UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Distribuições discretas zero-modificadas para modelar dados de contagem zeros faltantes

Isis Fernanda Mascarin

Dissertação de Mestrado do Programa Interinstitucional de Pós-Graduação em Estatística (PIPGEs)

ICMC USP SÃO CARLOS

Isis Fernanda Mascarin

Distribuições discretas zero-modificadas para modelar dados de contagem zeros faltantes

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP e ao Departamento de Estatística – DEs-UFSCar, como parte dos requisitos para obtenção do título de Mestra em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística. *VERSÃO REVISADA*

Área de Concentração: Estatística

Orientadora: Profa. Dra. Katiane Silva Conceição

USP – São Carlos
Fevereiro de 2020

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

M395d

Mascarin, Isis Fernanda
Distribuições discretas zero-modificadas para modelar dados de contagem zeros faltantes / Isis Fernanda Mascarin; orientador Katiane Silva Conceição. -- São Carlos, 2020.
139 p.

Dissertação (Mestrado - Programa Interinstitucional de Pós-graduação em Estatística) -- Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2020.

1. DADOS DE CONTAGEM. 2. DISTRIBUIÇÕES DISCRETAS. 3. SÉRIE DE POTÊNCIAS ZERO-MODIFICADAS. 4. DADOS ZERO-DEFLACIONADOS. 5. ZEROS FALTANTES. I. Conceição, Katiane Silva, orient. II. Título.

Bibliotecários responsáveis pela estrutura de catalogação da publicação de acordo com a AACR2:
Gláucia Maria Saia Cristianini - CRB - 8/4938
Juliana de Souza Moraes - CRB - 8/6176

Isis Fernanda Mascarin

Zero-modified discrete distributions for modeling missing zeros count data

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP and to the Department of Statistics – DEs-UFSCar, in partial fulfillment of the requirements for the degree of the Master Interagency Program Graduate in Statistics. *FINAL VERSION*

Concentration Area: Statistics

Advisor: Profa. Dra. Katiane Silva Conceição

USP – São Carlos
February 2020

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa Interinstitucional de Pós-Graduação em Estatística

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado da candidata Isis Fernanda Mascarin, realizada em 24/01/2020:

Profa. Dra. Katiane Silva Conceição
USP

Prof. Dr. Anderson Luiz Ara Souza
UFBA

Profa. Dra. Roseli Aparecida Leandro
ESALQ/USP

Dedico este trabalho a todos que estiveram à minha volta
com apoio, carinho e paciência.

AGRADECIMENTOS

Agradeço a Deus, que permitiu e acompanhou minha jornada até aqui. À minha família, por todo o suporte em minhas decisões e pelos estímulos para que eu busque sempre o meu melhor. À minha orientadora, Prof.ª Dr.ª Katiane Silva Conceição, por todo o conhecimento partilhado e pela dedicação, paciência e atenção. Aos colegas e amigos, pelo apoio. Ao Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP) e ao Departamento de Estatística (DEs) da Universidade Federal de São Carlos (UFSCar), pelas oportunidades de aprendizado e desenvolvimento.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Pesquisa desenvolvida com o auxílio dos recursos de Computação de Alto Desempenho (HPC)[1] disponibilizados pela Superintendência de Tecnologia da Informação (STI) da USP.

[1] Do inglês: *High-Performance Computing.*

"[...] Imagination is more important than knowledge.
For knowledge is limited,
whereas imagination embraces the entire world,
stimulating progress, giving birth to evolution.
It is, strictly speaking, a real factor in scientific research."
(Albert Einstein)

RESUMO

MASCARIN, ISIS F. **Distribuições discretas zero-modificadas para modelar dados de contagem zeros faltantes**. 2020. 139 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

A análise de dados de contagem ocupa um importante lugar dentro da estatística aplicada, uma vez que muitos problemas reais são expressos em termos de enumerações. Frequentemente, conjuntos de dados de contagem apresentam discrepâncias na frequência da observação zero, que pode ser alta ou baixa, e assim refere-se ao conjunto de dados como zero-inflacionado ou zero-deflacionado, respectivamente. Além disso, existem situações onde a observação zero não ocorre nos conjuntos de dados e, muitas vezes, modelos zero-truncados são inadequadamente considerados, visto que há uma probabilidade positiva (e não nula) para ocorrência de tal evento, embora este não tenha ocorrido. Esta dissertação tem como objetivo principal apresentar o procedimento de estimação dos parâmetros das distribuições zero-modificadas em situações em que a frequência da observação zero nos conjuntos de dados é nula e a probabilidade de ocorrência de tal valor é positiva (zero-deflacionada). A metodologia proposta considera a estimação de zeros faltantes no conjunto de dados formado apenas pelas observações positivas, tal que o conjunto de dados aumentados (adicionando-se os zeros estimados) pode ser explicado por uma distribuição tradicional. Métodos dos momentos e da máxima verossimilhança são considerados para o procedimento de estimação por meio do algoritmo de estimação-maximização. Estudos de simulação e com dados artificiais são utilizados para avaliação das propriedades dos estimadores e estimativas obtidas. Conjuntos de dados reais que apresentam diferentes casos de zero-modificação também são analisados.

Palavras-chave: distribuições zero-modificadas; dados zero-deflacionados; zeros faltantes; estimador de máxima verossimilhança; algoritmo EM.

ABSTRACT

MASCARIN, ISIS F. **Zero-modified discrete distributions for modeling missing zeros count data**. 2020. 139 p. Dissertação (Mestrado em Estatística – Programa Interinstitucional de Pós-Graduação em Estatística) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2020.

The analysis of count data takes an important place in applied statistics, since many real problems are expressed in terms of counts. Frequently, count data sets have discrepancies in the frequency of the zero observation, which may be high or low, and in these cases the set is referred as zero-inflated or zero-deflated, respectively. Besides, there are situations where the zero observation does not occur in the data set, and often zero-truncated models are inadequately considered, since there is a positive probability (and not a null one) for such event, although it has not occurred. The main aim of this dissertation is to present the procedure for parameter estimation of the zero-modified distributions in situations where the frequency of zero observation in the data set is zero and the occurrence probability of this same value is positive (zero-deflated). The proposed methodology considers the estimation of missing zeros in the data set consisting only of positive observations, such that the increased data set (with the estimated zeros included) can be explained by a traditional distribution. Moments and maximum likelihood methods are considered for the estimation procedure using the estimation-maximization algorithm. Simulation and artificial data studies are used to evaluate the properties of the estimators and estimates obtained. Real data sets with different cases of zero-modification are also analyzed.

Keywords: zero-modified distributions, zero-deflated data, missing zeros, maximum likelihood estimator, EM algorithm.

LISTA DE ILUSTRAÇÕES

Figura 1 – Amostras de classificação incorreta de imagens de pássaros, cujo algoritmo considera apenas características em nível de espécie. Nesta figura, as imagens nas caixas azuis são imagens classificadas incorretamente, e os rótulos previstos dessas imagens (em caixas azuis) são fornecidos por uma amostra (imagens em caixas vermelhas) nas linhas correspondentes. 28
Figura 2 – Sumário de procedimentos para modelagem de dados de contagem. 30
Figura 3 – Espaço paramétrico de p em função do parâmetro μ para algumas distribuições da família ZMPS. As regiões indicadas por \\\\ denotam valores de p em que se tem a distribuição ZIPS; e as indicadas por |||| representam valores de p em que a distribuição é a ZDPS. 38
Figura 4 – Comportamento de algumas distribuições da família ZMPS para diferentes valores dos parâmetros p e μ. 40
Figura 5 – Índice de dispersão em função do parâmetro μ para algumas distribuições da família ZMPS. Nos gráficos, — representa $p \approx 0$, $---$ denota $p = 1$ e $\cdots$ indica $p = 1/1 - \pi_{ZMPS}(0;\mu,\phi)$. 44
Figura 6 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ. 70
Figura 7 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ. 71
Figura 8 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ. 72
Figura 9 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ. . . . 73
Figura 10 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ. . . . 74
Figura 11 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ. . . 77
Figura 12 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ. . . 78
Figura 13 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ. . . 79
Figura 14 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ. 80

Figura 15 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ. 81
Figura 16 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ. . . . 84
Figura 17 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ. . . . 85
Figura 18 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ. . . . 86
Figura 19 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ. . . 87
Figura 20 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ. . . 88
Figura 21 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ. 91
Figura 22 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ. 92
Figura 23 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ. 93
Figura 24 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ. 94
Figura 25 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ. 95
Figura 26 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ. 98
Figura 27 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ. 99
Figura 28 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ. 100
Figura 29 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ. 101
Figura 30 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ. 102
Figura 31 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ. 105
Figura 32 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ. 106
Figura 33 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ. 107

Figura 34 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ. 108
Figura 35 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ. 109
Figura 36 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado. 113
Figura 37 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado. 115
Figura 38 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado. 117
Figura 39 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado. 119
Figura 40 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado. 121
Figura 41 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado. 123
Figura 42 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado. 124
Figura 43 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado. 125
Figura 44 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado. 127

LISTA DE TABELAS

Tabela 1 – Distribuições da família PS cujo suporte se inicia em zero. 34
Tabela 2 – Características da distribuição ZMPS. 41
Tabela 3 – Variâncias das distribuições PS cujo suporte se inicia em zero e das ZMPS correspondentes. 41
Tabela 4 – Relação entre p e $I(Y)$ para as distribuições ZMPS e PS associada. 43
Tabela 5 – Estimadores pelo Método dos Momentos das Distribuições ZMPS consideradas. 61
Tabela 6 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero para a distribuição Poisson. 62
Tabela 7 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero para a distribuição Geométrica. 62
Tabela 8 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero para a distribuição Binomial. 63
Tabela 9 – Valores dos parâmetros da distribuição ZDPS utilizados para a geração dos conjuntos de dados. 64
Tabela 10 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero da distribuição Poisson a partir de dados zero-deflacionados. 65
Tabela 11 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero da distribuição Geométrica a partir de dados zero-deflacionados. 65
Tabela 12 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero da distribuição Binomial a partir de dados zero-deflacionados. 66
Tabela 13 – Estimativas de n_0, μ e da probabilidade de zero da distribuição Poisson. . . 68
Tabela 14 – Estimativas de n_0, μ e da probabilidade de zero da distribuição Geométrica. 75
Tabela 15 – Estimativas de n_0, μ e da probabilidade de zero da distribuição Binomial, com $m = 10$. 82
Tabela 16 – Estimativas de n_0, μ e de $\pi_P(0;\mu)$ obtidas a partir de dados provenientes da distribuição Poisson Zero-Deflacionada. 90
Tabela 17 – Estimativas de n_0, μ e de $\pi_G(0;\mu)$ obtidas a partir de dados provenientes da distribuição Geométrica Zero-Deflacionada. 97
Tabela 18 – Estimativas de n_0, μ e de $\pi_G(0;\mu)$ obtidas a partir de dados provenientes da distribuição Binomial Zero-Deflacionada. 104
Tabela 19 – Distribuição de frequência e estatísticas descritivas do número mensal de incidentes de terrorismo internacional nos EUA entre 1968 e 1974. 112

Tabela 20 – Estimativas obtidas para o conjunto de incidentes de terrorismo internacional nos EUA. 112
Tabela 21 – Distribuição de frequência e estatísticas descritivas do número de vitórias do *Barcelona* que antecederam cada vitória do *Real Madrid* no período de março de 1916 a abril de 2014. 114
Tabela 22 – Estimativas obtidas para o número de vitórias do *Barcelona* que antecederam cada vitória do *Real Madrid* no período de março de 1916 a abril de 2014. . 114
Tabela 23 – Distribuição de frequência e estatísticas descritivas do número de ocorrências combinadas dos artigos *"the"*, *"a"* e *"an"* em amostras de 5 palavras de Macaulay (1909). 116
Tabela 24 – Estimativas obtidas para o número de ocorrências combinadas dos artigos *"the"*, *"a"* e *"an"* na amostra citada. 116
Tabela 25 – Distribuição de frequência e estatísticas descritivas do número de meses com variação positiva do dólar até a ocorrência de uma variação negativa, no período entre julho de 1994 e setembro de 2019. 118
Tabela 26 – Estimativas obtidas para o número de usuários de meses com variação positiva do dólar até a ocorrência de uma variação negativa. 118
Tabela 27 – Distribuição de frequência e estatísticas descritivas do número de usuários de metanfetamina com exatamente y_i visitas à clínicas de tratamento. 120
Tabela 28 – Estimativas obtidas para o número de visitas feitas por usuários de metanfetamina à clínicas de tratamento. 121
Tabela 29 – Distribuição de frequência e estatísticas descritivas do número de acidentes no período de 1998 a 2015 em países desenvolvidos e em países em desenvolvimento. 122
Tabela 30 – Estimativas obtidas para os conjuntos de acidentes químicos. 123
Tabela 31 – Distribuição de frequência e estatísticas descritivas do número de acidentes no período de 1998 a 2015. 124
Tabela 32 – Estimativas obtidas para o número de acidentes químicos. 125
Tabela 33 – Distribuição de frequência e estatísticas descritivas do número de acidentes por ano em indústrias petroquímicas no período de 1985 a 2002. 126
Tabela 34 – Estimativas obtidas para o número de acidentes por ano em indústrias petroquímicas entre 1985 e 2002. 127

LISTA DE ABREVIATURAS E SIGLAS

B	Binomial
BNP	*bootstrap* não paramétrico
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil
DEs	Departamento de Estatística
EM	Estimação-Maximização
EMV	estimador de máxima verossimilhança
EUA	Estados Unidos da América
G	Geométrica
GNB	Binomial Negativa Generalizada
GP	Poisson Generalizada
HPC	Computação de Alto Desempenho
ICMC	Instituto de Ciências Matemáticas e de Computação
MARS	Sistema Europeu de Notificação de Acidentes Graves
NB	Binomial Negativa
P	Poisson
PS	Série de Potência
STI	Superintendência de Tecnologia da Informação
UE	União Europeia
UFSCar	Universidade Federal de São Carlos
USP	Universidade de São Paulo
ZDPS	Série de Potência Zero-Deflacionada
ZIPS	Série de Potência Zero-Inflacionada
ZMPS	Série de Potência Zero-Modificada
ZTPS	Série de Potência Zero-Truncada

SUMÁRIO

1 INTRODUÇÃO 25
1.1 Motivação 27
1.2 Objetivo 30

2 CONCEITOS E NOTAÇÕES PRELIMINARES 33
2.1 Distribuições da Família PS 33
2.2 Distribuições da Família ZMPS 36
2.2.1 Caracterização da Distribuição ZMPS *41*
2.2.1.1 Índice de Dispersão *42*
2.2.2 Versão Hurdle da Distribuição ZMPS *43*
2.2.3 Inferência *46*
2.2.3.1 Método dos Momentos *47*
2.2.3.2 Procedimento de Máxima Verossimilhança *48*

3 UM CASO ESPECIAL DA DISTRIBUIÇÃO ZDPS 53
3.1 Distribuição ZMPS *versus* ZTPS 53
3.2 Distribuição ZDPS *versus* PS: Estimação via Algoritmo EM 55

4 ESTUDO COM DADOS SIMULADOS 59
4.1 Estudo de Simulação 59
4.2 Estudo com Dados Artificiais: Aplicações 67

5 APLICAÇÕES: DADOS REAIS 111

6 CONSIDERAÇÕES FINAIS E PROPOSTAS FUTURAS 129

REFERÊNCIAS 131

APÊNDICE A PROCEDIMENTO DE ESTIMAÇÃO BAYESIANO . . 137

CAPÍTULO

1

INTRODUÇÃO

Conjuntos de observações que apresentam-se como contagens são bastante comuns em problemas estatísticos aplicados à diversas áreas da ciência, fazendo-se necessário uma análise adequada, por meio de distribuições discretas, das informações que estes dados podem fornecer. A adoção de modelos, por exemplo a distribuição Poisson — descrita primeiramente em Poisson (1837) como caso limite da distribuição Binomial — para analisar dados de contagens, inclui desde trabalhos mais antigos até técnicas mais recentes, como pode ser visto em trabalhos como Richardson (1944), King (1989), Frome (1983) e Robinson, McCarthy e Smyth (2010).

Apesar de muito popular na análise de dados de contagem, o modelo Poisson pode não ter o melhor ajuste devido à sua suposição de igualdade entre média e variância, característica que é bastante restritiva e facilmente violada. Neste caso, o modelo Binomial Negativo, por exemplo, permite mais flexibilidade na acomodação da variabilidade e é frequentemente utilizado como alternativa (JIANG *et al.*, 2017). Os trabalhos de Paul e Plackett (1978), Gardner, Mulvey e Shaw (1995) e Lawless (1987) são exemplos de estudos sobre inferência utilizando misturas Poisson, especialmente representadas pela Binomial Negativa, de forma a acomodar sobredispersão (isto é, quando observa-se variabilidade maior que a média).

Com o reconhecimento de certas particularidades dos dados, uma ampla classe de distribuições pode ser derivada a partir de séries de potências, proporcionando a existência de outras famílias de distribuições de probabilidade discretas tais como a família de distribuições Série de Potência (PS)[1] e modificações desta (ver Gupta (1974), Consul (1990) e Consul e Famoye (2006)), podendo ser aplicadas em diversos casos (ver também Joshi (1975)).

Uma característica importante, verificada em diversos conjuntos de dados, é a discrepância na frequência da observação k, em que $k = 0, 1, 2, \ldots$. Em situações em que a frequência dos valores observados em uma amostra é diferente da esperada pela distribuição tradicional assumida para a modelagem dos dados, recomenda-se utilizar distribuições k-modificadas, que

[1] Do inglês *Power Series*.

são capazes de alterar as probabilidades — principalmente da observação k (CARVALHO, 2017). Para divergência na frequência da observação zero (isto é, para $k = 0$), a situação é denominada zero-modificada e é de interesse deste trabalho. Quando assume-se probabilidade alta de zero comparada à distribuição original, tem-se o caso de conjunto de dados de contagem zero-inflacionado; ao remover-se a probabilidade de ocorrência de zero, define-se o modelo zero-truncado; e no caso em que tal probabilidade é baixa, observa-se o caso zero-deflacionado. A discrepância que ocorre na classe de contagem de zero pode decorrer, segundo Dietz e Böhning (2000), das seguintes situações:

1. Nem todos os membros da população de estudo são afetados pelo processo, o que causa zero-inflação devido à resposta zero de membros não afetados.
2. Certos problemas inevitáveis decorrentes do processo de amostragem levam a uma chance aumentada ou diminuída de membros de zero-contagem da população serem selecionados na amostra, o que causa zero-inflação ou zero-deflação, respectivamente.
3. Um caso extremo da situação anterior, quando não há nenhuma chance de conseguir uma observação zero na amostra, que neste caso é truncada em zero, e cuja distribuição dos dados amostrais é uma distribuição positiva.
4. Uma combinação dos casos 1 e 3, em que se tem uma subpopulação gerando dados que possuem distribuição positiva, enquanto a população complementar não é afetada, fornecendo observações de contagem zero.

Casos de zero-modificação na distribuição Poisson, por exemplo, foram abordados em trabalhos como o de McKendrick (1926), Umbach (1981) e de Xie e Aickin (1997), no que diz respeito principalmente às situações zero-truncadas; por Lambert (1992) no caso zero-inflacionado e de forma mais geral por Dietz e Böhning (2000). Em Conceição (2013) é apresentada a família de distribuições Série de Potência Zero-Modificada (ZMPS)[2] que inclui um parâmetro de modificação das probabilidades em relação a uma PS tradicional, principalmente da probabilidade da observação zero, a qual apresenta como casos particulares as distribuições Série de Potência Zero-Inflacionada (ZIPS)[3], Série de Potência Zero-Truncada (ZTPS)[4] e Série de Potência Zero-Deflacionada (ZDPS)[5].

Zeros em excesso frequentemente são considerados como uma causa de sobredispersão. Modelos *hurdle* (ou com barreira), cuja proposta inicial foi feita por Cragg (1971), foram desenvolvidos para lidar com dados zero-inflacionados que apresentam subdispersão (variância menor que a média) ou sobredispersão. Capazes de tratar alta ocorrência de zero nos dados, são modelos com duas partes, que combinam os zeros gerados de uma distribuição com resposta

[2] Do inglês: *Zero-Modified Power Series.*
[3] Do inglês: *Zero-Inflated Power Series.*
[4] Do inglês: *Zero-Truncated Power Series.*
[5] Do inglês: *Zero-Deflated Power Series.*

binária (zero ou não zero) com as observações positivas geradas por uma segunda distribuição zero-truncada (CAMERON; TRIVEDI, 1998).

Dados de contagem que contém sobredispersão e excesso de zeros foram analisados nos diferentes trabalhos de Gurmu e Trivedi (1996), Mullahy (1986) e Dalrymple, Hudson e Ford (2003), considerando os modelos Poisson e Binomial Negativo e incluindo suas versões *hurdle*. Entre outras propostas que exploram modificações em distribuições de contagem tradicionais de forma a permitir que as probabilidades de realizações zero e não-zero sejam diferentes das esperadas pelas respectivas distribuições tradicionais, e capazes inclusive de modelar dados que apresentam sobre- ou subdispersão, podem ser citados Chin e Quddus (2003), Hu, Pavlicova e Nunes (2011) e Coly *et al.* (2016).

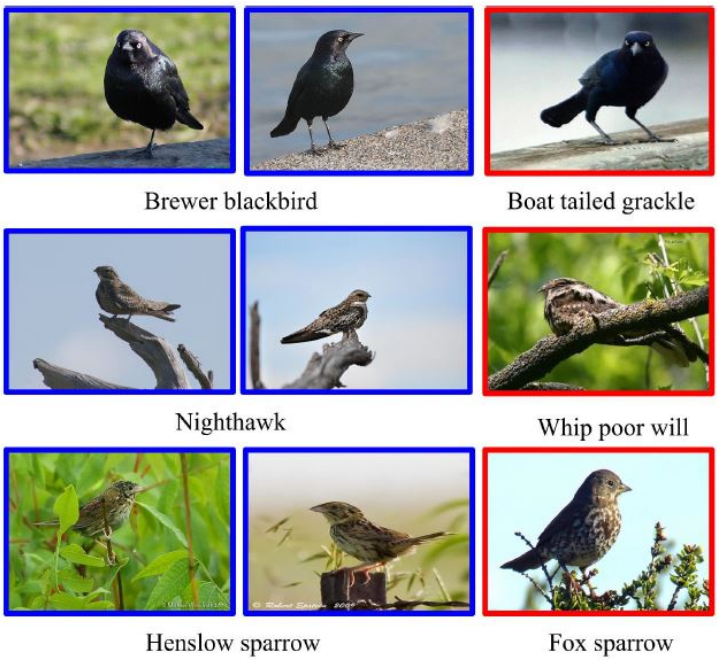
Em particular, existem situações em que a observação zero não ocorre nos conjuntos de dados. Na situação em que há uma probabilidade positiva (e não nula) para ocorrência de tal observação e esta não ocorreu, a suposição de modelos zero-truncados é inadequada. Trabalhar tal cenário é o foco deste trabalho.

1.1 Motivação

Há diversos cenários em que o conjunto de observações considerado constitui-se de dados de contagem. A não observação de zeros, quando este é um valor possível, pode ocorrer, por exemplo, em virtude de problemas de amostragem, tais quais variabilidade e/ou instabilidade dos dados ao longo do processo de amostragem, ou ainda inveracidade e incompletude das observações registradas; ou em consequência de erros não amostrais, que podem decorrer de não observação por dificuldades de cobertura ou coleta (BOLFARINE; BUSSAB, 2005). Outras situações que podem motivar a discrepância nas classes de contagem de zeros são descritas em Dietz e Böhning (2000).

O trabalho de Li, Zhang e Wang (2019) propõe um método computacional de aprendizado de máquina para classificação de imagens e contém uma aplicação que pode ser usada para exemplificar o caso em que a frequência de zero é nula devido a um erro de classificação. Os autores apresentam amostras cujas imagens foram classificadas erroneamente segundo as características (rótulos) utilizadas. Na Figura 1, as contagens das espécies representadas pelas caixas em azul foram positivas devido ao erro de classificação do algoritmo; na verdade as frequências observadas são zero, pois os pássaros pertencem às espécies representadas nas caixas em vermelho da figura. Assim, pode-se concluir a partir da Figura 1 que o conjunto de dados que consiste nas contagens de espécies de pássaros teve apenas frequências positivas, e o zero não apareceu na amostra devido a um erro de classificação.

Figura 1 – Amostras de classificação incorreta de imagens de pássaros, cujo algoritmo considera apenas características em nível de espécie. Nesta figura, as imagens nas caixas azuis são imagens classificadas incorretamente, e os rótulos previstos dessas imagens (em caixas azuis) são fornecidos por uma amostra (imagens em caixas vermelhas) nas linhas correspondentes.



Fonte: Adaptada de Li, Zhang e Wang (2019).

A adoção de procedimentos de estimação em que a probabilidade de zero é considerada inadequadamente como nula, quando esta é positiva, pode gerar estimativas viesadas ou tendenciosas, subestimando ou superestimando sistematicamente o parâmetro de interesse. Quando aplicados à problemas práticos, tal informação equivocada pode levar à decisões com impactos negativos (ver, por exemplo, Volo (2004), Kelly e Haidet (2007) e Rossi e Martins (2010)).

Como exemplo de aplicação para o procedimento desenvolvido neste trabalho, pode-se considerar uma pesquisa cujo interesse é registrar a frequência de observação de certa espécie de animal em uma região específica. É preciso levar em consideração na análise dos dados também a probabilidade positiva de ausência desta espécie, pois ainda que seja registrada somente sua presença, há a possibilidade, entre outros, de migrações sazonais ou captura. A estimação adequada dos parâmetros da distribuição de probabilidade permite fazer inferências corretas, por exemplo, sobre o número médio de exemplares da espécie em toda a região em diferentes estações, ou a identificação de espécies únicas em habitats particulares (veja Risely (2018)).

Outra situação a ser considerada se refere ao processo de contagem da frequência de acidentes em determinado trecho de rodovia. Tal como colocado por Shankar, Milton e Mannering (1997), existe a preocupação na distinção de trechos que são realmente seguros (aproximadamente zero acidentes) daqueles que são perigosos mas ocorre o registro de zero acidentes observados durante o período em consideração que dá origem à amostra, cuja análise

por meio de distribuições usuais pode produzir estimativas viesadas. Em uma amostra na qual somente há registros de frequências maiores que zero acidentes para o trecho de rodovia em questão, levar em consideração a probabilidade de não ocorrência de acidentes na correta modelagem do problema contribui, por exemplo, na predição da frequência de acidentes, que pode ser incorporada em medidas tomadas por parte do sistema de gerenciamento de segurança. Além disso, o "zero-acidente" pode ser consequência da severidade — pequenos acidentes podem sequer ser relatados. Por outro lado, apenas observações positivas de acidentes pode ocorrer por não ser levado em consideração o nível de gravidade e qualquer ocorrência seja registrada. Há também a possibilidade de "quase acidentes" que possam indicar um risco do trecho da estrada, embora nenhum acidente tenha sido verificado; ou seja, o "zero-acidente" pode ser verdadeiramente sua não ocorrência ou um acidente considerado de pouca gravidade ou apenas um "quase-acidente". Assim, a correta modelagem de zero-deflação, especialmente em um conjunto que consiste apenas de observações positivas, permite inferências adequadas para esta situação.

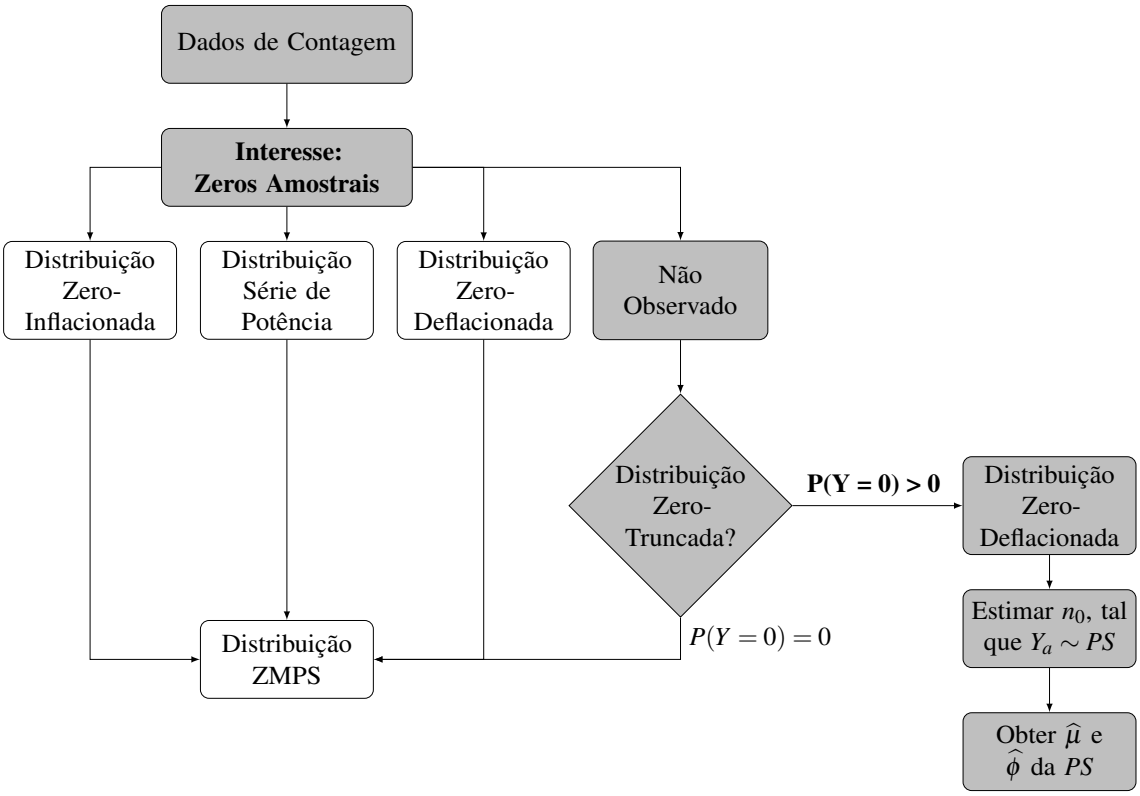
Pode-se citar também o trabalho de Heijden *et al.* (2003), que trata do problema de estimação do tamanho de uma população de interesse com base apenas em dados zero-truncados. É mencionada na referência a aplicação que busca estimar a população de imigrantes ilegais em certo país com base nos registros policiais, que contêm informações sobre aqueles que foram apreendidos pela polícia. Os dados são incompletos, pois os imigrantes ilegais que nunca foram apreendidos não são registrados. O mesmo ocorre com a população que sofre de determinada doença, para a qual a estimação é feita com base nos registros de pacientes que consultaram médicos. Outro exemplo é o que busca estimar o número de pesquisadores trabalhando em determinada área levando-se em conta o número de publicações ou patentes; ou ainda a estimação do número de clientes potenciais de uma rede de hotéis segundo o registro de reservas de clientes. Em todas as aplicações descritas, uma observação zero não pode ser registrada devido à natureza da coleta dos dados e são, desta forma, zero-truncados. Porém, a probabilidade da observação zero é positiva no contexto geral. Nestes casos, considerar modelos inadequados pode levar à subestimação do verdadeiro tamanho populacional. Outros trabalhos que discutem sobre o tema são, por exemplo, Puza *et al.* (2008), Böhning e Heijden (2009) e Heijden, Cruyff e Böhning (2014).

Há diversos outros exemplos que podem ser considerados em que a proposta deste trabalho se faz pertinente, buscando informar de maneira adequada a probabilidade de ocorrência de zero a partir de um conjunto no qual este valor não foi observado, mas poderia ser. A análise adequada para esta situação permite que as inferências realizadas sejam apropriadas nos contextos práticos.

1.2 Objetivo

Este trabalho tem como objetivo apresentar o procedimento de estimação dos parâmetros das distribuições zero-modificadas em situações em que a frequência de observações zero nos conjuntos de dados é nula e a probabilidade de ocorrência de tal valor é positiva (situação zero-deflacionada). A Figura 2 sumariza o procedimento de modelagem em dados de contagem e destaca o caso de interesse deste trabalho.

Figura 2 – Sumário de procedimentos para modelagem de dados de contagem.



Fonte: Elaborada pelo autor.

Foram considerados métodos clássicos para o procedimento de estimação dos parâmetros da distribuição ajustada, tais como o algoritmo Estimação-Maximização (EM)[6], utilizando distribuições zero-modificadas adequadas para cada caso analisado. O conteúdo inclui o estudo das famílias de distribuições PS e ZMPS, cuja definição se dá por modificar, principalmente, a probabilidade de ocorrência de zero das distribuições PS tradicionais (CONCEIÇÃO, 2013; CONCEIÇÃO *et al.*, 2017). Além disso, é apresentada sua versão *hurdle*, a partir da qual foram desenvolvidos os cálculos para obtenção dos estimadores de máxima verossimilhança dos parâmetros da distribuição ZMPS.

Buscando avaliar a eficácia do procedimento de estimação a partir do algoritmo EM, conjuntos de dados contendo observações zeros disponíveis na literatura também foram conside-

[6] Do inglês: *Estimation-Maximization*.

rados. Assim, é possível comparar a(s) estimativa(s) do(s) parâmetro(s), bem como o tipo de zero-modificação existente nos conjuntos de dados.

A estrutura do trabalho é tal qual a relacionada a seguir. No Capítulo 2, lista-se notações e definições preliminares da família de distribuições PS, e que são utilizadas ao longo do trabalho; descrição sobre o modelo zero-modificado proposto por Conceição (2013) (veja também Conceição *et al.* (2017)) também é apresentada, bem como a versão *hurdle* da distribuição ZMPS, suas características e aspectos inferenciais via procedimentos de máxima verossimilhança e método dos momentos. O Capítulo 3 destaca aspectos sobre a distribuição ZDPS e o caso de interesse deste trabalho, além de metodologia de estimação de máxima verossimilhança dos parâmetros deste modelo utilizando o algoritmo EM. Estudos de simulação considerando as distribuições Poisson, Binomial e Geométrica, e aplicação em conjuntos de dados reais são apresentados, respectivamente, nos Capítulos 4 e 5. Por fim, no Capítulo 6, são feitas as considerações quanto a metodologia desenvolvida neste trabalho bem como as conclusões obtidas.

A implementação computacional dos códigos empregados para a realização do estudo aqui apresentado foi feita utilizando-se o *Software R* (R Core Team, 2017). Alguns códigos computacionais foram executados em um computador pessoal com processador Intel(R) Core(TM) i5-5200U CPU @ 2,20GHz com dois núcleos e 8GB de RAM e sistema operacional *Windows* 10 de 64 bits. Os demais códigos foram executados com o auxílio dos recursos de HPC disponibilizados pela STI da USP, via *Cluster Aguia* constituído por 128 servidores físicos com 20 núcleos e 512 GB de RAM, processador Intel(R) Xeon(R) CPU E7- 2870 @ 2,40GHz e um *Filesystem* com 256 TB para arquivos temporários (HPC-STI-USP, 2019). A principal motivação para uso deste último recurso foi a possibilidade de execução de vários códigos em paralelo. Dada a natureza dos códigos, não foram verificadas diferenças significativas entre os tempos de execução para aqueles executadas no computador pessoal e aqueles executadas no *Cluster*.

CAPÍTULO

2

CONCEITOS E NOTAÇÕES PRELIMINARES

2.1 Distribuições da Família PS

A classe de distribuições PS é uma ampla classe de distribuições discretas, construídas a partir de uma série de potência, e inclui diversas das distribuições mais comuns (JOHNSON; KEMP; KOTZ, 2005). Definições e algumas características são introduzidas a seguir.

Seja Y uma variável aleatória com distribuição PS com parâmetro de média $\mu > 0$ e parâmetro de dispersão $\phi \geq 0$. A função massa de probabilidade de Y é definida por:

$$\pi_{PS}(y;\mu,\phi) = \frac{a(y,\phi)g(\mu,\phi)^y}{f(\mu,\phi)}, \qquad y \in \mathscr{A}_s, \tag{2.1}$$

cujo suporte $\mathscr{A}_s$ é o subconjunto dos inteiros $\{s, s+1, s+2, \dots\}$, com $s \geq 0$; $a(y,\phi)$ é uma função positiva, e $f(\mu,\phi)$ e $g(\mu,\phi)$ são funções positivas finitas e duas vezes diferenciáveis, com $f(\mu,\phi) = \sum_{y\in\mathscr{A}_s} a(y,\phi)g(\mu,\phi)^y$ (CORDEIRO; ANDRADE; CASTRO, 2009). A Tabela 1 apresenta as funções $a(y,\phi)$, $f(\mu,\phi)$ e $g(\mu,\phi)$ para algumas distribuições da família PS, cujo suporte se inicia em zero, a saber: distribuições Poisson (P), Poisson Generalizada (GP)[1], Geométrica (G), Binomial (B), Binomial Negativa (NB)[2] e Binomial Negativa Generalizada (GNB)[3].

A média e a variância de uma variável aleatória Y com distribuição PS são dadas por:

$$\mathbb{E}(Y) = \mu = \frac{f'(\mu,\phi)g(\mu,\phi)}{f(\mu,\phi)g'(\mu,\phi)} \qquad e \qquad \mathbb{V}(Y) = \sigma^2 = \frac{g(\mu,\phi)}{g'(\mu,\phi)}, \tag{2.2}$$

em que f' e g' referem-se às derivadas das funções com relação a μ (GUPTA, 1974).

[1] Do inglês: *Generalized Poisson.*
[2] Do inglês: *Negative Binomial.*
[3] Do inglês: *Generalized Negative Binomial.*

Tabela 1 – Distribuições da família PS cujo suporte se inicia em zero.

PS	**Distribuição**	$a(y,\phi)$	$f(\mu,\phi)$	$g(\mu,\phi)$	$\mathscr{A}_0$
P	Poisson	$\frac{1}{y!}$	e^{μ}	μ	$\{0,1,\dots\}$
GP	Poisson Generalizada	$\frac{(1+\phi y)^{y-1}}{y!}$	$e^{\mu(1+\mu\phi)^{-1}}$	$\frac{\mu e^{-\mu\phi(1+\mu\phi)^{-1}}}{1+\mu\phi}$	$\{0,1,\dots\}$
G	Geométrica	1	$1+\mu$	$\frac{\mu}{1+\mu}$	$\{0,1,\dots\}$
B	Binomial	$\binom{m}{y}$	$\left(\frac{m}{m-\mu}\right)^{m}$	$\frac{\mu}{m-\mu}$	$\{0,1,\dots,m\}$
NB	Binomial Negativa	$\frac{\Gamma(\phi+y)}{y!\Gamma(\phi)}$	$\left(\frac{\phi}{\mu+\phi}\right)^{-\phi}$	$\frac{\mu}{\mu+\phi}$	$\{0,1,\dots\}$
GNB	Binomial Negativa Generalizada	$\frac{\nu\Gamma(\phi y+\nu)}{y!\Gamma(\phi y-y+\nu+1)}$	$\left(\frac{\phi-1+\frac{\nu}{\mu}}{\phi+\frac{\nu}{\mu}}\right)^{-\nu}$	$\frac{1}{\phi+\frac{\nu}{\mu}}\left(\frac{\phi-1+\frac{\nu}{\mu}}{\phi+\frac{\nu}{\mu}}\right)^{\phi-1}$	$\{0,1,\dots\}$

Fonte: Conceição (2013).

Pode-se obter os momentos populacionais de ordem r ($r \geq 1$) para uma variável aleatória Y com distribuição PS fazendo-se:

$$\begin{aligned} m_{PS}^{(r)} = \mathbb{E}(Y^r) &= \sum_{y=0}^{\infty} y^r \pi_{PS}(y;\mu,\phi) \\ &= \sum_{y=0}^{\infty} y^r \frac{a(y,\phi)g(\mu,\phi)^y}{f(\mu,\phi)}. \end{aligned} \tag{2.3}$$

Ressalta-se que o momento populacional de ordem 1 é $m_{PS}^{(1)} = \mathbb{E}(Y) = \mu$.

Para obter os momentos recorrentes de ordem $r+1$ (GUPTA, 1974), pode-se reescrever a equação (2.3) como:

$$f(\mu,\phi)m_{PS}^{(r)} = \sum_{y=0}^{\infty} y^r a(y,\phi)g(\mu,\phi)^y.$$

Diferenciando com respeito a μ, segue que:

$$\begin{aligned} f'(\mu,\phi)m_{PS}^{(r)} + f(\mu,\phi)m'^{(r)}_{PS} &= \sum_{y=0}^{\infty} y^r a(y,\phi) y[g(\mu,\phi)]^{y-1} g'(\mu,\phi) \\ &= \sum_{y=0}^{\infty} y^{r+1} a(y,\phi)g(\mu,\phi)^y \frac{g'(\mu,\phi)}{g(\mu,\phi)}; \end{aligned}$$

ou ainda,

$$\frac{g(\mu,\phi)}{g'(\mu,\phi)}\left\{f'(\mu,\phi)m_{PS}^{(r)} + f(\mu,\phi)m'^{(r)}_{PS}\right\} = \sum_{y=0}^{\infty} y^{r+1} a(y,\phi)g(\mu,\phi)^y. \tag{2.4}$$

Ao multiplicar e dividir o segundo termo da equação (2.4) por $f(\mu,\phi)$, tem-se que:

$$\frac{g(\mu,\phi)}{g'(\mu,\phi)}\left\{f'(\mu,\phi)m_{PS}^{(r)} + f(\mu,\phi)m'^{(r)}_{PS}\right\} = \sum_{y=0}^{\infty} y^{r+1} \frac{a(y,\phi)g(\mu,\phi)^y}{f(\mu,\phi)} f(\mu,\phi);$$

isto é,

$$\frac{g(\mu,\phi)}{g'(\mu,\phi)f(\mu,\phi)}\left\{f'(\mu,\phi)m_{PS}^{(r)} + f(\mu,\phi)m'^{(r)}_{PS}\right\} = \sum_{y=0}^{\infty} y^{r+1} \frac{a(y,\phi)g(\mu,\phi)^y}{f(\mu,\phi)},$$

ou seja,

$$\begin{aligned}\mathbb{E}(Y^{r+1}) &= \frac{g(\mu,\phi)}{g'(\mu,\phi)f(\mu,\phi)}\left\{f'(\mu,\phi)m_{PS}^{(r)} + f(\mu,\phi)m'^{(r)}_{PS}\right\} \\ &= m_{PS}^{(1)}m_{PS}^{(r)} + \frac{g(\mu,\phi)}{g'(\mu,\phi)}\frac{d}{d\mu}m_{PS}^{(r)}. \end{aligned} \tag{2.5}$$

Usando-se as equações (2.2), pode-se reescrever os momentos recorrentes de ordem $r+1$ para uma variável Y com distribuição PS como:

$$m_{PS}^{(r+1)} = \mu m_{PS}^{(r)} + \sigma^2 m'^{(r)}_{PS}.$$

A versão zero-truncada da distribuição PS apresentada pela equação (2.1), definida como distribuição ZTPS (CONCEIÇÃO, 2013; CONCEIÇÃO *et al.*, 2017), é dada por:

$$\begin{aligned}\pi_{ZTPS}(y;\mu,\phi) =& \frac{\pi_{PS}(y;\mu,\phi)}{1-\pi_{PS}(0;\mu,\phi)}\{1-\mathbf{I}(y)\} \\ =& \frac{a(y,\phi)g(\mu,\phi)^y}{f(\mu,\phi)-a(0,\phi)}\{1-\mathbf{I}(y)\} \\ =& \frac{a(y,\phi)g(\mu,\phi)^y}{f(\mu,\phi)-1}\{1-\mathbf{I}(y)\},\end{aligned}$$

em que $\mathbf{I}(y)$ é a função indicadora tal que

$$\mathbf{I}(y) = \begin{cases} 1, & \text{se } y = 0 \\ 0, & \text{caso contrário.} \end{cases} \tag{2.6}$$

Nota-se que a distribuição ZTPS acarreta em probabilidade nula quando $y = 0$.

A média e a variância da variável aleatória Y com distribuição ZTPS são dadas por:

$$\begin{aligned}\mu_{ZTPS} =& \frac{\mu}{1-\pi_{PS}(0;\mu,\phi)} \\ =& \frac{f'(\mu,\phi)g(\mu,\phi)}{g'(\mu,\phi)(f(\mu,\phi)-1)} \\ =& \mu\frac{f(\mu,\phi)}{f(\mu,\phi)-1}\end{aligned}$$

e

$$\begin{aligned}\sigma^2_{ZTPS} =& \frac{g(\mu,\phi)}{g'(\mu,\phi)(f(\mu,\phi)-1)}\left[f(\mu,\phi) - \frac{g(\mu,\phi)}{g'(\mu,\phi)}\left(\frac{(f'(\mu,\phi))^2}{f(\mu,\phi)(f(\mu,\phi)-1)}\right)\right] \\ =& \frac{\sigma^2}{f(\mu,\phi)-1}\left(f(\mu,\phi) - \mu\frac{f'(\mu,\phi)}{f(\mu,\phi)-1}\right).\end{aligned}$$

No cenário de poucos ou muitos zeros na amostra (situações zero-deflacionada e zero-inflacionada, respectivamente), o uso da distribuição PS tradicional não é apropriado. Neste contexto, distribuições discretas mais gerais têm sido propostas. Conceição (2013) (ver também Conceição *et al.* (2017)) propôs a família de distribuições ZMPS, que é uma ampla família de distribuições e bastante flexível, a qual será apresentada a seguir.

2.2 Distribuições da Família ZMPS

Há situações em que a frequência observada de zeros em uma amostra é muito diferente (alta ou baixa) da esperada ao considerar uma distribuição discreta tradicional para explicar o comportamento destes dados. Quando o conjunto de dados apresenta alta (baixa) frequência de zero, este é dito ser zero-inflacionado (zero-deflacionado) e na situação em que o conjunto de dados não contém a observação zero, este pode ser zero-truncado ou até mesmo zero-deflacionado. Em tais situações, recomenda-se a utilização das distribuições zero-modificadas, capazes de produzir uma análise mais refinada das observações e explicar corretamente a ocorrência de zero. Em outras palavras, a família de distribuições ZMPS é capaz de acomodar de forma adequada dados de contagem com qualquer tipo de zero-modificação (ver Conceição (2013); Conceição *et al.* (2017)). Sua definição e características são apresentadas a seguir.

Considere Y uma variável aleatória definida nos inteiros não negativos, cuja distribuição pertence à família de distribuições ZMPS, com modificação principalmente na probabilidade de zero da distribuição PS tradicional. Então Y tem função massa de probabilidade definida por:

$$\pi_{ZMPS}(y;\mu,\phi,p) = (1-p)\mathbf{I}(y) + p\pi_{PS}(y;\mu,\phi), \qquad y \in \mathscr{A}_0,$$

em que o suporte $\mathscr{A}_0$ é o subconjunto dos inteiros $\{0,1,2,\dots\}$, p é o parâmetro de modificação das probabilidades da distribuição PS tradicional, satisfazendo a restrição

$$0 \leq p \leq \frac{1}{1-\pi_{PS}(0;\mu,\phi)},$$

e $\mathbf{I}(y)$ é a função indicadora tal qual definida pela equação (2.6).

A principal diferença entre a distribuição ZMPS e uma distribuição de mistura tradicional reside no fato de que o parâmetro p pode assumir valores maiores que 1 na distribuição ZMPS. Além disso, a ZMPS apresenta flexibilidade, dado que diferentes valores de p resultam em diferentes distribuições, ao levar-se em consideração a proporção de zeros "faltantes" ou "adicionais":

$$\begin{aligned}\pi_{ZMPS}(0;\mu,\phi,p) - \pi_{PS}(0;\mu,\phi) &= 1-p+p\pi_{PS}(0;\mu,\phi) - \pi_{PS}(0;\mu,\phi) \\ &= (1-p)(1-\pi_{PS}(0;\mu,\phi)). \end{aligned} \tag{2.7}$$

A partir da equação (2.7), para os possíveis valores de p, tem-se as seguintes situações:

- $p = 0$: $\pi_{ZMPS}(0;\mu,\phi,p) = 1$, e então $\pi_{ZMPS}(y;\mu,\phi,p)$ é uma distribuição degenerada no ponto com toda a massa no zero;
- $0 < p < 1$: $(1-p)(1-\pi_{PS}(0;\mu,\phi)) > 0$, e então $\pi_{ZMPS}(y;\mu,\phi,p)$ é uma distribuição ZIPS;

- $p = 1$: $\pi_{ZMPS}(0;\mu,\phi,p) = \pi_{PS}(0;\mu,\phi)$, e então $\pi_{ZMPS}(y;\mu,\phi,p)$ é uma distribuição PS usual;
- $1 < p < \frac{1}{1-\pi_{PS}(0;\mu,\phi)}$: $(1-p)(1-\pi_{PS}(0;\mu,\phi)) < 0$, e então $\pi_{ZMPS}(y;\mu,\phi,p)$ é uma distribuição ZDPS;
- $p = \frac{1}{1-\pi_{PS}(0;\mu,\phi)}$: $\pi_{ZMPS}(0;\mu,\phi,p) = 0$, e então $\pi_{ZMPS}(y;\mu,\phi,p)$ é uma distribuição ZTPS.

A Figura 3 ilustra os valores assumidos por p para diferentes valores de μ considerando as distribuições ZMP, ZMGP, ZMG, ZMB, ZMNB e ZMGNB[4], destacando os subconjuntos do espaço paramétrico em que as distribuições zero-deflacionada e zero-inflacionada estão definidas. Verifica-se que à medida em que o valor do parâmetro μ aumenta, o valor do parâmetro de modificação p diminui. Ou seja, para valores grandes de μ o limite superior de p se aproxima de 1 — que é a sua PS associada (CONCEIÇÃO, 2013).

Ressalta-se que, ao serem considerados outros valores para os parâmetros, principalmente para as distribuições que possuem também o parâmetro de dispersão ϕ, tem-se diferentes comportamentos do limite superior de p. Assim, no caso de uma distribuição ZDPS a escolha do valor do parâmetro p requer certo cuidado, devendo-se respeitar seu limite superior — isto é, no caso da zero-deflação, $1 < p < \frac{1}{1-\pi_{PS}(0;\mu,\phi)}$.

Um resultado que merece destaque e que não foi apresentado por Conceição *et al.* (2017) é a Proposição 1 a seguir.

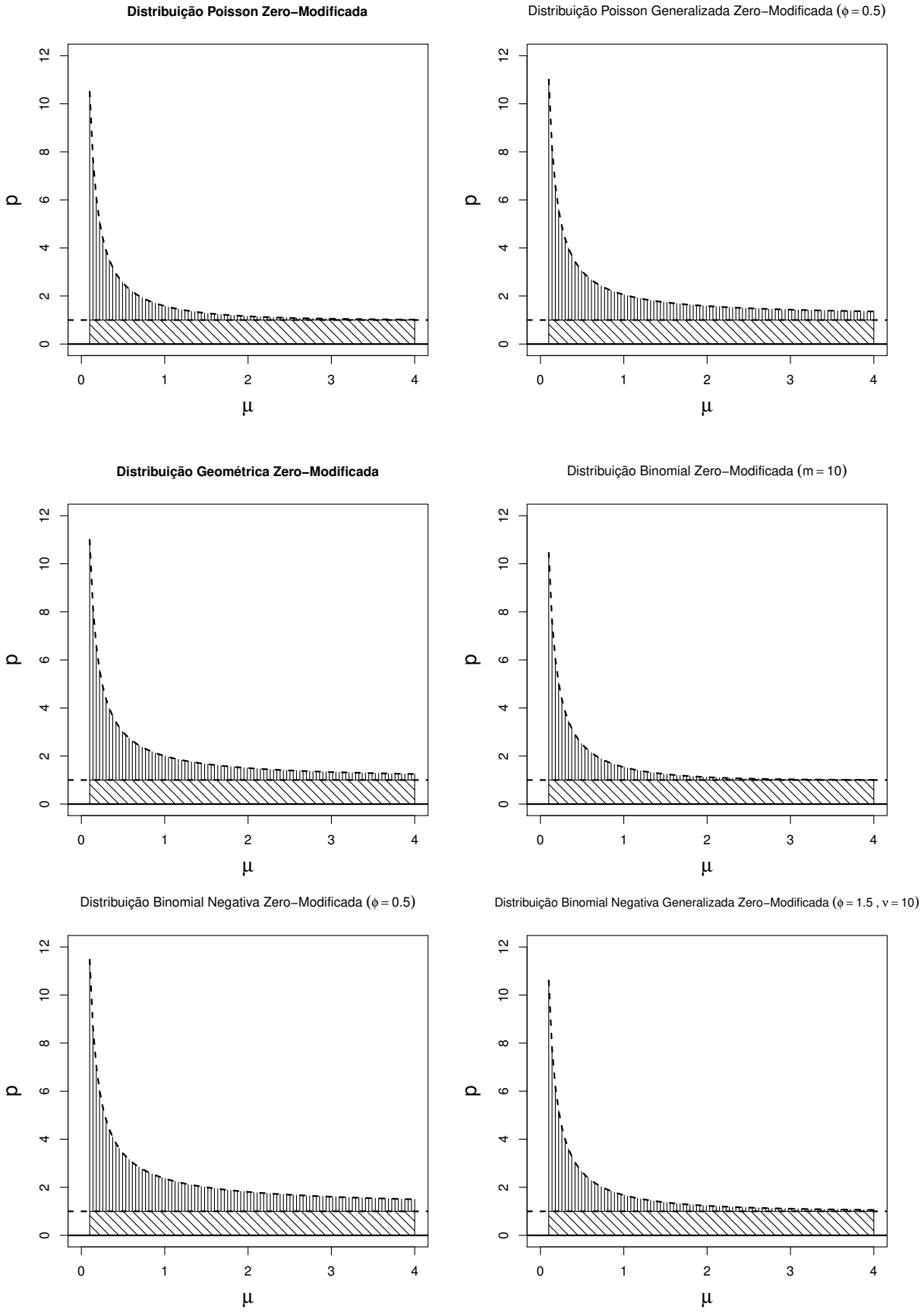
Proposição 1. Considere a distribuição ZMPS e sua distribuição PS associada. Para todo $y \in \mathscr{A}_s$, as probabilidades $\pi_{ZMPS}(y;\mu,\phi,p)$ e $\pi_{PS}(y;\mu,\phi)$ satisfazem as seguintes propriedades:

(**a**) Se $\pi_{ZMPS}(y;\mu,\phi,p)$ é zero-inflacionada (isto é, $0 < p < 1$), então $\pi_{ZMPS}(y;\mu,\phi,p) > \pi_{PS}(y;\mu,\phi)$ somente para $y = 0$ e $\pi_{ZMPS}(y;\mu,\phi,p) < \pi_{PS}(y;\mu,\phi)$ para todo $y \neq 0$.

(**b**) Se $\pi_{ZMPS}(y;\mu,\phi,p)$ é zero-deflacionada (isto é, $1 < p < 1 \,/\, (1-\pi_{PS}(0;\mu,\phi))$), então $\pi_{ZMPS}(y;\mu,\phi,p) < \pi_{PS}(y;\mu,\phi)$ somente para $y = 0$ e $\pi_{ZMPS}(y;\mu,\phi,p) > \pi_{PS}(y;\mu,\phi)$ para todo $y \neq 0$.

[4] Às siglas das distribuições da família PS (ver Tabela 1) acrescenta-se ZM para indicar que trata-se da família ZMPS, mais geral.

Figura 3 – Espaço paramétrico de p em função do parâmetro μ para algumas distribuições da família ZMPS. As regiões indicadas por \\\\ denotam valores de p em que se tem a distribuição ZIPS; e as indicadas por |||| representam valores de p em que a distribuição é a ZDPS.



Fonte: Conceição (2013).

Demonstração. Considere a diferença entre as probabilidades:

$$\begin{aligned}\pi_{ZMPS}(y;\mu,\phi,p) - \pi_{PS}(y;\mu,\phi) &= (1-p)\mathbf{I}(y) + p\pi_{PS}(y;\mu,\phi) - \pi_{PS}(y;\mu,\phi) \\ &= (1-p)\{\mathbf{I}(y) - \pi_{PS}(y;\mu,\phi)\}. \end{aligned} \tag{2.8}$$

Por meio da equação (2.8), demonstrar-se que:

(**a**) Se $\pi_{ZMPS}(y;\mu,\phi,p)$ é zero-inflacionada (isto é, $0 < p < 1$), então para $y = 0$ tem-se

$$\pi_{ZMPS}(0;\mu,\phi,p) - \pi_{PS}(0;\mu,\phi) > 0 \Rightarrow \pi_{ZMPS}(0;\mu,\phi,p) > \pi_{PS}(0;\mu,\phi).$$

Por outro lado, para todo $y \neq 0$,

$$\pi_{ZMPS}(y;\mu,\phi,p) - \pi_{PS}(y;\mu,\phi) < 0 \Rightarrow \pi_{ZMPS}(y;\mu,\phi,p) < \pi_{PS}(y;\mu,\phi).$$

(**b**) Se $\pi_{ZMPS}(y;\mu,\phi,p)$ é zero-deflacionada (isto é, $1 < p < 1/(1-\pi_{PS}(0;\mu,\phi))$), então para $y = 0$ tem-se

$$\pi_{ZMPS}(0;\mu,\phi,p) - \pi_{PS}(0;\mu,\phi) < 0 \Rightarrow \pi_{ZMPS}(0;\mu,\phi,p) < \pi_{PS}(0;\mu,\phi).$$

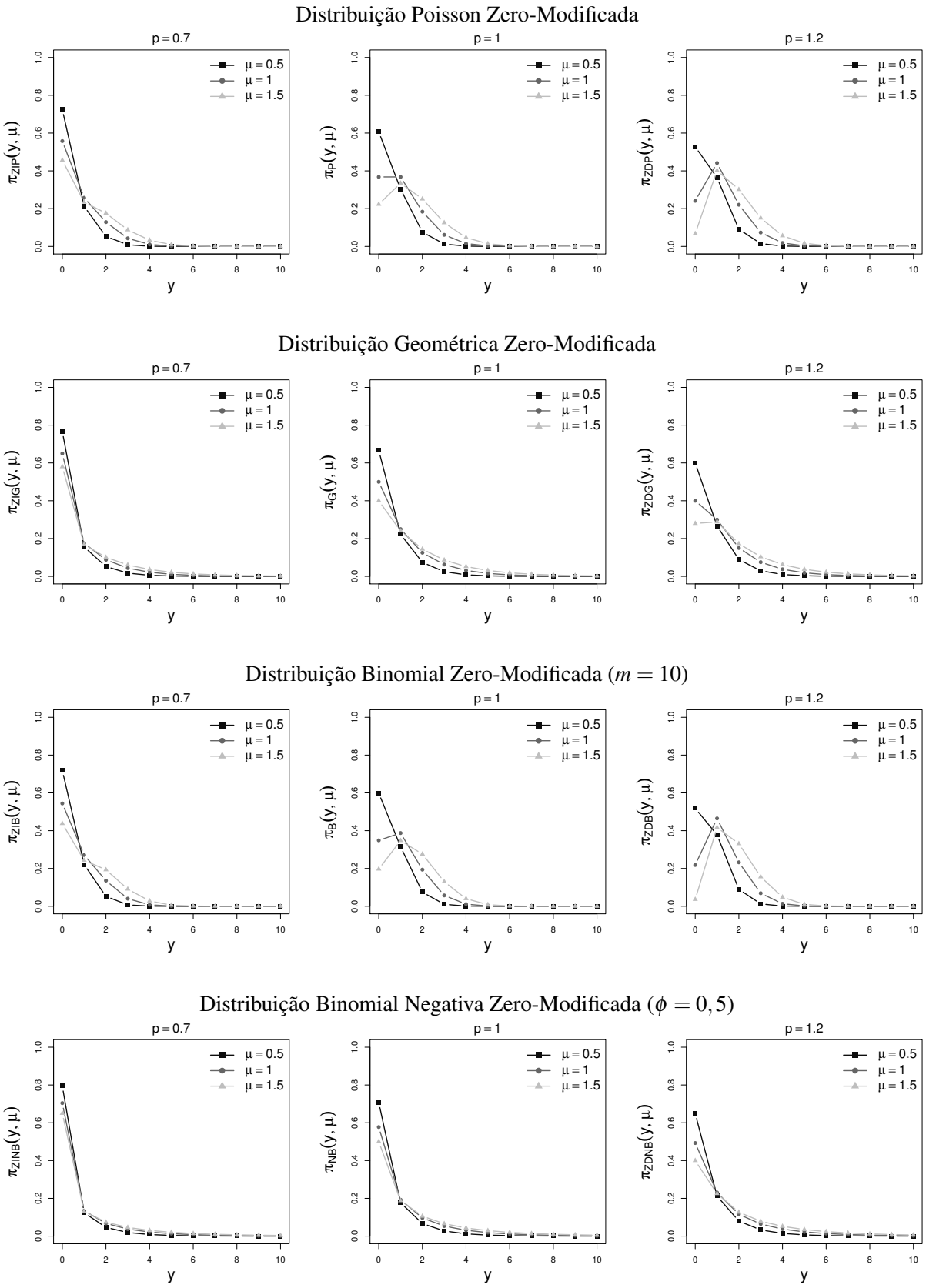
Por outro lado, para todo $y \neq 0$,

$$\pi_{ZMPS}(y;\mu,\phi,p) - \pi_{PS}(y;\mu,\phi) > 0 \Rightarrow \pi_{ZMPS}(y;\mu,\phi,p) > \pi_{PS}(y;\mu,\phi).$$

□

A Figura 4 permite verificar visualmente a Proposição 1, ilustrando o comportamento da função massa de probabilidade de algumas distribuições da família ZMPS (ZMP, ZMG, ZMB e ZMNB) para diferentes valores do parâmetro μ e valores de p que caracterizam a distribuição como zero-inflacionada, tradicional e zero-deflacionada, respectivamente. Verifica-se que no caso de zero-inflação ($p = 0.7$), as probabilidades calculadas com as distribuições ZMPS são maiores que as calculadas com as respectivas distribuições PS tradicionais somente para o valor zero e menores para os demais valores; no caso de zero-deflação ($p = 1.2$), a menor probabilidade das distribuições ZMPS em relação às calculadas com as respectivas distribuições PS tradicionais ocorre somente em zero, apresentando probabilidades maiores para os demais valores. Além disso, nota-se que quanto menor o valor do parâmetro μ, maior a probabilidade de zero, em qualquer situação.

Figura 4 – Comportamento de algumas distribuições da família ZMPS para diferentes valores dos parâmetros p e μ.



Fonte: Elaborada pelo autor.

2.2.1 Caracterização da Distribuição ZMPS

Para a classe ZMPS, as relações que caracterizam essa família de distribuições em termos da distribuição PS tradicional são listadas na Tabela 2. A demonstração de cada resultado pode ser encontrada em Conceição (2013) e Conceição *et al.* (2017).

Tabela 2 – Características da distribuição ZMPS.

Característica	**Função**
Função de Distribuição	$F_{ZMPS}(k) = 1 - p(1 - F_{PS}(k))$
Função Geradora de Probabilidade	$G_{ZMPS}(z) = 1 - p(1 - G_{PS}(z))$
Função Característica	$\varphi_{ZMPS}(t) = 1 - p(1 - \varphi_{PS}(t))$
Função Geradora de Momentos	$\mathbb{M}_{ZMPS}(t) = 1 - p(1 - \mathbb{M}_{PS}(t))$
r-ésimo Momento Populacional	$m^{(r)}_{ZMPS} = pm^{(r)}_{PS}$
Média	$\mu_{ZMPS} = p\mu$
Variância	$\sigma^2_{ZMPS} = p\sigma^2 + (1-p)\mu^2$

Fonte: Elaborada pelo autor.

Cita-se neste trabalho também as relações entre a função geradora de momentos ($\mathbb{M}_{ZMPS}$) e a função geradora de probabilidade (G_{ZMPS}), e entre a função geradora de cumulantes (Ψ_{ZMPS}) e a função geradora de momentos, dadas respectivamente por:

$$\mathbb{M}_{ZMPS}(t) = E(e^{tY}) = G_{ZMPS}(e^t) \qquad \text{e} \qquad \Psi_{ZMPS}(t) = \log(\mathbb{M}_{ZMPS}(t)) = \log(1 - p(1 - \mathbb{M}_{PS}(t))).$$

Além disso, o r-ésimo cumulante da classe ZMPS é obtido a partir do cálculo da r-ésima derivada de $\Psi_{ZMPS}(t)$ e fazendo-se $t = 0$.

A Tabela 3 apresenta as variâncias das distribuições da família PS e ZMPS.

Tabela 3 – Variâncias das distribuições PS cujo suporte se inicia em zero e das ZMPS correspondentes.

PS	**Distribuição**	σ^2	σ^2_{ZMPS}
P	Poisson	μ	$p\mu[1 + \mu(1-p)]$
GP	Poisson Generalizada	$\mu(1+\phi\mu)^2$	$p\mu[(1+\phi\mu)^2 + \mu(1-p)]$
G	Geométrica	$\mu(1+\mu)$	$p\mu[(1+\mu) + \mu(1-p)]$
B	Binomial	$\mu\left(1 - \frac{\mu}{m}\right)$	$p\mu\left[\left(1 - \frac{\mu}{m}\right) + \mu(1-p)\right]$
NB	Binomial Negativa	$\mu\left(1 + \frac{\mu}{\phi}\right)$	$p\mu\left[\left(1 + \frac{\mu}{\phi}\right) + \mu(1-p)\right]$
GNB	Binomial Negativa Generalizada	$\mu\left(1 + \frac{\phi\mu}{\nu}\right)\left(1 + \frac{\mu(\phi-1)}{\nu}\right)$	$p\mu\left[\left(1 + \frac{\phi\mu}{\nu}\right)\left(1 + \frac{\mu(\phi-1)}{\nu}\right) + \mu(1-p)\right]$

Fonte: Conceição (2013).

Destaca-se que para uma variável aleatória Y com distribuição ZMPS, tem-se que os momentos populacionais de ordem r ($r \geq 1$) obtidos por

$$\begin{aligned} m_{ZMPS}^{(r)} = \mathbb{E}(Y^r) &= \sum_{y=0}^{\infty} y^r \pi_{ZMPS}(y;\mu,\phi) \\ &= p m_{PS}^{(r)}, \end{aligned}$$

podem ser utilizados para obter os momentos recorrentes de ordem $r+1$. Assim, usando-se a equação (2.5), os momentos recorrentes de ordem $r+1$ de uma variável Y com distribuição ZMPS são dados por:

$$\begin{aligned} m_{ZMPS}^{(r+1)} = p m_{PS}^{(r+1)} &= p\left\{ m_{PS}^{(1)} m_{PS}^{(r)} + \frac{g(\mu,\phi)}{g'(\mu,\phi)} \frac{d}{d\mu} m_{PS}^{(r)} \right\} \\ &= p\left\{ \mu m_{PS}^{(r)} + \sigma^2 m'^{(r)}_{PS} \right\}. \end{aligned}$$

2.2.1.1 Índice de Dispersão

Zeros em excesso frequentemente são considerados como uma causa de sobredispersão (isto é, situação em que a variância excede a média dos dados). Na verdade, a abundância de zeros reduz a média de um conjunto de observações, causando inflação do índice de dispersão, o que resulta em chance maior de sobredispersão, mas não necessariamente garante que o conjunto é realmente sobredisperso. Assim, é preciso considerar distribuições flexíveis capazes não somente de lidar com excesso de zeros, mas também identificar os cenários de sobre ou subdispersão (SELLERS; RAIM, 2016).

O índice de dispersão é uma medida comumente utilizada para quantificar o quanto os dados estão dispersos ou concentrados em comparação a um modelo estatístico tradicional, e é dado por:

$$I(Y) = \frac{\mathbb{V}(Y)}{\mathbb{E}(Y)},$$

em que se $I(Y) = 0$, diz-se que não há dispersão no conjunto de observações; no caso em que $I(Y) = 1$, os dados satisfazem a suposição de equidispersão; e se $I(Y) < 1$ ou $I(Y) > 1$, fala-se na situação de subdispersão ou sobredispersão dos dados, respectivamente.

O índice de dispersão de uma variável aleatória Y com distribuição ZMPS é dado por:

$$\begin{aligned} I_{ZMPS} &= \frac{\sigma^2_{ZMPS}}{\mu_{ZMPS}} \\ &= \frac{\sigma^2}{\mu} + \mu(1-p) \\ &= I_{PS} + \mu(1-p), \end{aligned}$$

em que I_{PS} é o índice de dispersão para dados de uma distribuição PS tradicional e o termo $\mu(1-p)$ representa subdispersão ou sobredispersão causada por modificação na frequência de

zeros, relativamente à distribuição PS tradicional associada. Quando $I_{ZMPS} > 1$ (ou < 1), os dados exibem sobredispersão (ou subdispersão).

Para os diferentes valores que o parâmetro de modificação p assume, pode-se analisar o índice de dispersão com respeito à inflação ou deflação de zero nos dados, e assim, as distribuições ZMPS pode levar em consideração subdispersão ou sobredispersão além daquela causada por inflação ou deflação de zero. A Tabela 4 relaciona os casos da distribuição PS e as relações de seu índice de dispersão comparado ao da classe ZMPS.

Tabela 4 – Relação entre p e $I(Y)$ para as distribuições ZMPS e PS associada.

Distribuição PS	**Parâmetro de modificação**	**Índice de dispersão**
Zero-deflacionada	$1 < p < (1 - \pi_{PS}(0;\mu,\phi))^{-1}$	$I_{ZMPS} < I_{PS}$
Padrão	$p = 1$	$I_{ZMPS} = I_{PS}$
Zero-inflacionada	$0 < p < 1$	$I_{ZMPS} > I_{PS}$

Fonte: Adaptada de Conceição (2013).

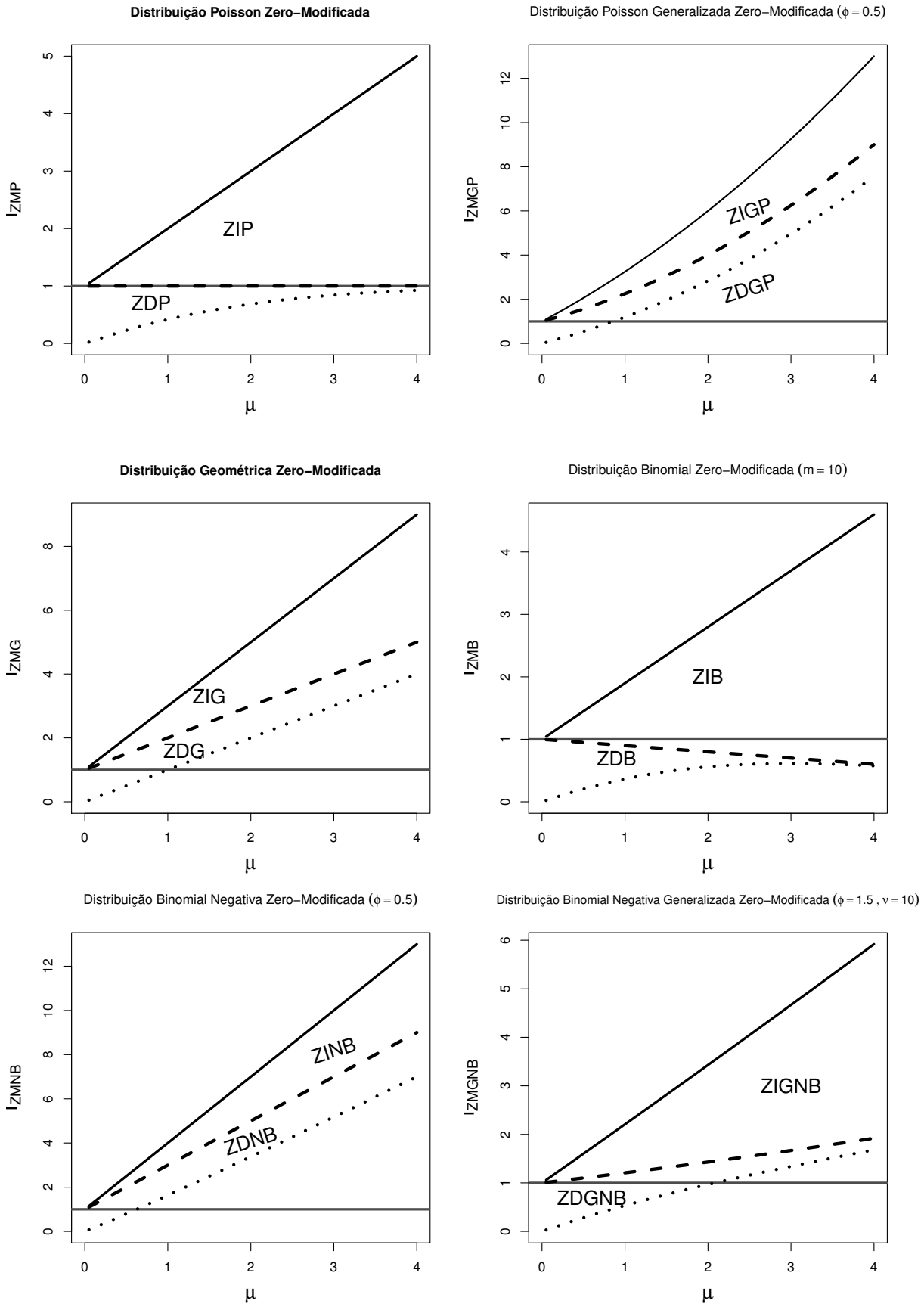
A Figura 5 apresenta os gráficos com o comportamento de I_{ZMPS} para diferentes valores de μ e p entre $[0, 1/(1 - \pi_{ZMPS}(0;\mu,\phi))]$ das distribuições ZMP, ZMGP, ZMG, ZMB, ZMNB e ZMGNB da família ZMPS, destacando os subconjuntos do espaço paramétrico em que as distribuições zero-deflacionada e zero-inflacionada estão definidas.

Verifica-se que, apenas para a distribuição ZMP, subdispersão implica zero-deflação, assim como sobredispersão implica zero-inflação. Para os valores de parâmetros considerados, no caso da distribuição ZMNB há situações nas quais tem-se sobredispersão e zero-deflação concomitantemente, algo que se repete também com as distribuições ZMGP, ZMG e ZMGNB. Na distribuição ZMB, há subdispersão para as situações que caracterizam zero-deflação e também para situações de zero-inflação.

2.2.2 Versão *Hurdle* da Distribuição ZMPS

Modelos *hurdle* foram desenvolvidos para lidar com dados zero-inflacionados que apresentam subdispersão ou sobredispersão. Capazes de tratar alta ocorrência de zero nos dados, diferem de um modelo zero-inflacionado na maneira como interpretam e analisam as observações zero (HU; PAVLICOVA; NUNES, 2011). A versão *hurdle* de uma distribuição é uma forma de abordar dados de contagem utilizando um modelo com duas partes. A primeira parte é responsável por modelar as observações zero e a segunda parte é responsável por modelar as observação positivas ($y > 0$). Apesar da semelhança entre os modelos *hurdle* e os modelos zero-inflacionados, estes últimos são apresentados na literatura com a suposição de que as observações zero são provenientes de dois processos. O primeiro gera apenas observações zero com certa probabilidade π_i, os chamados "zeros amostrais" (decorrentes do processo de amostragem); o segundo processo, por sua vez, fornece contagens zero com probabilidade $1 - \pi_i$ ("zeros estruturais") (KASSAHUN *et al.*, 2014).

Figura 5 – Índice de dispersão em função do parâmetro μ para algumas distribuições da família ZMPS. Nos gráficos, — representa $p \approx 0$, − − − denota $p = 1$ e $\cdots$ indica $p = 1/1 - \pi_{ZMPS}(0; \mu, \phi)$.



Fonte: Conceição (2013).

Considerando a parametrização $\omega = p(1-\pi_{PS}(0;\mu,\phi))$, a distribuição ZMPS pode ser representada como uma distribuição *hurdle*, com função massa de probabilidade dada por:

$$\pi_{ZMPS}(y;\mu,\phi,\omega) = (1-\omega)\mathbf{I}(y) + \omega\pi_{ZTPS}(y;\mu,\phi), y \in \mathscr{A}_0, \tag{2.9}$$

em que $0 \leq \omega \leq 1$ e esta versão será denotada por ZMPS(μ,ϕ,ω). Para mais detalhes ver Conceição (2013) e Conceição *et al.* (2017).

Desta forma, nota-se que a probabilidade de $Y = 0$ é $1-\omega$ e a probabilidade de ocorrência de uma observação y, $y > 0$, é $\omega\pi_{ZTPS}(y;\mu;\phi)$.

A distribuição ZMPS expressa como distribuição *hurdle* contém a distribuição ZTPS como um de seus componentes e pode ser interpretada como a união do processo que produz observações positivas provenientes de uma distribuição ZTPS e de outro que produz apenas observações zero.

Uma vez que $p = \frac{\omega}{1-\pi_{PS}(0;\mu,\phi)}$, obtém-se diretamente que a média e a variância de uma variável aleatória Y com função massa de probabilidade dada pela equação (2.9) são, respectivamente,

$$\mu_{ZMPS} = \frac{\omega\mu}{1-\pi_{PS}(0;\mu,\phi)} = \omega\mu_{ZTPS} \qquad e \qquad \sigma^2_{ZMPS} = \frac{\omega(\sigma^2+\mu^2)}{1-\pi_{PS}(0;\mu,\phi)} - (\omega\mu_{ZTPS})^2.$$

Similarmente, diferentes valores de ω resultam em diferentes modificações na frequência de zero da distribuição ZMPS em sua versão *hurdle*, ao levar-se em consideração que

$$\pi_{ZMPS}(0;\mu,\phi,\omega) - \pi_{PS}(0;\mu,\phi) = 1-\omega-\pi_{PS}(0;\mu,\phi). \tag{2.10}$$

Assim, para os possíveis valores de ω ($\omega \in [0;1]$) na equação (2.10), tem-se as seguintes situações:

- $\omega = 0$: $\pi_{ZMPS}(0;\mu,\phi,\omega) = 1$, e então $\pi_{ZMPS}(y;\mu,\phi,\omega)$ é uma distribuição degenerada no ponto com toda a massa no zero;
- $0 < \omega < 1-\pi_{PS}(0;\mu,\phi)$: $1-\omega-\pi_{PS}(0;\mu,\phi) > 0$, e então $\pi_{ZMPS}(y;\mu,\phi,\omega)$ é uma distribuição ZIPS;
- $\omega = 1-\pi_{PS}(0;\mu,\phi)$: $\pi_{ZMPS}(0;\mu,\phi,\omega) = \pi_{PS}(0;\mu,\phi)$, e então $\pi_{ZMPS}(y;\mu,\phi,\omega)$ é uma distribuição PS usual;
- $1-\pi_{PS}(0;\mu,\phi) < \omega < 1$: $1-\omega-\pi_{PS}(0;\mu,\phi) < 0$, e então $\pi_{ZMPS}(y;\mu,\phi,\omega)$ é uma distribuição ZDPS;
- $\omega = 1$: $\pi_{ZMPS}(0;\mu,\phi,\omega) = 0$, e então $\pi_{ZMPS}(y;\mu,\phi,\omega)$ é uma distribuição ZTPS.

2.2.3 Inferência

Como ferramenta importante para a análise de um conjunto de observações, a inferência estatística permite interpretar resultados apropriadamente, bem como obter deduções, a partir de uma amostra, sobre uma população de interesse.

Em um conjunto de observações em que o zero não ocorre, mas há possibilidade deste valor ocorrer no experimento em estudo, utilizar o procedimento de estimação de máxima verossimilhança diretamente para a distribuição ZMPS levaria a estimativas equivocadas, visto que se deve assumir uma distribuição ZDPS (e não uma ZTPS), pois $P(Y=0)>0$.

Os procedimentos de estimação por meio de método dos momentos e de máxima verossimilhança para os parâmetros das distribuições da família ZMPS são descritos a seguir. Comentários sobre o caso particular de interesse desse trabalho serão destacados a fim de apontar o problema existente em torno destes procedimentos de estimação considerados.

Para isso, seja Y uma variável aleatória com distribuição ZMPS. Considere a amostra aleatória $\boldsymbol{Y}=(Y_1,Y_2,\ldots,Y_n)$, que consiste de n realizações independentes e identicamente distribuídas de Y e $\boldsymbol{y}=(y_1,\ldots,y_n)$ um vetor de observações associado a $\boldsymbol{y}$. Considera-se $\mathscr{D}=(\boldsymbol{y},n,n_j)$ o vetor de informações de $\boldsymbol{y}$, em que n é o número total de observações na amostra e n_j o número de observações j em $\boldsymbol{y}$, $j=0,1,\ldots$.

Para a distribuição ZMPS parametrizada em p, a função de verossimilhança associada ao vetor $\boldsymbol{y}$ é dada por:

$$\begin{aligned}
L(\mu,\phi,p;\mathscr{D}) &= \prod_{i=1}^{n}\left\{(1-p)\mathbf{I}(y_i)+p\pi_{PS}(y_i;\mu,\phi)\right\}\\
&= \prod_{i=1}^{n}\left\{(1-p+p\pi_{PS}(0;\mu,\phi))^{\mathbf{I}(y_i)}(p\pi_{PS}(y_i;\mu,\phi))^{1-\mathbf{I}(y_i)}\right\}\\
&= (1-p(1-\pi_{PS}(0;\mu,\phi)))^{n_0}\prod_{j=1}^{\infty}\{(p\pi_{PS}(j;\mu,\phi))^{n_j}\}.
\end{aligned}$$

O logaritmo natural da função de verossimilhança é dada por:

$$\ell(\mu,\phi,p;\mathscr{D}) = n_0\log(1-p(1-\pi_{PS}(0;\mu,\phi)))+\sum_{j=1}^{\infty}n_j\log(p\pi_{PS}(j;\mu,\phi)), \tag{2.11}$$

que reescrita com as funções $a(y,\phi)$, $f(\mu,\phi)$ e $g(\mu,\phi)$, é expressa como:

$$\ell(\mu,\phi,p;\mathscr{D}) = n_0\log\left(1-p\frac{f(\mu,\phi)-1}{f(\mu,\phi)}\right)+\sum_{j=1}^{\infty}n_j\log\left(p\frac{a(j,\phi)g(\mu,\phi)^j}{f(\mu,\phi)}\right).$$

Por sua vez, para a ZMPS(μ, ϕ, ω), a equação de verossimilhança é dada por:

$$\begin{aligned} L(\mu,\phi,\omega;\mathscr{D}) =& \prod_{i=1}^{n}\{(1-\omega)\mathbf{I}(y_i)+\omega\pi_{ZTPS}(y_i;\mu,\phi)\} \\ =& \prod_{i=1}^{n}\left\{(1-\omega)^{\mathbf{I}(y_i)}(\omega\pi_{ZTPS}(y_i;\mu,\phi))^{1-\mathbf{I}(y_i)}\right\} \\ =& (1-\omega)^{n_0}\omega^{n-n_0}\prod_{j=1}^{\infty}\left\{\left(\frac{\pi_{PS}(j;\mu,\phi)}{1-\pi_{PS}(0;\mu,\phi)}\right)^{n_j}\right\} \end{aligned} \tag{2.12}$$

e, por sua vez, o logaritmo natural da função de verossimilhança $L(\mu, \phi, \omega; \mathscr{D})$ é dado por:

$$\begin{aligned} \ell(\mu,\phi,\omega;\mathscr{D}) =& n_0\log(1-\omega)+(n-n_0)\log(\omega)+\sum_{j=1}^{\infty}n_j\log\left(\frac{\pi_{PS}(j;\mu,\phi)}{1-\pi_{PS}(0;\mu,\phi)}\right) \\ =& \sum_{j=1}^{\infty}n_j\log(\pi_{PS}(j;\mu,\phi))-(n-n_0)\log(1-\pi_{PS}(0;\mu,\phi))+ \\ &+n_0\log(1-\omega)+(n-n_0)\log(\omega); \end{aligned} \tag{2.13}$$

isto é,

$$\ell(\mu,\phi,\omega;\mathscr{D}) = \ell_1(\mu,\phi;\mathscr{D})+\ell_2(\omega;\mathscr{D}), \tag{2.14}$$

em que

$$\ell_1(\mu,\phi;\mathscr{D}) = \sum_{j=1}^{\infty}n_j\log(\pi_{PS}(j;\mu,\phi))-(n-n_0)\log(1-\pi_{PS}(0;\mu,\phi)) \tag{2.15}$$

e

$$\ell_2(\omega;\mathscr{D}) = n_0\log(1-\omega)+(n-n_0)\log(\omega). \tag{2.16}$$

Assim, ao considerar-se a equação (2.13), verifica-se que a parametrização em ω é mais vantajosa, pois observa-se que ω é ortogonal a μ e ϕ (dado que $\ell_2(\omega;\mathscr{D})$ não possui termos que são função dos demais parâmetros), possibilitando a estimação independente de ω com relação a μ e ϕ.

A seguir são apresentados dois procedimentos de estimação dos parâmetros: método dos momentos e máxima verossimilhança.

2.2.3.1 Método dos Momentos

Pode-se obter estimadores para os parâmetros de uma distribuição ZMPS através do método dos momentos, que relaciona os momentos populacionais aos momentos amostrais correspondentes.

Seja Y uma variável aleatória com distribuição ZMPS. Seu r-ésimo momento amostral ($r \geq 1$), denotado por $M^{(r)}$, é definido por:

$$M^{(r)} = \frac{1}{n}\sum_{i=1}^{n}Y_i^r.$$

Igualando-se o momento populacional $m_{ZMPS}^{(r)}$ a seu respectivo momento amostral $M^{(r)}$, obtém-se estimadores para os parâmetros da distribuição ZMPS por meio do método dos momentos:

$$\begin{cases} m_{ZMPS}^{(1)} = pm_{PS}^{(1)} = M^{(1)} \\ m_{ZMPS}^{(2)} = pm_{PS}^{(2)} = M^{(2)} \\ m_{ZMPS}^{(3)} = pm_{PS}^{(3)} = M^{(3)}. \end{cases}$$

Assim, a solução do seguinte sistema de equações fornece os estimadores de momentos $\widetilde{\mu}$, $\widetilde{\phi}$ e $\widetilde{p}$ para os parâmetros da distribuição ZMPS:

$$\begin{cases} p\mu = \dfrac{1}{n}\sum\limits_{i=1}^{n} Y_i \\ p\left\{\mu^2+\sigma^2\right\} = \dfrac{1}{n}\sum\limits_{i=1}^{n} Y_i^2 \\ p\mu\left\{\mu^2+\sigma^2\right\} + p\sigma^2\left\{2\mu\left[\dfrac{d}{d\mu}\left(\sigma^2\right)\right]\right\} = \dfrac{1}{n}\sum\limits_{i=1}^{n} Y_i^3, \end{cases}$$

em que σ^2 é uma função de μ e ϕ como apresentado na Tabela 3.

O estimador de momentos para o parâmetro ω pode ser obtido por meio da relação $p = \frac{\omega}{1-\pi_{PS}(0;\mu,\phi)}$; isto é, $\widetilde{\omega} = \widetilde{p}(1-\pi_{PS}(0;\widetilde{\mu},\widetilde{\phi}))$.

2.2.3.2 *Procedimento de Máxima Verossimilhança*

No procedimento de estimação de máxima verossimilhança, o objetivo consiste em estimar os parâmetros do modelo mais plausíveis de terem fornecido o conjunto de dados observados.

Com a primeira derivada do logaritmo natural da função de verossimilhança em relação a cada parâmetro, define-se o vetor escore U. Para a função $\ell(\mu,\phi,p;\mathscr{D})$ expressa pela equação (2.11), os elementos do vetor escore são dados por:

$$\begin{aligned} U_p = \frac{\partial \ell(\mu,\phi,p;\mathscr{D})}{\partial p} =& \frac{-n_0(1-\pi_{PS}(0;\mu,\phi))}{1-p(1-\pi_{PS}(0;\mu,\phi))} + \frac{(n-n_0)}{p} \\ =& \frac{-n_0(f(\mu,\phi)-1)}{f(\mu,\phi)-p(f(\mu,\phi)-1)} + \frac{(n-n_0)}{p}; \end{aligned}$$

$$\begin{aligned} U_\mu = \frac{\partial \ell(\mu,\phi,p;\mathscr{D})}{\partial \mu} =& \frac{n_0 p(\frac{\partial}{\partial\mu}\pi_{PS}(0;\mu,\phi))}{1-p(1-\pi_{PS}(0;\mu,\phi))} + \sum_{j=1}^{\infty}\frac{n_j\left(\frac{\partial}{\partial\mu}\pi_{PS}(j;\mu,\phi)\right)}{\pi_{PS}(j;\mu,\phi)} \\ =& -\left(\frac{\partial}{\partial\mu}\log(f(\mu,\phi))\right) + \left[\frac{n_0 p}{f(\mu,\phi)-p(f(\mu,\phi)-1)} + (n-n_0)\right] + \\ &+ \left(\frac{\partial}{\partial\mu}\log(g(\mu,\phi))\right)\sum_{j=1}^{\infty} jn_j; \end{aligned}$$

$$
\begin{aligned}
U_\phi = \frac{\partial \ell(\mu,\phi,p;\mathscr{D})}{\partial \phi} =& \frac{n_0 p(\frac{\partial}{\partial \phi}\pi_{PS}(0;\mu,\phi))}{1-p(1-\pi_{PS}(0;\mu,\phi))} + \sum_{j=1}^{\infty} \frac{n_j\left(\frac{\partial}{\partial \phi}\pi_{PS}(j;\mu,\phi)\right)}{\pi_{PS}(j;\mu,\phi)} \\
=& -\left(\frac{\partial}{\partial \phi}\log(f(\mu,\phi))\right) + \left[\frac{n_0 p}{f(\mu,\phi)-p(f(\mu,\phi)-1)} + (n-n_0)\right] + \\
&+ \sum_{j=1}^{\infty} n_j\left(\frac{\partial}{\partial \phi}\left[\log(a(j,\phi)) + j\log(g(\mu,\phi))\right]\right).
\end{aligned}
$$

Para obter o estimador de máxima verossimilhança (EMV) para cada um dos parâmetros, iguala-se o vetor escore $U = (U_p, U_\mu, U_\phi)^T$ a zero. Apenas o parâmetro p possui forma explícita, dada por:

$$
\begin{aligned}
\widehat{p} =& \frac{n-n_0}{n(1-\pi_{PS}(0;\widehat{\mu},\widehat{\phi}))} \\
=& \frac{(n-n_0)f(\widehat{\mu},\widehat{\phi})}{n(f(\widehat{\mu},\widehat{\phi})-1)}.
\end{aligned} \tag{2.17}
$$

Ressalta-se que, para a situação em que o conjunto de dados não contém observações zero ($n_0 = 0$), tem-se que a estimativa de máxima verossimilhança do parâmetro de modificação p é:

$$
\begin{aligned}
\widehat{p}(\boldsymbol{y}) &= \frac{1}{1-\pi_{PS}(0;\widehat{\mu}(\boldsymbol{y}),\widehat{\phi}(\boldsymbol{y}))} \\
&= \frac{f(\widehat{\mu}(\boldsymbol{y}),\widehat{\phi}(\boldsymbol{y}))}{f(\widehat{\mu}(\boldsymbol{y}),\widehat{\phi}(\boldsymbol{y}))-1}
\end{aligned}
$$

que consiste no limitante superior do espaço paramétrico de p, indicando que o comportamento dos dados pode ser explicado por uma distribuição ZTPS, cuja probabilidade de ocorrência de zero é nula. Porém, tal fato pode não ser verdade — a probabilidade de ocorrência de zero pode ser positiva, embora tal valor não tenha ocorrido na amostra em questão —, e o objetivo deste trabalho é tratar adequadamente esta situação.

As estimativas para μ e ϕ podem ser obtidas utilizando-se $\widehat{p}$. Assim, obtém-se que o vetor escore

$$
\widetilde{U} = (\widetilde{U}_\mu, \widetilde{U}_\phi)^T
$$

com elementos

$$
\begin{aligned}
\widetilde{U}_\mu =& \frac{(n-n_0)\left(\frac{\partial}{\partial \mu}\pi_{PS}(0;\mu,\phi)\right)}{1-\pi_{PS}(0;\mu,\phi)} + \sum_{j=1}^{\infty} \frac{n_j\left(\frac{\partial}{\partial \mu}\pi_{PS}(j;\mu,\phi)\right)}{\pi_{PS}(j;\mu,\phi)} \\
=& -\left(\frac{\partial}{\partial \mu}\log(f(\mu,\phi))\right)\left[\frac{(n-n_0)f(\mu,\phi)}{f(\mu,\phi)-1}\right] + \left(\frac{\partial}{\partial \mu}\log(g(\mu,\phi))\right)\sum_{j=1}^{\infty} j n_j;
\end{aligned} \tag{2.18}
$$

$$\begin{aligned}\widetilde{U}_\phi =& \frac{(n-n_0)\left(\frac{\partial}{\partial\phi}\pi_{PS}(0;\mu,\phi)\right)}{1-\pi_{PS}(0;\mu,\phi)} + \sum_{j=1}^{\infty}\frac{n_j\left(\frac{\partial}{\partial\phi}\pi_{PS}(j;\mu,\phi)\right)}{\pi_{PS}(j;\mu,\phi)} \\ =& -\left(\frac{\partial}{\partial\phi}\log(f(\mu,\phi))\right)\left[\frac{(n-n_0)f(\mu,\phi)}{f(\mu,\phi)-1}\right] + \left(\frac{\partial}{\partial\phi}\log(g(\mu,\phi))\right)\sum_{j=1}^{\infty} jn_j + \\ &+ \sum_{j=1}^{\infty} n_j\left(\frac{\partial}{\partial\phi}\log(a(j,\phi))\right). \end{aligned} \tag{2.19}$$

Então, as estimativas numéricas para μ e ϕ podem ser obtidas por meio da resolução das equações $\widetilde{U}_\mu = 0$ e $\widetilde{U}_\phi = 0$. As equações (2.18) e (2.19) podem ser simplificadas por (CONCEIÇÃO, 2013):

$$\widetilde{U}_\mu = \bar{y}^+ - \frac{\mu f(\mu,\phi)}{f(\mu,\phi)-1} = 0,$$

isto é,

$$\bar{y}^+ = \frac{\mu f(\mu,\phi)}{f(\mu,\phi)-1} \tag{2.20}$$

e

$$(n-n_0)\left\{\frac{\bar{y}^+\left(\frac{\partial}{\partial\phi}g(\mu,\phi)\right)}{g(\mu,\phi)} - \frac{\frac{\partial}{\partial\phi}f(\mu,\phi)}{f(\mu,\phi)-1}\right\} + \sum_{j=1}^{\infty}\frac{n_j\left(\frac{\partial}{\partial\phi}a(j,\phi)\right)}{a(j,\phi)} = 0, \tag{2.21}$$

em que $\bar{y}^+ = \frac{1}{n-n_0}\sum_{i=1}^{n} y_i$ é a média das observações positivas do vetor $\mathbf{y}$.

Considerando-se a parametrização em ω, o procedimento de estimação por máxima verossimilhança utiliza a equação (2.13). Para obter os EMVs dos parâmetros ω, μ e ϕ, calcula-se vetor escore $U = (U_\omega, U_\mu, U_\phi)^T$, cujos elementos são dados por:

$$U_\omega = \frac{\partial\ell_2(\omega;\mathscr{D})}{\partial\omega} = \frac{-n_0}{1-\omega} + \frac{n-n_0}{\omega};$$

$$\begin{aligned}U_\mu =& \frac{\partial\ell_1(\mu,\phi;\mathscr{D})}{\partial\mu} = \frac{(n-n_0)\left(\frac{\partial}{\partial\mu}\pi_{PS}(0;\mu,\phi)\right)}{1-\pi_{PS}(0;\mu,\phi)} + \sum_{j=1}^{\infty}\frac{n_j\left(\frac{\partial}{\partial\mu}\pi_{PS}(j;\mu,\phi)\right)}{\pi_{PS}(j;\mu,\phi)} \\ =& -\left(\frac{\partial}{\partial\mu}\log(f(\mu,\phi))\right)\left[\frac{(n-n_0)f(\mu,\phi)}{f(\mu,\phi)-1}\right] + \\ &+ \left(\frac{\partial}{\partial\mu}\log(g(\mu,\phi))\right)\sum_{j=1}^{\infty} jn_j;\end{aligned}$$

$$
\begin{aligned}
U_\phi =& \frac{\partial \ell_1(\mu,\phi;\mathscr{D})}{\partial \phi} = \frac{(n-n_0)\left(\frac{\partial}{\partial \phi}\pi_{PS}(0;\mu,\phi)\right)}{1-\pi_{PS}(0;\mu,\phi)} + \sum_{j=1}^{\infty}\frac{n_j\left(\frac{\partial}{\partial \phi}\pi_{PS}(j;\mu,\phi)\right)}{\pi_{PS}(j;\mu,\phi)} \\
=& -\left(\frac{\partial}{\partial \phi}\log(f(\mu,\phi))\right)\left[\frac{(n-n_0)f(\mu,\phi)}{f(\mu,\phi)-1}\right] + \\
&+\left(\frac{\partial}{\partial \phi}\log(g(\mu,\phi))\right)\sum_{j=1}^{\infty} jn_j + \sum_{j=1}^{\infty} n_j\left(\frac{\partial}{\partial \phi}\log(a(j,\phi))\right),
\end{aligned}
$$

em que $\ell_1(\mu,\phi;\mathscr{D})$ e $\ell_2(\omega;\mathscr{D})$ são dadas, respectivamente, pelas equações (2.15) e (2.16).

Igualando-se cada um dos elementos do vetor U a zero, verifica-se que somente o estimador do parâmetro ω possui equação explícita, dada por:

$$
\widehat{\omega} = \frac{n-n_0}{n}. \tag{2.22}
$$

Novamente, destaca-se que para o caso em que o número de observações zero no conjunto de dados é igual a zero ($n_0 = 0$), tem-se que a estimativa de máxima verossimilhança do parâmetro ω é

$$
\widehat{\omega} = 1,
$$

e, mais uma vez tem-se como estimativa o limite superior do parâmero ω, indicando uma distribuição ZTPS.

Para obter as estimativas de máxima verossimilhança de μ e ϕ, procedimentos numéricos deverão ser considerados para a resolução do sistema de equações:

$$
-\left(\frac{\partial}{\partial \mu}\log(f(\mu,\phi))\right)\left[\frac{(n-n_0)f(\mu,\phi)}{f(\mu,\phi)-1}\right] + \left(\frac{\partial}{\partial \mu}\log(g(\mu,\phi))\right)\sum_{j=1}^{\infty} jn_j = 0 \tag{2.23}
$$

$$
\begin{aligned}
&-\left(\frac{\partial}{\partial \phi}\log(f(\mu,\phi))\right)\left[\frac{(n-n_0)f(\mu,\phi)}{f(\mu,\phi)-1}\right] + \left(\frac{\partial}{\partial \phi}\log(g(\mu,\phi))\right)\sum_{j=1}^{\infty} jn_j + \\
&+\sum_{j=1}^{\infty} n_j\left(\frac{\partial}{\partial \phi}\log(a(j,\phi))\right) = 0
\end{aligned} \tag{2.24}
$$

Como as equações (2.23) e (2.24) são iguais às equações (2.18) e (2.19) quando igualadas a zero, as estimativas de máxima verossimilhança de μ e ϕ das distribuições ZMPS(μ,ϕ,ω) podem ser obtidas, da mesma forma, por meio das equações (2.20) e (2.21).

Assim, com as estimativas de ω, μ e ϕ, denotadas respectivamente por $\widehat{\omega}(\boldsymbol{y})$, $\widehat{\mu}(\boldsymbol{y})$ e $\widehat{\phi}(\boldsymbol{y})$, pode-se obter a estimativa do parâmetro p por meio da relação:

$$
\widehat{p}(\boldsymbol{y}) = \frac{\widehat{\omega}(\boldsymbol{y})}{1-\pi_{PS}(0;\widehat{\mu}(\boldsymbol{y}),\widehat{\phi}(\boldsymbol{y}))}.
$$

Ainda sobre o procedimento de estimação dos parâmetros da distribuição ZMPS, poderia ser considerada uma abordagem bayesiana. Por não ser o enfoque do trabalho, este procedimento não será aprofundado; todavia no Apêndice A é apresentada uma breve descrição deste procedimento de estimação, que pode ser bastante eficiente por permitir incorporar algum conhecimento (informação, opinião) que um especialista tem (a priori) sobre os parâmetros.

CAPÍTULO

3

UM CASO ESPECIAL DA DISTRIBUIÇÃO ZDPS

Em um conjunto de observações em que o zero não ocorre, é preciso ter cuidado ao assumir a distribuição ZMPS como adequada para explicar o comportamento dos dados. Quando há possibilidade do zero ocorrer no experimento em estudo, embora não tenha sido observado, deve-se assumir uma distribuição ZDPS, pois a probabilidade de zero neste caso é positiva (e não nula). Utilizar o procedimento de estimação de máxima verossimilhança diretamente para a distribuição ZMPS nessa situação particular levaria a estimativas equivocadas, uma vez que, devido ao fato de ter-se que o número de zeros observados é igual a zero, $n_0 = 0$, a estimativa para o parâmetro p (ou ω; equações (2.17) e (2.22)) seria equivalente a seu limitante superior, indicando que o conjunto de dados é proveniente de uma distribuição ZTPS, o que na prática não condiz com a verdade.

A fim de realizar o procedimento de estimação dos parâmetros corretamente para a situação em que os dados de contagem a serem analisados correspondem a um conjunto que contém apenas observações positivas e com probabilidade de ocorrência de zero não nula, adota-se o algoritmo EM, capaz de lidar com os "zeros faltantes" para obter estimativas mais fidedignas.

3.1 Distribuição ZMPS *versus* ZTPS

Conjuntos de dados que não contém a observação zero podem ser distinguidos entre zero-deflacionado e zero-truncado por meio da avaliação do pesquisador sobre a probabilidade de ocorrência de tal observação: baixa ou nula, respectivamente. Dada uma variável aleatória Y com distribuição ZMPS em que $P(Y = 0) = 0$ (conhecida), o conjunto de dados obtidos a partir de n realizações independentes de Y pode ser visto como proveniente de uma distribuição ZTPS; por outro lado, quando $P(Y = 0) > 0$ no experimento em estudo, embora o zero não tenha sido

observado na amostra, deve-se assumir uma distribuição ZDPS. Ressalta-se que tanto a ZTPS quanto a ZDPS são casos particulares da distribuição ZMPS.

Amostras de dados de contagem zero-truncadas podem ocorrer devido à escolha do esquema amostral empregado; em casos no qual critério de elegibilidade para a amostra é ao menos uma ocorrência do evento, não se tem zeros registrados (GROGGER; CARSON, 1991). Situações práticas nas quais não há a observação zero dentre os dados de contagem são frequentes em estudos de ecologia que envolvem a abundância de espécies em determinada região, cenário em que a contagem de espécies é sempre maior que zero (ARRABAL; SILVA; BANDEIRA, 2014). Aplicações nas áreas de tráfego e transporte também são naturais, ao considerar-se, por exemplo, número de pessoas dentro de veículos em trechos de rodovia (ao menos um ocupante estará no veículo) ou no caso de um conjunto de dados que consiste no número de acidentes em seção de rodovia, coletados a partir de registros de ocorrências. Chowdhury *et al.* (2016) exemplifica citando dados de segurança rodoviária na Grã-Bretanha[1], cujos arquivos contêm informações detalhadas sobre acidentes ocorridos nas rodovias estaduais, incluindo o número de feridos, tipo de veículos envolvidos e outros dados, mas em que somente acidentes que envolvem feridos são reportados, gerando assim dados zero-truncados.

Como colocado por Conceição (2013), quando o parâmetro μ da distribuição ZMPS (ou PS associada) é grande (em relação ao valor máximo do espaço paramétrico restrito de p correspondente a cada distribuição ZDPS), a probabilidade de ocorrência de zero da PS é pequena. Consequentemente, dependendo do valor atribuído ao parâmetro de modificação p (e também ao parâmetro de dispersão ϕ, quando a distribuição o contém), a distribuição zero-deflacionada pode gerar conjuntos de dados com poucas observações zero ou até nenhuma. Nestas situações tornam-se mais difícil explicar corretamente o caso particular da distribuição ZMPS. Contudo, estimativas errôneas dos parâmetros μ e p são refletidas também nas probabilidades das observações positivas.

Contextualizando formalmente esta situação, suponha Y uma variável aleatória com distribuição ZMPS e $\boldsymbol{y} = (y_1, \ldots, y_n)$ o conjunto de observações provenientes de n realizações independentes de Y. Nota-se que n também denota o número total de observações na amostra. Define-se ainda n_j, $j = 0, 1, \ldots$, o número de observações j e n^+ o número total de observações positivas em $\boldsymbol{y}$, tal que $n^+ = \sum_{j=1}^{\infty} n_j = n - n_0$.

Definindo $\boldsymbol{y}^+$ o vetor que contém apenas observações positivas de $\boldsymbol{y}$, é razoável supor que $\boldsymbol{y}^+$ é um vetor de observações provenientes de n^+ realizações independentes da variável aleatória Y^+ que tem distribuição ZTPS com função massa de probabilidade dada por:

$$\pi_{ZTPS}(y^+; \mu, \phi) = \frac{\pi_{PS}(y^+; \mu, \phi)}{1 - \pi_{PS}(0; \mu, \phi)}, \qquad y^+ = 1, 2, \ldots. \tag{3.1}$$

[1] Disponíveis on-line em <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>.

A função de verossimilhança associada ao vetor $\boldsymbol{y}^+$ é dada por:

$$\begin{aligned} L(\mu,\phi;\boldsymbol{y}^+) &= \prod_{i=1}^{n^+}\left\{\frac{\pi_{PS}(y_i^+;\mu,\phi)}{1-\pi_{PS}(0;\mu,\phi)}\right\} \\ &= \frac{1}{(1-\pi_{PS}(0;\mu,\phi))^{n^+}}\prod_{j=1}^{\infty}\pi_{PS}(j;\mu,\phi)^{n_j} \end{aligned}$$

e a função log-verossimilhança é:

$$\ell(\mu,\phi;\boldsymbol{y}^+) = \sum_{j=1}^{\infty} n_j log(\pi_{PS}(j;\mu,\phi)) - n^+ log(1-\pi_{PS}(0;\mu,\phi)).$$

Ao derivar a função log-verossimilhança com relação aos parâmetros, obtém-se que o vetor escore $U = (U_\mu^+, U_\phi^+)^T$ tem elementos

$$U_\mu^+ = \frac{\partial\ell(\mu,\phi;\boldsymbol{y}^+)}{\partial\mu} = \frac{n^+\left(\frac{\partial}{\partial\mu}\pi_{PS}(0;\mu,\phi)\right)}{1-\pi_{PS}(0;\mu,\phi)} + \sum_{j=1}^{\infty}\frac{n_j\left(\frac{\partial}{\partial\mu}\pi_{PS}(j;\mu,\phi)\right)}{\pi_{PS}(j;\mu,\phi)}; \tag{3.2}$$

$$U_\phi^+ = \frac{\partial\ell(\mu,\phi;\boldsymbol{y}^+)}{\partial\phi} = \frac{n^+\left(\frac{\partial}{\partial\phi}\pi_{PS}(0;\mu,\phi)\right)}{1-\pi_{PS}(0;\mu,\phi)} + \sum_{j=1}^{\infty}\frac{n_j\left(\frac{\partial}{\partial\phi}\pi_{PS}(j;\mu,\phi)\right)}{\pi_{PS}(j;\mu,\phi)}. \tag{3.3}$$

Nota-se que as equações (3.2) e (3.3) coincidem com as equações (2.18) e (2.19), uma vez que $n^+ = n - n_0$, permitindo concluir que os EMVs dos parâmetros μ e ϕ da distribuição ZTPS são iguais aos obtidos para os parâmetros da distribuição ZMPS ao serem consideradas apenas as observações positivas de $\boldsymbol{y}$.

3.2 Distribuição ZDPS *versus* PS: Estimação via Algoritmo EM

Em problemas reais existem situações em que conjuntos de dados de contagem não contêm observações zeros, apesar da probabilidade positiva desta observação. Para estas situações, é razoável afirmar que estes conjuntos de dados são provenientes de uma distribuição zero-modificada (mais especificamente, zero-deflacionada).

Já é conhecido que, no procedimento de máxima verossimilhança, estimar os parâmetros μ e ϕ da distribuição ZMPS é equivalente a estimar os parâmetros μ e ϕ da distribuição ZTPS associada, considerando apenas as observações positivas. Com as estimativas destes parâmetros, é possível obter a estimativa do parâmetro p (ver equação 2.17), o qual informa o tipo de zero-modificação presentes nos dados.

Por outro lado, na situação em que um determinado conjunto de dados não contém observações zero, $n_0 = 0$, tem-se que $\boldsymbol{y} \equiv \boldsymbol{y}^+$. Isso permite afirmar que $\boldsymbol{y}$ é proveniente de uma

distribuição ZTPS, o que não é verdade, já que a observação zero tem uma probabilidade positiva. Essa afirmação pode ser mais facilmente compreendida ao obter-se a estimativa de máxima verossimilhança de p (equação 2.17) e verificar que esta resultará no seu limite superior, recaindo sobre o seu caso particular, a distribuição ZTPS. Na verdade, $\boldsymbol{y} \equiv \boldsymbol{y}^+$ é proveniente de uma distribuição ZDPS, uma vez que $P(Y = 0) > 0$.

Para contornar este problema, parte-se da ideia de que o particular conjunto de dados é incompleto (devido a algum problema do processo de amostragem ou limitações do processo de observação) e, desta forma, é introduzida uma variável latente com o objetivo de tornar os dados completos (aumentados), tal que a suposição de uma distribuição PS padrão seja adequada para explicar o comportamento do conjunto de dados aumentados. Assim, os EMVs dos parâmetros da distribuição suposta serão obtidos via algoritmo EM.

O algoritmo EM foi proposto por (DEMPSTER *et al.*, 1977) como um método iterativo para encontrar as estimativas de máxima verossimilhança dos parâmetros de uma distribuição para um conjunto de observações na presença de dados "faltantes" ou "incompletos". Cada iteração do algoritmo consiste em dois passos: o passo E e o passo M. No passo E, os dados faltantes são estimados segundo os dados observados, utilizando-se esperança condicional. No passo M, a função de verossimilhança é maximizada sob a hipótese de que os dados faltantes são conhecidos. A convergência é garantida pelo fato de que o algoritmo tem verossimilhança crescente em cada iteração (DEMPSTER *et al.*, 1977). O algoritmo EM possui extensões, diversas aplicações e variações, podendo ser estendido aos procedimentos de estimação de máxima verossimilhança penalizada e do máximo a posteriori (para mais detalhes, veja McLachlan e Krishnan (2008)).

Contextualizando o problema de interesse deste trabalho, considera-se o vetor de observações provenientes de realizações independentes da variável aleatória Y com distribuição ZMPS, o qual contém apenas observações positivas. Denotando por n^+ o número total de observações (todas positivas), representa-se o vetor de observações por $\boldsymbol{y}^+ = (y_1^+, \ldots, y_{n^+}^+)$, em que $y_i^+ > 0$, $\forall\, i = 1, \ldots, n^+$. Desta forma, a suposição já feita da distribuição ZMPS para os dados é equivalente a suposição de que $\boldsymbol{y}^+$ é um vetor de observações de provenientes de n^+ realizações independentes da variável aleatória Y^+ com distribuição ZTPS, cuja função massa de probabilidade é dada pela equação (3.1).

Como $\boldsymbol{y}^+$ tem zeros "faltantes", é preciso completá-lo de forma a obter o vetor de dados aumentado (completo). Para isso, considera-se N_0 uma variável aleatória latente ou auxiliar com distribuição NB, que representa a contagem de zeros "faltantes" (fracassos) e denotada por um vetor n_0-dimensional, até a ocorrência do vetor n^+-dimensional de observações positivas em $\boldsymbol{y}^+$ (ou seja, n^+ sucessos); desta forma, obtém-se o vetor de dados aumentados $(\boldsymbol{y}^+, \underbrace{0, \ldots, 0}_{n_0})$.

Assim, tem-se que $N_0|\boldsymbol{y}^+$ tem distribuição condicional $\text{NB}(n^+, 1 - \pi_{PS}(0; \mu, \phi))$, cuja

função massa de probabilidade é dada por:

$$\pi(N_0|\boldsymbol{y}^+;\mu,\phi) = \frac{\Gamma(n_0+n^+)}{\Gamma(n^+)\Gamma(n_0+1)}(1-\pi_{PS}(0;\mu,\phi))^{n^+}(\pi_{PS}(0;\mu,\phi))^{n_0}, \qquad n_0 = 0,1,...,$$

e número esperado de observações zero

$$\mathbb{E}[N_0|\boldsymbol{y}^+;\mu,\phi] = \frac{n^+\pi_{PS}(0;\mu,\phi)}{1-\pi_{PS}(0;\mu,\phi)}.$$

Então, tem-se:

$$\begin{aligned} Y^+ &\sim ZTPS(\mu,\phi), \\ N_0|\boldsymbol{y}^+ &\sim NB(n^+, 1-\pi_{PS}(0;\mu,\phi)). \end{aligned}$$

Uma vez que $\boldsymbol{y}^+$ é o vetor de dados incompletos e n_0 é tratado como um dado faltante, define-se $\boldsymbol{y}_a = (\boldsymbol{y}^+, \boldsymbol{0}_{[n_0]})$ o vetor de dados aumentados (completos), proveniente de realizações independentes da variável aleatória Y_a com distribuição PS. Ressalta-se que $\boldsymbol{y}_a$ é composto pelo vetor $\boldsymbol{y}^+$ de dimensão n^+ e pelo vetor de "zeros faltantes" com dimensão n_0 (a ser estimada); logo, a dimensão de $\boldsymbol{y}_a$ é $n = n^+ + n_0$, cuja estimativa é dada por $\widehat{n} = n^+ + \widehat{n}_0$.

Assim, a função de verossimilhança completa associada a $\boldsymbol{y}_a = (\boldsymbol{y}^+, \boldsymbol{0}_{[n_0]})$ é dada por:

$$\begin{aligned} L_a(\mu,\phi;\boldsymbol{y}_a) &= \prod_{i=1}^{n}\left\{\pi(y_i^+, n_0;\mu,\phi)\right\} \\ &= \prod_{i=1}^{n}\left\{\pi_{PS}(y_{a_i};\mu,\phi)\right\} \\ &= \big(\pi_{PS}(0;\mu,\phi)\big)^{n_0}\prod_{j=1}^{\infty}\left\{\left(\pi_{PS}(j;\mu,\phi)\right)^{n_j}\right\}, \end{aligned}$$

cuja função log-verossimilhança é

$$\ell_a(\mu,\phi;\boldsymbol{y}_a) = n_0\log\big(\pi_{PS}(0;\mu,\phi)\big) + \sum_{j=1}^{\infty} n_j\log\big(\pi_{PS}(j;\mu,\phi)\big).$$

Logo, o algoritmo EM primeiro encontra o valor esperado da função log-verossimilhança dos dados aumentados com relação às observações faltantes (N_0), condicionada aos valores observados e à estimativa atual do(s) parâmetro(s) (Passo E). Assim, define-se:

$$\begin{aligned} Q(\mu,\phi|\widehat{\mu}^{(k-1)},\widehat{\phi}^{(k-1)}) &= \mathbb{E}\left[\ell_a(\mu,\phi;\boldsymbol{y}_a)|\boldsymbol{y}^+;\widehat{\mu}^{(k-1)},\widehat{\phi}^{(k-1)}\right] \\ &= \widehat{n}_0\log\big(\pi_{PS}(0;\mu,\phi)\big) + \sum_{j=1}^{\infty} n_j\log\big(\pi_{PS}(j;\mu,\phi)\big), \end{aligned}$$

em que $\widehat{\mu}^{(k-1)}$ e $\widehat{\phi}^{(k-1)}$ são as estimativas atuais dos parâmetros para calcular a esperança $\widehat{n}_0 = \mathbb{E}[N_0|\boldsymbol{y}^+,\widehat{\mu}^{(k-1)},\widehat{\phi}^{(k-1)}]$.

Em seguida, a função $Q(\mu,\phi|\widehat{\mu}^{(k-1)},\widehat{\phi}^{(k-1)})$ deve ser maximizada para obter valores atuais $\widehat{\mu}^{(k)}$ e $\widehat{\phi}^{(k)}$ (Passo M), dados por

$$\widehat{\mu}^{(k)},\widehat{\phi}^{(k)} = \arg\max_{\mu,\phi} Q(\mu,\phi|\widehat{\mu}^{(k-1)},\widehat{\phi}^{(k-1)}),$$

que leva à seguinte solução analítica para $\widehat{\mu}^{(k)}$:

$$\widehat{\mu}^{(k)} = \frac{\sum_{i=1}^{n^+} y_i^+}{n^+ + \widehat{n}_0} = \frac{n^+\bar{y}^+}{n^+ + \widehat{n}_0},$$

em que $\bar{y}^+ = \frac{\sum_{i=1}^{n^+} y_i^+}{n^+}$ é a média amostral dos dados observados, $\boldsymbol{y}^+$.

O passo M, por sua vez, pode ser simplificado obtendo-se a solução da equação

$$\begin{aligned}\frac{\partial}{\partial\phi}Q(\widehat{\mu}^{(k)},\phi|\mu^{(k-1)},\phi^{(k-1)}) &= \widehat{n}_0\frac{\partial}{\partial\phi}\log\left(\pi_{PS}(0;\widehat{\mu}^{(k)},\phi)\right) + \\ &\quad \frac{\partial}{\partial\phi}\sum_{j=1}^{\infty} n_j \log\left(\pi_{PS}(j;\widehat{\mu}^{(k)},\phi)\right) = 0. \qquad (3.4)\end{aligned}$$

Logo, a estimativa $\widehat{\phi}^{(k)}$ pode ser obtida resolvendo-se iterativamente a seguinte equação:

$$\frac{n^+\bar{y}^+}{\widehat{\mu}^{(k)}}\left\{\frac{\partial}{\partial\phi}\log\left(f(\widehat{\mu}^{(k)},\phi)\right) - \widehat{\mu}^{(k)}\frac{\partial}{\partial\phi}\log\left(g(\widehat{\mu}^{(k)},\phi)\right)\right\} - \sum_{j=1}^{\infty} n_j\left\{\frac{\partial}{\partial\phi}\log\left(a(j,\phi)\right)\right\} = 0.$$

A estimativa de n é obtida da seguinte forma:

$$\begin{aligned}\widehat{n} &= n^+ + \mathbb{E}[N_0|\boldsymbol{y}^+,\widehat{\mu},\widehat{\phi}] \\ &= n^+ + \frac{n^+\pi_{PS}(0;\widehat{\mu},\widehat{\phi})}{1-\pi_{PS}(0;\widehat{\mu},\widehat{\phi})} \\ &= \frac{n^+}{1-\pi_{PS}(0;\widehat{\mu},\widehat{\phi})}.\end{aligned}$$

O procedimento de estimação dos parâmetros via algoritmo EM pode ser resumido, neste caso, por:

1. Dado o vetor de observações positivas $\boldsymbol{y}^+ = (y_1^+, y_2^+, \ldots, y_{n^+}^+)$; $k = 1$;
2. Obter as estimativas de μ e ϕ a partir dos respectivos estimadores de momentos $\widetilde{\mu}$ e $\widetilde{\phi}$ e fazer $\widetilde{\mu} = \widehat{\mu}^{(0)}$ e $\widetilde{\phi} = \widehat{\phi}^{(0)}$ (condição inicial);
3. Dados $\mu^{(k-1)}$ e $\phi^{(k-1)}$, calcular $\widehat{n}_0 = \mathbb{E}[N_0|\boldsymbol{y}^+,\widehat{\mu}^{(k-1)},\widehat{\phi}^{(k-1)}]$ (**Passo E**);
4. Obter $\widehat{\mu}^{(k)}$ e $\widehat{\phi}^{(k)}$ que maximizam a função $Q(\mu,\phi|\mu^{(k-1)},\phi^{(k-1)})$ (**Passo M**);
5. Fazer $k = k+1$ e repetir 2 e 3 até a convergência; isto é, até que a diferença entre as estimativas dos passos $k-1$ e k seja menor que um erro pré-estabelecido (geralmente muito pequeno).

CAPÍTULO

4

ESTUDO COM DADOS SIMULADOS

Para avaliar o desempenho do procedimento de estimação de parâmetros das distribuições da família ZMPS e também da probabilidade estimada para a ocorrência de zeros na situação particular tratada neste trabalho, foram realizados estudos de simulação e com dados artificialmente gerados, apresentados a seguir.

4.1 Estudo de Simulação

O estudo de simulação foi realizado considerando-se as distribuições ZMP, ZMG e ZMB (apenas as distribuições biparamétricas, uma vez que m é assumido conhecido). Para cada distribuição, o procedimento utilizado consistiu na geração de N amostras de tamanho n^+ (isto é, sem observações zero) para cada distribuição, com diferentes valores para o parâmetro de média μ.

Para cada particular amostra gerada, o método iterativo de estimação via algoritmo EM apresentado anteriormente foi considerado, estimando o valor esperado de zeros ($\widehat{n}_0$) que deveria ter ocorrido na amostra $\boldsymbol{y}^+$ para que o conjunto de dados aumentados $\boldsymbol{y}_a = (\boldsymbol{y}^+, \mathbf{0}_{[\widehat{n}_0]})$ pudesse ser explicado pela distribuição PS tradicional e, a partir dessa suposição, obteve-se a estimativa do parâmetro μ. Como critério de parada para o algoritmo foi estabelecido um erro de 10^{-4}; isto é, o algoritmo itera até que a diferença entre as estimativas atual e do passo anterior seja menor que o erro. Após a convergência, tem-se a estimativa do parâmetro μ e esta é utilizada para estimar a probabilidade de zero da respectiva distribuição PS. Assim, ao final do processo de simulação das N amostras geradas, tem-se um vetor $N-$dimensional das estimativas de μ e, consequentemente, outro vetor $N-$dimensional de estimativas de probabilidades de zero.

Esses vetores de estimativas foram utilizados para avaliar numericamente o desempenho dos estimadores com o método de estimação proposto. As medidas consideradas foram: o vício e a raiz quadrada da razão entre erro quadrático médio e variância. Tais medidas foram obtidas

por meio de estimativas de Monte Carlo do erro quadrático médio, variância e vício, denotadas respectivamente por $EQM(\theta)$, $Var(\theta)$ e $\mathscr{B}(\theta)$, e calculadas utilizando-se as seguintes equações:

$$EQM(\widehat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\theta}_i - \theta)^2; \quad Var(\widehat{\theta}) = \frac{1}{N-1}\sum_{i=1}^{N}(\widehat{\theta}_i - \overline{\theta})^2; \quad \mathscr{B}(\widehat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\theta}_i - \theta);$$

em que

$$\overline{\theta} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\theta}_i,$$

sendo $\theta = \{\mu, \phi, \pi_0\}$.

Além da avaliação dos estimadores, um procedimento de reamostragem do tipo *bootstrap* não paramétrico (BNP) (EFRON, 1979) foi realizado, considerando B réplicas para cada um dos N conjuntos, com o objetivo de obter intervalos *bootstrap* com um nível de confiança de 95%. Segundo Efron (1990), para os cálculos de vício e variância são necessárias entre 50 e 200 réplicas, e para a obtenção de intervalos de confiança é preciso algo entre 1000 e 2000 réplicas para se ter suficiente acurácia; porém, se os custos computacionais não fossem problema, o número ideal de réplicas seria infinito. Assim, neste trabalho, foram realizadas $B = 5000$ réplicas de forma a garantir o desempenho adequado do método.

Os intervalos de confiança obtidos foram utilizados para a obtenção da probabilidade de cobertura (denotado por PC, que é a razão entre a quantidade de vezes em que o verdadeiro valor do parâmetro pertence aos intervalos *bootstrap* e o número de replicações); o valor da PC é apresentado neste trabalho multiplicado por 100 para direta comparação com o valor nominal de 95%.

Para estimadores não viciados espera-se que, assintoticamente, a raiz da razão entre o erro quadrático médio e a variância se aproxime de 1 e o vício se aproxime de zero. Já para a probabilidade de cobertura é desejado que se aproxime do valor nominal estabelecido.

No estudo apresentado a seguir, foram considerados $N = 500$ conjuntos de dados de tamanhos $n^+ = 15,\ 25,\ 50,\ 100$ e 200, de forma a avaliar o desempenho do método para amostras pequenas e seus comportamentos assintóticos. Os valores considerados para μ foram $0,5,\ 1$ e $1,5$, e o parâmetro m da distribuição Binomial foi fixado em $m = 10$.

Como condição inicial para o algoritmo EM, foram consideradas para cada amostra estimativas de μ obtidas pelo método dos momentos. As equações para as distribuições ZMPS consideradas neste estudo são apresentadas na Tabela 5.

Para esse estudo de simulação foram considerados dois cenários para a geração dos conjuntos de dados da distribuições zero-modificadas: Cenário 1, dados gerados de uma distribuição zero-modificada com $p = 1$ (tradicional); Cenário 2: dados gerados de uma distribuição zero-modificada com $1 < p < 1/(1 - \pi_{PS}(0; \mu, \phi))$ (zero-deflacionada). Esses dois cenários foram necessários devido ao fato de que, além das diferenças nas probabilidade da observação zero, existem também diferenças nas probabilidades das observações positivas de cada cenário.

Tabela 5 – Estimadores pelo Método dos Momentos das Distribuições ZMPS consideradas.

PS	**Distribuição**	$\widetilde{\mu}$	$\widetilde{p}$
P	Poisson	$\left(\frac{\sum_{i=0}^{n} Y_i^2}{\sum_{i=0}^{n} Y_i}\right) - 1$	$\frac{\left(\sum_{i=0}^{n} Y_i\right)^2}{n\left(\sum_{i=0}^{n} Y_i^2 - \sum_{i=0}^{n} Y_i\right)}$
G	Geométrica	$\frac{1}{2}\left[\left(\frac{\sum_{i=0}^{n} Y_i^2}{\sum_{i=0}^{n} Y_i}\right) - 1\right]$	$2\frac{\left(\sum_{i=0}^{n} Y_i\right)^2}{n\left(\sum_{i=0}^{n} Y_i^2 - \sum_{i=0}^{n} Y_i\right)}$
B	Binomial	$\frac{m}{m-1}\left[\left(\frac{\sum_{i=0}^{n} Y_i^2}{\sum_{i=0}^{n} Y_i}\right) - 1\right]$	$\frac{m-1}{m}\frac{\left(\sum_{i=0}^{n} Y_i\right)^2}{n\left(\sum_{i=0}^{n} Y_i^2 - \sum_{i=0}^{n} Y_i\right)}$

Fonte: Elaborada pelo autor.

Cenário 1: Dados Gerados de uma Distribuição PS

Neste estudo as N amostras de tamanho n^+ foram geradas da respectiva distribuição PS (isto é, da distribuição ZMPS com parâmetro $p = 1$). Ressalta-se que cada observação amostral foi gerada uma a uma da distribuição PS, descartando os zeros gerados, até que a amostra tivesse as n^+ observações positivas.

A Tabela 6 apresenta o desempenho dos estimadores do parâmetro μ e da probabilidade de zero da distribuição Poisson. Observa-se que as colunas com as medidas de $\sqrt{\frac{EQM}{Var}}$ para os estimadores $\widehat{\mu}$ e $\pi_P(0;\widehat{\mu})$ apresentam valores bastante próximos ou mesmo iguais a $1,00$, assim como baixos valores do vício. Assim, verifica-se que os estimadores são assintoticamente não viciados. As PC's, por sua vez, estão em geral bastante próximas ao valor nominal de 95% para tamanhos amostrais maiores que 50. Ressalta-se que, mesmo para as amostras de tamanho 15 e 25, o desempenho dos estimadores foi satisfatório, apesar dos percentuais abaixo do nível nominal da probabilidade de cobertura.

O desempenho dos estimadores do parâmetro μ e da probabilidade de zero da distribuição Geométrica é apresentado na Tabela 7. As medidas de $\sqrt{\frac{EQM}{Var}}$ para os estimadores $\widehat{\mu}$ e $\pi_G(0;\widehat{\mu})$ apresentam valores bastante próximos ou mesmo iguais a $1,00$, e também apresentam valores baixos do vício, principalmente quando o tamanho da amostra é acrescido, indicando que os estimadores são assintoticamente não viciados. Observa-se também que as PC's estão próximas ao valor nominal e os valores das medidas também melhoram para tamanhos amostrais maiores.

Tabela 6 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero para a distribuição Poisson.

		$\widehat{\mu}$				$\pi_P(0;\widehat{\mu})$		
n^+	μ	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC	$\pi_P(0;\mu)$	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC
15	0,5	0,999	-0,006	88,800	0,607	1,009	0,021	88,800
	1,0	1,001	-0,021	91,400	0,368	1,022	0,026	91,400
	1,5	1,002	0,030	92,200	0,223	1,005	0,010	92,200
25	0,5	0,999	-0,006	89,800	0,607	1,006	0,014	89,800
	1,0	1,002	-0,017	92,000	0,368	1,016	0,017	92,000
	1,5	0,999	-0,006	91,400	0,223	1,012	0,011	91,400
50	0,5	1,010	-0,019	90,800	0,607	1,021	0,017	90,800
	1,0	1,000	-0,008	94,400	0,368	1,008	0,009	94,400
	1,5	1,000	-0,010	93,600	0,223	1,010	0,007	93,600
100	0,5	0,999	-0,001	93,200	0,607	1,001	0,003	93,200
	1,0	1,001	-0,008	93,600	0,368	1,007	0,006	93,600
	1,5	0,999	0,000	93,600	0,223	1,002	0,002	93,600
200	0,5	1,000	-0,003	94,600	0,607	1,002	0,003	94,600
	1,0	0,999	-0,001	94,800	0,368	1,001	0,002	94,800
	1,5	0,999	0,000	95,400	0,223	1,000	0,001	95,400

Fonte: Elaborada pelo autor.

Tabela 7 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero para a distribuição Geométrica.

		$\widehat{\mu}$				$\pi_G(0;\widehat{\mu})$		
n^+	μ	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC	$\pi_G(0;\mu)$	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC
15	0,5	0,999	0,003	90,600	0,667	1,006	0,011	90,600
	1,0	0,999	-0,008	87,800	0,500	1,018	0,018	87,800
	1,5	0,999	0,002	88,400	0,400	1,017	0,016	88,400
25	0,5	0,999	-0,005	91,800	0,500	1,009	0,011	91,800
	1,0	1,002	0,022	90,600	0,500	1,001	0,005	90,600
	1,5	1,000	0,014	92,200	0,400	1,006	0,007	92,200
50	0,5	0,999	-0,001	91,400	0,667	1,003	0,005	91,400
	1,0	1,001	0,013	93,800	0,500	1,000	0,002	93,800
	1,5	1,001	0,019	92,000	0,400	1,000	0,002	92,000
100	0,5	1,002	0,007	93,200	0,667	0,999	-0,001	93,200
	1,0	0,999	0,002	94,200	0,500	1,001	0,002	94,200
	1,5	0,999	0,003	94,000	0,400	1,001	0,002	94,000
200	0,5	0,999	0,000	94,600	0,667	1,000	0,001	94,600
	1,0	1,000	-0,003	95,600	0,500	1,002	0,002	95,600
	1,5	0,999	0,002	94,400	0,400	1,000	0,001	94,400

Fonte: Elaborada pelo autor.

A Tabela 8 apresenta os valores das medidas de desempenho dos estimadores obtidas para o parâmetro μ e probabilidade de zero da distribuição Binomial. Para este estudo, também pode-se notar que as colunas com $\sqrt{\frac{EQM}{Var}}$ têm números próximos ou ainda iguais a $1,00$ para $\widehat{\mu}$ e $\pi_B(0;\widehat{\mu})$. Já para $\mathscr{B}$, os valores obtidos foram baixos. Mais uma vez, tem-se que os estimadores são assintoticamente não viciados. Além disso, tem-se que as PC's apresentam-se próximas a 95%, principalmente quando o tamanho amostral aumenta.

Tabela 8 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero para a distribuição Binomial.

		$\widehat{\mu}$				$\pi_B(0;\widehat{\mu})$		
n^+	μ	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC	$\pi_B(0;\mu)$	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC
15	0,5	0,999	0,001	86,400	0,599	1,005	0,017	86,400
	1,0	1,000	-0,011	86,600	0,349	1,017	0,028	86,600
	1,5	1,000	0,014	90,200	0,197	1,009	0,013	90,200
25	0,5	1,000	-0,007	87,200	0,599	1,007	0,016	87,200
	1,0	0,999	-0,005	92,000	0,349	1,009	0,014	92,000
	1,5	1,000	-0,010	90,400	0,197	1,014	0,012	90,400
50	0,5	1,001	-0,009	91,200	0,599	1,007	0,011	91,200
	1,0	1,001	-0,010	91,200	0,349	1,009	0,011	91,200
	1,5	0,999	-0,003	93,200	0,197	1,006	0,005	93,200
100	0,5	0,999	-0,001	93,800	0,599	1,001	0,004	94,800
	1,0	0,999	0,003	94,400	0,349	1,000	0,002	94,400
	1,5	1,000	-0,005	94,000	0,197	1,005	0,004	94,000
200	0,5	1,000	-0,003	94,400	0,599	1,002	0,003	94,400
	1,0	1,003	0,007	95,600	0,349	1,000	-0,001	95,600
	1,5	0,999	0,001	94,600	0,197	1,000	0,001	94,600

Fonte: Elaborada pelo autor.

Cenário 2: Dados Gerados de uma Distribuição ZDPS

Neste estudo as N amostras de tamanho n^+ foram geradas da respectiva distribuição ZDPS (isto é, com parâmetro $1 < p < 1/(1-\pi_{PS}(0;\mu,\phi))$). O parâmetro p utilizado foi escolhido de tal forma a impactar em 95% de redução na probabilidade de zero em relação à distribuição PS tradicional. Vale ressaltar que outro percentual de redução poderia ser considerado, contudo este alto valor aborda um caso extremo de zero-deflação já que o valor de p será próximo do seu limite superior. A Tabela 9 apresenta os valores dos parâmetros das distribuições ZMPS (mais especificamente ZDPS) utilizados para geração dos dados neste estudo.

Tabela 9 – Valores dos parâmetros da distribuição ZDPS utilizados para a geração dos conjuntos de dados.

μ	p		
	ZDP	ZDG	ZDB ($m = 10$)
0,5	2,46	2,90	2,42
1,0	1,55	1,95	1,51
1,5	1,27	1,63	1,23

Fonte: Elaborada pelo autor.

A Tabela 10 apresenta o desempenho dos estimadores do parâmetro μ e da probabilidade de zero da distribuição Poisson a partir da análise de dados provenientes de uma distribuição ZDP. Observa-se comportamento similar ao verificado anteriormente para a distribuição Poisson (tradicional), em que as colunas com as medidas de $\sqrt{\frac{EQM}{Var}}$ para os estimadores $\widehat{\mu}$ e $\pi_P(0;\widehat{\mu})$ apresentam valores bastante próximos ou mesmo iguais a $1,00$, assim como baixos valores do vício. As medidas da tabela melhoram conforme o aumento do tamanho das amostras; assim, verifica-se que os estimadores são assintoticamente não viciados. As PC's, por sua vez, estão em geral bastante próximas ao valor nominal de 95%. Ressalta-se novamente que, para os conjuntos de menor tamanho amostral apresentados (15 e 25), o desempenho dos estimadores segundo às medidas propostas também foi razoável, apresentando valores de $\sqrt{\frac{EQM}{Var}}$ bastante próximos de 1, bem como valores de vício próximos de zero e PC's apenas um pouco mais baixas que o valor nominal. Este comportamento é verificado para as demais situações (distribuições tradicionais e deflacionadas) apresentadas neste estudo.

Já a Tabela 11 apresenta o desempenho dos estimadores de μ e da probabilidade de zero da distribuição Geométrica a partir da análise de dados provenientes de uma distribuição ZDG. As medidas de $\sqrt{\frac{EQM}{Var}}$ para os estimadores $\widehat{\mu}$ e $\pi_G(0;\widehat{\mu})$ apresentam valores bastante próximos ou mesmo iguais a $1,00$, e também valores baixos do vício. Observa-se também que as PC's estão próximas ao valor nominal considerado (de 95%). Os valores das medidas de avaliação dos estimadores melhoram conforme o tamanho das amostras aumenta, indicando que os estimadores são assintoticamente não viciados, bem como há melhora das probabilidades de cobertura.

Tabela 10 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero da distribuição Poisson a partir de dados zero-deflacionados.

		$\widehat{\mu}$				$\pi_P(0;\widehat{\mu})$		
n^+	μ	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC	$\pi_P(0;\mu)$	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC
15	0,5	1,000	-0,013	87,200	0,607	1,013	0,025	87,200
	1,0	1,000	-0,014	89,400	0,368	1,019	0,026	89,400
	1,5	1,000	0,014	91,600	0,223	1,009	0,014	91,600
25	0,5	1,000	0,008	90,800	0,607	1,000	0,006	90,800
	1,0	0,999	-0,005	91,800	0,368	1,009	0,013	91,800
	1,5	1,002	-0,023	92,000	0,223	1,022	0,015	92,000
50	0,5	1,004	-0,012	93,400	0,607	1,012	0,012	93,400
	1,0	1,001	-0,001	92,200	0,368	1,010	0,010	92,200
	1,5	0,999	-0,002	94,200	0,223	1,005	0,005	94,200
100	0,5	0,999	-0,002	93,600	0,607	1,001	0,004	93,600
	1,0	1,000	0,005	95,200	0,368	0,999	0,001	95,200
	1,5	1,000	-0,006	93,000	0,223	1,006	0,004	93,000
200	0,5	0,999	-0,001	94,200	0,607	1,000	0,002	94,200
	1,0	0,999	0,001	94,400	0,368	1,000	0,001	94,400
	1,5	0,999	-0,002	94,600	0,223	1,002	0,002	94,600

Fonte: Elaborada pelo autor.

Tabela 11 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero da distribuição Geométrica a partir de dados zero-deflacionados.

		$\widehat{\mu}$				$\pi_G(0;\widehat{\mu})$		
n^+	μ	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC	$\pi_G(0;\mu)$	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC
15	0,5	1,002	-0,017	87,600	0,667	1,023	0,022	87,600
	1,0	1,004	-0,035	86,400	0,500	1,034	0,024	86,400
	1,5	0,999	0,007	89,600	0,400	1,015	0,015	89,600
25	0,5	0,999	0,003	90,200	0,667	1,004	0,008	90,200
	1,0	0,999	0,000	91,000	0,500	1,008	0,010	91,000
	1,5	0,999	-0,009	92,600	0,400	1,013	0,010	92,600
50	0,5	1,000	-0,006	90,800	0,667	1,008	0,007	90,800
	1,0	0,999	-0,002	93,400	0,500	1,005	0,005	93,400
	1,5	0,999	0,001	92,800	0,400	1,004	0,004	92,800
100	0,5	0,999	-0,001	94,800	0,667	1,001	0,003	94,800
	1,0	0,999	0,000	92,600	0,500	1,002	0,003	92,600
	1,5	0,999	0,002	94,200	0,400	1,001	0,002	94,200
200	0,5	1,000	-0,003	96,000	0,667	1,003	0,002	96,000
	1,0	1,000	0,004	94,000	0,500	0,999	0,000	94,000
	1,5	0,999	0,002	94,200	0,400	1,000	0,001	94,200

Fonte: Elaborada pelo autor.

As medidas de desempenho dos estimadores de μ e da probabilidade de zero, bem como as probabilidades de coberturas, da distribuição Binomial obtidas a partir da análise de dados provenientes de uma distribuição ZDB são apresentados na Tabela 12. Os resultados são similares aos apresentados para a distribuição Binomial (tradicional). As colunas com $\sqrt{\frac{EQM}{Var}}$ têm números próximos ou ainda iguais a $1,00$. Os valores de $\mathscr{B}$ foram baixos. Há melhora das medidas de desempenho conforme o tamanho das amostras é acrescido, apontando que os estimadores são assintoticamente não viciados. Além disso, as PC's apresentam-se próximas do nível nominal estabelecido de 95%.

Tabela 12 – Medidas de desempenho dos estimadores de μ e da probabilidade de zero da distribuição Binomial a partir de dados zero-deflacionados.

		$\widehat{\mu}$				$\pi_B(0;\widehat{\mu})$		
n^+	μ	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC	$\pi_B(0;\mu)$	$\sqrt{\frac{EQM}{Var}}$	$\mathscr{B}$	PC
15	0,5	1,001	-0,016	85,200	0,599	1,015	0,028	85,200
	1,0	0,999	-0,001	88,800	0,349	1,010	0,019	88,800
	1,5	0,999	-0,002	89,600	0,197	1,015	0,018	89,600
25	0,5	1,000	-0,008	89,000	0,599	1,008	0,016	89,000
	1,0	1,000	-0,009	93,000	0,349	1,011	0,015	93,000
	1,5	1,002	-0,022	92,800	0,197	1,021	0,014	92,800
50	0,5	0,999	0,003	93,400	0,599	1,000	0,003	93,400
	1,0	0,999	-0,003	94,400	0,349	1,004	0,007	94,400
	1,5	0,999	0,000	95,200	0,197	1,003	0,004	95,200
100	0,5	1,001	-0,006	93,000	0,599	1,004	0,006	93,000
	1,0	0,999	-0,002	95,200	0,349	1,002	0,003	95,200
	1,5	0,999	0,001	95,800	0,197	1,001	0,002	95,800
200	0,5	0,999	-0,002	95,200	0,599	1,001	0,002	95,200
	1,0	0,999	0,003	95,400	0,349	0,999	0,000	95,400
	1,5	0,999	-0,003	96,200	0,197	1,002	0,002	96,200

Fonte: Elaborada pelo autor.

4.2 Estudo com Dados Artificiais: Aplicações

Com o intuito de ilustrar a eficácia do procedimento proposto para a estimação do parâmetro μ da distribuição ZMPS e, consequentemente, da probabilidade de zero da PS associada nas situações em que os conjuntos de dados não possuem a observação zero mas esta tem probabilidade positiva de ocorrer, foi realizado também um estudo que consiste na análise de um conjunto dados artificialmente gerados com cada valor de parâmetro μ e p considerados no estudo anterior (estudo de simulação), bem como para os tamanhos amostrais n^+, a fim de comparar as estimativas obtidas com os seus verdadeiros valores. Para este estudo, foram consideradas as distribuições ZMP, ZMG e ZMB, considerando também os dois cenários descritos anteriormente.

Adicionalmente para o procedimento *bootstrap*, além de fornecer o intervalo de confiança a partir de $B = 5000$ réplicas, as estimativas obtidas com cada amostra *bootstrap* foram utilizadas para obtenção de uma estimativa corrigida (ver Lemonte, Simas e Cribari-Neto (2008)), que consiste em:

$$\widehat{\theta}_C = 2\widehat{\theta} - \widehat{\theta}^{*(\cdot)},$$

em que

$$\widehat{\theta}^{*(\cdot)} = \frac{1}{B}\sum_{b=1}^{B} \widehat{\theta}^{*(b)},$$

sendo $\widehat{\theta}^{*(b)}$ a estimativa obtida de cada amostra *bootstrap*, $b = 1,\ldots,B$; e $\theta = \{\mu, \phi, \pi_0\}$.

Finalmente, as estimativas obtidas para os parâmetros foram comparadas com os seus verdadeiros valores, incluindo as estimativas corrigidas, e verificou-se também se os verdadeiros valores pertencem aos intervalos de confiança *bootstrap* para μ e $\pi_{PS}(0;\mu)$, respectivamente.

Cenário 1: Dados Gerados de uma Distribuição PS

Para cada valor de μ (μ =0,5; 1 e 1,5) e p ($p = 1$) uma amostra de tamanho n^+ (n^+ =15, 25, 50, 100 e 200) foi gerada de cada distribuição ZMPS. Ou seja, um conjunto de dados de cada respectiva distribuição PS considerada neste estudo. Vale frisar que cada observação amostral foi gerada uma a uma, descartando os zeros gerados, até a obtenção de n^+ observações positivas.

Na Tabela 13 são apresentados os resultados obtidos para a distribuição Poisson. Observa-se que as estimativas são próximas aos valores verdadeiros (μ e $\pi_p(0;\mu)$), bem como estes valores verdadeiros pertencem aos intervalos *bootstrap* com 95% de confiança. As estimativas corrigidas, calculadas utilizando-se as estimativas obtidas pelo algoritmo EM para as reamostras *bootstrap*, foram próximas àquelas resultantes da aplicação direta do método. São apresentados também nesta Tabela o valor de n_0 gerado (e desconsiderado no procedimento de estimação, que utiliza apenas os valores positivos), e a estimativa de n_0, a qual possibilitou a boa estimativa de μ e, consequentemente, da probabilidade de zero. Verifica-se que os valores gerados e estimados de n_0 foram próximos. Adicionalmente, constata-se que os valores estimados apresentados melhoram

conforme o aumento do tamanho das amostras; assim, verifica-se que o método produz boas estimativas.

Tabela 13 – Estimativas de n_0, μ e da probabilidade de zero da distribuição Poisson.

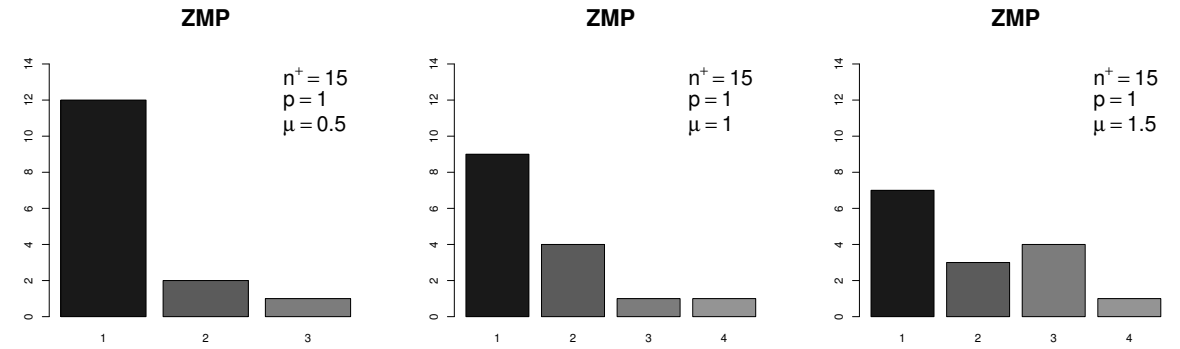
n^+	μ	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_P(0;\mu)$	$\pi_P(0;\widehat{\mu})$	$\pi_P(0;\widehat{\mu}_C)$	n_0	$\widehat{n}_0$
15	0,5	0,493 (0,001; 1,027)	0,508	0,607	0,611 (0,358; 1,000)	0,583	15	23,538
	1,0	1,027 (0,376; 1,682)	1,047	0,368	0,358 (0,186; 0,686)	0,330	8	8,365
	1,5	1,503 (0,822; 2,191)	1,515	0,223	0,222 (0,112; 0,440)	0,206	5	4,289
25	0,5	0,516 (0,231; 0,843)	0,518	0,607	0,597 (0,430; 0,794)	0,588	39	37,038
	1,0	1,087 (0,650; 1,540)	1,095	0,368	0,337 (0,214; 0,522)	0,325	25	12,723
	1,5	1,540 (1,026; 2,109)	1,543	0,223	0,214 (0,121; 0,359)	0,205	10	6,825
50	0,5	0,481 (0,268; 0,683)	0,483	0,607	0,618 (0,505; 0,765)	0,614	90	80,849
	1,0	1,057 (0,715; 1,402)	1,061	0,368	0,347 (0,246; 0,489)	0,341	35	26,622
	1,5	1,541 (1,087; 2,009)	1,541	0,223	0,214 (0,134; 0,337)	0,208	10	13,650
100	0,5	0,499 (0,305; 0,699)	0,501	0,607	0,607 (0,497; 0,737)	0,603	122	154,638
	1,0	0,997 (0,795; 1,204)	0,998	0,368	0,369 (0,300; 0,451)	0,367	62	58,476
	1,5	1,526 (1,262; 1,792)	1,527	0,223	0,217 (0,167; 0,283)	0,215	31	27,776
200	0,5	0,507 (0,368; 0,658)	0,507	0,607	0,602 (0,518; 0,692)	0,601	326	302,643
	1,0	0,982 (0,804; 1,160)	0,982	0,368	0,375 (0,313; 0,448)	0,373	106	119,801
	1,5	1,485 (1,283; 1,687)	1,487	0,223	0,226 (0,185; 0,277)	0,225	64	58,551

Fonte: Elaborada pelo autor.

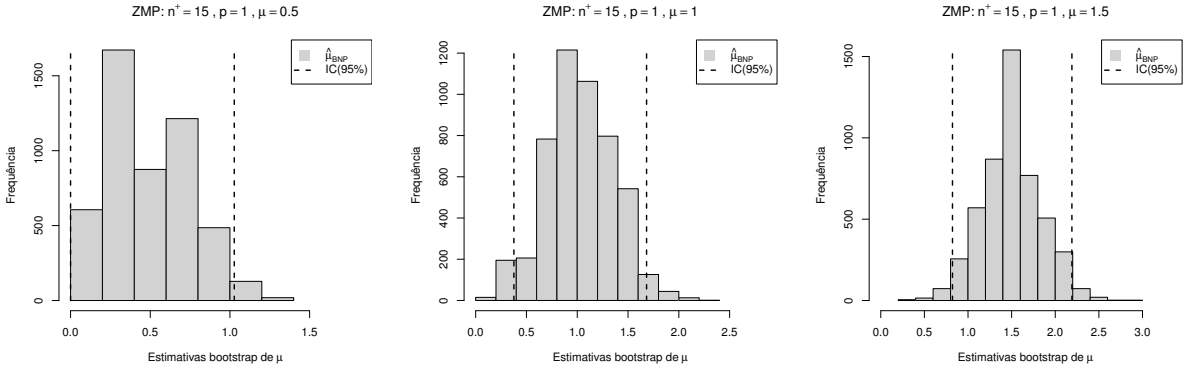
As Figuras 6 - 10 apresentam as análises gráficas para as amostras geradas (sem o valor zero) a partir da distribuição Poisson, com os diferentes tamanhos amostrais e valores do parâmetro μ, cujas estimativas foram apresentadas na Tabela 13. Nestas Figuras, os gráficos apresentados em (a) são as distribuições de frequências observada dos conjuntos de dados, nas quais verificam-se visualmente, dentre outros aspectos, o aumento na amplitude dos dados à medida em que o valor de μ aumenta. A escala em degradê em tons de cinza ressalta essa amplitude, atribuindo uma cor diferente para cada observação particular da amostra. É possível notar também para todos os casos que baixa (alta) amplitude ocasiona maiores (menores) frequências observadas. Ainda considerando estas Figuras, nos gráficos em (b), observa-se as distribuições empíricas para o estimador do parâmetro μ, obtidas via BNP, com os percentis de $2,5\%$ e $97,5\%$ em destaque. Essas apresentam-se aproximadamente simétricas para todos os conjuntos amostrais, com maior concentração em torno dos valores verdadeiros do parâmetro e verifica-se também pouca variabilidade em torno deste. Por sua vez, nos gráficos em (c), são comparadas as frequências observadas (f_i) e as frequências esperadas (f_e) obtidas a partir da distribuição tradicional ajustada aos valores positivos dos conjuntos. Devido à baixa discrepância, tem-se fortes evidências da aderência da distribuição ajustada aos conjuntos de dados estudados.

Destaca-se que os comportamentos acima descritos são observados para todos os conjuntos amostrais e valores de parâmetros apresentados, verificando-se a melhora nas características com o aumento dos tamanhos amostrais, embora para pequenas amostras os resultados também sejam satisfatórios. Assim, evidencia-se a adequabilidade da metodologia considerada às situações estudadas.

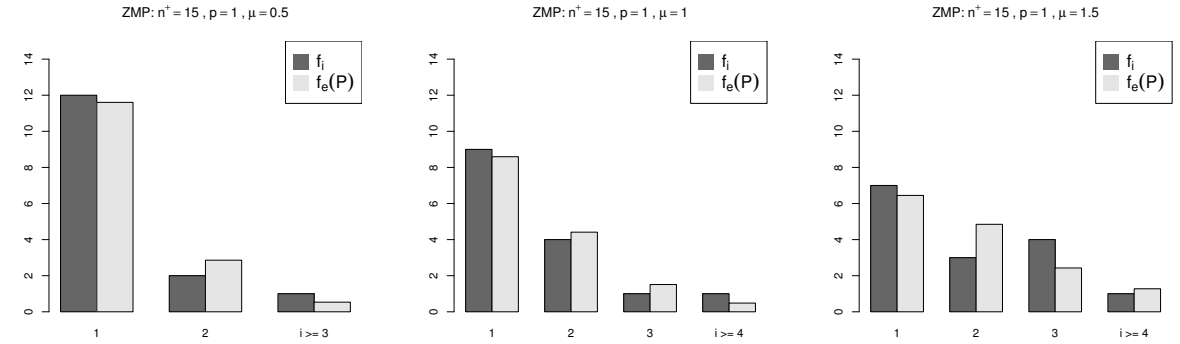
Figura 6 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



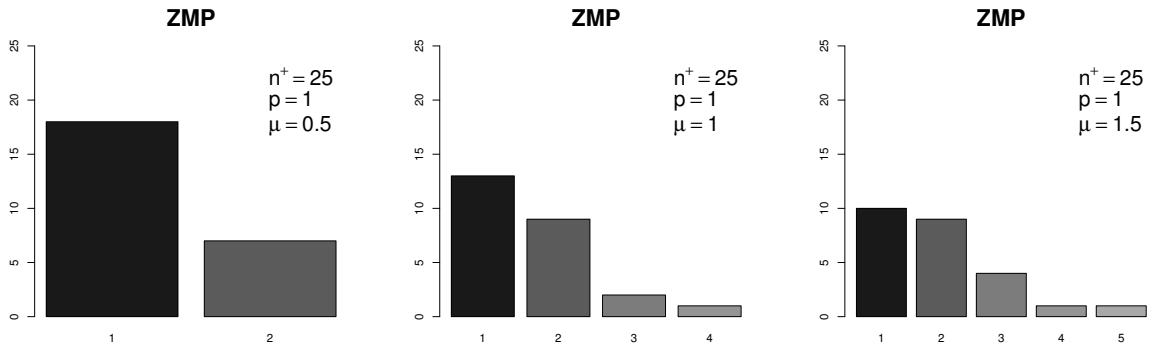
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



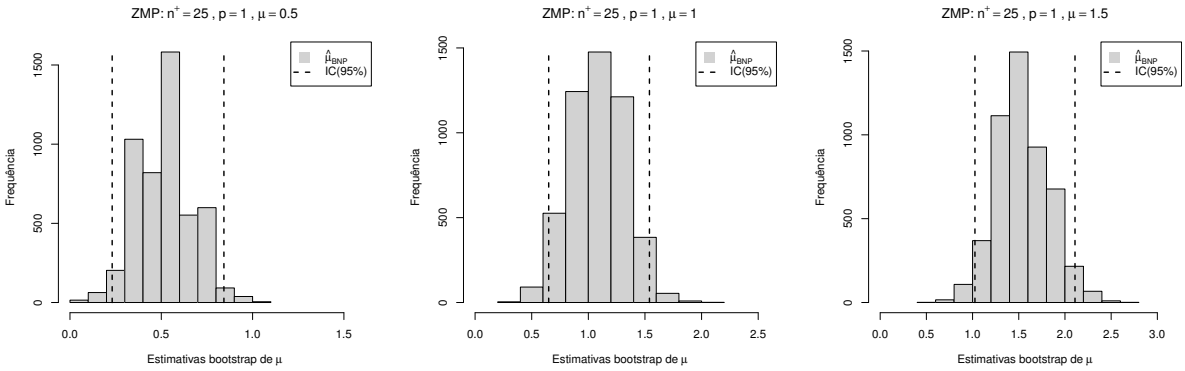
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

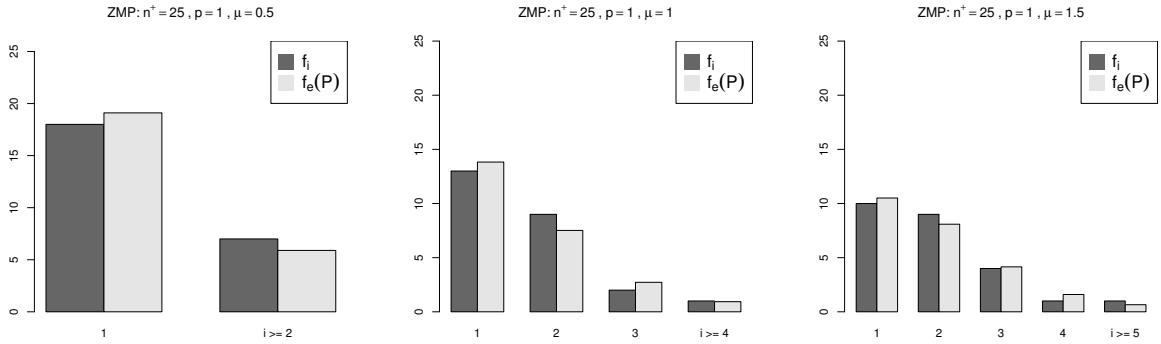
Figura 7 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



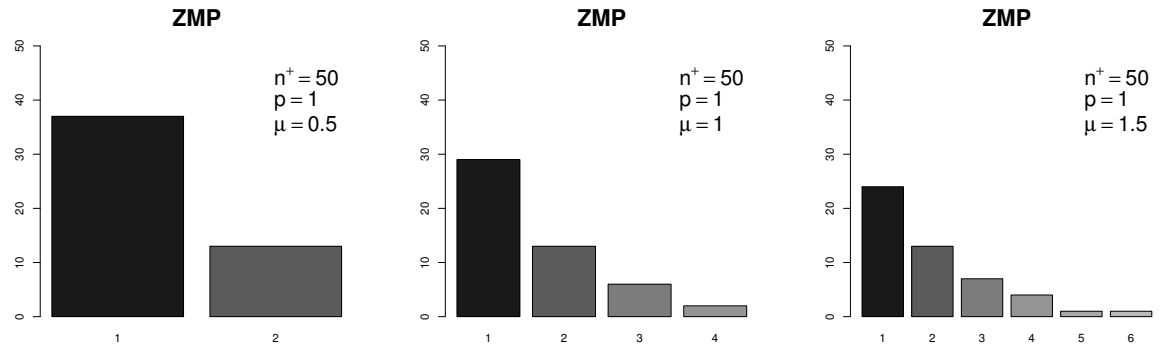
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



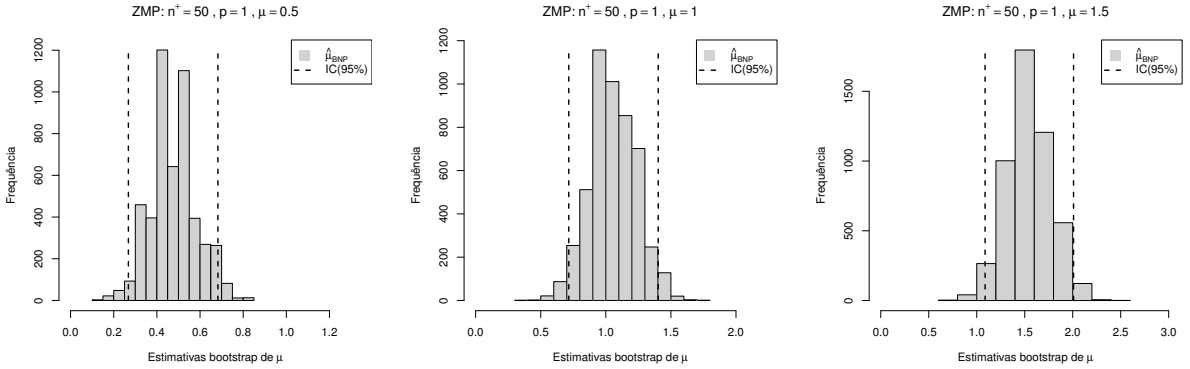
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

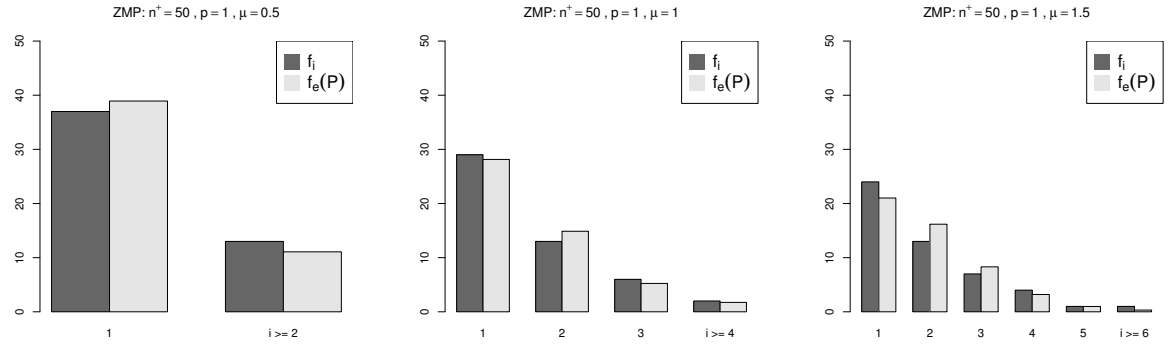
Figura 8 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



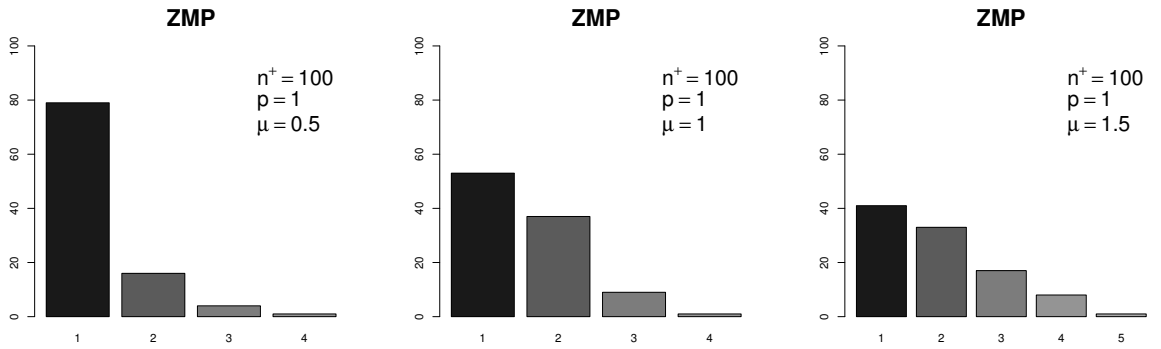
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



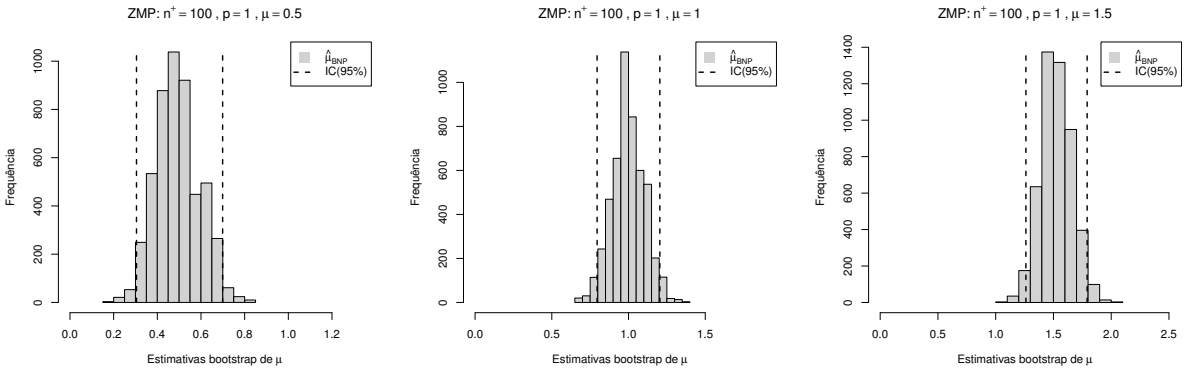
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

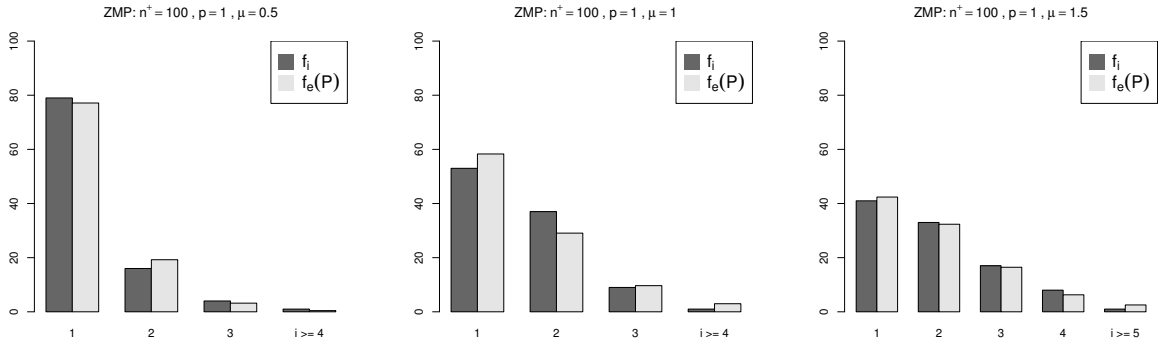
Figura 9 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



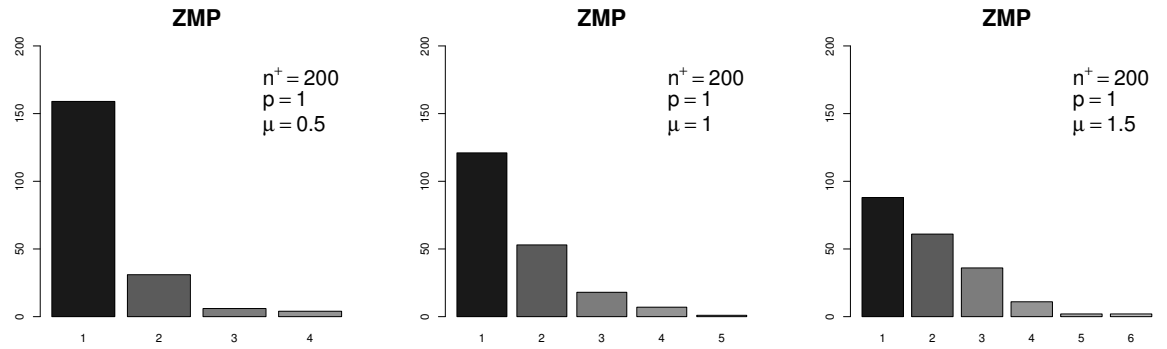
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



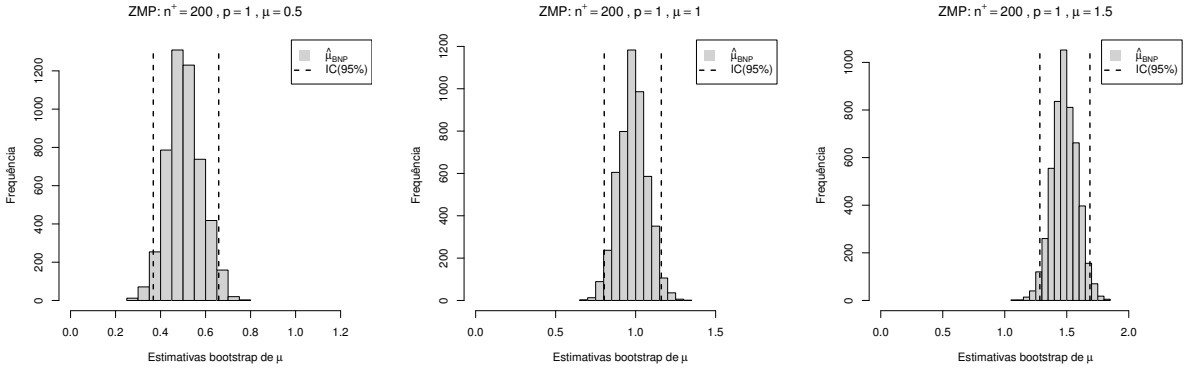
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

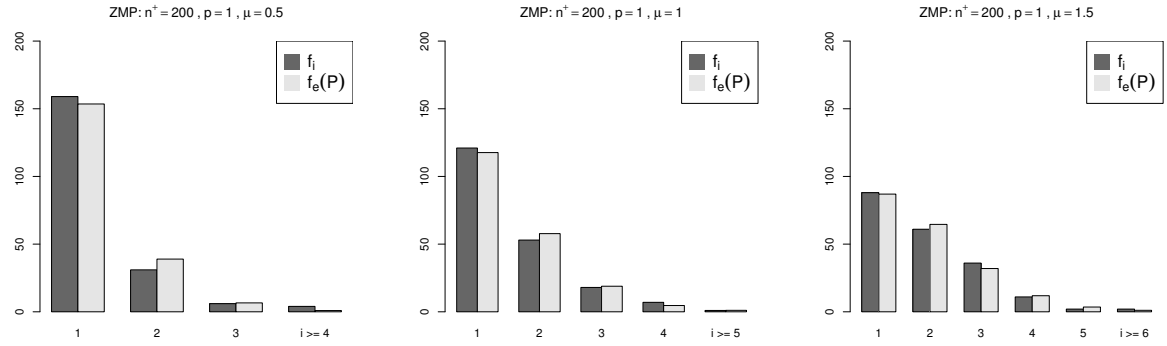
Figura 10 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Poisson (ZMP, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

Os resultados obtidos para a distribuição Geométrica são apresentados na Tabela 14. Verifica-se que as estimativas são próximas dos valores verdadeiros (μ e $\pi_G(0;\mu)$) considerados para geração dos dados, assim como estes valores verdadeiros pertencem aos respectivos intervalos de confiança *bootstrap*. Similarmente, as estimativas corrigidas (que utilizaram os valores calculados via aplicação do procedimento às réplicas *bootstrap*) foram próximas àquelas obtidas via aplicação direta algoritmo EM. Além disso, tem-se os valores gerados (e desconsiderados na estimação) e estimados de n_0 para cada amostra, que contribuíram para a obtenção de boas estimativas e foram próximos aos gerados. Por fim, nota-se também erros menores nas estimativas de μ e $\pi_G(0;\mu)$ quando os tamanhos amostrais são acrescidos.

Tabela 14 – Estimativas de n_0, μ e da probabilidade de zero da distribuição Geométrica.

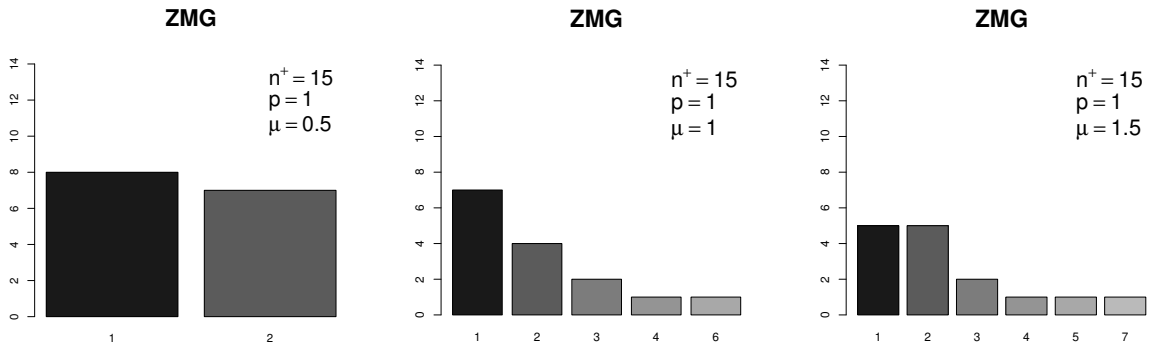
n^+	μ	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_G(0;\mu)$	$\pi_G(0;\widehat{\mu})$	$\pi_G(0;\widehat{\mu}_C)$	n_0	$\widehat{n}_0$
15	0,5	0,467 (0,200; 0,733)	0,470	0,667	0,682 (0,577; 0,833)	0,675	38	32,144
	1,0	1,067 (0,467; 1,800)	1,068	0,500	0,484 (0,357; 0,682)	0,469	22	14,063
	1,5	1,467 (0,733; 2,400)	1,468	0,400	0,405 (0,294; 0,577)	0,392	9	10,227
25	0,5	0,520 (0,280; 0,760)	0,520	0,667	0,658 (0,568; 0,781)	0,653	45	48,079
	1,0	1,000 (0,600; 1,400)	1,001	0,500	0,500 (0,417; 0,625)	0,494	33	25,000
	1,5	1,480 (0,920; 2,120)	1,477	0,400	0,403 (0,321; 0,521)	0,397	15	16,892
50	0,5	0,500 (0,260; 0,780)	0,503	0,667	0,667 (0,562; 0,794)	0,660	84	99,995
	1,0	0,980 (0,680; 1,320)	0,980	0,500	0,510 (0,431; 0,595)	0,502	37	51,020
	1,5	1,540 (0,960; 2,201)	1,538	0,400	0,394 (0,312; 0,510)	0,388	23	32,467
100	0,5	0,490 (0,340; 0,650)	0,491	0,667	0,671 (0,606; 0,746)	0,669	189	204,072
	1,0	1,040 (0,780; 1,320)	1,035	0,500	0,490 (0,431; 0,562)	0,489	105	96,153
	1,5	1,480 (1,120; 1,870)	1,485	0,400	0,403 (0,348; 0,472)	0,400	57	67,567
200	0,5	0,500 (0,395; 0,610)	0,501	0,667	0,667 (0,621; 0,717)	0,665	370	399,977
	1,0	1,005 (0,825; 1,190)	1,005	0,500	0,499 (0,457; 0,548)	0,498	184	199,002
	1,5	1,505 (1,235; 1,805)	1,502	0,400	0,399 (0,357; 0,447)	0,398	133	132,889

Fonte: Elaborada pelo autor.

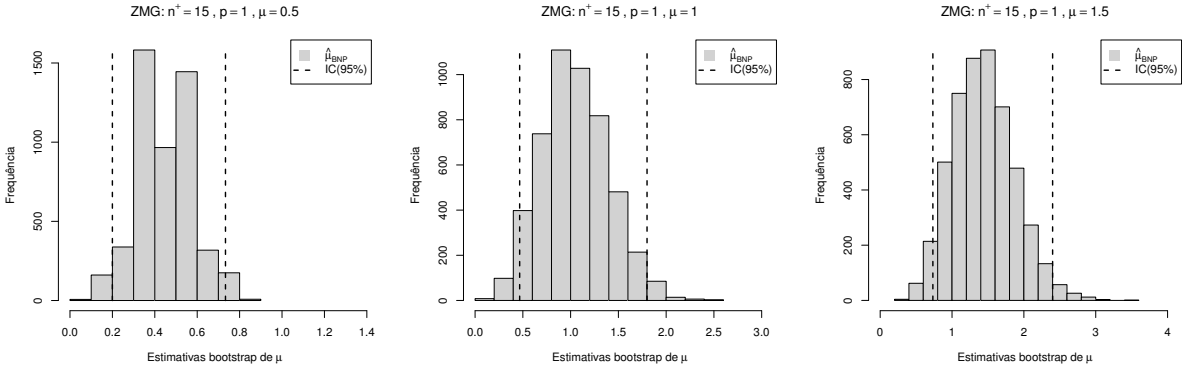
As Figuras 11 - 15 apresentam as análises gráficas para as amostras geradas (sem o valor zero) a partir da distribuição Geométrica, com os diferentes tamanhos amostrais e valores do parâmetro μ, cujas estimativas foram apresentadas na Tabela 14. Nas Figuras, os gráficos em (a) são as distribuições de frequências para os conjuntos. Com a análise dos gráficos, verifica-se o aumento da amplitude dos dados à medida em que o valor da média aumenta; novamente, degradê em tons de cinza ressalta a amplitude e observa-se para todos os casos que baixa amplitude ocasiona em maiores frequências observadas bem como alta amplitude ocasiona em menores frequências. Nos gráficos apresentados em (b) ainda nestas Figuras, observa-se as distribuições empíricas para o estimador do parâmetro μ, obtidas via BNP, com os percentis de $2,5\%$ e $97,5\%$ em destaque. Verifica-se que há uma leve assimetria para todos os conjuntos amostrais, e que há maior concentração das frequências em torno dos valores verdadeiros do parâmetro; constata-se também pouca variabilidade em torno deste. Por sua vez, nos gráficos em (c), as frequências observadas (f_i) e esperadas (f_e) pela distribuição tradicional ajustada para os valores positivos dos conjuntos são comparadas. Observa-se que as barras f_i e f_e têm alturas próximas, evidenciando que a suposição da distribuição foi adequada aos conjuntos de dados estudados.

Mais uma vez, ressalta-se que os comportamentos acima descritos são observados para todos os tamanhos de amostras e valores de parâmetros considerados neste estudo, verificando-se a melhora nas características para maiores conjuntos amostrais, embora para aqueles de menor tamanho os resultados também sejam razoáveis, corroborando a qualidade da metodologia considerada neste trabalho.

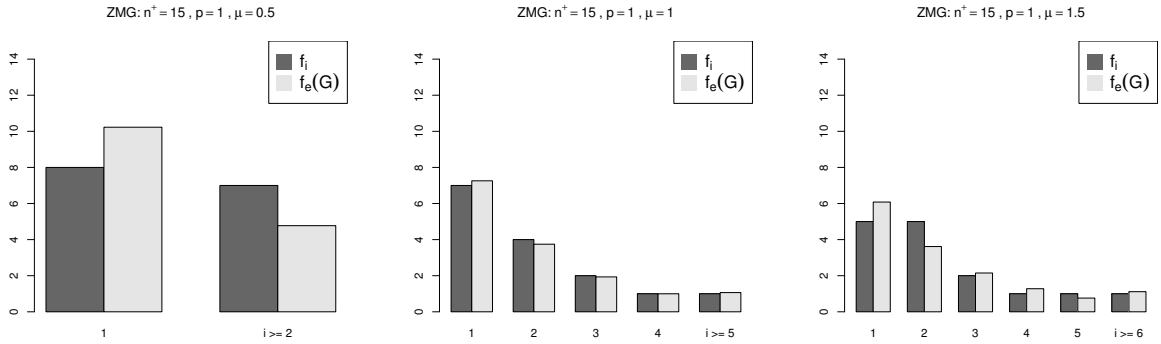
Figura 11 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



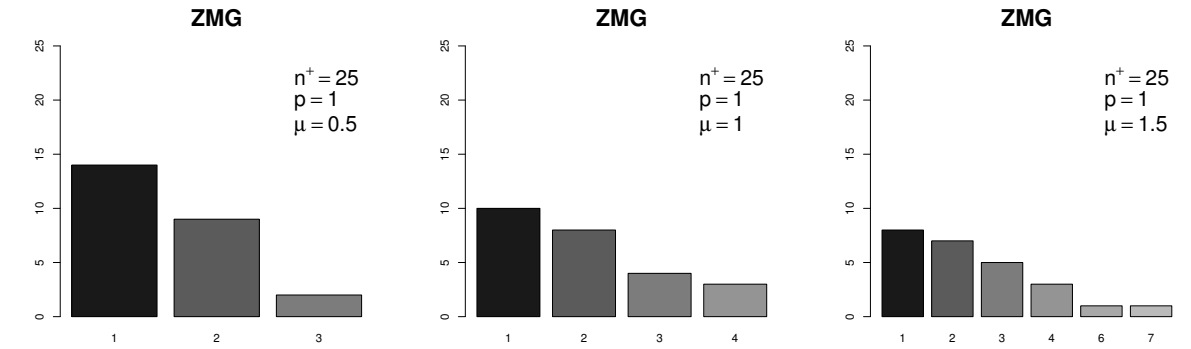
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



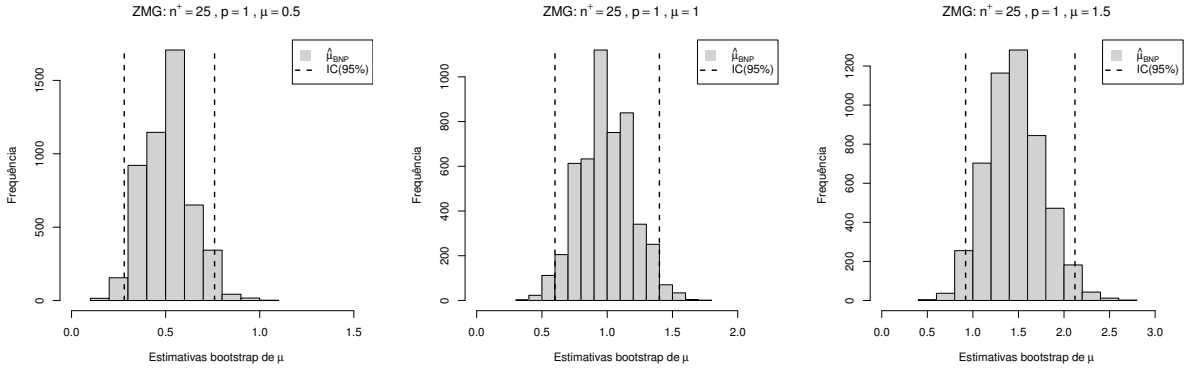
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

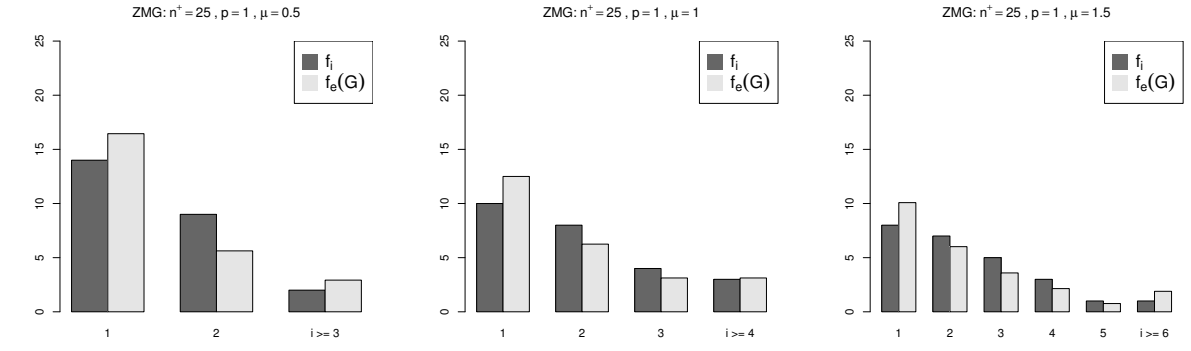
Figura 12 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



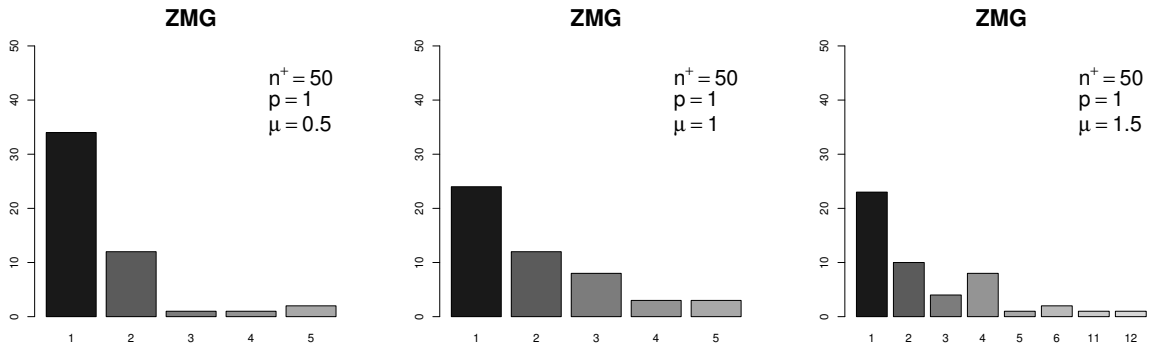
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



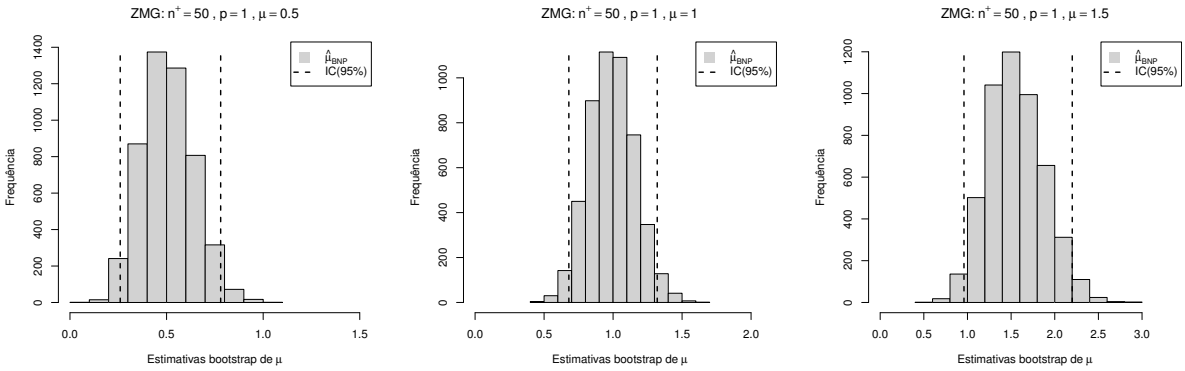
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

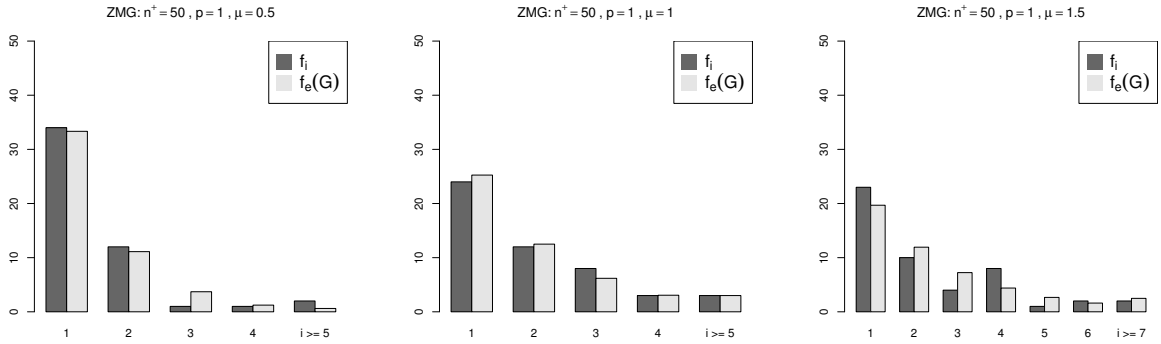
Figura 13 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



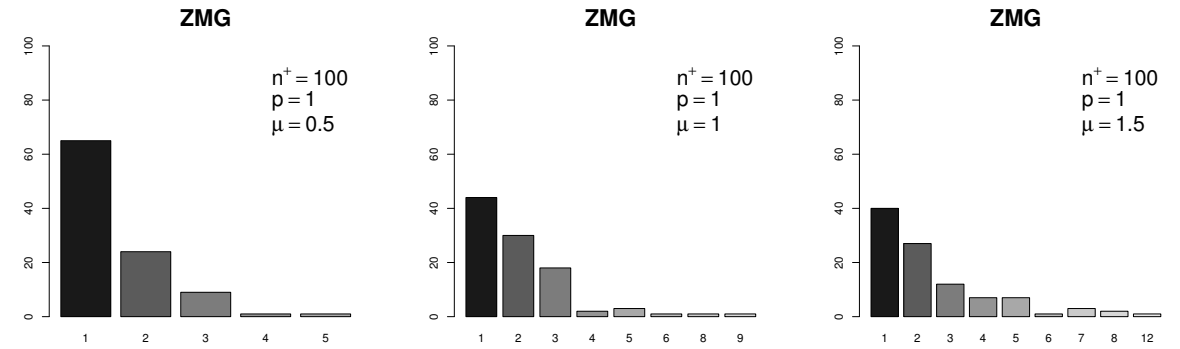
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



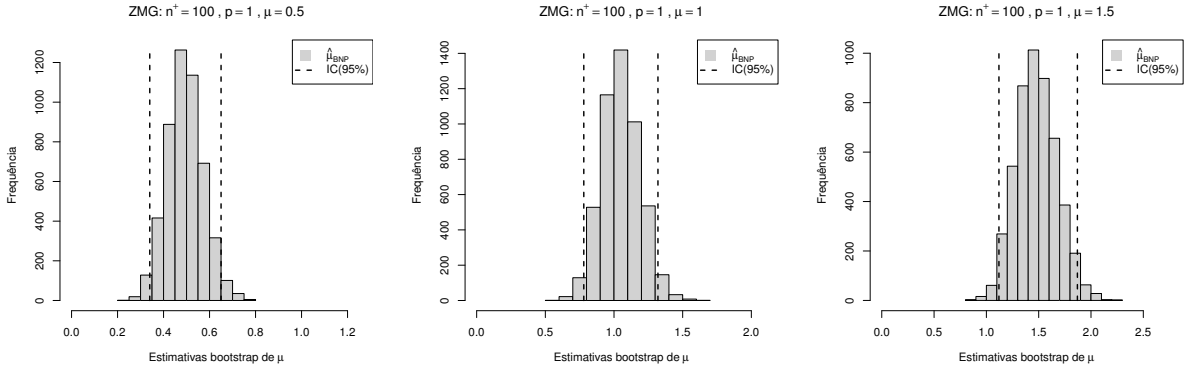
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

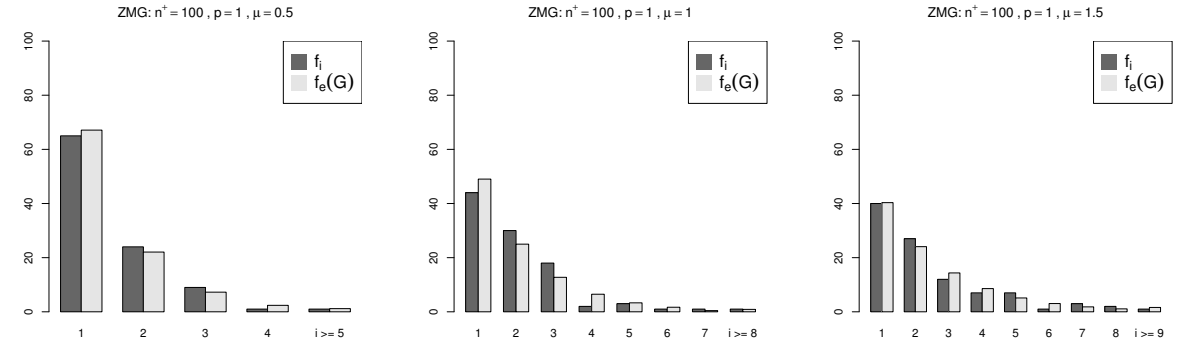
Figura 14 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



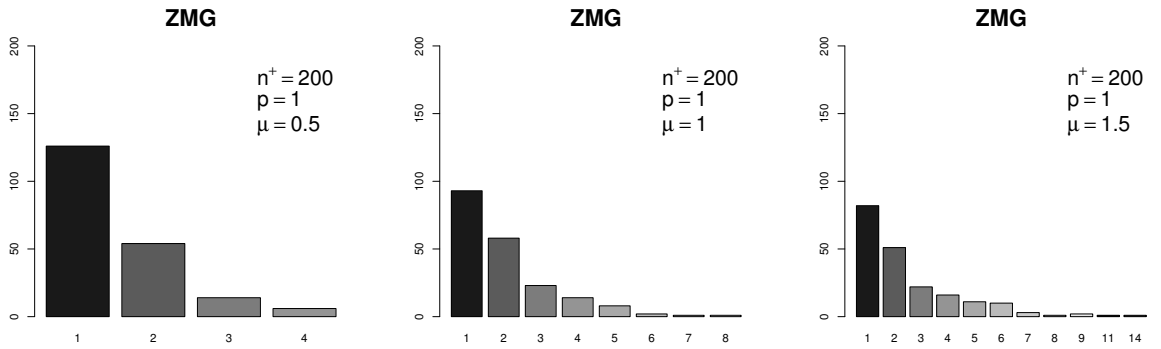
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



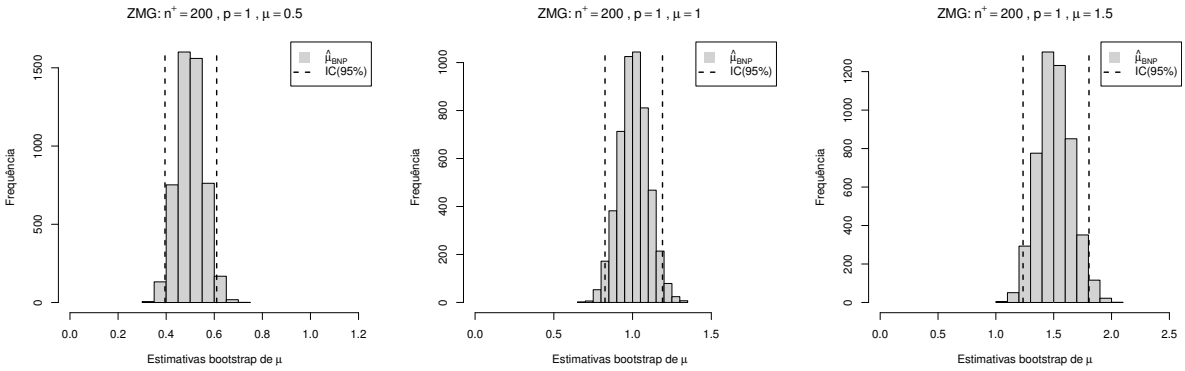
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

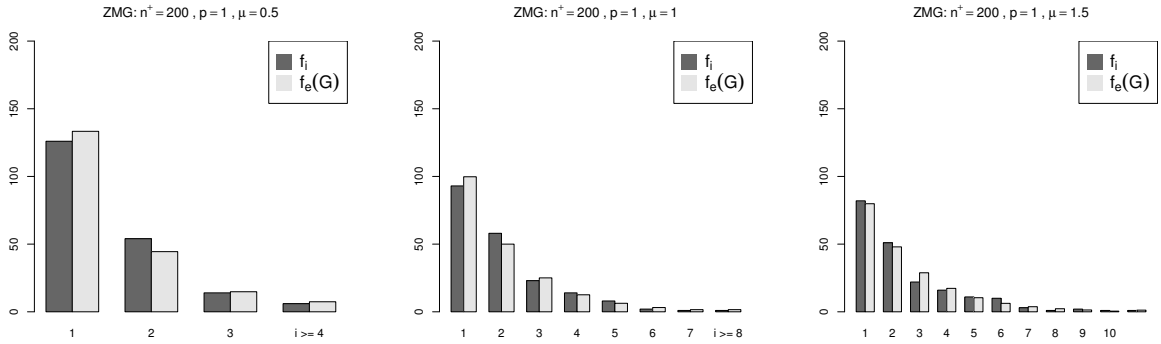
Figura 15 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Geométrica (ZMG, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

A Tabela 15 apresenta as estimativas obtidas de n_0, μ e $\pi_B(0;\mu)$ para a distribuição Binomial. Verifica-se que os valores resultantes da aplicação do método proposto são próximos dos valores verdadeiros (μ e $\pi_B(0;\mu)$) e estes valores verdadeiros pertencem aos respectivos intervalos *bootstrap* com 95% de confiança. As estimativas corrigidas (obtidas a partir das réplicas *bootstrap*) tiveram resultados próximos aos obtidos pela aplicação direta do algoritmo EM às amostras. Os valores de n_0 gerados para as amostras e os estimados pelo procedimento ao considerar-se apenas as observações positivas foram próximos. Para maiores tamanhos de amostra, há a melhora nas estimativas, corroborando que o método de estimação considerado produz bons resultados.

Tabela 15 – Estimativas de n_0, μ e da probabilidade de zero da distribuição Binomial, com $m = 10$.

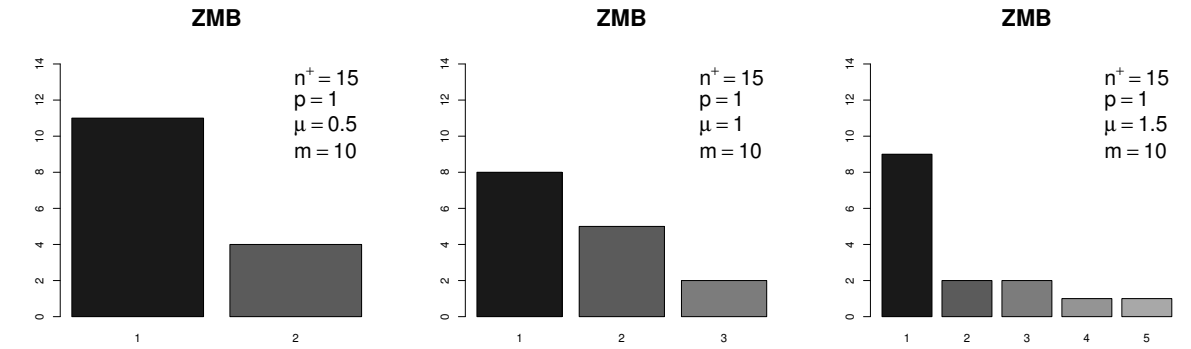
n^+	μ	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_B(0;\mu)$	$\pi_B(0;\widehat{\mu})$	$\pi_B(0;\widehat{\mu}_C)$	n_0	$\widehat{n}_0$
15	0,5	0,538 (0,144; 0,888)	0,548	0,599	0,575 (0,395; 0,865)	0,556	23	20,300
	1,0	1,102 (0,538; 1,683)	1,115	0,349	0,311 (0,158; 0,575)	0,292	5	6,769
	1,5	1,499 (0,538; 2,362)	1,532	0,197	0,197 (0,068; 0,575)	0,161	3	3,685
25	0,5	0,489 (0,172; 0,843)	0,497	0,599	0,606 (0,414; 0,841)	0,590	45	38,446
	1,0	1,039 (0,489; 1,573)	1,049	0,349	0,334 (0,181; 0,606)	0,315	14	12,521
	1,5	1,517 (0,975; 2,100)	1,533	0,197	0,193 (0,095; 0,358)	0,179	10	5,975
50	0,5	0,526 (0,295; 0,775)	0,527	0,599	0,583 (0,446; 0,741)	0,578	68	69,792
	1,0	1,007 (0,706; 1,315)	1,010	0,349	0,346 (0,244; 0,481)	0,340	38	26,429
	1,5	1,489 (1,134; 1,817)	1,494	0,197	0,199 (0,135; 0,300)	0,194	13	12,455
100	0,5	0,489 (0,295; 0,689)	0,492	0,599	0,606 (0,490; 0,741)	0,600	120	153,783
	1,0	1,055 (0,792; 1,330)	1,058	0,349	0,328 (0,240; 0,438)	0,323	47	48,775
	1,5	1,503 (1,241; 1,764)	1,504	0,197	0,196 (0,144; 0,266)	0,194	23	24,398
200	0,5	0,507 (0,374; 0,644)	0,508	0,599	0,594 (0,514; 0,683)	0,592	315	292,791
	1,0	0,983 (0,818; 1,157)	0,982	0,349	0,355 (0,292; 0,426)	0,354	89	110,171
	1,5	1,489 (1,300; 1,677)	1,490	0,197	0,199 (0,160; 0,248)	0,198	41	49,819

Fonte: Elaborada pelo autor.

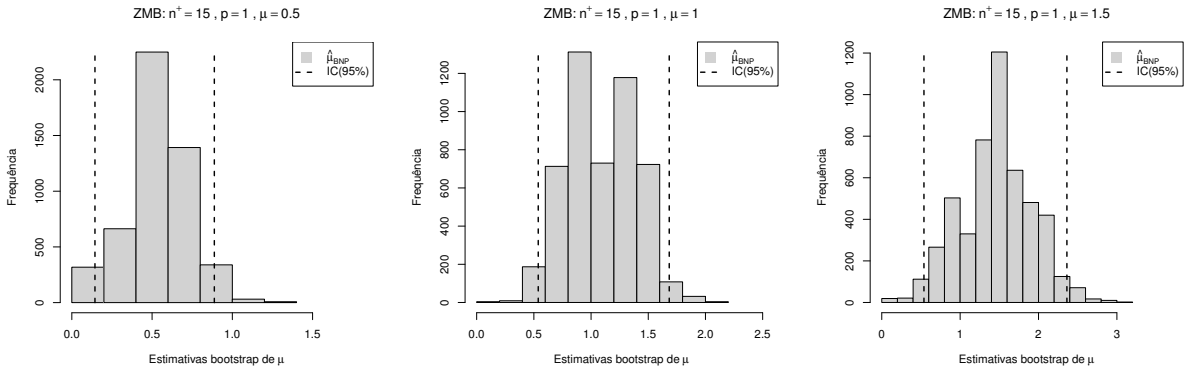
As Figuras 16 - 20 apresentam as análises gráficas para as amostras geradas (sem o valor zero) a partir da distribuição Binomial, com os diferentes tamanhos amostrais e valores do parâmetro μ, cujas estimativas foram apresentadas na Tabela 15. As distribuições de frequências para os conjuntos, ilustradas pelos gráficos (a) das Figuras, permitem verificar que há aumento na amplitude dos dados conforme o aumento do parâmetro μ, ressaltada pela escala degradê em tons de cinza. Também é possível notar que baixa (alta) amplitude ocasiona maiores (menores) frequências observadas em todos os casos. As distribuições empíricas para o estimador do parâmetro μ, obtidas via BNP, com os percentis de $2,5\%$ e $97,5\%$ em destaque, podem ser observadas nos gráficos (b) destas Figuras. Além da constatação de uma leve assimetria para todos os conjuntos amostrais e da verificação de maior concentração das frequências em torno dos valores verdadeiros do parâmetro, verifica-se a pouca variabilidade em torno deste. As frequências observadas (f_i) e esperadas (f_e) obtidas a partir da distribuição tradicional ajustada aos valores positivos dos conjuntos são, por sua vez, comparadas nos gráficos em (c). Observa-se que há pouca diferença entre as alturas das barras f_i e f_e, evidenciando boa aderência da distribuição aos conjuntos de dados estudados.

Salienta-se que para todos os tamanhos de amostras e valores de parâmetros considerados neste estudo, os comportamentos acima descritos são observados. A melhora nas características para maiores conjuntos amostrais também verifica-se, embora os resultados sejam aceitáveis ainda para aqueles de menor tamanho, validando as qualidades do método proposto.

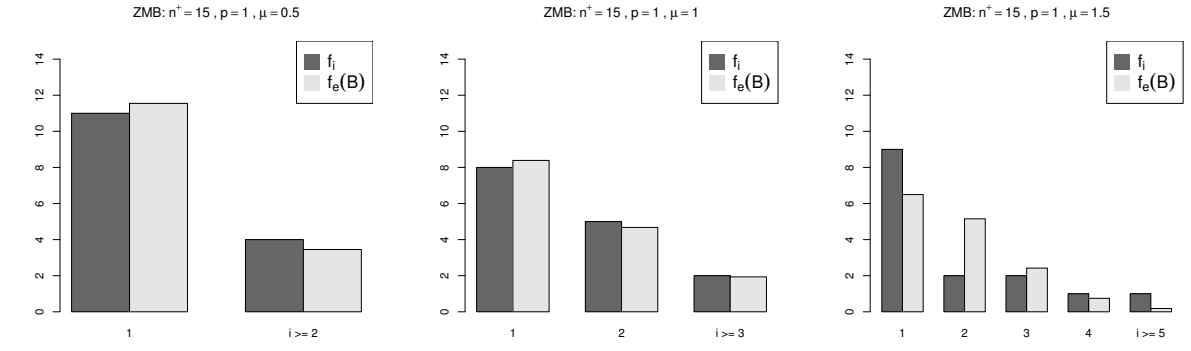
Figura 16 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



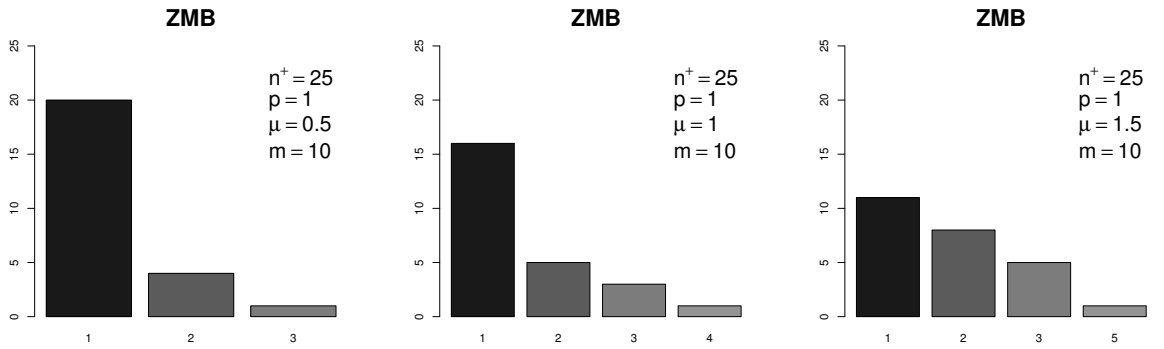
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



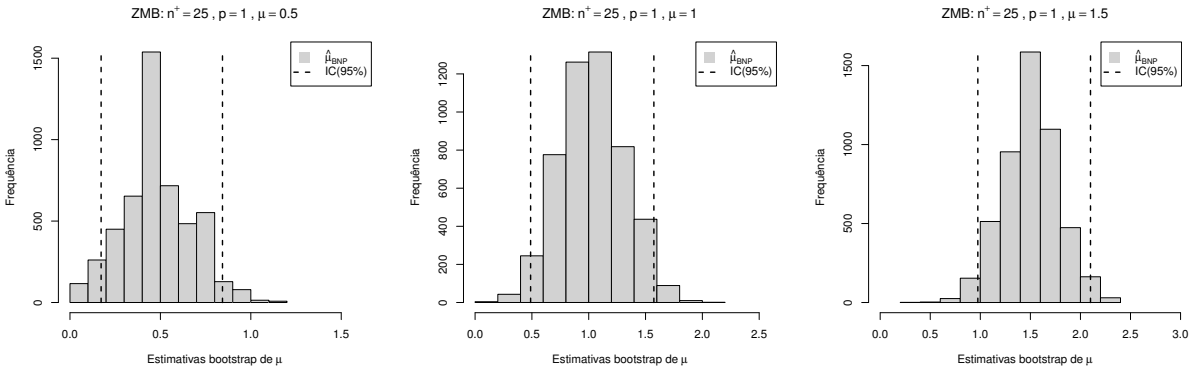
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

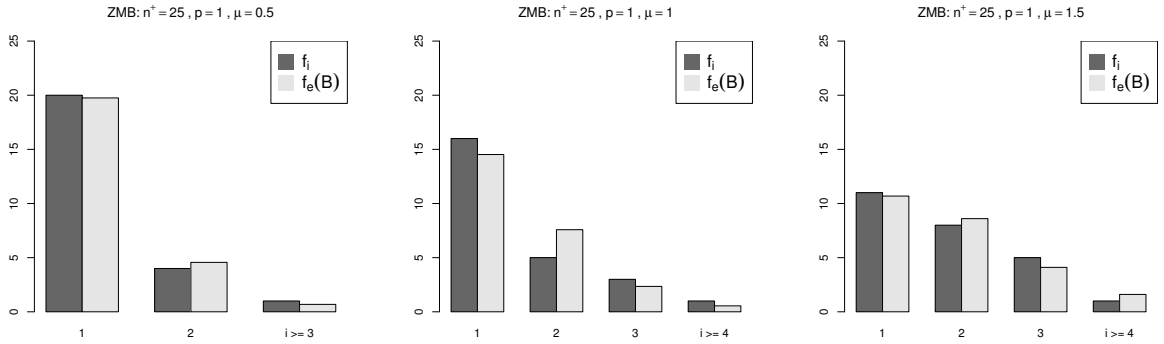
Figura 17 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



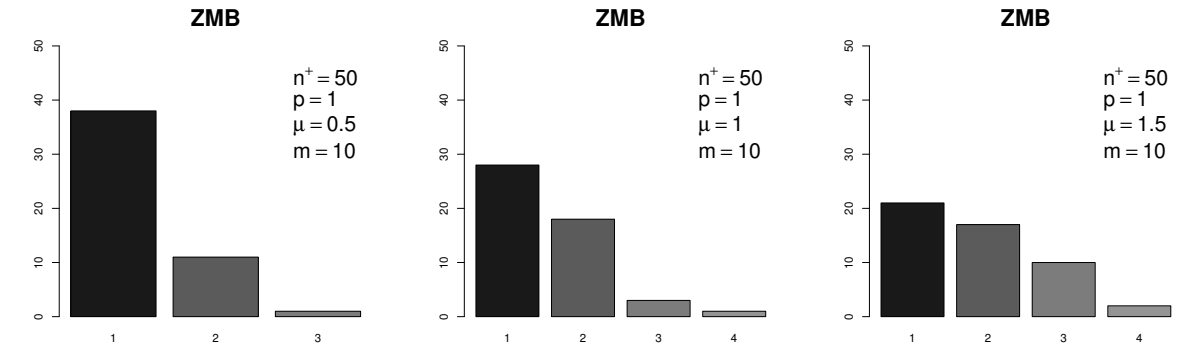
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



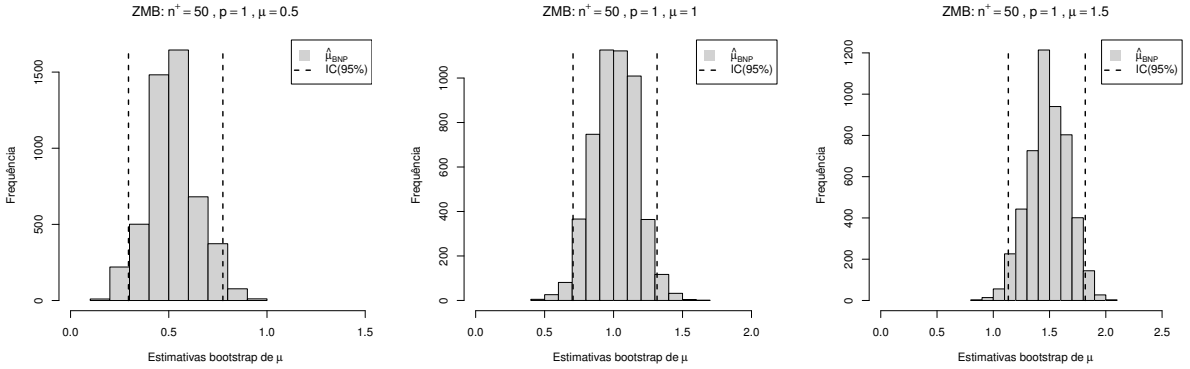
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

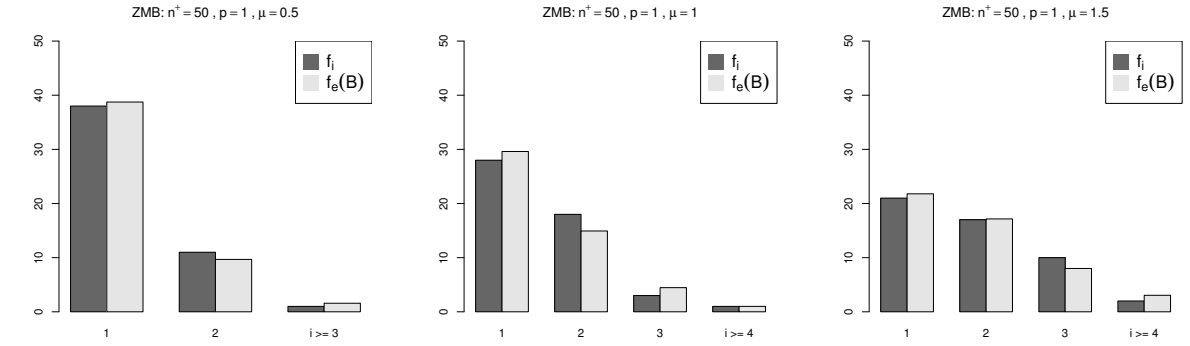
Figura 18 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



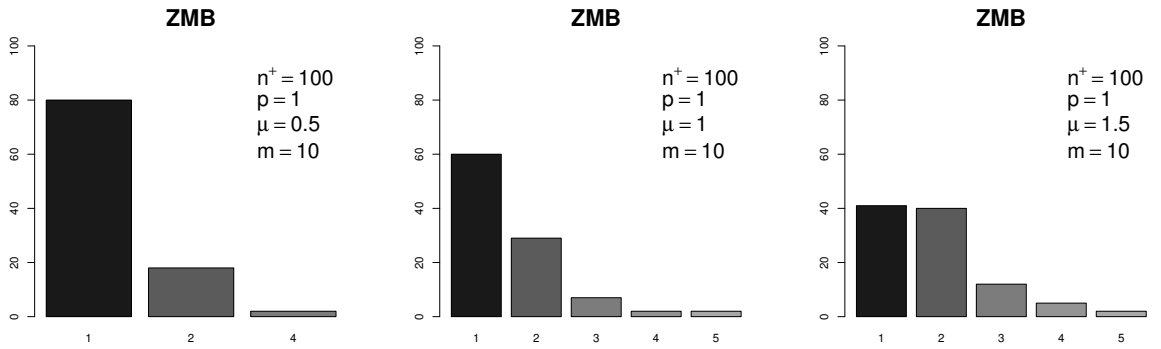
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



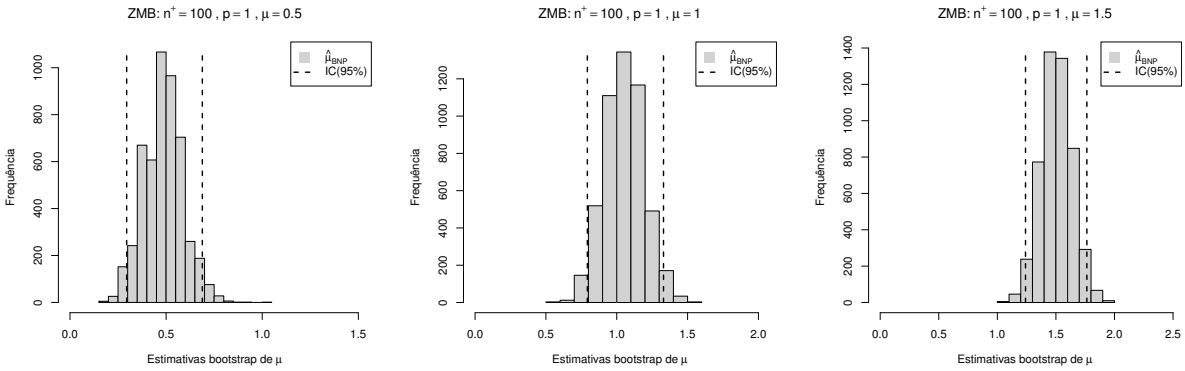
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

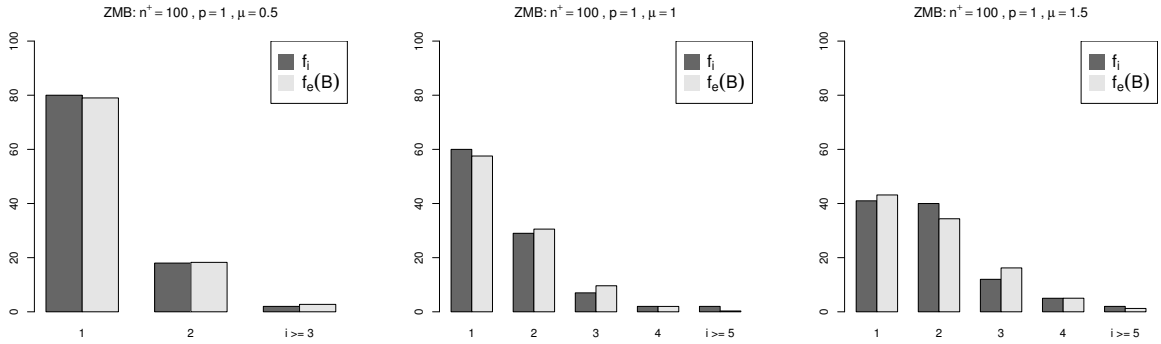
Figura 19 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



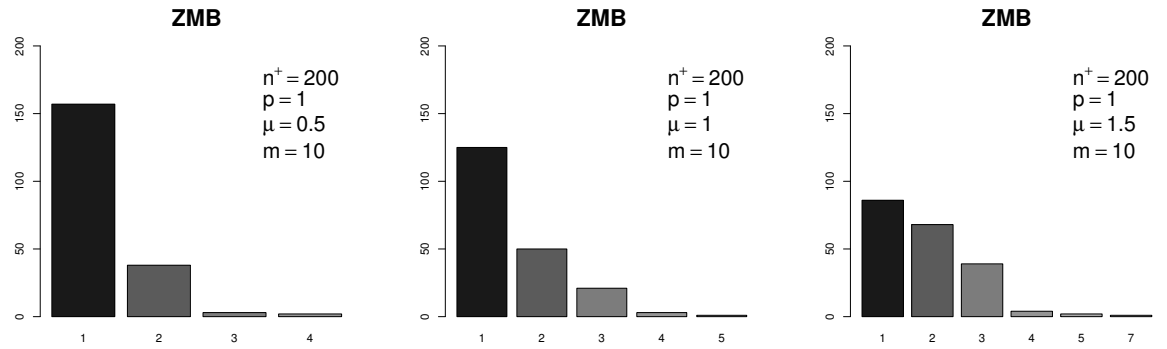
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



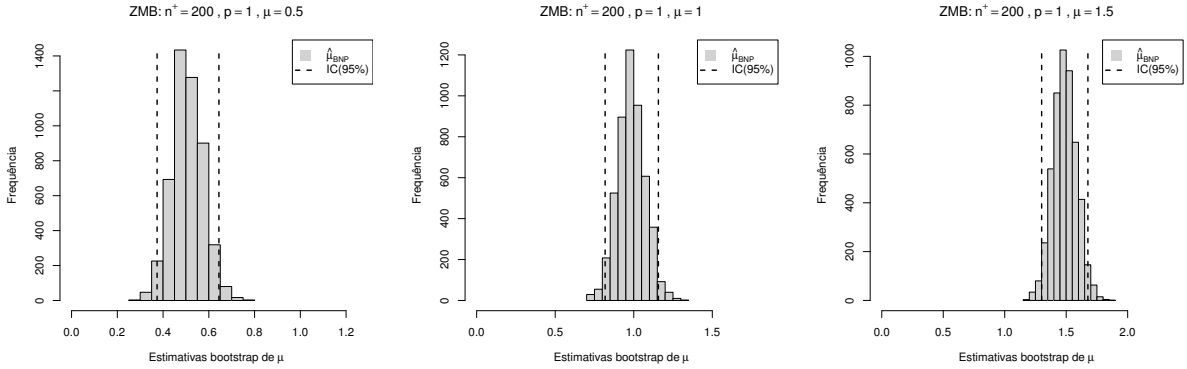
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

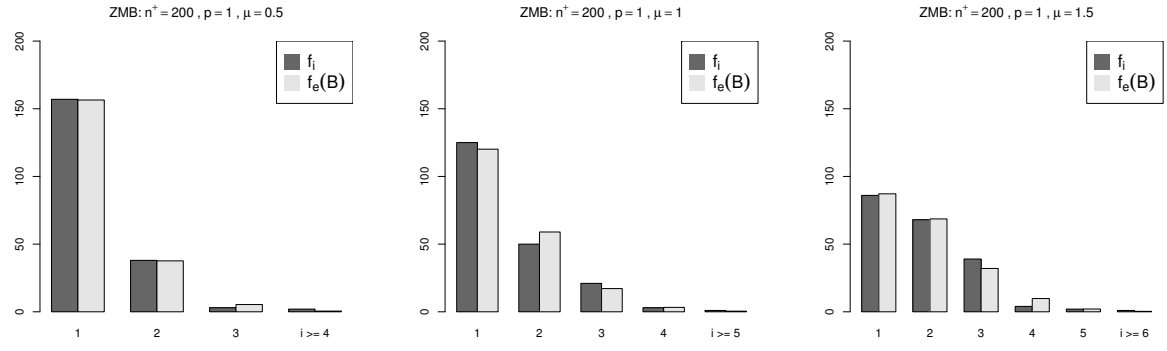
Figura 20 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Binomial (ZMB, com $p = 1$) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

Cenário 2: Dados Gerados de uma Distribuição ZDPS

Neste estudo, uma amostra de tamanho n^+ foi gerada da respectiva distribuição ZDPS (isto é, com parâmetro $1 < p < 1/(1-\pi_{PS}(0;\mu,\phi))$) considerando-se os diferentes valores de μ e p (ver Tabela 9).

Na Tabela 16 são apresentados os resultados obtidos a partir da análise de dados provenientes de uma distribuição ZDP. Observa-se que as estimativas são próximas aos valores verdadeiros (μ e $\pi_P(0;\mu)$), bem como estes valores verdadeiros pertencem aos intervalos *bootstrap* com 95% de confiança. As estimativas corrigidas, calculadas utilizando-se as estimativas obtidas pelo algoritmo EM para as reamostras *bootstrap*, foram próximas àquelas resultantes da aplicação direta do método. São apresentados também nesta Tabela o valor de n_0 gerado (e desconsiderado no procedimento de estimação, que utiliza apenas os valores positivos), e a estimativa de n_0, a qual possibilitou a boa estimativa de μ e, consequentemente, da probabilidade de zero. Verifica-se que os valores gerados e estimados de n_0 foram bastante diferentes, evidenciando que o conjunto de dados é zero-deflacionado uma vez que o valor estimado é superior ao gerado pela distribuição ZDP (com cerca de 95% observações zero a menos do que o esperado por uma distribuição tradicional). Adicionalmente, constata-se que os valores estimados apresentados melhoram conforme o aumento do tamanho das amostras; assim, verifica-se que o método produz boas estimativas.

As Figuras 21 - 25 apresentam as análises gráficas para as amostras geradas (sem o valor zero) a partir da distribuição ZDP, com os diferentes tamanhos amostrais e valores do parâmetro μ, cujas estimativas foram apresentadas na Tabela 16. Nos gráficos apresentados em (a) estão as distribuições de frequências para os conjuntos, nas quais verificam-se visualmente, dentre outros aspectos, o aumento na amplitude dos dados à medida em que o valor de μ aumenta. A escala em degradê em tons de cinza ressalta essa amplitude, atribuindo uma cor diferente para cada observação particular da amostra. É possível notar também para todos os casos que baixa (alta) amplitude ocasiona maiores (menores) frequências observadas. Ainda nestas Figuras, os gráficos em (b), observa-se as distribuições empíricas para o estimador do parâmetro μ, obtidas via BNP, com os percentis de $2,5\%$ e $97,5\%$ em destaque. Essas apresentam-se aproximadamente simétricas para todos os conjuntos amostrais, com maior concentração em torno dos valores verdadeiros do parâmetro e verifica-se também pouca variabilidade em torno destes. Por sua vez, os gráficos em (c) apresentam as comparações entre as frequências observadas (f_i) e as esperadas (f_e), obtidas a partir da distribuição tradicional ajustada aos valores positivos dos conjuntos. Devido à baixa discrepância, tem-se evidência para a adequabilidade da distribuição suposta aos conjuntos de dados estudados.

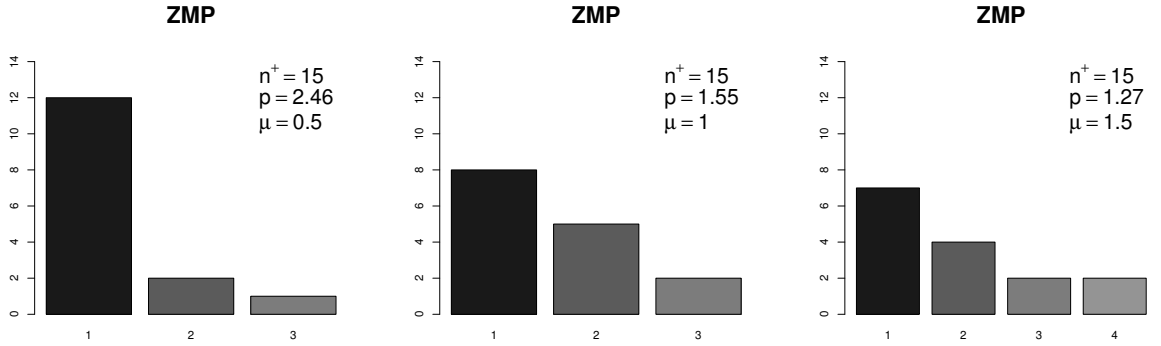
Destaca-se que os comportamentos acima descritos são observados para todos os conjuntos amostrais e valores de parâmetros apresentados, verificando-se a melhora nas características com o aumento dos tamanhos amostrais, embora para pequenas amostras os resultados também sejam satisfatórios. Assim, evidencia-se o bom desempenho do procedimento.

Tabela 16 – Estimativas de n_0, μ e de $\pi_P(0;\mu)$ obtidas a partir de dados provenientes da distribuição Poisson Zero-Deflacionada.

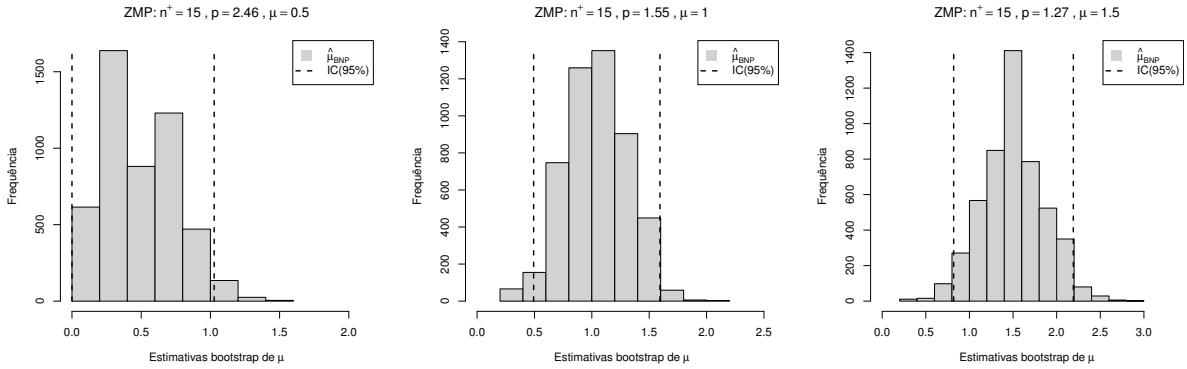
n^+	μ	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_P(0;\mu)$	$\pi_P(0;\widehat{\mu})$	$\pi_P(0;\widehat{\mu}_C)$	n_0	$\widehat{n}_0$
15	0,5	0,493 (0,000; 1,027)	0,506	0,607	0,611 (0,358; 1,000)	0,584	2	23,538
	1,0	1,027 (0,493; 1,594)	1,038	0,368	0,358 (0,203; 0,611)	0,340	0	8,365
	1,5	1,503 (0,819; 2,191)	1,511	0,223	0,222 (0,112; 0,441)	0,205	0	4,289
25	0,5	0,516 (0,156; 0,905)	0,522	0,607	0,597 (0,404; 0,856)	0,581	2	37,030
	1,0	1,087 (0,650; 1,540)	1,084	0,368	0,337 (0,214; 0,522)	0,329	1	12,723
	1,5	1,485 (0,967; 2,009)	1,495	0,223	0,226 (0,134; 0,380)	0,216	0	7,319
50	0,5	0,516 (0,305; 0,715)	0,521	0,607	0,597 (0,489; 0,737)	0,591	3	74,075
	1,0	0,997 (0,617; 1,375)	1,007	0,368	0,369 (0,253; 0,540)	0,358	1	29,236
	1,5	1,513 (1,146; 1,882)	1,517	0,223	0,220 (0,152; 0,318)	0,216	1	14,132
100	0,5	0,481 (0,323; 0,650)	0,481	0,607	0,618 (0,522; 0,724)	0,616	2	161,694
	1,0	1,087 (0,843; 1,347)	1,090	0,368	0,337 (0,260; 0,430)	0,334	1	50,889
	1,5	1,485 (1,218; 1,779)	1,486	0,223	0,226 (0,169; 0,296)	0,224	0	29,275
200	0,5	0,516 (0,394; 0,650)	0,517	0,607	0,597 (0,522; 0,674)	0,595	8	269,301
	1,0	1,035 (0,858; 1,204)	1,036	0,368	0,355 (0,300; 0,424)	0,354	4	110,239
	1,5	1,499 (1,297; 1,693)	1,502	0,223	0,223 (0,184; 0,273)	0,222	4	57,528

Fonte: Elaborada pelo autor.

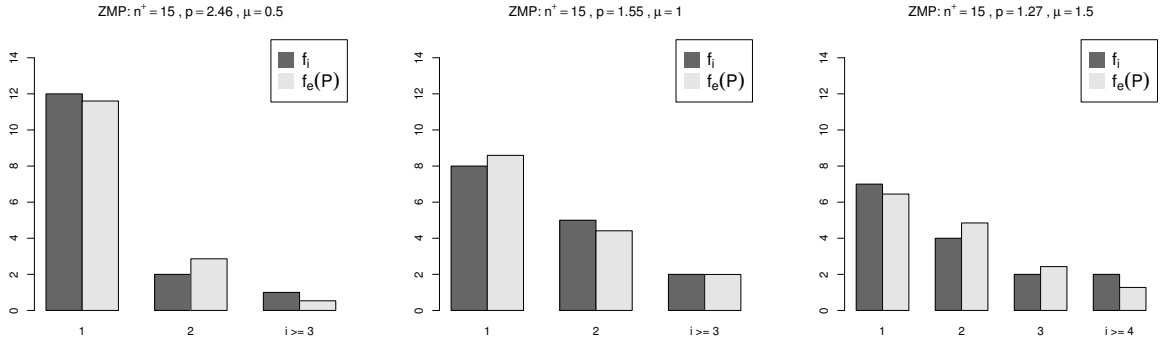
Figura 21 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



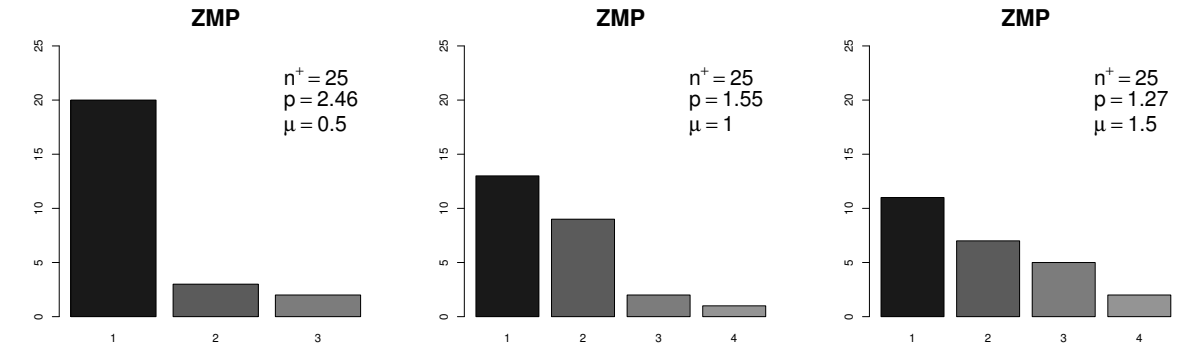
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



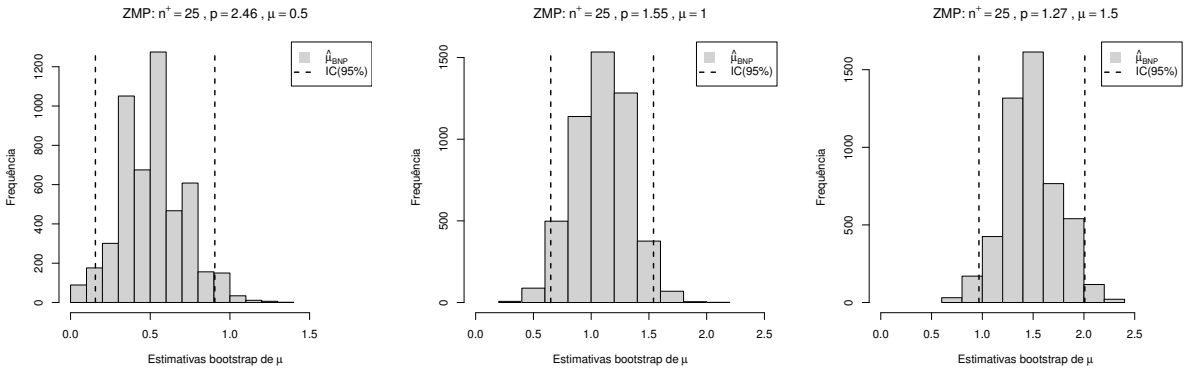
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

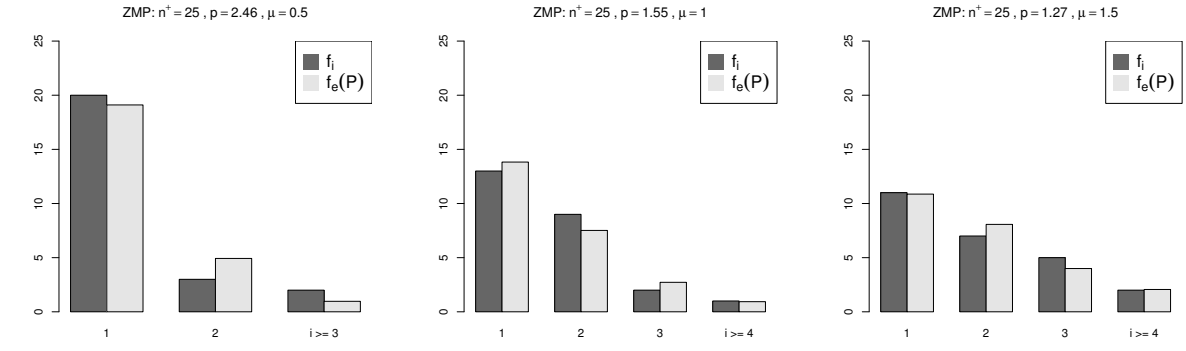
Figura 22 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



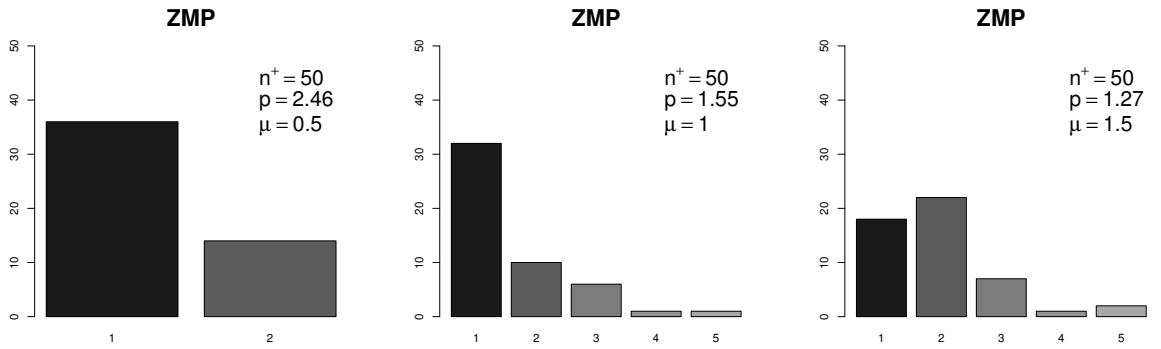
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



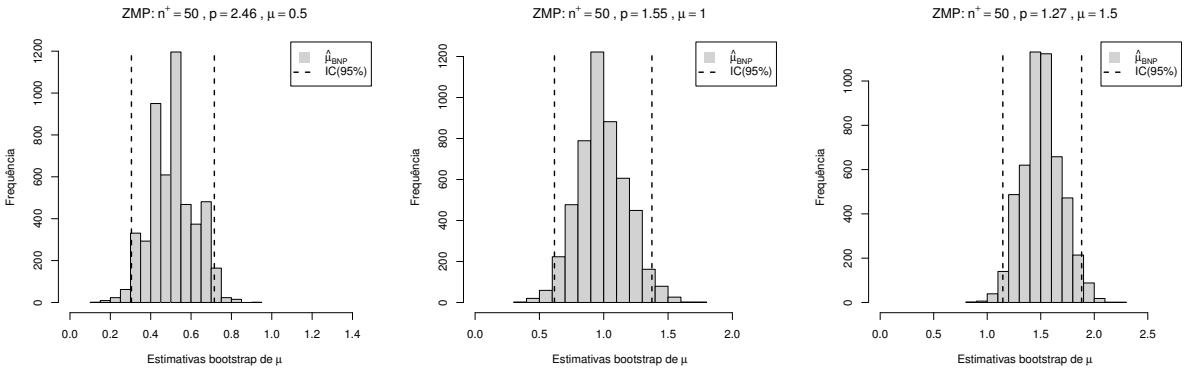
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

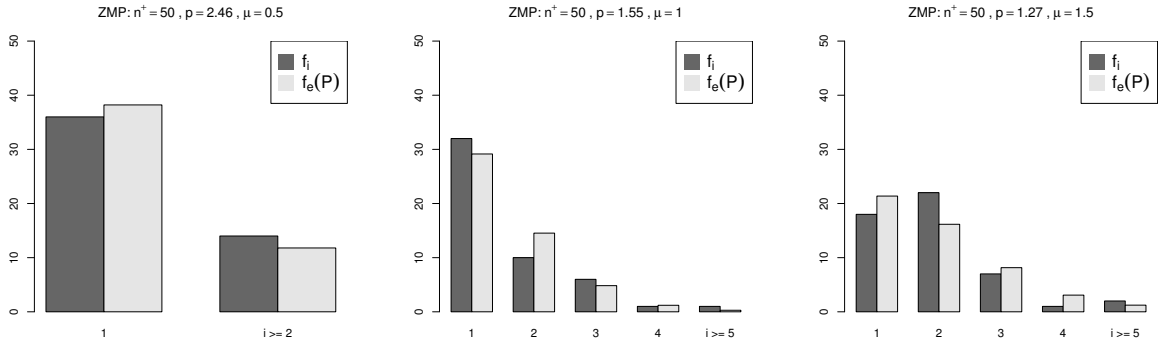
Figura 23 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



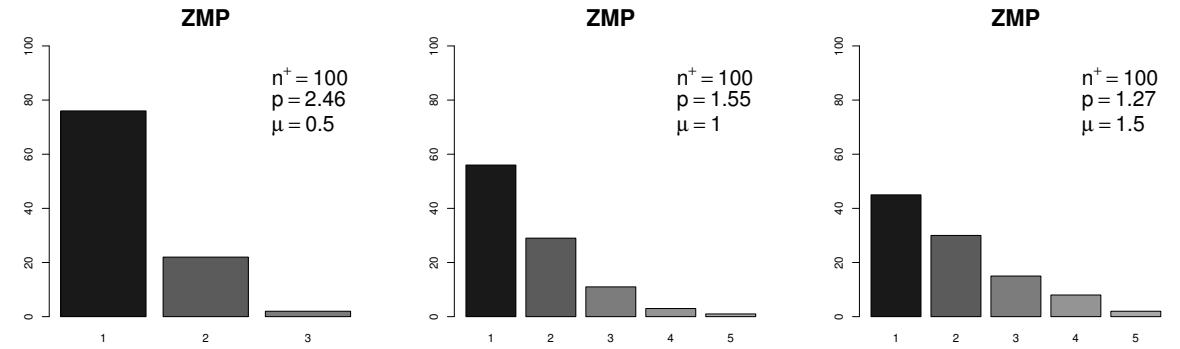
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



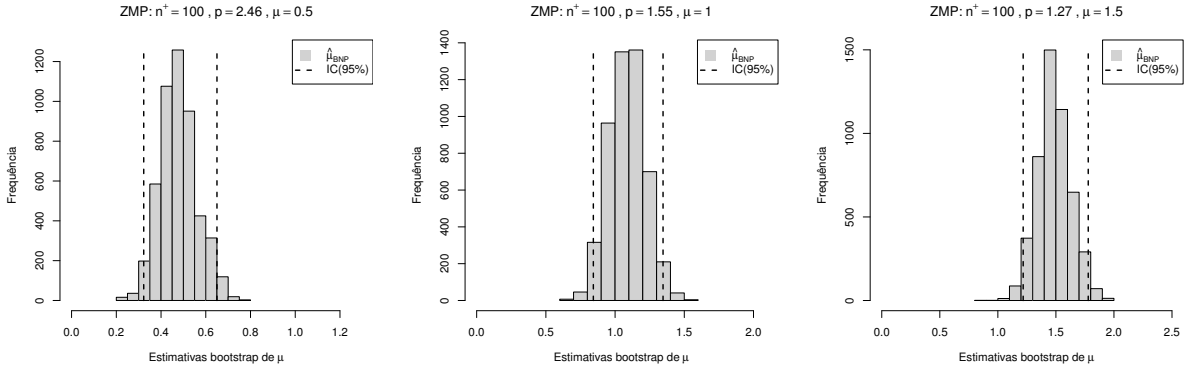
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

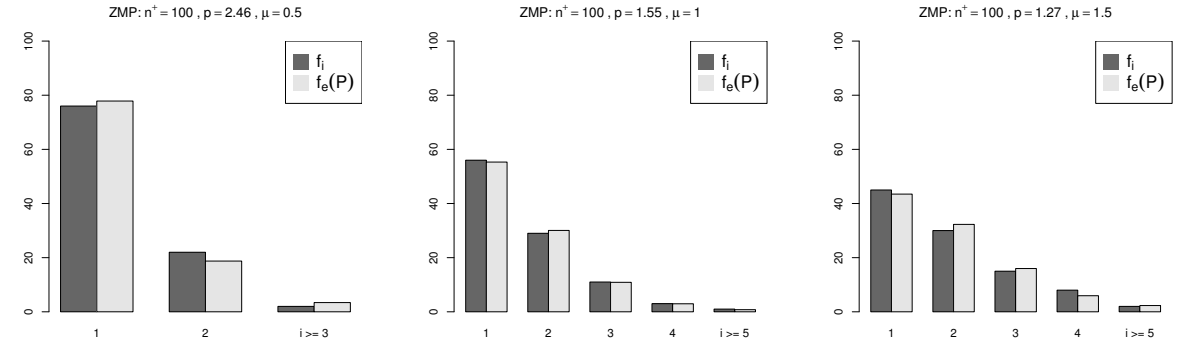
Figura 24 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



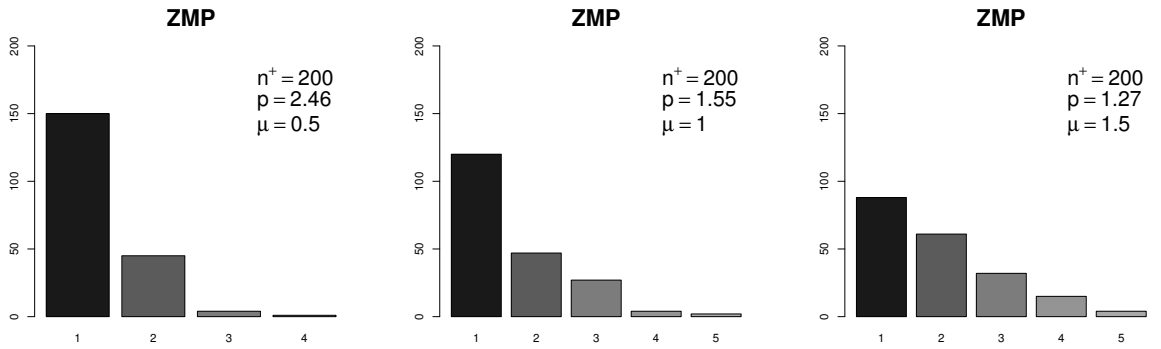
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



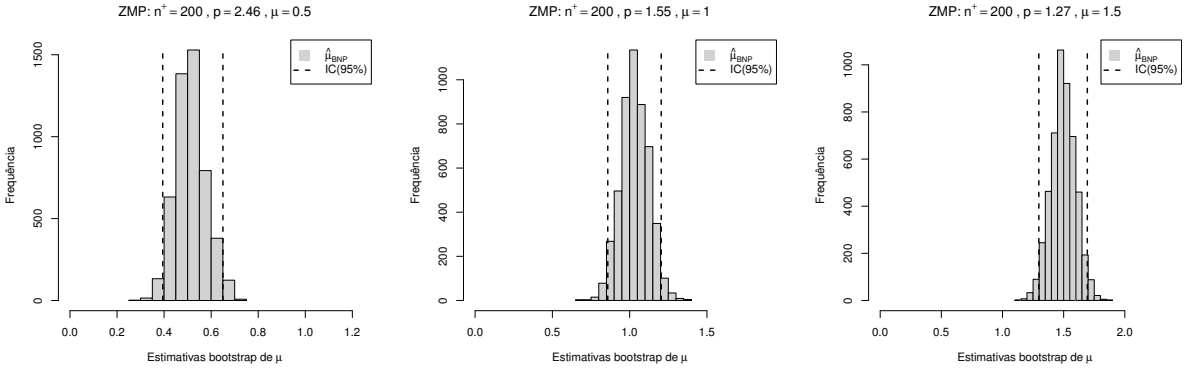
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

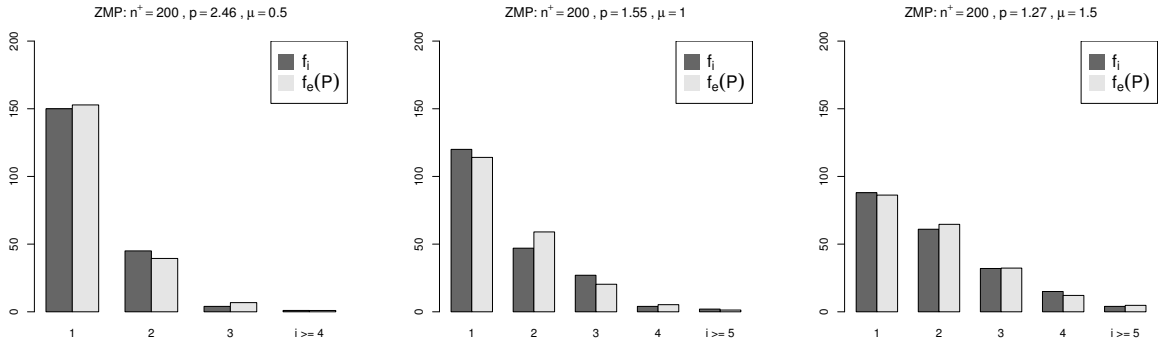
Figura 25 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Poisson (ZDP) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

Na Tabela 17, que apresenta os resultados obtidos com a análise de dados provenientes de uma distribuição ZDG, verifica-se que as estimativas são próximas dos valores verdadeiros (μ e $\pi_G(0;\mu)$) considerados para geração dos dados, assim como estes valores verdadeiros pertencem aos respectivos intervalos de confiança *bootstrap*. Similarmente, as estimativas corrigidas (que utilizaram os valores calculados via aplicação do procedimento às réplicas *bootstrap*) foram próximas àquelas obtidas via aplicação direta aos dados com o algoritmo EM. Além disso, tem-se que os valores estimados de n_0 foram bem acima das verdadeiras quantidades geradas desta observação para as amostras, fato extremamente aceitável uma vez que, de fato, os dados foram gerados de uma distribuição ZDG. Por fim, nota-se também erros menores nas estimativas de μ e $\pi_G(0;\mu)$ quando os tamanhos amostrais aumentam.

As Figuras 26 - 30 apresentam as análises gráficas para as amostras geradas (sem o valor zero) a partir da distribuição ZDG, com os diferentes tamanhos amostrais e valores do parâmetro μ, cujas estimativas encontram-se na Tabela 17. Nas Figuras, os gráficos em (a) estão as distribuições de frequências para os conjuntos e verifica-se o aumento da amplitude dos dados à medida em que o valor da média aumenta; novamente, degradê em tons de cinza ressalta a amplitude e observa-se para todos os casos que baixa amplitude ocasiona em maiores frequências observadas bem como alta amplitude ocasiona em menores frequências. Ainda considerando estas Figuras, nos gráficos em (b) observa-se as distribuições empíricas para o estimador do parâmetro μ, obtidas via BNP, com os percentis de $2,5\%$ e $97,5\%$ em destaque. Verifica-se que há pouca assimetria para todos os conjuntos amostrais e que há maior concentração das frequências em torno dos valores verdadeiros do parâmetro; constata-se também pouca variabilidade em torno deste. Por sua vez, os gráficos em (c) apresentam as comparações entre as frequências observadas (f_i) e esperadas (f_e) pela distribuição tradicional ajustada aos valores positivos dos conjuntos de dados. Observa-se que as barras f_i e f_e têm alturas próximas, dando indícios de que a suposição da distribuição é adequada para explicar o comportamento dos conjuntos de dados estudados.

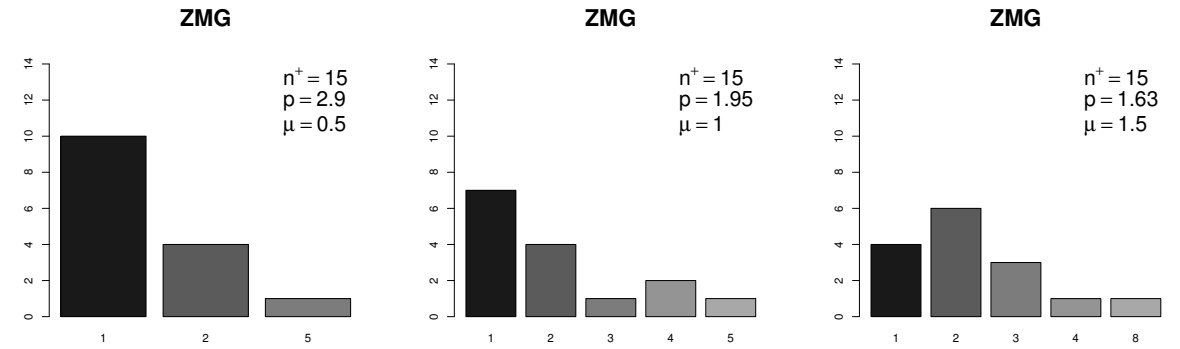
Mais uma vez, ressalta-se que os comportamentos acima descritos são observados para todos os tamanhos de amostras e valores de parâmetros considerados neste estudo, verificando-se a melhora nas características para maiores conjuntos amostrais, embora para aqueles de menor tamanho os resultados também sejam razoáveis, corroborando a qualidade da metodologia considerada.

Tabela 17 – Estimativas de n_0, μ e de $\pi_G(0;\mu)$ obtidas a partir de dados provenientes da distribuição Geométrica Zero-Deflacionada.

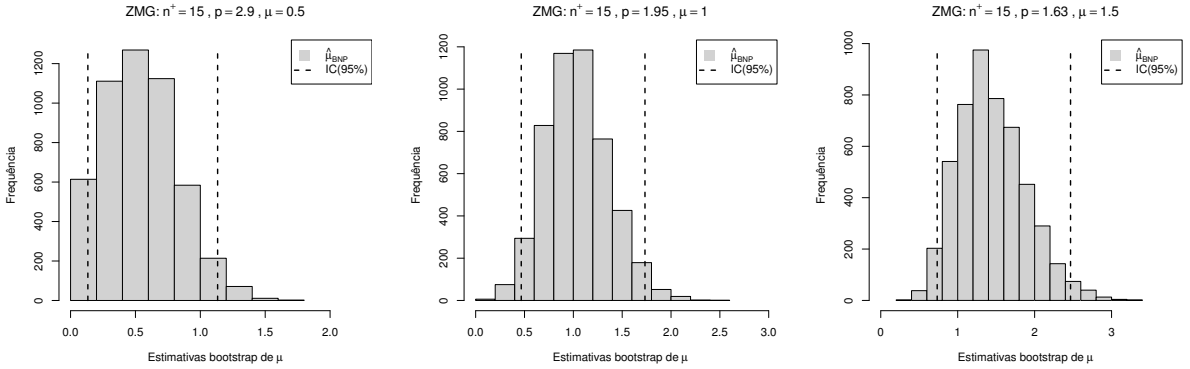
n^+	μ	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_G(0;\mu)$	$\pi_G(0;\widehat{\mu})$	$\pi_G(0;\widehat{\mu}_C)$	n_0	$\widehat{n}_0$
15	0,5	0,533 (0,133; 1,133)	0,535	0,667	0,652 (0,469; 0,882)	0,633	0	28,124
	1,0	1,067 (0,467; 1,733)	1,071	0,500	0,484 (0,366; 0,682)	0,470	0	14,063
	1,5	1,467 (0,733; 2,467)	1,465	0,400	0,405 (0,288; 0,577)	0,393	1	10,227
25	0,5	0,520 (0,200; 0,880)	0,520	0,667	0,658 (0,532; 0,833)	0,650	0	48,079
	1,0	1,040 (0,600; 1,560)	1,04	0,500	0,490 (0,391; 0,625)	0,483	2	24,039
	1,5	1,520 (0,840; 2,320)	1,520	0,400	0,397 (0,301; 0,543)	0,387	0	16,447
50	0,5	0,520 (0,280; 0,780)	0,521	0,667	0,658 (0,562; 0,781)	0,653	3	96,149
	1,0	1,100 (0,680; 1,600)	1,097	0,500	0,476 (0,385; 0,595)	0,471	3	45,454
	1,5	1,520 (1,000; 2,060)	1,526	0,400	0,397 (0,327; 0,500)	0,391	0	32,894
100	0,5	0,490 (0,320; 0,680)	0,491	0,667	0,671 (0,595; 0,758)	0,668	2	204,070
	1,0	1,070 (0,810; 1,340)	1,071	0,500	0,483 (0,427; 0,552)	0,481	5	93,456
	1,5	1,510 (1,120; 1,950)	1,511	0,400	0,398 (0,339; 0,472)	0,395	5	66,225
200	0,5	0,505 (0,395; 0,630)	0,506	0,667	0,664 (0,613; 0,717)	0,663	5	396,024
	1,0	1,050 (0,855; 1,260)	1,052	0,500	0,488 (0,442; 0,539)	0,486	8	190,474
	1,5	1,505 (1,245; 1,785)	1,507	0,400	0,399 (0,359; 0,445)	0,398	5	132,889

Fonte: Elaborada pelo autor.

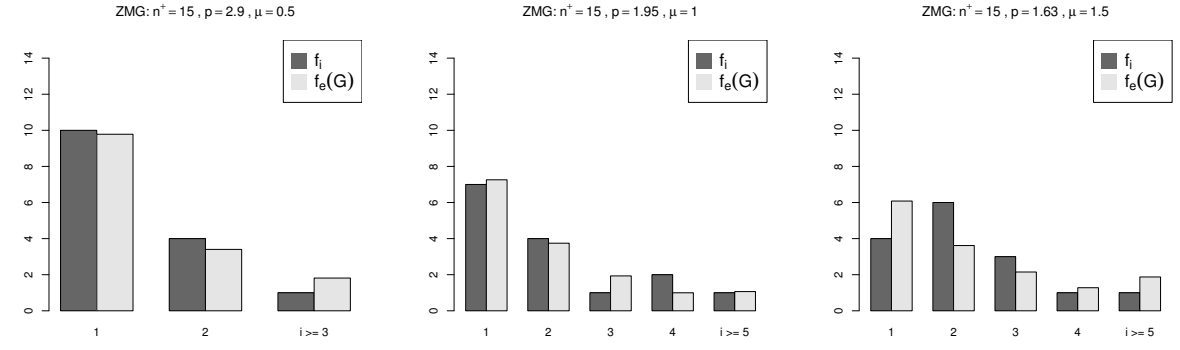
Figura 26 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



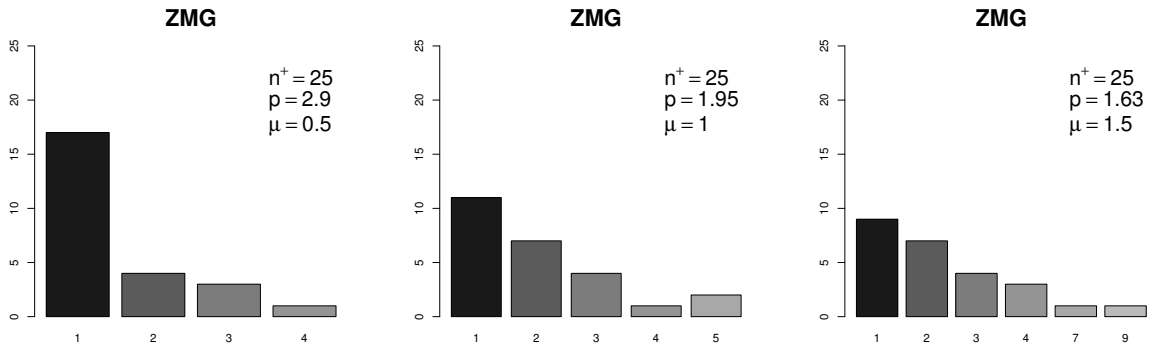
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



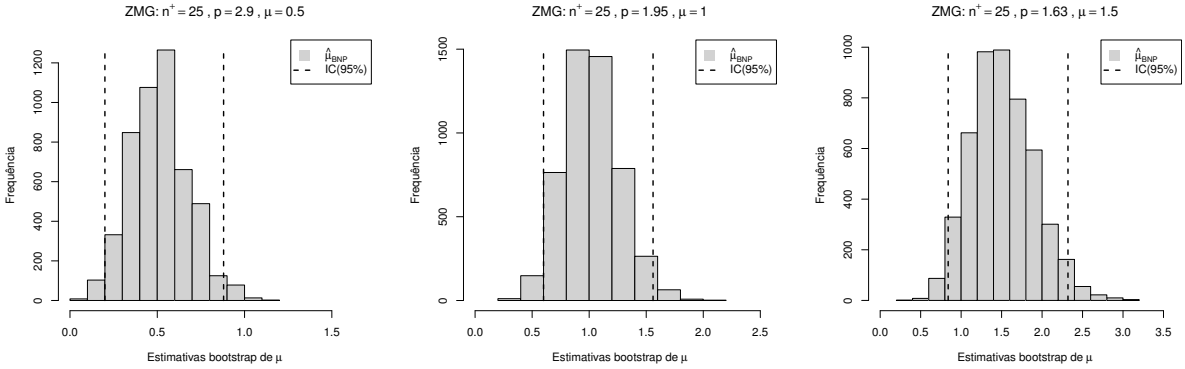
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

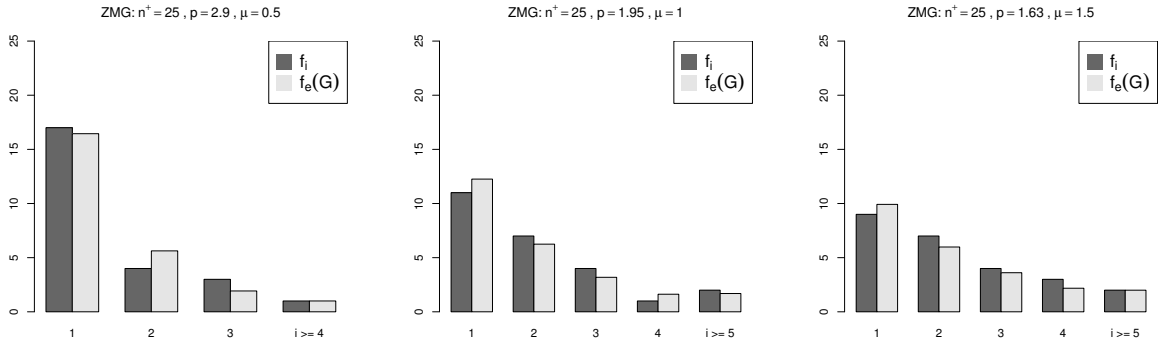
Figura 27 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



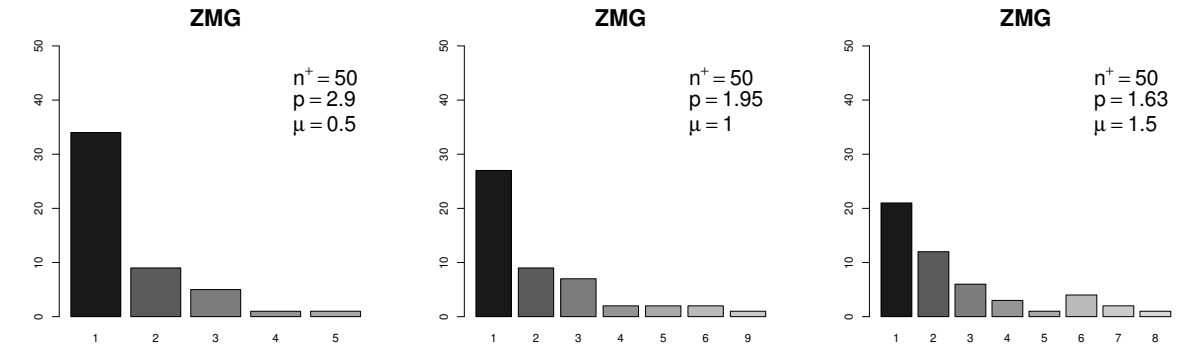
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



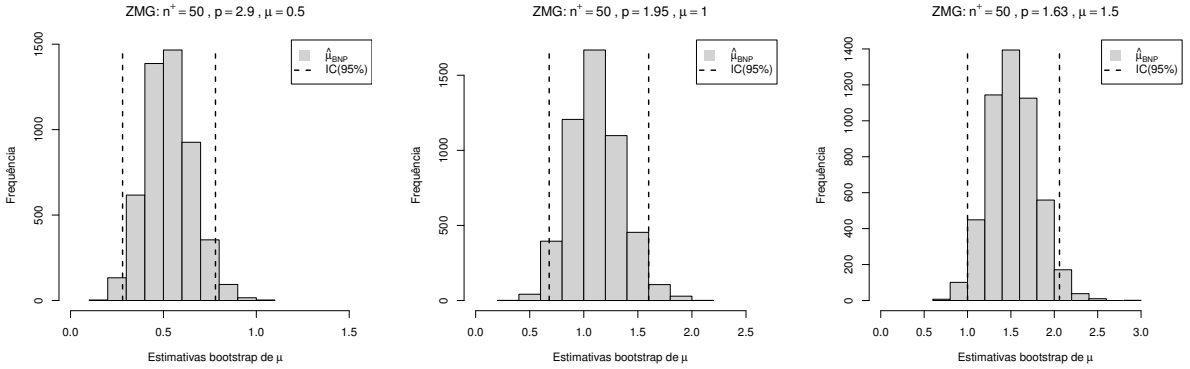
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

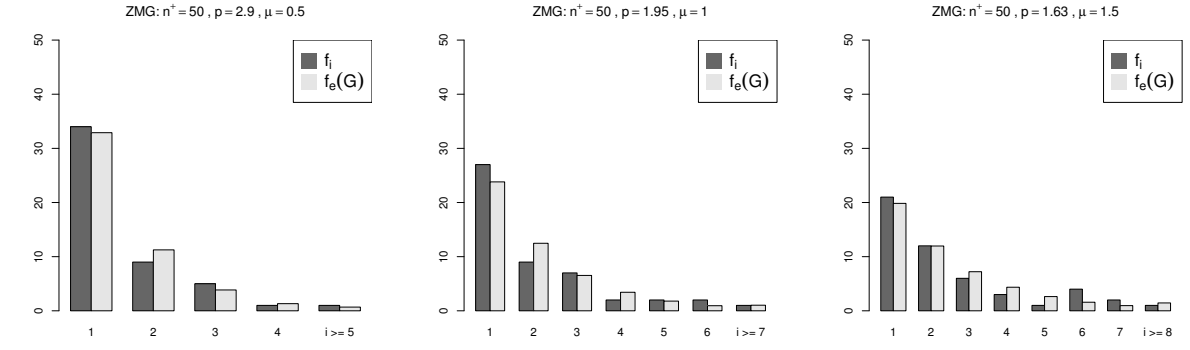
Figura 28 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



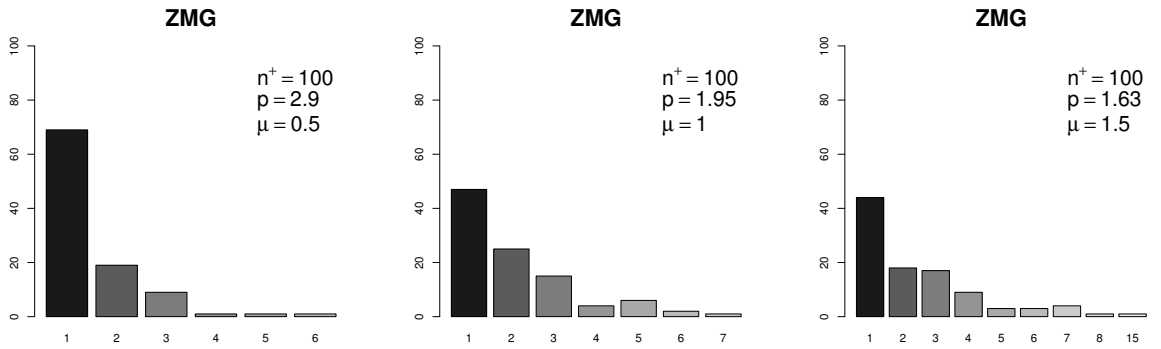
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



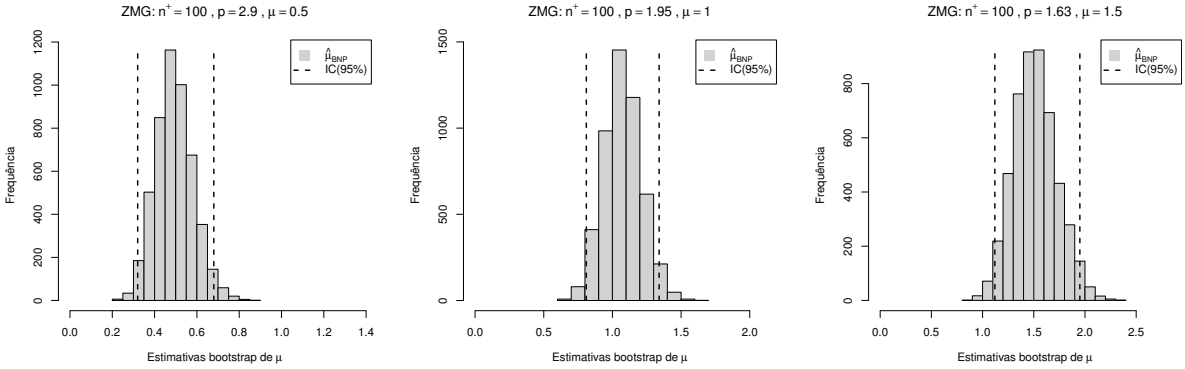
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

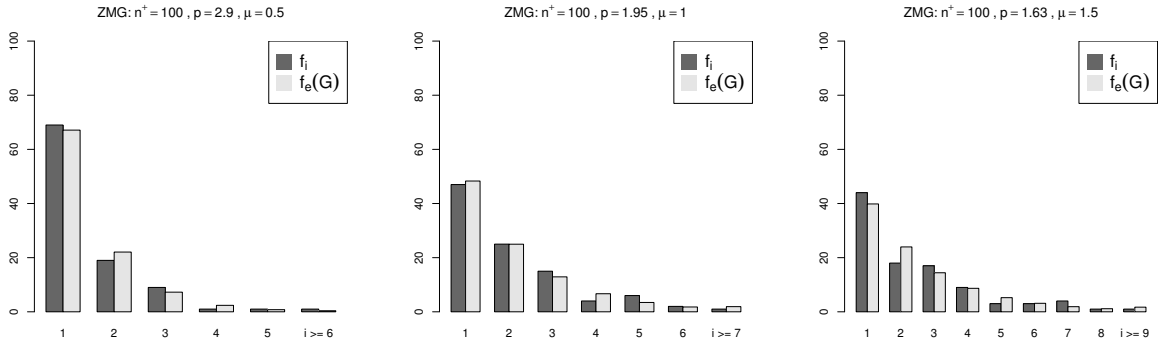
Figura 29 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



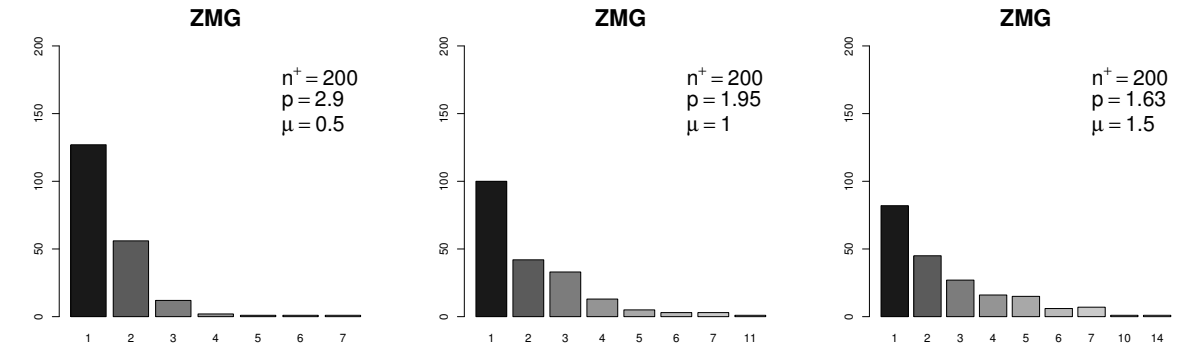
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



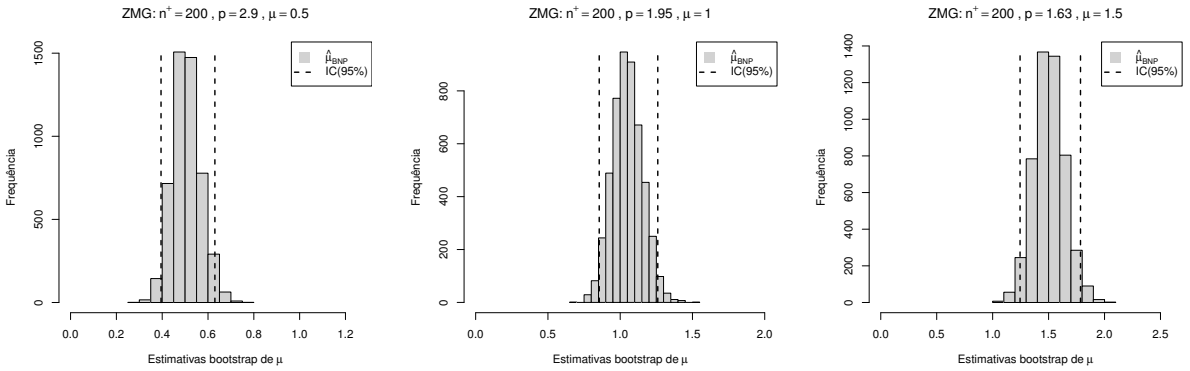
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

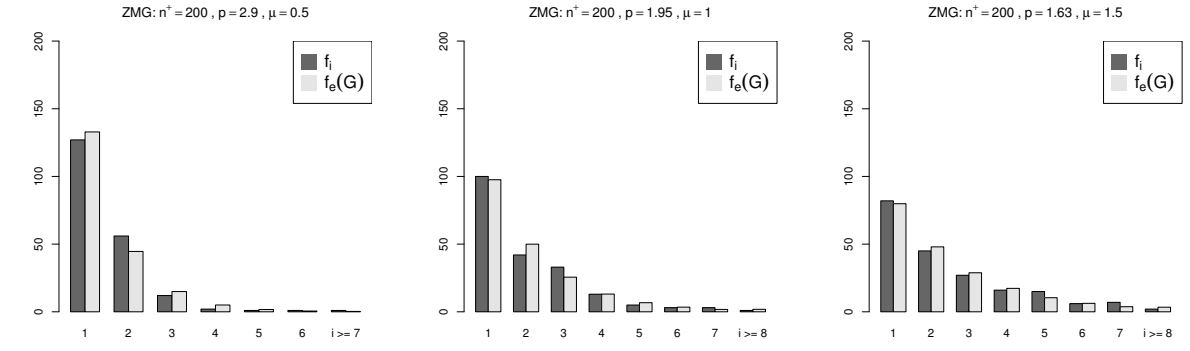
Figura 30 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Geométrica (ZDG) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

A Tabela 18 apresenta as estimativas obtidas de n_0, μ e $\pi_B(0;\mu)$ para a distribuição ZDB, com $m = 10$. Pode-se observar que os valores resultantes da aplicação do método proposto são próximos dos valores verdadeiros (μ e $\pi_B(0;\mu)$), e além disso, que estes valores verdadeiros pertencem aos respectivos intervalos *bootstrap* com 95% de confiança. As estimativas corrigidas (obtidas a partir das réplicas *bootstrap*) tiveram resultados próximos aos obtidos pela aplicação direta do algoritmo EM às amostras. Os valores de n_0 gerados para as amostras e os estimados pelo procedimento ao considerar-se apenas as observações positivas foram bastante diferentes, evidenciando que a estimativa é cerca de 95% maior que o número gerado pela amostra deflacionada. Para maiores tamanhos de amostra, há a melhora nas estimativas, corroborando que o método de estimação considerado produz bons resultados.

As Figuras 31 - 35 apresentam as análises gráficas para as amostras geradas (sem o valor zero) a partir da distribuição ZDB, com os diferentes tamanhos amostrais e valores do parâmetro μ, cujas estimativas encontram-se na Tabela 18. As distribuições de frequências para os conjuntos, apresentadas em (a) destas Figuras, permitem verificar que há aumento na amplitude dos dados conforme o aumento do parâmetro μ, ressaltada pela escala degradê em tons de cinza. Também é possível notar que baixa (alta) amplitude ocasiona maiores (menores) frequências observadas em todos os casos. As distribuições empíricas para o estimador do parâmetro μ, obtidas via BNP, com os percentis de $2,5\%$ e $97,5\%$ em destaque, podem ser observadas ainda nestas Figuras, gráficos (b). Além da constatação de pouca assimetria para todos os conjuntos amostrais e da verificação de maior concentração das frequências em torno dos valores verdadeiros do parâmetro, verifica-se a pouca variabilidade em torno deste. As frequências observadas (f_i) e esperadas (f_e) pela distribuição tradicional ajustada aos valores positivos dos conjuntos são, por sua vez, comparadas nos gráficos em (c) destas Figuras. Observa-se que há pouca diferença entre as alturas das barras f_i e f_e, evidenciando boa aderência da distribuição aos conjuntos de dados estudados.

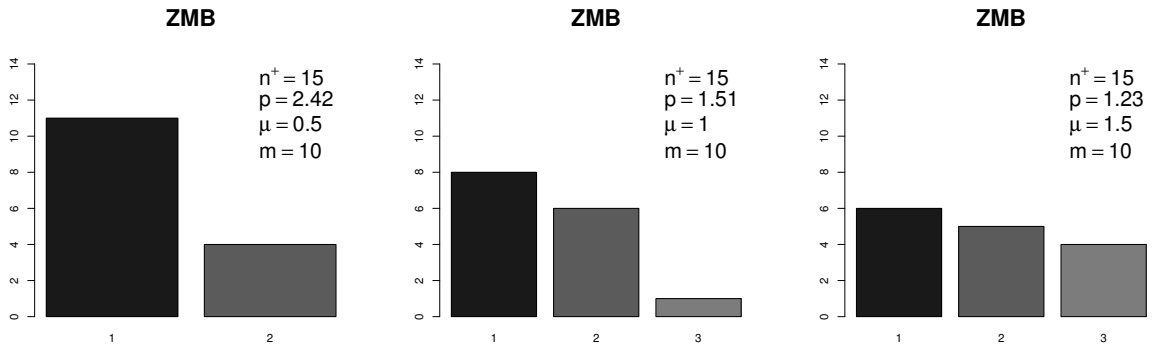
Salienta-se que para todos os tamanhos de amostras e valores de parâmetros considerados neste estudo, os comportamentos acima descritos são observados. A melhora nas características para maiores conjuntos amostrais também verifica-se, embora os resultados sejam aceitáveis ainda para aqueles de menor tamanho, validando a qualidade do método considerado neste trabalho.

Tabela 18 – Estimativas de n_0, μ e de $\pi_G(0;\mu)$ obtidas a partir de dados provenientes da distribuição Binomial Zero-Deflacionada.

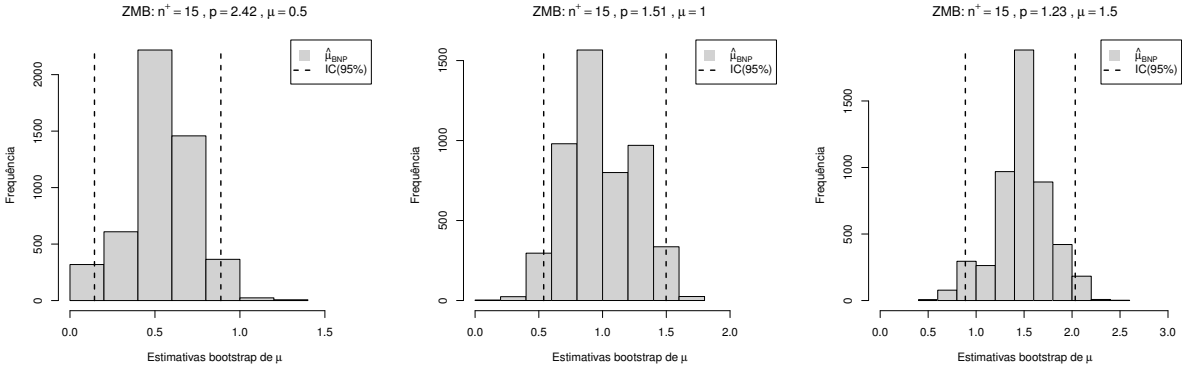
n^+	μ	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_B(0;\mu)$	$\pi_B(0;\widehat{\mu})$	$\pi_B(0;\widehat{\mu}_C)$	n_0	$\widehat{n}_0$
15	0,5	0,538 (0,144; 0,888)	0,544	0,599	0,575 (0,395; 0,865)	0,559	1	20,3
	1,0	0,997 (0,538; 1,499)	0,999	0,349	0,350 (0,197; 0,575)	0,336	0	8,075
	1,5	1,499 (0,888; 2,033)	1,515	0,197	0,197 (0,103; 0,395)	0,182	0	3,685
25	0,5	0,489 (0,087; 0,843)	0,496	0,599	0,606 (0,414; 0,916)	0,590	2	38,446
	1,0	1,039 (0,563; 1,461)	1,043	0,349	0,334 (0,206; 0,560)	0,323	0	12,521
	1,5	1,573 (0,975; 2,150)	1,585	0,197	0,181 (0,089; 0,358)	0,167	0	5,507
50	0,5	0,526 (0,255; 0,810)	0,529	0,599	0,583 (0,430; 0,773)	0,575	4	69,794
	1,0	1,071 (0,706; 1,432)	1,077	0,349	0,322 (0,213; 0,481)	0,314	1	23,757
	1,5	1,517 (1,134; 1,896)	1,522	0,197	0,193 (0,122; 0,300)	0,187	0	11,949
100	0,5	0,507 (0,335; 0,688)	0,508	0,599	0,594 (0,490; 0,712)	0,591	3	146,397
	1,0	0,991 (0,775; 1,210)	0,992	0,349	0,352 (0,275; 0,446)	0,349	3	54,328
	1,5	1,503 (1,225; 1,778)	1,506	0,197	0,196 (0,141; 0,271)	0,193	2	24,398
200	0,5	0,498 (0,364; 0,635)	0,498	0,599	0,600 (0,519; 0,690)	0,598	8	300,024
	1,0	1,031 (0,868; 1,187)	1,032	0,349	0,337 (0,282; 0,403)	0,335	2	101,515
	1,5	1,496 (1,323; 1,670)	1,497	0,197	0,198 (0,161; 0,242)	0,197	3	49,304

Fonte: Elaborada pelo autor.

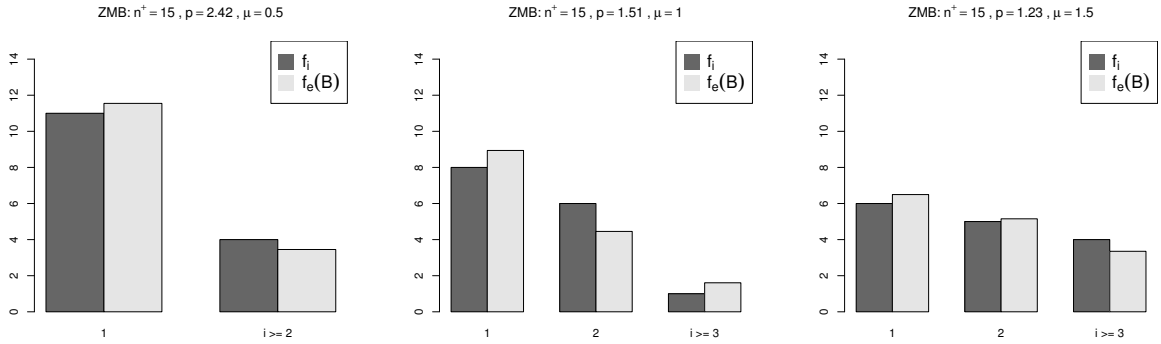
Figura 31 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 15$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



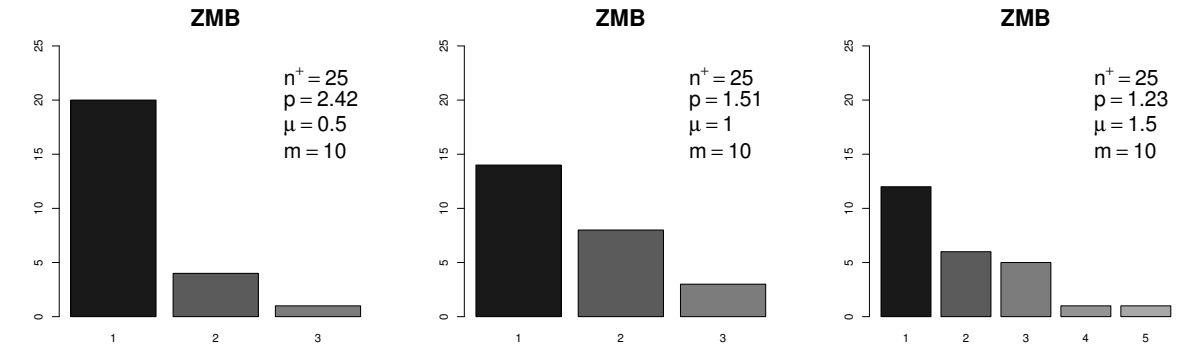
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



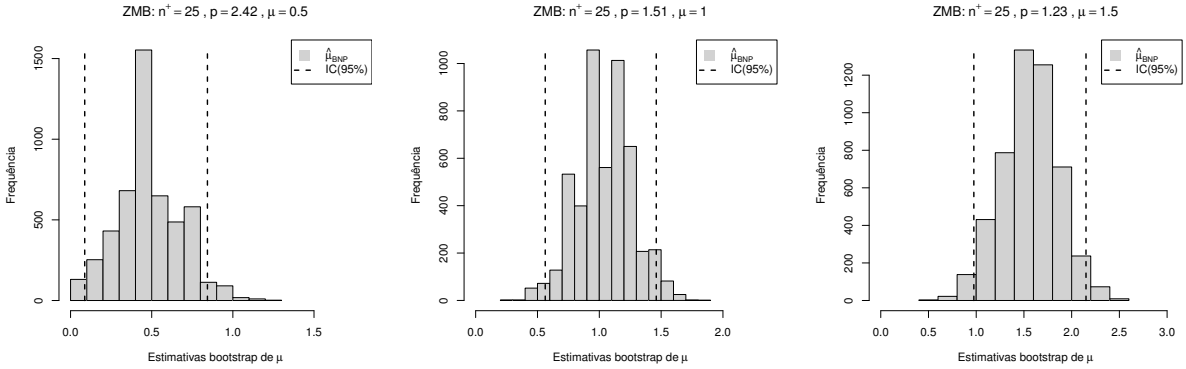
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

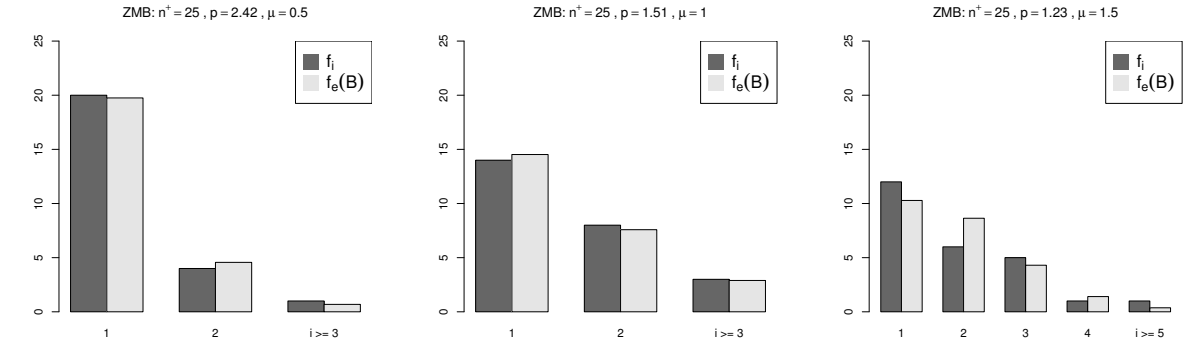
Figura 32 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 25$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



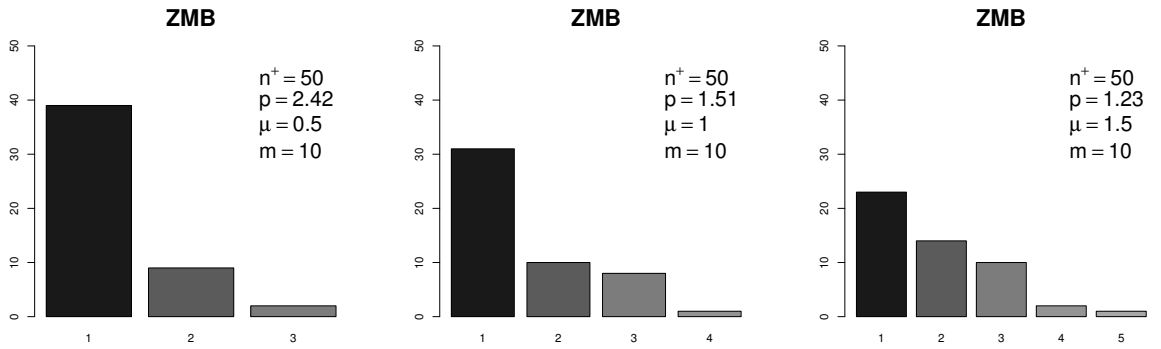
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



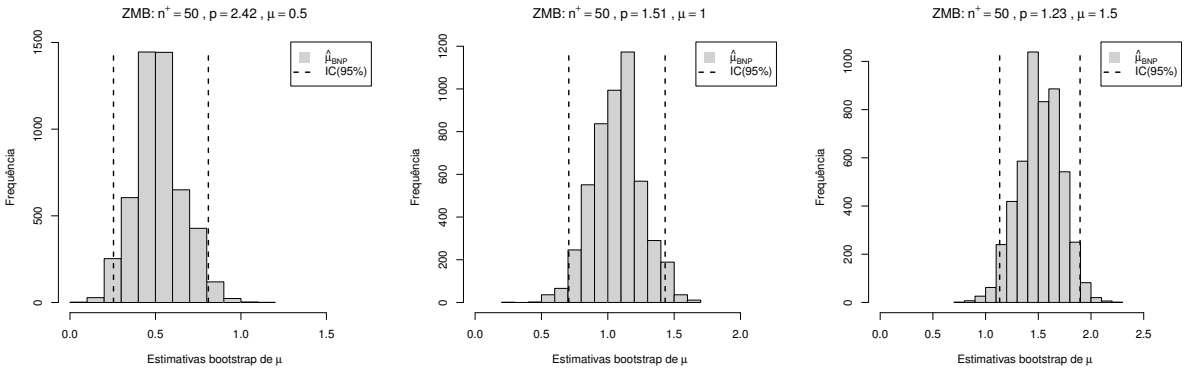
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

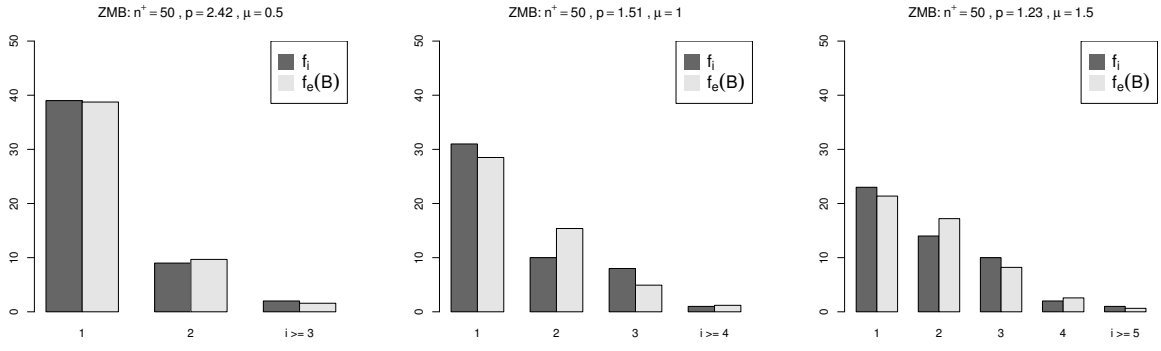
Figura 33 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 50$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



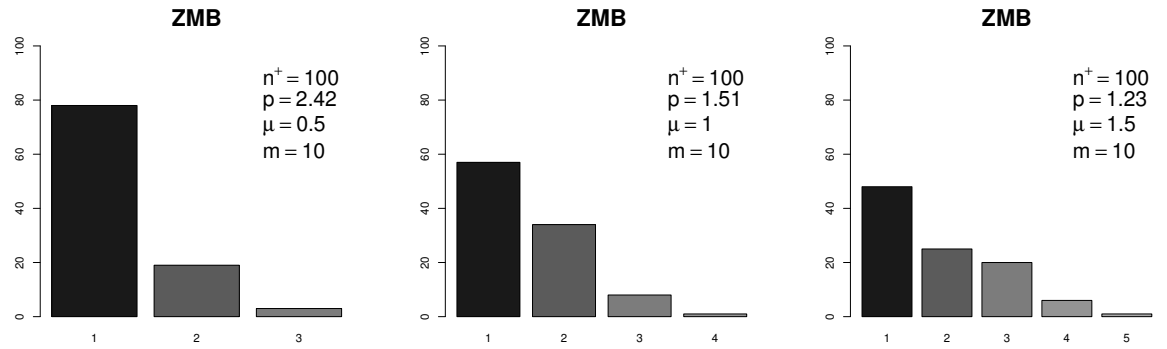
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



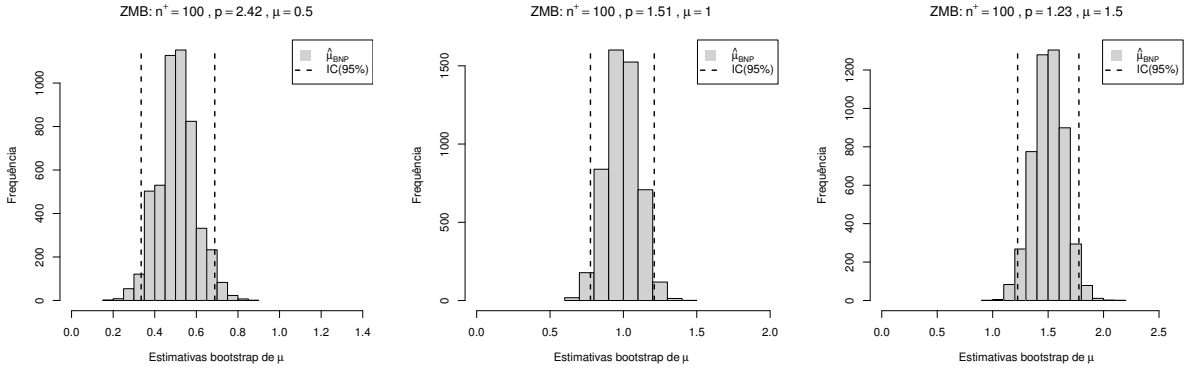
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

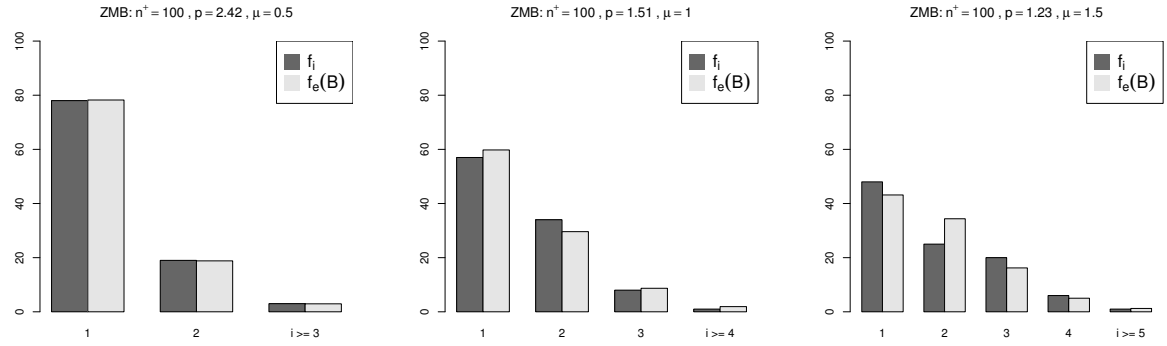
Figura 34 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 100$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



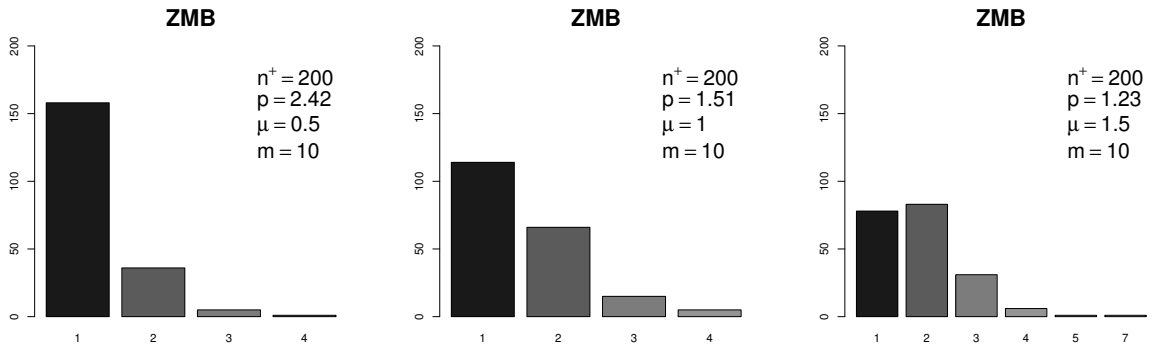
(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



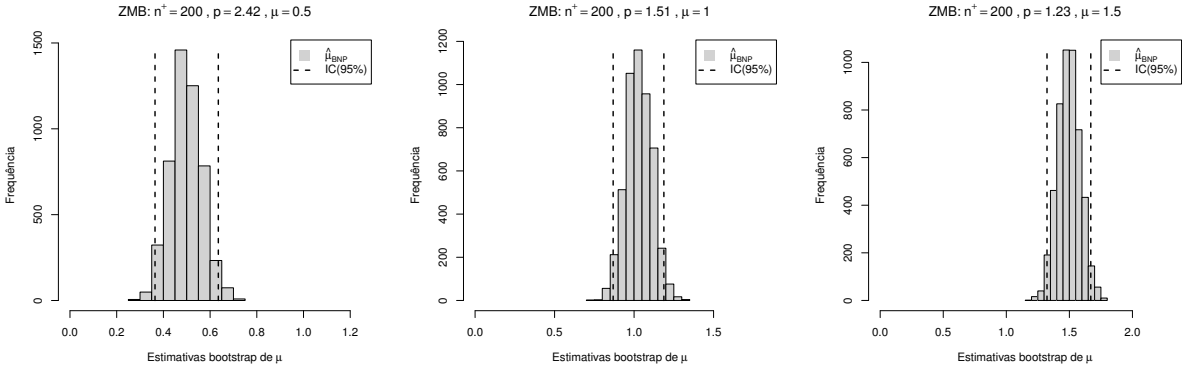
(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

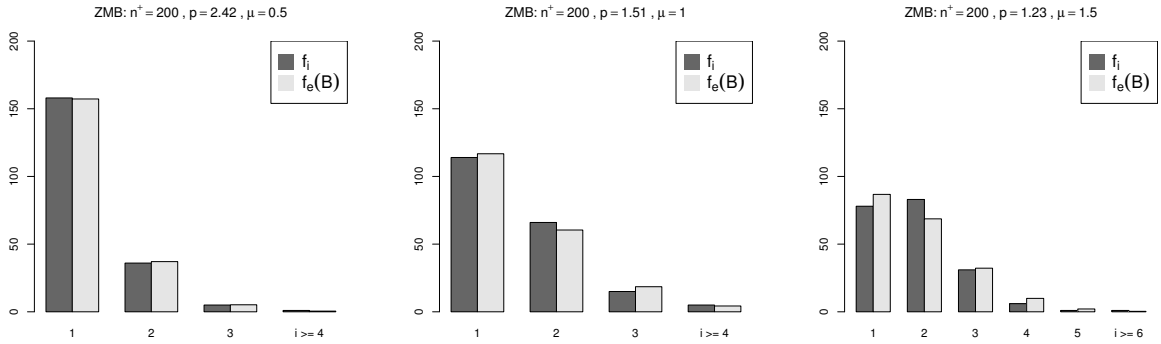
Figura 35 – Análises gráficas para os conjuntos de dados de tamanho $n^+ = 200$ gerados da distribuição Binomial (ZDB) para diferentes valores de μ.



(a) Distribuições de frequência dos conjuntos gerados.



(b) Distribuições empíricas dos estimadores do parâmetro μ obtidas via bootstrap não paramétrico.



(c) Distribuições de frequências observadas e esperadas pela distribuição ajustada.

Fonte: Elaborada pelo autor.

Com o conhecimento da eficiência do procedimento de estimação via algoritmo EM, verificada através de um estudo de simulação e também de um estudo com aplicações em conjuntos de dados gerados artificialmente, a metodologia proposta será aplicada em conjuntos de dados reais. Os resultados destas aplicações são apresentados no Capítulo 5 a seguir.

CAPÍTULO

5

APLICAÇÕES: DADOS REAIS

O método de estimação via algoritmo EM apresentado no Capítulo 3 foi aplicado para estimar as quantidades de interesse a partir do estudo de conjuntos de dados reais. Para ilustrar essas aplicações, foram considerados conjuntos de dados que contêm observações zero, incluindo situações anteriormente estudadas quanto à zero-modificação e cujos resultados já foram apresentados em outros trabalhos científicos para fins de comparação (Situação 1). Além destes, foi considerada também a análise de conjuntos que de fato não contêm zeros dentre suas observações, cuja investigação sobre a zero-modificação (zero-deflação, mais especificamente) foi efetuada neste trabalho (Situação 2). A seguir, são apresentadas as análises realizadas que incluem obtenção e comparação de estimativas, representação gráfica da distribuição empírica do estimador do parâmetro μ obtida via BNP e comparação gráfica das frequências observadas e esperadas para os conjuntos de dados considerados. Ressalta-se que poderiam ser aplicados ainda testes estatísticos para verificar a aderência das distribuições propostas aos dados estudados (como o teste χ^2 de aderência, por exemplo). Porém, tal análise adicional não faz parte do escopo deste trabalho.

Situação 1: Conjuntos ZMPS com zeros observados

A seguir, são apresentadas aplicações do procedimento de estimação proposto a conjuntos de dados reais que contêm zeros dentre suas observações, incluindo problemas para os quais é possível fazer a comparação com resultados quanto à zero-modificação, obtidos em trabalhos da literatura (Problemas 1.1 - 1.3) e um problema estudado neste trabalho de forma inédita (Problema 1.4).

Problema 1.1: Incidentes de terrorismo internacional nos EUA

Os dados, disponibilizados em Conigliani, Castro e O'Hagan (2000), consistem no número mensal de incidentes de terrorismo internacional nos Estados Unidos da América (EUA) entre os anos de 1968 e 1974. Os impactos do terrorismo internacional são de grande interesse e estudos sobre o tema permitem que as principais instituições nacionais possam agir de modo a prevenir e punir atos terroristas, além de realizar análises sobre seu funcionamento, financiamento e organização estrutural e estratégica, o que possibilita que o país tenha aparato institucional para lidar com a problemática.

Este conjunto de dados foi analisado por Bayarri, Berger e Datta (2008), que verificaram o ajuste do modelo Poisson aos dados como adequado, algo também confirmado por Conceição *et al.* (2017), que consideram um modelo mais geral, o modelo ZMP (cujas estimativas obtidas para os parâmetros foram $\hat{\mu} = 0,72$ e $\hat{p} = 0,96$ — próximo de 1.)

A Tabela 19 apresenta a distribuição de frequência do número de incidentes, e inclui as estatísticas descritivas como as médias amostrais ($\bar{y}$ e $\bar{y}^+$) e desvios-padrões amostrais ($sd(\boldsymbol{y})$ e $sd(\boldsymbol{y}^+)$), calculadas considerando-se o conjunto de dados completos e também o conjunto que consiste apenas das observações positivas.

Tabela 19 – Distribuição de frequência e estatísticas descritivas do número mensal de incidentes de terrorismo internacional nos EUA entre 1968 e 1974.

Número de incidentes	y_i	0	1	2	3	4	Total	$\bar{y}$	$\bar{y}^+$	$sd(y)$	$sd(y^+)$
Número de meses	f_i	38	26	8	2	1	75	0,693	1,405	0,870	0,725

Fonte: Adaptada de Conceição *et al.* (2017).

Para a utilização do procedimento de estimação proposto, foi considerado apenas o conjunto de observações positivas de tamanho $n^+ = 37$. As estimativas (incluindo as corrigidas) e os intervalos de confiança *bootstrap* ao nível de 95% para μ e $\pi_P(0;\mu)$ são apresentadas na Tabela 20 a seguir, além da estimativa de n_0.

Tabela 20 – Estimativas obtidas para o conjunto de incidentes de terrorismo internacional nos EUA.

n^+	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_P(0;\widehat{\mu})$	$\pi_P(0;\widehat{\mu}_C)$	$\widehat{n}_0$
37	0,724 (0,357; 1,100)	0,729	0,485 (0,333; 0,700)	0,474	34,805

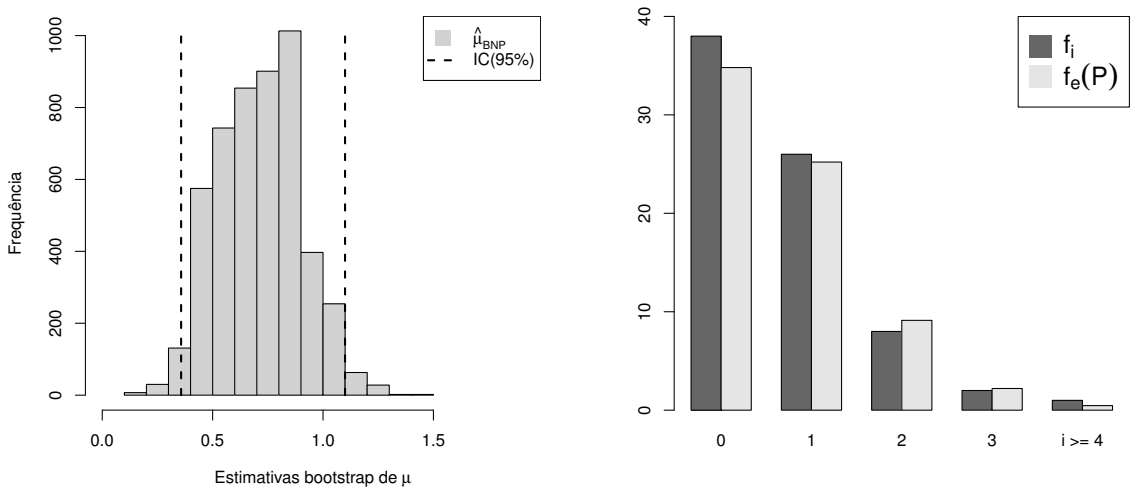
Fonte: Elaborada pelo autor.

Verifica-se que a estimativa $\widehat{\mu}$ obtida pelo algoritmo EM descrito neste trabalho foi bastante próxima ao valor estimado por Conceição *et al.* (2017), indicando uma média aproximada abaixo de um incidente mensal no período em estudo. Além disso, o número estimado de zeros foi $\widehat{n}_0 \approx 35$, também próximo ao valor real da amostra, corroborando o fato de que cerca de 50% dos meses não tiveram nenhum ataque (nota-se que a probabilidade de zero estimada também é

próxima a esse valor). Assim, é possível constatar que o procedimento de estimação é apropriado para o conjunto de dados sobre os incidentes de terrorismo por mês nos EUA, indicando que boas inferências podem ser obtidas e aplicadas, por exemplo, na tomada de decisão quanto à medidas de redução, prevenção e controle de tais ocorrências indesejadas.

A Figura 36 (a) ilustra graficamente a distribuição empírica do estimador do parâmetro μ, destacando os percentis do intervalo de confiança de 95%. Nota-se que tal distribuição é aproximadamente simétrica. A Figura 36 (b) mede visualmente a aderência entre a distribuição teórica especificada (Poisson) e o conjunto de dados amostrais estudado. Pode-se notar que há pouca diferença entre as frequências observada e esperada, o que ressalta a adequabilidade da suposição.

Figura 36 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado.



(a) (b)

Fonte: Elaborada pelo autor.

Problema 1.2: Partidas de futebol entre os times Real Madrid e Barcelona

Os dados, extraídos de Conceição *et al.* (2016), consistem no número de vitórias do *Barcelona* que antecederam cada vitória do *Real Madrid* entre março de 1916 a abril de 2014. Chamado de *El Clásico* (em português, O Clássico), a disputa é considerada o maior clássico de futebol de toda Espanha, sendo o confronto entre dois times que foi mais vezes disputado. A rivalidade transcende os campos de futebol e tem raízes históricas e políticas entre as regiões que os clubes representam (barcelonas.com, 2019).

Considerando-se as partidas de futebol disputadas entre os times *Barcelona* e *Real Madrid* no período de março de 1916 a abril de 2014, foram 227 jogos. Destes, 179 partidas resultaram em vitória de um time (88 para o *Barcelona* e 91 para *Real Madrid*). Levando-se em conta apenas estes jogos com um time vitorioso, o interesse é a contagem no número de vitórias do *Barcelona* que antecederam cada vitória do *Real Madrid*, sugerindo uma distribuição Geométrica. Isto é, o "fracasso" e o "sucesso" associados ao ensaio Bernoulli são, respectivamente, a vitória do *Barcelona* (ou a derrota do *Real Madrid*) e a vitória do *Real Madrid* (ou derrota do *Barcelona*).

Este conjunto de dados foi analisado por Conceição *et al.* (2016), que ajustaram o modelo ZMG aos dados e verificaram a presença de deflação de observações zero (cujas estimativas obtidas foram $\hat{\mu} = 0,72$ e $\hat{p} = 1,40$ — ou seja, uma distribuição ZDG).

A Tabela 21 apresenta a distribuição de frequência do número de vitórias do *Barcelona* que antecederam cada vitória do *Real Madrid* nas partidas consideradas, incluindo também as médias amostrais e desvios-padrões, calculadas com o conjunto de dados completos ($\bar{y}$ e $sd(\boldsymbol{y})$) e também com o conjunto que consiste apenas das observações positivas ($\bar{y}^+$ e $sd(\boldsymbol{y}^+)$).

Tabela 21 – Distribuição de frequência e estatísticas descritivas do número de vitórias do *Barcelona* que antecederam cada vitória do *Real Madrid* no período de março de 1916 a abril de 2014.

Número de vitórias	y_i	0	1	2	3	4	5	Total	$\bar{y}$	$\bar{y}^+$	$sd(y)$	$sd(y^+)$
Frequência	f_i	39	31	11	6	3	1	91	0,967	1,692	1,140	1,020

Fonte: Adaptada de Conceição *et al.* (2016).

Apenas o conjunto de observações positivas de tamanho $n^+ = 52$ foi considerado e utilizou-se o procedimento de estimação proposto para obter as estimativas sobre as quantidades de interesse. A Tabela 20 apresenta as estimativas (incluindo as corrigidas) e os intervalos *bootstrap* com 95% de confiança para μ e $\pi_G(0;\mu)$. Observa-se também nesta tabela a estimativa de n_0, que corresponde a quantidade de observações zero que deve ser adicionada ao conjunto de dados para que a distribuição Geométrica tradicional explique adequadamente o comportamento dos dados.

Tabela 22 – Estimativas obtidas para o número de vitórias do *Barcelona* que antecederam cada vitória do *Real Madrid* no período de março de 1916 a abril de 2014.

n^+	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_G(0;\widehat{\mu})$	$\pi_G(0;\widehat{\mu}_C)$	$\widehat{n}_0$
52	0,692 (0,423; 0,981)	0,691	0,591 (0,505; 0,703)	0,587	75,113

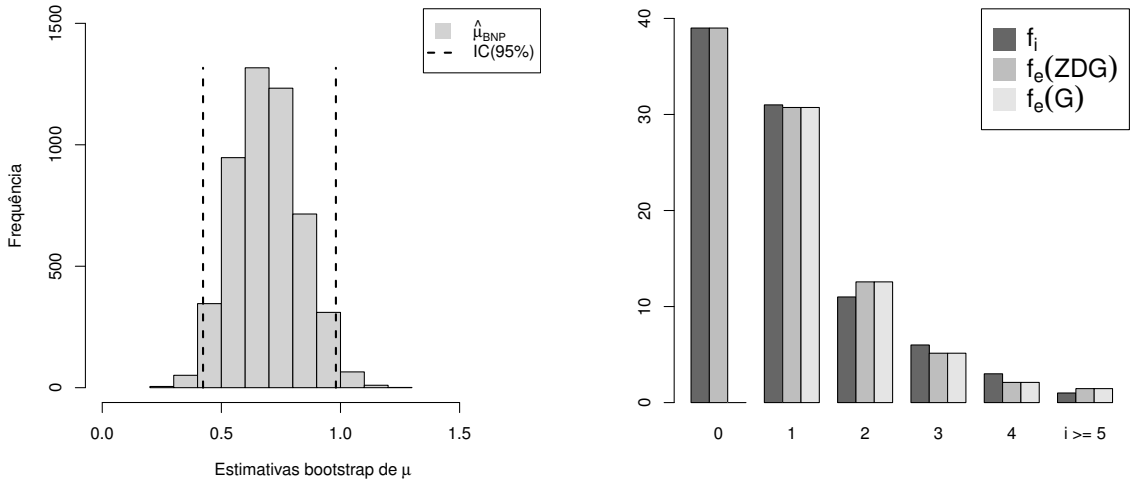
Fonte: Elaborada pelo autor.

Verifica-se que a estimativa $\widehat{\mu}$ obtida pelo algoritmo EM descrito neste trabalho foi próxima ao valor estimado por Conceição *et al.* (2016), indicando que uma média de aproximadamente $0,7$ vitórias do *Barcelona* antecederam cada vitória do *Real Madrid*. Além disso, o valor estimado de n_0, $\widehat{n}_0 \approx 76$, é maior do que o valor observado efetivamente no conjunto de

dados (39 observações zero), reforçando o fato concluído por Conceição *et al.* (2016) sobre a existência de um processo que deflacionou a amostra na observação zero (zero-deflação). Assim, supõe-se que o time do *Barcelona* deixou de ser o vitorioso do confronto menos vezes do que o suposto pela distribuição.

A Figura 37 (a) ilustra graficamente a distribuição empírica do estimador do parâmetro μ, destacando os percentis do intervalo de confiança de 95%. Nota-se que tal distribuição é aproximadamente simétrica. A Figura 37 (b) mede visualmente a aderência entre a distribuição teórica especificada (Geométrica) e o conjunto de dados amostrais estudado, destacando ainda as frequências esperadas para a distribuição zero-deflacionada. Pode-se notar que há pouca diferença entre as frequências observada e esperada, o que ressalta a adequabilidade da suposição. Destaca-se que para a observação zero, a frequência observada é bastante próxima à esperada pela distribuição zero-deflacionada. Ressalta-se que a barra que corresponde à frequência de zero para a distribuição Geométrica tradicional não é apresentada (havendo em vez disso apenas uma linha representativa), pois trata-se da frequência obtida para o conjunto de dados aumentados, que é superior às demais (como evidenciado pela estimativa $\widehat{n}_0$).

Figura 37 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado.



(a) (b)

Fonte: Elaborada pelo autor.

Problema 1.3: Estilo de escrita de um autor

Em estudos que visam caracterizar o estilo de escrita de um autor, toma-se amostras de n palavras e conta-se o número de palavras funcionais (usadas apenas para estabelecer relações entre elementos do discurso) em cada amostra. Os dados aqui utilizados, estão disponíveis em Bailey (1990), e consistem no número de ocorrências combinadas dos artigos *"the"*, *"a"* e *"an"* em amostras (não sobrepostas) de 5 palavras retiradas do livro de Macaulay (1909), tomadas a

partir de palavras iniciais de duas linhas escolhidas aleatoriamente de cada uma de 50 páginas do texto impresso. Assim, $m = 5$ é o número de palavras e Y é o número de ocorrências combinadas dos artigos *"the"*, *"a"* e *"an"*. Bailey (1990) verificou o ajuste do modelo B aos dados, que apresentam evidências de subdispersão. O mesmo conjunto de dados foi analisado por Conceição *et al.* (2017), que ajustaram o modelo ZMB e concluíram que há um processo de zero-deflação (cujas estimativas obtidas foram $\hat{\mu} = 0,31$ e $\hat{p} = 1,99$ — ou seja, uma distribuição ZDB).

A Tabela 23 apresenta a distribuição de frequência do número de ocorrências combinadas dos artigos *"the"*, *"a"* e *"an"* na amostra, as médias amostrais ($\bar{y}$ e $\bar{y}^+$) e desvios-padrões amostrais ($sd(\boldsymbol{y})$ e $sd(\boldsymbol{y}^+)$).

Tabela 23 – Distribuição de frequência e estatísticas descritivas do número de ocorrências combinadas dos artigos *"the"*, *"a"* e *"an"* em amostras de 5 palavras de Macaulay (1909).

Número de ocorrências	y_i	0	1	2	Total	$\bar{y}$	$\bar{y}^+$	$sd(y)$	$sd(y^+)$
Frequência	f_i	45	49	6	100	0,610	1,109	0,601	0,314

Fonte: Adaptada de Conceição *et al.* (2017).

Para aplicação do procedimento de estimação, foi considerado apenas o conjunto formado pelas observações positivas, de tamanho $n^+ = 55$. As estimativas (incluindo as corrigidas) de μ e $\pi_B(0;\mu)$, juntamente com os respectivos intervalos *bootstrap* com 95% de confiança são apresentadas na Tabela 24. Adicionalmente, observa-se também nesta Tabela a estimativa de n_0, indicando a quantidade de observações zero que deve ser adicionada ao conjunto de dados para que a distribuição Binomial tradicional explique adequadamente o comportamento dos dados.

Tabela 24 – Estimativas obtidas para o número de ocorrências combinadas dos artigos *"the"*, *"a"* e *"an"* na amostra citada.

n^+	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_B(0;\widehat{\mu})$	$\pi_B(0;\widehat{\mu}_C)$	$\widehat{n}_0$
55	0,259 (0,089; 0,456)	0,259	0,767 (0,620; 0,914)	0,763	180,622

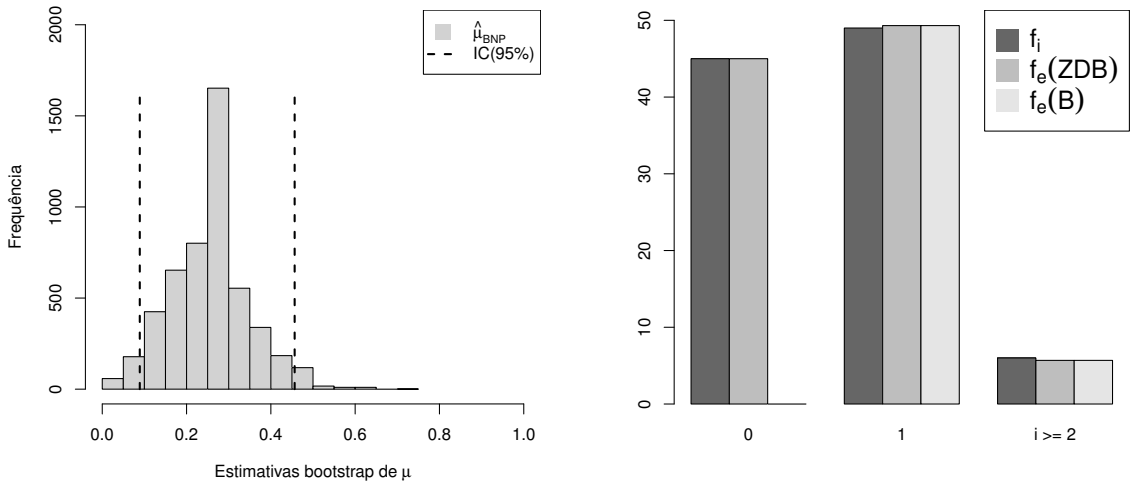
Fonte: Elaborada pelo autor.

Tem-se que a estimativa $\widehat{\mu}$ obtida pelo algoritmo EM descrito neste trabalho foi próxima ao valor estimado por Conceição *et al.* (2017) e que $\widehat{n}_0 \approx 181$ é maior do que o valor observado efetivamente na amostra (45 observações zero), reforçando o fato concluído por Conceição *et al.* (2017) sobre a existência de um processo que removeu a observação zero (zero-deflação). Assim, observa-se que o autor utiliza-se poucas vezes do uso combinado dos três artigos a cada 5 palavras — em média, aproximadamente $0,3$ vezes. A frequência do evento de não ocorrência foi menor que a estimada segundo o ajuste da distribuição (dado que $\widehat{n}_0$ é consideravelmente maior que o número de zeros observado), cuja probabilidade é em torno de $0,8$.

A Figura 38 (a) ilustra graficamente a distribuição empírica do estimador do parâmetro μ, destacando os percentis do intervalo de confiança de 95%. Nota-se que tal distribuição é

aproximadamente simétrica. A Figura 38 (b) mede visualmente a aderência entre a distribuição teórica especificada (Binomial) e o conjunto de dados amostrais estudado, destacando ainda as frequências esperadas para a distribuição zero-deflacionada. Pode-se notar que há pouca diferença entre as frequências observada e esperada, o que ressalta a adequabilidade da suposição. Destaca-se que para a observação zero, a frequência observada é bastante próxima à esperada pela distribuição zero-deflacionada. Novamente, ressalta-se que a barra que corresponde à frequência de zero para a distribuição Binomial tradicional não é apresentada (havendo em vez disso apenas uma linha representativa), pois trata-se da frequência obtida para o conjunto de dados aumentados, que é superior às demais (como evidenciado pela estimativa $\widehat{n}_0$).

Figura 38 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado.



(a) (b)

Fonte: Elaborada pelo autor.

Problema 1.4: Variação do Dólar

Os dados, obtidos em Investing.com (2019), consistem na variação mensal do valor do Dólar americano em Reais brasileiros, entre julho de 1994 e setembro de 2019. O dólar é uma moeda que tem muita credibilidade e é utilizado para efetuar transações internacionais, o que faz com que variações em seu valor influenciem preços de produtos importados e exportados no mercado interno e afetem diretamente os consumidores do país.

Quando o valor do dólar aumenta, é mais vantajoso exportar os produtos de produção nacional, além de interessar aos importadores a realização de compras nas indústrias brasileiras, devido à possibilidade de pagar menos pelo produto. Com isso, o consumo dos produtos nacionais aumenta no exterior; por outro lado aumenta também o valor deles no mercado interno.

Assim, com o intuito de fazer inferências sobre a variação do Dólar, mais especificamente nas ocorrências de eventos de queda em seu valor, supõe-se que as observações são decorrentes de um processo Bernoulli, cujo sucesso é a queda no valor do Dólar. A partir desse processo, define-se uma variável aleatória Y que representa o número de meses consecutivos de aumento na moeda norte-americana (fracasso) até a ocorrência de uma queda em seu valor (sucesso). Assim, faz-se a suposição de uma distribuição Geométrica para a variável aleatória Y. A Tabela 25 apresenta a distribuição de frequência do número de meses com aumento do Dólar que antecederam cada queda, incluindo também as médias amostrais e desvios-padrões, calculadas com o conjunto de dados completos ($\bar{y}$ e $sd(\boldsymbol{y})$) e também com o conjunto que consiste apenas das observações positivas ($\bar{y}^+$ e $sd(\boldsymbol{y}^+)$).

Tabela 25 – Distribuição de frequência e estatísticas descritivas do número de meses com variação positiva do dólar até a ocorrência de uma variação negativa, no período entre julho de 1994 e setembro de 2019.

Número de meses	y_i	0	1	2	3	4	5	6	7	44	Total	$\bar{y}$	$\bar{y}^+$	$sd(y)$	$sd(y^+)$
Frequência	f_i	66	38	8	3	7	5	1	2	1	131	1,374	2,769	4,073	5,457

Fonte: Adaptada de Investing.com (2019).

Apenas o conjunto de observações positivas de tamanho $n^+ = 65$ foi considerado e utilizou-se o procedimento de estimação proposto para obter as estimativas sobre as quantidades de interesse. A Tabela 26 apresenta as estimativas (incluindo as corrigidas) e os intervalos *bootstrap* com 95% de confiança para μ e $\pi_{_G}(0;\mu)$. Observa-se também nesta tabela a estimativa de n_0, que corresponde a quantidade de observações zero que deve ser adicionada ao conjunto de dados para que a distribuição Geométrica tradicional explique adequadamente o comportamento dos dados.

Tabela 26 – Estimativas obtidas para o número de usuários de meses com variação positiva do dólar até a ocorrência de uma variação negativa.

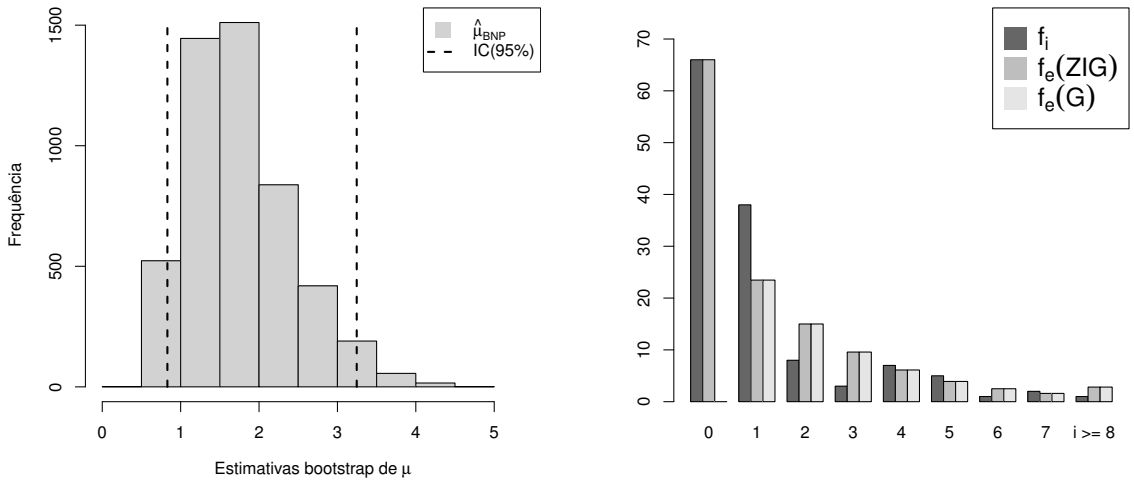
n^+	$\widehat{\mu}$	$\widehat{\mu}_{_C}$	$\pi_{_G}(0;\widehat{\mu})$	$\pi_{_G}(0;\widehat{\mu}_{_C})$	$\widehat{n}_0$
65	1,769 (0,831; 3,215)	1,777	0,361 (0,237; 0,546)	0,340	36,739

Fonte: Elaborada pelo autor.

Verifica-se que a estimativa $\widehat{\mu}$ obtida pelo algoritmo EM descrito neste trabalho foi de $1,777$, indicando que uma média de aproximadamente 2 meses de aumento do Dólar antecederam cada mês de queda de seu valor. Além disso, o valor estimado de n_0, $\widehat{n}_0 \approx 37$, é menor do que o valor observado de fato no conjunto de dados (66 observações zero), indicando a possibilidade de existência de um processo que inflacionou a amostra na observação zero (zero-inflação). Além disso, o valor estimado para o parâmetro p (utilizando a estimativa de μ) foi $0,324$ (também um indicativo de zero-inflação). Assim, supõe-se que o valor do Dólar deixou de sofrer aumentos mais vezes do que o suposto pela distribuição.

A Figura 39 (a) ilustra graficamente a distribuição empírica do estimador do parâmetro μ, destacando os percentis do intervalo de confiança de 95%. Nota-se que tal distribuição é aproximadamente simétrica. A Figura 39 (b) mede visualmente a aderência entre a distribuição teórica especificada (Geométrica) e o conjunto de dados amostrais estudado, destacando ainda as frequências esperadas para a distribuição zero-inflacionada. Pode-se notar que há pouca diferença entre as frequências observada e esperada, o que ressalta a adequabilidade da suposição. Destaca-se que para a observação zero, a frequência observada é bastante próxima à esperada pela distribuição zero-inflacionada. Ressalta-se que a barra correspondente à frequência de zero para a distribuição Geométrica tradicional não é apresentada (havendo em vez disso apenas uma linha representativa), pois trata-se da frequência obtida para o conjunto de dados aumentados, que é inferior às demais (como evidenciado pela estimativa $\widehat{n}_0$).

Figura 39 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado.



(a) (b)

Fonte: Elaborada pelo autor.

Situação 2: Conjuntos ZMPS sem zeros observados

A seguir, são apresentadas aplicações do procedimento de estimação proposto a conjuntos de dados reais que não contêm zeros dentre suas observações, incluindo um problema para o qual é possível fazer a comparação com resultados obtidos em estudos anteriores (Problema 2.1) e problemas estudados originalmente neste trabalho, no que diz respeito à zero-modificação (Problemas 2.2 e 2.3).

Problema 2.1: População de usuários de drogas em Bancoque

Os dados, estudados em Böhning e Heijden (2009), consistem no número de contatos feitos por usuários de metanfetamina à clínicas de tratamento em Bancoque, com base nos registros feitos em todos os 61 centros de saúde na região metropolitana da cidade, entre 1º de outubro a 31 de dezembro de 2001.

Considerada um problema sério na Tailândia, a metanfetamina em forma de pílula é a principal droga sintética ilícita em uso no país (United Nations Office on Drugs and Crime, 2012). Estudos que consideram a população dos usuários são usados, por exemplo, para obter informações sobre drogas emergentes e para tomada de decisões sobre mudanças na legislação visando resolver o problema das drogas sintéticas.

Como salientado por Böhning e Heijden (2009), considerando-se Y como a variável aleatória que representa o número de contatos que um usuário de drogas teve com as instituições de tratamento, e derivando-se uma lista de usuários de drogas a partir do registro de todos os contatos feitos a tais instituições, um usuário só é observado se houve um número positivo de contatos com as instituições de tratamento. Assim, a lista não terá a observação $y = 0$, refletindo portanto, uma variável de contagem truncada em zero. Com base na frequência das procuras por tratamento de cada paciente, surge uma distribuição de contagem que pode ser ajustada por meio de um modelo Poisson. Usando esse modelo, uma estimativa para o número de usuários de drogas não observados pode ser construída, obtendo-se, assim, uma estimativa do tamanho da população de interesse.

A Tabela 27 apresenta a distribuição de frequência do número usuários de metanfetamina com exatamente y_i visitas à clínicas de tratamento, bem como a média amostral ($\bar{y}^+$) e desvio-padrão amostral ($sd(\boldsymbol{y}^+)$) para o conjunto de observações. Os autores de Böhning e Heijden (2009) obtiveram, sob a suposição do modelo de contagem Poisson homogêneo[1] para este conjunto de dados, a estimativa $\widehat{n} = 15325$ (com intervalo com 95% de confiança dado por $(13989; 16661)$.

Tabela 27 – Distribuição de frequência e estatísticas descritivas do número de usuários de metanfetamina com exatamente y_i visitas à clínicas de tratamento.

Número de visitas	y_i	1	2	3	4	5	6	7	8	9	10	12	Total	$\bar{y}^+$	$sd(y^+)$
Frequência	f_i	3114	163	23	20	9	3	3	3	4	3	1	3346	1,128	0,664

Fonte: Adaptada de Böhning e Heijden (2009).

As estimativas (incluindo as corrigidas) de μ e $\pi_P(0;\mu)$, juntamente com os respectivos intervalos *bootstrap* com 95% de confiança são apresentadas na Tabela 28. Ainda nesta Tabela,

[1] Os autores sugerem que sob a suposição de heterogeneidade e de um modelo de mistura para os dados (zero-truncados), um estimador mais robusto a ser utilizado para inferências sobre a frequência de zeros e tamanho populacional deve ser o proposto por Zelterman (1988), que é um limitante superior para a estimativa do tamanho populacional, em vez do EMV simples. As estimativas obtidas segundo esta abordagem podem ser encontradas no artigo Böhning e Heijden (2009).

observa-se a estimativa de n_0, indicando a quantidade de observações zero que deve ser adicionada ao conjunto de dados para que a distribuição Poisson tradicional explique adequadamente o comportamento do conjunto de dados. Pode-se notar que o valor estimado de n_0 — $\widehat{n}_0 \approx 11976$ — é bastante alto, indicando que há muito mais usuários de metanfetamina do que aqueles registrados por sua procura por tratamentos em clínicas. Isto é, a estimativa para o total populacional é, neste caso, $\widehat{n} = 15321,250$, o que reforça a evidência de que o conjunto de dados é deflacionado de zeros.

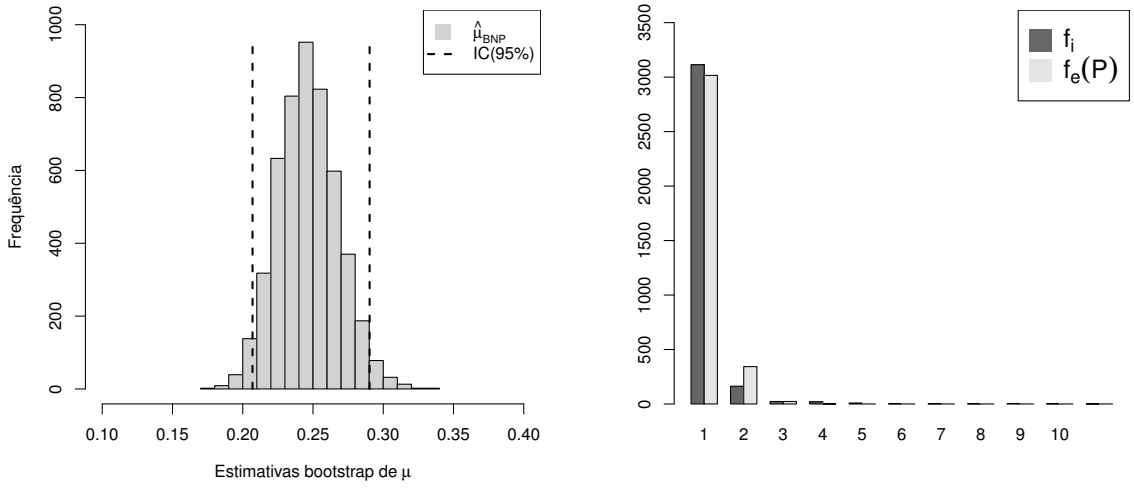
Tabela 28 – Estimativas obtidas para o número de visitas feitas por usuários de metanfetamina à clínicas de tratamento.

n^+	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_P(0;\widehat{\mu})$	$\pi_P(0;\widehat{\mu}_C)$	$\widehat{n}_0$
3346	0,246 (0,206; 0,290)	0,246	0,782 (0,748; 0,814)	0,781	11975,250

Fonte: Elaborada pelo autor.

A Figura 40 (a) ilustra graficamente a distribuição empírica do estimador do parâmetro μ, destacando os percentis do intervalo de confiança de 95%. Nota-se que tal distribuição é aproximadamente simétrica. A Figura 40 (b) mede visualmente a aderência entre a distribuição teórica especificada (Poisson) e o conjunto de dados amostrais estudado. Pode-se notar que há pouca diferença entre as frequências observada e esperada, o que ressalta a adequabilidade da suposição da distribuição.

Figura 40 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado.



(a) (b)

Fonte: Elaborada pelo autor.

Problema 2.2: Acidentes em indústrias químicas

Os dados, extraídos de Reddy e Yarrakula (2016), consistem em registros de acidentes ocorridos em indústrias de processos químicos no período de 1998 a 2015, em 29 países. O aumento no número desses acidentes e o potencial dos danos é motivo de preocupação em todo o mundo, fazendo-se necessária a realização de estudos capazes de fornecer informações que permitam conhecer melhor os acidentes passados e aprender lições sobre eles, de forma a desenvolver estratégias de prevenção e mitigação.

Duas características de interesse consideradas pelos autores de Reddy e Yarrakula (2016) para análise foram a frequência de acidentes e o número de pessoas feridas e mortas, em países desenvolvidos e em desenvolvimento. Os pesquisadores enfatizam que o número de mortes é consideravelmente menor nos países desenvolvidos que nos em desenvolvimento, mas o número de feridos é maior no primeiro grupo. Além disso, em geral, as indústrias em questão são relutantes em revelar as causas dos acidentes e o número de vítimas; assim, devido a condições políticas, sociais e econômicas, muitos acidentes não foram reportados pelos bancos de dados e mídias internacionais.

As indústrias de processos químicos, segundo Swedish Chemicals Agency (2011), têm sua maior produção concentrada em países da Europa, no Japão e nos EUA — considerados países desenvolvidos —, embora haja expressivo crescimento de países em desenvolvimento neste cenário, como a China, por exemplo. A Tabela 29 apresenta as contagens do número de acidentes para o período de 1998 a 2015 em 15 países desenvolvidos e em 14 países em desenvolvimento. Ressalta-se a ausência da observação zero nos dois conjuntos apresentados. A Tabela apresenta ainda a média amostral ($\bar{y}^+$) e desvio-padrão amostral ($sd(\boldsymbol{y}^+)$).

Tabela 29 – Distribuição de frequência e estatísticas descritivas do número de acidentes no período de 1998 a 2015 em países desenvolvidos e em países em desenvolvimento.

Países	y_i	1	2	3	5	11	15	*Total*	$\bar{y}^+$	$sd(y^+)$
Desenvolvidos	f_i	6	6	1	1	-	1	15	2,733	3,555
Em desenvolvimento	f_i	10	1	2	-	1	-	14	2,071	2,674

Fonte: Adaptada de Reddy e Yarrakula (2016).

Para cada conjunto de dados (países desenvolvidos e em desenvolvimento), o procedimento de estimação foi considerado sob a suposição de uma distribuição Poisson. As estimativas (incluindo as corrigidas) de μ e $\pi_P(0;\mu)$, juntamente com os respectivos intervalos *bootstrap* com 95% de confiança são apresentadas na Tabela 30. Ainda nesta Tabela, observa-se a estimativa de n_0, indicando a quantidade de observações zero que deve ser adicionada ao conjunto de dados para que a distribuição Poisson tradicional explique adequadamente o comportamento de cada conjunto de dados. Para os países desenvolvidos, pode-se notar o baixo valor estimado de n_0, $\widehat{n}_0 \approx 2$, indicando que a distribuição Poisson é adequada para explicar o comportamento dos dados até mesmo sem a inclusão das observações zero estimadas, uma vez que as estimativas $\widehat{\mu}$ e $\widehat{\mu}_C$ são próximas da média amostral $\bar{y}^+$ (que é o EMV de μ da distribuição Poisson).

Tabela 30 – Estimativas obtidas para os conjuntos de acidentes químicos.

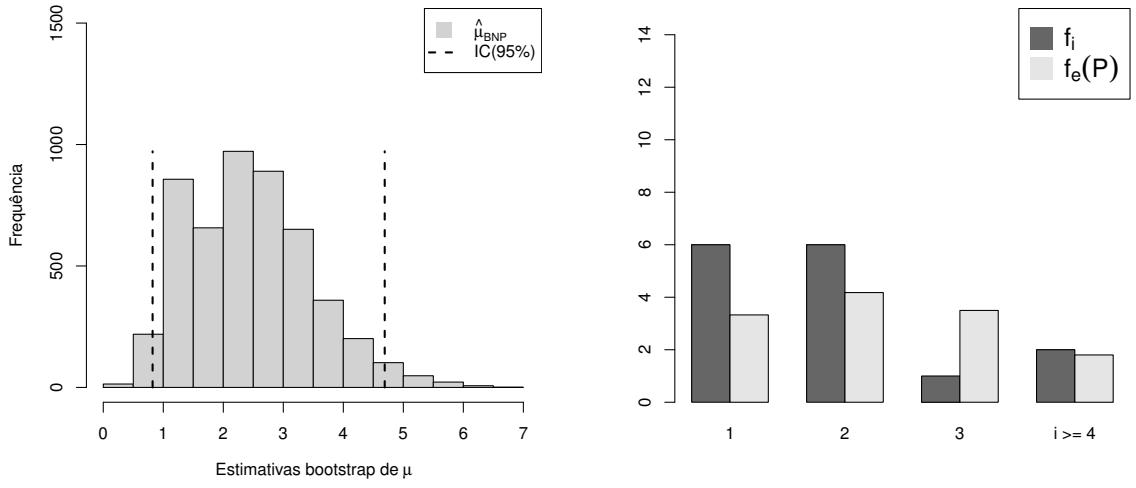
Países	n^+	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_P(0;\widehat{\mu})$	$\pi_P(0;\widehat{\mu}_C)$	$\widehat{n}_0$
Desenvolvidos	15	2,512 (0,822; 4,690)	2,587	0,081 (0,009; 0,440)	0,026	1,325
Em desenvolvimento	14	1,689 (0,273; 3,459)	1,771	0,185 (0,031; 0,761)	0,087	3,173

Fonte: Elaborada pelo autor.

Nota-se que a média estimada de acidentes é mais baixa para países em desenvolvimento ($1,689$ contra $2,512$ para países desenvolvidos), possível reflexo da concentração da produção ser maior no outro grupo de países. Como destacado também em Reddy e Yarrakula (2016), os países em desenvolvimento sofrem da falta de legislação rigorosa, conscientização pública e mais regulamentos de segurança, medidas mais consolidadas nos demais países e que podem resultar em mais acidentes cujas ocorrências foram efetivamente registradas — algo que também pode influenciar na maior média estimada.

As Figuras 41 (a) e 42 (a) ilustram graficamente a distribuição empírica do estimador do parâmetro μ, destacando os percentis do intervalo de confiança de 95%, para o conjunto que consiste do acidentes nos países desenvolvidos e nos países em desenvolvimento, respectivamente. Verifica-se que as distribuições apresentam leve assimetria. As Figuras 41 (b) e 42 (b) medem visualmente a aderência entre a distribuição teórica especificada (Poisson) e os conjuntos de dados amostrais estudados. Pode-se notar que não há diferenças muito grandes entre as frequências observada e esperada, apesar dos tamanhos amostrais serem pequenos, o que ressalta a adequabilidade da suposição da distribuição.

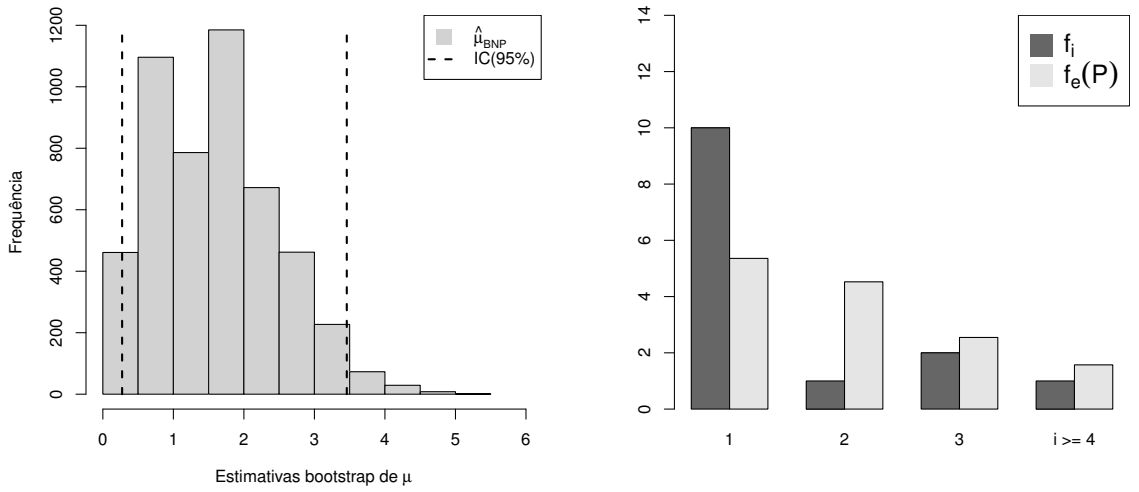
Figura 41 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado.



(a) (b)

Fonte: Elaborada pelo autor.

Figura 42 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado.



(a) (b)

Fonte: Elaborada pelo autor.

Pode-se realizar também a análise dos dados de acidentes em indústrias de processos químicos no período entre 1998 e 2015, considerando-se as observações como um único conjunto. A Tabela 31 contém as frequências e estatísticas descritivas para esse outro cenário, mais geral.

Tabela 31 – Distribuição de frequência e estatísticas descritivas do número de acidentes no período de 1998 a 2015.

Número de acidentes	y_i	1	2	3	5	11	15	Total	$\bar{y}^+$	$sd(y^+)$
Frequência	f_i	16	7	3	1	1	1	29	2,414	3,123

Fonte: Adaptada de Reddy e Yarrakula (2016).

Para esse conjunto de dados, o procedimento de estimação foi considerado também sob a suposição de uma distribuição Poisson. As estimativas (incluindo as corrigidas) de μ e $\pi_P(0;\mu)$, juntamente com os respectivos intervalos *bootstrap* com 95% de confiança são apresentadas na Tabela 32. Ainda nesta Tabela, observa-se a estimativa de n_0, indicando a quantidade de observações zero que deve ser adicionada ao conjunto de dados para que a distribuição Poisson tradicional explique adequadamente o comportamento do conjunto de dados. Pode-se notar que o valor estimado de n_0, $\widehat{n_0} \approx 4$, foi bastante próximo ao obtido levando-se em conta apenas as observações sobre acidentes em países em desenvolvimento. A média estimada, por sua vez, teve valor similar ao da média amostral do conjunto unificado, e também mais próximo ao estimado para o conjunto dos países desenvolvidos.

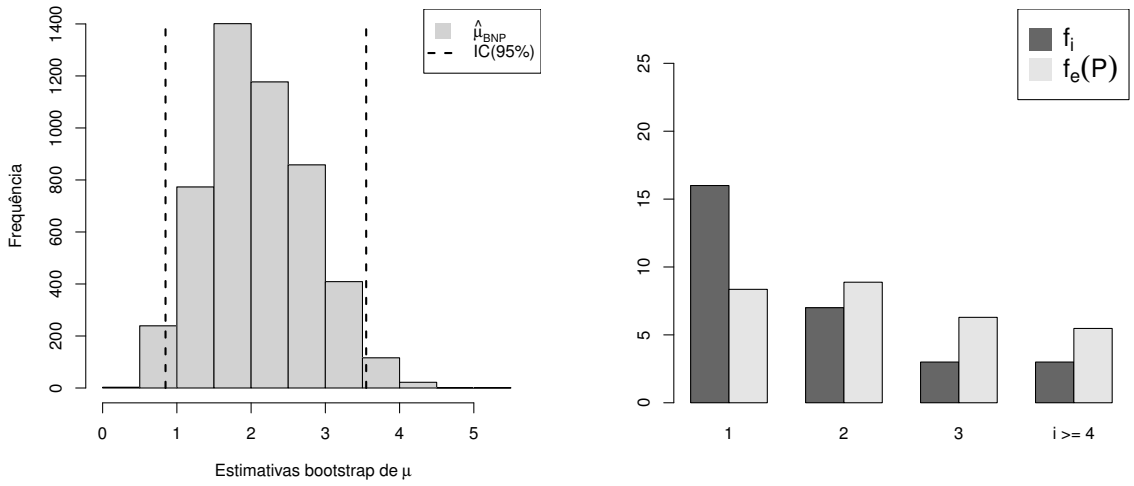
Tabela 32 – Estimativas obtidas para o número de acidentes químicos.

n^+	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_P(0;\widehat{\mu})$	$\pi_P(0;\widehat{\mu}_C)$	$\widehat{n}_0$
29	2,126 (0,847; 3,550)	2,164	0,119 (0,029; 0,429)	0,083	3,930

Fonte: Elaborada pelo autor.

Na Figura 43 (a) ilustra-se graficamente a distribuição empírica do estimador do parâmetro μ, destacando-se os percentis do intervalo de confiança de 95%. Verifica-se que a distribuição do estimador é aproximadamente simétrica (mais do que se comparada àquelas obtidas para os estimadores dos subconjuntos de países segundo seu desenvolvimento). A Figura 43 (b) mede visualmente a aderência entre a distribuição teórica especificada (Poisson) e o conjunto de dados amostrais estudado. As diferenças entre as frequências observada e esperada são menores que as apresentadas para os subconjuntos, mais uma vez podendo-se supor a adequabilidade da distribuição pressuposta, e levando a presumir que a análise feita considerando-se o conjunto dos 29 países permite inferências mais adequadas para o cenário geral.

Figura 43 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado.



(a) (b)

Fonte: Elaborada pelo autor.

Problema 2.3: Acidentes em indústrias petroquímicas na União Europeia

Os dados, apresentados em Nivolianitou, Konstandinidou e Michalis (2006), consistem no número de acidentes por ano que ocorreram em instalações petroquímicas e refinarias, entre os anos de 1985 e 2002, registrados no Sistema Europeu de Notificação de Acidentes Graves (MARS)[2].

O MARS foi criado na União Europeia (UE) para registrar todos os grandes acidentes industriais notificados às autoridades da UE pelos estados membros. A análise estatística dessas ocorrências oferece dados significativos para a compreensão e prevenção de acidentes químicos industriais. Destaca-se que as instalações petroquímicas são caracterizadas por altos níveis de risco devido à natureza das substâncias inflamáveis processadas e da gravidade das consequências, no caso de um acidente grave nesses estabelecimentos, que podem afetar muitas pessoas dentro e fora das instalações, além do ambiente circundante.

A Tabela 33 apresenta a frequência do número de acidentes por ano em indústrias petroquímicas entre os anos de 1985 e 2002, bem como a média amostral ($\bar{y}^+$) e desvio-padrão amostral ($sd(\boldsymbol{y}^+)$). Nota-se que houve apenas registro de números maiores que zero acidentes. Os autores de Nivolianitou, Konstandinidou e Michalis (2006) destacam que maiores números de acidentes reportados não significa necessariamente maior número de ocorrências; tal fato pode ser explicado como uma consequência de uma maior aceitação do MARS, que levou à notificação de quaisquer eventos acidentais considerados pelas autoridades competentes.

Tabela 33 – Distribuição de frequência e estatísticas descritivas do número de acidentes por ano em indústrias petroquímicas no período de 1985 a 2002.

Número de acidentes	y_i	1	2	3	4	5	6	7	8	9	Total	$\bar{y}^+$	$sd(y^+)$
Frequência	f_i	1	3	3	2	2	2	2	2	1	18	4,722	2,421

Fonte: Adaptada de Nivolianitou, Konstandinidou e Michalis (2006).

Considerando-se a distribuição Poisson para este conjunto de dados de contagem, o procedimento de estimação via algoritmo EM apresentado foi aplicado. As estimativas obtidas (incluindo as corrigidas) de μ e $\pi_P(0;\mu)$, juntamente com os respectivos intervalos *bootstrap* com 95% de confiança e a estimativa de n_0 são apresentadas na Tabela 34.

Verifica-se que o valor estimado para o parâmetro de média μ foi bastante próximo ao valor da média amostral do conjunto, indicando a adequabilidade da distribuição (o EMV da Poisson é $\bar{y}$). Além disso, a quantidade de zeros estimada foi baixa (com probabilidade também bastante baixa). Assim, pode-se inferir que o zero-acidente tem reduzida probabilidade de ocorrência, e ocorreram, em média, $4,7$ acidentes por ano no período estudado. Como ressaltado por Nivolianitou, Konstandinidou e Michalis (2006), o seguimento petroquímico tem

[2] Do inglês: *European Major Accident Reporting System*.

alta porcentagem de grandes acidentes dentre os demais setores químico-industriais, indicando que é preciso investir em medidas para reduzir o número de ocorrências e minimizar suas consequências.

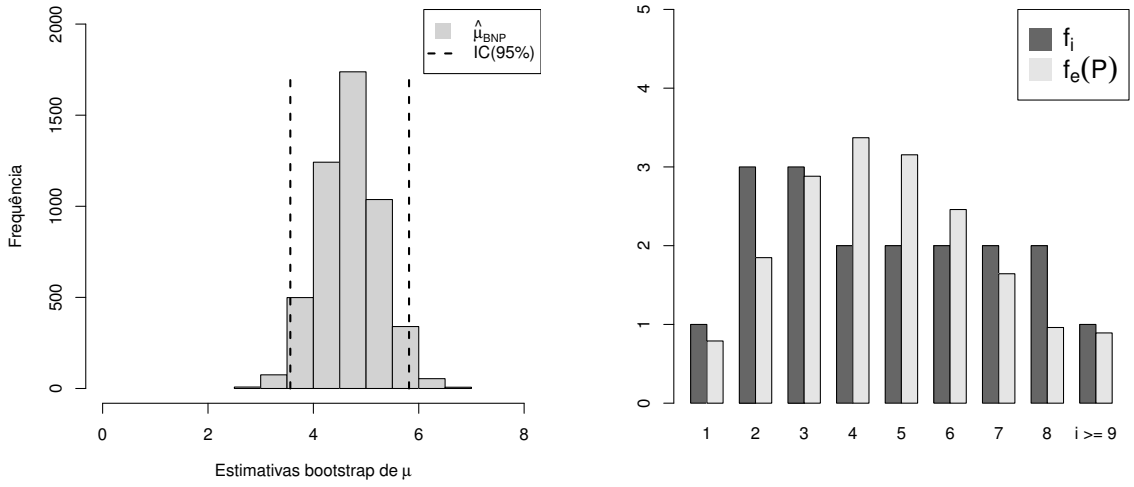
Tabela 34 – Estimativas obtidas para o número de acidentes por ano em indústrias petroquímicas entre 1985 e 2002.

n^+	$\widehat{\mu}$	$\widehat{\mu}_C$	$\pi_P(0;\widehat{\mu})$	$\pi_P(0;\widehat{\mu}_C)$	$\widehat{n}_0$
18	4,678 (3,563; 5,816)	4,676	0,009 (0,003; 0,028)	0,008	0,169

Fonte: Elaborada pelo autor.

A Figura 44 (a) ilustra graficamente a distribuição empírica do estimador do parâmetro μ, destacando os percentis do intervalo de confiança de 95%. Nota-se que tal distribuição é aproximadamente simétrica. A Figura 44 (b) mede visualmente a aderência entre a distribuição teórica especificada (Poisson) e o conjunto de dados amostrais estudado. Pode-se notar que há pouca diferença entre as frequências observada e esperada, mesmo apesar do pequeno tamanho amostral, o que ressalta a adequabilidade da suposição.

Figura 44 – (a): Distribuição empírica do estimador do parâmetro μ obtida via *bootstrap* não paramétrico. (b): Distribuição de frequências observada e esperada para o conjunto de dados considerado.



(a) (b)

Fonte: Elaborada pelo autor.

CAPÍTULO

6

CONSIDERAÇÕES FINAIS E PROPOSTAS FUTURAS

Com o objetivo de explicar adequadamente o comportamento de dados de contagem principalmente na situação em que não se observa zeros, mas há uma probabilidade positiva deste valor ocorrer, a apresentação de algumas definições se fez necessária. Para esta situação específica, foi apresentado neste trabalho um procedimento de estimação dos parâmetros das distribuições zero-modificadas, que são distribuições amplas e bastante flexíveis já que contemplam várias distribuições discretas tradicionais como casos particulares.

Inicialmente, apresentou-se notações e definições preliminares da família de distribuições PS, que são utilizadas ao longo do trabalho. Além disso, foi feita a descrição da família ZMPS (incluindo sua versão *hurdle*), que foi proposta por Conceição (2013) como alternativa aos casos em que ocorre zero-modificação no conjunto de dados de contagem. Mais geral, esta distribuição traz em sua forma a modificação, principalmente, da probabilidade de ocorrência de zero e, consequentemente, das probabilidades de observações positivas; tal fato foi evidenciado por meio de uma proposição apresentada neste trabalho. Características e aspectos inferenciais para os estimadores dos parâmetros da distribuição ZMPS via procedimentos de máxima verossimilhança e método dos momentos foram desenvolvidos ao longo deste trabalho.

Conjuntos de dados que não possuem a observação zero podem ser distinguidos entre zero-deflacionado e zero-truncado por meio da avaliação do pesquisador sobre a probabilidade de ocorrência de tal observação: baixa ou nula, respectivamente, sendo que dados de contagem zero-truncadas podem ocorrer devido à escolha do esquema amostral empregado. Para estes casos, a suposição de uma distribuição mais geral, a distribuição ZMPS, implicaria em probabilidade nula para a ocorrência de zero que em certos problemas práticos nem sempre é verdade. Para contornar este problema, foi proposto um procedimento de estimação via algoritmo EM, de forma a "aumentar" os dados para que uma distribuição mais simples, uma distribuição discreta tradicional, explique adequadamente o comportamento, principalmente, dos dados que de fato

foram observados (observações positivas). Assim, foi dado destaque à distribuição ZDPS e o caso de interesse deste trabalho, além de metodologia de estimação de máxima verossimilhança dos parâmetros deste modelo utilizando o algoritmo EM.

Para avaliar o desempenho do procedimento de estimação proposto, foi realizado um estudo de simulação e um estudo com conjuntos de dados artificiais, considerando as distribuições Poisson, Binomial e Geométrica, para os quais verificou-se que o procedimento produz boas estimativas, mesmo para tamanhos amostrais pequenos e os estimadores têm o comportamento assintótico desejado. Além disso, foi possível verificar que os valores gerados e estimados para a quantidade de zeros também evidencia se o conjunto de dados é proveniente de uma distribuição tradicional ou zero-modificada.

Para ilustrar a metodologia apresentada neste trabalho, foram consideradas algumas aplicações a conjuntos de dados reais com e sem zeros dentre suas observações, incluindo casos anteriormente analisados por outros autores quanto à zero-modificação, para fins de comparação. A análise dos resultados permitiu identificar características dos dados quanto aos aspectos de zero-modificação, dentre os quais ressalta-se a verificação de zero-inflação no Problema que avalia a variação do Dólar, algo que influencia diretamente o mercado nacional e suas transações internacionais; e a comprovação da zero-deflação constatada por Conceição *et al.* (2016) no que se refere ao número de vitórias do time *Barcelona* que antecederam cada vitória de seu rival clássico *Real Madrid*, cuja avaliação das estimativas permite fazer inferências sobre o desempenho de ambas as equipes. Ressalta-se ainda que, por meio da estimação do número esperado de zeros para conjuntos que não apresentaram este valor, foi possível fazer inferências sobre o tamanho populacional, tendo-se verificado, por exemplo, no Problema 2.1 que o número de usuários de metanfetamina é bastante superior ao número daqueles que procuram por clínicas de tratamento em Bancoque. Assim, dentre as situações analisadas, que apresentaram diferentes tipos de zero-modificação (incluindo o caso de zero-inflação), concluiu-se o bom desempenho do método e sua capacidade de estimar de forma satisfatória a quantidade de zeros e sua probabilidade de ocorrência, além das demais quantidades de interesse.

Como trabalhos futuros, é possível considerar a aplicação e extensão do método à distribuições que além do parâmetro μ apresentam o parâmetro de dispersão ϕ (como exemplo, a Binomial Negativa e a Poisson Generalizada); e também a realização um estudo aplicando o procedimento bayesiano para estimação das quantidades de interesse também no contexto de dados aumentados.

REFERÊNCIAS

ARRABAL, C. T.; SILVA, K. Paula dos S.; BANDEIRA, L. Zero-truncated negative binomial applied to nonlinear data. **JP Journal of Biostatistics**, v. 11, p. 55–67, 06 2014. Citado na página 54.

BAILEY, B. J. R. A model for function word counts. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, [Wiley, Royal Statistical Society], v. 39, n. 1, p. 107–114, 1990. ISSN 00359254, 14679876. Disponível em: <http://www.jstor.org/stable/2347816>. Citado nas páginas 115 e 116.

barcelonas.com. **El Clásico: History of Rivalry between FC Barcelona and Real Madrid CF**. 2019. Disponível em: <https://www.barcelonas.com/el-clasico.html>. Citado na página 113.

BAYARRI, M. J.; BERGER, J. O.; DATTA, G. S. Objective Bayes testing of Poisson versus inflated Poisson models. In: **Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh**. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2008. p. 105–121. ISBN 978-0-940600-75-1. Disponível em: <http://projecteuclid.org/euclid.imsc/1209398464>. Citado na página 112.

BÖHNING, D.; HEIJDEN, P. G. M. van der. A covariate adjustment for zero-truncated approaches to estimating the size of hidden and elusive populations. **The Annals of Applied Statistics**, Institute of Mathematical Statistics, v. 3, n. 2, p. 595–610, 6 2009. ISSN 1932-6157. Disponível em: <http://projecteuclid.org/euclid.aoas/1245676187>. Citado nas páginas 29 e 120.

BOLFARINE, H.; BUSSAB, W. d. O. **Elementos de amostragem**. [S.l.]: Edgard Blücher, 2005. 274 p. ISBN 8521203675. Citado na página 27.

CAMERON, A. C.; TRIVEDI, P. K. **Regression analysis of count data**. Cambridge University Press, 1998. 411 p. ISBN 0521635675. Disponível em: <http://cameron.econ.ucdavis.edu/racd/count.html>. Citado na página 27.

CARVALHO, S. O. d. **Distribuições k-modificadas da família série de potência uniparamétrica**. Tese (Doutorado) — Universidade de São Paulo; Universidade Federal de São Carlos, São Carlos, 9 2017. Disponível em: <http://www.teses.usp.br/teses/disponiveis/104/104131/tde-11092017-143703/>. Citado na página 26.

CHIN, H. C.; QUDDUS, M. A. Modeling Count Data with Excess Zeroes. **Sociological Methods & Research**, SAGE Publications, v. 32, n. 1, p. 90–116, 8 2003. ISSN 0049-1241. Disponível em: <http://journals.sagepub.com/doi/10.1177/0049124103253459>. Citado na página 27.

CHOWDHURY, R. I.; ISLAM, M. A.; CHOWDHURY, R. I.; ISLAM, M. A. Zero Truncated Bivariate Poisson Model: Marginal-Conditional Modeling Approach with an Application to Traffic Accident Data. **Applied Mathematics**, Scientific Research Publishing, v. 07, n. 14, p. 1589–1598, aug 2016. ISSN 2152-7385. Disponível em: <http://www.scirp.org/journal/doi.aspx?DOI=10.4236/am.2016.714137>. Citado na página 54.

COLY, S.; YAO, A.-F.; ABRIAL, D.; GARRIDO, M. Distributions to model overdispersed count data. 01 2016. Citado na página 27.

CONCEIÇÃO, K.; TOMAZELLA, V.; ANDRADE, M.; LOUZADA, F. Biparametric zero-modified power series distributions: Bayesian analysis under a reference prior approach. **Communications in Statistics - Theory and Methods**, v. 46, 10 2016. Citado nas páginas 113, 114, 115 e 130.

CONCEIÇÃO, K. S. **Modelos Série de Potência Zero-Modificados**. 134 p. Tese (Doutorado) — UFSCar, 2013. Citado nas páginas 26, 30, 31, 34, 35, 36, 37, 38, 41, 43, 44, 45, 50, 54 e 129.

CONCEIÇÃO, K. S.; LOUZADA, F.; ANDRADE, M. G.; HELOU, E. S. Zero-modified power series distribution and its Hurdle distribution version. **Journal of Statistical Computation and Simulation**, Taylor & Francis, v. 87, n. 9, p. 1842–1862, 6 2017. ISSN 0094-9655. Disponível em: <https://www.tandfonline.com/doi/full/10.1080/00949655.2017.1289529>. Citado nas páginas 30, 31, 35, 36, 37, 41, 45, 112 e 116.

CONIGLIANI, C.; CASTRO, J. I.; O'HAGAN, A. Bayesian assessment of goodness of fit against nonparametric alternatives. **The Canadian Journal of Statistics / La Revue Canadienne de Statistique**, [Statistical Society of Canada, Wiley], v. 28, n. 2, p. 327–342, 2000. ISSN 03195724. Disponível em: <http://www.jstor.org/stable/3315982>. Citado na página 112.

CONSUL, P. New class of location-parameter discrete probability distributions and their characterizations. **Communications in Statistics - Theory and Methods**, Marcel Dekker, Inc., v. 19, n. 12, p. 4653–4666, 1 1990. ISSN 0361-0926. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/03610929008830465>. Citado na página 25.

CONSUL, P. C.; FAMOYE, F. **Lagrangian Probability Distributions**. Boston: Birkhauser, 2006. 352 p. ISBN 10 0-8176-4365-6. Citado na página 25.

CORDEIRO, G. M.; ANDRADE, M. G.; CASTRO, M. de. Power series generalized nonlinear models. **Computational Statistics & Data Analysis**, North-Holland, v. 53, n. 4, p. 1155–1166, 2 2009. ISSN 0167-9473. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167947308004866>. Citado na página 33.

CRAGG, J. G. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. **Econometrica**, v. 39, n. 5, p. 829, 9 1971. ISSN 00129682. Disponível em: <https://www.jstor.org/stable/1909582?origin=crossref>. Citado na página 26.

DALRYMPLE, M.; HUDSON, I.; FORD, R. Finite Mixture, Zero-inflated Poisson and Hurdle models with application to SIDS. **Computational Statistics & Data Analysis**, North-Holland, v. 41, n. 3-4, p. 491–504, 1 2003. ISSN 0167-9473. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167947302001871>. Citado na página 27.

DEMPSTER, A. P.; DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the EM algorithm. **JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B**, v. 39, n. 1, p. 1—-38, 1977. Disponível em: <http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.4884>. Citado na página 56.

DIETZ, E.; BÖHNING, D. On estimation of the Poisson parameter in zero-modified Poisson models. **Computational Statistics & Data Analysis**, North-Holland, v. 34, n. 4, p. 441–459, 10 2000. ISSN 0167-9473. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167947399001115>. Citado nas páginas 26 e 27.

EFRON, B. Bootstrap Methods: Another Look at the Jackknife. **The Annals of Statistics**, Institute of Mathematical Statistics, v. 7, n. 1, p. 1–26, jan 1979. ISSN 0090-5364. Disponível em: <http://projecteuclid.org/euclid.aos/1176344552>. Citado na página 60.

____. More efficient bootstrap computations. **Journal of the American Statistical Association**, v. 85, n. 409, p. 79–89, 1990. ISSN 01621459. Disponível em: <http://www.jstor.org/stable/2289528>. Citado na página 60.

FROME, E. L. The Analysis of Rates Using Poisson Regression Models. **Biometrics**, International Biometric Society, v. 39, n. 3, p. 665, 9 1983. ISSN 0006341X. Disponível em: <https://www.jstor.org/stable/2531094?origin=crossref>. Citado na página 25.

GARDNER, W.; MULVEY, E. P.; SHAW, E. C. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. **Psychological Bulletin**, v. 118, n. 3, p. 392–404, 11 1995. ISSN 1939-1455. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/7501743http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.118.3.392>. Citado na página 25.

GROGGER, J. T.; CARSON, R. T. Models for truncated counts. **Journal of Applied Econometrics**, Wiley, v. 6, n. 3, p. 225–238, 1991. ISSN 08837252, 10991255. Disponível em: <http://www.jstor.org/stable/2096628>. Citado na página 54.

GUPTA, R. C. **Modified Power Series Distribution and Some of Its Applications**. Indian Statistical Institute, 1974. 288–298 p. Disponível em: <https://www.jstor.org/stable/25051912>. Citado nas páginas 25, 33 e 34.

GURMU, S.; TRIVEDI, P. K. Excess Zeros in Count Models for Recreational Trips. **Journal of Business & Economic Statistics**, v. 14, n. 4, p. 469, 10 1996. ISSN 07350015. Disponível em: <https://www.jstor.org/stable/1392255?origin=crossref>. Citado na página 27.

HEIJDEN, P. G. M. van der; CRUYFF, M.; BÖHNING, D. Capture Recapture to Estimate Criminal Populations. In: **Encyclopedia of Criminology and Criminal Justice**. New York, NY: Springer New York, 2014. p. 267–276. Disponível em: <http://link.springer.com/10.1007/978-1-4614-5690-2_662>. Citado na página 29.

HEIJDEN, P. G. van der; BUSTAMI, R.; CRUYFF, M. J.; ENGBERSEN, G.; HOUWELINGEN, H. C. van. Point and interval estimation of the population size using the truncated Poisson regression model. **Statistical Modelling: An International Journal**, v. 3, n. 4, p. 305–322, 12 2003. ISSN 1471-082X. Disponível em: <http://journals.sagepub.com/doi/10.1191/1471082X03st057oa>. Citado na página 29.

HPC-STI-USP. **HPC - Universidade de São Paulo**. 2019. <https://hpc.usp.br/>. Acessado em: 2019-10-09. Citado na página 31.

HU, M.-C.; PAVLICOVA, M.; NUNES, E. V. Zero-Inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. **The American Journal of Drug and Alcohol Abuse**, v. 37, n. 5, p. 367–375, 9 2011. ISSN 0095-2990. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/21854279>. Citado nas páginas 27 e 43.

Investing.com. **USD/BRL - Dólar Americano Real Brasileiro Dados Históricos**. 2019. Disponível em: <https://br.investing.com/currencies/usd-brl-historical-data>. Citado nas páginas 117 e 118.

JIANG, Y.; HOUSE, L. A.; JIANG, Y.; HOUSE, L. A. Comparison of the Performance of Count Data Models under Different Zero-Inflation Scenarios Using Simulation Studies. Agricultural and Applied Economics Association, 7 2017. Disponível em: <https://econpapers.repec.org/paper/agsaaea17/258342.htm>. Citado na página 25.

JOHNSON, N. L.; KEMP, A. W.; KOTZ, S. **Univariate discrete distributions.** Wiley, 2005. 646 p. ISBN 9780471272465. Disponível em: <https://www.wiley.com/en-us/Univariate+Discrete+Distributions%2C+3rd+Edition-p-9780471272465>. Citado na página 33.

JOSHI, S. W. Some Recent Advances with Power Series Distributions. In: **A Modern Course on Statistical Distributions in Scientific Work**. Dordrecht: Springer Netherlands, 1975. p. 9–17. Disponível em: <http://link.springer.com/10.1007/978-94-010-1842-5_2>. Citado na página 25.

KASSAHUN, W.; NEYENS, T.; MOLENBERGHS, G.; FAES, C.; VERBEKE, G. Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. **Statistics in Medicine**, v. 33, n. 25, p. 4402–4419, 11 2014. ISSN 02776715. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/24957791>. Citado na página 43.

KELLY, P. A.; HAIDET, P. Physician overestimation of patient literacy: A potential source of health care disparities. **Patient Education and Counseling**, Elsevier, v. 66, n. 1, p. 119–122, apr 2007. ISSN 0738-3991. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0738399106003466>. Citado na página 28.

KING, G. Event Count Models for International Relations: Generalizations and Applications. **International Studies Quarterly**, Oxford University Press, v. 33, n. 2, p. 123, 6 1989. ISSN 00208833. Disponível em: <https://academic.oup.com/isq/article-lookup/doi/10.2307/2600534>. Citado na página 25.

LAMBERT, D. Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. **Technometrics**, Taylor & Francis, Ltd.American Statistical AssociationAmerican Society for Quality, v. 34, n. 1, p. 1, 2 1992. ISSN 00401706. Disponível em: <http://www.jstor.org/stable/1269547?origin=crossref>. Citado na página 26.

LAWLESS, J. F. Negative binomial and mixed poisson regression. **Canadian Journal of Statistics**, John Wiley & Sons, Ltd, v. 15, n. 3, p. 209–225, 9 1987. ISSN 03195724. Disponível em: <http://doi.wiley.com/10.2307/3314912>. Citado na página 25.

LEMONTE, A. J.; SIMAS, A. B.; CRIBARI-NETO, F. Bootstrap-based improved estimators for the two-parameter birnbaum–saunders distribution. In: . [S.l.: s.n.], 2008. Citado na página 67.

LI, A.-X.; ZHANG, K.-X.; WANG, L.-W. Zero-shot Fine-grained Classification by Deep Feature Learning with Semantics. **International Journal of Automation and Computing**, Institute of Automation, Chinese Academy of Sciences, v. 16, n. 5, p. 563–574, 10 2019. ISSN 1476-8186. Disponível em: <http://link.springer.com/10.1007/s11633-019-1177-8>. Citado nas páginas 27 e 28.

MACAULAY, T. **Macaulay's Essay on Milton**. Macmillan, 1909. Disponível em: <https://books.google.com.br/books?id=hSJbAAAAMAAJ>. Citado nas páginas 20, 115 e 116.

MCKENDRICK, A. Applications of Mathematics to Medical Problems. **Proc. Edinburg Math. Soc.**, v. 4, p. 98–103, 1926. Disponível em: <http://people.cs.vt.edu/badityap/classes/cs6604-Fall13/readings/kendrick-1926.pdf>. Citado na página 26.

MCLACHLAN, G.; KRISHNAN, T. **The EM algorithm and extensions**. 2. ed. ed. Hoboken, NJ: Wiley, 2008. (Wiley series in probability and statistics). ISBN 978-0-471-20170-0. Disponível em: <http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+52983362X&sourceid=fbw_bibsonomy>. Citado na página 56.

MULLAHY, J. Specification and testing of some modified count data models. **Journal of Econometrics**, North-Holland, v. 33, n. 3, p. 341–365, 12 1986. ISSN 0304-4076. Disponível em: <https://www.sciencedirect.com/science/article/pii/0304407686900023>. Citado na página 27.

NIVOLIANITOU, Z.; KONSTANDINIDOU, M.; MICHALIS, C. Statistical analysis of major accidents in petrochemical industry notified to the major accident reporting system (MARS). **Journal of Hazardous Materials**, v. 137, n. 1, p. 1–7, 9 2006. ISSN 03043894. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0304389406003037>. Citado na página 126.

PAUL, S. R.; PLACKETT, R. L. Inference Sensitivity for Poisson Mixtures. **Biometrika**, v. 65, n. 3, p. 591, 12 1978. ISSN 00063444. Disponível em: <https://www.jstor.org/stable/2335911?origin=crossref>. Citado na página 25.

POISSON, S. D. **Recherches sur la probabilité des jugements en matière criminelle et en matière civile precédées des règles générales du calcul des probabilités par SD Poisson**. [S.l.]: Bachelier, 1837. Citado na página 25.

PUZA, B. D.; JOHNSON, H. L.; O'NEILL, T. J.; BARRY, S. C. Bayesian Truncated Poisson Regression with Application to Dutch Illegal Immigrant Data. **Communications in Statistics - Simulation and Computation**, Taylor & Francis Group, v. 37, n. 8, p. 1565–1577, 8 2008. ISSN 0361-0918. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/03610910802117073>. Citado na página 29.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>. Citado na página 31.

REDDY, K. G.; YARRAKULA, K. Analysis of Accidents in Chemical Process Industries in the period 1998-2015. **Internationa Journal of ChemTech Research**, v. 9, p. 15, 2016. Disponível em: <https://www.semanticscholar.org/paper/Analysis-of-Accidents-in-Chemical-Process-in-the-Reddy-Yarrakula/cec6943f673077ce2189c2c137f1063f0b9ca717>. Citado nas páginas 122, 123 e 124.

RICHARDSON, L. F. The Distribution of Wars in Time. **Journal of the Royal Statistical Society**, WileyRoyal Statistical Society, v. 107, n. 3/4, p. 242, 1944. ISSN 09528385. Disponível em: <https://www.jstor.org/stable/10.2307/2981216?origin=crossref>. Citado na página 25.

RISELY, K. Zeros Matter. **Bird Table - BTO - British Trust for Ornithology**, 2018. Disponível em: <https://www.bto.org/volunteer-surveys/gbw/publications>. Citado na página 28.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, Oxford University Press, v. 26, n. 1, p. 139–140, 1 2010. ISSN 1367-4803. Disponível em: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp616>. Citado na página 25.

ROSSI, R. M.; MARTINS, E. N. Influência das coletas sistemáticas e parciais na seleção de codornas por meio de curvas de probabilidade de postura Influence of systematic and partial collections on quail selection through curves of laying probability. n. 8, p. 1699–1707, 2010. ISSN 1806-9290. Disponível em: <www.sbz.org.br>. Citado na página 28.

SELLERS, K. F.; RAIM, A. A flexible zero-inflated model to address data dispersion. **Computational Statistics & Data Analysis**, North-Holland, v. 99, p. 68–80, 7 2016. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167947316000165>. Citado na página 42.

SHANKAR, V.; MILTON, J.; MANNERING, F. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. **Accident; analysis and prevention**, v. 29, n. 6, p. 829–37, nov 1997. ISSN 0001-4575. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/9370019>. Citado na página 28.

Swedish Chemicals Agency. **Chemical industry from an economic perspective Development trends in the world, the EU and Sweden in 2010**. Stockholm, 2011. 43 p. Disponível em: <www.kemikalieinspektionen.se>. Citado na página 122.

UMBACH, D. On inference for a mixture of a poisson and a degenerate distribution. **Communications in Statistics - Theory and Methods**, Marcel Dekker, Inc., v. 10, n. 3, p. 299–306, 1 1981. ISSN 0361-0926. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/03610928108828039>. Citado na página 26.

United Nations Office on Drugs and Crime. **Global SMART (Synthetics Monitoring: Analyses, Reporting and Trends) Programme - UPDATE REPORT 2012**. [S.l.], 2012. Disponível em: <https://www.unodc.org/documents/scientific/Global_SMART_Update_8_E_web.pdf>. Citado na página 120.

VOLO, S. A journey through tourism statistics: accuracy and comparability issues across local, regional and national levels. In: SCORUS. **SCORUS 24th Biennial Conference on Regional and Urban Statistics: Understanding Change**. [S.l.], 2004. p. 210–216. Citado na página 28.

XIE, T.; AICKIN, M. A truncated poisson regression model with applications to occurrence of adenomatous polyps. **Statistics in medicine**, v. 16, n. 16, p. 1845–57, 8 1997. ISSN 0277-6715. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed/9280037>. Citado na página 26.

ZELTERMAN, D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. **Journal of Statistical Planning and Inference**, North-Holland, v. 18, n. 2, p. 225–237, 2 1988. ISSN 0378-3758. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/0378375888900079>. Citado na página 120.

APÊNDICE

A

PROCEDIMENTO DE ESTIMAÇÃO BAYESIANO

O procedimento de estimação bayesiano permite fazer inferências sobre os parâmetros com base nas distribuições a posteriori marginais, obtidas através da integração da distribuição a posteriori conjunta.

Em um conjunto de observações em que o zero não ocorre, mas há possibilidade deste valor ocorrer no experimento em estudo, utilizar o procedimento de estimação de máxima verossimilhança diretamente para a distribuição ZMPS levaria a estimativas errôneas, visto que uma distribuição ZDPS deveria ser suposta (e não uma ZTPS), pois $P(Y=0)>0$. Por outro lado, o procedimento bayesiano de estimação consegue contornar o problema por meio do uso de informação a priori adequada, e este pode ser utilizado inclusive para comparação com os resultados obtidos via procedimentos clássicos de estimação.

Para simplificações[1], considere Y uma variável aleatória com distribuição ZMPS(μ,ϕ,ω) e vetor de observações $\boldsymbol{y}=(y_1,\ldots,y_n)$. Já é conhecido que a função de verossimilhança associada ao vetor $\boldsymbol{y}$ é dada pela equação (2.12), que pode ser reescrita utilizando-se a equação (2.14) da seguinte forma:

$$L(\mu,\phi,\omega;\mathscr{D})=\exp\{\ell_1(\mu,\phi;\mathscr{D})+\ell_2(\omega;\mathscr{D})\}. \tag{A.1}$$

Por meio da equação (A.1), fica clara a suposição razoável de independência das distribuições a priori $\pi(\mu,\phi)$ e $\pi(\omega)$ para os parâmetros (μ,ϕ) e ω, respectivamente. Assim, tem-se que a distribuição a priori conjunta é dada por:

$$\pi(\mu,\phi,\omega)=\pi(\mu,\phi)\pi(\omega).$$

[1] A parametrização da distribuição dada por ZMPS(μ,ϕ,ω) permite a suposição de independência a priori dos parâmetros μ,ϕ e ω.

A abordagem bayesiana pode então ser feita por meio da densidade a posteriori conjunta:

$$\pi(\mu,\phi,\omega|\mathscr{D}) \propto L(\mu,\phi,\omega;\mathscr{D})\pi(\mu,\phi,\omega),$$

isto é,

$$\begin{aligned}\pi(\mu,\phi,\omega|\mathscr{D}) &\propto \exp\{\ell_1(\mu,\phi;\mathscr{D})+\ell_2(\omega;\mathscr{D})\}\pi(\mu,\phi,\omega)\\ &\propto \exp\{\ell_1(\mu,\phi;\mathscr{D})\}\exp\{\ell_2(\omega;\mathscr{D})\}\pi(\mu,\phi)\pi(\omega). \qquad \text{(A.2)}\end{aligned}$$

Por sua vez, as densidades a posteriori marginais para os parâmetros são obtidas por meio da integração de (A.2):

$$\begin{aligned}\pi(\mu|\mathscr{D}) &\propto \int_\omega\int_\phi \pi(\mu,\phi,\omega|\mathscr{D})d\phi d\omega\\ &\propto \int_\phi \exp\{\ell_1(\mu,\phi;\mathscr{D})\}\pi(\mu,\phi)d\phi; \qquad \text{(A.3)}\\ \pi(\phi|\mathscr{D}) &\propto \int_\omega\int_\mu \pi(\mu,\phi,\omega|\mathscr{D})d\mu d\omega;\\ &\propto \int_\mu \exp\{\ell_1(\mu,\phi;\mathscr{D})\}\pi(\mu,\phi)d\mu; \qquad \text{(A.4)}\\ \pi(\omega|\mathscr{D}) &\propto \int_\phi\int_\mu \pi(\mu,\phi,\omega|\mathscr{D})d\mu d\phi\\ &\propto \exp\{\ell_2(\omega;\mathscr{D})\}\pi(\omega). \qquad \text{(A.5)}\end{aligned}$$

Devido à complexidade da densidade a posteriori $\pi(\mu,\phi,\omega|\mathscr{D})$, para resolver as integrais (A.3) e (A.4), pode-se utilizar técnicas de Monte Carlo com Cadeia de Markov (MCMC)[2]. Para isso, considera-se as densidades condicionais para os parâmetros μ e ϕ dadas, respectivamente, por:

$$\begin{aligned}\pi(\mu|\phi;\mathscr{D}) &\propto \exp\{\ell_1(\mu,\phi;\mathscr{D})\}\pi(\mu|\phi);\\ \pi(\phi|\mu;\mathscr{D}) &\propto \exp\{\ell_1(\mu,\phi;\mathscr{D})\}\pi(\phi|\mu).\end{aligned} \qquad \text{(A.6)}$$

Inferências sobre o parâmetro ω podem ser feitas diretamente considerando-se a densidade a posteriori marginal $\pi(\omega|\mathscr{D})$. Respeitando-se o espaço de variação do parâmetro ω, diversas prioris podem ser consideradas. Por simplificação, neste trabalho, opta-se por uma priori conjugada. Assim, ao considerar a densidade a priori de ω, $\pi(\omega)$, como uma Beta com hiperparâmetros a e b ($a,b \in (0,+\infty)$), isto é, $\omega \sim Beta(a;b)$, tem-se a densidade a posteriori dada por:

$$\pi(\omega|\mathscr{D}) \propto \omega^{n-n_0+a-1}(1-\omega)^{n_0+b-1},$$

ou seja, $\omega|\mathscr{D} \sim Beta(n-n_0+a;n_0+b)$.

[2] Outra sugestão é o uso de Monte Carlo Hamiltoniano (HMC).

Para a situação em que não há observações zero no conjunto de dados ($n_0 = 0$), ressalta-se que, ao considerar uma densidade a priori "vaga" para ω (isto é, escolhendo-se os hiperparâmetros $a = b = 1$, tal que $\mathbb{V}(\omega) \to \infty$) e supor uma função de perda quadrática, o estimador bayesiano que consiste na média a posteriori será superestimado, aproximando-se de 1 à medida que n cresce. Isso implica na subestimação da probabilidade de zero, que será aproximadamente zero. Ao considerar esta abordagem, essa subestimação da probabilidade de zero pode ser contornada com o uso de prioris informativas, que podem ser obtidas, por exemplo, com base no conhecimento de um especialista.


ICMC
USP
SÃO CARLOS