além disso, quando a dimensão do espaço paramétrico aumenta a tratabilidade desses métodos diminui consideravelmente e sua aplicação torna-se limitada. Os métodos numéricos baseados em amostragem envolvem os *métodos de Monte Carlo Simples, de Reamostragem por Importância e os métodos de Monte Carlo via Cadeias de Markov (MCMC).* Este

último é preferido aos primeiros porque são métodos iterativos, gerando cadeias de Markov que convergem para a distribuição de equilíbrio. (ver Souza, 1999)

A aplicabilidade desta técnica aumentou após a década de 90 com o avanço computacional e mostrou eficiência em problemas de dados aumentados, variáveis latentes, econometria e outros. Atualmente há contínuos refinamentos e extensões para o algoritmo como, por exemplo, o algoritmo Metropolis-Hastings e o amostrador de Gibbs, este último foi utilizado neste trabalho.

Para os valores simulados para a distribuição a *posteriori* via MCMC se faz necessária a análise de convergência, visto que existem alguns problemas associados ao uso destes procedimentos, tais como: *os valores gerados podem ser influenciados pelos valores atribuídos inicialmente; a determinação do número de iterações necessárias para atingir a convergência e a possível existência de correlação entre os parâmetros.* Mas, não há uma técnica geral para resolver esses possíveis eventos, a verificação pode ser feita através das propriedades da Cadeia de Markov.

A convergência das cadeias geradas pode ser averiguada através de técnicas gráficas e numéricas. As principais técnicas gráficas utilizadas são os gráficos de sequência, a superposição de histogramas ou de densidades ou através dos gráficos quantil-quantil. As principais técnicas númericas são as de Geweke e a de Gelman e Rubin. (ver Saito, 2005)

Neste trabalho foi utilizado o diagnóstico de Gelman-Rubin que é baseado na análise de variância, comparando-a intra e entre as cadeias geradas. Pela proposta a convergência é verificada através do fator:

$$\sqrt{R} = \sqrt{\frac{Var\,(X)}{D}},$$

em que, D representa a variância dentro das cadeias.

Sob condição de convergência $\sqrt{R} \to 1$ quando o número de iterações tende ∞. Na prática a convergência se dá quando $\sqrt{R} < 1,01$.

3.3.1 Distribuição a Priori de Jeffreys

Em inferência bayesiana é utilizada informações disponíveis, por exemplo de experimentos passados e, através delas, determina-se a chance a *priori* para os parâmetros de interesse. A distribuição a *priori*, único elemento novo na análise bayesiana, é o ponto mais crítico e o mais criticado pelos frequentistas. Esta distribuição tem por intuito representar bem o conhecimento sobre a quantidade θ desconhecida, antes de realizar o experimento.

Uma alternativa é determinar a *priori* objetivamente baseando-se no mínimo de informação subjetiva a *priori*. Uma das razões em se usar *prioris* não informativas é que se espera que o experimento seja mais informativo que a *priori*. A princípio as distribuições uniformes são sujeridas para representar situações onde não se dispõe ou não se deseja utilizar informações já existentes. Mas, geralmente, esta distribuição a *priori* é imprópria.

A classe proposta por Jeffreys de *prioris* não informativas é invariante, porém, eventualmente imprópria e baseia-se na medida de informação de Fisher.

Para determinar a matriz de informação de Fisher, considere X com função de (densidade) probabilidade $p\left(x|\theta\right)$. A matriz de informação esperada de Fisher de θ através de X é definida como

$$J(\theta) = E\left[-\frac{\partial^2 logp(x|\theta)}{\partial\theta^2}\right]$$

Se θ for um vetor paramétrico defini-se então a matriz de informação esperada de Fisher de θ através de X como

$$J(\theta) = E\left[-\frac{\partial^2 logp(x|\theta)}{\partial\theta\partial\theta'}\right]$$

Quanto maior o valor do núcleo $J(\theta)$ maior é a informação contida na verossimilhança. Toma-se o negativo da esperança porque se espera que a curvatura da função seja negativa. Além disso, ela pode ser considerada como uma medida de informação global.

Agora, seja uma observação X com função de (densidade) probabilidade $p(x|\theta)$. A distribuição a *priori* não-informativa de Jeffreys tem função

de densidade dada por $p(\theta) \propto \left[J(\theta)^{1/2}\right]$, se θ for um vetor paramétrico então $p(\theta) \propto \left|detJ(\theta)^{1/2}\right|$.

É importante ressaltar que, como esta classe de distribuição a *priori* não informativa depende da distribuição amostral ela viola o princípio da verossimilhança.

3.3.2 Inferência Bayesiana para a Classe Distribuições Série de Potências Inflacionadas

Função de Verossimilhança baseada nos Dados Aumentados

Como a distribuição *série de potências zero inflacionadas* é uma mistura de duas distribuições, então a função de verossimilhança (3.22) pode ser simplificada com a utilização de um procedimento baseado em dados ampliados com variáveis latentes. Estas variáveis, por sua vez, auxilia na obtenção da distribuição a *posteriori*, no sentido em que torna as distribuições condicionais completas conhecidas, facilitando a implementação do amostrador de Gibbs.

Neste tipo de situação é natural definir a variável I na qual equivale a 1, se o valor amostral for $y_i = 0$, admitindo probabilidade $p(\theta, \omega)$ e I equivale a 0, se o valor amostral for $y_i \neq 0$, com probabilidade $[1 - p(\theta, \omega)]$.(ver Rodrigues, 2003). Assim,

$$I_i = \begin{cases} 1, & com \;\; probabilidade \quad p(\theta, \omega) \\ 0, & com \;\; probabilidade \; 1 - p(\theta, \omega), \end{cases} \tag{3.22}$$

$i = 1, ..., m$ e

$$p(\theta, \omega) = \frac{\omega}{\omega + (1 - \omega) p(0|\theta)}.$$

ou seja, esta variável indica se o elemento da i-ésima posição de A é tirado do *primeiro* componente de (3.20) ou não. Assim, a função de verossimilhança

baseada nos dados aumentados $D = \{Y, I\}$ é

$$\begin{aligned} L(\theta,\omega|D) &= L(\theta,\omega)\prod_{i=1}^{m} p(\theta,\omega)^{I_i}\left(1-p(\theta,\omega)\right)^{1-I_i} \\ &= \left\{[\omega+(1-\omega)p(0|\theta)]^m(1-\omega)^{n-m}\prod_{y_i\notin A}p(y_i|\theta)\right\}[p(\theta,\omega)]^{\sum I_i}[1-p(\theta,\omega)]^{m-\sum I_i} \\ &= \frac{\omega^S[(1-\omega)p(0|\theta)]^{m-S}}{[\omega+(1-\omega)p(0|\theta)]^{S+m-S}}[\omega+(1-\omega)p(0|\theta)]^m(1-\omega)^{n-m}\prod_{y_i\notin A}p(y_i|\theta) \\ &= \omega^S(1-\omega)^{n-S}p(0|\theta)^{m-S}\prod_{y_i\notin A}p(y_i|\theta) \\ &\propto \omega^S(1-\omega)^{n-S}p(0|\theta)^{m-S}\frac{g(\theta)^Z}{\prod_{y_i\notin A}f(\theta)}, \end{aligned} \tag{3.23}$$

em que, $Z = \sum_{y_i\notin A} y_i$ e $S = \sum_{i=1}^{m} I_i \sim Bin[m, p(\theta,\omega)]$. Para valores fixos de θ, $L(\theta,\omega|D)$ pode ser vista como uma função somente de ω e a função de verossimilhança pode ser interpretada como um modelo *Beta*. Similarmente, para um valor fixo de ω, $L(\theta,\omega|D)$ é vista como uma função somente de θ e, consequentemente, a função de verossimilhança pode ser considerada como uma densidade da família exponencial. (Ghosh, 2006).

A distribuição a *priori* conjunta é dada por $\pi(\theta,\omega)$ e a distribuição a *posteriori* conjunta de (θ,ω), dado D, é denotada por

$$\pi(\theta,\omega|D) \propto L(\theta,\omega|D)\,\pi(\theta,\omega).$$

Assuma que a distribuição a *priori* para ω é independente e as seguintes distribuições a *priori* condicionais conjugadas:

$$\omega \sim Beta(b_1, b_2) \text{ e } \theta \sim \pi(\theta),$$

sendo que $\pi(\theta) \propto g(\theta)^{a_1}/[f(\theta)]^{a_2}$ é uma distribuição a *priori* conjugada para a família série de potências. Os hiperparâmetros a_1, a_2, b_1, b_2, são assumidos conhecidos. Em particular, $b_1 = b_2 = 1$ determinam a distribuição a *priori* uniforme no intervalo $(0, 1)$ para ω. Para pequenos valores de a_2 resultam em uma distribuição a *priori* não informativa para θ.

O processo deve ser repetido até a convergência da distribuição a *posteriori* de $(\theta,\omega|D)$. Para a maioria dos procedimentos inferenciais, a sequência,

$\left(\left\{\theta^{(l)}, \omega^{(l)}\right\}\right)$, $l = 1..., N$, para N suficientemente grande, baseado em algum critério obtemos a convergência da distribuição.

Após a obtenção da convergência esta amostra pode ser usada para fazer inferências sobre (θ, ω) e qualquer outra função, como por exemplo $Pr(Y = 0) = \omega + (1 - \omega)\left[\frac{a(0)}{f(\theta)}\right]$.

No cenário bayesiano as distribuições a *priori* não informativas são importantes, principalmente quando não se tem conhecimento ou não se deseja utilizá-los. Uma classe de distribuições a *priori* foi desenvolvida por Fisher e está apresentada na próxima Seção.

Matriz de Informação de Fisher para a Classe de Distribuições Série de Potências Inflacionadas

A matriz de informação de Fisher para a classe de distribuições série de potências inflacionadas foi apresentada em Murat e Szynal (1998) que consideraram a seguinte reparametrização:

$$\delta = (1 - \omega)\left(1 - \frac{a(s)\left[g(\theta)\right]^s}{f(\theta)}\right). \tag{3.24}$$

Logo o modelo (3.9) pode ser reescrito como:

$$P[Y = y] = \begin{cases} 1 - \delta, \text{ se } y = s \\ \frac{\delta a(y)[g(\theta)]^y}{f(\theta) - a(s)[g(\theta)]^s}, \text{ se } y \neq s. \end{cases} \tag{3.25}$$

Seja $Y_1, Y_2, ..., Y_n$ uma amostra aleatória da distribuição (3.25) e n_y o número de observações que são iguais a y, tal que $n = \sum_{y \geq 0} n_y$. A Tabela 3.3 expressa esta idéia

TABELA 3.3: Frequência observada para os valores da variável Y

Y	0	1	2	3	$\cdots$	k	$k+1$	$\cdots$	Total
Frequência	n_0	n_1	n_2	n_3	$\cdots$	n_k	n_{k+1}	$\cdots$	n

Logo, a função de verossimilhança pode ser obtida como

$$L = \prod_{y=0}^{\infty} \left(P\left[Y = y\right]\right)^{n_y}.$$

Tomando o logarítmo da função de verossimilhança temos

$$\ln L = n_s \ln\left(1-\delta\right) + \sum_{y=0, y\neq s}^{\infty} n_y \ln\left[\frac{\delta a\left(y\right)\left[g\left(\theta\right)\right]^y}{f\left(\theta\right) - a\left(s\right)\left[g\left(\theta\right)\right]^s}\right]. \tag{3.26}$$

É necessário determinar as derivadas de primeira ordem em relação aos parâmetros do modelo, que são dadas por

$$\frac{\partial}{\partial\delta}\ln L = -\frac{n_s}{1-\delta} + \sum_{y=0, y\neq s}^{\infty}\frac{n_y}{\delta} = -\frac{n_s}{1-\delta} + \frac{n-n_s}{\delta}, \tag{3.27}$$

e

$$\frac{\partial}{\partial\theta}\ln L = \frac{g^{'}\left(\theta\right)}{g\left(\theta\right)}\sum_{y=0, y\neq s}^{\infty} y n_y - \frac{f^{'}\left(\theta\right) - sa\left(s\right)\left[g\left(\theta\right)\right]^{s-1} g^{'}\left(\theta\right)}{f\left(\theta\right) - a\left(s\right)\left[g\left(\theta\right)\right]^s}\left(n - n_s\right). \tag{3.28}$$

As derivadas de segunda ordem

$$\frac{\partial^2}{\partial\delta^2}\ln L = -\frac{n_s}{(1-\delta)^2} - \frac{n-n_s}{\delta^2}, \tag{3.29}$$

consequentemente,

$$\frac{\partial^2}{\partial\theta\partial\delta}\ln L = 0, \tag{3.30}$$

e

$$\begin{aligned}
\frac{\partial^2 \ln L}{\partial\theta^2} &= \left[\frac{g''\left(\theta\right)}{g\left(\theta\right)} - \left(\frac{g'\left(\theta\right)}{g\left(\theta\right)}\right)^2\right]\sum_{y=1, y\neq s}^{\infty} y n_y \\
&\quad - \left(n-n_s\right)\left\{\frac{f''\left(\theta\right)}{f\left(\theta\right) - a\left(s\right)\left[g\left(\theta\right)\right]^s} - \left(\frac{f'\left(\theta\right)}{f\left(\theta\right) - a\left(s\right)\left[g\left(\theta\right)\right]^s}\right)^2\right\} \\
&\quad + \left(n-n_s\right)\left\{\frac{sa\left(s\right)\left[g\left(\theta\right)\right]^{s-2}\left[\left(s-1\right)\left[g'\left(\theta\right)\right]^2 + g\left(\theta\right)g''\left(\theta\right)\right]}{f\left(\theta\right) - a\left(s\right)\left[g\left(\theta\right)\right]^s}\right\} \\
&\quad - \left(n-n_s\right)\left\{\frac{2f'\left(\theta\right)sa\left(s\right)\left[g\left(\theta\right)\right]^{s-1}g'\left(\theta\right)}{\left(f\left(\theta\right) - a\left(s\right)\left[g\left(\theta\right)\right]^s\right)^2} - \left(\frac{sa\left(s\right)\left[g\left(\theta\right)\right]^{s-1}g'\left(\theta\right)}{f\left(\theta\right) - a\left(s\right)\left[g\left(\theta\right)\right]^s}\right)^2\right\}.
\end{aligned} \tag{3.31}$$

Para determinar a matriz de informação de Fisher J é necessário encontrar a esperança das derivadas segundas. Calculamos,

$$
\begin{aligned}
J_{11} &= -E\left(\frac{\partial^2}{\partial\theta^2}\ln L\right) \qquad (3.32)\\
&= n\delta\left\{\frac{f''(\theta)}{f(\theta)-a(s)[g(\theta)]^s}-\left(\frac{f'(\theta)}{f(\theta)-a(s)[g(\theta)]^s}\right)^2\right\}\\
&\quad -n\delta\left\{\left[\frac{g''(\theta)}{g(\theta)}-\left(\frac{g'(\theta)}{g(\theta)}\right)^2\right]\frac{\mu_1(\theta)f(\theta)-sa(s)[g(\theta)]^s}{f(\theta)-a(s)[g(\theta)]^s}\right\}\\
&\quad -n\delta\frac{sa(s)[g(\theta)]^{s-2}}{(f(\theta)-a(s)[g(\theta)]^s)^2}\left[g'(\theta)\left(sf(\theta)g'(\theta)-2f'(\theta)g(\theta)\right)\right]\\
&\quad -n\delta\frac{sa(s)[g(\theta)]^{s-2}}{(f(\theta)-a(s)[g(\theta)]^s)^2}\left[\left(f(\theta)-a(s)[g(\theta)]^s\right)\left(g''(\theta)-[g'(\theta)]^2\right)\right],
\end{aligned}
$$

em que, $\mu_1(\theta)=\frac{g(\theta)f'(\theta)}{g'(\theta)f(\theta)}$.

$$J_{22}=-E\left(\frac{\partial^2}{\partial\delta^2}\ln L\right)=\frac{n}{\delta(1-\delta)}; \qquad (3.33)$$

$$J_{12}=I_{21}=-E\left(\frac{\partial^2}{\partial\theta\partial\delta}\ln L\right)=0. \qquad (3.34)$$

Como ressaltado, a classe de distribuições série de potências pode ser para qualquer distribuição série de potências $p(y|\theta)$. Logo, as distribuições binomial, binomial negativa e Poisson se adequam a esta situação.

No próximo capítulo será considerado um caso particular da classe de distribuições série de potências inflacionadas, denominada como distribuições zero inflacionadas (ZI). Estas distribuições são muito utilizadas para modelagem de dados discretos, no cenário bayesiano. Além disso, as distribuições a *priori* não informativas serão utilizadas, justificando o estudo da matriz de informação de Fisher apresentada neste capítulo.

3.4 Seleção de Modelos

O processo para a seleção de modelos é reconhecidamente essencial em inferência, pois o ajuste de um modelo a um conjunto de dados envolve a discriminação entre diferentes modelos competitivos.

A seleção bayesiana de modelos, em geral, é baseada no cálculo de probabilidades a *posteriori* para os modelos em questão e não apresenta dificuldades na comparação entre modelos com estruturas diferentes.

Quando, a *priori*, os modelos são igualmente prováveis, a chance a posteriori de um modelo em relação a outro se reduz ao fator de Bayes, interpretado com um resumo da evidência fornecida pelos dados, em favor de uma teoria científica quando comparado a outra. No entanto, existem dificuldades com esta abordagem quando a informação a *priori* sobre os parâmetros dos modelos é imprópria.

Pereira e Stern (2008) propuseram a medida de evidência bayesiana denominada teste de significância completamente bayesiano - *Full Bayesian Significanc Test* - aplicada em testes de significância para hipóteses precisas. Ele traduz a qualidade da teoria de decisão Bayesiana.

Estes critérios de seleção de modelos são apresentados a seguir.

3.4.1 Fator de Bayes

Para a obtenção do fator de Bayes considere $p\left(\mathbf{x}|\theta_{\mathbf{j}}, M_j\right)$ a distribuição conjunta dos dados, $p\left(\theta_{\mathbf{j}}|M_j\right)$ a distribuição a *priori*, $p\left(\theta_{\mathbf{j}}|x, M_j\right)$ a distribuição a *posteriori* de θ_j sob o modelo M_j, $j = 1, 2$.

Considere ainda

$p\left(M_1\right)$ e $p\left(M_2\right) = 1 - p\left(M_1\right)$ a distribuição a *priori* para os modelos M_1 e M_2, respectivamente.

Além disso,

$p\left(M_1|\mathbf{x}\right)$ e $p\left(M_2|\mathbf{x}\right) = 1 - p\left(M_1|\mathbf{x}\right)$ a distribuição a *posteriori* para tais modelos.

Utilizando o teorema de Bayes temos,

$$p\left(M_j|\mathbf{x}\right) = \frac{p\left(\mathbf{x}|M_j\right) p\left(M_j\right)}{p\left(\mathbf{x}|M_1\right) p\left(M_1\right) + p\left(\mathbf{x}|M_2\right) p\left(M_2\right)}.$$

A chance a *posteriori* do modelo M_1 contra M_2 pode ser escrita como

$$\frac{p\left(M_1|\mathbf{x}\right)}{p\left(M_2|\mathbf{x}\right)} = \frac{\frac{p(\mathbf{x}|M_1)p(M_1)}{p(\mathbf{x}|M_1)p(M_1)+p(\mathbf{x}|M_2)p(M_2)}}{\frac{p(\mathbf{x}|M_2)p(M_2)}{p(\mathbf{x}|M_1)p(M_1)+p(\mathbf{x}|M_2)p(M_2)}} = \frac{p\left(\mathbf{x}|M_1\right)p\left(M_1\right)}{p\left(\mathbf{x}|M_2\right)p\left(M_2\right)} = B_{12}\frac{p\left(M_1\right)}{p\left(M_2\right)}, \quad (3.35)$$

onde B_{12} é denominado fator de Bayes.

Quando $p\left(M_1\right) = p\left(M_2\right)$ a chance a *posteriori* de M_1 em relação a M_2 se resume ao fator de Bayes.

O fator de Bayes é obtido através da verossimilhança marginal. Assim

$$p\left(\mathbf{x}|M_j\right) = \int p\left(\mathbf{x}|\theta_\mathbf{j}, M_j\right)p\left(\theta_\mathbf{j}|M_j\right)d\theta_\mathbf{j}.$$

Para interpretar o fator de Bayes, foi apresentado por Jeffreys (1961) *apud* Souza (1999) um critério, conforme os valores dispostos na Tabela 3.4.

TABELA 3.4: Interpretação para diferentes valores do fator de Bayes

B_{21}	$log_{10}\left(B_{21}\right)$	$log_e\left(B_{21}\right)$	Evidência contra H_1
1.0 - 3.2	0.0 - 0.5	0.0 - 1.2	fraca
3.2 - 10.0	0.5 - 1.0	1.2 - 2.3	substancial
10.0 - 100.0	1.0 - 2.0	2.3 - 4.6	forte
> 100.0	> 2	> 4.6	decisiva

3.4.2 Teste de Significância Completamente Bayesiano

Pereira e Stern (2008) desenvolveram o FBST para testar a significância de uma hipótese precisa, assunto este rico em discussões e controvérsias, este teste mostra a qualidade da teoria de decisão bayesiana. Um dos benefícios em utilizá-la é por não apresentar problemas quando são atribuídas distribuições a *priori* impróprias, como acontece no fator de Bayes.

Para o cálculo da medida de evidência Ev é necessário somente conhecer a distribuição a *posteriori*, e não apresenta complicações quando a dimensionalidade do parâmetro e do espaço amostral são grandes. Computacionalmente Ev utiliza somente a otimização e a integração numérica, não se baseando em resultados assintóticos.

O teste de significância, baseado em Ev, não exige a adoção de uma distribuição a *priori* que atribua probabilidade positiva para o subconjunto referente à hipótese nula precisa. Esta é a característica de coerência mais pertinente do teste de significância complentamente bayesiano sobre o fator de Bayes para hipóteses nulas precisas.

Segundo Pereira e Stern (2008) muitos autores discutiram que, nos casos em que a distribuição a *posteriori* é absolutamente contínua e a hipótese nula é precisa, o uso do fator de Bayes para testar significância pode ser controverso, além disso, há recomendações para o uso dele quando se utiliza distribuições a *priori* vagas.

Para determinar Ev é necessário considerar uma hipótese precisa $H_0 : \theta \in \Theta_0$ então

$$g^* = sup_{H_0} g_x(\theta) \;\; e \;\; T = \{\theta \in \Theta : g_x(\theta) > g^*\},$$

em que

$$g_x(\theta) = g(\theta|x) \propto L_x(\theta)\, g(\theta).$$

A medida de evidência bayesiana contra H_0 é definida como a probabilidade a *posteriori* do conjunto tangencial, isto é,

$$\bar{Ev} = Pr(\theta \in T|x) = \int_T g_x(\theta) d\theta.$$

O valor da medida de evidência que apoia H_0, $Ev = 1 - \bar{Ev}$, não é uma evidência contra a hipótese alternativa. Equivalentemente, Ev não é evidência a favor da alternativa, embora esteja contra H_0.

Definição 3.1: O teste de significância completamente bayesiano é o procedimento que rejeita H_0 sempre que Ev é pequeno. Ou similarmente, não rejeita H_0 quando Ev for grande. (Pereira e Stern, 2008).

A única polêmica que pode surgir, na prática, é determinar qual valor de Ev leva a rejeição de H_0.

Rodrigues (2006) ressaltou que a medida de evidência bayesiana para hipóteses precisas é fácil implementar e pode ser aplicado para dados ampliados.

Esta é encontrada em dois passos: o primeiro é a *otimização* e o outro *integração numérica*.

Para isto considere, H_0 sob $\mathbf{\Theta}$, $p(\theta|x)$ uma densidade a posteriori de θ, dada a amostra x e o seguinte conjunto definido no espaço paramétrico

$$T_x = \{\theta \in \mathbf{\Theta} : p(\theta|x) > sup_{\Theta_0} p(\theta|x)\}.$$

A medida de evidência é definida como

$$Ev(\Theta_0, x) = 1 - P(\theta \in T_x|x) = P\left(\theta \in T_x^C|x\right). \tag{3.36}$$

Não rejeita H_0 sempre que a $Ev(\Theta_0, x)$ for grande, ou rejeita H_0 quando $Ev(\Theta_0, x)$ por pequena.

Capítulo 4

Modelos para Dados Inflacionados de Zeros

Em muitos estudos envolvendo dados discretos é comum encontrar uma grande quantidade de zeros nos dados. Logo os modelos usuais não conseguem acomodar os zeros excessivos requerendo a determinação de modelos mais flexíveis que consigam melhorar a análise estatística para o problema.

A teoria para dados inflacionados de zeros, na última década foi bastante desenvolvida, principalmente devido a sua aplicação ser de interesse em diversas áreas como atuária, ecologia, controle de qualidade, agronomia, biomédicas, ciências do comportamento humano, dentre outras.

Mendes (2007) também comentou sobre modelos zero inflacionados. Relata que uma abordagem possível para dados de contagem é considerar o modelo de Poisson. Quando há a presença de zeros em excesso é interessante atribuir a distribuição de Poisson inflacionada de zeros (ZIP). E sugere três tipos de abordagens como solução: modelos mistos (com a separação de efeitos fixos e aleatórios para as variáveis explicativas); estimação por quase-verossimilhança (incluindo um fator de dispersão diferente da unidade ou uma função de variância diversa) ou através da abordagem bayesiana (assumindo que o parâmetro do modelo possui uma distribuição de probabilidade). Neste trabalho será utilizada esta última abordagem.

Este capítulo está organizado como segue. Na seção 1 está apresentada a distribuição de Poisson zero inflacionada e na seção 2 é distribuição binomial negativa inflacionada, nelas são apresentadas as propriedades, o processo para simulação da distribuição a *posteriori*, a seleção de modelos e a matriz de informação de Fisher para os respectivos modelos.

4.1 A Distribuição de Poisson Zero Inflacionada

O modelo de Poisson é um dos modelos discretos mais conhecidos e utilizados na literatura. Ele foi publicado em 1837 e descrito por Siméon Denis Poisson, que estudava a probabilidade de julgamentos criminais e civis. (ver Saito, 2005).

Se $X \sim Po(\theta)$ a distribuição de probabilidade é descrita como

$$p(x|\theta) = \frac{\theta^x e^{-\theta}}{x!},\ x = 0, 1, 2, ...,$$

com $\theta > 0$, além disso, a média e variância desta distribuição são iguais a θ.

Quando nos dados analisados ocorre a presença excessiva de zeros é comum, na literatura, utilizar a distribuição de Poisson zero inflacionada (ZIP), que é um caso particular da classe de distribuições série de potências inflacionadas descritos em (3.19), com expressão

$$Pr(Y = y|\theta, \omega) = \omega I_0(y) + (1 - \omega)\frac{\theta^y e^{-\theta}}{y!},\ y = 0, 1, 2, ... \tag{4.1}$$

A esperança é obtida de (3.13)

$$E(Y) = (1 - \omega)E(X) = (1 - \omega)\theta. \tag{4.2}$$

Conforme (3.17), a variância será

$$Var(Y) = (1 - \omega)\left[Var(X) + \omega\mu_1(\theta)^2\right] = (1 - \omega)\left(\theta + \omega\theta^2\right). \tag{4.3}$$

Assumindo $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ um vetor aleatório oriundo da distribuição ZIP e utilizando as variáveis latentes sugeridas por Rodrigues (2003) a função de verossimilhança com dados aumentados será

$$L(\theta, \omega|D) \propto \omega^S (1-\omega)^{n-S} \theta^Z e^{-(n-S)\theta}, \tag{4.4}$$

em que $S \sim Bin[m, p(\theta, \omega)]$, $Z = \sum_{y_i \notin A} y_i$ e

$$p(\theta, \omega) = \frac{\omega}{\omega + (1-\omega) e^{-\theta}}.$$

As distribuições a *priori* serão da seguinte forma:

$$\pi(\theta) \sim Gamma(a, b) \quad \text{e} \quad \pi(\omega) \sim Beta(c, d).$$

Assim, a distribuição a *posteriori* conjunta para (θ, ω), dado $D = \{Y, I\}$, é

$$\pi(\theta, \omega|D) \propto \omega^{S+c-1} (1-\omega)^{n-S+d-1} \theta^{Z+a-1} e^{-(n-S+b)\theta}. \tag{4.5}$$

Para a obtenção das distribuições a *posteriori* dos parâmetros envolvidos no modelo, faz-se necessário utilizar métodos computacionais para aproximar a distribuição a *posteriori*. O algoritmo de Gibbs foi utilizado na simulação, através do *Software* R e está descrito a seguir.

4.1.1 Simulação da Posteriori do Modelo ZIP

Para a simulação da posteriori do modelo ZIP o procedimento foi o seguinte:

Passo 1: Dado $\left(\theta^{(j-1)}, \omega^{(j-1)}\right)$ corresponde ao estágio $(j-1)$, então $S^{(j)}$ é tal que

$$S^{(j)} \sim Binomial\left[m, p\left(\theta^{(j-1)}, \omega^{(j-1)}\right)\right].$$

Passo 2: Dado uma observação obtida da geração de $S^{(j)}$ então $\left(\theta^{(j)}, \omega^{(j)}\right)$ tem a seguinte densidade:

$$\omega^{(j)}|S^{(j)} \sim Beta\left[S^{(j)} + c, n - S^{(j)} + d\right];$$
$$\theta^{(j)}|S^{(j)} \sim Gamma\left[Z + a, n - S^{(j)} + b\right].$$

O processo é iterado até a obtenção da convergência.

Um dos objetivos deste trabalho é selecionar o modelo que representa melhor os dados através de diferentes critérios de seleção. Nas próximas subseções são apresentados os critérios de seleção aqui considerados.

4.1.2 Fator de Bayes para o modelo ZIP *versus* Poisson

Para a determinação do fator de Bayes é necessária a especificação das distribuições a *priori*. Nas aplicações deste trabalho foram utilizados peso iguais aos modelos Poisson e ZIP.

Como em Datta et al. (2007), atribuímos para a distribuição a *priori* do parâmetro θ envolvido no modelo de Poisson, $P(\theta) \propto \theta^{-1/2}$, e a mesma para a distribuição a *priori* para o modelo ZIP, $P(\omega, \theta) \propto \theta^{-1/2}$.

Considerando uma amostra aleatória $Y_1, Y_2, \cdots, Y_n$, m o número de zeros contidos na amostra e r é a soma dos valores diferente de zero. Assim

$$m = \sum_{i=1}^{n} I_{[Y_i=0]} \text{ e } \ r = \sum_{i=1}^{n} Y_i.$$

Se considerarmos que os dados são provinientes de uma distribuiçao de Poisson M_0, a função de verossimilhança é denotada por,

$$L_0(\theta) = \frac{e^{-n\theta}\theta^r}{\prod y_i!},$$

e a verossimilhança para o modelo M_1, ou seja, os dados provêm de uma distribuição de Poisson zero inflacionada é dada por

$$L_1(\omega, \theta) = [\omega + (1-\omega)e^{-\theta}]^m[(1-\omega)e^{-\theta}]^{n-m}\frac{\theta^r}{\prod_{i\notin A} y_i!}.$$

Como $p(M_0) = p(M_1)$, a chance a posteriori de M_1 em relação a M_0 se resume ao fator de Bayes

$$B_{10=}\frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_0)}.$$

Para encontrar o fator de Bayes é necessário determinar as distribuições

preditivas

$$\begin{aligned} p\left(\mathbf{y}|M_0\right) &= \int_0^\infty L_0\left(\theta\right)\theta^{-1/2}d\theta \qquad (4.6)\\ &= \int_0^\infty \frac{e^{-n\theta}\theta^r}{\prod y_i!}\theta^{-1/2}d\theta \\ &= \int_0^\infty \frac{e^{-n\theta}\theta^{r+1/2-1}}{\prod y_i!}\frac{n^{r+1/2}}{\Gamma\left(r+1/2\right)}d\theta\frac{\Gamma\left(r+1/2\right)}{n^{r+1/2}} \\ &= \frac{\Gamma\left(r+1/2\right)n^{-(r+1/2)}}{\prod y_i!}, \end{aligned}$$

e

$$\begin{aligned} p\left(\mathbf{y}|M_1\right) &= \int_0^\infty\int_0^1 L_1\left(\omega,\theta\right)p(\omega,\theta)d\omega d\theta \qquad (4.7)\\ &= \int_0^\infty\int_0^1 [\omega+(1-\omega)e^{-\theta}]^m[(1-\omega)e^{-\theta}]^{n-m}\frac{\theta^r}{\prod y_i!}\theta^{-1/2}d\omega d\theta \\ &= \int_0^\infty\int_0^1\sum_{j=0}^m \binom{m}{j}\left((1-\omega)e^{-\theta}\right)^{m-j}\omega^j[(1-\omega)e^{-\theta}]^{n-m}\frac{\theta^{r-1/2}}{\prod y_i!}d\omega d\theta \\ &= \sum_{j=0}^m \binom{m}{j}\int_0^\infty\int_0^1 \omega^j[(1-\omega)e^{-\theta}]^{n-j}\frac{\theta^{r-1/2}}{\prod y_i!}d\omega d\theta \\ &= \sum_{j=0}^m \binom{m}{j}\int_0^\infty\int_0^1 \omega^j[(1-\omega)]^{n-j}\frac{B\left(j+1,n-j+1\right)}{B\left(j+1,n-j+1\right)}d\omega\frac{e^{-\theta(n-j)}\theta^{r-1/2}}{\prod y_i!}d\theta \\ &= \sum_{j=0}^m \frac{m!}{(m-j)!j!}\frac{j!\,(n-j)!}{(n+1)!}\int_0^\infty \frac{(n-j)^{r+1/2}\,e^{-\theta(n-j)}\theta^{r-1/2+1-1}}{\Gamma\left(r+1/2\right)}d\theta\frac{\Gamma\left(r+1/2\right)}{(n-j)^{r+1/2}\prod y_i!} \\ &= \sum_{j=0}^m \frac{m!}{(m-j)!}\frac{(n-j)!}{(n+1)!}\frac{\Gamma\left(r+1/2\right)}{(n-j)^{r+1/2}\prod y_i!}, \end{aligned}$$

que é finita. Logo o fator de Bayes se reduz a

$$B_{10}=\frac{p\left(\mathbf{y}|M_1\right)}{p\left(\mathbf{y}|M_0\right)}=\frac{\sum_{j=0}^m \frac{m!}{(m-j)!}\frac{(n-j)!}{(n+1)!}\frac{\Gamma(r+1/2)}{(n-j)^{r+1/2}\prod y_i!}}{\frac{\Gamma(r+1/2)n^{-(r+1/2)}}{\prod y_i!}}=\frac{m!}{(n+1)!}\sum_{j=0}^m\frac{(n-j)!}{(m-j)!}\left(1-\frac{j}{n}\right)^{-(r+1/2)}.$$

Datta et al. (2007) estudaram o comportamento do fator de Bayes para diferentes tipos de distribuições a *priori*. Quando a densidade marginal sob o modelo M_1 é baseada em uma distribuição a *priori* imprópria e quando a maioria dos dados são zeros, o fator de Bayes não é bem-definido. Além disso, quando os zeros dominam a amostra, as informações não estimam precisamente os valores dos parâmetros do modelo ZIP.

Uma *priori* própria natural para θ é a distribuição *Gama(a, b)*. Se $a = b = 1$, a distribuição se reduz ao modelo *Exponencial(1)* e o fator de Bayes será $B_{10} = \sum_{j=0}^{n}(j+1)^{-1}$ quando as observações zero são dominantes nas contagens. Quando o tamanho da amostra é grande ($n \to \infty$) o fator de Bayes pode ser aproximado por $B_{10} \approx log(n+1)$.

Existem várias críticas relacionadas ao uso do fator de Bayes como critério de seleção. Uma delas é quando atribui-se distribuições a *priori* vagas aos parâmetros, além disso, o fator de Bayes utiliza uma escala arbitrária de interpretação do valor obtido.

Na literatura, existem outros critérios de seleção de modelos, podemos citar o critério de informação de Akaike - AIC, o critério de informação bayesiana - BIC. Mas, como alternativa ao fator de Bayes propõe-se utilizar a medida de evidência *Full Bayesian Significance Test*, denonimada teste de significância completamente bayesiano, considerada uma medida fácil de interpretar e será descrita a seguir.

4.1.3 Teste de Significância Completamente Bayesiano para a Distribuição de Poisson Zero Inflacionada

Rodrigues (2006) apresentou que o teste de significância completamente bayesiano pode ser aplicado para situações onde os dados são ampliados.

Esta medida de evidência é encontrada em dois passos. O primeiro é obtido através da otimização e o outro através da integração numérica da distribuição a *posteriori* do parâmetro de locação θ.

Para isto considere, $H_0 : \omega = 0$, ou seja, o modelo M_0 é adequado e a hipótese $H_1 : \omega > 0$, o modelo M_1 é adequado. Se a fatoração da função de verossimilhança for obtida, o teste de significância completamente bayesiano é baseado na verossimilhança marginal de θ. Como visto em (3.24) os parâmetros ω e θ não são correlacionados e obtemos a fatoração neste caso.

Passo de Otimização - Encontre a moda θ_0 da distribuição a *posteriori*

$\pi_0(\theta|Y)$ sob $H_0 : \theta \in \Theta_0$, onde $\Theta_0 = \{\theta : \omega = 0\}$, em que a densidade a *posteriori*, $\pi_0(\theta|Y)$, é dada por

$$\pi_0(\theta|Y) = \pi_0(\theta, \omega = 0) \prod_{i=1}^{n} p(y_i|\theta).$$

Equivale encontrar a moda θ_0 de $\pi_0(\theta|Y)$ sob H_0 dada por

$$\theta_0 = \frac{Z + a - 1}{n + b},$$

em que $\pi_0(\theta|Y)$ é a distribuição Gama$(Z + a, n + b)$.

Passo de Integração - como caso particular de (3.36) temos

$$Ev(H_0|D) = 1 - \Pr[\theta \in Z^\star(D)|D] = 1 - \int_0^\infty \int_{Z^\star(D)}^\infty \pi(\theta, \omega|D) d\theta d\omega,$$

com, $[Z^\star(D)] = \{\theta : \pi(\theta|D) \geq \pi_0(\theta|D)\}$.

A densidade marginal, $\pi(\theta|D)$, corresponde à distribuição Gama com parâmetros $(\sum_{y_i \notin A} y_i + a, n - S + b)$ com máximo em $\frac{Z+a-1}{n-S+b}$.

Na situação de dados ampliados Rodrigues (2006) disse que é equivalente testar $H_0 : \theta = \frac{Z+a-1}{n+b}$.

A medida de evidência é $Ev(H_0|D)$ é igual a 1 se, e somente se, $\omega = 0$, ou seja, não se rejeita a hipótese nula de que os dados se ajustam melhor ao modelo de Poisson. Se $Ev(H_0|D)$ for pequena, tem-se que $\omega > 0$, logo rejeita-se H_0 em favor de H_1 e o modelo ZIP representa melhor os dados.

4.1.4 Matriz de Informação de Fisher para o Modelo ZIP

Como já foi ressaltado, a matriz de informação de Fisher é importante, no cenário bayesiano, pois a classe de distribuições a *priori* não informativas proposta por Jeffreys se baseia no cálculo desta matriz.

O modelo ZIP é um caso particular da classe de distribuições série potências inflacionadas. Logo a parametrização utilizada por Murat e Szynal (1998) apresentada em (3.24) é considerada:

$$\delta = (1 - \omega)\left(1 - e^{-\theta}\right), \tag{4.8}$$

conforme o modelo apresentado em (3.25) e de acordo com a Tabela 3.1 a densidade do modelo ZIP é:

$$P(Y=y)=\begin{cases} 1-\delta, & se\ \ y=0 \\ \delta\left(\frac{e^{-\theta}\theta^{y}}{y!\left(1-e^{-\theta}\right)}\right), & se\ \ y\neq 0. \end{cases} \tag{4.9}$$

O logarítmo da função de verossimilhança do modelo ZIP, de acordo com (3.26), será

$$\ln L = n_0 \ln(1-\delta) + \sum_{y=1}^{\infty} n_y \ln\left[\frac{\delta\theta^{y}}{y!\,(e^{\theta}-1)}\right]. \tag{4.10}$$

As derivadas de primeira ordem em relação aos parâmetros serão

$$\frac{\partial \ln L}{\partial \delta} = \frac{-n_0}{1-\delta} + \frac{n-n_0}{\delta},$$

e

$$\frac{\partial \ln L}{\partial \theta} = \frac{1}{\theta}\sum_{y=1}^{\infty} y n_y - (n-n_0)\frac{e^{\theta}}{e^{\theta}-1}.$$

As derivadas de segunda ordem

$$\frac{\partial^2 \ln L}{\partial \delta^2} = \frac{-n_0}{(1-\delta)^2} - \frac{(n-n_0)}{\delta^2},$$

e

$$\frac{\partial^2 \ln L}{\partial \theta^2} = \frac{-1}{\theta^2}\sum_{y=1}^{\infty} y n_y + (n-n_0)\frac{e^{\theta}}{(e^{\theta}-1)^2}.$$

A matriz de informação de Fisher é obtida fazendo

$$J(\theta,\delta) = -E\begin{bmatrix} \frac{\partial^2 \ln L}{\partial \theta^2} & \frac{\partial^2 \ln L}{\partial \theta \partial \delta} \\ \frac{\partial^2 \ln L}{\partial \theta \partial \delta} & \frac{\partial^2 \ln L}{\partial \delta^2}. \end{bmatrix}$$

Logo,

$$J_{11}(\theta,\delta) = -E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = n\delta\left\{\frac{-\theta e^{\theta} + e^{2\theta} - e^{\theta}}{\theta\,(e^{\theta}-1)^2}\right\},$$

e

$$J_{22}(\theta,\delta) = -E\left[\frac{\partial^2 \ln L}{\partial \delta^2}\right] = \frac{n}{\delta\,(1-\delta)},$$

e

$$J_{12}(\theta,\delta) = J_{21}(\theta,\delta) = -E\left[\frac{\partial^2 \ln L}{\partial \delta \partial \theta}\right] = 0.$$

4.2 A Distribuição de Binomial Negativa Generalizada Zero Inflacionada

Os dados selecionados em pesquisas biológicas pode ser bem ajustado através da distribuição de Poisson. Mas eventualmente, os dados apresentam superdispersão, que pode ser resultado dos organismos estarem acumulados, agrupados ou agregados em mesmo espaço ou tempo.

Ross e Preece (2006), disserram que quando a superdispersão ocorre, a distribuição de binomial negativa é freqüentemente apropriada para os dados.

A seguir, é apresentada a distribuição binomial negativa generalizada.

$$p\left(x|\theta\right)=\frac{N\Gamma\left(N+\beta x\right)}{x!\Gamma\left(N+\beta x-x+1\right)}\left(\theta\left(1-\theta\right)^{\beta-1}\right)^{x}\left(1-\theta\right)^{N},\ x=0,1,2,...,\ (4.11)$$

com $0<\theta<1$, $|\beta\theta|<1$, $\beta=0$ ou $\beta\geq1$ e β conhecido.

A esperança matemática foi apresentada na Tabela 3.2 e é dada por

$$E\left(X\right)=\mu_{1}\left(\theta\right)=\frac{N\theta}{1-\theta\beta}$$

Se o parâmetro $\beta=0$ temos que o modelo (4.2) se reduz ao modelo Binomial, então $E\left(X\right)=N\theta$. Se $\beta=1$ temos o modelo Binomial Negativo, logo esperança será $E\left(X\right)=\frac{N\theta}{1-\theta}$.

Quando há a presença de valores s excessivos pode-se utilizar a distribuição binomial negativa generalizada inflacionada, que é um caso particular das distribuições série de potências inflacionadas descrita em (3.9). Então

$$P\left[Y=y|\Theta=(\theta,\omega)\right]=\begin{cases}\omega+(1-\omega)\frac{N\Gamma(N+\beta y)}{y!\Gamma(N+\beta y-y+1)}\left(\theta\left(1-\theta\right)^{\beta-1}\right)^{y}\left(1-\theta\right)^{N},\ \text{se } y=s\\(1-\omega)\frac{N\Gamma(N+\beta y)}{y!\Gamma(N+\beta y-y+1)}\left(\theta\left(1-\theta\right)^{\beta-1}\right)^{y}\left(1-\theta\right)^{N},\ \text{se } y\neq s.\end{cases}\quad(4.12)$$

Neste caso,

$$g\left(\theta\right)=\theta\left(1-\theta\right)^{\beta-1}$$

$$f\left(\theta\right)=\left(1-\theta\right)^{-N}$$

$$a\left(y\right)=\frac{N\Gamma\left(N+\beta y\right)}{y!\Gamma\left(N+\beta y-y+1\right)}$$

Se o parâmetro $\beta = 0$ o modelo (4.12) se reduz à distribuição binomial inflacionada

$$P\left[Y = y|\Theta = (\theta, \omega)\right] = \begin{cases} \omega + (1-\omega)\begin{pmatrix} N \\ y \end{pmatrix}\theta^y (1-\theta)^{(N-y)}, & \text{se } y = s \\ (1-\omega)\begin{pmatrix} N \\ y \end{pmatrix}\theta^y (1-\theta)^{(N-y)}, & \text{se } y \neq s. \end{cases} \tag{4.13}$$

Se o parâmetro $\beta = 1$ o modelo (4.12) se reduz à distribuição binomial negativa inflacionada

$$P\left[Y = y|\Theta = (\theta, \omega)\right] = \begin{cases} \omega + (1-\omega)\begin{pmatrix} N+y-1 \\ y \end{pmatrix}\theta^y (1-\theta)^{N}, & \text{se } y = s \\ (1-\omega)\begin{pmatrix} N+y-1 \\ y \end{pmatrix}\theta^y (1-\theta)^{N}, & \text{se } y \neq s. \end{cases} \tag{4.14}$$

Se o ponto inflacionado for $s = 0$. A esperança matemática é obtida de (3.13), logo

$$Se,\ \ \beta = 0 \Rightarrow E\left(Y\right) = (1-\omega)\, N\theta,$$
$$Se,\ \ \beta = 1 \Rightarrow E\left(Y\right) = (1-\omega)\,\frac{N\theta}{1-\theta}.$$

Conforme (3.17) a variância será

$$Var\left(Y\right) = (1-\omega)\left[Var\left(X\right) + \omega\mu_1\left(\theta\right)^2\right]$$

$$Se,\ \ \beta = 0 \Rightarrow Var\left(Y\right) = (1-\omega)\left[N\theta\left(1-\theta\right) + \omega\left(N\theta\right)^2\right],$$
$$Se,\ \ \beta = 1 \Rightarrow Var\left(Y\right) = (1-\omega)\left[\frac{N\theta}{\left(1-\theta\right)^2} + \omega\left(\frac{N\theta}{1-\theta}\right)^2\right].$$

Assumindo $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$ um vetor aleatório oriundo da distribuição ZIBNG, considerando $\beta = 1$ e utilizando o procedimento de variáveis latentes, a função de verossimilhança com dados aumentados será

$$L\left(\theta, \omega|D\right) \propto \omega^S \left(1-\omega\right)^{n-S} \theta^Z \left(1-\theta\right)^{N(n-S)},$$

em que $S \sim Bin\left[m, p\left(\theta, \omega\right)\right]$ e

$$p(\theta,\omega) = \frac{\omega}{\omega + (1-\omega)(1-\theta)^N}.$$

As distribuições a *priori* serão

$$\pi(\theta) \sim Beta(a,b) \quad \text{e} \quad \pi(\omega) \sim Beta(c,d).$$

Assim, a distribuição a *posteriori* conjunta para $(\theta,\omega|D)$ é

$$\pi(\theta,\omega|D) \propto \omega^{S+c-1}(1-\omega)^{n-S+d-1}\theta^{Z+a-1}(1-\theta)^{N(n-S)+b-1}.$$

O algoritmo utilizado para a obtenção das distribuições a *posteriori* está descrito seguir.

4.2.1 Simulação da Posteriori do Modelo ZIBN

Para a simulação das distribuições a *posteriori* envolvidas no modelo ZIBN foram utilizados o seguinte procedimento

Passo 1: Dado $\left(\theta^{(j-1)},\omega^{(j-1)}\right)$ corresponde ao estágio $(j-1)$, então $S^{(j)}$ é tal que

$$S^{(j)} \sim Binomial\left[m, p\left(\theta^{(j-1)},\omega^{(j-1)}\right)\right].$$

Passo 2: Dado uma observação obtida da geração de $S^{(j)}$ então $\left(\theta^{(j)},\omega^{(j)}\right)$ tem a seguinte densidade:

$$\omega^{(j)}|S^{(j)} \sim Beta\left[S^{(j)}+c, n-S^{(j)}+d\right];$$
$$\theta^{(j)}|S^{(j)} \sim Beta\left[Z+a, N\left(n-S^{(j)}\right)+b\right].$$

O mesmo raciocínio pode ser utilizado para a distribuição binomial zero inflacionada.

4.2.2 Fator de Bayes para o modelo ZIBN *versus* Binomial Negativo

Como ressaltado, para determinar o fator de Bayes devemos especificar o peso atribuído aos modelos competitivos. Nesta situação consideramos pesos

iguais aos modelos binomial negativo e ZIBN. Além disso, como distribuições a *priori* para os respectivos modelos atribuímos $\theta^{-1/2}$.

Se considerarmos que em M_2 os dados provêm da distribuição binomial negativa, então a função de verossimilhança é denotada

$$L_2(\theta) = \theta^r (1-\theta)^{Nn} \prod_{i=1}^{n} \begin{pmatrix} N+y_i-1 \\ y_i \end{pmatrix}. \tag{4.15}$$

Se o modelo M_3, considera que os dados provêm de uma distribuição Binomial Negativa Zero Inflacionada, então a função de verossimilhança é denotada por

$$L_3(\omega,\theta) = [\omega+(1-\omega)(1-\theta)^N]^m \left[(1-\omega)(1-\theta)^N\right]^{n-m} \theta^r \prod_{y_i \notin A} \begin{pmatrix} N+y_i-1 \\ y_i \end{pmatrix}. \tag{4.16}$$

Como os pesos atribuídos ao modelo são iguais, a chance a *posteriori* de M_3 em relação a M_2 se resume ao fator de Bayes. Então necessitamos determinar as distribuições preditivas para a distribuição binomial negativa e para a ZIBN. Temos respectivamente,

$$\begin{aligned}
p(\mathbf{y}|M_2) &= \int_0^1 L_2(\theta)\,\theta^{-1/2} d\theta \\
&= \int_0^1 \theta^{r-1/2}(1-\theta)^{Nn} \prod_{i=1}^{n} \begin{pmatrix} N+y_i-1 \\ y_i \end{pmatrix} d\theta \\
&= \int_0^1 \theta^{r-1/2}(1-\theta)^{Nn} \frac{B(r+1/2, Nn+1)}{B(r+1/2, Nn+1)} d\theta \prod_{i=1}^{n} \begin{pmatrix} N+y_i-1 \\ y_i \end{pmatrix} \\
&= \frac{\Gamma(r+1/2)\,\Gamma(Nn+1)}{\Gamma(Nn+r+3/2)} \prod_{i=1}^{n} \begin{pmatrix} N+y_i-1 \\ y_i \end{pmatrix}
\end{aligned}$$

e

$$
\begin{aligned}
p(\mathbf{y}|M_3) &= \int_0^1\int_0^1 L_3(\omega,\theta)\,p(\omega,\theta)d\omega d\theta \\
&= \int_0^1\int_0^1 [\omega+(1-\omega)(1-\theta)^N]^m[(1-\omega)(1-\theta)^N]^{n-m}\theta^r \prod_{y_i\notin A}\binom{N+y_i-1}{y_i}\theta^{-1/2}d\omega d\theta \\
&= \int_0^1\int_0^1\sum_{j=0}^m \binom{m}{j}\left((1-\omega)(1-\theta)^N\right)^{m-j}\omega^j[(1-\omega)(1-\theta)^N]^{n-m}\theta^{r-1/2}\prod_{y_i\notin A}\binom{N+y_i-1}{y_i}d\omega d\theta \\
&= \sum_{j=0}^m \binom{m}{j}\int_0^1\int_0^1 \omega^j(1-\omega)^{n-j}\frac{B(j+1,n-j+1)}{B(j+1,n-j+1)}d\omega\,(1-\theta)^{N(n-j)}\theta^{r-1/2}\prod_{y_i\notin A}\binom{N+y_i-1}{y_i}d\theta \\
&= \sum_{j=0}^m \frac{m!}{(m-j)!j!}\frac{j!\,(n-j)!}{(n+1)!}\int_0^1\theta^{r-1/2}(1-\theta)^{N(n-j)}d\theta\prod_{y_i\notin A}\binom{N+y_i-1}{y_i} \\
&= \sum_{j=0}^m \frac{m!}{(m-j)!}\frac{(n-j)!}{(n+1)!}\frac{\Gamma(r+1/2)\,\Gamma(N(n-j)+1)}{\Gamma(N(n-j)+r+3/2)}\prod_{y_i\notin A}\binom{N+y_i-1}{y_i},
\end{aligned}
$$

Logo o fator de Bayes se reduz a

$$
\begin{aligned}
B_{32} &= \frac{p(y|M_3)}{p(y|M_2)} \\
&= \frac{\sum_{j=0}^m \frac{m!}{(m-j)!}\frac{(n-j)!}{(n+1)!}\frac{\Gamma(r+1/2)\Gamma(N(n-j)+1)}{\Gamma(N(n-j)+r+3/2)}\prod_{y_i\notin A}\binom{N+y_i-1}{y_i}}{\frac{\Gamma(r+1/2)\Gamma(Nn+1)}{\Gamma(Nn+r+3/2)}\prod_{i=1}^n\binom{N+y_i-1}{y_i}} \\
&= \sum_{j=0}^m \frac{m!}{(m-j)!}\frac{(n-j)!}{(n+1)!}\frac{\Gamma(N(n-j)+1)\,\Gamma(Nn+r+3/2)}{\Gamma(N(n-j)+r+3/2)\,\Gamma(Nn+1)} \\
&= \frac{m!}{(n+1)!}\frac{\Gamma(Nn+r+3/2)}{\Gamma(Nn+1)}\sum_{j=0}^m \frac{(n-j)!}{(m-j)!}\frac{\Gamma(N(n-j)+1)}{\Gamma(N(n-j)+r+3/2)}.
\end{aligned}
$$

4.2.3 Fator de Bayes para o modelo ZIBN *versus* ZIP

Se considerarmos que M_1 os dados são oriundos da distribuição ZIP, então a preditiva é como em (4.8).

Se o modelo M_3, considera que os dados provêm de uma distribuição binomial negativa zero inflacionada, o fator de Bayes se reduz a

$$
\begin{aligned}
B_{31} &= \frac{p(y|M_2)}{p(y|M_1)} \\
&= \frac{\sum_{j=0}^m \frac{m!}{(m-j)!}\frac{(n-j)!}{(n+1)!}\frac{\Gamma(r+1/2)\Gamma(N(n-j)+1)}{\Gamma(N(n-j)+r+3/2)}\prod_{y_i\notin A}\binom{N+y_i-1}{y_i}}{\sum_{j=0}^m \frac{m!}{(m-j)!}\frac{(n-j)!}{(n+1)!}\frac{\Gamma(r+1/2)}{(n-j)^{r+1/2}\prod_{y_i\notin A}y_i!}} \\
&= \left[\prod_{y_i\notin A}\frac{(N+y_i-1)!}{(N-1)!}\right]\sum_{j=0}^m \frac{\Gamma(N(n-j)+1)}{\Gamma(N(n-j)+r+3/2)}\left(1-\frac{j}{n}\right)^{(r+1/2)}.
\end{aligned}
$$

4.2.4 Teste de Significância Completamente Bayesiano para a Distribuição Binomial Negativa Zero Inflacionada

Considere, $H_0 : \omega = 0$, ou seja, o modelo binomial negativo M_0 é adequado e a hipótese $H_1 : \omega > 0$, o modelo ZIBN M_1 é adequado.

Se a fatoração da função de verossimilhança for obtida, o teste de significância completamente bayesiano é baseado na verossimilhança marginal de θ.

Passo de Otimização - Encontre a moda θ_0 da distribuição a *posteriori* $\pi_0(\theta|Y)$ sob $H_0 : \theta \in \Theta_0$, onde $\Theta_0 = \{\theta : \omega = 0\}$. Que é equivalente a encontrar a moda θ_0 de $\pi_0(\theta|Y)$ sob H_0 dada por

$$\theta_0 = \frac{Z + a - 1}{Z + Nn + a + b - 2},$$

em que $\pi_0(\theta|Y)$ é a distribuição Beta$(Z + a, Nn + b)$.

Passo de Integração

$$Ev(H_0|D) = 1 - \Pr\left[\theta \in Z^\star(D)|D\right] = 1 - \int_0^\infty \int_{Z^\star(D)}^\infty \pi(\theta, \omega|D) d\theta d\omega,$$

com, $[Z^\star(D)] = \{\theta : \pi(\theta|D) \geq \pi_0(\theta|D)\}$.

A densidade marginal, $\pi(\theta|D)$, corresponde à distribuição Beta com parâmetros $(Z+a, N(n-S)+b)$ com máximo em $\frac{Z+a-1}{Z+N(n-S)+a+b-2}$. E é equivalente testar $H_0 : \theta = \frac{Z+a-1}{Z+Nn+a+b-2}$.

Não se rejeita a hipótese nula de que os dados se ajustam melhor ao modelo binomial negativo se, a medida de evidência é $Ev(H_0|D)$ é igual a 1. A hipótese nula é rejeitada se $Ev(H_0|D)$ for pequena, o modelo ZIBN representa melhor os dados.

4.2.5 Matriz de Informação de Fisher para o Modelo ZIBNG

Os resultados para a matriz de informação de Fisher para a distribuição binomial negativa zero inflacionada foram obtidas em Murat e Szynal (1998).

A parametrização apresentada em (3.24) é dada por

$$\delta = (1-\omega)\,(1-\theta)^N\,, \tag{4.17}$$

Utilizando (3.32), (3.33) e (3.34) temos

$$J_{11}\,(\theta,\delta) = \left\{ \frac{nN\,(1-\omega)\left[\sum_{i=0}^{N-1}\left[(1-\theta)^i - (1-\theta)^{N-1}\right] + N\theta\beta\,(1-\theta)^{N-1}\right]}{(1-\theta)\,(1-\theta\beta)\left(1-(1-\theta)^N\right)} \right\},$$

e

$$J_{22}\,(\theta,\delta) = \frac{n}{(1-\omega)\left(1-(1-\theta)^N\right)\left[\omega\left(1-(1-\theta)^N\right) + (1-\theta)^N\right]},$$

e

$$J_{12}\,(\theta,\delta) = J_{21}\,(\theta,\delta) = -E\left[\frac{\partial^2 \ln L}{\partial\delta\partial\theta}\right] = 0.$$

No próximo capítulo serão apresentadas algumas aplicações envolvendo a distribuição de Poisson zero inflacionada (ZIP) e a distribuição binomial negativa zero inflacionada (ZIBN). O principal objetivo foi verificar a qualidade de sua modelagem quando comparada à outras distribuições. Para a obtenção das distribuições a *posteriori*, distribuições a *priori* não informativas foram atribuídas aos seus respectivos parâmetros, fazendo a uso da matriz de informação de Fisher apresentada nesta seção.

Capítulo 5

Aplicações envolvendo as Distribuições Zero Inflacionadas

Neste capítulo são apresentadas aplicações que, concretamente, envolvem esta metodologia. Três conjuntos de dados utilizados foram extraídos de Datta et al. (2007), eles consideraram seleção de modelos considerando como critério de seleção o fator de Bayes. A proposta aqui, foi selecionar o melhor modelo entre o Poisson e ZIP, mas considerando o fator de Bayes e o teste de significância completamente bayesiano e, com isto, verificar quais são os benefícios em considerar tais medidas de evidências.

O conjunto de dados da quarta aplicação foi retirado de Ghosh et al. (2006), nesta aplicação eles consideraram os modelos de Poisson, ZIP, binomial negativo e ZIBN e os comparou através do *deviance*, nesta aplicação também utilizamos estes modelos e o comparamos através do fator de Bayes e do teste de significância completamente bayesiano.

Na seção 1 apresentamos uma aplicação sobre o número de infecções urinárias em homens com HIV. Na seção 2 um estudo sobre o número de pessoas com cólera em domicílios da Índia. Na seção 3 é apresentado um estudo sobre o número de atentados terroristas contra os EUA. Na seção 4 apresentamos uma aplicação sobre o número de defeitos em carros.

5.1 Infecção Urinária em Homens com HIV

Datta et al. (2007) apresentaram uma aplicação envolvendo o estudo de infecção urinária em homens com HIV. O ensaio clínico considerou 98 homens dos quais 81 não haviam apresentado infecção. O objetivo desta aplicação foi verificar se os homens não tinham infecção urinária, ou seja, se o modelo ZIP ajusta melhor os dados. Os dados estão resumidos na tabela 5.1.

TABELA 5.1: Número de infecções urinárias apresentadas por homens com HIV.

Y	0	1	2	3	Total
Frequência	81	9	7	1	98

Este conjunto de dados apresentou média amostral $\bar{x} = 0,2653$ e variância $s^2 = 0,400$. Como os dados são resultado de uma contagem o modelo de Poisson seria adequado, no entanto, observamos que a exigência entre a igualdade da média e da variância deste modelo não é cumprida. Como a variância é superior a média, temos indício de superdispersão e a classe de distribuições série de potências inflacionadas poderá ser adequada neste caso.

Considerando a metodologia descrita na seção (3.3.2) devemos determinar as aproximações das distribuições a *posteriori* e para isto, é necessário especificar as distribuições a *priori* para os respectivos parâmetros. Para média θ do modelo de Poisson foi atribuído uma distribuição a *priori* Gama com hiperparâmetros $a = b = 10^{-4}$. A distribuição a *posteriori* de θ é, também, uma distribuição Gama com parâmetros atualizados. Foram geradas duas cadeias com 50000 amostras utilizando um período de aquecimento de 5000 observações e saltos de iterações entre as amostras. A convergência do procedimento Monte Carlo via cadeias de Markov (MCMC) destes exemplos foi monitoradas pelo diagnóstico de Gelman-Rubin (1992), que consiste basicamente em uma análise de variância dos valores gerados.

Para a obtenção das distribuições a *posteriori* do modelo ZIP também foi considerada para θ uma priori Gama com os mesmos hiperparâmetros atribuídos ao modelo de Poisson e para ω foi atribuída uma distribuição a *priori* Beta, que neste caso coincide com a de Jeffreys *(Be(0.5,0.5))*. Considerou-se o mesmo

período de aquecimento e número de simulações. Além disso, através do critério de Gelman-Rubin, $R = 1$, podemos dizer que houve convergência das cadeias.

Na Tabela 5.2 estão apresentadas a média a *posteriori*, o desvio-padrão, e os quantis de $2,5\%$ e $97,5\%$ para o parâmetro θ do modelo de Poisson e para θ e ω do modelo ZIP.

TABELA 5.2: Resumo a posteriori para o parâmetro θ do modelo de Poisson, e para os parâmetros θ e ω do modelo ZIP.

Modelo	Parâmetro	Média	sd	2,5%	97,5%
Poisson	θ	0,2653	0,0521	0,1730	0,3762
ZIP	θ	0,8568	0,2824	0,3785	1,4846
	ω	0,6630	0,1167	0,3758	0,8306

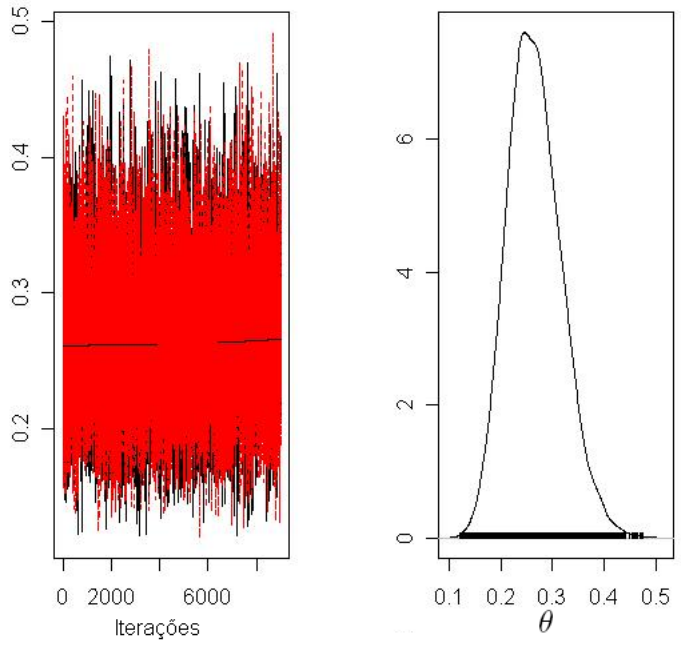

FIGURA 5.1: Comportamento das cadeias ao longo das iterações e a distribuição estimada o para o parâmetro θ do modelo de Poisson.

Na Figura 5.1 são apresentadas a distribuição estimada e o comportamento das cadeias ao longo das iterações para os parâmetros do modelo de Poisson.

Na Figura 5.2 são apresentadas as distribuições estimadas e o comportamento das cadeias ao longo das iterações para os parâmetros do modelo ZIP.

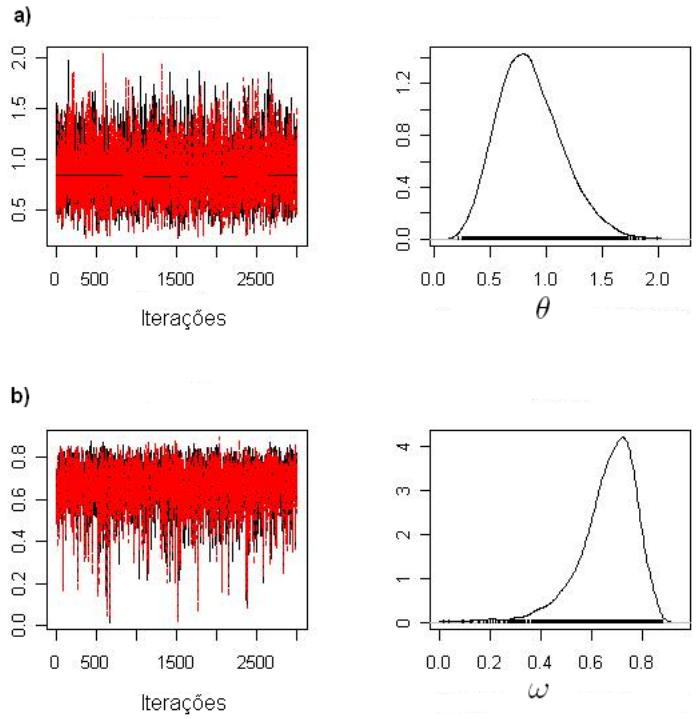

FIGURA 5.2: a) Comportamento das cadeias e distribuição estimada para o parâmetro θ do modelo ZIP. b) Comportamento das cadeias e distribuição estimada para o parâmetro ω do modelo ZIP.

Considerando os dados da Tabela 5.3 podemos observar que o valor obtido para o parâmetro θ do modelo de Poisson é muito ínfimo. Portanto, considerando este valor esperamos um número elevado de zeros na estimativa, ou seja, o número de homens sem infecção urinária estimado está em torno de 75. Espera-se que este valor seja mais próximo de 81. Se observamos as frequências esperadas para os demais valores de y observamos que as estimativas ficaram distantes da frequência observada. Então, como alternativa, serão vistas as estimativas para os parâmetros do modelos ZIP

Ao considerar o modelo ZIP verifica-se que a estimativa do parâmetro θ aumentou quando comparada à distribuição de Poisson e o valor obtido para ω é, possivelmente, significativo, ou seja, há fortes evidências sobre a existência de zeros excessivos nos dados. Considerando estas estimativas são esperados

79 homens sem infecção urinária, valor próximo ao amostral, além disso, se observamos as frequências esperadas para os demais valores de y observamos que as estimativas ficaram próximas à frequência observada.

Para verificar qual modelo se ajusta melhor aos dados é necessário utilizar critérios para a seleção de modelos.

TABELA 5.3: Valores observados e esperados segundo os modelos ZIP e Poisson

Y	Frequência observada	Esperada - ZIP	Esperada - Poisson
0	81	79	75
1	9	12	20
2	7	6	3
3	1	1	0
Total	98	-	-

Visualmente o modelo ZIP estima melhor os dados, mas para comprovar essa evidência é necessário verificar através do fator de Bayes. Datta et al. (2007) obteveram $B_{10} = 223,13$ (este foi confirmado no programa desenvolvido em R). O fator de Bayes utiliza uma tabela arbitrária para interpretação e existem críticas associadas às escalas utilizadas.

Diferentemente de Datta et al. (2007) foi considerado outro critério para seleção de modelos: *o teste de significância completamente bayesiano*. Esta medida de evidência foi implementada utilizando o programa em *Winbugs* apresentado por Rodrigues (2006). O valor obtido foi $Ev = 0.01$, este resultado é mais realista que o fator de Bayes, principalmente na explicação da medida, ou seja, ela pode ser considerada como um peso que favorece a hipótese alternativa se Ev for próximo de zero.

Devido à magnitude de θ (valor muito próximo de zero), o modelo de Poisson teve um ajustamento razoável para o valor $Y = 0$, já para os outros valores de Y o ajustamento foi aquém do desejado.

5.2 Incidência de Cólera

Datta et al. (2007) apresentaram dados sobre a incidência de cólera em uma aldeia da Índia em 1920, onde foi considerado o número de infectados por casa. Esta aplicação teve o objetivo de verificar qual modelo se adequa melhor aos dados. Estes estão listados na Tabela a seguir.

TABELA 5.4: Número de pessoas com cólera

Y	0	1	2	3	4	**Total**
Frequência	168	32	16	6	1	223

A Tabela 5.4 apresenta um número de zeros bastante elevado, correspondendo a $75,34\%$ do total de observações.

O conjunto de dados apresentou média amostral $\bar{x} = 0,3856$ e variância $s^2 = 0,5957$. Como a igualdade da média e da variância deste modelo não se verificou, o modelo ZIP também será considerado. Mantendo o mesmo objetivo de verificar quais dos modelos Poisson ou ZIP se ajustam melhor aos dados, as estimativas das distribuições a *posteriori* para os respectivos parâmetros foram obtidas e estão na Tabela 5.5.

TABELA 5.5: Resumo a posteriori para o parâmetro θ do modelo de Poisson e para os parâmetros θ e ω do modelo ZIP

Modelo	Parâmetro	Média	(sd)	$2,5\%$	$97,5\%$
Poisson	θ	0,3857	0,0417	0,3083	0,4712
ZIP	θ	0,9581	0,1625	0,6782	1,3140
	ω	0,5887	0,0669	0,4464	0,7020

Nota-se que o valor obtido para θ da distribuição Poisson é baixo, logo o modelo estima um número de 151 zeros muito distante de 168. (ver Tabela 5.6)

Considerando o modelo ZIP, observamos que o valor do parâmetro θ cresceu consideravelmente quando comparado ao obtido para a distribuição de Poisson e a estimativa de ω é relativamente grande, indicando a existência zeros que não são acomodados pelo modelo de Poisson.

Na Figura 5.3 são apresentadas as distribuições estimadas e o comportamento das cadeias ao longo das iterações para os parâmetros do modelo de Poisson.

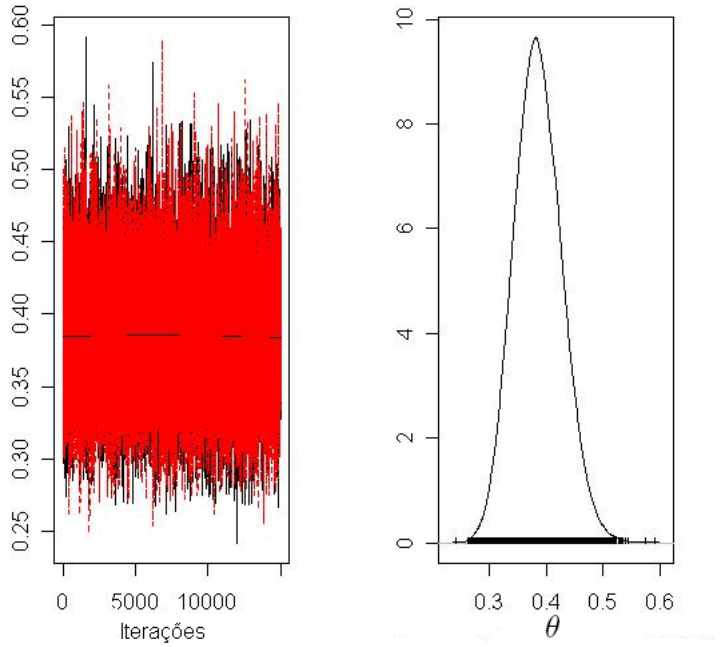

FIGURA 5.3: Comportamento das cadeias e distribuição estimada para o parâmetro θ do modelo de Poisson.

A partir do gráfico das amostras geradas observa-se que existe um indício de convergência já que os valores amostrados para as duas cadeias se dispõem aleatoriamente em torno de uma reta constante paralela ao eixo das abscissas.

Para determinar a distribuição a *posteriori* de ω considerou-se a distribuição a *priori* Beta(0.5,0.5). Na Figura 5.4 são apresentadas as trajetórias dos valores amostrados e as densidades estimadas para os parâmetros do modelo ZIP. O gráfico mostra um indício de convergência para os valores gerados, $(R = 1)$.

Observando a Tabela 5.6 pode-se verificar que modelo ZIP estima melhor os dados, mas para comprovar essa evidência é necessário verificar, através de um critério de seleção, qual modelo melhor se ajusta aos dados. O fator de Bayes obtido foi $B_{10} = 238.090$ e a medida de evidência obtida foi $Ev = 0.0$.

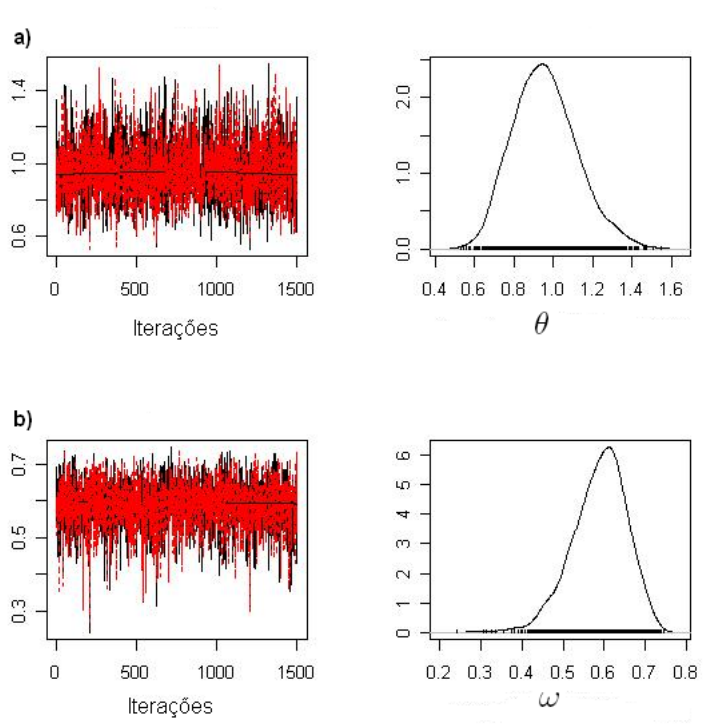

FIGURA 5.4: a) Comportamento das cadeias e distribuição estimada para o parâmetro θ do modelo ZIP. b) Comportamento das cadeias e distribuição estimada para o parâmetro ω do modelo ZIP.

TABELA 5.6: Valores observados e esperados segundo os modelos ZIP e Poisson

Y	Frequência Observada	Esperada - ZIP	Esperada - Poisson
0	168	167	151
1	32	34	59
2	16	16	11
3	6	5	2
4	1	1	0
Total	223	-	-

Considerando os critérios de seleção apresentados verifica-se que o modelo ZIP ajusta melhor os dados sobre a incidência de cólera em uma aldeia da Índia, ou seja, podemos dizer que existe excesso de zeros nos dados sobre a incidência de cólera em uma aldeia da Índia (a doença recaiu a poucos indivíduos da aldeia).

5.3 Número de Atentados Terroristas por Mês contra os EUA

Datta et al. (2007) apresentaram uma aplicação sobre o número de atentados terroristas por mês contra o Estados Unidos, entre os anos de 1968 e 1974. Nesta aplicação pretende-se verificar se os atentados terroristas não aconteceram, com alta frequência, durante o tempo pesquisado, ou seja, verificar se o valor $Y = 0$ é excessivo, justificando a utilização do modelo ZIP. A Tabela 5.7 apresenta as informações sobre o número de atentados terroristas por mês contra os Estados Unidos.

TABELA 5.7: Número de atentados terroristas por mês nos EUA

Y	0	1	2	3	4	Total
Frequência	38	26	8	2	1	75

Nota-se que há aproximadamente 50% de zeros nas observações, a média amostral $\bar{x} = 0,69$ e a variância amostral $s^2 = 0,76$, visualmente a diferença entre a média e a variância não é tão grande para podermos dizer que há superdispersão nos dados. Mas, verificaremos a estimativa do parâmetro ω do modelo ZIP para verificar se há zeros excessivos na amostra.

Serão determinadas as distribuições a *posteriori* dos parâmetros dos modelos. A Figura 5.5 apresentada as distribuições estimadas e o comportamento das cadeias ao longo das iterações para o parâmetro do modelo de Poisson.

A Tabela 5.8 apresenta o resumo estatístico para as distribuições a *posteriori*.

TABELA 5.8: Resumo estatístico para o parâmetro θ do modelo de Poisson e para os parâmetros θ e ω do modelo ZIP

Modelos	Parâmetro	Média	sd	2,5%	97,5%
Poisson	θ	0,6953	0,0972	0,5164	0,8971
ZIP	θ	0,7688	0,1366	0,5455	1,0826
	ω	0,0948	0,0920	0,0002	0,3189

Considerando estimativa para θ do modelo de Poisson espera-se que nos EUA tenham 37 meses sem atentados terroristas, de um total de 75 meses analisados. Este valor obtido foi muito próximo ao real.

Considerando os valores obtidos para os parâmetros do modelo ZIP esperamos que em 38 meses de um total 75 meses não tenham sido cometidos atentados terroristas contra os EUA, valor exatamente igual ao dado real. O bom ajustamento das distribuições de Poisson e ZIP pode ser comprovados na Tabela 5.9.

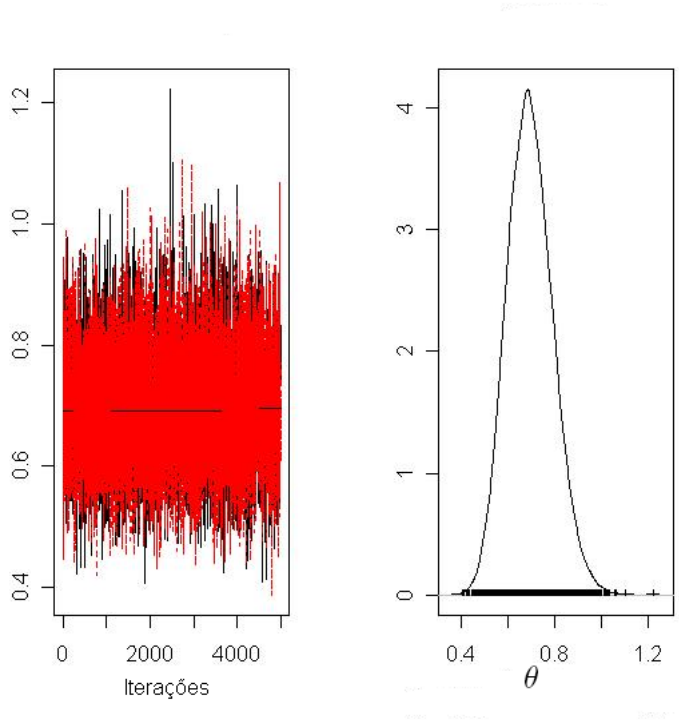

FIGURA 5.5: Comportamento das cadeias e a distribuição estimada para o parâmetro θ do modelo de Poisson.

As densidades estimadas para os parâmetros do modelo ZIP e as trajetórias das cadeias estão apresentadas na Figura 5.6.

Nota-se que os modelos de Poisson e ZIP estimam bem os dados sobre atentados terroristas aos EUA. Para seleção de modelos utilizou-se o fator de Bayes ($B_{10} = 0.28$, ou seja, há uma fraca evidência a favor do modelo ZIP). Este resultado pode ser consequência do valor de θ obtido para a distribuição de

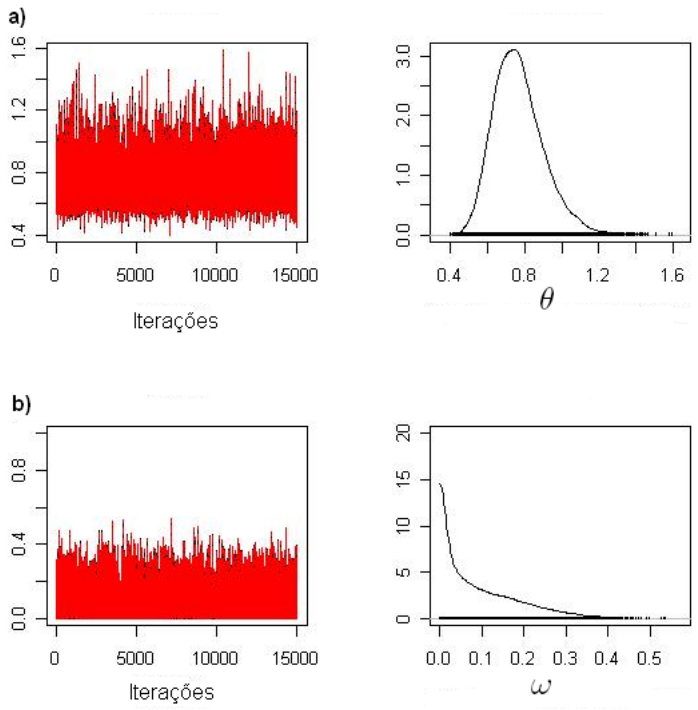

FIGURA 5.6: Comportamento das cadeias e as distribuições estimadas para os parâmetros θ do modelo ZIP. b) Comportamento das cadeias e as distribuições estimadas para os parâmetros ω do modelo ZIP.

TABELA 5.9: Valores observados e esperados segundo os modelos ZIP e Poisson

Y	Frequência Observada	Esperada - ZIP	Esperada Poisson
0	38	38	37
1	26	24	25
2	8	10	9
3	2	2	2
4	1	0	0
Total	75	-	-

Poisson ($\theta = 0,6953$). A medida de evidência $FBST$ obtida foi $Ev = 0.59$ (este não rejeita H_0, mas o valor obtido considera que o modelo ZIP também ajusta bem os dados, isto devido às magnitudes de θ e ω obtidas, ou seja, existem 9.48% de zeros excessivos. Como o valor de θ é relativamente baixo é esperado um alto número de zeros); portanto, resultado justifica a adequabilidade tanto da Poisson

como do modelo ZIP.

5.4 Número de Defeitos em Carros

Ghosh et al. (2006), apresentaram um conjunto de dados sobre o número de defeitos em carros. Em determinado período foram verificados 54 carros e classificados quanto aos número de não conformidades. O interesse pelo conjunto de dados foi pela grande quantidade de zeros na amostra e por considerar a mesma metodologia sugerida neste trabalho.

Ajustou-se os modelos de Poisson, ZIP, binomal negativo e ZIBN com o objetivo de verificar quais deles ajustam melhor dados com zeros em excesso. As informações estão dispostas a seguir.

TABELA 5.10: Número defeitos apresentados pelos carros

Y	0	1	2	3	Total
Frequência	42	8	2	2	54

Sabemos que em empresas que utilizam regras rígidas de controle de qualidade é esperado que os carros não apresentem defeitos, ou seja, deseja-se que o valor zero formem uma subpopulação. Nestas situações misturas de modelos são recomendadas, logo o modelo ZIP ou o ZIBN se ajustariam a essa situação.

Como o objetivo principal é fazer a seleção de modelos, na Tabela 5.11 são apresentadas as estimativas para a distribuição a *posteriori* de θ sob o modelo de Poisson e binomial negativo e as de θ e ω para os modelos ZIP e ZIBN.

As Figuras 5.7 e 5.8 apresentam o comportamento das cadeias a distribuição estimada para os parâmetros do modelo de Poisson e binomial negativo, ZIP e ZIBN.

Considerando os gráficos apresentados para as amostras geradas observa-se que existe um indício de convergência que foi comprovado através do critério de Gelman-Rubin ($R_{Poisson} = 0,999$ e $R_{BN} = 1$).

Considerando os gráficos apresentados na Figura 5.8 observa-se que existe

TABELA 5.11: Resumo estatístico para o parâmetro θ do modelo de Poisson e do binomial negativo e para os parâmetros θ, ω dos modelos ZIP e ZIBN

Modelo	Parâmetro	Média	sd	2,5%	97,5%
Poisson	θ	0,3332	0,0791	0,1930	0,5056
ZIP	θ	0,7746	0,3293	0,2930	1,5757
	ω	0,5097	0,1908	0,0338	0,7852
Binomial Negativo	θ	0,1088	0,0233	0,0670	0,1555
ZIBN	θ	0,1779	0,0647	0,0810	0,3244
	ω	0,4275	0,2067	0,0111	0,7497

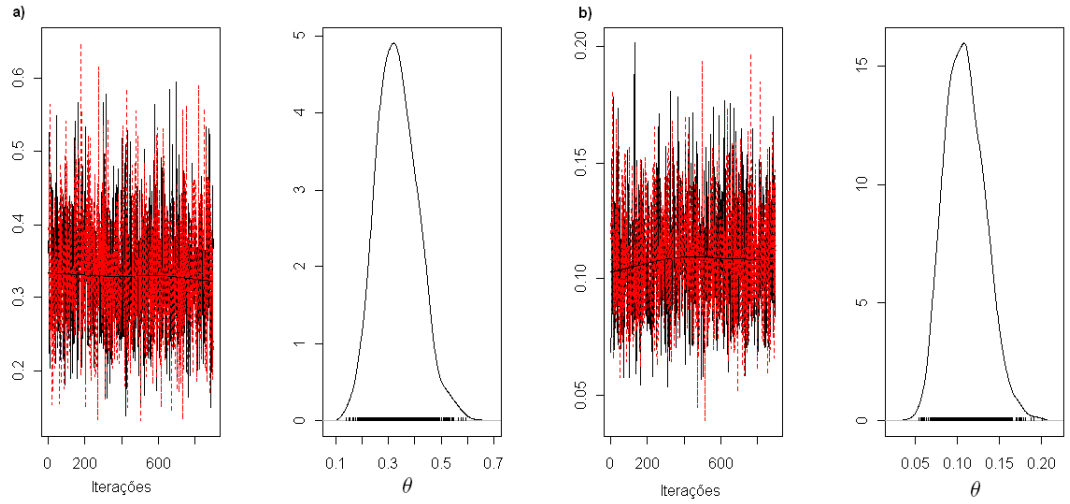

FIGURA 5.7: a) Comportamento das cadeias a distribuição estimada para o parâmetro θ do modelo de Poisson. b) Comportamento das cadeias a distribuição estimada para o parâmetro θ do modelo binomial negativo.

um indício de convergência que comprovou-se através do critério de Gelman-Rubin ($R_{ZIP} = 1$ e $R_{ZIBN} = 1$).

Pela Tabela 5.12 podemos notar que há evidências que o modelo ZIBN e ZIP ajustam melhor os dados. Na Tabela a seguir são apresentados os valores esperados segundo cada modelo.

Através dos valores esperados notamos que o modelo ZIBN e ZIP foram os modelos que mais se aproximaram da frequência observada.

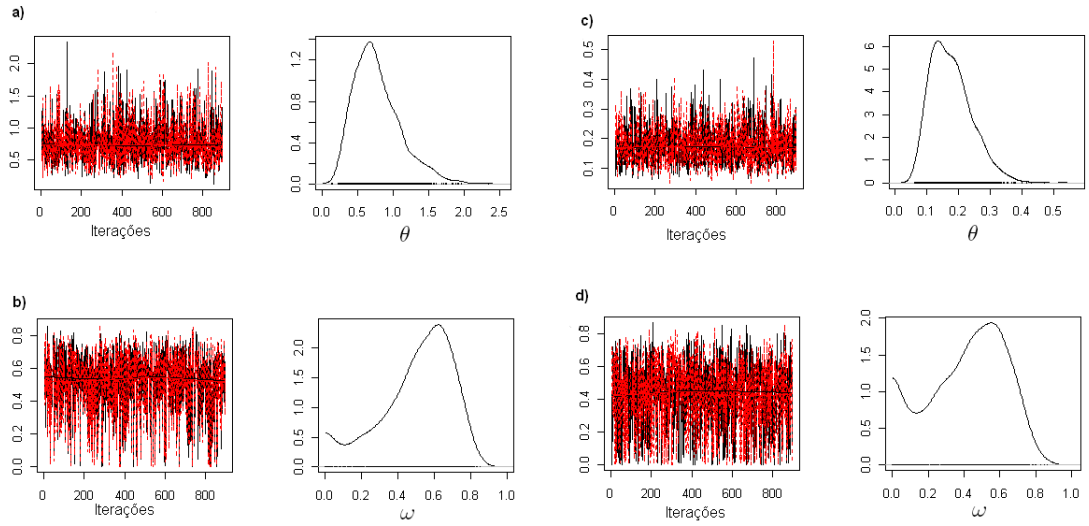

FIGURA 5.8: a) Comportamento das cadeias a distribuição estimada para o parâmetro θ do modelo de ZIP. b) Comportamento das cadeias a distribuição estimada para o parâmetro ω do modelo de ZIP. c) Comportamento das cadeias a distribuição estimada para o parâmetro θ do modelo de ZIBN. d) Comportamento das cadeias a distribuição estimada para o parâmetro ω do modelo de ZIBN.

TABELA 5.12: Valores observados e esperados segundo os modelos de Poisson, ZIP, binomial negativo e ZIBN

Y	Frequência Observada	Poisson	ZIP	BN	ZIBN
0	42	38.6978	39.7264	38.2225	40.2618
1	8	12.8941	9.4521	12.4758	9.1673
2	2	2.1482	3.6608	2.7147	3.2617
3	2	0.2386	0.9452	0.4923	0.9671
Total	54	-	-	-	-

O fator de Bayes do modelo ZIP *versus* Poisson é $B_{10} = 10,2213$, indicando uma evidência forte contra o modelo de Poisson. Ou seja, o modelo ZIP é preferido ao Poisson. A medida de evidência obtida $Ev = 0,04$ também comprova esta evidência.

O fator de Bayes encontrado para o modelo ZIBN *versus* binomial negativo é $B_{32} \approx \infty$, como Datta et al. (2007) ressaltaram o fator de Bayes pode ser

problemático quando a grande maioria dos dados são zero e quando atribuímos distribuições a *priori* impróprias. No entanto, a medida de evidência obtida $Ev = 0.01$, isso significa que o modelo ZIBN ajusta melhor os dados sobre o número de defeitos em carros.

Ghosh et al. (2006) concluiram que o modelo ZIP ajusta melhor os dados sobre defeitos em carros quando comparado ao modelo de Poisson e ao binomial negativo, considerando como critério de seleção o *deviance*. Além disso, ressaltaram que estimativa de $Pr(Y = 0) = 0,791$ da distribuição ZIP é mais próxima da porcentagem observada nos dados que é 0,78. Isto indica um melhor ajuste da distribuição de ZIP quando comparada ao modelo de Poisson e binomial negativa simples. As estimativas encontradas nesta aplicação mostraram-se similares às encontradas em Ghosh et al. (2006). Por isso, podemos tirar as mesmas conclusões a respeito dos modelos ZIP, Poisson e Binomial Negativa.

Se observarmos as estimativas do modelo ZIBN e ZIP podemos concluir que ambos modelos ajustam bem os dados sobre o número de defeitos apresentados por carros. Isto era esperado, pois as empresas que utilizam controle de qualidade buscam chegar a um estado de zero defeitos, logo, o número de defeitos tende a ser o menor possível.

Capítulo 6

Considerações Finais

Esta classe de modelos apresenta propriedades interessantes que foram exploradas neste trabalho. Além disso, esta classe engloba os modelos discretos comumente utilizados. São eles, o Poisson, binomial, binomial negativo dentre outros.

Neste trabalho estudamos os casos particulares da classe - da *distribuição de Poisson zero inflacionada* e a *distribuição binomal negativa zero inflacionada*, fizemos uma revisão bibliográfica sobre o assunto, analisamos as propriedades relevantes e verificamos a aplicabilidade em situações práticas utilizando o contexto bayesiano.

Para ilustrar esta metodologia foram apresentados três exemplos que tinham por objetivo selecionar o modelo que melhor se ajustava aos dados, ou seja, selecionar entre o modelo de Poisson e o modelo ZIP. Para isto, considerou-se como critério de seleção de modelos o fator de Bayes (problemático em algumas situações) e o teste de significância completamente bayesiano proposto por Pereira e Stern em 1999.

O fator de Bayes torna-se difícil de ser obtido quando o valor de n cresce, como ressaltaram Datta et al. (2007), pois tende ao infinito. O mesmo não acontece com a medida de evidência Ev. Além disso, quando os zeros dominam a amostra há a incerteza não temos confiabilidade sobre os parâmetros estimados.

Pode-se concluir que nas nas duas primeiras aplicações o modelo ZIP se ajustou melhor os dados. Já na terceira pode-se verificar através do fator de Bayes e da medida de evidência Ev que tanto o modelo Poisson como o ZIP ajustaram bem os dados.

A última aplicação comparou os modelos de Poisson, binomial negativo, ZIP e ZIBN. Verificou-se que tanto o modelo ZIP quanto o modelo ZIBN ajustaram bem os dados sobre o número de defeitos em carros superando os modelos de Poisson e binomial negativo.

Portanto, pode-se concluir que os modelos pertencentes à classe de distribuições série de potências inflacionadas é uma alternativa eficaz para o modelo discretos comuns quando existem zeros excessivos e, mesmo quando ω é relativamente pequeno. Além disso, verificou-se que Ev é mais fácil de interpretar como medida de evidência que o fator de Bayes.

Além disso, o tema deste trabalho é de grande relevância, visto que tem grande aplicabilidade em diversas situações práticas e nas mais diversas áreas. Logo, este trabalho pode auxiliar pesquisadores na modelagem de dados de contagem com valores inflacionados.

Como propostas futuras para pesquisas sugerimos os seguintes projetos:

- Estudar a classe de distribuições série de potências inflacionadas utilizando aplicações com valores excessivos diferentes de zero.

- Estudar a classe de distribuições série de potências inflacionadas na presença de covariáveis.

- Estudar a probabilidade de cobertura para os parâmetros θ e ω das distribuições série de potências inflacionadas. Verificar se a cobertura média se aproxima da cobertura nominal.

- Avaliar as técnicas bayesianas através de um contexto frequentista, utilizando distribuições a *priori* não informativas.

- Considerar o mesmo estudo utilizando outras distribuições a prioris e assim, avaliar a sensibilidade da análise.

- Avaliar o comportamento desta técnica sob a abordagem clássicaa, por exemplo, considerar o teste de razão de verossimilhança.

Referências Bibliográficas

[1] BORGATTO, A. F. Modelos para Proporções com Superdispersão e Excesso de Zeros: Um Procedimento Bayesiano. **Tese** (Doutorado), Escola Superior de Agricultura Luiz de Queiroz - Univerdidade de São Paulo, Piracicaba, 2004.

[2] DATTA, G. S., BAYARRY, S. e BERGER, J. *Model Selection for Count Data: ZIP It?* **Apresentação no 6^0 Workshop de Inferência Bayesiana Objetiva**. University of Rome, La Sapienza, 2007.

[3] ECHAVARRIA, L. E. O., *Semiparametric Bayesian Count Data Models*. **Dissertação** (Mestrado). Munich, 2004.

[4] GELMAN, A. B. e RUBIN, D. B. *Inference from Iterative Simulation using Multiple*. Statistical Science, v. 7, pp. 457-511, 1992.

[5] GEWEKE, J. *Contemporary Bayesian Econometrics and Statistics*. New Jersey: Wiley & Sons, 2005.

[6] GHOSH, S. K., MUKHOPADHYAY, P. e LU, J. C. *Bayesian Analysis of Zero-Inflated Regression Models*, **Journal of Statistical Planning and Inference**, v. 136, p. 1360-1375, 2006.

[7] GUPTA, P. L., GUPTA, R. C. e TRIPATHI, R. C. *Inflated Modified Power Series Distributions with Applications*, **Communications in Statistics - Theory and Methods**, v. 24, p. 2355-2374, 1995.

[8] MARTIN, T. G., Wintle, B. A., RHODES, J. R., KUHNERT, P. M., FIELD, S. A., LOW-CHOY, S. J., TYRE, A. J. e POSSINGHAM, H. P. *Zero Tolerance Ecology: Improving Ecological Inference by Modelling the Source of Zero Observations*. **Ecology Letters**, p. 12351246, (2005).

[9] MENDES, C. C. Modelos para Dados de Contagem e Aplicações. **Dissertação** (Mestrado), Programa Pós Graduação em Estatística - Universidade de Campinas, Campinas 2007.

[10] MURAT, M. e SZYNAL, D. *Non-Zero Inflated Modified Power Series Distributions*, **Communications in Statistics - Theory and Methods**, v. 27, p. 3047-3064, 1998.

[11] PAULA, G. A. *Modelos de Regressão com Apoio Computacional*. **Instituto de Matemática e Estatística Universidade de São Paulo**, 2004.

[12] PEREIRA, C. A. e STERN, J. M. e WECHSLER, S. *Can a Significance Teste be Genuinely Bayesian*. **International Society for Bayesian Analysis**, v. 2, n. 4, p. 1-22, 2008.

[13] RODRIGUES, J. *Bayesian Analysis of Zero-Inflated Distributions*. **Communications in Statistics - Theory and Methods**, v. 32, n. 2, p. 281-289, 2003.

[14] RODRIGUES, J. *Full Bayesian Significance Test for Zero Inflated Distribution*. **Communications in Statistics - Theory and Methods**, v. 35, p. 1-9, 2006.

[15] SAITO M. Y. *Inferência Bayesiana para Dados Discretos com Excesso de Zero e Uns*. **Dissertação** (Mestrado), Programa Pós Graduação em Estatística - Departamento de Estatística, Universidade Federal de São Carlos, São Carlos, 2005.

[16] SOUZA, A. D. P. *Métodos Aproximados em Modelos Hierárquicos Dinâmicos Bayesianos*. **Tese** (Doutorado), Programa de Engenharia de Produção - COPPE/ Univerdidade Federal do Rio de Janeiro, Rio de Janeiro, 1999.

[17] ROSS, G. J. S. e PREECE, D. A. *The Negative Binomial Distribution*. The Statistician, vol. 34, $n^0$3 pp. 323-335, 1985.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo