Análise empírica de dados multinomiais

Renata Pelissari

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura:_____

Análise empírica de dados multinomiais [1]

Renata Pelissari

Orientador: ***Prof. Dr. Dorival Leão P. Júnior***

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional.

USP – São Carlos
Julho/2009

[1] Este trabalho contou com o apoio financeiro do CNPq e da Fapesp.

Aos meus pais,

Antonio e Sandra

com amor ...

Agradecimentos

Agradeço a Deus por me capacitar e me dar força para concluir mais essa etapa na minha vida.

Aos meus pais, Antonio Pelissari e Sandra Barbosa Pelissari, pelo esforço, dedicação, apoio incondicional e educação. Sempre priorizaram meus estudos, independente das circunstâncias, e graças a eles cheguei até aqui. À minha irmã Fernanda e ao meu cunhado Rafael, pela ajuda em todos as áreas da minha vida. Amo vocês!

Ao Prof. Dorival Leão, pela orientação deste trabalho, pela amizade e disponibilidade, sempre quando precisei. Agradeço também pelo conhecimento que adquiri com você e pela sua influência intensa e positiva na minha vida profissional.

Ao meu namorado Kleber, pelo incentivo, carinho, compreensão e atenção, e por sempre ter uma palavra de conforto e estímulo nos momentos que mais precisava. Obrigada também por me ajudar nas programações e nas correções.

Aos meus professores de graduação, especialmente à Prof. Sandra Godoy pela orientação de iniciação científica durante minha graduação e pela amizade.

Aos meus amigos de longe e de perto, novos ou antigos, que de alguma forma contribuíram com este trabalho.

Aos professores de estatística do ICMC, Mariana Cúri, Reiko Aoki, Mário de Castro, Marinho Andrade, Vicente Garibay, pela atenção e ajuda durante o mestrado. Aos funcionários do ICMC pela dedicação.

À Fapesp e ao CNPq pelo apoio financeiro durante o desenvolvimento deste trabalho.

Muito obrigada!

Resumo

Em diversas análises estatísticas, nos deparamos com dados multinomiais, dos quais precisamos analisar o comportamento ao longo do tempo e sua relação com fatores determinantes. Os métodos clássicos para modelos de regressão multinomiais consistem em utilizar a estrutura de modelos lineares generalizados para desenvolver tais modelos McCullagh & Nelder (1989). No entanto, este enfoque apresenta algumas desvantagens como não admiter a incidência de zeros em nenhuma categoria, a hipótese da proporcionalidade da razão de chances e o fato de não serem modelos adequados para análise de dados censurados. Com o objetivo de analisar dados multinomiais com essas características propomos um modelo que é uma extensão do modelo de intensidade multiplicativo desenvolvido por Aalen (1978) e apresentado em Fleming & Harrington (2005), para variáveis aleatórias multinomiais. Com isso, ao invés de modelarmos as probabilidades associadas às categorias, como nos métodos clássicos, modelamos a função intensidade associada à variável aleatória multinomial. Através do critério martingale, estimamos os parâmetros do modelo ajustado e propomos testes de hipóteses para estes parâmetros para uma e duas populações. O teste para comparação de duas populações é baseado na estatística de *logrank*.

Abstract

In several applications, we want to analyze the behavior of multinomial datas over the time and its relationship with important factors. The classic methods commonly used for multinomial regression models are based in the generalized linear model framework. However, this models presents some disadvantages such that: it does not admit the incidence of zeros in any category, the assumption of proportionality of odds ratio and the fact that they are not appropriate models to analyze censored data. For multinomial data analyses with this characteristics, we propose a model that it is an extension of the multiplicative intensity model developed by Aalen to random multinomial variables. Therefore, instead of modeling the categorical probabilities, as in the classics methods, we modeled the intensity fuction associated with the multinomial variable. Using the martingale criterion, we estimate the models parameters and propose hypothesis testing for these parameters for one and two populations. The test for comparing two populations is based in the logrank statistics.

Sumário

1 Introdução **1**

1.1 Motivação 2

1.2 Estrutura da Dissertação 4

2 Modelo Multinomial **5**

2.1 Introdução 5

2.2 Espaço de Cantor 5

2.3 Variável Aleatória 10

2.4 Produto Finito do Espaço Mensurável $(S^{\infty}, \mathcal{A})$ 12

2.5 Probabilidade Condicional 13

2.6 Cálculo Estocástico 15

2.6.1 Martingale Array Difference 19

2.7 Integral Estocástica 19

2.8 Decomposição de Doob-Meyer 21

2.8.1 Processo Variação Quadrática 23

3 Modelo de Intensidade Multiplicativo **27**

3.1 Introdução 27

3.2 Desenvolvimento do Modelo 27

3.3 Inferência no Modelo de Intensidade Multiplicativo 33

3.3.1 Propriedades dos estimadores para amostras finitas 37

3.3.2 Propriedades assintóticas 42

4 Distribuição Assintótica de Integrais Estocásticas **45**

4.1 Amostra Única 45

4.2 Duas Amostras 52

5 Testes de Hipóteses **63**

5.1 Introdução 63

5.2 Teste de Hipóteses para uma População 63

5.3 Teste de Hipóteses para Duas amostras 65

6 Estudo de simulação **69**

6.1 Introdução 69

6.2 Modelo na ausência de censura 69

6.2.1 Avaliação da qualidade dos estimadores 71

6.2.2 Teste de hipótese 75

6.3 Análise na presença de censura 78

6.3.1 Avaliação da qualidade dos estimadores 80

6.3.2 Teste de hipótese 83

6.4 Conclusão 86

7 Aplicação **87**

7.1 Introdução 87

7.2 Análise Descritiva da Evasão dos Alunos de Graduação da USP 87

7.3 Aplicação do Modelo de Intensidade Multiplicativo 95

7.3.1 Comparação entre anos 96

7.3.2 Comparação entre sexos 104

7.3.3 Comparação entre períodos 106

7.3.4 Comparação entre áreas 108

7.3.5 Portal Evasão 112

7.4 Conclusão 114

8 Conclusão e Trabalhos Futuros **117**

A Espaço de Cantor **119**

Referências Bibliográficas **122**

Capítulo 1

Introdução

Em diversas situações práticas temos o interesse de analisar o comportamento de dados multinomiais (*polytomous ou polytochomous*) ao longo do tempo e sua relação com fatores determinantes. Uma variável aleatória é multinomial, conforme McCullagh & Nelder (1989), se esta assume um número finito de valores, aos quais nos referimos como categorias ou classes.

Modelos de regressão multinomial são utilizados para identificar possíveis relações entre a variável resposta multinomial e os fatores associados. Métodos clássicos para modelos de regressão multinomial, como podemos encontrar em McCullagh & Nelder (1989) e Hosmer & Lemeshow (1989), consistem em utilizar a estrutura de modelos lineares generalizados para desenvolver modelos de regressão multinomiais. Dependendo da classificação dada à variável multinomial (intervalar, nominal ou ordinal) são propostas diferentes funções de ligação e estratégias para a locação de efeitos.

No entanto, este enfoque apresenta algumas desvantagens, sendo uma delas o fato de que o modelo logístico multinomial não admite a incidência de zeros em nenhuma categoria. Uma possível solução é a forma diferenciada de se lidar com o modelo multinomial proposta por Begg & Gray (1984) e estendida por Rom & Cohen (1995) que consiste em ajustar diversos modelos binários, ou seja, os parâmetros de uma regressão logísitca multinomial são estimados utilizando regressões binárias.

Outra hipótese importante assumida na modelagem clássica é a proporcionalidade da razão de chances (*odds ratio*). Contudo, em muitos conjuntos de dados esta hipótese é inadequada, o que nos leva a desenvolver métodos apropriados para tratar o caso em que a razão de chances

não são proporcionais. Outra desvantagen dos métodos clássicos de regressão multinomial importante neste estudo é o fato de que esses modelos não são adequados para análise de dados censurados.

Com o objetivo de analisar dados multinomiais com as características descritas acima, categorias nulas, razão de chances não proporcionais e dados censurados, propomos um modelo que é uma extensão do modelo de intensidade multiplicativo desenvolvido por Aalen (1978) e apresentado em Fleming & Harrington (2005), para variáveis aleatórias multinomiais. O modelo sugerido é também adequado para análise de dados multinomiais que atendem às hipóteses necessárias no modelo de regressão multinomial clássico.

Estimamos os parâmetros do modelo utilizando o critério de martingale e desenvolvemos um estudo das propriedades assintóticas e de amostras finitas. Os estimadores encontrados correspondem aos estimadores empíricos da função de distribuição acumulada, que coincide com o estimador de *Kaplan-Meier*. Propomos testes de hipóteses para comparação dos parâmetros do modelo por meio da teoria assintótica desenvolvida. Realizamos também estudos de simulação para avaliar as propriedades dos estimadores e a qualidade dos testes de hipóteses. Finalmente, aplicamos o modelo proposto para analisar o comportamento da evasão dos alunos de graduação da Universidade de São Paulo.

1.1 Motivação

Utilizamos o modelo proposto neste trabalho para analisar o conjunto de dados reais referente a alunos de graduação da Universidade de São Paulo. O interesse é estudar o tempo até a evasão dos alunos de graduação. Para isso, consideremos o conjunto de dados da graduação da USP (Fonte: JúpiterWeb) referente aos alunos ingressantes por vestibular entre 1999 e 2008.

Neste conjunto de dados temos informações sobre o aluno ingressante e sobre o curso no qual ele está matriculado. A variável de interesse é a que determina a última ocorrência do aluno até o fechamento do segundo semestre de 2008, que é a data do fim do estudo. A última ocorrência pode ser dividida em 5 classes: conclusão, evasão, em andamento, falecimento, migração. Na Tabela 1.1 temos a quantidade de alunos em cada uma dessas classes por ano e o número de ingressantes em cada ano.

Tabela 1.1: *Quantidade de alunos em cada uma das classes de última ocorrência (em andamento, conclusão, evasão, falecimento e migração) por ano e número de ingressantes em cada ano.*

	Última ocorrência					Número de
Ano de ingresso	Em andamento	Conclusão	Evasão	Falecimento	Migração	ingressantes
1999	126	5096	1560	6	282	7070
2000	255	5148	1481	12	311	7207
2001	422	5445	1248	7	336	7458
2002	921	5465	1278	11	318	7993
2003	1900	5055	1179	9	381	8524
2004	3932	3343	1102	5	388	8770
2005	6970	1205	1180	6	416	9777
2006	8848	13	877	6	425	10169
2007	9169	2	671	4	343	10189
2008	9657	1	485	1	169	10313

O termo evasão foi definido pela Pró-Reitoria de graduação da USP e se refere ao aluno que se desligou da USP a seu pedido ou por motivos administrativos (3 semestres sem matrícula, jubilamento, transferência externa, outros). Migração se refere ao aluno que efetuou transferência interna e o termo em adamento se refere ao aluno ativo no fim do estudo.

O objetivo deste trabalho é analisar o número de semestres até a ocorrência de evasão. Para isso, definimos que o início da contagem do número de semestres até a evasão será sempre o ano em que o aluno ingressou na USP e, a data limite, quando o estudo se encerra, é o 2° semestre de 2008.

O evento de interesse é a evasão. Os eventos migração, transferência, falecimento e conclusão, correspondem aos eventos de censura. A cada aluno associamos um variável aleatória multinomial que determina o número de semestres até a última ocorrência. Iremos então avaliar o comportamento desta variável aleatória multinomial ao longo do tempo de forma a estudar o tempo até a evasão e sua relação com atributos associados ao aluno, como sexo, período e área.

1.2 Estrutura da Dissertação

A presente dissertação de mestrado está dividida em oito capítulos. No segundo capítulo, contruímos um modelo probabilístico para acomodar uma variável aleatória multinomial e apresentamos definições e resultados de probabilidade, esperança condicional, cálculo estocástico, integral estocástica e decomposição de Doob-Meyer, que serão utilizados nos capítulos posteriores.

No terceiro capítulo, apresentamos o modelo de intensidade multiplicativo proposto neste trabalho, obtemos os estimadores dos parâmetros desse modelo e estudamos suas propriedades de amostras finitas e assintóticas. O quarto capítulo contém um estudo da distribuição assintótica de integrais estocásticas. Este estudo foi de fundamental importância para o desenvolvimento de testes de hipóteses para comparações dos parâmetros do modelo, obtidos no quinto capítulo.

No sexto capítulo, avaliamos numericamente a qualidade das aproximações utilizadas nas inferências sobre os parâmetros do modelo de intensidade multiplicativo por meio um estudo de simulação. O interesse nesse capítulo é comprovar, através do estudo de simulação, as proriedades dos estimadores provadas no capítulo anterior, além de avaliar a qualidade dos testes de hipóteses.

No sétimo capítulo, analisamos o conjunto de dados de alunos de graduação da Universidade de São Paulo utilizando o modelo de intensidade multiplicativo. Finalmente, o oitavo capítulo apresenta conclusões dos resultados obtidos nesse trabalho e propostas de trabalhos futuros.

Capítulo 2

Modelo Multinomial

2.1 Introdução

Se a resposta associada a um indivíduo ou ítem assume um número finito de valores, temos que a resposta é multinomial (McCullagh & Nelder (1989)). Os valores que a resposta assume são denominados categorias ou intervalos de classe. As técnicas utilizadas para analisar dados multinomiais dependem do tipo de dado analisado. Os principais tipos de dados multinomiais são: nominal, ordinal e intervalar. O modelo aqui desenvolvido pode ser aplicado para qualquer um destes tipos.

A seguir será contruido um modelo probabilístico para acomodar uma variável aleatória multinomial.

2.2 Espaço de Cantor

Consideremos η uma função de distribuição de probabilidade tomando valores em $\{1, 2, \ldots, k\}$, para $k \in \mathbb{N}$, tal que

- $\eta(i) \geq 0, \ \forall i = 1, \ldots, k;$
- $\eta(1) + \ldots + \eta(k) = 1.$

Na seqüência, construímos uma variável aleatória multinomial a partir da distribuição η. Seja $S = \{0, 1\}$ o espaço formado por dois elementos, 0 e 1, e

$$S^\infty = S \times S \times \ldots.$$

O espaço S^∞ é conhecido como espaço de Cantor. Uma descrição detalhada deste espaço pode ser encontrada em Leao *et al.* (2005). Seja $\pi_k : S^\infty \to S$ a projeção na k-ésima coordenada ($k \in \mathbb{N}$). A partir da projeção coordenada, defimos a classe de cilindros de S^∞ com base em S

$$\mathcal{H} = \{\pi_k^{-1}(\{\omega_k\}) : \omega_k \in S \ \text{ e } \ k \in \mathbb{N}\}.$$

Da definição de $\mathcal{H}$ obtemos os seguintes resultados:

1. $\pi_k^{-1}(\{0\}) \cap \pi_k^{-1}(\{1\}) = \emptyset, \ \forall k \in \mathbb{N}$.

2. $\pi_k^{-1}(\{0\}) \cup \pi_k^{-1}(\{1\}) = S^\infty, \ \forall k \in \mathbb{N}$.

Na classe de cilindros de dimensão 1, $\mathcal{H}$, definimos a função de probabilidade $P : \mathcal{H} \to [0, 1]$ como

$$P[\pi_k^{-1}(\{\omega_k\})] = \frac{1}{2}.$$

A partir de $\mathcal{H}$, construímos a classe Δ

$$\begin{aligned}\Delta &= \{\pi_{v_1}^{-1}(\{\omega_{v_1}\}) \cap \ldots \cap \pi_{v_n}^{-1}(\{\omega_{v_n}\}) : (\omega_{v_1}, \ldots, \omega_{v_n}) \in S^n, \ v_i \in \mathbb{N}, \ i = 1, \ldots, n \text{ e } n \in \mathbb{N}\} \cup \{\emptyset, S^\infty\} \\ &= \{D \subset S^\infty : D = B_1 \cap B_2 \cap \ldots \cap B_n, \quad B_i \in \mathcal{H}, \quad i = 1, \ldots, n, \quad n \in \mathbb{N}; \ \emptyset, \ S^\infty\}.\end{aligned}$$

Lema 2.1. *A classe Δ é uma semi-álgebra, ou seja,*

(i) $\emptyset$ e $S^\infty \in \Delta$;

(ii) Δ é fechada pela interseção finita;

(iii) Se $C \in \Delta$, então $C^C \notin \Delta$. Porém,

$$C^c = D_1 \cup D_2 \cup \ldots \cup D_n, \quad D_i \cap D_l = \emptyset, \quad i \neq l \ \text{ e } D_i \in \Delta.$$

Demonstração. (i) Definição.

(ii) Sejam $\Delta_1, \Delta_2, \ldots, \Delta_n \in \Delta$ tal que

$$\begin{aligned}
\Delta_1 &= B_1^1 \cap B_2^1 \cap \ldots \cap B_{k_1}^1 \\
\Delta_2 &= B_1^2 \cap B_2^2 \cap \ldots \cap B_{k_2}^2 \\
&\vdots \\
\Delta_n &= B_1^n \cap B_2^n \cap \ldots \cap B_{k_n}^n, \quad B_l^i \in \mathcal{H}, \quad l = 1, \ldots, k_i, \quad i = 1, \ldots, n, \quad k_i, n \in \mathbb{N}.
\end{aligned}$$

Assim,

$$\bigcap_{i=1}^{n} \Delta_i = \bigcap_{i=1}^{n} \bigcap_{l=1}^{k_i} B_l^i.$$

Portanto, $\bigcap_{i=1}^{n} \Delta_i$ pode ser escrita como interseção finita de elementos de $\mathcal{H}$. Logo, $\bigcap_{i=1}^{n} \Delta_i \in \Delta$.

(iii) Consideremos $C \in \Delta$. Então,

$$C = B_1 \cap B_2 \cap \ldots \cap B_n, \quad B_i \in \mathcal{H}, \quad i = 1, \ldots, n, \quad n \in \mathbb{N}.$$

Assim,

$$C^c = S^\infty - \{B_1 \cap B_2 \cap \ldots \cap B_n\}.$$

Como $B_i \in \mathcal{H}$, para determinado $\omega_{k_i} \in S, \quad k_i \in \mathbb{N}, \quad i = 1, \ldots, n$ e $n \in \mathbb{N}$, obtemos

$$B_i = \pi_{k_i}^{-1}(\{\omega_{k_i}\}).$$

Portanto,

$$\begin{aligned}
C^c &= S^\infty - \{\pi_{k_1}^{-1}(\{\omega_{k_1}\}) \cap \ldots \pi_{k_n}^{-1}(\{\omega_{k_n}\})\} \\
&= \{\mathbf{x} = (x_n)_n \in S^\infty : \exists j, \quad j \in \{1, \ldots, n\}, \text{ tal que } x_{k_j} \neq \omega_{k_j}\}
\end{aligned}$$

Consideremos agora o conjunto

$$\Gamma = S^n - \{(\omega_{k_1}, \omega_{k_2}, \ldots, \omega_{k_n})\}.$$

Daí temos que

$$C^c = \bigcup_{\mathfrak{H} \in \Gamma} \{\mathbf{x} = (x_n)_n \in S^\infty : \pi_{k_i}(\mathbf{x}) = \mathfrak{H}_i, \quad i = 1, \ldots, n, \quad n \in \mathbb{N}\}.$$

Portanto, C^c é união finita e disjunta de elementos de Δ.

$\square$

Podemos estender a probabilidade P para a semi-álgebra de cilindros Δ, $P : \Delta \to [0,1]$ de tal forma que $P(\emptyset) = 0$, $P(S^\infty) = 1$ e

$$P(B) \;=\; \prod_{i=1}^{j} P[\pi_{k_i}^{-1}(\{\omega_{k_i}\})] = \left(\frac{1}{2}\right)^j,$$

para todo $(\omega_{v_1}, \ldots, \omega_{v_j}) \in S^j$ e $C \in \Delta$.

Para obtermos uma álgebra gerada pelos cilindros, basta aplicar as operações de conjunto definidas no item (iii) do Lema 2.1 aos elementos de Δ. Assim, denotamos por $\mathcal{A}$ a classe formada pela união finita disjunta de elementos de Δ,

$$\mathcal{A} = \{A \in S^\infty : A = \bigcup_{i=1}^{n} C_i, \quad C_i \in \Delta, \quad C_k \cap C_j = \emptyset, \quad k, j = 1, \ldots, n, \quad k \neq j\}.$$

Lema 2.2. *A classe $\mathcal{A}$ é a álgebra gerada por Δ, ou seja,*

(i) $S^\infty, \emptyset \in \mathcal{A}$.

(ii) A classe $\mathcal{A}$ é fechada pela complementação.

(iii) A classe $\mathcal{A}$ é fechada pela interseção finita.

(iv) Toda álgebra $\mathcal{B}$ de conjuntos de S^∞ que contém Δ também contém $\mathcal{A}$.

Demonstração. (i) $S^\infty, \emptyset \in \mathcal{A}$, pois $\Delta \in \mathcal{A}$.

(ii) Mostremos que se $A \in \mathcal{A}$ então $A^c \in \mathcal{A}$.

Considere $A \in \mathcal{A}$. Assim,

$$A = \bigcup_{i=1}^{n} C_i, \quad C_i \in \Delta, \quad C_k \cap C_j = \emptyset, \quad k, j = 1, \ldots, n, \quad k \neq j.$$

Conseqüentemente,

$$A^c = \bigcap_{i=1}^{n} C_i^c.$$

Pelo item (iii) do Lema 2.1 temos que

$$C_i^c = \bigcup_{j_i=1}^{m_i} D_{j_i}^i, \; D_{j_i}^i \in \Delta, \; D_l^i \cap D_s^i = \emptyset, \; l \neq s, \; j_i = 1, \ldots, m_i, \; m_i \in \mathbb{N}, \; i = 1, \ldots, n.$$

Portanto,

$$\begin{aligned} A^c &= \bigcap_{i=1}^{n}\left(\bigcup_{j_i=1}^{m_i} D_j^i\right) = \{\bigcup_{j_1=1}^{m_1} D_{j_1}^1\} \cap \ldots \cap \{\bigcup_{j_n=1}^{m_n} D_{j_n}^n\} \\ &= \bigcup_{\{j_1,\ldots,j_n\}\in M} (D_{j_1}^1 \cap \ldots \cap D_{j_n}^n) \end{aligned}$$

sendo $M = \{1, \ldots, m_1\} \times \ldots \times \{1, \ldots, m_n\}$.

Observe que os conjuntos

$$D_{j_1}^1 \cap \ldots \cap D_{j_n}^n,$$

com $(j_1, j_2, \ldots, j_n) \in M$, são disjuntos. Além disso, como Δ é fechada pela interseção, obtemos que

$$D_{j_1}^1 \cap \ldots \cap D_{j_n}^n \in \Delta$$

para $(j_1, \ldots, j_n) \in M$.

Assim, temos que A^c é união finta disjunta de elementos de Δ, ou seja, $A^c \in \mathcal{A}$.

(iii) Sejam $A, B \in \mathcal{A}$. Então,

$$A = \bigcup_{i=1}^{n_A} C_i^A, \qquad B = \bigcup_{j=1}^{n_B} C_i^C$$

com $C_i^A \in \Delta$, $C_j^B \in \Delta$, $C_s^L \cap C_l^L = \emptyset$ para $s \neq l$, $s, l \in \{1, \ldots, n_L\}$, $L = A, B$.

Assim temos que

$$A \cap B = \left[\bigcup_{i=1}^{n_A} C_i^A\right] \cap \left[\bigcup_{j=1}^{n_B} C_j^B\right] = \bigcup_{i=1}^{n_A}\bigcup_{j=1}^{n_B} \left[C_i^A \cap C_j^B\right].$$

Note que $[C_{i_1}^A \cap C_{j_1}^B] \cap [C_{i_2}^A \cap C_{j_2}^B] = \emptyset$, se $i_1 \neq i_2$ ou $j_1 \neq j_2$. Além disso, Δ é fechada pela interseção finita, portanto, $A \cap B$ é união finita disjunta de elementos de Δ. Logo, $A \cap B \in \mathcal{A}$. Portanto, $\mathcal{A}$ é fechada pela interseção finita.

□

Na seqüência construímos uma distribuição de probabilidade sobre a álgebra gerada pelos cilindros. Consideremos a probabilidade definida em 2.1, $P : \Delta \to [0, 1]$ tal que

$$P[\pi_{v_1}^{-1}(\{\omega_{v_1}\}) \cap \ldots \cap \pi_{v_j}^{-1}(\{\omega_{v_j}\})] = \left(\frac{1}{2}\right)^j,$$

Assim, tomando $C \in \mathcal{A}$, sabemos que existe uma coleção finita $\{D_i\}_{i=1}^n \subset \Delta$ tal que

(i) $D_i \cap D_l = \emptyset,\ i \neq l$

(ii) $\bigcup_{i=1}^{n} D_i = C$.

Desta forma, a função de conjunto P é estendida sobre a álgebra de cilindros $\mathcal{A}$ por

$$P(C) = \sum_{i=1}^{n} P(D_i). \tag{2.1}$$

Observemos que a função P está bem definida. De fato, consideremos $\sum_{i=1}^{n} D_i = \sum_{j=1}^{m} B_j$ para $m, n \in \mathbb{N}$ e $\{D_i\}_{i=1}^{n}, \{B_i\}_{j=1}^{m} \in \Delta$. Então,

$$\sum_{i=1}^{n} P(D_i) = \sum_{i=1}^{n}\sum_{j=1}^{m} P[D_i \cap B_j] = \sum_{j=1}^{m}\sum_{i=1}^{n} P[D_i \cap B_j] = \sum_{j=1}^{m} B_j.$$

Finalmente, via o Teorema de extensão de Caratheodory, obtemos uma única probabilidade sobre a $\sigma-$álgebra $\mathcal{F}$ gerada pela álgebra $\mathcal{A}$. Como as variáveis aleatórias que serão estudadas nesta dissertação são multinomiais, todas as análises realizadas na seqüência serão desenvolvidas sobre a álgebra $\mathcal{A}$. Com isso, denotamos por $(S^{\infty}, \mathcal{A}, P)$ nosso espaço de probabilidade.

2.3 Variável Aleatória

Nesta seção construímos uma variável aleatória associada à distribuição de probabilidade η. Primeiramente, definimos que uma partição $\mathcal{K}$ do S^{∞} é uma classe finita de conjuntos $K_1, \ldots, K_n$ satisfazendo

(i) $K_i \in \mathcal{A}, \quad i = 1, \ldots, n$;

(ii) $\cup_{i=1}^{n} K_i = S^{\infty}$;

(iii) $K_j \cap K_l = \emptyset, \quad j, l = 1, \ldots, n$ e $j \neq l$.

Denotamos por $S(\mathcal{A})$ conjunto de todas as funções $X : S^{\infty} \to \mathbb{R}$ tais que

$$X = \sum_{i=1}^{n} c_i \mathbb{1}_{\{K_i\}}$$

com $K_1, \ldots, K_n$ uma partição do S^{∞} e $c_i \in \mathbb{R}, \quad i = 1, \ldots, n$. Os elementos de $S(\mathcal{A})$ são denominados variáveis aleatórias (ou funções mensuráveis).

Teorema 2.1. *Consideremos o espaço de probabilidade* $(S^\infty, \mathcal{A}, P)$.

(i) Dada uma variável aleatória X *assumindo valores no espaço finito* $\{c_1, \ldots, c_n : c_i \in \mathbb{R}$ *e* $n \in \mathbb{N}\}$ *existe uma função* η *associada à* X *tal que*

$$P[\omega : X(\omega) = c_i] = \eta(i) = \frac{j_i}{2^j},$$

para $j_i \in \{0, 1, \ldots, 2^j\}$ *e* $j \in \mathbb{N}$.

(ii) Dada uma distribuição de probabilidade discreta η *sobre o espaço finito* $\{c_1, \ldots, c_n : c_i \in \mathbb{R}$ *e* $n \in \mathbb{N}\}$ *satisfazendo* $\eta(i) = \dfrac{j_i}{2^j}$ *para* $j_i \in \{0, 1, \ldots, 2^j\}$ *e* $j \in \mathbb{N}$, *existe uma variável aleatória* $X : S^\infty \to \mathbb{R}$ *satisfazedo*

$$P[\omega : X(\omega) = c_i] = \eta(i)$$

Demonstração. Consideremos $\{B_1, B_2, \ldots, B_{2^j}\}$ uma enumeração qualquer tal que

$$\Delta_j = \{\pi_1^{-1}(\{\omega_1\}) \cap \ldots \cap \pi_1^{-1}(\{\omega_j\}) : (\omega_1, \omega_2, \ldots, \omega_J) \in S^j\} = \{B_1, B_2, \ldots, B_{2^j}\}.$$

Tome,

$$\begin{aligned} A_1 &= B_1 \cup \cdots \cup B_{j_1} \\ &\vdots \\ A_i &= B_{j_1+\cdots+j_{i-1}} \cup \cdots \cup B_{j_1+\cdots+j_i}, \end{aligned}$$

para $i = 1, 2, \ldots, n$. Assim, obtemos uma variável aleatória X

$$X(\omega) = \sum_{i=1}^{n} i \mathbb{1}_{\{A_i\}}(\omega)$$

tal que

$$\eta(i) = P[w \in S^\infty : X(w) = i] \;=\; P(A_i) \;=\; \frac{j_i}{2^j}. \tag{2.2}$$

A probabilidade imagem P definida em (2.2) é a distribuição de probabilidade discreta $\eta(i)$ associada à variável aleatória X definida anteriormente.

Por outro lado, dado $X = \sum_{i=1}^{n} c_i \mathbb{1}_{\{K_i\}}$ uma variável aleatória obtemos que $\eta(i) = P[K_i]$.

□

Observe que a representação da variável aleatória X não é única: qualquer permutação dos elementos de Δ_j nos conduz a uma variável aleatória distinta. Por isso, fixamos a enumeração definida por Δ_j.

2.4 Produto Finito do Espaço Mensurável $(S^\infty, \mathcal{A})$

A classe de variáveis aleatória que vamos estudar evolui ao longo do tempo através de uma estrutura que envolve espaços mensuráveis produto. Assim, vamos contruir esses espaços e estudar algumas de suas propriedades. Denotamos por

$$(S^\infty)^n = \underbrace{S^\infty \times \ldots \times S^\infty}_{n\ vezes},$$

o produto cartesiano do S^∞ com ele mesmo n vezes.

Considere $\pi_i : (S^\infty)^n \to S^\infty$ a projeção na $i-$ésima coordenada, para $i = 1, \ldots, n$. Seja A um conjunto arbitrário de $(S^\infty)^n$, $L \subset \{1, 2, \cdots, n\}$ e $\omega_u \in S^\infty$, para todo $u \in L$. Definimos a seção de A em $\{\omega_u : u \in L\}$, por

$$A_{\{\omega_u; u \in L\}} = \{\{\omega_i; i \in L^c\} : (\omega_1, \cdots, \omega_n) \in A\}$$

Para todo $\{\omega_u : u \in L\}$ fixos, a transformação $A \to A_{\{\omega_u : u \in L\}}$ preserva as operações de união e interseção finitas, e complementação. Na realidade, se tomarmos $\{A_k\}$ uma família finita qualquer de subconjuntos de $(S^\infty)^n$, temos que

$$\left(\bigcap_{i=1}^{k} A_i\right)_{\{\omega_u : u \in L\}} = \bigcap_{i=1}^{k} (A_i)_{\{\omega_u : u \in L\}}$$

com o mesmo sendo válido para as operações de união finita e complementaçãao. Por exemplo, se tomarmos $n = 2$, isto é, $(S^\infty)^2$ e $A \subset (S^\infty)^2$, obtemos que

$$A_{\omega_2} \ = \ \{\omega_1 \in S^\infty : (\omega_1, \omega_2) \in A\}.$$

com $\omega_2 \in S^\infty$. Consideremos o conjunto $L \subset \{1, 2, \cdots, n\}$ e $\alpha(L)$ a cardinalidade deste conjunto. Outra operação que utilizaremos é a projeção de subconjuntos $A \subset (S^\infty)^n$ sobre $(S^\infty)^{\alpha(L)}$, que definiremos por

$$\mathrm{Proj}_{(S^\infty)^{\alpha(L)}}[A] \ = \ \left\{\{\omega_u : u \in L\} : A_{\{\omega_u; u \in L\}} \ \neq \ \emptyset\right\}.$$

A construção de uma álgebra sobre o espaço $(S^\infty)^n$ segue as mesmas linhas de contrução da topologia produto, trocando-se continuidade por mensurabilidade. Para isto, dizemos que um retângulo em $(S^\infty)^n$ é um subconjunto, na forma

$$R \ = \ R_1 \times R_2 \times \cdots \times R_n \ = \prod_{i=1}^{n} R_i = \{(\omega_1, \cdots, \omega_n) : \omega_i \in R_i\}$$

com $R_i \subset S^\infty$, $i = 1, \ldots, n$. Observamos que R é vazio se, e somente se, $R_i = \emptyset$ para algum $i = 0, 1, \cdots, n$. Ao aplicarmos operação de seção em um retângulo R de $(S^\infty)^n$, obtemos que

$$R_{\{\omega_u : u \in L\}} = \begin{cases} \prod_{j \in S^c} R_j, & ; \quad \{\omega_u; u \in L\} \in \prod_{u \in L} R_u \\ \\ \emptyset & ; \quad \{\omega_u; u \in L\} \notin \prod_{u \in L} R_u \end{cases}$$

Um retângulo $R = R_1 \times R_2 \times \cdots \times R_n$ em $(S^\infty)^n$ é denominado mensurável se $R_i \in \mathcal{A}_i$ para $1 \leq i \leq n$. Na seqüência, vamos mostrar que a classe de retângulos mensuráveis é uma semi-álgebra.

Lema 2.3. *Os retângulos mensuráveis de $(S^\infty)^n$ formam uma semi-álgebra.*

A partir desta estrutura de semi-álgebra sobre a classe de retângulos mensuráveis podemos construir a álgebra produto, utilizando resultados clássicos da teoria de conjuntos. Se tomarmos uniões finitas disjuntas (2 a 2) de retângulos mensuráveis obtemos uma álgebra $\mathcal{A}^n$ de sobconjuntos de $(S^\infty)^n$ [Neveu (1965); prop. I.6.1, pp. 25].

2.5 Probabilidade Condicional

Nesta seção definimos conceitos de probabilidade condicional e esperança condicional importantes para o estudo de processos estocásticos em geral. Consideremos $P : \mathcal{A} \to [0, 1]$ a probabilidade como definida na equação 2.1. Sejam $A, B \in \mathcal{A}$ com $P(B) > 0$, então a probabilidade condicional de A dado B é dada por

$$P[A|B] = \frac{P[A \cap B]}{P(B)}.$$

Seja $\mathcal{K} = \{K_1, \ldots, K_n\}$ uma partição. Definimos a probabilidade condicional de $A \in \mathcal{A}$ dado a partição $\mathcal{K}$, $P[A|\mathcal{K}] : S^\infty \to [0, 1]$ como

$$\begin{aligned} P[A|\mathcal{K}](\omega) &= \sum_{i=1}^{n} P[A|K_i] \mathbb{1}_{\{k_i\}}(\omega) \\ &= \sum_{i=1}^{n} \frac{P[A \cap K_i]}{P(K_i)} \mathbb{1}_{\{K_i\}}(\omega), \quad \omega \in S^\infty. \end{aligned}$$

Seja $X \in S(\mathcal{A})$ uma variável aleatória. Definimos a esperança condicional da variável X dado à partição $\mathcal{K}$ como

$$E[X \mid \mathcal{K}](\omega) = \sum_{i=1}^{n} E[X \mid K_i] 1\!\!1_{\{K_i\}}(\omega), \quad \omega \in S^{\infty}$$

no qual

$$E[X \mid K_i] = \frac{E[X 1\!\!1_{\{K_i\}}]}{E[1\!\!1_{\{K_i\}}]} = \frac{E[X 1\!\!1_{\{K_i\}}]}{P(K_i)}.$$

A partir da partição $\mathcal{K}$, obtemos a álgebra $\mathcal{A}(\mathcal{K})$ cujos elementos são formados pela união finita disjunta de elementos de $\mathcal{K}$ mais o conjunto vazio $\emptyset$ e o conjunto S^{∞}. Por outro lado, dado uma sub-álgebra finita $\mathcal{G} \subset \mathcal{A}$, obtemos uma partição $\mathcal{K} \subset \mathcal{A}$ tal que $\mathcal{G} = \mathcal{A}(\mathcal{K})$. Com isso, para toda variável $X \in S(\mathcal{A})$, verificamos que

$$E[X \mid \mathcal{A}(\mathcal{K})](\omega) = E[X \mid \mathcal{K}](\omega), \quad \omega \in S^{\infty}.$$

Desta forma, observe que se $X(\omega) = 1\!\!1_{\{A\}}(\omega)$ para A um conjunto qualquer, então

$$\begin{aligned} E[X|\mathcal{A}(\mathcal{K})](\omega) &= E[1\!\!1_{\{A\}}|\mathcal{K}](\omega) = \sum_{i=1}^{n} E[1\!\!1_{\{A\}}|K_i] 1\!\!1_{\{K_i\}}(\omega) \\ &= \sum_{i=1}^{n} \frac{P[A \cap K_i]}{P(K_i)} 1\!\!1_{\{K_i\}}(\omega) \\ &= P[A|\mathcal{K}](\omega) \end{aligned}$$

Outras propriedades da esperança condicional:

1. Sejam $X_1, X_2 \in S(\mathcal{A})$ e $a \in \mathbb{R}$, então

$$E[aX_1 + X_2 \mid \mathcal{K}] = aE[X_1 \mid \mathcal{K}] + E[X_2 \mid \mathcal{K}].$$

2. Se $\mathcal{K}_1$ e $\mathcal{K}_2$ são partições independetes, então para $X \in S(\mathcal{K}_1)$ temos que

$$E[X \mid \mathcal{K}_2] = E[X].$$

3. Seja $X \in S(\mathcal{A}(\mathcal{K}))$. Então

$$E[X \mid \mathcal{A}(\mathcal{K})](\omega) = X(\omega).$$

2.6 Cálculo Estocástico

Seja $\mathbb{N}_0$ o conjunto formado pelos números naturais e pelo elemento zero. Um processo estocástico é uma família de variáveis aleatórias $Z = \{Z(n,\cdot) : n \in \mathbb{N}_0\}$ definidas sobre o espaço de probabilidade $(S^\infty, \mathcal{A}, P)$. Uma das principais estratégias para se estudar processos estocásticos, proposta por Doob (1953), consiste em definir uma seqüência de álgebras que correspondem à quantidade de informações que se tem ao longo do tempo (n). Denominamos essa seqüência de álgebras de filtragem.

Definição 2.1. *Uma filtragem $\{\mathcal{E}_n\}$ no espaço de probabilidade $(S^\infty, \mathcal{A}, P)$ é definida como uma seqüência de álgebras encaixantes tais que*

(i) $\mathcal{E}_n \subset \mathcal{A}$ *e*

(ii) $\mathcal{E}_n \subset \mathcal{E}_{n+1},\ n \in \mathbb{N}$.

Uma das filtragens mais usuais consiste em considerar os cilindros de dimensão finita. Para isto, definimos uma classe de conjuntos Δ_n que contempla os cilindros com base em dimensão menor ou igual a n,

$$\Delta_n = \{\pi_1^{-1}(\{\omega_1\}) \cap \ldots \cap \pi_k^{-1}(\{\omega_k\}) : (\omega_1, \ldots, \omega_k) \in S^k,\ k \leq n\} \cup \{\emptyset, S^\infty\},$$

para $n \in \mathbb{N}$. No espaço de probabilidade $(S^\infty, \mathcal{A}, P)$, considere a filtragem $\{A_n\}$

$$\begin{aligned}
\mathcal{A}_0 &= \{\emptyset, S^\infty\} \\
\mathcal{A}_1 &= \{\emptyset, \pi_1^{-1}(\{0\}), \pi_1^{-1}(\{1\}), S^\infty\} \\
&\vdots \\
\mathcal{A}_n &= \text{ união finita disjunta de elementos de } \Delta_n, \quad n \in \mathbb{N}.
\end{aligned}$$

Neste caso, $\{\mathcal{A}_n\}$ é a filtragem gerada pela seqüência de partições $\{\Delta_n\}$.

Definição 2.2. *Um processo estocástico Z é adaptado para a filtragem $\{\mathcal{E}_n\}$ se*

$$Z(n,.) \in S(\mathcal{E}_n),\ \forall n \in \mathbb{N}.$$

Um dos principais processos estudados nesta dissertação é o processo de contagem. Para isso, tomamos X uma variável aleatória assumindo valores $\{1, 2, \ldots, k\}$, com representação

$$X(\omega) = \sum_{i=1}^{k} i \mathbb{1}_{\{A_i\}}$$

no qual $A_1, \ldots, A_k$ é uma partição do S^∞. O processo de contagem associado à variável aleatória X é definido por

$$N(i, \omega) = \mathbb{1}_{\{X(\omega) \leq i\}}, \ i = 1, \ldots, k \text{ e } \omega \in S^\infty. \tag{2.3}$$

A seguir, definimos uma seqüência de partições associada ao processo de contagem N,

$$\begin{aligned}
\Delta_0^N &= \{\emptyset, S^\infty\} \\
\Delta_1^N &= \{\{N(1,.) = 0\}; \{N(1,.) = 1\}\} \\
&= \{A_1; A_1^c\} \\
&\vdots \\
\Delta_n^N &= \{\{N(1,.) = 0, \ldots, N(n,.) = 0\}, \{N(1,.) = 1, \ldots, N(n,.) = 0\}, \ldots, \{N(1,.) = 1, \ldots, N(n,.) = 1\}\} \\
&= \{(A_1 \cup A_2 \cup \ldots \cup A_n)^c\}, \{A_1^c \cup A_2 \cup \ldots \cup A_n\}, \ldots, \{A_1 \cup A_2 \cup \ldots \cup A_n\}\}.
\end{aligned}$$

A filtragem $\{\mathcal{A}_n^N\}$, também denominada filtragem interna do processo de contagem N, é a filtragem gerada pela seqüência de partições $\{\Delta_n^N\}$.

Definição 2.3. *Seja Z um processo estocástico e $\{\mathcal{E}_n\}$ uma filtragem definida no espaço de probabilidade $(S^\infty, \mathcal{A}, P)$. Z é um martingale se*

(i) $Z(n,.)$ é adaptado para a filtragem $\{\mathcal{E}_n\}$;

(ii) $|E[Z(n,.)]| < \infty, \quad \forall n \in \mathbb{N}$;

(iii) $E[Z(n+1,.) \mid \mathcal{E}_n] = Z_n$.

Z é um supermartingale se forem válidas as propriedades (i), (ii) e

$$E[Z(n+1,.) \mid \mathcal{E}_n] \leq Z_n.$$

Z é um submartingale se forem válidas (i), (ii) e

$$E[Z(n+1,.) \mid \mathcal{E}_n] \geq Z_n.$$

Se um processo estocástico Z é um supermartingale, então $-Z$ é um submartingale. Com isso, se um processo estocástico é um supermartingale e um submartingale, então ele é um martingale. Além disso, se Z é supermartingale a esperança de $Z(n,\cdot)$ é não crescente, isto é,

$$\mathrm{E}[Z(n,\cdot)] \geq \mathrm{E}[Z(m,\cdot)],\ n \leq m.$$

Se Z é um martingale a esperança de $Z(n,\cdot)$ é constante para todo $n \in \mathbb{N}$.

Exemplo 2.1. Sejam $X_1, \ldots, X_n$ variáveis independentes e identicamente distribuídas com $E(X_i) = 0,\ i = 1, \ldots, n$. Considere

$$S(k,\omega) = \sum_{i=1}^{k} X_i(\omega), \quad k \in \{1, \ldots, n\} \text{ e } \omega \in (S^\infty)^n.$$

Definimos a filtragem interna $\{\mathcal{F}_k\}$

$$\begin{aligned}
\mathcal{F}_0 &= \{\emptyset, (S^\infty)^n\} \\
\mathcal{F}_1 &= \text{álgebra gerada por } \{X_1\} \\
\mathcal{F}_2 &= \text{álgebra gerada por } \{X_1, X_2\} \\
&\vdots \\
\mathcal{F}_k &= \text{ álgebra gerada por } \{X_1, X_2, \ldots, X_k\}.
\end{aligned}$$

Com isso S é um processo estocático adaptado para a filtragem $\mathcal{F}_k$ e, mais ainda, S é um martingale. De fato,

$$\begin{aligned}
E[S(k+1,\cdot)|\mathcal{F}_k] &= E[X_{k+1} + S(k,\cdot)|\mathcal{F}_k] \\
&= E[X_{k+1}|\mathcal{F}_k] + S(k,\cdot) = E[X_{k+1}] + S(k,\cdot) = S(k,\cdot).
\end{aligned}$$

□

O processo de contagem N definido em (2.3) tem tem trajetórias não descrescentes. Assim, verificamos facilmente que N é um submartingale com relação à filtragem interna $\mathcal{A}_n^N$. De fato,

$$\begin{aligned}
E[N(i+1,.)|\mathcal{A}_i^N] &= E[N(i+1,.) - N(i,.) + N(i,.)|\mathcal{A}_i^N] \\
&= E[N(i+1,.) - N(i,.)|\mathcal{A}_i^N] + E[N(i,.)|\mathcal{A}_i^N] \geq N(i,.)
\end{aligned}$$

para $i = 1, \ldots, n$.

Proposição 2.1. *Seja $Z = \{Z(n,.) : n \in \mathbb{N}_0\}$ um martingale com respeito a filtragem $\mathcal{E}_n$ e $k \in \mathbb{N}$ tal que $1 \leq k \leq n$. Então, para $n > p$,*

$$E[Z^2(n,.)] - E[Z^2(p,.)] = \sum_{j=p}^{n-1} E[(Z(j+1,.) - Z(j,.))^2]$$

Demonstração. Seja $e_j = Z(j+1,.) - Z(j,.)$. Então, para $m \leq j$,

$$\begin{aligned} E(e_j|\mathcal{E}_m) &= E[E(Z(j+1,.) - Z(j,.)|\mathcal{E}_j)|\mathcal{E}_m] \\ &= E[-Z(j,.) + E(Z(j+1,.)|\mathcal{E}_j)|\mathcal{E}_m] \\ &= E[-Z(j,.) + Z(j,.)|\mathcal{E}_m] = 0 \end{aligned} \tag{2.4}$$

Podemos escrever $Z(n,.) = Z(p,.) + \sum_{j=p}^{n-1} e_j$. Assim,

$$E[Z^2(,.)] = E[Z^2(p,.)] + 2\sum_{j=p}^{n-1} E[Z(p,.)e_j] + E[(\sum_{j=p}^{n-1} e_j)^2].$$

Mas,

$$\left(\sum_{j=p}^{n-1} e_j\right)^2 = \sum_{j=p}^{n-1} e_j^2 + \sum_{j,j' j\neq j'} e_j e_{j'} \tag{2.5}$$

Logo,

$$E[Z^2(n,.)] = E[Z^2(p,.)] + \sum_{j=p}^{n-1} E(e_j^2) + \sum_{j,j' \ j\neq j'} E(e_j e_{j'}) + 2\sum_{j=p}^{n-1} + 2\sum_{j=p}^{n-1} E[Z(p,.)e_j].$$

Vamos mostrar que todos os termos que aparecem nas últimas duas somas são zero. Assuma que $j < j'$. Então,

$$E(e_j e_j^{'}) = E(E(e_j e_j^{'})|\mathcal{E}_{j+1}) = e_j E(e_j|\mathcal{E}_{j+1})$$

Como $j \leq j'$, temos, por 2.4, que $E(e_{j'}) = 0$, então $E(e_j e_j^{'}|\mathcal{E}_{j+1}) = 0$. De forma análoga, obtemos que $E[Z(p,.)e_j] = E[E(Z(p,.)e_j|\mathcal{E}_j)] = 0$. □

2.6.1 Martingale Array Difference

Nesta seção apresentamos a definição de martingale *array difference* e propriedades relacionadas. Estas informações serão necessárias para o estudo de convergência realizado no Capítulo 4.

Definição 2.4. *Se $Z = \{Z(n,.) : n \in \mathbb{N}_0\}$ é um processo estocástico adaptado a filtragem $\{\mathcal{E}_n\}$ com*

$$E[Z(n,.)|\mathcal{E}_{n-1}] = 0 \text{ para } n \geq 1, \tag{2.6}$$

então, Z é um martingale array difference.

Consideremos a soma parcial $U(n,.) = \sum_{j=1}^{n} Z(j,.)$. O seguinte resultado é válido.

Proposição 2.2. *Seja $\{U(n,.) : n \in \mathbb{N}_0\}$ um processo estocástico adptado à filtragem $\{\mathcal{E}_n\}$. Seja $Z(0,.) = 0$ e $Z(n,.) = U(n,.) - U(n-1,.)$ para $n \leq 1$. Então, o processo $\{U(n,.) : n \in \mathbb{N}_0\}$ é um martingale se, e somente se, $\{Z(n,.) : n \in \mathbb{N}_0\}$ é um martingale array difference.*

Demonstração. Temos que $Z(n,.) = U(n,.) - U(n-1,.)$. Então, $Z(n,.)$ é um martingale *array difference* se, e somente se,

$$E[Z(n,.)|\mathcal{E}_{n-1}] = E[U(n,.)|\mathcal{E}_{n-1}] - U(n-1,) = 0.$$

Mas $E[U(n,.)|\mathcal{E}_{n-1}] - U(n-1,) = 0$ se, e somente se, $U(n,.)$ é um martingale. □

2.7 Integral Estocástica

Nesta seção, introduzimos o conceito de integral estocástica, fundamental para o estudo de propriedades de estimadores. Tomemos $(S^\infty, \mathcal{A}, P)$ um espaço de probabilidade, $\{\mathcal{E}_n\}$ uma filtragem e denotemos por $(S^\infty, \mathcal{A}, \{\mathcal{E}_n\}, P)$ a base estocástica associada. Um processo estocástico Z está definido sobre a base estocástica $(S^\infty, \mathcal{A}, \{\mathcal{E}_n\}, P)$ se $Z(n,.)$ são variáveis aleatórias definidas sobre $(S^\infty, \mathcal{A}, P)$ e Z é adaptado para a filtragem $\{\mathcal{E}_n\}$.

Considere Y e V dois processos definidos na base estocástica $(S^\infty, \mathcal{A}, \{\mathcal{E}_n\}, P)$. Definimos a integral estocástica como sendo

$$(V.Y)(n,\cdot) = V(0,\cdot)Y(0,\cdot) + \sum_{i=1}^{n} V(i,\cdot)\Delta Y(i,\cdot)$$

no qual $\Delta Y(i,\cdot) = Y(i,\cdot) - Y(i-1,\cdot)$, para todo $n \in \mathbb{N}$.

Teorema 2.2. *Seja Y um martingale e V um processo estocástico, ambos definidos na base estocástica $(S^\infty, \mathcal{A}, \{\mathcal{E}_n\}, P)$. Então, as seguintes afirmações são equivalentes:*

(i) O processo V é previsível, isto é, $V(n,.) \in S(\mathcal{E}_{n-1})$.

(ii) A integral estocástica $(V.Y)$ é martingale.

Demonstração. Consideremos Y um martingale e V um processo previsível. Assim, temos que

$$\begin{aligned}
E[(V.Y)(n+1,.)|\mathcal{E}_n] &= E[V(0,.).Y(0,.) + \sum_{i=1}^{n} V(i,.)\Delta Y(i,.) + V(n+1,.)\Delta Y(n+1,.)|\mathcal{E}_n] \\
&= V(0,.)Y(0,.) + \sum_{i=1}^{n} V(i,.)\Delta Y(i,.) + E[V(n+1,.)\Delta Y(n+1,.)|\mathcal{E}_n] \\
&= (V.Y)(n,.) + E[V(n+1)\Delta Y(n+1,.)|\mathcal{E}_n] \\
&= (V.Y)(n,.) + V(n+1)E[\Delta Y(n+1,.)|\mathcal{E}_n] \\
&= (V.Y)(n,.) + V(n+1)E[Y(n+1,.)|\mathcal{E}_n] - Y(n,.) \\
&= (V.Y)(n,.) + V(n+1)[Y(n,.) - Y(n,.)] = (V.Y)(n,.)
\end{aligned}$$

Então, $E[(V.Y)(n+1,.)|\mathcal{E}_n] = (V.Y)(n,.)$, ou seja, $(V.Y)$ é uma martingale.

Por outro lado, como

$$E[(V.Y)(n+1,.)|\mathcal{E}_n] = (V.Y)(n,.) + E[V(n+1)\Delta Y(n+1,.)|\mathcal{E}_n],$$

e $(V.Y)$ é um martingale, $E[V(n+1)\Delta Y(n+1,.)|\mathcal{E}_n]$ deve ser igual a zero. No entanto,

$$E[V(n+1)\Delta Y(n+1,.)|\mathcal{E}_n] = 0 \Leftrightarrow V(n+1) \text{ e } \Delta Y(n+1,.) \text{ são independentes.}$$

Para realizar esta demonstração consideramos o fato de $\Delta Y(n+1,.)$ ser $\mathcal{A}_{n+1}$- mensurável e $V(n,.)$ ser $\mathcal{A}_n$- mensurável. □

Da mesma forma, mostramos que se Y é um supermartingale (submartingale) e V é um processo previsível, então, a integral estocástica $(V.Y)$ é um supermartingale (submartingale).

2.8 Decomposição de Doob-Meyer

A decomposição de Doob-Meyer é um resultado fundamental na teoria dos processos estocásticos, pois garante que um processo estocástico possa ser decomposto como a soma de um martingale e um processo previsível. Este fato permite a decomposição do processo de contagem N, associado a uma variável aleatória X, em um componente ruído (martingale) e um componente estimável (previsível). Utilizando deste componente previsível, vamos estabelecer o modelo de intensidade mltiplicativo.

Teorema 2.3. *Seja $Z : \mathbb{N}_0 \times S^\infty \to \mathbb{R}$ um processo estocástico definido na base estocástica $(S^\infty, \mathcal{A}, \{\mathcal{E}_n\}, P)$ com $Z(0,.) = 0$. O processo Z pode ser decomposto de forma única em*

$$Z = Y + A,$$

no qual Y é um martingale e A um processo previsível, satisfazendo

$$Y(n,.) = Z(0,.) + \sum_{l=1}^{n} [Z(l,.) - E(Z(l,.) \mid \mathcal{E}_{l-1})],\ n \geq 1 \tag{2.7}$$

e

$$A(n,.) = \sum_{l=1}^{n} [E[Z(l,.) \mid \mathcal{E}_{l-1}] - Z(l-1,.)],\ n \geq 1, \tag{2.8}$$

com $Y(0,.) = Z(0,.) = 0$ e $A(0,.) = 0$.

Demonstração. (i) Temos que $Z = Y + A$:

$$\begin{aligned}
Z(0,.) &= Y(0,.) + A(0,.) \\
Z(1,.) &= Z(0,.) + Z(1,.) - E[Z(1,.) \mid \mathcal{E}_0] + E[Z(1,.) \mid \mathcal{E}_0] - Z(0,.) \\
&= Y(1,.) + A(1,.) \\
Z(2,.) &= Z(0,.) + Z(1,.) - E[Z(1,.) \mid \mathcal{E}_0] + E[Z(1,.) \mid \mathcal{E}_0] - Z(0,.) \\
&+ Z(2,.) - E[Z(2,.) \mid \mathcal{E}_1] + E[Z(2,.) \mid \mathcal{E}_1] - Z(1,.) \\
&= Y(2,.) + A(2,.)
\end{aligned}$$

Assim, de forma indutiva,

$$Z(n,.) = Y(n,.) + A(n,.).$$

(ii) Da a equação 2.7, podemos ver que Y é $\mathcal{A}_n$-mensuráel. Além disso,

$$\begin{aligned} E[Y(n,.)|\mathcal{E}_{n-1}] &= Z_0 + \sum_{l=1}^{n-1}[Z(l,.) - E[Z(l,.)|\mathcal{E}_{l-1}]] + E[Z(n,.)|\mathcal{E}_{n-1}] - E[Z(n,.)|\mathcal{E}_{n-1}] \\ &= Y(n-1,.) \end{aligned}$$

Portanto Y é um martingale.

Observando a equação (2.8), verificamos que A é um processo previsível.

(iii) Considere, agora, que Z tem duas representações

$$Z = Y + A \quad \text{e} \quad Z = Y^{'} + A^{'}$$

com Y e $Y^{'}$ martingales e A e $A^{'}$ processos previsíveis. Então,

$$A^{'}(n+1,.) - A^{'}(n,.) = [A(n+1,.) - A(n,.)] - [Y(n+1,.) - Y(n,.)] + [Y^{'}(n+1,.) - Y^{'}(n,.)].$$

Agora,

$$\begin{aligned} E[A^{'}(n+1,.) - A^{'}(n,.)|\mathcal{E}_n] &= E[A(n+1,.) - A(n,.)|\mathcal{E}_n] + E[Y(n+1,.) - Y(n,.)|\mathcal{E}_n] \\ &+ E[Y^{'}(n+1,.) - Y^{'}(n,.)|\mathcal{E}_n] \end{aligned}$$

Mas, observe que, como Y é um martingale temos que

$$E[Y(n+1,.) - Y(n,.)|\mathcal{E}_n] = -Y(n,.) + E[Y(n+1,.)|\mathcal{E}_n] = -Y(n,.) + Y(n,.) = 0$$

Da mesma forma, como $Y^{'}$ também é um martingale, obtemos que $E[Y^{'}(n+1,.) - Y^{'}(n,.)|\mathcal{E}_n] = 0$. Assim,

$$E[A^{'}(n+1,.) - A^{'}(n,.)|\mathcal{E}_n] = E[A(n+1,.) - A(n,.)|\mathcal{E}_n].$$

Como A e $A^{'}$ são previsíveis

$$A^{'}(n+1,.) - A^{'}(n,.) = A(n+1,.) - A(n,.) \Rightarrow A^{'} = A.$$

Assim, temos também que $Y^{'} = Y$. Portanto, a decomposição de Doob é única.

$\square$

2.8.1 Processo Variação Quadrática

O processo variação quadrática corresponde a uma das principais aplicações da decomposição de Doob-Meyer. Dado um martingale Z com respeito a base estocástica $(S^\infty, \mathcal{A}, \{\mathcal{E}_n\}, P)$, o processo variação quadrática associado a Z pode ser interpretado como sua volatilidade, e será utilizado para calcularmos a variância de um martingale. Desde que Z é um martingale, obtemos da desigualdade de Jensen que Z^2 é um submartingale. Na realidade,

$$\begin{aligned} E[Z^2(n+1,.)|\mathcal{E}_n] &\geq (E[Z(n+1,.)|\mathcal{E}_n])^2 \\ &= Z^2(n+1,.). \end{aligned}$$

Pela decomposição de Doob-Meyer, existe um único processo previsível crescente denotado por $< Z, Z >$ tal que $Z^2 - < Z, Z >$ é um martingale. O processo $< Z, Z >$ é denominado variação quadrática (ou compensador) e é definido como

$$< Z, Z >= \sum_{l=1}^{i} E[(\Delta Z(l,.))^2 | \mathcal{E}_{l-1}^n], \quad i \leq k.$$

Ao considerar X uma variável aleatória, $N(i,.) = \mathbb{1}_{\{X \leq i\}}$ o processo de contagem associado a X com sua filtragem interna, sabemos que

$$N(i,.) = Y(i,.) + A(i,.),$$

com Y um martingale e A um processo previsível crescente.

Na seqüência, vamos calcular o processo variação quadrática de Y. Inicicalmente, temos que

$$[Y(i,.)]^2 = 2\sum_{l=1}^{i} Y(l-1,.)\Delta Y(l,.) + \sum_{l=1}^{i} [\Delta Y(l,.)]^2.$$

Além disso,

$$[\Delta Y(i,.)]^2 = [\Delta N(i,.) - \Delta A(i,.)]^2 = [\Delta N(i,.)]^2 - 2\Delta N(i,.)\Delta A(i,.) + [\Delta A(i,.)]^2.$$

Assim,

$$
\begin{aligned}
[Y(i,.)]^2 &= 2\sum_{l=1}^{i} Y(l-1,.)\Delta Y(l,.) + \sum_{l=1}^{i}[\Delta N(l,.)]^2 - 2\sum_{l=1}^{i}\Delta N(l,.)\Delta A(l,.) + \sum_{l=1}^{i}[\Delta A(l,.)]^2 \\
&= 2\sum_{l=1}^{i} Y(l-1,.)\Delta Y(l,.) + \sum_{l=1}^{i}\Delta N(l,.) - 2\sum_{l=1}^{i}\Delta N(l,.)\Delta A(l,.) + \sum_{l=1}^{i}[\Delta A(l,.)]^2 \\
&= 2\sum_{l=1}^{i} Y(l-1,.)\Delta Y(l,.) + \sum_{l=1}^{i}\Delta Y(l,.) + \sum_{l=1}^{i}\Delta A(l,.) \\
&- 2\sum_{l=1}^{i}\Delta A(l,.)\Delta Y(l,.) + \sum_{l=1}^{i}\Delta A(l,.)\Delta A(l,.) - 2\sum_{l=1}^{i}\Delta A(l,.)A(l,.) \\
&= 2\sum_{l=1}^{i} Y(l-1,.)\Delta Y(l,.) + \sum_{l=1}^{i}\Delta Y(l,.) + \sum_{l=1}^{i}\Delta A(l,.) \\
&- 2\sum_{l=1}^{i}\Delta A(l,.)\Delta Y(l,.) - \sum_{l=1}^{i}\Delta A(l,.)\Delta A(l,.) \\
&= \sum_{l=1}^{i}[2Y(l-1,.)+1-2\Delta A(l,.)]\Delta Y(l,.) + \sum_{l=1}^{i}[1-\Delta A(l,.)]\Delta A(l,.) \\
&= (G.Y)(i,.) + ((1-\Delta A).A)(i,.),
\end{aligned}
$$

no qual $G(l,.) = 2Y(l-1,.)+1-2\Delta A(l,.)$ para $l = 1, \cdots, k$. Desde que G é previsível, obtemos a decomposição de Y^2 em um martingale mais um processo previsível crescente. Assim, como a decomposição de Doob-Meier é única, obtemos que a variação quadrática é dada por

$$
< Y, Y > (i,.) = \sum_{l=1}^{i}[1-\Delta A(l,.)]\Delta A(l,.) = [(1-\Delta A).A](i,.). \tag{2.9}
$$

Proposição 2.3. *Sejam Z um martingale e V um processo previsível com respeito a base estocástica $(S^\infty, \mathcal{A}, \{\mathcal{E}_n\}, P)$. Assim, temos que*

$$
< (V.Z), (V.Z) >= (V^2. < Z, Z >) \tag{2.10}
$$

Demonstração.

$$
\begin{aligned}
< (V.Z), (V.Z) > (i,.) &= \sum_{l=1}^{i} E[(\Delta(V.Z)(l,.))^2 | \mathcal{E}_{l-1}] = \sum_{l=1}^{i} E[(V(l,.)\Delta Z(l,.))^2 | \mathcal{E}_{l-1}] \\
&= \sum_{l=1}^{i} E[V^2(l,.)(\Delta Z(l,.)^2) | \mathcal{E}_{l-1}] = \sum_{l=1}^{n} V^2(l,.) E[(\Delta Z(l,.))^2 | \mathcal{E}_{l-1}] \\
&= \sum_{l=1}^{i} V^2(l,.)\Delta < Z, Z > (l,.) = (V^2. < Z, Z >)(i,.).
\end{aligned}
$$

□

Capítulo 3

Modelo de Intensidade Multiplicativo

3.1 Introdução

O modelo de intensidade multiplicativo foi proposto por Aalen (1978) para o caso de variáveis aleatórias com distribuição de probabilidade absolutamente contínua. Para uma descrição detalhada deste modelo ver Fleming e Harrington (1990). Neste trabalho fizemos uma adaptação do modelo de intensidade multiplicativo para tratar o caso de variáveis aleatórias à valores em um espaço finito que é introduzido neste capítulo.

Para um maior entendimento das definições apresentadas e hipóteses assumidas neste capítulo, utilizamos como motivação os dados de alunos de graduação da Universidade de São Paulo apresentados na Seção 1.1.

Na segunda seção apresentamos o modelo de intensidade multiplicativo para variáveis aleatórias multinomais. Na terceira seção introduzimos o estudo de inferência no modelo de intensidade multiplicativo, obtemos os estimadores dos parâmetros do modelo e estudamos as propriedades finitas e assintóticas desses estimadores.

3.2 Desenvolvimento do Modelo

Consideremos W e C variáveis aleatórias multinomiais independentes assumindo valores $\{1, \cdots, k, k+1\}$, $k \in \mathbb{N}$. Aos valores assumidos por essas variáveis nos referimos como categorias ou classes. Quando a variável W assumir valores $\{1, \ldots, k\}$ ela representa a categoria em que ocorreu o

evento de interesse. Caso o evento de interesse não tenha ocorrido, W assume valor $k+1$ e a variável C representa a categoria em que ocorreu o evento de censura. De forma resumida temos que

$$W = \begin{cases} i, & \text{se ocorreu o evento de interesse na categoria } i, i = 1, \ldots, k; \\ k+1, & \text{se não ocorreu o evento de interesse até a categoria } k. \end{cases}$$

$$C = \begin{cases} i, & \text{se ocorreu o evento de censura na categoria } i, i = 1, \ldots, k; \\ k+1, & \text{se não ocorreu evento de censura até a categoria } k. \end{cases}$$

Definimos a variável multinomial X como sendo o mínimo entre W e C,

$$X = W \wedge C.$$

Denotamos por η a distribuição de probabilidade associada à variável aleatória X e por ζ à distribuição de probabilidade associada à variável W.

Consideremos o conjunto de dados apresentado na Seção 1.1. Associamos a cada aluno a variável aleatória multinomial X que representa o número de semestres até a última ocorrência. O evento de interesse é ter como última ocorrência evasão. Se o aluno evadiu no semestre i, $C = k+1$ e $X = W = i$, para $i = 1, \ldots, k$. Se o aluno migrou, faleceu ou concluiu no semestre i, eventos considerados de censura, $W = k+1$ e $X = C = i$, para $i = 1, \ldots, k$. Se o aluno estiver em andamento até o fim do estudo, $X = W = C = k+1$.

Observe que o valor k é o número de semestres analisados. Por exemplo, ao estudarmos os alunos ingressantes em 2004, teremos no máximo 10 semestres para serem considerados. Neste caso, $k = 10$ e, portanto, C e W assumem o valor 11 se o aluno estiver ativo no fim do estudo.

Em notação de eventos, como a evasão é o evento de interesse na análise dos alunos de graduação da USP, o evento $\{X = i, X = W\}$ determina a evasão do aluno no semestre i, para $i = 1, \ldots, k$, o evento $\{X = i, X = C\}$ determina se o aluno migrou, faleceu ou concluiu no semestre i, $i = 1, \ldots, k$, e o evento $\{X = k+1, X = W = C\}$ determina se o aluno está ativo até o semestre k.

Definimos o processo estocástico $N : \{0, 1, 2, \cdots, k\} \times S^{\infty} \to \{0, 1\}$ conforme a equação 3.1

$$N(i, .) = \mathbb{1}_{\{X \leq i, X = W\}}, \quad i = 1, \cdots, k. \tag{3.1}$$

O processo N é um processo de contagem, com trajetórias não decrescentes e a variável aleatória $N(i,.)$ determina se o evento de interesse ocorreu até a categoria i. Definimos a álgebra $\mathcal{B}_i^N$ associada ao processo N conforme

$$\begin{aligned}
\mathcal{B}_0^N &= \text{álgebra gerada por } \{\emptyset, S^\infty\} \\
\mathcal{B}_1^N &= \text{álgebra gerada por } \{N(1,.), C\} \\
\mathcal{B}_2^N &= \text{álgebra gerada por } \{N(1,.), N(2,.), C\} \\
&\vdots \\
\mathcal{B}_i^N &= \text{álgebra gerada por } \{N(1,.), N(2,.), N(3,.), \ldots, N(i,.), C\}
\end{aligned}$$

Denotamos por $\{\mathcal{B}^N\}$ a filtragem interna associada ao processo N formada pelas álgebras encaixantes $\mathcal{B}_i^N$, $0 \leq i \leq k$.

Pela decomposição de Doob-Meyer 2.8, sabemos que o processo de contagem N pode ser representado por

$$N(i,.) = Y(i,.) + A(i,.),$$

para $i = 1, \cdots, k$ no qual Y é martingale e A é um processo previsível com respeito a base estocástica $(S^\infty, \mathcal{A}, \{\mathcal{B}^N\}, P)$, com $A(0,.) = 0$ e $Y(0,.) = 0$. Desde que N é um submartingale, concluímos que A tem trajetórias não decrescentes, satisfazendo

$$\begin{aligned}
Y(i,.) &= \sum_{l=1}^{i} \left[N(l,.) - E[N(l,.)|\mathcal{B}_{l-1}^N]\right] \\
A(i,.) &= \sum_{l=1}^{i} \left[E[N(l,.)|\mathcal{B}_{l-1}^N] - N(l-1,.)\right]
\end{aligned}$$

com $A(0,.) = 0$ e $Y(0,.) = 0$. Na seqüência, são calculados os processos estocásticos relacionados com a decomposição de Doob-Meyer.

$$\begin{aligned}
A(1,.) &= E[N(1,.)|\mathcal{B}_0^N] - N(0,.) = E[N(1,.)] = E[\mathbb{1}_{\{X \leq 1, X = W\}}] = P[W = 1] \\
A(2,.) &= A(1,.) + E[N(2,.)|\mathcal{B}_1^N] - N(1,.) \\
&= P[W = 1] + E[N(2,.) - N(1,.)|\mathcal{B}_1^N] = P[W = 1] + E[N(2,.) - N(1,.)|N(1,.), C]
\end{aligned}$$

Observe que se $N(1,.) = 1$, então

$$N(2,.) - N(1,.) = 1\!\!1_{\{X=2, X=W\}} = 0,$$

e portanto, $E[N(2,.) - N(1,.)|N(1,.), C] = 0$. Se $N(1,.) = 0$ e $C = 1$, $E[N(2,.) - N(1,.)|N(1,.), C] = 0$. E, finalmente, se $N(1,.) = 0$ e $C \geq 2$, $E[N(2,.) - N(1,.)|N(1,.), C] \neq 0$. Desde que W e C são independentes, concluímos que

$$\begin{aligned}
E[N(2,.) - N(1,.)|N(1,.), C] &= E[N(2,.) - N(1,.)|N(1,.) = 0, C \geq 2]1\!\!1_{\{N(1,.)=0, C\geq 2\}} \\
&= E[1\!\!1_{\{X\leq 2, X=W\}} - 1\!\!1_{\{X\leq 1, X=W\}}|N(1,.) = 0, C \geq 2]1\!\!1_{\{N(1,.)=0, C\geq 2\}} \\
&= E[1\!\!1_{\{W=2\}}|N(1,.) = 0, C \geq 2]1\!\!1_{\{N(1,.)=0, C\geq 2\}} \\
&= \frac{P[W = 2; N(1,.) = 0; C \geq 2]}{P[N(1,.) = 0; C \geq 2]} 1\!\!1_{\{N(1,.)=0, C\geq 2\}} \\
&= \frac{P[W = 2]}{P[W \geq 2]} 1\!\!1_{\{N(1,.)=0, C\geq 2\}}
\end{aligned}$$

Com isso,

$$A(2,.) = A(1,.) + \frac{P[W = 2]}{P[W \geq 2]} 1\!\!1_{\{N(1,.)=0, C\geq 2\}}.$$

Realizando cálculos análogos a estes, obtemos

$$\begin{aligned}
A(k,.) &= A(k-1,.) + E[N(k,.)|\mathcal{A}^N_{k-1}] - N(k-1,.) \\
&= A(k-1,.) + E[N(k,.) - N(k-1,.)|\mathcal{A}^N_{k-1}] \\
&= A(k-1,.) + E[N(k,.) - N(k-1,.)|N(k-1,.) = 0, C \geq k]1\!\!1_{\{N(k-1,.)=0, C\geq k\}} \\
&= A(k-1,.) + E[1\!\!1_{\{W=k\}}|N(k-1,.) = 0, C \geq k]1\!\!1_{\{N(k-1,.)=0, C\geq k\}} \\
&= A(k-1) + \frac{P[W = k]}{P[W \geq k]} 1\!\!1_{\{N(k-1,.)=0, C\geq k\}}.
\end{aligned}$$

De forma geral, obtemos

$$A(i,.) = \sum_{l=1}^{i} \frac{P[W = l]}{P[W \geq l]} 1\!\!1_{\{N(l-1,.)=0, C\geq l\}} = \sum_{l=1}^{i} \frac{P[W = l]}{P[W \geq l]} 1\!\!1_{\{X\geq l\}} \tag{3.2}$$

para $i = 1, \ldots, k+1$.

Finalmente, concluímos que

$$Y(i,.) = N(i,.) - \sum_{l=1}^{i} \frac{P[W = l]}{P[W \geq l]} 1\!\!1_{\{X\geq l\}}.$$

para $i = 1, \ldots, k+1$.

Definição 3.1. *A função intensidade* $h : \{0, 1, 2, \cdots, k\} \to [0, 1]$, *associada ao processo de contagem* N, *é definida por*

$$h(i) = \frac{P[W = i]}{P[W \geq i]}, \quad i = 1, \cdots, k, \tag{3.3}$$

com $h(0) = 0$ *e* $h(k) \leq 1$.

A partir da definição a função intensidade h, obtemos que a distribuição de probabilidade associada à variável aleatória multinomial é dada por

$$\begin{aligned} \zeta(i) = P[W = i] &= [1 - h(1)][1 - h(2)] \ldots [1 - h(i-1)]h(i, .) \\ &= h(i) \prod_{l=1}^{i-1} [1 - h(l)] \quad ; \quad i = 1 \cdots, k. \end{aligned} \tag{3.4}$$

De fato, pela equação 3.3, temos que

$$h(1) = P[W = 1];$$

$$\begin{aligned} h(2) &= \frac{P[W = 2]}{P[W \geq 2]} = \frac{P[W = 2]}{1 - P[W = 1]} \\ &\Rightarrow P[W = 2] = h(2)(1 - P[W = 1]) = (1 - h(1))h(2); \end{aligned}$$

$$\begin{aligned} h(3) &= \frac{P[W = 3]}{P[W \geq 3]} = \frac{P[W = 3]}{1 - (P[W = 1] + P[X = 2])} \\ &\Rightarrow P[W = 3] = h(3)[1 - (P[W = 2] + P[W = 1])] = (1 - h(1))(1 - h(2))h(3). \end{aligned}$$

Por indução sobre k, obtemos que

$$\begin{aligned} h(k) &= \frac{P[W = k]}{P[W \geq k]} = \frac{P[W = k]}{1 - (\sum_{l=1}^{k-1} P[W = l])} \\ &\Rightarrow P[W = k] = h(k)[1 - \sum_{l=1}^{k-1} P[W = l]] \\ &= h(k)[1 - [(1 - h(1)) \ldots (1 - h(k-2))h(k-1)] + \ldots + (1 - h(1))h(2) + h(1)] \\ &= h(k)(1 - h(1))(1 - h(2))(1 - h(3)) \ldots (1 - h(k-1)) \end{aligned}$$

Assim, concluímos que a relação apresentada na equação 3.4 é válida.

Defininamos o processo previsível $V(i, .) = \mathbb{1}_{\{X \geq i\}}$, em que

$$\{X \geq i\} = \{X \geq i, X = W\} \cup \{X \geq i, X = C\}, \quad i = 1, \ldots, k.$$

Com isso, podemos denotar a distribuição de probabilidade ζ definda na equação 3.5 da seguinte maneira

$$\zeta(i) = P[W = i] = \prod_{l=1}^{k} [1 - h(i)]^{V(l,.)-\Delta N(l,.)} [h(i)]^{V(l,.)}. \tag{3.5}$$

A partir da função intensidade, o processo previsível A pode ser representado na forma

$$A(i,.) = \sum_{l=1}^{i} h(l) 1\!\!1_{\{X \geq l\}} \;=\; \sum_{l=1}^{i} V(l,.)\Delta H(l),$$

no qual H é a função intensidade acumulada, dada por $H(i) = \sum_{l=1}^{i} h(l)$. Assim, concluímos que o processo previsível A pode ser representado na forma multiplicativa

$$A(i,.) = \sum_{l=1}^{i} V(l,\cdot)\Delta H(l) = (V.H)(i,.),$$

na qual $\Delta H(l) = H(l) - H(l-1) = h(l)$. Como V é previsível, o processo A é determinado pela função intensidade h (ou, pela função intensidade acumulada H). Então, dado uma variável aleatória W, existe uma única função intensidade h satisfazendo (3.3). A seguir, vamos mostrar que a recíproca também é válida.

Teorema 3.1. *As seguintes afirmações são válidas:*

a. Dado W uma variável aleatória multinomial definida sobre $(S^{\infty}, \mathcal{A}, P)$ com distribuição de probabilidade $\zeta(i) = P[W = i]$. Então, existe uma única função $h : \{0, 1, 2, \cdots, k\} \to [0, 1]$, tal que

$$h(i) \;=\; \frac{P[W = i]}{P[W \geq i]}, \tag{3.6}$$

com $h(0) = 0$, $i = 1, \cdots, k$.

b. Dado uma função h definida no item (a.), existe uma única variável aleatória W sobre $(S^{\infty}, \mathcal{A}, P)$ cuja distribuição de probabilidade é unicamente determinada pela função intensidade h e satisfaz

$$\zeta(i) = P[W = i] \;=\; h(i) \prod_{l=1}^{i-1} [1 - h(l)] = \prod_{l=1}^{k} [1 - h(i)]^{V(l,.)-\Delta N(l,.)} [h(i)]^{V(l,.)},$$

na qual $V(l,.) = 1\!\!1_{\{X \geq l\}}$ e $N(i,.) = 1\!\!1_{\{X \leq l, X = W\}}$, para todo $i = 1, \cdots, k$.

Demonstração. Basta aplicarmos o Teorema 2.1. □

Com este Teorema, relacionamos (de forma única) a classe de variáveis aleatórias W e classe de funções intensidades h. Com isso, ao desenvolvermos um método para estimar a função intensidade h, estimamos também a distribuição de probabilidade ζ.

3.3 Inferência no Modelo de Intensidade Multiplicativo

Consideremos uma amostra aleatória simples

$$((W_1^{(p)}, C_1^{(p)}), \ldots, (W_{n_p}^{(p)}, C_{n_p}^{(p)})), \tag{3.7}$$

para $p = 1, \ldots, J$ e $n_p, J \in \mathbb{N}$. Além disso, consideremos que existe indpendência entre as variáveis $W_m^{(p)}$ e $C_m^{(p)}$, para $\forall m = 1, \ldots, n_p$.

As variáveis $W_m^{(p)}$ e $C_m^{(p)}$ são variáveis aleatórias multinomiais assumindo valores $\{1, \ldots, k+1\}$. A variável $W_m^{(p)}$ representa a categoria de ocorrência do evento de interesse para o item m da amostra p quando assumir valores $1, \ldots, k$. Se $W_m^{(p)} = k+1$, o evento de interesse não ocorreu para o item m da amostra p e, portanto, a variável $C_m^{(p)}$ representa a categoria em que ocorreu o evento de censura para este item. Denotamos por $\zeta^{(p)} : \{0, 1, \ldots, k+1\} \to [0, 1]$ a distribuição de probabilidade associada à variável $W_m^{(p)}, m = 1, \ldots, n_p$.

De forma análoga, definimos a variável $X_m^{(p)}$ como sendo o mínimo entre as variáveis $W_m^{(p)}$ e $C_m^{(p)}$, coforme a equação 3.8,

$$X_m^{(p)} = W_m^{(p)} \wedge C_m^{(p)} \tag{3.8}$$

tendo $X_m^{(p)}, m = 1, \ldots, n_p$, distribuição de probabilidade $\eta^{(p)} : \{0, 1, \ldots, k+1\} \to [0, 1]$.

Em relação ao conjunto de dados referente aos alunos de graduação da USP apresentado na Seção 1.1, temos que p corresponde ao ano de ingresso do aluno, n_p corresponde ao número de ingressantes no ano p e cada uma das variáveis da amostra $X^{(p)}$ está associada a cada um dos alunos que ingressou no ano p, representando o semestre de última ocorrência do aluno.

O evento $\{X_m^{(p)} = i, X_m^{(p)} = W_m^{(p)}\}$ representa se o aluno m ingressante no ano p, para $m = 1, \ldots, n_p$ e $p = 1, \ldots, J$, evadiu até o semestre i, para $i = 1, \ldots, k$. Se $W_m^{(p)} = k+1$, então o aluno m ingressante no ano p migrou, faleceu, concluiu ou ainda está ativo. Para

identificar qual desses possíveis eventos ocorreu, utilizamos os eventos $\{X_m^{(p)} = i, X_m^{(p)} = C_m^{(p)}\}$ e $\{X_m^{(p)} = k+1, X_m^{(p)} = C_m^{(p)}\}$, que indicam que o aluno faleceu, migrou ou concluiu no semestre i e que o aluno ainda está ativo ao fim do estudo, respectivamente.

Definimos o processo associado à variável aleatória $X_m^{(p)}$

$$N_m^{(p)}(i,.) = \mathbb{1}_{\{X_m^{(p)} \leq i, X_m^{(p)} = W_m^{(p)}\}}, \; m = 1, \cdots, n_p. \tag{3.9}$$

Este processo indica se o evento de interesse ocorreu para o item m até a categoria i. Definimos também o processo de contagem N^{n_p} associado a amostra aleatória $X^{(p)}$ conforme a equação 3.10

$$N^{n_p}(i,.) = \sum_{m=1}^{n_p} N_m^{(p)}(i,.). \tag{3.10}$$

A variável aleatória $N^{n_p}(i,.)$ indica quantos vezes o evento de interesse ocorreu para a amostra $X^{(p)}$ até a categoria i, $i = 1, \ldots, k$. Desde que N^{n_p} é soma de processos de contagem, concluímos que N^{n_p} é um submartingale definido sobre $\{1, 2, \cdots, k\} \times (S^\infty)^{n_p}$ com valores em $\{0, 1, 2, \cdots, n_p\}$.

Considerando os dados dos alunos de graduação da USP, temos que as variáveis aleatórias $N_m^{(p)}(i,.)$ e $N^{n_p}(i,.)$ indicam, respectivamente, se o aluno m evadiu até o semestre i e quantos alunos ingressantes no ano p evadiram até o semestre i.

A seguir definimos as álgebras encaixantes $\mathcal{B}_i^{N^{n_p}}$, para $i = 1, \ldots, k$,

$$\begin{aligned}
\mathcal{B}_0^{N^{n_p}} &= \text{álgebra gerada por } \{\emptyset, S^\infty\} \\
\mathcal{B}_1^{N^{n_p}} &= \text{álgebra gerada por } \{N^{n_p}(1,.), C_1^{(p)}, \ldots, C_{n_1}^{(p)}\} \\
&\vdots \\
\mathcal{B}_k^{N^{n_p}} &= \text{álgebra gerada por } \{N^{n_p}(1,.), \ldots, N^{n_p}(k,.), C_1^{(p)}, \ldots, C_{n_p}^{(p)}\}.
\end{aligned}$$

A filtragem interna, $\{\mathcal{B}^{N^{n_p}}\}$, associada ao processo N^{n_p} é a união das álgebras encaixantes $\mathcal{B}_i^{N^{n_p}}, 0 \leq i \leq k$. Pela decomposição de Doob-Meyer, o processo de contagem N^{n_p} pode ser representado da seguinte forma

$$N^{n_p}(i,.) = Y^{n_p}(i,.) + A^{n_p}(i,.)$$

no qual A^{n_p} é um processo previsível com trajetórias não decrescentes, Y^{n_p} é um martingale, com $A^{n_p}(0,.) = Y^{n_p}(0,.) = 0$, dados respectivamente por

$$A(i,.) = \sum_{l=1}^{i} \left[E[N(l,.)|\mathcal{B}_{l-1}^{N}] - N(l-1,.)\right]$$

$$Y(i,.) = \sum_{l=1}^{i} \left[N(l,.) - E[N(l,.)|\mathcal{B}_{l-1}^{N}]\right].$$

Na seção anterior, foram calculados os processos A e Y para uma única variável. Realizando cálculos análogos para o caso de uma amostra, obtemos que

$$A^{n_p}(i,.) = \sum_{l=1}^{i} \sum_{m=1}^{n_p} \mathbb{1}_{\{X_m^{(p)} \geq l\}} h^{(p)}(l),$$

em que

$$h^{(p)}(i) = \frac{P[W_m^{(p)} = i]}{P[W_m^{(p)} \geq i]},$$

é a função intensidade associada ao processo de contagem $N^{n_p}(i,.)$, para $i = 1, \ldots, k$ e $\forall m, p$, com $m = 1, \ldots, n_p$ e $p = 1, \ldots, J$. Conseqüentemente, obtemos que

$$Y^{n_p}(i,.) = N^{n_p}(i,.) - \sum_{l=1}^{i} \sum_{m=1}^{n_p} \mathbb{1}_{\{X_m^{(p)} \geq l\}} h^{(p)}(l)$$

Consideremos o processo previsível V^{n_p} e a função intensidade acumulada $H^{(p)}$ dados, respectivamente, por

$$V^{n_p}(l,.) = \sum_{m=1}^{n_p} \mathbb{1}_{\{X_m^{(p)} \geq l\}} \tag{3.11}$$

$$H^{(p)}(i) = \sum_{l=1}^{i} h^{(p)}(l). \tag{3.12}$$

Com isso, obtemos o modelo de intensidade multiplicativo

$$A^{n_p}(i,.) = \sum_{l=1}^{i} V^{n_p}(l,.) h^{(p)}(l) = \sum_{l=1}^{i} V^{n_p}(l,.) \Delta H^{(}p) = (V^{n_p}.H^{(p)})(l,.).$$

Portanto, como V^{n_p} é um processo previsível, o processo A^{n_p} é unicamente determinado pela função intensidade $h^{(p)}$. Nosso objetivo é estimar a função intensidade $h^{(p)}$ para determinarmos o modelo de intensidade multiplicativo.

Um estimador para a função intensidade $h^{(p)}$ é um processo estocástico $\hat{h}^{n_p} : \{0, 1, 2, \cdots, k\} \to [0, 1]$ definido na base estocástica produto $((S^\infty)^{n_p}, \mathcal{B}^{n_p}, \{\mathcal{B}^{N^{n_p}}\}, P^{n_p})$, no qual P^{n_p} corresponde à probabilidade produto. Neste trabalho, um critério martingale é proposto para determinarmos um estimador $\hat{h}^{n_p}$ ótimo para a função intensidade. Desde que o martingale representa o ruído (parte não previsível), igualamos o martingale Y^{n_p} a zero. Assim, para $i = 1$

$$\begin{aligned} Y^{n_p}(1,.) &= N^{n_p}(1,.) - \hat{h}^{n_p}(1) \sum_{m=1}^{n_p} \mathbb{1}_{\{X_m^{(p)} \geq 1\}} = 0 \\ \Rightarrow \hat{h}^{n_p}(1) &= \frac{N^{n_p}(1,.)}{\sum_{m=1}^{n_p} \mathbb{1}_{\{X_m^{(p)} \geq 1\}}} = \frac{N^{n_p}(1,.)}{V^{n_p}(1,.)}. \end{aligned}$$

Para $i = 2$,

$$\begin{aligned} Y^{n_p}(2,.) &= N^{n_p}(2,.) - \left[\hat{h}^{n_p}(2) \sum_{m=1}^{n_p} \mathbb{1}_{\{X_m^{(p)} \geq 2\}} - \hat{h}^{n_p}(1) \sum_{m=1}^{n_p} \mathbb{1}_{\{X_m^{(p)} \geq 1\}}\right] = 0 \\ Y^{n_p}(2,.) &= N^{n_p}(2,.) - \left[\hat{h}^{n_p}(2) V^{n_p}(2,.) - \hat{h}^{n_p}(1) V^{n_p}(1,.)\right] = 0 \\ Y^{n_p}(2,.) &= N^{n_p}(2,.) - \left[\hat{h}^{n_p}(2) V^{n_p}(2,.) - \frac{N^{n_p}(1,.)}{V^{n_p}(1,.)} V^{n_p}(1,.)\right] = 0 \\ \Rightarrow \hat{h}^{n_p}(2) &= \frac{\Delta N^{n_p}(2,.)}{V^{n_p}(2,.)}, \end{aligned}$$

em que $\Delta N(2,.) = N(2,.) - N(1,.)$. De forma geral, obtemos o estimador de *Kaplan-Meier*

$$\hat{h}^{n_p}(i) = \frac{\Delta N^{n_p}(i,.)}{V^{n_p}(i,.)}, \tag{3.13}$$

em que $\Delta N(i+1,.) = N(i+1,.) - N(i,.)$, para $i = 1, \ldots, k$, com $\hat{h}^{n_p}(0) = 0$. O estimador de $h^{(p)}$ está bem definido, pois podemos assumir, sem perda de generalidade, que $V^{n_p}(k,.) > 0$ e, como $V^{n_p}(i,.)$ é um processo decrescente para $i = 1, \ldots, k$, $V^{n_p}(i,.) > 0 \ \forall i$.

A partir daí, obtemos o estimador da distribuição de probabilidade da variável $W_m^{(p)}, m = 1, \ldots, n_p$,

$$\begin{aligned} \hat{\zeta}^{n_p}(i) &= \hat{h}^{n_p}(i) \prod_{l=1}^{i-1} [1 - \hat{h}^{n_p}(l)] \\ &= \prod_{l=1}^{i} [1 - \hat{h}^{n_p}(l)]^{V^{n_p}(l,.) - \Delta N^{n_p}(l,.)} [\hat{h}^{n_p}(l)]^{V^{n_p}(l,.)}. \end{aligned} \tag{3.14}$$

para todo $i \in \{1, \ldots, k\}$ e também o estimador da função intensidade acumulada $H^{(p)}$

$$\hat{H}^{n_p}(i) = \sum_{l=1}^{i} \hat{h}^{n_p}(l) = \sum_{l=1}^{i} \frac{\Delta N^{n_p}(l,.)}{V^{n_p}(l,.)} \tag{3.15}$$

3.3.1 Propriedades dos estimadores para amostras finitas

Nesta seção estudadamos propriedades para amostras finitas dos estimadores $\hat{H}^{n_p}$, $\hat{h}^{n_p}$ e $\hat{\zeta}^{n_p}$. Inicialmente, observe que $\Delta N_m^{(p)}(i,.)$ pode assumir apenas dois valores, 0 ou 1. A probabilidade dele assumir o valor 1 é

$$P[\Delta N_m^{(p)}(i,.) = 1] = P[X_m^{(p)} = 1, X_m^{(p)} = W_m^{(p)}] = \zeta^{(p)}(i).$$

Logo, $\Delta N_m^{(p)}(i,.)$ tem distribuição de Bernoulli com parâmetro $\zeta^{(p)}(i)$. Considerando que $\Delta N^{n_p}(i,.) = \sum_{m=1}^{n_p} \Delta N_m^{(p)}(i,.)$, obtemos que

$$\Delta N^{n_p}(i,.) \sim \text{Binomial } (n_p, \zeta^{(p)}(i)), \ i = 1, \ldots, k.$$

Lema 3.1. *O estimador $\hat{H}^{n_p}$ da função intensidade acumulada $H^{(p)}$ pode ser escrito como*

$$\hat{H}^{n_p}(i) = \left(\frac{1}{V^{n_p}}.Y^{n_p}\right)(i,.) + H^{(p)}(i). \tag{3.16}$$

Demonstração. Como apresentado na equação (3.15), o estimador da função de intensidade acumulada é

$$\hat{H}^{n_p}(i) = \sum_{l=1}^{i} \hat{h}^{n_p}(l) = \sum_{l=1}^{i} \frac{\Delta N^{n_p}(l,.)}{V^{n_p}(l,.)} \tag{3.17}$$

Pela Decomposição de Doob-Meyer, obtemos

$$\begin{aligned}
\sum_{l=1}^{i} \frac{\Delta N^{n_p}(l,.)}{V^{n_p}(l,.)} &= \sum_{l=1}^{i} \frac{\Delta[Y^{n_p}(l,.) + A^{n_p}(l,.)]}{V^{n_p}(l,.)} \\
&= \sum_{l=1}^{i} \frac{\Delta Y^{n_p}(l,.)}{V^{n_p}(l,.)} + \sum_{l=1}^{i} \frac{\Delta A^{n_p}(l,.)}{V^{n_p}(l,.)} \\
&= \left(\frac{1}{V^{n_p}}.Y^{n_p}\right)(i,.) + \sum_{l=1}^{i} h^{(p)}(l) \\
&= \left(\frac{1}{V^{n_p}}.Y^{n_p}\right)(i,.) + H^{(p)}(i).
\end{aligned} \tag{3.18}$$

□

Com esses resultados, conseguimos provar os próximos teoremas.

Teorema 3.2. *Considere o estimador $\hat{H}^{n_p}$ da função intensidade acumualada $H^{(p)}$ e o estimador $\hat{h}^{n_p}$ da função intensidade $h^{(p)}$. Temos que*

$$E\left[\hat{H}^{n_p}(i)\right] = H^{(p)}(i),$$

e

$$E[\hat{h}^{n_p}(i)] = h^{(p)}(i),$$

para $i = 1, \ldots, k$.

Demonstração. (a) Por definição, $V^{n_p}(i,.)$, $i = 1, \ldots, k$ é um processo previsível, logo $\frac{1}{V^{n_p}(i,.)}$ também é um processo previsível. Assim, pelo Teorema 2.2 obtemos que $(\frac{1}{V^{n_p}}.Y^{n_p})(i,.)$, $i = 1, \ldots, k$ é um martingale, o que implica que $\mathrm{E}[(\frac{1}{V^{n_p}}.Y^{n_p})(i,.)] = 0$, $i = 1, \ldots, k$. Com isso, obtemos do Lema 3.1 que

$$\mathrm{E}\left[\hat{H}^{n_p}(i,.)\right] = H^{(p)}(i,.),\ i \leq k,$$

ou seja, o estimador $\hat{H}^{n_p}$ é um estimador não-viciado da função $h^{(p)}$

(b) Pelas equações (3.12) e (3.15) temos que

$$\hat{H}^{n_p}(i) - H^{(p)}(i) = \sum_{l=1}^{i} \hat{h}^{n_p}(l) - \sum_{l=1}^{i} h^{(p)}(l) = \sum_{l=1}^{i} [\hat{h}^{n_p}(l) - h^{(p)}(l)].$$

Pelo item (a) $\mathrm{E}[\hat{H}^{n_p}(1)] = H^{(p)}(1)$, o que implica que $\mathrm{E}[\hat{h}^{n_p}(1)] = h^{(p)}(1)$. Da mesma forma, $\mathrm{E}[\hat{H}^{n_p}(2)] = H^{(p)}(2)$ e, portanto,

$$\mathrm{E}\left[\sum_{l=1}^{2} \hat{h}^{n_p}(l)\right] = \sum_{l=1}^{2} h^{(p)}(l) \Rightarrow \sum_{l=1}^{2} \mathrm{E}\left[\hat{h}^{n_p}(l)\right] = \sum_{l=1}^{2} h^{(p)}(l) \Rightarrow \mathrm{E}\left[\hat{h}^{n_p}(2)\right] = \sum_{l=1}^{2} h^{(p)}(l)$$

Assumamos que vale para $k-1$, ou seja,

$$\mathrm{E}[\hat{H}^{n_p}(k-1)] = H^{(p)}(k-1).$$

Provando por indução sobre k, temos que $\mathrm{E}[\hat{H}^{n_p}(k)] = H^{(p)}(k)$, o que implica que

$$\mathrm{E}\left[\sum_{l=1}^{k} \hat{h}^{n_p}(l)\right] = \sum_{l=1}^{k} h^{(p)}(l) \Rightarrow \mathrm{E}\left[\sum_{l=1}^{k-1} \hat{h}^{n_p}(l)\right] + \mathrm{E}\left[\hat{h}^{n_p}(k)\right] = \sum_{l=1}^{k-1} h^{(p)}(l) + h^{(p)}(k)$$

e, portanto,

$$E[\hat{h}^{n_p}(k)] = h^{(p)}(k).$$

Assim, conluímos que

$$E[\hat{h}^{n_p}(l)] = h^{(p)}(l),\ l = 1, \ldots, k$$

ou seja, o estimador $\hat{h}^{n_p}$ é um estimador não-viciado para a função $h^{(p)}$.

□

Para dados censurados, $\hat{\zeta^{n_p}}$ é um estimador viciado para $\zeta^{(p)}$. Porém, se tivermos dados não censurados, obtemos

$$\begin{aligned}
\hat{\zeta}^{n_p}(i) &= \hat{h}^{n_p}(i)\prod_{l=1}^{i-1}[1-\hat{h}^{n_p}(l)] \\
&= \prod_{l=1}^{i}[1-\hat{h}^{n_p}(l)]^{V^{n_p}(l,.)-\Delta N^{n_p}(l,.)}[\hat{h}^{n_p}(l)]^{V^{n_p}(l,.)} \\
&= \frac{\Delta N^{(n_p)(i,.)}}{n_p}, \qquad (3.19)
\end{aligned}$$

que corresponde ao estimador empírico da probabilidade $\zeta^{(p)}$. Para este caso, podemos mostrar que o estimador é não-viciado.

Teorema 3.3. *Para dados não censurados, o estimador $\hat{\zeta^{n_p}}$ da função de distribuição de probabilidade $\zeta^{(p)}$ é um estimador não-viciado.*

Demonstração. Como visto anteriormente, $\Delta N^{n_p}(l,.) \sim$ Binomial $(n_p, \zeta^{(p)}(l))$. Consequentemente, temos que $E[\Delta N^{n_p}(l,.)] = n_p\zeta_l^{(p)}$. Utilizando esta informação e a equação 3.19, temos que

$$\mathrm{E}[\hat{\zeta}^{n_p}(l)] = \mathrm{E}\left[\frac{\Delta N^{n_p}(l,.)}{n_p}\right] = \zeta^{(p)}(l).$$

Portanto, o estimador $\hat{\zeta^{n_p}}$ da função de distribuição de probabilidade $\zeta^{(p)}$ é um estimador não-viciado.

□

Na seqüência são determinadas as variâncias dos estimadores $\hat{h}^{n_p}$ e $\hat{H}^{n_p}$. Para determinarmos a variância do estimador $\hat{H}^{n_p}$ utilizamos o processo variação quadrática associado ao martingale $\left(\frac{1}{V^{n_p}}.Y^{n_p}\right)$.

Teorema 3.4. *As variâncias dos estimadores $\hat{H}^{n_p}$ e $\hat{h}^{n_p}$ são dadas, respectivamente, por*

$$Var[\hat{H}^{n_p}(i)] = \sum_{l=1}^{i}[h^{(p)}(l) - {h^{(p)}}^2(l)]E\left[\frac{1}{V^{n_p}(l,.)}\right],\ i \in \{1,\ldots,k\} \tag{3.20}$$

e

$$Var[\hat{h}^{n_p}(l)] = [h^{(p)}(l) - {h^{(p)}}^2(l)]E\left[\frac{1}{V^{n_p}(l,.)}\right],\ l \leq i. \tag{3.21}$$

Demonstração. Utilizando o Lema 3.1, podemos escrever a diferença entre o estimador $\hat{H}^{n_p}(i)$ e a variável aleatória $H^{(p)}(i)$ como

$$\hat{H}^{n_p}(i,.) - H^{(p)}(i,.) = \left(\frac{J^{n_p}}{V^{n_p}}.Y^{n_p}\right)(i,.).$$

Utilizando esta igualdade, a definição de processo variação quadrática e a Proposição 2.3, obtemos que

$$\begin{aligned} E[\hat{H}^{n_p}(i,.) - H^{(p)}(i,.)]^2 &= E\left[\left(\frac{1}{V^{n_p}}.Y^{n_p}\right)(i,.)\right]^2 = E\left[<\left(\frac{1}{V^{n_p}}.Y^{n_p}\right);\left(\frac{1}{V^{n_p}}.Y^{n_p}\right)>\right] \\ &= E\left[\left(\left(\frac{1}{V^{n_p}}\right)^2.<Y^{n_p},Y^{n_p}>\right)(i,.)\right] \end{aligned}$$

No entanto, temos que

$$\begin{aligned} &\left(\left(\frac{1}{V^{n_p}}\right)^2.<Y^{n_p},Y^{n_p}>\right)(i,.) = \\ &= \sum_{l=1}^{i}\left(\frac{1}{V^{n_p}(l,.)}\right)^2\Delta<Y^{n_p},Y^{n_p}>(l,.) = \sum_{l=1}^{i}\left(\frac{1}{V^{n_p}(l,.)}\right)^2\Delta\sum_{m=1}^{n_p}<Y_m^{(p)},Y_m^{(p)}>(l,.) \\ &= \sum_{l=1}^{i}\left(\frac{1}{V^{n_p}(l,.)}\right)^2\sum_{m=1}^{n_p}[\Delta A_m^{(p)}(l,.) - \Delta {A_m^{(p)}}^2(l,.)] \\ &= \sum_{l=1}^{i}\frac{1}{{V^{n_p}}^2(l,.)}\sum_{m=1}^{n_p}\left[h^{(p)}(l)V_m^{(p)}(l,.) - {h^{(p)}}^2(l){V_m^{(p)}}^2(l,.)\right] \\ &= \sum_{l=1}^{i}\frac{1}{{V^{n_p}}^2(l,.)}[h^{(p)}(l)V^{n_p}(l,.) - {h^{(p)}}^2(l)V^{n_p}(l,.) \\ &= \sum_{l=1}^{i}\frac{1}{V^{n_p}(l,.)}[h^{(p)}(l) - {h^{(p)}}^2(l)] \end{aligned} \tag{3.22}$$

Assim, a variância do estimador $\hat{H}^{n_p}$ é dado por

$$\mathrm{Var}[\hat{H}^{n_p}(i)] = \sum_{l=1}^{i}[h^{(p)}(l) - {h^{(p)}}^2(l)]\mathrm{E}\left[\frac{1}{V^{n_p}(l,.)}\right]. \tag{3.23}$$

Para obtermos a variância do estimador $\hat{h}^{n_p}$ consideremos o processo $S^{n_p}(l)$ tal que $S^{n_p}(l) = \hat{H}^{n_p}(l) - H^{(p)}(l)$. Observe que $S^{n_p}(l)$ é um martingale, $l \leq i$. Além disso, temos que

$$E[\hat{h}^{n_p}(l) - h^{(p)}(l)]^2 = E[S^{n_p}(l) - S^{n_p}(l-1)]^2.$$

Por outro lado, temos que

$$\begin{aligned} E[S^{n_p}(l) - S^{n_p}(l-1)]^2 &= E\left[\left(\frac{1}{V^{n_p}}.Y^{n_p}\right)(l,.) - \left(\frac{1}{V^{n_p}}.Y^{n_p}\right)(l-1,.)\right] \\ &= [h^{(p)}(l) - {h^{(p)}}^2(l)]\mathrm{E}\left[\frac{1}{V^{n_p}(l,.)}\right]. \end{aligned}$$

Portanto,

$$\mathrm{Var}[\hat{h}^{n_p}(l)] = [h^{(p)}(l) - {h^{(p)}}^2(l)]\mathrm{E}\left[\frac{1}{V^{n_p}(l,.)}\right].$$

□

Proposição 3.1. *Consideremos as variáveis aleatórias* $V_m^{(p)}(l,.) = \mathbb{1}_{\{X_m^{(p)} \geq l\}}$ *e* $V^{n_p}(l,.) = \sum_{m=1}^{n_p} V_m^{(p)}(l,.)$. *Denotamos por*

$$\theta^{(p)}(l) = \eta^{(p)}(l) + \eta^{(p)}(l+1) + \cdots + \eta^{(p)}(k+1)$$

$\theta^{(p)}(l) > 0, \forall l, p.$

(a) A variável aleatória $V_m^{(p)}(l,.)$ *tem distribuição de Bernoulli com parâmetro* $\theta^{(p)}(l)$.

(b) A variável aleatória $V^{n_p}(l,.)$ *tem distribuinção Binomial com parâmetros* n_p *e* $\theta^{(p)}(l)$.

Demonstração. (a) A variável aleatória $V_m^{(p)}$ pode assumir apenas dois valores, 0 ou 1. A probabilidade dele assumir o valor 1 é

$$\begin{aligned} P[V_m^{(p)}(l,.) = 1] &= P[X_m^{(p)} \geq l] \\ &= 1 - (\eta^{(p)}(1) + \eta^{(p)}(2) + \cdots + \eta^{(p)}(l-1)) > 0 \\ &= \theta^{n_p}(l) \end{aligned}$$

ou seja, o processo $V_m^{(p)}(l,.)$ tem distribuição de Bernoulli com parâmetro $\theta_l^{(p)}$.

(b) Como $V^{n_p}(l,.) = \sum_{m=1}^{n_p} V_m^{(p)}(l,.)$ obtemos do item anterior que

$$V^{n_p}(l,.) \sim \text{Binomial } (n_p, \theta^{(p)}(l)).$$

□

Utilizando este resultado, será estudada a distribuição exata do estimador $\hat{h}^{n_p}$. Como a distribuição condicional de $\Delta N^{n_p}(i,.)/V^{n_p}(i,.)$ é uma Binomial com parâmetros $V^{n_p}(i,.)$ e $h^{(p)}(i)$, concluímos que para todo $0 < a \leq b \leq n$,

$$\begin{aligned} P[\Delta N^{(n_p)}(i,.) = a; V^{n_p}(i,.) = b] &= P[\Delta N^{n_p}(i,.) = a/V^{n_p}(i,.) = b]P[V^{n_p}(i,.) = b] \\ &= \begin{pmatrix} b \\ a \end{pmatrix} \left(h^{(p)}(i)\right)^a \left(1 - h^{(p)}(i)\right)^{b-a} \begin{pmatrix} n \\ b \end{pmatrix} (\theta^{(p)}(i))^b [1 - \theta^{(p)}(i)]^{n-b} \end{aligned}$$

Com isso, a distribuição exata do estimador $\hat{h}^{n_p}(i)$ é obtida por

$$P\left[\hat{h}^{n_p} = \frac{\Delta N^{n_p}(i,.)}{V^{n_p}(i,.)} = \frac{a}{b}\right] = P[\Delta N^{n_p}(i,.) = a; V^{n_p}(i,.)]. \tag{3.24}$$

3.3.2 Propriedades assintóticas

Nesta seção, vamos estudar algumas propriedades assintóticas relacionadas ao estimador $\hat{H}^{n_p}$ do processo estocástico H^{n_p} e, conseqüetemente, do estimador de *Kaplan-Meier*. Primeiramente, vamos verificar alguns resultados necessários para o estudo das propriedades assintóticas dos estimadores. Depois, iremos mostrar a consistência do estimador.

Teorema 3.5. *Dada uma seqüência $\hat{H}^{n_p}$ de estimadores da função intensidade $H^{(p)}$ temos que*

$$P\left[\sup_{l \leq i} |\hat{H}^{n_p}(l) - H^{(p)}(l)|\right] \to 0,$$

ou seja, $\hat{H}^{n_p}$ é um estimador consistente para a função $H^{(p)}$.

Demonstração. Pela desigualdade de Doob (Teorema 1, pp. 464 de Shiryayev (1984)) e desigualdade de Burkholder (Shiryayev (1984), pp. 469) obtemos que

$$P\left[\sup_{l \leq i} |\hat{H}^{n_p}(l) - H^{(p)}(l)|\right] \leq \frac{c}{\lambda^2} \sum_{l=1}^{i} [h^{(p)}(l) - h^{(p)^2}(l)] \mathrm{E}\left[\frac{1}{V^{n_p}(l,.)}\right]$$

Por outro lado, pela Proposição 3.1 temos que $V_m^{(p)}(l,.) \sim$ Bernoulli $(\theta^{n_p}(l))$. Definamos o conjunto $S^{n_p} = \{V_m^{(p)}(l,.) = 1\}$. Como $P[S^{n_p}] = \theta_l^{n_p}$,

$$\sum_{n_p=1}^{\infty} P[S^{n_p}] = +\infty,$$

para todo $p = 1, \ldots J$. Além disso, $S^{n_1}, \cdots, S^{n_p}$ são independentes. Logo, pelo Lema de Borel-Cantelli, temos que

$$P\left[\limsup_{n_p} S^{n_p}\right] = 1,$$

ou seja, a probabilidade de ocorrer S^{n_p} é 1, exceto num número finito de pontos. Portanto,

$$V^{n_p}(l,.) = \sum_{m=1}^{n_p} V_m^{(p)}(l,.) \to +\infty, \quad P - q.c., \quad \text{quando } n_p \to \infty.$$

Conseqüentemente, $\frac{1}{V^{n_p}(l,.)}$ converge quase certamente para zero. Portanto, pelo Teorema da Convergência Monótona, $E\left[\frac{1}{V^{n_p}(l,.)}\right] \to 0$. Assim, temos que $\hat{H}^{n_p}$ é um estimador consistente para a função intensidade acumulada $H^{(p)}$. □

Teorema 3.6. *Dada uma seqüência $\hat{h}^{n_p}$ de estimadores da função intensidade $h^{(p)}$ temos que*

$$P\left[|\hat{h}^{n_p}(l) - h^{(p)}(l)|\right] \to 0,$$

ou seja, $\hat{h}^{n_p}$ é um estimador consistente para a função $h^{(p)}$.

Demonstração. Consideremos $l = 1$. Neste caso, temos que

$$|\hat{H}^{n_p}(1) - H^{(p)}(1)| = |\hat{h}^{n_p}(1) - h^{(p)}(1)|$$

e portanto, pelo Teorema 3.5, concluímos que

$$P\left[|\hat{h}^{n_p}(1) - h^{(p)}(1)|\right] \to 0.$$

Por outro lado temos que

$$\begin{aligned} \hat{H}^{n_p}(i) - H^{(p)}(i) &= \sum_{l=1}^{i} \hat{h}^{n_p}(l) - \sum_{l=1}^{i} h^{(p)}(l) \\ &= \sum_{l=1}^{i-1} [\hat{h}^{n_p}(l) - h^{(p)}(l)] + [\hat{h}^{n_p}(i) - h^{(p)}(i)] \end{aligned}$$

$$\Rightarrow [\hat{h}^{n_p}(i) - h^{(p)}(i)] = [\hat{H}^{n_p}(i) - H^{(p)}(i)] - \sum_{l=1}^{i-1}[\hat{h}^{n_p}(l) - h^{(p)}(l)]$$

Portanto,

$$|\hat{h}^{n_p}(i) - h^{(p)}(i)| \leq |\hat{H}^{n_p}(i) - H^{(p)}(i)| + \sum_{l=1}^{i-1}|\hat{h}^{n_p}(l) - h^{(p)}(l)|$$

Assumindo que vale para $i-1$ e utilizando o Teorema 3.5, concluímos por indução que vale para i. Portanto,

$$P\left[|\hat{h}^{n_p}(l) - h^{(p)}(l)|\right] \to 0, l = 1, \ldots, i.$$

□

Teorema 3.7. *O estimador $\hat{\zeta}^{n_p}$ é um estimador consistente para a função $\zeta^{(p)}$.*

Demonstração. Pela equação 3.14 temos que o estimador $\hat{\zeta}^{n_p}$ é função contínua do estimador da função intensidade, $\hat{h}^{n_p}$. Como $\hat{h}^{n_p}$ é estimador consistente para $h^{(p)}$, concluímos que $\hat{\zeta}^{n_p}$ é estimador consistente para $\zeta^{(p)}$. □

Capítulo 4

Distribuição Assintótica de Integrais Estocásticas

Nesta seção desenvolvemos um teorema do limite central para as integrais estocásticas associadas ao processo de contagem N^{n_p}. Essa teoria foi baseada nos resultados de Helland (1982). Esses resultados são importantes para que possamos desenvolver os testes de hipótese.

Para que possamos desenvolver o teorema do limite central, precisamos definir uma filtragem em relação ao número de itens (ou indivíduos), pois o limite é tomado neste número.

4.1 Amostra Única

Consideremos uma amostra aleatória simples

$$((W_1^{(1)}, C_1^{(1)}), \ldots, (W_{n_1}^{(1)}, C_{n_1}^{(1)})), \tag{4.1}$$

com as mesmas características da amostra 3.7 definida na Seção 3.3. Para $n_1 \in \mathbb{N}$ e $l \in \{1, \ldots, i\}$ fixos, com $i \in \{1, \ldots, k\}$ e $k \in \mathbb{N}$, definimos a álgebra

$$\vartheta_l(0) = \text{álgebra gerada por}\{V^{n_1}(l,.); V_1^{(1)}(l,.)\},$$

$$\vartheta_l(q) = \text{álgebra gerada por}\{V^{n_1}(l,.); V_1^{(1)}(l,.), \ldots, V_{q+1}^{(1)}(l,.); \Delta N_1^{(1)}(l,.), \ldots, \Delta N_q^{(1)}(l,.); C_1^{(1)}, \ldots, C_{n_1}^{(1)}\}$$

para $q \in \{1, \ldots, n_1 - 1\}$ fixo e

$$\vartheta_l(n_1) = \text{álgebra gerada por}\{V^{n_1}(l,.); V_1^{(1)}(l,.), \ldots, V_{n_1}^{(1)}(l,.); \Delta N_1^{(1)}(l,.), \ldots, \Delta N_{n_1}^{(1)}(l,.); C_1^{(1)}, \ldots, C_{n_1}^{(1)}\}$$

Definimos a álgebra $\gamma_i(j)$, para $0 \leq j \leq n_1$, como sendo o acumulado da álgebra $\vartheta_l(j)$ até a categoria i. Assim,

$$\gamma_i(0) = \text{álgebra gerada por } \{\vartheta_1(0), \vartheta_2(0), \ldots, \vartheta_i(0)\}$$

$$\gamma_i(q) = \text{álgebra gerada por } \{\vartheta_1(q), \vartheta_2(q), \ldots, \vartheta_i(q)\}$$

para $q = 1, \ldots, n_1 - 1$ fixo e

$$\gamma_i(n_1) = \text{álgebra gerada por } \{\vartheta_1(n_1), \vartheta_2(n_1), \ldots, \vartheta_i(n_1)\}.$$

Consideremos $\{\gamma_i\}$ a filtragem formada pelas álgebras encaixantes $\gamma_i(j)$, $0 \leq j \leq n_1$. Seja $U^{n_1}(l,.)$ uma variável aleatória com respeito à álgebra $\gamma_i(0)$ para $l = 1, \ldots i$ fixo, de tal forma que as seguintes propriedades são válidas

(a) Para n_1 fixo o processo U^{n_1} é previsível na filtragem $\{\mathcal{B}^{N^{n_1}}\}$.

(b) $\mathrm{E}[|U^{n_1}(l,.)|] \to 0$ quando $n_1 \to \infty$, $\forall l = 1, \ldots, i$.

(c) $\mathrm{E}\left[U^{n_1 2}(l,.)V^{n_1}(l,.)\right] \to G_l$ quando $n_1 \to \infty$, para algum $G_l > 0$ e $\forall l = 1, \ldots, i$.

Seja $a_m = \sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y_m^{(1)}(l,.)$, para $0 \leq m \leq n_1$. Definimos a somas parciais $R_j(i) = \sum_{m=1}^{j} a_m$, para $0 \leq j \leq n_1$ e $i \in \{1, \ldots, k\}$ fixo, com $R_0(i) = 0$. Definimos o processo estocástico

$$R(i) = \{R_0(i), R_1(i), \ldots, R_{n_1}(i)\} \tag{4.2}$$

que é adaptado à fitragem $\{\gamma_i\}$ por construção.

Proposição 4.1. *O processo estocástico $R(i)$, para $i \in \{1, \ldots, k\}$ fixo, é um martingale com respeito à filtragem $\{\gamma_i\}$.*

Demonstração. O processo $R(i)$ é adaptado à filtragem $\{\gamma_i\}$. Para demonstrar que esse processo é um martingale, basta provar que $\mathrm{E}[R_j(i)|\gamma_i(j-1)] = R_{j-1}(i)$, conforme a Definição 2.3, que é o mesmo que mostrar que $\mathrm{E}[R_j(i) - R_{j-1}|\gamma_i(j-1)] = 0$. Mas observe que

$$R_j(i) - R_{j-1}(i) = \sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y_j^{(1)}(l,.) = a_j(i).$$

Portanto, basta mostrar que

$$E\left[a_j(i)|\gamma_i(j-1)\right]=0$$

para $i \in \{1,\ldots,k\}$ fixo e $j=1,\ldots,n_1$. Inicialmente, temos que

$$\begin{aligned}
&E\left[\sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y_j^{(1)}(l,.)|\gamma_i(j-1)\right] \\
&\qquad = \sum_{l=1}^{i} U^{n_1}(l,.)\mathrm{E}\left[\Delta N_j^{(1)}(l.)-\Delta A_j^{(1)}(l.)|\gamma_i(j-1)\right] \\
&\qquad = \sum_{l=1}^{i} U^{n_1}(l,.)\left\{\mathrm{E}\left[\Delta N_j^{(1)}(l.)|\gamma_i(j-1)\right]+\mathrm{E}\left[\Delta A_j^{(1)}(l,.)|\gamma_i(j-1)\right]\right\}
\end{aligned}$$

Observe que $\Delta A_j^{(1)}(l,.)$ é mensurável em $\gamma_i(j-1)$ para $j=1,\ldots,n_1$ e $l=1,\ldots,i$. Além disso, temos ques $\Delta N_j^{(1)}(l,.)$ condicionado a $V_j^{(1)}(l,.)$ é independente de todos os outros termos da filtragem, $\gamma_i(j-1)$, também para $j=1,\ldots,n_1$ e $l=1,\ldots,i$. Com isso,

$$\begin{aligned}
&E\left[\sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y_j^{(1)}(l,.)|\gamma_i(j-1)\right] \\
&\qquad = \sum_{l=1}^{i} U^{n_1}(l,.)\left\{\mathrm{E}\left[\Delta N_j^{(1)}(l.)|V_j^{(1)}(l,.)\right]-\Delta A_j^{(1)}(l,.)\right\} \\
&\qquad = \sum_{l=1}^{i} U^{n_1}(l,.)\left\{\mathrm{E}\left[\Delta N_j^{(1)}(l.)|V_j^{(1)}(l,.)\right]-V_j^{(1)}(l,.)h^{(1)}(l)\right\} \\
&\qquad = 0
\end{aligned}$$

pois $\Delta N_j^{(1)}(l,.)|V_j^{(1)}(l,.)$ tem distribuição binomial com número de ensaios $V_j^{(1)}(l,.)$ e probabilidade $h^{(1)}(l)$, para $j=1,\ldots,n_1$ e $l=1,\ldots,i$. Assim, concluímos que $R(i)$ é um martingale em relação a filtragem $\{\gamma_i\}$. □

Da Proposição 2.2, concluímos que o processo $\{a_j(i), 0 \leq j \leq n_1\}$ é um martingale *array difference* em relaçao a filtragem $\{\gamma_i\}$.

Lema 4.1. *A seguinte propriedade é válida para os termos do processo $R(i)$*

$$E\left[\max_{1\leq j\leq n_1}|a_j(i)|\right]\to 0, \quad \textit{quando } n_1\to\infty. \tag{4.3}$$

Demonstração. Por propriedades de máximo e módulo temos que

$$\begin{aligned}\max_{1\le j\le n_1}|a_j(i)| = \max_{1\le j\le n_1}\left|\sum_{l=1}^{i}U^{n_1}(l,.)\Delta Y_j^{(1)}(l.)\right| &\le \max_{1\le j\le n_1}\sum_{l=1}^{i}\left|U^{n_1}(l,.)\Delta Y_j^{(1)}(l.)\right| \\ &\le \sum_{l=1}^{i}\max_{1\le j\le n_1}\left|U^{n_1}(l,.)\Delta Y_j^{(1)}(l.)\right|\end{aligned}$$

Com isso,

$$\mathrm{E}\left[\max_{1\le j\le n_1}\left|\sum_{l=1}^{i}U^{n_1}(l,.)\Delta Y_j^{(1)}(l.)\right|\right] \le \mathrm{E}\left[\sum_{l=1}^{i}\max_{1\le j\le n_1}\left|U^{n_1}(l,.)\Delta Y_j^{(1)}(l.)\right|\right]. \tag{4.4}$$

Por outro lado, como $\Delta N_j^{(1)}(l,.) \le 1$ e $A_j^{(1)}(l,.) \le 1$, temos que

$$\begin{aligned}\left|U^{n_1}(l,.)\Delta Y_j^{(1)}(l.)\right| &= \left|U^{n_1}(l,.)[\Delta N_j^{(1)}(l,.) - \Delta A_j^{(1)}(l,.)]\right| \\ &\le \left|U^{n_1}(l,.)\Delta N_j^{(1)}(l,.)\right| + \left|U^{n_1}(l,.)\Delta A_j^{(1)}(l,.)\right| \\ &\le |U^{n_1}(l,.)| + |U^{n_1}(l,.)| = 2|U^{n_1}(l,.)|\end{aligned} \tag{4.5}$$

Das equações 4.4 e 4.5, obtemos que

$$\mathrm{E}\left[\max_{1\le j\le n_1}\left|\sum_{l=1}^{i}U^{n_1}(l,.)\Delta Y_j^{(1)}(l.)\right|\right] \le \mathrm{E}\left[\max_{1\le j\le n_1}2iU^{n_1}(l,.)\right] = 2i\mathrm{E}[U^{n_1}(l,.)] \tag{4.6}$$

para $i \in \{1,\ldots,k\}$ fixo, com $k \in \mathbb{N}$. Por hipóteste $E[|U^{n_1}(l,.)|] \to 0$ quando $n_1 \to \infty$, $\forall l$. Portanto,

$$\mathrm{E}\left[\max_{1\le j\le n_1}\left|\sum_{l=1}^{i}U^{n_1}(l,.)\Delta Y_j^{(1)}(l.)\right|\right] \to 0, \text{quando } n_1 \to \infty. \tag{4.7}$$

□

Lema 4.2. *Seja* $\sigma^2(i) = \sum_{l=1}^{i}[h^{(1)}(l) - {h^{(1)}}^2(l)]G_l$. *Então,*

$$\sum_{j=1}^{n_1}[a_j(i)]^2 \to \sigma^2(i)$$

Demonstração. Pela preposição 4.1, o processo $R(i)$ é um martingale e, portanto, pela Proposição 2.1 temos que

$$\mathrm{E}[{R_j}^2(i)] - \mathrm{E}[{R_0}^2(i)] = \sum_{j=0}^{n_1-1}\mathrm{E}[(\Delta R_{j+1}(i))^2]. \tag{4.8}$$

Os termos do processo $R(i)$ podem ser escritos da seguinte forma $R_j(i) = \sum_{m=1}^{j} \Delta R_m(i)$. Assim, podemos reescrever a equação 4.8

$$\mathrm{E}\left[\left(\sum_{j=1}^{n_1} \Delta R_j(i)\right)^2\right] = \sum_{j=0}^{n_1-1} \mathrm{E}\left[(\Delta R_{j+1}(i))^2\right] = \sum_{j=1}^{n_1} \mathrm{E}\left[(\Delta R_j(i))^2\right].$$

Por outro lado, observe que $\Delta R_j(i) = a_j(i)$, para todo $0 \leq j \leq n_1$, e, como $a_j(i) = \sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y_j^{(1)}(l,.)$, obtemos que

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{n_1}[a_j(i)]^2\right] &= \mathrm{E}\left[\left(\sum_{j=1}^{n_1} a_j(i)\right)^2\right] \\
&= E\left[\left(\sum_{j=1}^{n_1}\sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y_j^{(1)}(l,.)\right)^2\right] \\
&= E\left[\left(\sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y^{n_1}(l,.)\right)^2\right]
\end{aligned}$$

Utilizando a definição de integral estocástica e a Proposição 2.3, temos que

$$\begin{aligned}
E\left[\left(\sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y^{n_1}(l,.)\right)^2\right] &= E\left[(U^{n_1}.Y^{n_1})^2(i,.)\right] \\
&= E[< (U^{n_1}.Y^{n_1}); (U^{n_1}.Y^{n_1}) > (i,.)] \\
&= E\left[(U^{n_1 2}. < Y^{n_1}, Y^{n_1} >)(i,.)\right].
\end{aligned}$$

Utilizando a equação 2.9,

$$\begin{aligned}
(U^{n_1 2}. < Y^{n_1}, Y^{n_1} >)(i,.) &= \sum_{l=1}^{i} U^{n_1 2}(l,.)\Delta < Y^{n_1}, Y^{n_1} > (l,.) \\
&= \sum_{l=1}^{i} U^{n_1 2}(l,.)\Delta \sum_{j=1}^{n_1} < Y_j^{(1)}, Y_j^{(1)} > (l,.) \\
&= \sum_{l=1}^{i} U^{n_1 2}(l,.) \sum_{j=1}^{n_1} \left[\Delta A_j^{(1)}(l,.) - \Delta A_j^{(1)^2}(l,.)\right] \\
&= \sum_{l=1}^{i} U^{n_1 2}(l,.) \sum_{j=1}^{n_1} \left[V_j^{(1)}(l,.)h^{(1)}(l) - V_j^{(1)^2}(l,.)h^{(1)^2}(l)\right] \\
&= \sum_{l=1}^{i} U^{n_1 2}(l,.) \left[h^{(1)}(l,.)V^{n_1}(l,.) - h^{(1)^2}(l,.)V^{n_1}(l,.)\right] \\
&= \sum_{l=1}^{i} U^{n_1 2}(l,.)V^{n_1}(l,.)h^{(1)}(l)[1 - h^{(1)}(l)]
\end{aligned}$$

Assim,

$$\begin{aligned}
E\left[(U^{n_1 2}. < Y^{n_1}, Y^{n_1} >)(i,.)\right] &= E\left[\sum_{l=1}^{i} U^{n_1 2}(l,.)V^{n_1}(l,.)h^{(1)}(l)[1 - h^{(1)}(l)]\right] \\
&= \sum_{l=1}^{i} h^{(1)}(l)[1 - h^{(1)}(l)]E\left[U^{n_1 2}(l,.)V^{n_1}(l,.)\right]
\end{aligned}$$

Se G_l denota o limite de $E\left[U^{n_1 2}V^{n_1}(l,.)\right]$ quando $n_1 \to \infty$, para $l = 1, \ldots, i$, então obtemos que

$$E\left[\sum_{j=1}^{n_1} [a_j(i)]^2\right] \to \sum_{l=1}^{i} [h^{(1)}(l) - h^{(1)^2}(l)]G_l, \quad \text{quando} \quad n_1 \to \infty. \tag{4.9}$$

Desde que o processo $R(i)$ é um martingale uniformemente integrável concluímos que

$$\sum_{j=1}^{n_1} [a_j(i)]^2 \to \sum_{l=1}^{i} [h^{(1)}(l) - h^{(1)^2}(l)]G_l \quad \text{quando} \quad n_1 \to \infty.$$

□

Teorema 4.1. *Temos que*

$$R_{n_1}(i) = \sum_{j=1}^{n_1} a_j(i) = \sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y^{n_1}(l,.)$$

converge em distribuição para uma variávela aleatória normal com média 0 e variância $\sigma^2(i)$.

Demonstração. Pela Proposição 4.1 concluímos que o processo $\{a_j(i) : 0 \leq j \leq n_1\}$ é um martingale *array difference*. Além disso, utilizando os Lemas 4.1 e 4.2, concluímos pelo Teorema 2.5, item (b), de Helland (1982) que $R_{n_1}(i)$ converge em distribuição para uma variável aleatória normal com média 0 e variância $\sigma^2(i)$. □

Corolário 4.1. *Ao denotarmos*

$$U^{n_1}(l,.) = \begin{cases} \dfrac{\sqrt{n_1}}{V^{n_1}(s,.)}, & l = s; \\ 0, & l \neq s \end{cases}$$

para $l, s = 1, \ldots, i$, *obtemos que* $\sqrt{n_1}[\hat{h}^{n_1}(s) - h^{(1)}(s)]$ *converge em distribuição para uma variável aleatória normal com média 0 e variância*

$$\sigma_h^2(s) = \frac{[h^{(1)}(s) - {h^{(1)}}^2(s)]}{\theta^{(1)}(s)}. \tag{4.10}$$

Demonstração. Para $U^{n_1}(s,.) = \dfrac{\sqrt{n_1}}{V^{n_1}(s,.)}$ e $U^{n_1}(l,.) = 0$ para todo $l \neq s$, o processo $R_{n_1}(s)$ é dado por

$$\begin{aligned} R_{n_1}(s) = U^{n_1}(s,.)\Delta Y^{n_1}(s,.) &= \frac{\sqrt{n_1}}{V^{n_1}(s,.)}\Delta Y^{n_1}(s,.) = \frac{\sqrt{n_1}}{V^{n_1}(s,.)}[\Delta N^{n_1}(s,.) - \Delta A^{n_1}(s,.)] \\ &= \frac{\sqrt{n_1}}{V^{n_1}(s,.)}[\Delta N^{n_1}(s,.) - V^{n_1}(s,.)h^{(1)}(l)] \\ &= \sqrt{n_1}\left[\frac{\Delta N^{n_1}(s,.)}{V^{n_1}(s,.)} - h^{(1)}(s)\right] \\ &= \sqrt{n_1}[\hat{h}^{n_1}(s) - h^{(1)}(s)] \end{aligned}$$

Além disso,

$$E[{U^{n_1}}^2(s,.)V^{n_1}(s,.)] = E\left[\frac{n_1}{V^{n_1}(s,.)}\right] \to \frac{1}{\theta_s^{(1)}}.$$

Portanto, pelo Teorema 4.1 temos que

$$\sqrt{n_1}[\hat{h}^{n_1}(s) - h^{(1)}(s)]$$

converge em distribuição para uma variável aleatória com média 0 e variância $\dfrac{[h^{(1)}(s) - {h^{(1)}}^2(s)]}{\theta^{(1)}(s)}$.

□

Corolário 4.2. *Tomando* $U^{n_1}(l,.) = \dfrac{\sqrt{n_1}}{V^{n_1}(l,.)}$, *para* $l = 1, \ldots, i$, *obtemos que* $\sqrt{n_1}[\hat{H}^{n_1}(l) - H^{(1)}(l)]$ *converge em distribuição para uma variável aleatória com média 0 e variância*

$$\sigma_H^2(i) = \sum_{l=1}^{i} \frac{[h^{(1)}(l) - {h^{(1)}}^2(l)]}{\theta^{(1)}(l)}. \tag{4.11}$$

Demonstração. Para $U^{n_1}(l,.) = \dfrac{\sqrt{n_1}}{V^{n_1}(s,.)}$ para todo $l = 1, \ldots, i$, o processo $R_{n_1}(i)$ é dado por

$$\begin{aligned}
R_{n_1}(i) = \sum_{l=1}^{i} U^{n_1}(l,.)\Delta Y^{n_1}(l,.) &= \sum_{l=1}^{i} \frac{\sqrt{n_1}}{V^{n_1}(l,.)}\Delta Y^{n_1}(l,.) = \sum_{l=1}^{i} \frac{\sqrt{n1}}{V^{n_1}(l,.)}[\Delta N^{n_1}(l,.) - \Delta A^{n_1}(l,.)] \\
&= \sum_{l=1}^{i} \frac{\sqrt{n_1}}{V^{n_1}(l,.)}[\Delta N^{n_1}(l,.) - V^{n_1}(l,.)h^{(1)}(l)] \\
&= \sum_{l=1}^{i} \sqrt{n_1}\left[\frac{\Delta N^{n_1}(l,.)}{V^{n_1}(l,.)} - h^{(1)}(l)\right] \\
&= \sum_{l=1}^{i} \sqrt{n_1}[\hat{h}^{n_1}(l) - h^{(1)}(l)] \\
&= \sqrt{n_1}[\hat{H}^{n_1}(i) - H^{(1)}(i)]
\end{aligned}$$

Além disso,

$$E[{U^{n_1}}^2(l,.)V^{n_1}(l,.)] = E\left[\frac{n_1}{V^{n_1}(l,.)}\right] \to \frac{1}{\theta^{(1)}(l)}, \text{ quando } n_1 \to \infty, \ \forall l = 1, \ldots, i.$$

Portanto, pelo Teorema 4.1 temos que

$$\sqrt{n_1}[\hat{H}^{n_1}(i) - H^{(1)}(i)]$$

converge em distribuição para uma variável aleatória com média 0 e variância $\displaystyle\sum_{l=1}^{i} \frac{[h^{(1)}(l) - {h^{(1)}}^2(l)]}{\theta^{(1)}(l)}$.

□

4.2 Duas Amostras

Consideremos duas amostras aleatórias simples independentes de tamanho n_1 e n_2, dadas, respectivamente, por $((W_1^{(1)}, C_1^{(1)}), \ldots, (W_{n_p}^{(1)}, C_{n_1}^{(1)}))$ e $((W_1^{(2)}, C_1^{(2)}), \ldots, (W_{n2}^{(2)}, C_{n2}^{(2)}))$.

A seguir, iremos desenvolver o processo e a filtragem correspondente utilizados no estudo de convergência. O processo será construído termo a termo. A filtragem será formada por álgebras encaixantes construídas conforme é construído cada termo do processo.

Suponha, sem perda de generalidade, que $n_1 \geq n_2$ e considere $i \in \{1, \ldots, k\}, k \in \mathbb{N}$, fixo. No ponto inicial (zero), a álgebra é dada por

$$\begin{aligned}\tau_i(0) = \text{álgebra gerada por}\{ \quad & V^{n_1}(1,.) \quad , V^{n_2}(1,.), V_1^{(1)}(1,.), V_1^{(2)}(1,.); \ldots; \\ & V^{n_1}(i,.) \quad , V^{n_2}(i,.), V_1^{(1)}(i,.), V_1^{(2)}(i,.)\}\end{aligned}$$

e o termo correspondente ao ponto zero é nulo ($a_0 = 0$). Assim, consideraremos a_1 como sendo o primeiro termo do processo que será construído. Seja $U^{n_1,n_2}(l,.)$, $l = 1, \ldots, i$, uma variável aleatória com respeito à álgebra $\tau_i(0)$, com as seguintes propriedades

(a) Para n_1 e n_2 fixos o processo U^{n_1,n_2} é previsível na filtragem produto $\{\mathcal{B}^{N^{n_1}}\} \times \{\mathcal{B}^{N^{n_2}}\}$.

(b) $\mathrm{E}\left[\left|\dfrac{U^{n_1,n_2}(l,.)}{V^{n_p}(l,.)}\right|\right] \to 0$, quando $n_1, n_2 \to \infty$, para $p = 1, 2$ e $l = 1, \ldots, i$.

(c) $\mathrm{E}\left[\dfrac{U^{n_1,n_2}{}^2(l,.)}{V^{n_p}(l,.)}\right] \to G_l^{(p)}$, quando $n_1, n_2 \to \infty$, para $G^{(p)} > 0$, $p = 1, 2$, $l = 1, \ldots, i$.

De forma geral, assumimos que

$$\frac{n_1 + n_2}{n_p} \to u_p, \tag{4.12}$$

para $p = 1, 2$ e $u_p \in \mathbb{R}$, quando $n_1, n_2 \to \infty$. Definimos o primeiro termo do processo

$$a_1 = \sum_{l=1}^{i} U^{n_1,n_2}(l,.) \left[\frac{\Delta Y_1^{(1)}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y_1^{(2)}(l,.)}{V^{n_2}(l,.)}\right]$$

e a álgebra nesse ponto é

$$\begin{aligned}\tau_i(1) = \text{álgebra gerada por}\{ \quad & V^{n_1}(1,.), V^{n_2}(1,.), V_1^{(1)}(1,.), V_2^{(1)}(1,.), V_1^{(2)}(1,.), V_2^{(2)}(1,.), \\ & \Delta N_1^{(1)}(1,.), \Delta N_1^{(2)}(1,.); \ldots; V^{n_1}(i,.), V^{n_2}(i,.), V_1^{(1)}(i,.), \\ & V_2^{(1)}(i,.), V_1^{(2)}(i,.), V_2^{(2)}(i,.), \Delta N_1^{(1)}(i,.), \Delta N_1^{(2)}(i,.); \\ & C_1^{(1)}, \ldots, C_{n_1}^{(1)}, C_1^{(2)}, \ldots, C_{n_2}^{(2)}\}.\end{aligned}$$

O segundo termo é

$$a_2 = \sum_{l=1}^{i} U^{n_1,n_2}(l,.) \left[\frac{\Delta Y_2^{(1)}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y_2^{(2)}(l,.)}{V^{n_2}(l,.)}\right]$$

e a álgebra nesse ponto é

$$\begin{aligned}\tau_i(2) = \text{álgebra gerada por}\{ \quad & V^{n_1}(1,.), V^{n_2}(1,.), V_1^{(1)}(1,.), V_2^{(1)}(1,.), V_3^{(1)}(1,.), V_1^{(2)}(1,.),\\ & V_2^{(2)}(1,.), V_3^{(2)}(1,.), \Delta N_1^{(1)}(1,.), \Delta N_2^{(1)}(1,.), \Delta N_1^{(2)}(1,.), \Delta N_2^{(2)}(1,.); \ldots;\\ & V^{n_1}(i,.), V^{n_2}(i,.), V_1^{(1)}(i,.), V_2^{(1)}(i,.), V_3^{(1)}(i,.), V_1^{(2)}(i,.), V_2^{(2)}(i,.),\\ & V_3^{(2)}(i,.), \Delta N_1^{(1)}(i,.), \Delta N_2^{(1)}(i,.), \Delta N_1^{(2)}(i,.), \Delta N_2^{(2)}(i,.);\\ & C_1^{(1)}, \ldots, C_{n_1}^{(1)}, C_1^{(2)}, \ldots, C_{n_2}^{(2)}\}.\end{aligned}$$

Definimos de forma análoga todos os termos até o termo de número n_2. Assim, para $s = 1, \ldots, n_2$, os termos do processo são definidos conforme a equação 4.13

$$a_s = \sum_{l=1}^{i} U^{n_1,n_2}(l,.) \left[\frac{\Delta Y_s^{(1)}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y_s^{(2)}(l,.)}{V^{n_2}(l,.)} \right] \tag{4.13}$$

A álgebra também é definida analogamente até o ponto n_2-1. Portanto, para $s = 1, \ldots, n_2-1$,

$$\begin{aligned}\tau_i(s) = \text{álgebra gerada por}\{ \quad & V^{n_1}(1,.), V^{n_2}(1,.), V_1^{(1)}(1,.), \ldots, V_{s+1}^{(1)}(1,.), V_1^{(2)}(1,.), \ldots, V_{s+1}^{(2)}(1,.),\\ & \Delta N_1^{(1)}(1,.), \ldots, \Delta N_s^{(1)}(1,.), \Delta N_1^{(2)}(1,.), \ldots, \Delta N_s^{(2)}(1,.); \ldots;\\ & V^{n_1}(i,.), V^{n_2}(i,.), V_1^{(1)}(i,.), \ldots, V_{s+1}^{(1)}(i,.), V_1^{(2)}(i,.), \ldots, V_{s+1}^{(2)}(i,.),\\ & \Delta N_1^{(1)}(i,.), \ldots, \Delta N_s^{(1)}(i,.), \Delta N_1^{(2)}(i,.), \ldots, \Delta N_s^{(2)}(i,.);\\ & C_1^{(1)}, \ldots, C_{n_1}^{(1)}, C_1^{(2)}, \ldots, C_{n_2}^{(2)}\}\end{aligned}$$

No entanto, no ponto n_2, a álgebra é definida de forma diferente

$$\begin{aligned}\tau_i(n_2) = \text{álgebra gerada por}\{ \quad & V^{n_1}(1,.), V^{n_2}(1,.), V_1^{(1)}(1,.), \ldots, V_{n_2+1}^{(1)}(1,.), V_1^{(2)}(1,.), \ldots, V_{n_2}^{(2)}(1,.),\\ & \Delta N_1^{(1)}(1,.), \ldots, \Delta N_{n_2}^{(1)}(1,.), \Delta N_1^{(2)}(1,.), \ldots, \Delta N_{n_2}^{(2)}(1,.); \ldots;\\ & V^{n_1}(i,.), V^{n_2}(i,.), V_1^{(1)}(i,.), \ldots, V_{n_2+1}^{(1)}(i,.), V_1^{(2)}(i,.), \ldots, V_{n_2}^{(2)}(i,.),\\ & \Delta N_1^{(1)}(i,.), \ldots, \Delta N_{n_2}^{(1)}(i,.), \Delta N_1^{(2)}(i,.), \ldots, \Delta N_{n_2}^{(2)}(i,.);\\ & C_1^{(1)}, \ldots, C_{n_1}^{(1)}, C_1^{(2)}, \ldots, C_{n_2}^{(2)}\}.\end{aligned}$$

Como $n_1 > n_2$, a partir do ponto n_2+1 não existem elementos que correspondem à amostra $X^{(2)}$. Portanto, para $t = n_2+1, \ldots, n_1$,

$$a_t = \sum_{l=1}^{i} U^{n_1,n_2}(l,.) \left[\frac{\Delta Y_t^{(1)}(l,.)}{V^{n_1}(l,.)} \right].$$

Para $t = n_2+1, \ldots, n_1-1$, a álgebra é dada conforme a equação 4.14. A álgebra para o último termo difere das demais e é definida conforme a equação 4.16.

$$\begin{aligned}\tau_i(t) = \text{álgebra gerada por}\{ \quad & V^{n_1}(1,.), V^{n_2}(1,.), V_1^{(1)}(1,.), \ldots, V_{n_2}^{(1)}(1,.), V_1^{(2)}(1,.), \ldots, V_{t+1}^{(2)}(1,.), \\ & \Delta N_1^{(1)}(1,.), \ldots, \Delta N_{n_2}^{(1)}(1,.), \Delta N_1^{(2)}(1,.), \ldots, \Delta N_t^{(2)}(1,.); \ldots; \\ & V^{n_1}(i,.), V^{n_2}(i,.), V_1^{(1)}(i,.), \ldots, V_{n_2}^{(1)}(i,.), V_1^{(2)}(i,.), \ldots, V_{t+1}^{(2)}(i,.), \\ & \Delta N_1^{(1)}(i,.), \ldots, \Delta N_{n_2}^{(1)}(i,.), \Delta N_1^{(2)}(i,.), \ldots, \Delta N_t^{(2)}(i,.); \\ & C_1^{(1)}, \ldots, C_{n_1}^{(1)}, C_1^{(2)}, \ldots, C_{n_2}^{(2)}\}. \end{aligned} \tag{4.14}$$

$$\begin{aligned}\tau_i(n_1) = \text{álgebra gerada por}\{ \quad & V^{n_1}(1,.), V^{n_2}(1,.), V_1^{(1)}(1,.), \ldots, V_{n_1}^{(1)}(1,.), V_1^{(2)}(1,.), \ldots, V_{n_1}^{(2)}(1,.), \\ & \Delta N_1^{(1)}(1,.), \ldots, \Delta N_{n_1}^{(1)}(1,.), \Delta N_1^{(2)}(1,.), \ldots, \Delta N_{n_1}^{(2)}(1,.); \ldots; \\ & V^{n_1}(i,.), V^{n_2}(i,.), V_1^{(1)}(i,.), \ldots, V_{n_1}^{(1)}(i,.), V_1^{(2)}(i,.), \ldots, V_{n_1}^{(2)}(i,.), \\ & \Delta N_1^{(1)}(i,.), \ldots, \Delta N_{n_1}^{(1)}(i,.), \Delta N_1^{(2)}(i,.), \ldots, \Delta N_{n_1}^{(2)}(i,.); \end{aligned} \tag{4.15}$$

$$C_1^{(1)}, \ldots, C_{n_1}^{(1)}, C_1^{(2)}, \ldots, C_{n_2}^{(2)}\}. \tag{4.16}$$

Consideremos as soma parciais $P_j^{n_1,n_2}(i)$ como sendo a soma dos termos $a_0, a_1, \ldots, a_j$ para $0 \leq j \leq n_1$ e $i \in \{1, \ldots, k\}$ fixo,

$$P_j^{n_1,n_2}(i) = a_0 + a_1 + \ldots + a_j \tag{4.17}$$

e seja $\{\tau_i\}$ a filtragem formada pelas álgebras encaixantes $\tau_i(j)$ com $0 \leq j \leq n_1$. Com isso, para $i \in \{1, \ldots, k\}$ fixo, definimos o processo estocástico

$$P^{n_1,n_2}(i) = \{P_0^{n_1,n_2}(i), P_1^{n_1,n_2}(i), \ldots, P_{n_1}^{n_1,n_2}(i)\} \tag{4.18}$$

que é adaptado à filtragem $\{\tau_i\}$ por construção.

Proposição 4.2. *O processo estocástico $P^{n_1,n_2}(i)$, para $i \in \{1, \ldots, k\}$ fixo, é um martingale com respeito à filtragem $\{\tau_i\}$.*

Demonstração. Por construção o processo $P^{n_1,n_2}(i)$ é adaptado à filtragem $\{\tau_i\}$. Com isso, basta provar que

$$\mathrm{E}[P_m^{n_1,n_2}(i)|\tau_i(m-1)] = P_{m-1}^{n_1,n_2}(i). \tag{4.19}$$

No entanto, pela definição das somas parciais dada pela equação 4.17, $P_j^{n_1,n_2}(i) - P_{j-1}^{n_1,n_2}(i) = a_j$, para todo $j = 1, \ldots n_1$. Com isso, basta mostrar que

$$\mathrm{E}[a_j|\tau_i(j-1)] = 0.$$

Para $1 \leq j \leq n_2$, temos que

$$\begin{aligned}
\mathrm{E}[a_j|\tau_i(j-1)] &= \mathrm{E}\left[\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left(\frac{\Delta Y_j^{(1)}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right)|\tau_i(j-1)\right] \\
&= \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left\{\frac{1}{V^{n_1}(l,.)}\mathrm{E}\left[\Delta N_j^{(1)}(l.) - \Delta A_j^{(1)}(l,.)|\tau_i(j-1)\right]\right. \\
&- \left.\frac{1}{V^{n_2}(l,.)n_1}\mathrm{E}\left[\Delta N_j^{(2)}(l.) - \Delta A_j^{(2)}(l,.)|\tau_i(j-1)\right]\right\}
\end{aligned}$$

Observe que $\Delta A_j^{(p)}(l,.)$ é mensurável em $\tau_i(j-1)$ e $\Delta N_j^{(p)}(l,.)$ condicionado a $V_j^{(p)}(l,.)$ é independente de todos os outros termos da filtragem, $\tau_i(j-1)$ para $1 \leq j \leq n_2$ e $p = 1, 2$. Utilizando essas propriedades, temos que

$$\begin{aligned}
&E\left[\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left(\frac{\Delta Y_j^{(1)}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right)|\tau_i(j-1)\right] \\
&= \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left\{\frac{1}{V^{n_1}(l,.)n_2}\mathrm{E}\left[\Delta N_j^{(1)}(l.)|V_j^{(1)}(l,.)\right] - V_j^{(1)}(l,.)h^{(1)}(l)\right\} \\
&- \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left\{\frac{1}{V^{n_2}(l,.)n_1}\mathrm{E}\left[\Delta N_j^{(2)}(l.)|V_j^{(2)}(l,.)\right] - V_j^{(2)}(l,.)h^{(2)}(l)\right\} \\
&= 0
\end{aligned}$$

pois $\Delta N_j^{(p)}(l,.)|V_j^{(p)}(l,.)$ tem distribuição binomial com número de ensaios $V_j^{(p)}(l,.)$ e probabilidade $h^{(p)}(l)$, para $p = 1, 2$. Para $n_2 < j \leq n_1$ os termos do processo são dados por

$$a_j = \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\frac{\Delta Y_j^{(1)}(l,.)}{V^{n_1}(l,.)}$$

As propriedades verificadas para o primeiro caso também são válidas para esse caso. Assim, de forma análoga ao realizado anteriormente, temos que

$$\mathrm{E}\left[\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\frac{\Delta Y_j^{(1)}(l,.)}{V^{n_1}(l,.)}|\tau_i(j-1)\right] = 0$$

□

Da Proposição 2.2, concluímos que o processo $\{a_j : 0 \leq j \leq n_1\}$ é um martingale *array difference* em relaçao a filtragem $\{\tau_i\}$.

Lema 4.3. *Considere os termos a_j, $j = 1, \ldots, n_1$. A seguinte propriedade é válida*

$$E\left[\max_{1\leq j\leq n_1} |a_j|\right] \to 0, \quad \textit{quando } n_1, n_2 \to \infty. \tag{4.20}$$

Demonstração. Para $1 \leq j \leq n_2$,

$$a_j = \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_j^{(1)}(l.)}{V^{n_1}(l,.)} - \frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right].$$

Utilizando propriedades de módulo e máximo, temos que

$$\begin{aligned}
\max_{1\leq j\leq n_2}\left|\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_j^{(1)}(l.)}{V^{n_1}(l,.)} - \frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right]\right| &\leq \max_{1\leq j\leq n_2}\sum_{l=1}^{i}\left|U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_j^{(1)}(l.)}{V^{n_1}(l,.)} - \frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right]\right| \\
&\leq \sum_{l=1}^{i}\max_{1\leq j\leq n_2}\left|U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_j^{(1)}(l.)}{V^{n_1}(l,.)} - \frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right]\right|
\end{aligned}$$

Com isso,

$$\begin{aligned}
&\mathrm{E}\left[\max_{1\leq j\leq n_2}\left|\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_j^{(1)}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right]\right|\right] \\
&\qquad \leq \mathrm{E}\left[\sum_{l=1}^{i}\max_{1\leq j\leq n_2}\left|U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_j^{(1)}(l.)}{V^{n_1}(l,.)} - \frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right]\right|\right]
\end{aligned} \tag{4.21}$$

Por outro lado,

$$\begin{aligned}
&\left|U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_j^{(1)}(l.)}{V^{n_1}(l,.)} - \frac{\Delta Y_j^{(2)}(l.)}{V^{n_2}(l,.)}\right]\right| \\
&\quad = \left|U^{n_1,n_2}(l,.)\left[\frac{\Delta N_j^{(1)}(l,.) - \Delta A_j^{(1)}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta N_j^{(2)}(l.) - \Delta A_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right]\right| \\
&\quad \leq \left|U^{n_1,n_2}(l,.)\frac{\Delta N_j^{(1)}(l.) - \Delta A_j^{(1)}(l,.)}{V^{n_1}(l,.)}\right| + \left|U^{n_1,n_2}(l,.)\frac{\Delta N_j^{(2)}(l.) - \Delta A_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right| \\
&\quad \leq \left|\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)}\right|\left(\left|\Delta N_j^{(1)}(l.)\right| + \left|\Delta A_j^{(1)}(l,.)\right|\right) + \left|\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)}\right|\left(\left|\Delta N_j^{(2)}(l.)\right| + \left|\Delta A_j^{(2)}(l,.)\right|\right) \\
&\quad \leq 2\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)}\right| + 2\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_2}(l,.)}\right|
\end{aligned} \tag{4.22}$$

Das equações 4.24 e 4.24, obtemos que

$$\begin{aligned}
&\mathrm{E}\left[\max_{1\le j\le n_2}\left|\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_j^{(1)}(l,.)}{V^{n_1}(l,.)}-\frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right]\right|\right] \\
&\qquad\le \mathrm{E}\left[\max_{1\le j\le n_2} 2i\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)}\right|+2i\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_2}(l,.)}\right|\right] \\
&\qquad= 2i\mathrm{E}\left[\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)}\right|\right]+2i\mathrm{E}\left[\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_2}(l,.)}\right|\right]
\end{aligned}$$

para $i\in\{1,\dots,k\}$ fixo, com $k\in\mathbb{N}$. Por hipóteste $E\left[\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)}\right|\right]\to 0$ e $E\left[\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_2}(l,.)}\right|\right]\to 0$ quando $n_1,n_2\to\infty$, $\forall l$. Portanto,

$$\mathrm{E}\left[\max_{1\le j\le n_2}\left|\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_j^{(1)}(l.)}{V^{n_1}(l,.)}-\frac{\Delta Y_j^{(2)}(l,.)}{V^{n_2}(l,.)}\right]\right|\right]\to 0, \text{quando } n_1,n_2\to\infty. \tag{4.23}$$

Para $n_2<j\le n_1$,

$$a_j=\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\frac{\Delta Y_j^{(1)}(l,.)}{V^{n_1}(l,.)}.$$

De forma forma análoga ao caso anteorior obtemos as seguintes desigualdades

$$\mathrm{E}\left[\max_{1\le j\le n_2}\left|\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\frac{\Delta Y_j^{(1)}(l,.)}{V^{n_1}(l,.)}\right|\right]\le \mathrm{E}\left[\sum_{l=1}^{i}\max_{1\le j\le n_2}\left|U^{n_1,n_2}(l,.)\frac{\Delta Y_j^{(1)}(l.)}{V^{n_1}(l,.)}\right|\right]$$

e

$$\left|U^{n_1,n_2}(l,.)\frac{\Delta Y_j^{(1)}(l.)}{V^{n_1}(l,.)}\right| \le 2\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)} \tag{4.24}$$

$\forall l=1,\dots,i$.

Assim, concluímos que

$$\mathrm{E}\left[\max_{1\le j\le n_2}\left|\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\frac{\Delta Y_j^{(1)}(l,.)}{V^{n_1}(l,.)}\right|\right]\to 0, \quad \text{quando} \quad n_1,n_2\to\infty.$$

□

Lema 4.4. *Consideremos o processo* $\{a_j : 0\le j\le n_1\}$ *um martingale* array difference *como definimos anteriormente. Temos que*

$$\sum_{j=1}^{n_1} {a_j}^2(i)\to\sigma^2(i)$$

em probabilidade, quando $n_1,n_2\to\infty$, *em que*

$$\sigma^2(i)=\sum_{l=1}^{i}[1-h^{(1)}(l)]h^{(1)}(l)G_l^1+\sum_{l=1}^{i}[1-h^{(2)}(l)]h^{(2)}(l)G_l^2.$$

Demonstração. Observe que a $\sum_{j=1}^{n_1} a_j(i)$ pode ser escrito da seguinte forma

$$\sum_{j=1}^{n_1} a_j(i) = \sum_{j_1=1}^{n_1}\sum_{j_2=1}^{n_2}\left(\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_{j_1}^{(1)}(l,.)}{V^{n_1}(l,.)n_2} - \frac{\Delta Y_{j_2}^{(2)}(l,.)}{V^{n_2}(l,.)n_1}\right]\right) \tag{4.25}$$

Conforme a Proposição 2.1 e pela equação 4.25 temos que

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{n_1} {a_j}^2(i)\right] &= \mathrm{E}\left[\left(\sum_{l=1}^{n_1} a_j(i)\right)^2\right] \\
&= \mathrm{E}\left[\left(\sum_{m_1=1}^{n_1}\sum_{m_2=1}^{n_2}\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y_m^{(1)}(l,.)}{V^{n_1}(l,.)n_2} - \frac{\Delta Y_m^{(2)}(l,.)}{V^{n_2}(l,.)n_1}\right]\right)^2\right] \\
&= \mathrm{E}\left[\left(\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y^{n_1}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y^{n_2}(l,.)}{V^{n_2}(l,.)}\right]\right)^2\right]
\end{aligned}$$

Denotando por

$$\Delta F(l,.) = \frac{\Delta Y^{n_1}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y^{n_2}(l,.)}{V^{n_2}(l,.)},$$

temos que

$$\begin{aligned}
\mathrm{E}\left[\left(\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y^{n_1}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y^{n_2}(l,.)}{V^{n_2}(l,.)}\right]\right)^2\right] &= E\left[\left(\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\Delta F(l,.)\right)^2\right] \\
&= E\left[(U^{n_1,n_2}.F)^2(i,.)\right]] \\
&= E\left[< (U^{n_1,n_2}.F),(U^{n_1,n_2}.F) > (i,.)\right] \\
&= E\left[(U^{n_1,n_2}{}^2. < F,F >)(i,.)\right]
\end{aligned}$$

Por outro lado, temos que

$(U^{n_1,n_2}{}^2; < F,F >)(i,.)$

$$
\begin{aligned}
&= \sum_{l=1}^{i} U^{n_1,n_2 2}(l,.)\Delta < F,F > (l,.) \\
&= \sum_{l=1}^{i} U^{n_1,n_2 2}(l,.)\left[\frac{1}{V^{n_1 2}(l,.)}\Delta < Y^{n_1},Y^{n_1} > (l,.) + \frac{1}{V^{n_2 2}(l,.)}\Delta < Y^{n_2},Y^{n_2} > (l,.)\right] \\
&= \sum_{l=1}^{i} U^{n_1,n_2 2}(l,.)\left[\frac{1}{V^{n_1 2}(l,.)}\Delta \sum_{m_1=1}^{n_1} < Y^{(1)}_{m_1},Y^{(1)}_{m_1} > (l,.) + \frac{1}{V^{n_2 2}(l,.)}\Delta \sum_{m_2=1}^{n_2} < Y^{(2)}_{m_2},Y^{(2)}_{m_2} > (l,.)\right] \\
&= \sum_{l=1}^{i} U^{n_1,n_2 2}(l,.)\left[\sum_{m_1=1}^{n_1} \frac{\Delta A^{(1)}_{m_1}(l,.) - \Delta A^{(1)^2}_{m_1}(l,.)}{V^{n_1 2}(l,.)} + \frac{\sum_{m_2=1}^{n_2} \Delta A^{(2)}_{m_2}(l,.) - \Delta A^{(2)^2}_{m_2}(l,.)}{V^{n_2 2}(l,.)}\right] \\
&= \sum_{l=1}^{i} \frac{U^{n_1,n_2 2}(l,.)}{V^{n_1}(l,.)}[1-h^{(1)}(l)]h^{(1)}(l) + \sum_{l=1}^{i} \frac{U^{n_1,n_2 2}(l,.)}{V^{n_2}(l,.)}[1-h^{(2)}(l)]h^{(2)}(l)
\end{aligned}
$$

Assim,

$$
\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{n_1} a_j{}^2(i)\right] &= \mathrm{E}\left[\sum_{l=1}^{i} \frac{U^{n_1,n_2 2}(l,.)}{V^{n_1}(l,.)}[1-h^{(1)}(l)]h^{(1)}(l) + \sum_{l=1}^{i} \frac{U^{n_1,n_2 2}(l,.)}{V^{n_2}(l,.)}[1-h^{(2)}(l)]h^{(2)}(l)\right] \\
&= \sum_{l=1}^{i}[1-h^{(1)}(l)]h^{(1)}(l)\mathrm{E}\left[\frac{U^{n_1,n_2 2}(l,.)}{V^{n_1}(l,.)}\right] + \sum_{l=1}^{i}[1-h^{(2)}(l)]h^{(2)}(l)\mathrm{E}\left[\frac{U^{n_1,n_2 2}(l,.)}{V^{n_2}(l,.)}\right]
\end{aligned}
$$

Se G^1_l denota o limite de $\mathrm{E}\left[\frac{U^{n_1,n_2 2}(l,.)}{V^{n_1}(l,.)}\right]$ e G^2_l denota o limite de $\mathrm{E}\left[\frac{U^{n_1,n_2 2}(l,.)}{V^{n_2}(l,.)}\right]$, quando $n_1, n_2 \to \infty$, então obtemos que

$$
\mathrm{E}\left[\sum_{j=1}^{n_1} a_j{}^2(i)\right] \to \sum_{l=1}^{i}[1-h^{(1)}(l)]h^{(1)}(l)G^1_l + \sum_{l=1}^{i}[1-h^{(2)}(l)]h^{(2)}(l)G^2_l,
$$

quando $n_1, n_2 \to \infty$. Conseqüentemente, pelo Teorema 2.7.2 de Stout (1970), temos que

$$
\sum_{j=1}^{n_1} a_j{}^2(i) \to \sum_{l=1}^{i}[1-h^{(1)}(l)]h^{(1)}(l)G^1_l + \sum_{l=1}^{i}[1-h^{(2)}(l)]h^{(2)}(l)G^2_l \tag{4.26}
$$

em probabilidade, quando $n_1, n_2 \to \infty$.

□

Teorema 4.2. *Temos que*

$$
P^{n_1,n_2}_{n_1}(i) = \sum_{j=1}^{n_1} a_j(i) = \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y^{n_1}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y^{n_2}(l,.)}{V^{n_2}(l,.)}\right]
$$

converge em distribuição para uma variável aleatória normal com média 0 e variância $\sigma^2(i)$, *quando* $n_1, n_2 \to \infty$.

Demonstração. Pela Proposição 4.2 concluímos que o processo $\{a_j(i) : 0 \leq j \leq n_1\}$ é um martingale *array difference*. Além disso, utilizando os Lemas 4.1 e 4.2, concluímos pelo Teorema 2.5, item (b), de Helland (1982) que $P_{n_1}^{n_1,n_2}(i)$ converge em distribuição para uma variável aleatória com média 0 e variância $\sigma^2(i)$. □

Corolário 4.3. *Ao tomarmos*

$$U^{n_1,n_2}(l,.) = \left(\frac{n_1+n_2}{n_1 n_2}\right)^{1/2} \frac{V^{n_1}(l,.)V^{n_2}(l,.)}{V^{n_1}(l,.)+V^{n_2}(l,.)},$$

para $l = 1, \ldots, i$, *obtemos que*

$$\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\left(\hat{h}^{n_1}(l) - h^{(1)}(l)\right) - \left(\hat{h}^{n_2}(l) - h^{(2)}(l)\right)\right], \tag{4.27}$$

converge em distribuição para uma variável aleatória com média 0 e variância $\sigma^2(i)$ *dada por*

$$\sigma^2(i) = \sum_{l=1}^{i}\left([1-\hat{h}^{n_1}(l)]\hat{h}^{n_1}(l)G_l^1 + [1-\hat{h}^{n_2}(l)]\hat{h}^{n_2}(l)G_l^2\right).$$

Demonstração. Para $U^{n_1,n_2}(l,.)$ definido como na equação 4.28, o processo $P_{n_1}^{n_1,n_2}(l)$ é dado por

$$\begin{aligned}
P_{n_1}^{n_1,n_2}(i) &= \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta Y^{n_1}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y^{n_2}(l,.)}{V^{n_2}(l,.)}\right] \\
&= \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta N^{n_1}(l,.) - \Delta A^{n_1}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta N^{n_2}(l,.) - \Delta A^{n_2}(l,.)}{V^{n_2}(l,.)}\right] \\
&= \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\frac{\Delta N^{n_1}(l,.) - V^{n_1}(l,.)h^{(1)}(l)}{V^{n_1}(l,.)} - \frac{\Delta N^{n_2}(l,.) - V^{n_2}(l,.)h^{(2)}(l)}{V^{n_2}(l,.)}\right] \\
&= \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[(\hat{h}^{n_1}(l) - h^{(1)}(l)) - (\hat{h}^{n_2}(l) - h^{(2)}(l))\right]
\end{aligned}$$

Além disso,

$$\begin{aligned}
\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)}\right| &= \left(\frac{n_1+n_2}{n_1 n_2}\right)^{1/2} \frac{V^{n_2}(l,.)}{V^{n_1}(l,.)+V^{n_2}(l,.)} \\
&\leq \left(\frac{n_1+n_2}{n_1 n_2}\right)^{1/2} \\
&= \left(\frac{1}{n_2} + \frac{1}{n_1}\right)^{1/2}
\end{aligned}$$

Assim, $|\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)}|$ é limitado e converge para 0 em probabilidade quando $n_1, n_2 \to \infty$. Portanto,

$$\mathrm{E}\left[\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_1}(l,.)}\right|\right] \to 0, \text{ quando } n_1, n_2 \to \infty.$$

Da mesma forma, temos que

$$\mathrm{E}\left[\left|\frac{U^{n_1,n_2}(l,.)}{V^{n_2}(l,.)}\right|\right] \to 0, \text{ quando } n_1, n_2 \to \infty.$$

Por outro lado temos que

$$\begin{aligned}
\frac{U^{n_1,n_2}{}^2(l,.)}{V^{n_1}(l,.)} &= \frac{n_1+n_2}{n_1 n_2}\frac{V^{n_1}(l,.)V^{n_2}{}^2(l,.)}{(V^{n_1}(l,.)+V^{n_2}(l,.))^2} \\
&= \frac{n_1+n_2}{n_2}\frac{V^{n_1}(l,.)}{n_1}\frac{V^{n_2}{}^2(l,.)}{(V^{n_1}(l,.)+V^{n_2}(l,.))^2} \\
&\leq L\frac{n_1+n_2}{n_2}\frac{V^{n_1}(l,.)}{n_1}.
\end{aligned}$$

Como assumimos que

$$\frac{n_1+n_2}{n_p} \to u_p,$$

$p = 1, 2$ e $u_p \in \mathbb{R}$, quando $n_1, n_2 \to \infty$, ou seja, existe uma proporcionalidade entre n_1 e n_2, obtemos que existe o limite de

$$\mathrm{E}\left[\frac{U^{n_1,n_2}{}^2(l,.)}{V^{n_1}(l,.)}\right].$$

Da mesma forma temos que existe o limite de

$$\mathrm{E}\left[\frac{U^{n_1,n_2}{}^2(l,.)}{V^{n_2}(l,.)}\right].$$

Assim, como a Proposição 4.2 e o Lema 4.3 são satisfeitos, obtemos do Teorema 4.2 que

$$\sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\left(\hat{h}^{n_1}(l) - h^{(1)}(l)\right) - \left(\hat{h}^{n_2}(l) - h^{(2)}(l)\right)\right], \tag{4.28}$$

converge em distribuição para uma variável aleatória com média 0 e variância $\sigma^2(i)$. □

Capítulo 5

Testes de Hipóteses

5.1 Introdução

Neste capítulo apresentamos testes de hipóteses para os parâmetros do modelo de intensidade multiplicativo para uma e duas populações. Os testes foram desenvolvidos utilizando o estudo de distribuição assintótica de integrais estocásticas apresentado no Capítulo 4.

A estatística de teste para comparação de duas populações com dados censurados, apresentada neste capítulo, foi proposta por Mantel (1966) e é baseada na estatística de Mantel-Haenszel para tabelas de contigências estratificadas. Esta estatística ficou conhecida como estatística de *logrank* e é provavelmente a mais utilizada como estatística de teste para comparação de duas populações com dados censurados.

5.2 Teste de Hipóteses para uma População

Nesta seção, apresentamos testes de hipóteses para os parâmetros do modelo de intensidade multiplicativo para uma população.

Consideremos uma amostra aleatória simples de tamanho n, $X^{(1)} = ((W_1^{(1)}, C_1^{(1)}), \ldots, (W_n^{(1)}, C_n^{(1)}))$, e a teoria definida e desenvolvida no Capítulo 3. Como resultado do Corolário 4.1, temos que

$$\sqrt{n}\left[\hat{h}^n(l) - h^{(1)}(l)\right]$$

tem distribuição assintóticamente normal com média 0 e variância $\dfrac{[h^{(1)}(l) - {h^{(1)}}^2(l)]}{\theta^{(1)}(l)}$, para

$l = 1, \ldots, i$ e $i \in \{1, \ldots, k\}$, em que $\theta^{(1)}(l) = \eta^{(1)}(l) + \ldots + \eta^{(1)}(k)$. Portanto, para testarmos as hipóteses

$$\begin{cases} H_0 : h^{(1)}(l) = h_0^{(1)}(l) \\ H_1 : h^{(1)}(l) \neq h_0^{(1)}(l) \end{cases} \tag{5.1}$$

em que $h_0^{(1)}(l) \in [0, 1]$ utilizamos a seguinte estatística de teste

$$W^h(l) = \frac{\sqrt{n}\left[\hat{h}^n(l) - h^{(1)}(l)\right]}{\sqrt{[h^{(1)}(l) - h^{(1)^2}(l)]/\theta^{(1)}(l)}}$$

que sob H_0 é dada por

$$W_0^h(l) = \frac{\sqrt{n}\left[\hat{h}^n(l) - h_0^{(1)}(l)\right]}{\sqrt{[h^{(1)}(l) - h^{(1)^2}(l)]/\theta^{(1)}(l)}} \tag{5.2}$$

e tem distribuição assintóticamente normal com média 0 e variância 1. No lugar dos parâmetros $h^{(1)}$ e $\theta^{(1)}$, utilizamos os estimadores $\hat{h}^n$ e $\hat{\theta}^n$, pois tratam-se de estimadores consistentes. Com isso, a estatística de teste utilizada é dada por

$$W_0^h(l) = \frac{\sqrt{n}\left[\hat{h}^n(l) - h_0^{(1)}(l)\right]}{\sqrt{[\hat{h}^n(l) - \hat{h}^{n^2}(l)]/\hat{\theta}^{(1)}(l)}} \tag{5.3}$$

tem distribuição assintóticamente normal com média 0 e variância 1. Observe que ao testarmos as hipóteses 5.1, estamos testando se a função intensidade na categoria l é igual a um determinado valor.

Para testarmos hipóteses relacionadas a função intensidade acumulada, por exemplo,

$$\begin{cases} H_0 : H^{(1)}(i) = H_0^{(1)}(i) \\ H_1 : H^{(1)}(i) \neq H_0^{(1)}(i) \end{cases} \tag{5.4}$$

em que $H_0^{(1)}(i) \in \mathbb{R}$, utilizaremos a estatística

$$W^H(l) = \frac{\sqrt{n}\left[\hat{H}^n(i) - H^{(1)}(i)\right]}{\sqrt{\sum_{l=1}^{i}[h^{(1)}(l) - h^{(1)^2}(l)]/\theta^{(1)}(l)}}$$

que sob H_0 é dada por

$$W_0^H(l) = \frac{\sqrt{n}\left[\hat{H}^n(i) - H_0^{(1)}(i)\right]}{\sqrt{\sum_{l=1}^{i}[h^{(1)}(l) - h^{(1)^2}(l)]/\theta^{(1)}(l)}}$$

e tem distribuição assintóticamente normal com média 0 e variância 1. Da mesma forma que discutido anteriormente, substituímos os parâmetros pelos seus respectivos estimadores, pois se tratam de estimadores consistentes. Portanto, a estatística utilizada dada por

$$W_0^H(l) = \frac{\sqrt{n}\left[\hat{H}^n(i) - H_0^{(1)}(i)\right]}{\sqrt{\sum_{l=1}^{i}[\hat{h}^n(l) - \hat{h}^{n^2}(l)]/\hat{\theta}^{(1)}(l)}} \tag{5.5}$$

tem distribuição assintótica normal com média 0 e variância 1.

Observe que ao testarmos as hipóteses 5.4, estamos testando se a função intensidade acumulada até a categoria i, com $i \in \{1, \ldots, k\}$ é igual a um determinado valor.

5.3 Teste de Hipóteses para Duas amostras

Agora iremos apresentar a estatística para comparação de duas populações. Consideremos uma amostra aleatória simples de tamanho n_1, $X^{(1)} = ((W_1^{(1)}, C_1^{(1)}), \ldots, (W_{n_1}^{(1)}, C_{n_1}^{(1)}))$, e uma amostra aleatória de tamanho n_2, $X^{(2)} = ((W_1^{(2)}, C_1^{(2)}), \ldots, (W_{n_1}^{(2)}, C_{n_1}^{(2)}))$, com $X^{(1)}$ e $X^{(2)}$ independentes.

Para comparar as curvas de duas populações iremos introduzir o teste de *logrank.* Consideremos l categorias, $l = 1, \ldots, i$ com $i \in \{1, \ldots, k\}$, que são os possíveis valores que a variável aleatória pode assumir. Utilizando a notação introduzida no Capítulo 3, temos que $\Delta N^{n_p}(l, .)$ e $V^{n_p}(l, .)$, $p = 1, 2$, denotam o número de observações ou itens em que ocorreu o evento de interesse e o número de observações ou itens sob risco, respectivamente, na categoria l. Denotamos por $\Delta N(l, .) = \Delta N^{n_1}(l, .) + \Delta N^{n_2}(l, .)$ e $V(l, .) = V^{n_1}(l, .) + V^{n_2}(l, .)$. Os dados em uma categoria expecífica l podem ser resumidos conforme a Tabela 5.1.

Tabela 5.1: *Número de itens em que ocorreu o evento de interesse e número de itens sob risco na categoria l, por amostra.*

	Amostras		
Ocorrência do evento de interesse	1	2	Total
Sim	$\Delta N^{n_1}(l, .)$	$\Delta N^{n_2}(l, .)$	$\Delta N(l, .)$
Não	$V^{n_1}(l, .) - \Delta N^{n_1}(l, .)$	$V^{n_2}(l, .) - \Delta N^{n_2}(l, .)$	$V(l, .) - \Delta N(l, .)$
Total	$V^{n_1}(l, .)$	$V^{n_2}(l, .)$	$V(l, .)$

Sabemos que $\Delta N^{n_p}(l,.)/V^{n_p}(l,.)$ tem distribuição binomial com número de enasaios $V^{n_p}(l,.)$ e probabilidade $h^{(p)}(l)$, $p = 1, 2$. Assim, obtemos que $\Delta N^{n_p}(l,.)/V^{n_p}(l,.); \Delta N(l,.)$ tem distribuição hipergeométrica com parâmetros $V^{n_p}(l,.)$, $V(l,.)$ e $\Delta N(l,.)$. Com isso, a esperança e a variância da variável $\Delta N^{n_1}(l,.)/V^{n_1}(l,.); \Delta N(l,.)$ são dadas por

$$\mathrm{E}^{(1)}(l) = \mathrm{E}\left[\Delta N^{n_1}(l,.)/V^{n_1}(l,.); \Delta N(l,.)\right] = \Delta N(l,.)\frac{V^{n_1}(l,.)}{V(l,.)}$$

e

$$\mathrm{Var}^{(1)}(l) = \Delta N(l,.)\frac{V^{n_1}(l,.)V^{n_2}(l,.)}{V^2(l,.)}\frac{V(l,.) - \Delta N(l,.)}{V(l,.) - 1}$$

Consideremos o vetor formado pela diferença entre o número de itens observados em que ocorreram o evento de interesse e o número esperado em cada categoria, relacionado a amostra 1,

$$\{\Delta N^{n_1}(1,.) - E_1^1, \Delta N^{n_1}(2,.) - E_2^1, \ldots, \Delta N^{n_1}(i,.) - E_i^1\},$$

para $i = 1, \ldots, k$, e Q a soma de cada um dos termos deste vetor.

$$
\begin{aligned}
W &= \sum_{l=1}^{i} \left(\Delta N^{n_1}(l,.) - E_l^{(1)} \right) \\
&= \sum_{l=1}^{i} \Delta N^{n_1}(l,.) - \sum_{l=1}^{i} \frac{\Delta N(l,.) V^{n_1}(l,.)}{V(l,.)} \\
&= \sum_{m=1}^{n_1} \sum_{l=1}^{i} \Delta N_m^{(1)}(l,.) - \sum_{p=1}^{2} \sum_{m=1}^{n_p} \sum_{l=1}^{i} \Delta N_m^{(p)}(l,.) \frac{V^{n_1}(l,.)}{V(l,.)} \\
&= \sum_{m=1}^{n_1} \sum_{l=1}^{i} \Delta N_m^{(1)}(l,.) - \left[\sum_{m=1}^{n_1} \sum_{l=1}^{i} \Delta N_m^{(1)}(l,.) \frac{V^{n_1}(l,.)}{V(l,.)} + \sum_{m=1}^{n_2} \sum_{l=1}^{i} \Delta N_m^{(2)}(l,.) \frac{V^{n_1}(l,.)}{V(l,.)} \right] \\
&= \sum_{m=1}^{n_1} \sum_{l=1}^{k} \left[1 - \frac{V^{n_1}(l,.)}{V(l,.)} \right] \Delta N_m^{(1)}(l,.) - \sum_{m=1}^{n_2} \sum_{l=1}^{k} \Delta N_m^{(2)}(l,.) \frac{V^{n_1}(l,.)}{V(l,.)} \\
&= \sum_{m=1}^{n_1} \sum_{l=1}^{i} \frac{V^{n_2}(l,.)}{V(l,.)} \Delta N_m^{(1)}(l,.) - \sum_{m=1}^{n_2} \sum_{l=1}^{i} \frac{V^{n_1}(l,.)}{V(l,.)} \Delta N_m^{(2)}(l,.) \\
&= \sum_{m=1}^{n_1} \sum_{l=1}^{i} \frac{V^{n_2}(l,.) V^{n_1}(l,.)}{V(l,.)} \{V^{n_1}(l,.)\}^{-1} \left[\Delta N_m^{(1)}(l,.) - V_m^{(1)}(l,.) \Delta H^{(1)}(l) \right] \\
&- \sum_{m=1}^{n_2} \sum_{l=1}^{i} \frac{V^{n_1}(l,.) V^{n_2}(l,.)}{V(l,.)} \{V^{n_2}(l,.)\}^{-1} \left[\Delta N_m^{(2)}(l,.) - V_m^{(2)}(l,.) \Delta H^{(2)}(l) \right] \\
&= \sum_{m=1}^{n_1} \sum_{l=1}^{i} \frac{V^{n_2}(l,.)}{V(l,.)} V^{n_1}(l,.) \{V^{n_1}(l,.)\}^{-1} \left[\Delta N_m^{(1)}(l,.) - \Delta A_m^{(1)}(l,.) \right] \\
&- \sum_{m=1}^{n_2} \sum_{l=1}^{i} \frac{V^{n_1}(l,.) V^{n_2}(l,.)}{V(l,.)} \{V^{n_2}(l,.)\}^{-1} \left[\Delta N_m^{(2)}(l,.) - \Delta A_m^{(1)}(l,.) \right] \\
&= \sum_{l=1}^{i} \frac{V^{n_2}(l,.) V^{n_1}(l,.)}{V(l,.)} \left[\sum_{m=1}^{n_1} \frac{\Delta Y_m^{(1)}(l,.)}{V^{n_1}(l,.)} - \sum_{m=1}^{n_2} \frac{\Delta Y_m^{(2)}(l,.)}{V^{n_2}(l,.)} \right] \\
&= \sum_{l=1}^{i} \frac{V^{n_2}(l,.) V^{n_1}(l,.)}{V(l,.)} \left[\frac{\Delta Y^{n_1}(l,.)}{V^{n_1}(l,.)} - \frac{\Delta Y^{n_2}(l,.)}{V^{n_2}(l,.)} \right]
\end{aligned}
$$

Esta é derivação de Mantel da estatística de *logrank* (Fleming & Harrington (2005)). A seguir, utilizaremos a versão da estatística de *logrank* proposta por Fleming & Harrington

(2005), conforme abaixo,

$$\begin{aligned}
W &= \sum_{l=1}^{i}\left(\frac{n_1+n_2}{n_1n_2}\right)^{1/2}\frac{V^{n_2}(l,.)V^{n_1}(l,.)}{V(l,.)}\left[\frac{\Delta Y^{n_1}(l,.)}{V^{n_1}(l,.)}-\frac{\Delta Y^{n_2}(l,.)}{V^{n_2}(l,.)}\right]\\
&= \sum_{l=1}^{i}\left(\frac{n_1+n_2}{n_1n_2}\right)^{1/2}\frac{V^{n_2}(l,.)V^{n_1}(l,.)}{V(l,.)}\left[\left(\frac{\Delta N^{n_1}(l,.)}{V^{n_1}(l,.)}-\Delta H^{(1)}(l)\right)-\left(\frac{\Delta N^{n_2}(l,.)}{V^{n_2}(l,.)}-\Delta H^{(2)}(l)\right)\right]\\
&= \sum_{l=1}^{i}\left(\frac{n_1+n_2}{n_1n_2}\right)^{1/2}\frac{V^{n_2}(l,.)V^{n_1}(l,.)}{V(l,.)}\left[\left(\hat{h}^{n_1}(l)-h^{(1)}(l)\right)-\left(\hat{h}^{n_2}(l)-h^{(2)}(l)\right)\right]
\end{aligned}$$

Sob a hipótese nula $H_0 : H^{n_1} = H^{n_2}$, dada por

$$W = \sum_{l=1}^{i}\left(\frac{n_1+n_2}{n_1n_2}\right)^{1/2}\frac{V^{n_2}(l,.)V^{n_1}(l,.)}{V(l,.)}\left[\hat{h}^{n_1}(l)-\hat{h}^{n_2}(l)\right] \tag{5.6}$$

e tem distribuição normal com média 0 e variância

$$\sigma^2(i) = \sum_{l=1}^{i}\left([1-h^{(1)*}(l)]h^{(1)}(l)G_l^1 + [1-h^{(2)}(l)]h^{(2)}(l)G_l^2\right).$$

Um estimador para a variância $\sigma^2(i)$, conforme proposto por Fleming & Harrington (2005), é dado por

$$\hat{\sigma}^2(i) = \sum_{l=1}^{i}\frac{U^{n_1,n_2 2}(l,.)}{V^{n_1}(l,.)}[1-\hat{h}^{n_1}(l)]\hat{h}^{n_1}(l) + \sum_{l=1}^{i}\frac{U^{n_1,n_2 2}(l,.)}{V^{n_2}(l,.)}[1-\hat{h}^{n_2}(l)]\hat{h}^{n_2}(l).$$

Observe que com a estatística de teste W obtida, ao testarmos a hipótese $H_0 : H^{(1)} = H^{(2)}$, estamos testando a igualdade entre as curvas das duas populações.

Capítulo 6

Estudo de simulação

6.1 Introdução

Neste capítulo avaliamos numericamente a qualidade das aproximações assintóticas utilizadas nas inferências sobre os parâmetros do modelo de intensidade multiplicativo. Particularmente, o interesse é avaliar a consistências dos estimadores da função intensidade $h^{(p)}$, da função intensidade acumulada $H^{(p)}$ e da distribuição de probabilidade $\zeta^{(p)}$, além de avaliar a qualidade dos testes apresentados nesse trabalho.

As simulações computacionais deste trabalho foram realizadas usando a versão 2.8.0 do software estatístico R. Os resultados numéricos apresentados neste capítulo baseiam-se no modelo de intensidade multiplicativo proposto neste trabalho e na teoria desenvolvida a partir dele, apresentados nos capítulos 3, 4 e 5.

O modelo proposto pode ser utilizado para analisar dados multinomiais na presença ou não de censura. Foram realizados estudos de simulação para os dois casos.

6.2 Modelo na ausência de censura

Inicialmente, consideramos dados multinomiais na ausência de censura. Foram realizadas simulações considerando $k = 3, 7$ e 10 categorias. Os passos para a simulação são descritos a seguir.

1. Definimos valores para a probabilidade de ocorrência do evento de interesse em cada uma das categorias, $\zeta^{(p)}(1), \ldots, \zeta^{(p)}(k)$, de forma que a soma da distribuição de probabilidade some 1. Estes valores são apresentados na Tabela 6.1.

2. A partir desses valores, obtemos os valores da função intensidade $h^{(p)}$ e da função intensidade acumulada $H^{(p)}$ utilizando as equações

$$h^{(p)}(i) = \frac{\zeta^{(p)}(i)}{\zeta^{(p)}(i) + \ldots + \zeta^{(p)}(k)}$$

e

$$H^{(p)}(i) = \sum_{l=1}^{i} h^{(p)}(i),$$

para $i = 1, \ldots, k$. Os valores obtidos da função intensidade $h^{(p)}$ e da função intensidade acumulada $H^{(p)}$ também são apresentados na Tabela 6.1.

Tabela 6.1: *Valores da distribuição de probabilidade ζ, da função intensidade h e da função intensidade acumulada H para 3, 7 e 10 categorias.*

	3 categorias			7 categorias			10 categorias		
Categoria	$\zeta^{(p)}$	$h^{(p)}$	$H^{(p)}$	$\zeta^{(p)}$	$h^{(p)}$	$H^{(p)}$	$\zeta^{(p)}$	$h^{(p)}$	$H^{(p)}$
1	0,2	0,200	0,200	0,16	0,160	0,160	0,10	0,100	0,100
2	0,5	0,625	0,825	0,35	0,417	0,577	0,25	0,278	0,378
3	0,3	1,000	1,825	0,20	0,408	0,985	0,17	0,262	0,639
4				0,06	0,207	1,192	0,09	0,188	0,827
5				0,08	0,348	1,540	0,13	0,333	1,160
6				0,05	0,333	1,873	0,04	0,192	1,353
7				0,03	0,300	2,173	0,03	0,205	1,557
8				0,07	1,000	3,173	0,05	0,479	2,036
9							0,06	0,690	2,726
10							0,077	1,000	3,726

3. Consideramos amostras de tamanhos $20, 50, 80, 100$ para avaliarmos a qualidade dos estimadores e $20, 50, 80, 100, 200, 500$ para avaliarmos a qualidade dos testes de hipóteses.

4. O número de iterações foi fixado em 10000.

5. Para cada iteração geramos uma amostra aleatória da variável multinomial X utilizando a distribuição uniforme com parâmetros 0 e 1.

6. Calculamos os estimadores da função intensidade $h^{(p)}$, da função intensidade acumulada $H^{(p)}$ e da função distribuição de probabilidade $\zeta^{(p)}$ em cada categoria i, $i = 1, \ldots, k$, para cada iteração, utilizado as seguintes equações

$$\hat{h}^{n_p}(i) = \frac{\Delta N^{n_p}(i,.)}{V^{n_p}(i,.)}, \quad \hat{H}^{n_p}(i) = \sum_{l=1}^{i} \frac{\Delta N^{n_p}(l,.)}{V^{n_p}(l,.)}$$

$$\hat{\zeta}^{n_p}(i) \quad = \quad \hat{h}^{n_p}(i) \prod_{l=1}^{i-1} [1 - \hat{h}^{n_p}(l)].$$

7. Calculamos os erros quadráticos da função intensidade $h^{(p)}$, da função intensidade acumulada $H^{(p)}$ e da função distribuição de probabilidade $\zeta^{(p)}$ para cada uma das iterações, dados respectivamente, por

$$e_h(i) = [\hat{h}^{n_p}(i) - h^{(p)}(i)]^2, \quad e_H(i) = [\hat{H}^{n_p}(i) - H^{(p)}(i)]^2$$

$$e_\zeta(i) = [\hat{\zeta}^{n_p}(i) - \zeta^{(p)}(i)]^2$$

8. Ao final das 10000 iterações, calculamos a média por categoria dos erros quadráticos, os erros quadráticos médios (EQM).

6.2.1 Avaliação da qualidade dos estimadores

Nas Tabelas 6.2, 6.3 e 6.4 apresentamos o erros quadráticos médios da função intensidade, da função intensidade acumulada e da distribuição de probabilidade. Observamos que o erro quadrático médio dos estimadores dos parâmetros do modelo de intensidade multiplicativo apresentam uma diferença em relação aos verdadeiros valores a partir da segunda casa decimal. Essa conclusão é obtida no caso de 3 categorias para amostras pequenas e grandes. Já para 7 e 10 categorias, a diferença dos estimadores em relação aos verdadeiros valores chega a quarta casa decimal para amostras grandes.

Com isso, concluímos que a partir de amostras de tamanho 80 obtemos boas aproximações dos valores dos estimadores dos parâmetros do modelo de intensidade multiplicativo aos valores reais, independente do número de categorias.

Tabela 6.2: *Valores dos erros quadráticos médios associados à função intensidade, à função intensidade acumulada e à distribuição de probabilidade para 3 categorias - sem censura.*

			3 Categorias					
Categoria	n_p	EQM_h	EQM_H	EQM_ζ	n_p	EQM_h	EQM_H	EQM_ζ
1		0,00807	0,00807	0,00807		0,00198	0,00198	0,00198
2	20	0,01514	0,02308	0,01275	80	0,00365	0,00567	0,00308
3		0	0,02290	0,01059		0	0,00567	0,00261
1		0,00320	0,00320	0,00320		0,00159	0,00159	0,00159
2	50	0,00598	0,00926	0,00502	100	0,00290	0,00448	0,00248
3		0	0,00926	0,00428		0	0,00448	0,00207

Tabela 6.3: *Valores dos erros quadráticos médios associados à função intensidade, à função intensidade acumulada e à distribuição de probabilidade para 7 categorias - sem censura.*

				7 Categorias				
Categoria	n_p	EQM_h	EQM_H	EQM_ζ	n_p	EQM_h	EQM_H	EQM_ζ
1		0,00684	0,00684	0,00684		0,00167	0,00167	0,00167
2		0,01491	0,02163	0,01158		0,00364	0,00523	0,00288
3		0,02726	0,04901	0,00951		0,00638	0,01161	0,00232
4	20	0,04819	0,09578	0,00279	80	0,01010	0,02167	0,00071
5		0,08340	0,15973	0,00317		0,01748	0,03854	0,00083
6		0,11977	0,20869	0,00251		0,03217	0,07111	0,00070
7		0,00000	0,12129	0,00147		0,00000	0,06026	0,00055
1		0,00271	0,00271	0,00271		0,00136	0,00136	0,00136
2		0,00590	0,00869	0,00460		0,00296	0,00439	0,00230
3		0,01053	0,01945	0,00376		0,00505	0,00945	0,00186
4	50	0,01702	0,03694	0,00115	100	0,00810	0,01760	0,00056
5		0,02949	0,06614	0,00131		0,01423	0,03194	0,00065
6		0,05605	0,11829	0,00112		0,02476	0,05757	0,00056
7		0,00000	0,07165	0,00074		0,00000	0,05335	0,00046

Tabela 6.4: *Valores dos erros quadráticos médios associados à função intensidade, à função intensidade acumulada e à distribuição de probabilidade para 10 categorias - sem censura.*

				10 Categorias				
Categoria	n_p	EQM_h	EQM_H	EQM_ζ	n_p	EQM_h	EQM_H	EQM_ζ
1		0,00453	0,00453	0,00453		0,00113	0,00113	0,00113
2		0,01114	0,01584	0,00929		0,00275	0,00393	0,00230
3		0,01538	0,03100	0,00704		0,00370	0,00784	0,00173
4		0,01711	0,04883	0,00414		0,00398	0,01158	0,00104
5	20	0,03114	0,07927	0,00561	80	0,00724	0,01858	0,00143
6		0,03005	0,10674	0,00189		0,00652	0,02496	0,00048
7		0,03569	0,12934	0,00161		0,00772	0,03227	0,00041
8		0,06569	0,17366	0,00236		0,01367	0,04583	0,00058
9		0,10688	0,22260	0,00265		0,02476	0,06916	0,00071
10		0,00000	0,11938	0,00225		0,00000	0,06725	0,00085
1		0,00179	0,00179	0,00179		0,00090	0,00090	0,00090
2		0,00439	0,00623	0,00366		0,00225	0,00311	0,00190
3		0,00608	0,01247	0,00281		0,00297	0,00620	0,00139
4		0,00657	0,01895	0,00165		0,00324	0,00953	0,00082
5	50	0,01191	0,03056	0,00231	100	0,00581	0,01533	0,00115
6		0,01059	0,04127	0,00077		0,00507	0,02017	0,00038
7		0,01287	0,05391	0,00065		0,00583	0,02600	0,00031
8		0,02379	0,07825	0,00095		0,01082	0,03668	0,00047
9		0,04216	0,11571	0,00114		0,01907	0,05593	0,00056
10		0,00000	0,09659	0,00128		0,00000	0,05532	0,00070

6.2.2 Teste de hipótese

Nesta seção iremos avaliar a qualidade dos desempenhos dos testes de hipóteses, para uma e duas populações.

Uma população

O desempenho do teste para uma população é avaliado em função da proporção de vezes em que não rejeitamos a hipótese nula, dado que simulamos sob a hipótese $H_0 : H^{(p)}(i) = H_0^{(p)}(i)$, ao nível de confiança de 95%.

Para realizar este estudo de simulação, devemos seguir as indicações apresentadas na seção anterior até o sexto passo. A partir daí, o procedimento é o seguinte:

7. Utilizando os estimadores da função intensidade acumulada e da distribuição de probabilidade, $\hat{H}^{n_p}$ e $\hat{\zeta}^{n_p}$, calculamos as estatísticas do teste para uma população para cada iteração. A estatística é dada por

$$W_0^H = \frac{\sqrt{n_p}\left[\hat{H}^{n_p}(i) - H_0^{(n_p)}(i)\right]}{\sqrt{\sum_{l=1}^{i}[1-\hat{h}^{(n_p)}(l)]\hat{h}^{(n_p)}(l)/\hat{\theta}_l^{(n_p)}}},$$

em que $\hat{\theta}_l^{(n_p)} = \hat{\zeta}^{n_p}(l) + \ldots + \hat{\zeta}^{n_p}(k)$. A estatística W_0^H tem distribuição assintótica normal com média 0 e variância 1.

8. Com nível de significância de 5%, calculamos a porporção de vezes em que não rejeitamos a hipótese H_0.

Na Tabela 6.11 temos os resultados dos testes de hipótese para 3, 7 e 10 categorias e para diferentes tamanhos de amostras (n_p). Observamos que os níveis de confianças empíricos dos testes aproximam-se do nível de confiância adotado de 95% com o aumento do tamanho amostral, conforme esperado.

Tabela 6.5: *Teste de hipótese para a função acumualada até a terceira, sétima e décima categoria - uma amostra.*

	Categorias		
Amostra	3	7	10
20	92,85	90,49	92,36
50	94,32	94,11	94,40
80	94,66	94,69	94,27
100	94,69	94,04	94,53
200	94,80	94,65	94,65
500	95,01	94,79	95,25

Duas populações

Realizamos testes de comparação entre o comportamento das funções intensidades acumuladas das duas populações até a 3°, 7° e 10° categorias. O desempenho do teste para comparação de duas populações é avaliado em função da proporção de vezes em que não rejeitamos a hipótese nula, dado que simulamos sob a hipótese $H_0 : H^{(p)} = H_0^{(p)}$, ao nível de confiança de 95%.

Os passos para este estudo de simulação são apresentados a seguir.

1. Definimos valores para a probabilidade de ocorrência do evento de interesse em cada uma das categorias para as duas populações, $\zeta(1), \ldots, \zeta(k)$.

2. Consideramos amostras de tamanhos $20, 50, 80, 100, 200$ e 500 para as duas populações. Neste estudo de simulação, comparamos apenas amostras de tamanhos iguais.

3. O número de iterações foi fixado em 10000.

4. Para cada iteração geramos duas amostras aleatórias utilizando a distribuição uniforme com parâmetros 0 e 1 e a distribuição de probabilidade.

5. Calculamos os estimadores das funções intensidades $h^{(1)}$ e $h^{(2)}$, da função intensidade acumulada $H^{(1)}$ e $H^{(2)}$ por categoria para cada uma das iterações utilizando as seguintes

equações

$$\hat{h}^{n_p}(i) = \frac{\Delta N^{n_p}(i,.)}{V^{n_p}(i,.)}, \quad \hat{H}^{n_p}(i) = \sum_{l=1}^{i} \frac{\Delta N^{n_p}(l,.)}{V^{n_p}(l,.)}$$

para $p = 1, 2$.

6. Utilizando os estimadores calculamos o valor da estatística de teste para comparação entre as duas populações, dada por

$$W = \frac{\sum_{l=1}^{i} \left(\frac{n_1+n_2}{n_1 n_2}\right)^{1/2} \frac{V^{n_2}(l,.)V^{n_1}(l,.)}{V(l,.)} \left[\hat{h}^{n_1}(l) - \hat{h}^{n_2}(l)\right]}{\sqrt{\hat{\sigma}^2(i)}} \tag{6.1}$$

em que $V(l,.) = V^{n_2}(l,.) + V^{n_2}(l,.)$ e

$$\hat{\sigma}^2(i) = \sum_{l=1}^{i} \frac{U^{n_1,n_2 2}(l,.)}{V^{n_1}(l,.)}[1 - \hat{h}^{n_1 *}(l)]\hat{h}^{n_1 *}(l) + \sum_{l=1}^{i} \frac{U^{n_1,n_2 2}(l,.)}{V^{n_2}(l,.)}[1 - \hat{h}^{n_2 *}(l)]\hat{h}^{n_2 *}(l).$$

A estatística W tem distribuição assintótica normal com média 0 e variância 1.

7. Com nível de significância de 5%, calculamos a porporção de vezes em que não rejeitamos a hipótese H_0.

Na Tabela 6.12 temos os resultados dos testes de hipótese para 3, 7 e 10 categorias e para diferentes tamanhos de amostras ($n_1 = n_2$). Observa-se que os níveis de confianças empíricos dos testes aproximam-se do nível de confiância adotado de 95% com o aumento do tamanho amostral, conforme esperado.

Tabela 6.6: *Teste de hipótese para a comparação entre duas populações até a terceira, sétima e décima categoria - duas amostras.*

	Categorias		
$n_1 = n_2$	3	7	10
20	92,79	90,49	92,36
50	94,33	95,11	95,4
80	95,13	94,69	94,27
100	94,85	94,04	94,53
200	94,88	94,65	94,65
500	94,89	94,79	95,25

6.3 Análise na presença de censura

Nesta seção apresentamos o estudo de simulação para dados multinomiais na presença de censura. Os cálculos dos estimadores e as estatística de testes são iguais aos efetuados na análise sem censura. Portanto, ao descrever os passos para a realização do estudo de simualação na presença de censura estaremos sendo repetitivos em algumas descrições.

1. Definimos valores para a probabilidade de ocorrência do evento de interesse $\zeta^{(p)}(1), \ldots, \zeta^{(p)}(k)$ e para a probabilidade de ocorrência do evento de censura $\eta^{(p)}(1) - \zeta^{(p)}(1), \ldots, \eta^{(p)}(k) - \zeta^{(p)}(k)$, em cada uma das categorias. Estes valores são apresentados na Tabela 6.7. Definimos também valores para a probabilidade do item "sobreviver" até o fim do estudo $\eta^{(p)}(k+1)$: $\eta(4) = 0,8262$, $\eta(8) = 0,7697$ e $\eta(11) = 0,4909$.

2. A partir desses valores, obtemos os valores verdadeiros da função intensidade $h^{(p)}$ e da função intensidade acumulada $H^{(p)}$ para cada categoria utilizando as equações

$$h^{(p)}(i) = \frac{\zeta^{(p)}(i)}{\zeta^{(p)}(i) + \ldots + \zeta^{(p)}(k)}$$

e

$$H^{(p)}(i) = \sum_{l=1}^{i} h^{(p)}(i),$$

Tabela 6.7: *Valores da distribuição de probabilidade ζ, da probabilidade de censura, $\eta - \zeta$, da função intensidade h e da função intensidade acumulada H para 3, 7 e 10 categorias - com censura.*

	3 categorias				7 categorias				10 categorias			
	$\zeta^{(p)}$	$\eta^{(p)} - \zeta^{(p)}$	$h^{(p)}$	$H^{(p)}$	$\zeta^{(p)}$	$\eta^{(p)} - \zeta^{(p)}$	$h^{(p)}$	$H^{(p)}$	$\zeta^{(p)}$	$\eta^{(p)} - \zeta^{(p)}$	$h^{(p)}$	$H^{(p)}$
1	0,05	0,0020	0,05	0,05	0,05	0,0020	0,05	0,05	0,05	0,0020	0,05	0,05
2	0,090	0,0018	0,0949	0,1449	0,09	0,0018	0,0949	0,1449	0,09	0,0018	0,0949	0,1449
3	0,025	0,0050	0,0292	0,1741	0,025	0,0050	0,0292	0,1741	0,025	0,0050	0,0292	0,1741
4					0,02	0,0040	0,0242	0,1983	0,02	0,0040	0,0242	0,1983
5					0,01	0,0030	0,0125	0,2108	0,01	0,0030	0,0125	0,2108
6					0,006	0,0045	0,0076	0,2184	0,006	0,0045	0,0076	0,2184
7					0,005	0,0040	0,0064	0,2248	0,005	0,0040	0,0064	0,2248
8									0,0034	0,1100	0,0044	0,2293
9									0,0034	0,1100	0,0052	0,2344
10									0,002	0,0500	0,2381	0,2381

para $i = 1, \ldots, k$. Os valores obtidos da função intensidade $h^{(p)}$ e da função intensidade acumulada $H^{(p)}$ também são apresentados na Tabela 6.7.

3. Consideramos amostras de tamanhos $20, 50, 80, 100$ para avaliarmos a qualidade dos estimadores e $20, 50, 80, 100, 200, 500$ para avaliarmos a qualidade dos testes de hipóteses.

4. O número de iterações foi fixado em 10000.

5. Para cada iteração geramos uma amostra aleatória da variável multinomial X utilizando a distribuição uniforme com parâmetros 0 e 1.

6. Calculamos os estimadores da função intensidade $h^{(p)}$, da função intensidade acumulada $H^{(p)}$, da função distribuição de probabilidade $\zeta^{(p)}$ e de $\eta^{(p)} - \zeta^{(p)}$.

7. Calculamos o erro quadrático desses parâmetros para cada uma das iterações.

8. Ao final das 10000 iterações, calculamos a média por categoria dos erros quadráticos, obtendo os erros quadráticos médios (EQM).

6.3.1 Avaliação da qualidade dos estimadores

Nas Tabelas 6.8, 6.9 e 6.10 apresentamos o erros quadráticos médios relacionados aos parâmetros do modelo de intensidade multiplicativo. Observamos que o erro quadrático médio destes parâmetros apresentam uma diferença em relação aos valores assumidos a partir da segunda casa decimal. Quando aumentamos o tamanho da amostra de 20 para 80, essa diferença chega à quinta casa decimal. Com isso, concluímos que a partir de amostras de tamanho 80 obtemos boas aproximações dos valores dos estimadores dos parâmetros do modelo de intensidade multiplicativo aos valores assumidos e, assim, verificamos a consistência dos estimadores.

Tabela 6.8: *Valores dos erros quadráticos médios associados à função intensidade, à função intensidade acumulada e à distribuição de probabilidade para 3 categorias - com censura.*

				3 Categorias						
n_p	Categoria	EQM_h	EQM_H	EQM_ζ	$EQM_{\eta-\zeta}$	n_p	EQM_h	EQM_H	EQM_ζ	$EQM_{\eta-\zeta}$
20	1	0,00237	0,00237	0,00237	8,60E-05	80	0,00059	0,00059	0,00059	2,10E-05
	2	0,00448	0,00680	0,00406	8,50E-05		0,00112	0,00170	0,00102	2,10E-05
	3	0,00167	0,00846	0,00124	0,000272		0,00042	0,00211	0,00031	6,90E-05
	4				0,008682					0,002203
50	1	0,00094	0,00094	0,00094	3,20E-05	100	0,00046	0,00046	0,00046	1,70E-05
	2	0,00182	0,00276	0,00164	3,40E-05		0,00091	0,00137	0,00083	1,70E-05
	3	0,00067	0,00340	0,00049	0,000109		0,00033	0,00170	0,00025	5,60E-05
	4				0,003441					0,001717

Tabela 6.9: *Valores dos erros quadráticos médios associados à função intensidade, à função intensidade acumulada e à distribuição de probabilidade para 7 categorias - com censura.*

				7 Categorias						
n_p	Categoria	EQM_h	EQM_H	EQM_ζ	$EQM_{\eta-\zeta}$	n_p	EQM_h	EQM_H	EQM_ζ	$EQM_{\eta-\zeta}$
	1	0,0024	0,0024	0,0024	8,10E-05		0,0006	0,0006	0,0006	2,00E-05
	2	0,0045	0,0070	0,0041	7,80E-05		0,0011	0,0017	0,0010	2,00E-05
	3	0,0017	0,0087	0,0012	0,00027		0,0004	0,0022	0,0003	6,60E-05
20	4	0,0015	0,0102	0,0010	0,00025	80	0,0004	0,0025	0,0003	5,90E-05
	5	0,0008	0,0110	0,0005	0,00018		0,0002	0,0027	0,0001	4,60E-05
	6	0,0005	0,0115	0,0003	0,00030		0,0001	0,0028	0,0001	6,80E-05
	7	0,0004	0,0119	0,0003	0,00024		0,0001	0,0029	0,0001	6,10E-05
	8				0,0113					0,0029
	1	0,0009	0,0009	0,0009	3,20E-05		0,0005	0,0005	0,0005	1,60E-05
	2	0,0018	0,0027	0,0016	3,40E-05		0,0009	0,0014	0,0008	1,70E-05
	3	0,0007	0,0034	0,0005	0,000105		0,0003	0,0017	0,0002	5,50E-05
50	4	0,0006	0,0039	0,0004	9,10E-05	100	0,0003	0,0020	0,0002	4,50E-05
	5	0,0003	0,0042	0,0002	6,70E-05		0,0002	0,0022	0,0001	3,50E-05
	6	0,0002	0,0044	0,0001	0,00011		9,40E-05	0,002291	6,10E-05	5,60E-05
	7	0,0002	0,0046	0,0001	0,00010		8,20E-05	0,002382	5,20E-05	5,00E-05
	8				0,0044					0,0022

Tabela 6.10: *Valores dos erros quadráticos médios associados à função intensidade, à função intensidade acumulada e à distribuição de probabilidade para 10 categorias - com censura.*

				10 Categorias						
n_p	Categoria	EQM_h	EQM_H	EQM_ζ	$EQM_{\eta-\zeta}$	n_p	EQM_h	EQM_H	EQM_ζ	$EQM_{\eta-\zeta}$
	1	0,00236	0,00236	0,00236	8,50E-05		0,00059	0,00059	0,00059	2,00E-05
	2	0,00453	0,00682	0,00411	7,10E-05		0,00112	0,00170	0,00102	2,00E-05
	3	0,00169	0,00849	0,00125	0,00026		0,00041	0,00209	0,00030	6,70E-05
	4	0,00146	0,01001	0,00101	0,00022		0,00035	0,00244	0,00025	5,80E-05
	5	0,00077	0,01084	0,00050	0,00017		0,00019	0,00261	0,00012	4,30E-05
20	6	0,00048	0,01133	0,00030	0,00029	80	0,00012	0,00272	7,70E-05	7,00E-05
	7	0,00041	0,01175	0,00026	0,00025		0,00010	0,00284	6,50E-05	6,30E-05
	8	0,00029	0,01197	0,00018	0,00630		7,00E-05	0,00292	4,40E-05	0,00156
	9	0,00038	0,01234	0,00024	0,00633		9,20E-05	0,00302	5,70E-05	0,00155
	10	0,00029	0,01262	0,00018	0,00305		7,00E-05	0,00308	4,40E-05	0,00077
	11				0,01604					0,00401
	1	0,00093	0,00093	0,00093	3,10E-05		0,00047	0,00047	0,00047	1,50E-05
	2	0,00179	0,00271	0,00162	3,10E-05		0,00091	0,00140	0,00083	1,60E-05
	3	0,00066	0,00338	0,00049	0,00011		0,00033	0,00174	0,00025	5,40E-05
	4	0,00056	0,00396	0,00039	9,10E-05		0,00029	0,00203	0,00020	4,40E-05
	5	0,00031	0,00429	0,00020	7,30E-05		0,00016	0,00219	0,00010	3,50E-05
50	6	0,00019	0,00451	0,00012	0,00011	100	9,60E-05	0,00227	6,20E-05	5,40E-05
	7	0,00016	0,00467	1,00E-04	1,00E-04		8,00E-05	0,00234	5,10E-05	4,70E-05
	8	0,00011	0,00477	6,90E-05	0,00246		5,70E-05	0,00241	3,60E-05	0,00123
	9	0,00015	0,00492	9,20E-05	0,00244		7,20E-05	0,00248	4,50E-05	0,00123
	10	0,00012	0,00502	7,70E-05	0,00123		5,70E-05	0,00252	3,60E-05	0,00062
	11				0,00630					0,00315

6.3.2 Teste de hipótese

Nesta seção iremos avaliar a qualidade dos desempenhos dos testes de hipóteses, para uma e duas populações.

Uma população

Da mesmas forma que no caso sem censura, o desempenho do teste para comparação de duas populações é avaliado em função da proporção de vezes em que não rejeitamos a hipótese nula, dado que simulamos sob a hipótese $H_0 : H^{(p)} = H_0^{(p)}$, ao nível de confiança de 95%.

Para realizar este estudo de simulação, devemos seguir as indicações apresentadas na seção anterior até o sexto passo. A partir daí, o procedimento é o seguinte.

7. Utilizando os estimadores $\hat{H}^{n_p}$ e $\hat{\eta}^{n_p}$ calculamos as estatísticas do teste para uma população para cada iteração. A estatística é dada por

$$W_0^H = \frac{\sqrt{n_p}\left[\hat{H}^{n_p}(i) - H_0^{(n_p)}(i)\right]}{\sqrt{\sum_{l=1}^{i}[1-\hat{h}^{(n_p)}(l)]\hat{h}^{(n_p)}(l)/\hat{\theta}_l^{(n_p)}}},$$

em que $\hat{\theta}^{(n_p)}(l) = \hat{\eta}^{n_p}(l) + \ldots + \hat{\eta}^{n_p}(k+1)$. A estatística W_0^H tem distribuição normal com média 0 e variância 1.

8. Com nível de significância de 5%, calculamos a porporção de vezes em que não rejeitamos a hipótese H_0.

Na Tabela 6.11 temos os resultados dos testes de hipótese para 3, 7 e 10 categorias e para diferentes tamanhos de amostras (n_p). Observa-se que os níveis de confianças empíricos dos testes aproximam-se do nível de confiância adotado de 95% com o aumento do tamanho amostral, conforme esperado.

Tabela 6.11: *Teste de hipótese para a função acumualada até a terceira, sétima e décima categoria - uma amostra.*

	Categorias		
n_p	3	7	10
20	92,85	90,49	92,36
50	94,32	94,11	94,40
80	94,66	94,69	94,27
100	94,69	94,04	94,53
200	94,80	94,65	94,65
500	95,01	94,79	95,25

Duas populações

Realizamos testes de comparação entre o comportamento das funções intensidades acumuladas das duas populações até a 3°, 7° e 10° categorias. Os passos para este estudo de simulação são apresentados a seguir:

1. Assumimos valores para a probabilidade de ocorrência do evento de interesse em cada uma das categorias, para a probabilidade de ocorrência dos eventos de censura e para a probabilidade de um item "sobreviver" até o fim do estudo. Os valores são os mesmos assumidos para a avaliação do estimadores.

2. Consideramos amostras de tamanhos $20, 50, 80, 100, 200$ e 500 para as duas populações. Neste estudo de simulação, comparamos apenas amostras de tamanhos iguais.

3. O número de iterações foi fixado em 10000.

4. Para cada iteração geramos duas amostras aleatórias utilizando a distribuição uniforme com parâmetros 0 e 1.

5. Calculamos os estimadores das funções intensidades $h^{(1)}$ e $h^{(2)}$, da função intensidade acumulada $H^{(1)}$ e $H^{(2)}$ por categoria para cada uma das iterações utilizando as seguintes

equações

$$\hat{h}^{n_p}(i) = \frac{\Delta N^{n_p}(i,.)}{V^{n_p}(i,.)}, \quad \hat{H}^{n_p}(i) = \sum_{l=1}^{i} \frac{\Delta N^{n_p}(l,.)}{V^{n_p}(l,.)}$$

para $p = 1, 2$.

6. Utilizando os estimadores calculamos o valor da estatística de teste para comparação entre as duas populações, dada por

$$W = \frac{\sum_{l=1}^{i} \left(\frac{n_1+n_2}{n_1 n_2}\right)^{1/2} \frac{V^{n_2}(l,.)V^{n_1}(l,.)}{V(l,.)} \left[\hat{h}^{n_1}(l) - \hat{h}^{n_2}(l)\right]}{\sqrt{\hat{\sigma}^2(i)}} \tag{6.2}$$

em que $V(l,.) = V^{n_2}(l,.) + V^{n_2}(l,.)$ e

$$\hat{\sigma}^2(i) = \sum_{l=1}^{i} \frac{U^{n_1,n_2 2}(l,.)}{V^{n_1}(l,.)}[1 - \hat{h}^{n_1}(l)]\hat{h}^{n_1}(l) + \sum_{l=1}^{i} \frac{U^{n_1,n_2 2}(l,.)}{V^{n_2}(l,.)}[1 - \hat{h}^{n_2}(l)]\hat{h}^{n_2}(l).$$

A estatística W tem distribuição assintótica normal com média 0 e variância.

7. Com nível de significância de 5%, calculamos a porporção de vezes em que não rejeitamos a hipótese H_0.

Na Tabela 6.12 temos os resultados dos testes de hipótese para 3, 7 e 10 categorias e para diferentes tamanhos de amostras ($n_1 = n_2$). Observa-se que os níveis de confianças empíricos dos testes aproximam-se do nível de confiância adotado de 95% com o aumento do tamanho amostral, conforme esperado.

Tabela 6.12: *Teste de hipótese para a comparação entre duas populações até a terceira, sétima e décima categoria - duas amostras.*

	Categorias		
$n_1 = n_2$	3	7	10
20	92,79	90,49	92,36
50	94,33	95,11	95,4
80	95,13	94,69	94,27
100	94,85	94,04	94,53
200	94,88	94,65	94,65
500	94,89	94,79	95,25

6.4 Conclusão

Os resultados obtidos indicam que os estimadores dos parâmetros do modelo de intensidade mutiplicativo têm vício desprezível comprovando a consistência dos estimadores. Notamos também que os testes de hipóteses para uma e duas populações forneceram taxas de não rejeição da hipótese nula bem próximas ao nível de confiança de 95%.

Capítulo 7

Aplicação

7.1 Introdução

Consideremos o conjunto de dados de alunos de graduação da USP (Fonte: JúpiterWeb) descrito na Seção 1.1. O interesse neste capítulo é analisar o tempo de evasão (em semestres) desses alunos. Para isso associamos a cada aluno uma variável aleatória multinomial que determina o número de semestres até a evasão, caso o aluno tenha evadido, ou o número de semestres até a ocorrência de algum dos eventos de censura, migração, falecimento ou conclusão, e analisamos seu comportamento ao longo do tempo e sua relação com fatores determinantes, como sexo, área e período do curso.

Na primeira seção foram realizadas análises descritivas. Na segunda seção os dados foram analisados utilizando o modelo de intensidade multiplicativo e a teoria de testes de hipóteses desenvolvidos nessa dissertação e apresentados nos Capítulos 3 e 5, respectivamente.

7.2 Análise Descritiva da Evasão dos Alunos de Graduação da USP

Consideremos apenas os alunos ingressantes por vestibular entre 2000 e 2004. Conforme visto na Seção 1.1, os alunos foram divididos em 5 classes, segundo a última ocorrência observada no sistema Júpter ao fim do estudo, que se deu em dezembro de 2008. As 5 últimas ocorrências

são evasão, conclusão, falecimento, migração e em andamento. Na Tabela 7.1 temos o número de alunos que se econtravam em cada uma dessas categorias ao fim do segundo semestre de 2008.

Tabela 7.1: *Quantidade de alunos em cada uma das classes de última ocorrência (em andamento, conclusão, evasão, falecimento, migração) por ano.*

	Ano de ingresso					
Última ocorrência	2000	2001	2002	2003	2004	Total
Em andamento	255	422	921	1900	3932	7430
Conclusão	5148	5445	5465	5055	3343	24456
Evasão	1481	1248	1278	1179	1102	6288
Falecimento	12	7	11	9	5	44
Migração	311	336	318	381	388	1734
Nro ingressantes	7207	7458	7993	8524	8770	39952

Podemos observar que, dos alunos ingressantes em 2000, 5148 concluíram, 1481 evadiram, 12 faleceram, 311 migraram e 255 estavam em andamento, até o fim do segundo semestre de 2008.

O evento de interesse dentre as últimas ocorrências é a evasão e um dos objetivos é estudar seu comportamento ao longo do tempo. Para isso, observe que um aluno ingressante em 2004 teve menos semestres para evadir do que um aluno ingressante em 2000, 2001, 2002 ou 2003. Desta forma, para realizarmos uma análise comparativa da evasão ao longo dos anos devemos observar o comportamento dos alunos apenas até o décimo semestre. Portanto, não é correto utilizarmos os dados como apresentados na Tabela 7.1 para verificarmos o comportamento da evasão ao longo do tempo. Na Tabela 7.2 temos o número de ingressantes e a quantidade de evasão até o 10° semestre por ano, e na Figura 7.1 temos o gráfico da proporção de evasão até o 10° semestre por ano de ingresso.

Tabela 7.2: *Número e proporção de alunos ingressantes entre 2000 e 2004 que evadiram até o* 10° *semestre após o ingresso.*

Ano de ingresso	2000	2001	2002	2003	2004
Evasão	1222	1022	1133	1123	1094
Ingressantes	7207	7458	7993	8524	8770
Proporção	0,1696	0,1370	0,1418	0,1318	0,1247

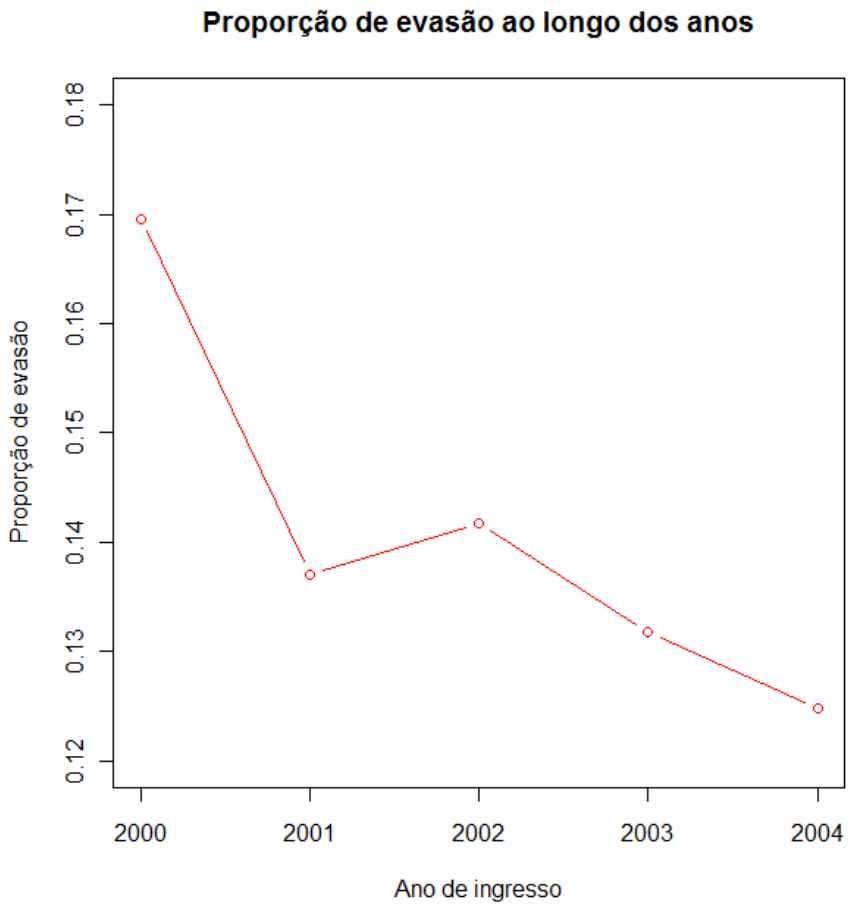

Figura 7.1: *Proporção de evasão dos alunos ingressantes em 2000, 2001, 2002, 2003 e 2004 até o* 10° *semestre.*

Pelo gráfico apresentado na Figura 7.1, verificamos que a evasão vem diminuindo ao longo dos anos. Para verificar se a diferença na evasão ao longo dos anos é significativa, consideremos uma modelo de regressão logística. Os resultados estão na Tabela 7.3. Podemos ver vemos que

existe diferença significativa entre o ano 2000 e todos os demais, e que a evasão é maior no ano 2000, se comparado com os demais anos.

Tabela 7.3: *Resultado do modelo de regressão logística para comparação da evasão ao longo dos anos.*

	Estimador	Desvio Padrão	Estatística do teste	Valor descritivo do teste
Intercepto	-1,58877	0,03139	-50,612	2,00E-16
2001	-0,25138	0,04604	-5,461	4,75E-08
2002	-0,21207	0,04488	-4,726	2,29E-06
2003	-0,29684	0,04484	-6,619	3,61E-11
2004	-0,35949	0,04505	-7,979	1,47E-15

Na Tabela 7.4 temos o número de alunos ingressantes em 2000, 2001, 2002, 2003 ou 2004 que evadiram em cada um dos 10 semestres observados. O semestre com maior número de evasão, dentre os 10 que o aluno tem para evadir, é o 2º semestre, com 29,7% da evasão total dos 10 semestres. Aproximadamente 67% dos alunos evadiram até o 3º semestre. Na Figura 7.2, um gráfico de pareto apresenta essas informações.

Tabela 7.4: *Número de alunos que evadiram por semestre.*

Semestre	*Freqüência*	*Porcentagem*	*Porc. Acumalada*
2	1772	29,70	29,70
1	1616	27,08	56,78
3	560	9,38	66,16
5	416	6,97	73,14
4	398	6,67	79,81
7	312	5,23	85,03
6	307	5,14	90,18
9	220	3,69	93,87
8	213	3,57	97,44
10	153	2,56	100,00

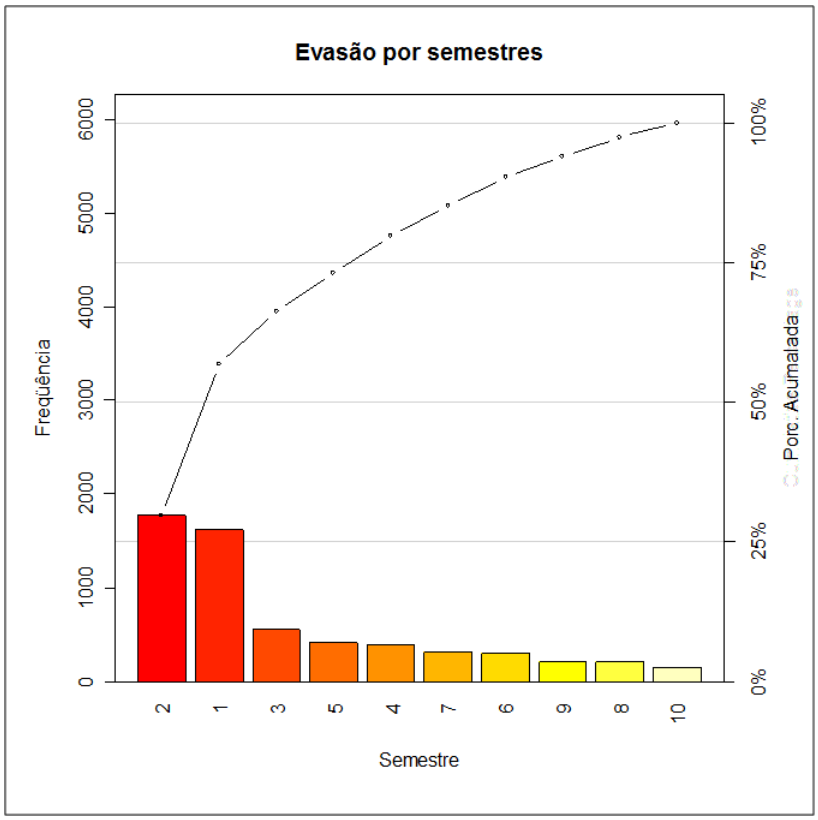

Figura 7.2: *Evasão*

Há evidência de que variáveis como sexo, período, campus, área estão associadas a ocorrência da evasão e, por isso, serão analisadas a seguir.

Inicialmente analisamos a evasão por sexo. Na Tabela 7.5 temos a proporção de evasão por sexo. A proporção foi calculada em relação ao número de ingressantes daquele sexo no respectivo ano. Na Figura 7.3 vemos que a evasão entre as mulheres é proporcionalmente menor do que entre os homens durante os anos estudados e, em ambos os casos, a evasão diminui ao longo do tempo.

Tabela 7.5: *Proporção de evasão por sexo ao longo dos anos.*

	Feminino			Masculino		
Ano de ingresso	Ingressantes	Evasão	Proporção	Ingressantes	Evasão	Proporção
2000	3029	507	0,1674	4178	974	0,2331
2001	3262	399	0,1223	4196	849	0,2023
2002	3467	420	0,1211	4562	858	0,1896
2003	3750	397	0,1059	4774	782	0,1638
2004	3684	360	0,0977	5086	742	0,1459

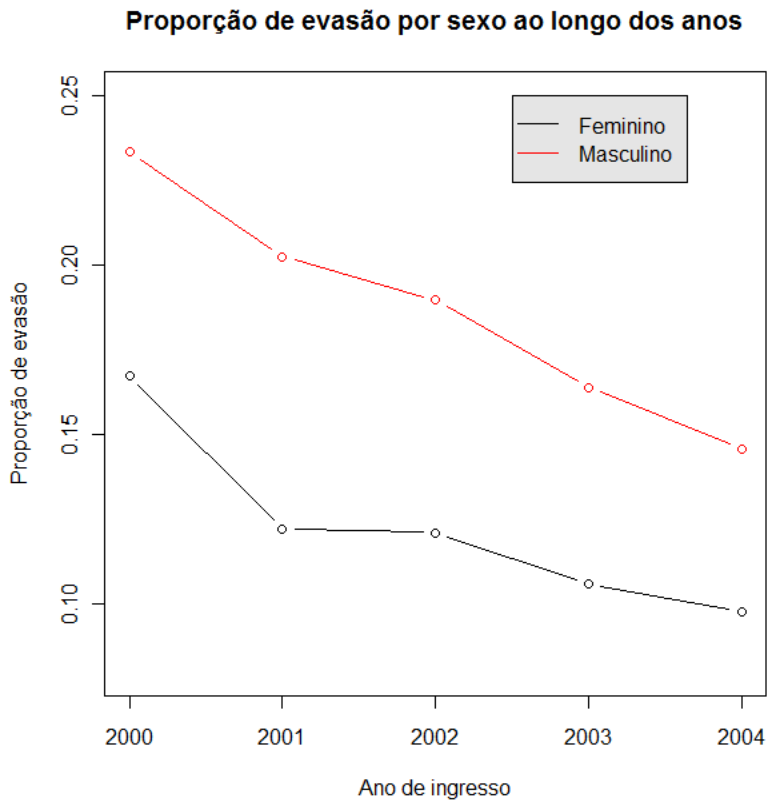

Figura 7.3: *Proporção de evasão ao longo dos anos - sexo.*

Na Tabela 7.6 apresentamos os valores relativos a evasão por área. Podemos ver que a evasão entre os alunos de biólogicas é proporcionalmente menor do que entre os alunos de humanas ou alunos de exatas.

A proporção de evasão por ano relacionada ao período do curso se encontra na Tabela 7.7.

Tabela 7.6: *Proporção de evasão por área, biológicas, exatas e humanas ao longo dos anos.*

	Biológicas			Exatas			Humanas		
Ano de ingresso	Ingres.	Evasão	Prop.	Ingres.	Evasão	Prop.	Ingres.	Evasão	Prop.
2000	1753	139	0,0793	1994	419	0,2101	3460	923	0,2668
2001	1880	104	0,0553	2063	476	0,1823	3515	768	0,2185
2002	2006	133	0,0663	2262	382	0,1689	3725	763	0,2048
2003	2065	101	0,0489	2687	385	0,1433	3772	693	0,1837
2004	2129	130	0,0611	2828	371	0,1312	3813	601	0,1576

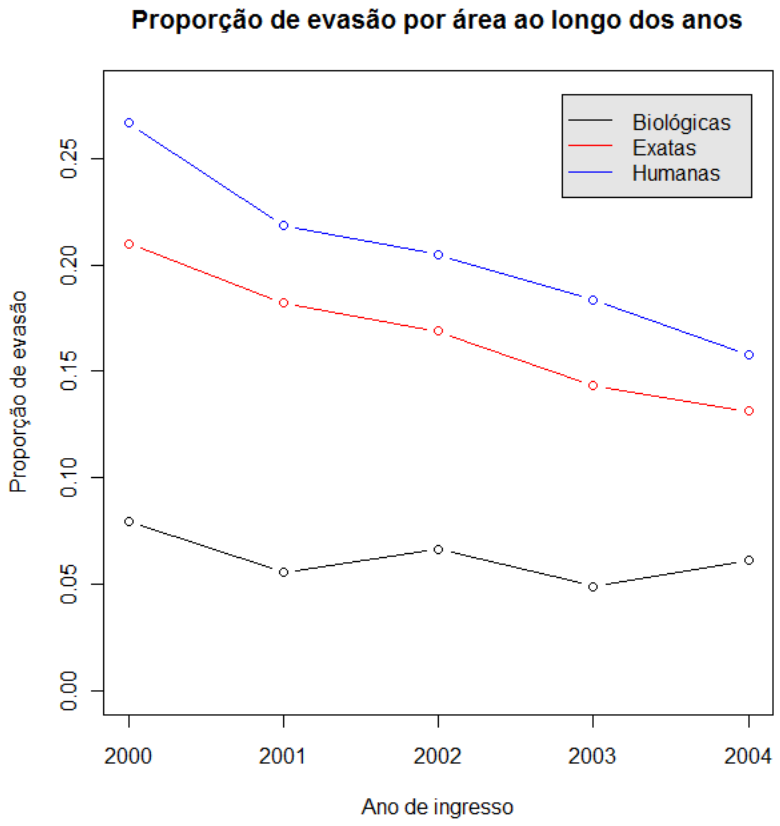

Figura 7.4: *Proporção de evasão ao longo dos anos - área.*

Da mesma forma, a proporção foi calculada em relação ao número de ingressantes em cada período no respectivo ano. Na Figura 7.5 apresentamos o gráfico da proporção de evasão por período ao longo dos anos.

Tabela 7.7: *Proporção de evasão por período ao longo dos anos.*

	Diurno			Integral			Matutino		
Ano de ingresso	Ingres.	Evasão	Prop.	Ingres.	Evasão	Prop.	Ingres.	Evasão	Prop.
2000	643	159	0,2473	3032	309	0,1019	763	157	0,2058
2001	688	125	0,1817	3035	259	0,0853	760	155	0,2039
2002	740	127	0,1716	3145	242	0,0769	772	120	0,1554
2003	948	128	0,1350	3390	228	0,0673	767	115	0,1499
2004	1111	110	0,0990	3442	251	0,0729	735	100	0,1361
	Noturno			Vespertino					
Ano de ingresso	Ingressantes	Evasão	Proporção	Ingressantes	Evasão	Proporção			
2000	2432	757	0,3113	337	106	0,3145			
2001	2627	612	0,2330	348	97	0,2787			
2002	2960	685	0,2314	376	104	0,2766			
2003	3067	617	0,2012	352	91	0,2585			
2004	3115	546	0,1753	367	95	0,2589			

Vimos em cada um dos gráficos e tabelas apresentados nesta seção uma aparente diferença na proporção de evasão por sexo, área e período. No entanto, apenas com as informações apresentadas não podemos verificar se essas diferença são significativas. Para isso, na próxima seção, aplicamos o modelo de intensidade multiplicativo ao conjunto de dados de alunos de graduação da USP e realizamos testes de hipóteses para verificar se as diferenças são significativas.

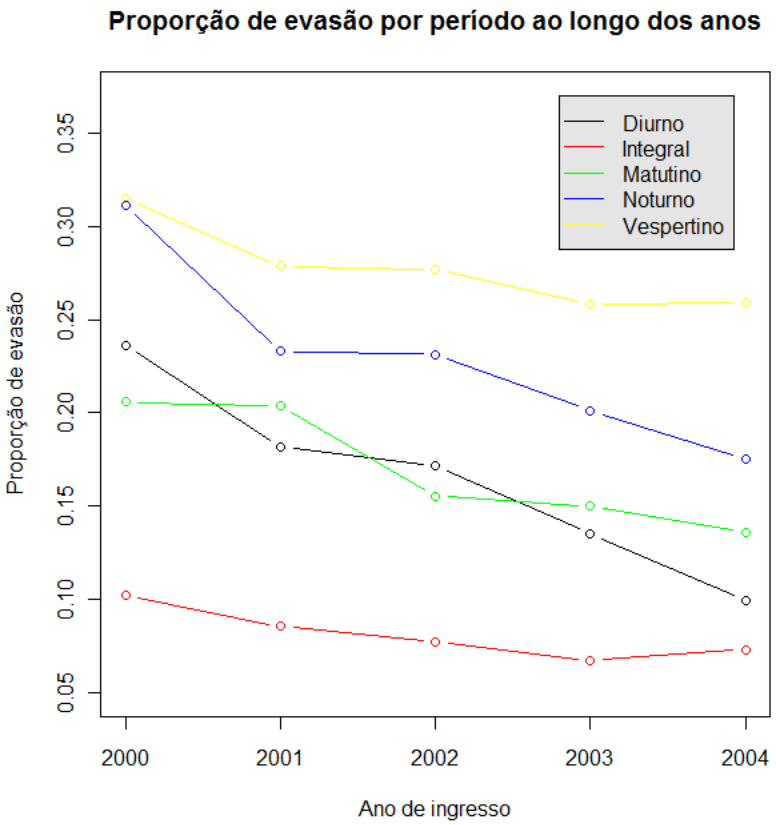

Figura 7.5: *Proporção de evasão ao longo dos anos - período.*

7.3 Aplicação do Modelo de Intensidade Multiplicativo

Nesta seção aplicamos o modelo de intensidade multiplicativo apresentado na Seção 3 aos dados dos alunos de graduação da Universidade de São Paulo. Na seção anterior podemos verificar através do modelo clássido de regressão logística que existe diferença significativa da evasão entre os anos de ingresso e que a evasão vem diminuindo significativamente ao longo dos anos. Utilizando o modelo de intensidade multiplicativo não só conseguimos confirmar esses resultados como também conseguimos identificar se essa diferença entre os anos de ingresso é resultado de uma queda geral da evasão ou se é resultado de uma mudança de comportamento entre semestres. Enquanto o modelo de regressão logística comprara apenas a evasão geral acumulada até o último semestre observado, o modelo de intensidade multiplicativo compara o comportamento da evasão ao longo dos semestres.

Consideremos os alunos ingressantes por vestibular nos anos 2000, 2001, 2003 e 2004. Como

discutido na Seção 7.2, para fazermos uma análise comparativa da evasão ao longo dos anos consideramos apenas os alunos que evadiram até o 10° semestre.

A variável aleatória multinomial $X^{(p)}$, assumindo valores $\{1, \ldots, 11\}$ com distribução de probabilidade $\eta^{(p)}$, representa o semestre da última ocorrência de um aluno ingressante no ano p, para $p = 1, \ldots, 5$ correspondendo aos anos 2000, 2001, 2002, 2003, 2004. Consideremos uma amostra aleatória de tamanho n_p desta variável, $X_1^{(1)}, \ldots X_{n_p}^{(p)}$, em que n_p é o número de alunos ingressantes no ano p. A variável $X_m^{(p)}$, com $m = 1, \ldots, n_p$, representa o número de semestres até a última ocorrência do aluno m ingressante no ano p e pode ser escrita da seguinte forma

$$X_m^{(p)} = W_m^{(p)} \wedge C_m^{(p)}$$

em que $W_m^{(p)}$ e $C_m^{(p)}$ são variáveis aleatórias multinomiais assumindo valores $\{1, \ldots, 11\}$. Essas variáveis são caracterizadas da seguinte forma,

$$W_m^{(p)} = \begin{cases} i, & \text{se o aluno evadiu no semestre } i, \quad i = 1, \ldots, 10; \\ 11, & \text{se o aluno não evadiu até o } 10^\circ \text{ semestre.} \end{cases}$$

$$C_m^{(p)} = \begin{cases} i, & \text{se o aluno migrou, concluiu ou faleceu no semestre } i, \ i = 1, \ldots, 10; \\ 11, & \text{se não ocorreu nenhum dos eventos de censura até } 10^\circ \text{ semestre.} \end{cases}$$

Para esta aplicação, o processo $N^{n_p}(i, .)$ definido na Seção 3.3 pela seguinte equação

$$N^{n_p}(i, .) = \sum_{m=1}^{n_p} N_m^{(p)}(i, .) \tag{7.1}$$

indica quantos alunos ingressantes no ano p evadiram até o semestre i, $i = 1, \ldots, 10$. Portanto, $\Delta N^{n_p}(i, .) = N^{n_p}(i, .) - N^{n_p}(i-1, .)$ indica o número de alunos que evadiram no semestre i. O processo $V^{n_p}(i, .)$ representa o número de alunos ingressantes no ano p sob risco no $i-$ésimo semestre.

A seguir, utilizando o modelo de intensidade multiplicativo, comparamos a evasão ao longo dos anos e sua relação com as variáveis sexo, período e área.

7.3.1 Comparação entre anos

Nas Tabelas 7.8 e 7.9 as colunas intituladas "Sob risco" e "Evasão" correspondem, respectivamente, ao número de alunos sob risco e que evadiram em cada semestre. A coluna "Outros"

corresponde a quantidade de vezes por semestre em que ocorreu os eventos de censura migração, falecimento e transferência. Utilizando esse valores, obtemos os estimadores dos parâmetros do modelo de intensidade multiplicativo.

Para exclarecer como são efetuados os cálculos para se obter os estimadores dos parâmetros do modelo de intensidade multiplicativo, consideremos como exemplo o ano de 2003 ($p = 4$). Temos que $\Delta N^{(4)}(1,.) = 212$ é o número de alunos que evadiram no 1° semestre e $\Delta N^{(4)}(2,.) = 228$ é o número de alunos que evadiram no 2° semestre. $V^{(4)}(1,.) = 8522$ é o número de alunos sob risco no semestre 1. No primeiro semestre um evento de censura ocorreu apenas uma única vez. Portanto, o número de alunos sob risco no segundo semestre é $V^{(4)}(2,.) = 8522-(212+1) = 8309$.

Através destes valores e da equação 3.13 obtemos o valor estimado da função intensidade $\hat{h}^{n_4}$ referente ao ano de 2003 e ao semestre 1,

$$\hat{h}^{n_4}(1) = \frac{\Delta N^{n_4}(1,.)}{V^{n_4}(1,.)} = \frac{212}{8522} = 0,02487 \tag{7.2}$$

Para os demais anos e semestres os cálculos são efetuados de forma análoga.

Depois de calculados os valores de $\hat{h}^{n_p}$, obtemos as estimativas da função intensidade acumulada e da distribuição de probabilidade, utilizando as equações

$$\hat{H}^{n_p}(i) = \sum_{l=1}^{i} \hat{h}^{n_p}(i)$$

e

$$\hat{\zeta}^{n_p}(i) = \hat{h}^{n_p}(i) \prod_{l=1}^{i-1} \hat{h}^{n_p}(l)$$

Nas Tabelas 7.10, 7.11 e 7.12 temos as estimativas da função intensidade, da função intensidade acumulada, da probabilidade de evasão e da probabilidade acumulada por ano para cada um dos 10 semestres analisados.

Tabela 7.8: *Número de alunos que estavam sob risco, que evadiram, ou que tiveram outro tipo de última ocorrência em cada um dos 10 semestres analisados, por ano de ingresso.*

	Ingressantes em 2000			Ingressantes em 2001			Ingressantes em 2002		
Semestre	Sob risco	Evasão	Outros	Sob risco	Evasão	Outros	Sob risco	Evasão	Outros
1	7207	285	0	7458	179	2	7993	179	3
2	6922	257	31	7277	206	26	7811	297	22
3	6634	117	50	7045	105	79	7492	125	69
4	6467	110	40	6861	101	54	7298	92	42
5	6317	119	65	6706	108	60	7164	103	48
6	6133	93	43	6538	86	39	7013	67	39
7	5997	68	40	6413	71	40	6907	82	48
8	5889	47	790	6302	58	859	6777	69	919
9	5052	63	330	5385	61	349	5789	62	423
10	4659	63	1954	4975	47	2024	5304	57	2092
Em andamento	2642			2904			3155		

Tabela 7.9: *Número de alunos que estavam sob risco, que evadiram, ou que tiveram outro tipo de última ocorrência em cada um dos 10 semestres analisados, por ano de ingresso.*

	Ingressantes em 2003			Ingressantes em 2004		
Semestre	Sob risco	Evasão	Outros	Sob risco	Evasão	Outros
1	8524	212	1	8770	194	0
2	8311	228	33	8576	279	13
3	8050	117	90	8284	103	82
4	7843	104	60	8099	98	44
5	7679	100	52	7957	130	65
6	7527	60	45	7762	92	45
7	7422	105	48	7625	90	67
8	7269	79	1021	7468	54	1009
9	6169	79	405	6405	47	433
10	5685	39	2141	5925	7	1956
Em andamento	3505			3964		

Tabela 7.10: *Estimativas da função intensidade, da função intensidade acumulada, da probabilidade de evasão e da probabilidade acumulada em cada um dos 10 semestres analisados, por ano de ingresso, 2000 e 2001.*

	Ingressantes em 2000				Ingressantes em 2001			
Semestre (l)	$h(l)$	$H(l)$	$\zeta(l)$	Acum.	$h(l)$	$H(l)$	$\zeta(l)$	Acum.
1	0,0395	0,0395	0,0395	0,0395	0,0240	0,0240	0,0240	0,0240
2	0,0371	0,0767	0,0357	0,0752	0,0283	0,0523	0,0276	0,0516
3	0,0169	0,0936	0,0156	0,0908	0,0145	0,0668	0,0138	0,0654
4	0,0156	0,1092	0,0142	0,1050	0,0139	0,0807	0,0130	0,0784
5	0,0169	0,1261	0,0151	0,1201	0,0148	0,0955	0,0136	0,0920
6	0,0132	0,1393	0,0117	0,1318	0,0118	0,1073	0,0107	0,1027
7	0,0096	0,1489	0,0083	0,1401	0,0097	0,1170	0,0087	0,1114
8	0,0066	0,1556	0,0057	0,1458	0,0079	0,1249	0,0070	0,1184
9	0,0099	0,1654	0,0084	0,1543	0,0093	0,1342	0,0082	0,1267
10	0,0092	0,1747	0,0078	0,1621	0,0067	0,1409	0,0058	0,1325

Tabela 7.11: *Estimativas da função intensidade, da função intensidade acumulada, da probabilidade e da probabilidade acumulada em cada um dos 10 semestres analisados, por ano de ingresso, 2002 e 2003.*

	Ingressantes em 2002				Ingressantes em 2003			
Semestre (l)	$h(l)$	$H(l)$	$\zeta(l)$	Acum.	$h(l)$	$H(l)$	$\zeta(l)$	Acum.
1	0,0224	0,0224	0,0224	0,0224	0,0249	0,0249	0,0249	0,0249
2	0,0380	0,0604	0,0372	0,0596	0,0274	0,0523	0,0268	0,0516
3	0,0163	0,0767	0,0153	0,0749	0,0142	0,0665	0,0134	0,0651
4	0,0118	0,0885	0,0109	0,0858	0,0125	0,0790	0,0117	0,0767
5	0,0131	0,1016	0,0120	0,0978	0,0120	0,0909	0,0110	0,0878
6	0,0085	0,1102	0,0077	0,1055	0,0072	0,0981	0,0065	0,0943
7	0,0104	0,1205	0,0093	0,1148	0,0125	0,1106	0,0113	0,1056
8	0,0088	0,1293	0,0078	0,1226	0,0094	0,1200	0,0084	0,1141
9	0,0089	0,1382	0,0078	0,1303	0,0106	0,1306	0,0094	0,1235
10	0,0076	0,1458	0,0066	0,1369	0,0049	0,1355	0,0043	0,1277

Tabela 7.12: *Estimativas da função intensidade, da função intensidade acumulada, da probabilidade e da probabilidade acumulada em cada um dos 10 semestres analisados, por ano de ingresso, 2004.*

	Ingressantes em 2004			
Semestre (l)	$h(l)$	$H(l)$	$\zeta(l)$	Acum.
1	0,0221	0,0221	0,0221	0,0221
2	0,0325	0,0547	0,0318	0,0539
3	0,0121	0,0668	0,0115	0,0654
4	0,0114	0,0782	0,0107	0,0761
5	0,0151	0,0933	0,0139	0,0900
6	0,0107	0,1040	0,0098	0,0998
7	0,0104	0,1144	0,0094	0,1092
8	0,0063	0,1207	0,0056	0,1147
9	0,0061	0,1268	0,0054	0,1201
10	0,0008	0,1277	0,0007	0,1209

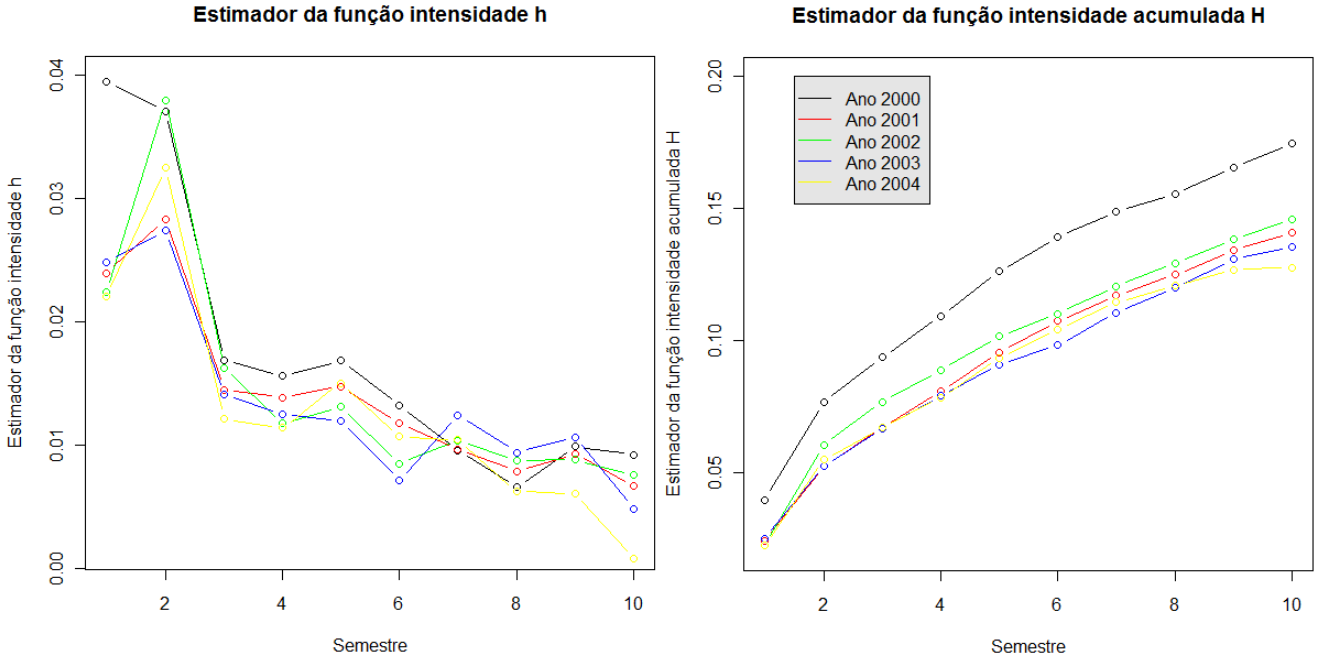

Figura 7.6: *Estimadores da função intensidade e da função intensidade acumualada por semestres ao longo dos anos.*

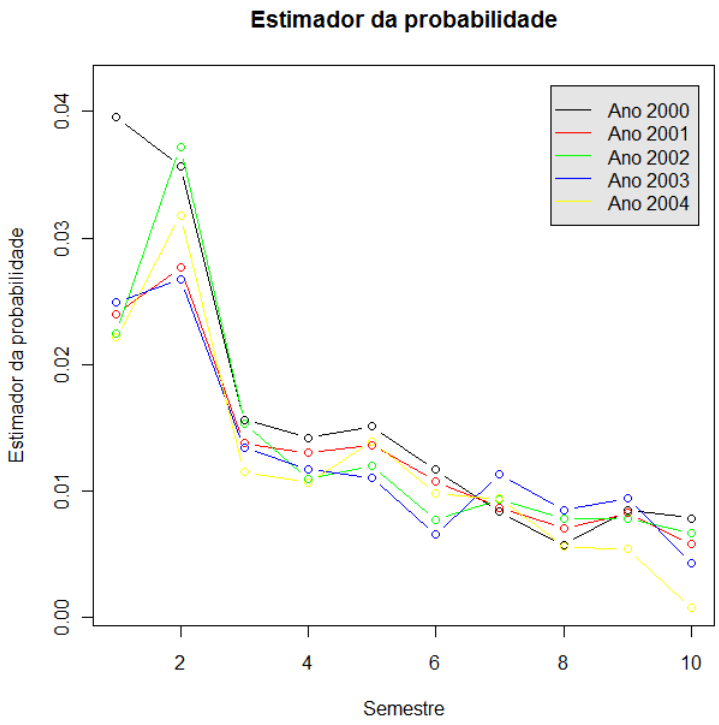

Figura 7.7: *Probabilidade de evasão por semestre de alunos ingressantes em 2000, 2001, 2002 e 2003.*

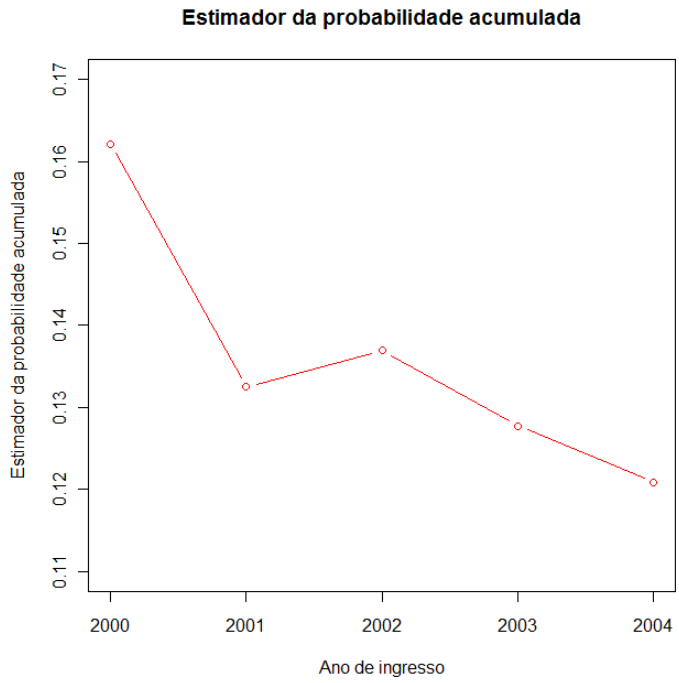

Figura 7.8: *Probabilidade de evasão por semestre de alunos ingressantes em 2000, 2001, 2002 e 2003.*

Para compararmos a evasão entre os anos de ingresso, utilizamos a estatística de *logrank* apresentada na Capítulo 5 para comparação de duas populações. Nosso interesse é comparar o comportamento da evasão entre alunos ingressantes em dois anos diferentes. Assim, nosso interesse é testar as seguintes hipóteses

$$\begin{cases} H_0 : H^{(t)} = H^{(p)} \\ H_1 : H^{(t)} \neq H^{(p)} \end{cases}$$

A estatística de teste utilizada é dada por

$$W = \sum_{l=1}^{i} U^{n_t,n_p}(l,.) \left[\hat{h}^{n_t}(l) - \hat{h}^{n_p}(l)\right] \tag{7.3}$$

em que

$$U^{n_t,n_p}(l,.) = \sqrt{\left(\frac{n_t + n_p}{n_t n_p}\right)} \frac{V^{n_t}(l,.)V^{n_p}(l,.)}{V(l,.)}$$

A estatística W tem distribuição normal com média 0 e variância

$$\hat{\sigma}^2(i) = \sum_{l=1}^{i} \frac{U^{n_t,n_p 2}(l,.)}{V^{n_t}(l,.)}[1 - \hat{h}^{n_t}(l)]\hat{h}^{n_t}(l) + \sum_{l=1}^{i} \frac{U^{n_t,n_p 2}(l,.)}{V^{n_p}(l,.)}[1 - \hat{h}^{n_p}(l)]\hat{h}^{n_p}(l).$$

Com isso, obtemos as estatísticas de testes resultantes da comparação entres os anos e os respectivos valores descritivos do teste. Estes valores são apresentados na Tabela 7.13.

Tabela 7.13: *Testes de hipóteses de comparação dos anos.*

	Estatística	Valor descritivo do teste
2000-2001	5,17	2,31E-07
2000-2002	4,49	7,29E-06
2000-2003	6,21	5,09E-10
2000-2004	7,51	5,98E-14
2001-2002	-0,79	0,427
2001-2003	0,92	0,358
2001-2004	2,20	0,028
2002-2003	1,76	0,079
2002-2004	3,06	0,0022
2003-2004	1,32	0,187

Como realizamos 10 testes de comparação dois a dois, utilizaremos o critério de Bonferroni para definirmos o nível de significância das comparações. Com isso, admitimos um nível de significância de 0,005.

Com este nível de significância, concluímos que existe diferença siginificativa na evasão entre o ano 2000 e todos os demais anos, pois os valores descritivos dos testes são menores que 0,005. Além disso, as estatísticas de teste são positivas, o que implica que a evasão é sempre maior para alunos ingressantes em 2000 quando comparada a alunos ingressantes nos demais anos. Observando a Figura 7.6, podemos verificar que a diferença da evasão entre 2000 e os demais anos não é só resultante de uma queda na evasão ao longo dos anos, mas também de uma mudança de comportamento da evasão durante os 10 semestres analisados. Observando a Figura 7.9, podemos ver que para o ano 2000 temos o mais alto índice de evasão no primeiro semestre, decaindo no segundo e mais ainda no terceiro. No entanto, para os anos 2001 e 2004, a evasão no segundo semestre é maior do que no primeiro.

Há diferença siginificativa também na evasão entre os anos 2002 e 2004 pois o valor descritivo do teste é menor do que 0,005. Além disso, como a estatística de teste é positiva, concluímos

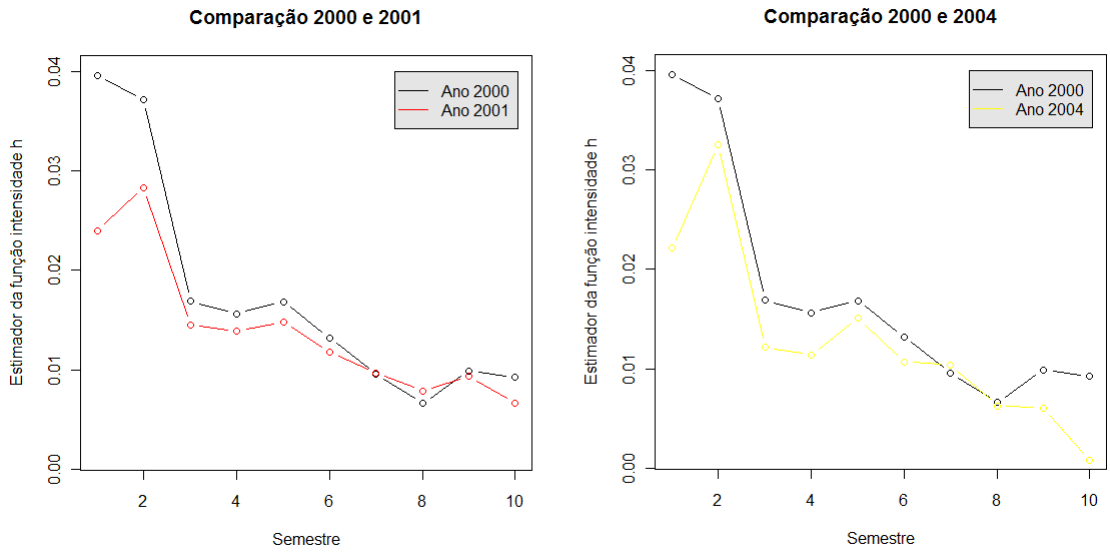

Figura 7.9: *Probabilidade de evasão por semestre de alunos ingressantes em 2000, 2001, 2002 ou 2003.*

que a evasão decaiu de 2002 para 2004 e, observando a Figura 7.6, confirmamos que a diferença é resultante apenas dessa queda da evasão pois não há mudança de comportamento ao longo dos 10 semestres para os ingressos em 2002 e 2004. Entre os demais anos, não existe diferença significativa.

7.3.2 Comparação entre sexos

Realizamos agora uma análise comparativa da evasão entre sexo feminino e masculino. Na Tabela 7.14 temos as estimativas da função intesnsidade, da função intensidade acumulada e da distribuição de probabilidade para sexo masculino e feminino em cada uma das 10 categorias. Na Figura 7.10, temos os gráficos da função intensidade e da função intensidade acumulada para os sexos.

Nosso interesse é verificar se existe diferença significativa da evasão entre homens e mulheres. Para isso, vamos testar as seguintes hipóteses

$$\begin{cases} H_0 : H^{(1)} = H^{(2)} \\ H_1 : H^{(1)} \neq H^{(2)} \end{cases}$$

em que $H^{(1)}$ é a função intensidade referente à população de mulheres e $H^{(2)}$ é a função inten-

Tabela 7.14: *Estimadores da função intensidade, da função intensidade acumulada, da probabilidade e da probabilidade acumulada em cada um dos 10 semestres analisados por sexo.*

	Feminino				Masculino			
l	$h(l)$	$H(l)$	$\zeta(l)$	Acumulada	$h(l)$	$H(l)$	$\zeta(l)$	Acumulada
1	0,01849	0,01849	0,01849	0,01849	0,02949	0,02949	0,02949	0,02949
2	0,02270	0,04119	0,02228	0,04077	0,03962	0,06911	0,03845	0,06794
3	0,01239	0,05358	0,01189	0,05265	0,01664	0,08575	0,01551	0,08345
4	0,00760	0,06118	0,00720	0,05986	0,01283	0,09858	0,01176	0,09521
5	0,00942	0,07061	0,00886	0,06872	0,01430	0,11287	0,01294	0,10814
6	0,00565	0,07626	0,00526	0,07398	0,00850	0,12138	0,00759	0,11573
7	0,00564	0,08190	0,00522	0,07920	0,00912	0,13050	0,00807	0,12379
8	0,00461	0,08651	0,00425	0,08345	0,00603	0,13653	0,00529	0,12908
9	0,00462	0,09113	0,00423	0,08768	0,00633	0,14286	0,00552	0,13459
10	0,00321	0,09434	0,00293	0,09061	0,00437	0,14723	0,00378	0,13837

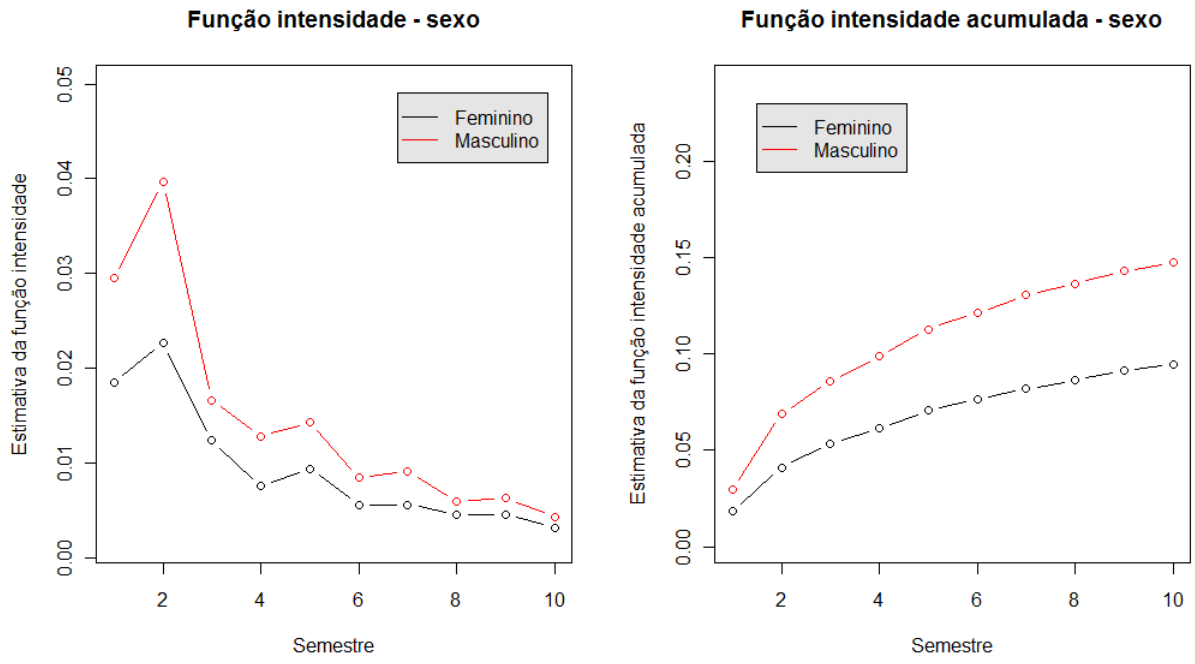

Figura 7.10: *Estimadores da função intensidade e da função intensidade acumualada por semestres - sexo.*

sidade acumulada referente à população de homens. Para testar estas hipóteses utilizamos a estatística de teste

$$W = \sum_{l=1}^{i} U^{n_1,n_2}(l,.)\left[\hat{h}^{n_1}(l) - \hat{h}^{n_2}(l)\right] \tag{7.4}$$

em que

$$U^{n_1,n_2}(l,.) = \sqrt{\left(\frac{n_1+n_2}{n_1 n_2}\right)}\frac{V^{n_1}(l,.)V^{n_2}(l,.)}{V(l,.)}$$

A estatística W que tem distribuição normal com média 0 e variância

$$\hat{\sigma}^2(i) = \sum_{l=1}^{i}\frac{U^{n_1,n_2}{}^2(l,.)}{V^{n_1}(l,.)}[1-\hat{h}^{n_1}(l)]\hat{h}^{n_1}(l) + \sum_{l=1}^{i}\frac{U^{n_1,n_2}{}^2(l,.)}{V^{n_2}(l,.)}[1-\hat{h}^{n_2}(l)]\hat{h}^{n_2}(l).$$

Obtemos o resultado apresentado na Tabela 7.15. Concluímos, portanto, que existe diferença significativa da evasão entre homens e mulheres, e que a evasão entre mulheres é menor do que entre os homens. Além disso, observando o grafico 7.10, vemos que não há mudança de comportamento ao longo dos 10 semestres entre homens e mulheres. Portanto, a diferença está apenas no fato da evasão ser menor entre mulheres.

Tabela 7.15: *Teste de comparação entre sexos.*

Sexo	Estatística	Valor descritivo do teste
Feminino/Maculino	-22,24	1,29E-109

7.3.3 Comparação entre períodos

Agora, realizamos teste de comparação entre os períodos de cursos. Os períodos existentes são:

- Diurno: curso pode ser ministrado de manhã e de tarde.
- Integral: curso pode ser ministrado durante todo o dia: manhã, tarde e noite.
- Matutino: curso ministrado apenas de manhã.
- Noturno: curso ministrado apenas a noite.
- Vespertino: curso ministrado apenas a tarde.

Nas Tabelas 7.16, 7.17 e 7.18 temos as estimativas da função intensidade, da função intensidade acumulada e da probabilidade para cada um dos períodos.

Tabela 7.16: *Estimativas da função intensidade, da função intensidade acumulada, da probabilidade e da probabilidade acumulada em cada um dos 10 semestres analisados por período - Diurno, integral.*

	Diurno				Integral			
l	$h(l)$	$H(l)$	$\zeta(l)$	Acumulada	$h(l)$	$H(l)$	$\zeta(l)$	Acumulada
1	0,0211	0,0211	0,0211	0,0211	0,0086	0,0086	0,0086	0,0086
2	0,0311	0,0522	0,0305	0,0516	0,0145	0,0230	0,0143	0,0229
3	0,0132	0,0655	0,0126	0,0641	0,0090	0,0320	0,0088	0,0317
4	0,0081	0,0736	0,0076	0,0717	0,0051	0,0371	0,0049	0,0366
5	0,0126	0,0862	0,0117	0,0834	0,0075	0,0446	0,0072	0,0438
6	0,0064	0,0926	0,0059	0,0893	0,0034	0,0480	0,0033	0,0471
7	0,0070	0,0996	0,0064	0,0957	0,0046	0,0526	0,0043	0,0514
8	0,0051	0,1046	0,0046	0,1002	0,0027	0,0553	0,0026	0,0540
9	0,0059	0,1105	0,0053	0,1055	0,0035	0,0588	0,0033	0,0573
10	0,0048	0,1154	0,0043	0,1099	0,0023	0,0611	0,0022	0,0595

O interesse aqui é verificar se existe diferença significativa entre os períodos e identificar em quais períodos a evasão é maior. Para isso, testamos as hipóteses

$$\begin{cases} H_0 : H^{(t)} = H^{(p)} \\ H_1 : H^{(t)} \neq H^{(p)} \end{cases}$$

em que t e p correspondem a quaisquer diferentes períodos. Utilizando a estatística já apresentada nesta seção para comparação dos anos, obtemos os resultados apresentados na Tabela 7.19.

Utilizando o critério de Bonferroni, adotamos um nível de significância de 0,005. Com isso, concluímos que apenas entre os período noturno e vespertino não existe diferença significativa. Entre todos os outros há diferença significativa. Observando a Figura 7.11 concluímos que as diferenças existentes são resultantes apenas da queda de evasão entre os período e não de uma variação no comportamento ao longo dos semestres. Concluímos também que a evasão é maior nos períodos noturno e vespertino, seguido dos períodos matutino, diurno e integral.

Tabela 7.17: *Estimadores da função intensidade, da função intensidade acumulada, da probabilidade e da probabilidade acumulada em cada um dos 10 semestres analisados por período - Matutino, vespertino.*

	Matutino				Vespertino			
l	$h(l)$	$H(l)$	Probabilidade	Acumulada	$h(l)$	$H(l)$	Probabilidade	Acumulada
1	0,0302	0,0302	0,0302	0,0302	0,0446	0,0446	0,0446	0,0446
2	0,0381	0,0683	0,0370	0,0672	0,0458	0,0905	0,0438	0,0884
3	0,0149	0,0832	0,0139	0,0811	0,0297	0,1202	0,0271	0,1155
4	0,0117	0,0949	0,0107	0,0918	0,0173	0,1375	0,0153	0,1308
5	0,0133	0,1082	0,0121	0,1039	0,0159	0,1534	0,0139	0,1447
6	0,0080	0,1162	0,0071	0,1110	0,0096	0,1630	0,0082	0,1529
7	0,0053	0,1215	0,0047	0,1157	0,0115	0,1746	0,0098	0,1627
8	0,0048	0,1262	0,0042	0,1200	0,0084	0,1829	0,0070	0,1697
9	0,0051	0,1313	0,0045	0,1244	0,0081	0,1910	0,0067	0,1764
10	0,0028	0,1341	0,0024	0,1268	0,0064	0,1974	0,0052	0,1816

Tabela 7.18: *Estimativas da função intensidade, da função intensidade acumulada, da probabilidade e da probabilidade acumulada em cada um dos 10 semestres analisados por período - Noturno.*

	Noturno			
l	$h(l)$	$H(l)$	$\zeta(l)$	Acumulada
1	0,0384	0,0384	0,0384	0,0384
2	0,0483	0,0868	0,0465	0,0849
3	0,0195	0,1063	0,0179	0,1028
4	0,0162	0,1225	0,0145	0,1173
5	0,0164	0,1389	0,0145	0,1318
6	0,0114	0,1503	0,0099	0,1417
7	0,0114	0,1617	0,0098	0,1515
8	0,0085	0,1702	0,0072	0,1587
9	0,0077	0,1779	0,0065	0,1652
10	0,0053	0,1832	0,0045	0,1696

7.3.4 Comparação entre áreas

Nesta seção comparamos a evasão entre as áreas biológicas, humanas e exatas. Nas Tabelas 7.20 e 7.21 temos as estimativas da função intensidade, da função intensidade acumulada e da

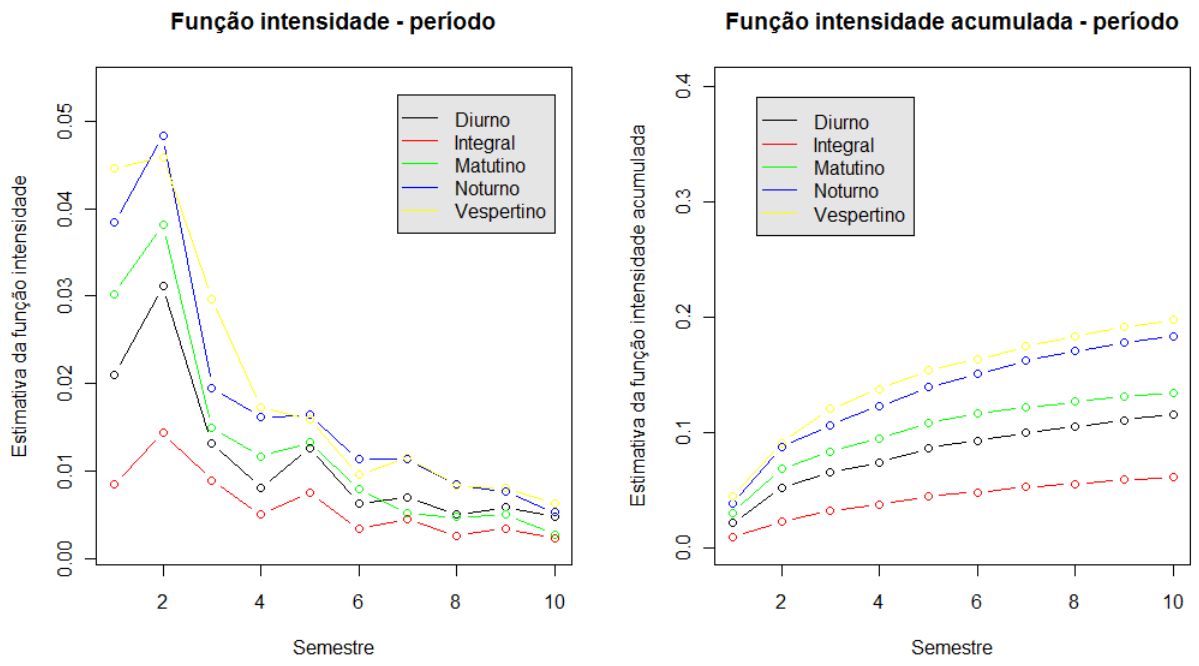

Figura 7.11: *Estimadores da função intensidade e da função intensidade acumualada por semestres - período.*

Tabela 7.19: *Teste de comparação entre períodos.*

Período	Estatística	Valor descritivo do teste
Diurno/Integral	14,41	4,44E-47
Diurno/Matutino	-3,81	0,00014177
Diurno/Noturno	-15,90	6,57E-57
Diurno/Vespertino	-11,03	2,68E-28
Integral/Matutino	-17,53	8,98E-69
Integral/Noturno	-42,94	0,00E+00
Integral/Vespertino	-20,19	1,12E-90
Matutino/Noturno	-10,18	2,41E-24
Matutino/Vespertino	-8,02	1,06E-15
Noturno/Vespertino	-2,04	4,17E-02

distribuição de probabilidade para cada um dos 10 semestres.

Tabela 7.20: *Estimativas da função intensidade, da função intensidade acumulada, da probabilidade e da probabilidade acumulada em cada um dos 10 semestres analisados por área - Biológicas, humanas.*

	Biológicas				Humanas			
Semestre	$h(l)$	$H(l)$	$\zeta(l)$	Acumulada	$h(l)$	$H(l)$	$\zeta(l)$	Acumulada
1	0,00795	0,00795	0,00795	0,00795	0,03588	0,03588	0,03588	0,03588
2	0,01214	0,02009	0,01204	0,01999	0,03979	0,07566	0,03836	0,07423
3	0,00750	0,02758	0,00735	0,02734	0,01700	0,09267	0,01574	0,08998
4	0,00444	0,03202	0,00432	0,03166	0,01382	0,10649	0,01258	0,10256
5	0,00615	0,03818	0,00596	0,03761	0,01448	0,12097	0,01299	0,11555
6	0,00256	0,04074	0,00246	0,04008	0,00881	0,12978	0,00780	0,12335
7	0,00338	0,04411	0,00324	0,04332	0,00890	0,13868	0,00780	0,13114
8	0,00152	0,04564	0,00146	0,04478	0,00691	0,14559	0,00600	0,13715
9	0,00234	0,04798	0,00224	0,04701	0,00701	0,15260	0,00605	0,14320
10	0,00196	0,04994	0,00187	0,04888	0,00503	0,15763	0,00431	0,14751

Tabela 7.21: *Estimativas da função intensidade, da função intensidade acumulada, da probabilidade e da probabilidade acumulada em cada um dos 10 semestres analisados por área - Exatas*

	Exatas			
Semestre	$h(l)$	$H(l)$	$\zeta(l)$	Acumulada
1	0,02116	0,02116	0,02116	0,02116
2	0,03711	0,05827	0,03632	0,05748
3	0,01744	0,07570	0,01644	0,07392
4	0,01065	0,08635	0,00986	0,08378
5	0,01370	0,10005	0,01255	0,09633
6	0,00889	0,10894	0,00803	0,10436
7	0,00922	0,11816	0,00825	0,11262
8	0,00653	0,12468	0,00579	0,11841
9	0,00606	0,13074	0,00534	0,12375
10	0,00363	0,13437	0,00318	0,12693

O interesse é testar as seguintes hipóteses

$$\begin{cases} H_0 : H^{(t)} = H^{(p)} \\ H_1 : H^{(t)} \neq H^{(p)} \end{cases}$$

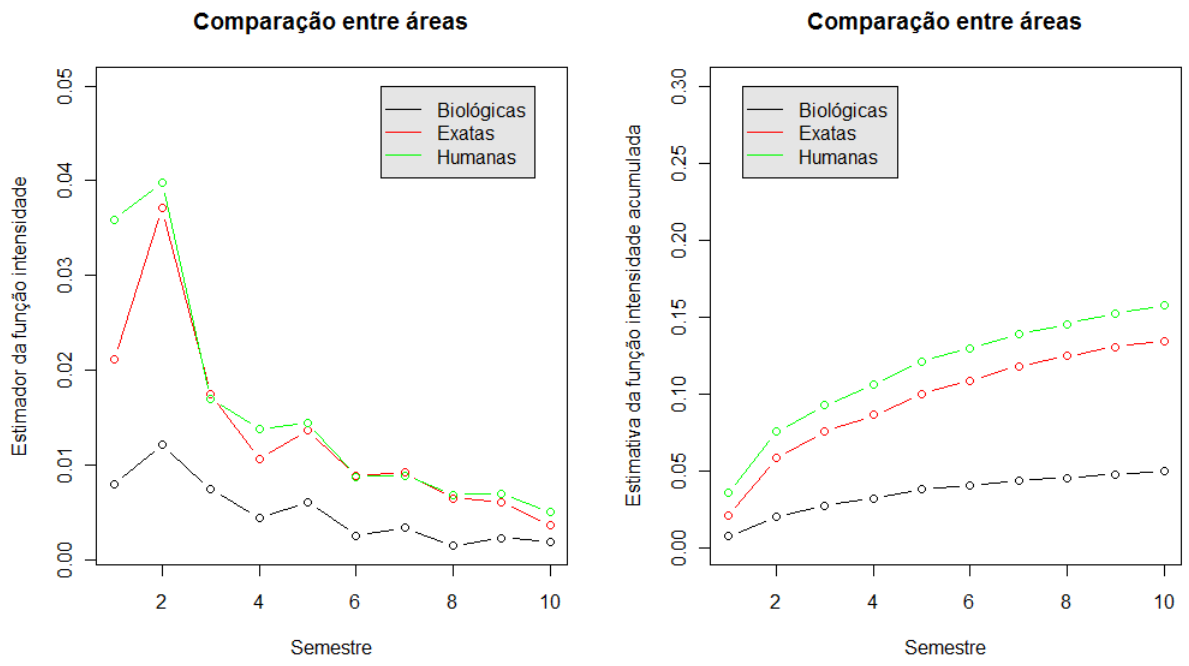

Figura 7.12: *Estimadores da função intensidade e da função intensidade acumualada por semestres - área.*

Para isso, utilizamos a estatística de teste apresentada nesta seção para comparação entre os anos. Os resultados encontram-se na Tabela 7.22. Utilizando o critério de Bonferroni, adotamos um nível de significância de 0,017. Conluímos que existe diferença significativa entre todas as áreas e que a evasão é maior em humanas, seguido de exatas e por último biológicas. Observando a Figura 7.12, concluímos que a diferença resultante entre humanas e exatas é também conseqüência de uma mudança de comportamento entre os semestres, além de uma queda na evasão. Para humanas a evasão no primeiro semestre é alta sendo pouca a diferença entre esta e a evasão no segundo semestre. Para exatas a evasão no primeiro semestre é bem menor do que no segundo semestre.

Entre biológicas e humanas, e biológicas e exatas a diferença reside especificamente na alta queda da evasão.

Tabela 7.22: *Teste de comparação entre áreas.*

Área	Estatística	Valor descritivo do teste
Biológicas/Exatas	-30,19	3,55E-200
Biológicas/Humanas	-42,02	0,00E+00
Exatas/Humanas	-7,81	5,91E-09

7.3.5 Portal Evasão

Ao decorrer deste trabalho desenvolvemos o Portal Evasão. O Portal Evasão, é um sistema interno da USP, onde podemos efetuar análises estatísticas descritivas e específicas a respeito dos alunos de graduação da USP. As análises são efetuadas em linguagem R. O desenvolvimento do sistema ocorreu junto a alunos de pós-graduação da área de Ciências de Computação do ICMC.

No Portal Evasão, até o momento, temos análises básicas como distribuição de freqüência, gráfico de pareto e tabela cruzada, e análises específicas como, tendência da evasão, motivo da evasão e tempo até a conclusão do curso.

O sistema entrou no ar em outubro de 2008 e abaixo temos algumas imagens de sua primeira versão.

Figura 7.13: *Página de login.*

Figura 7.14: *Página principal.*

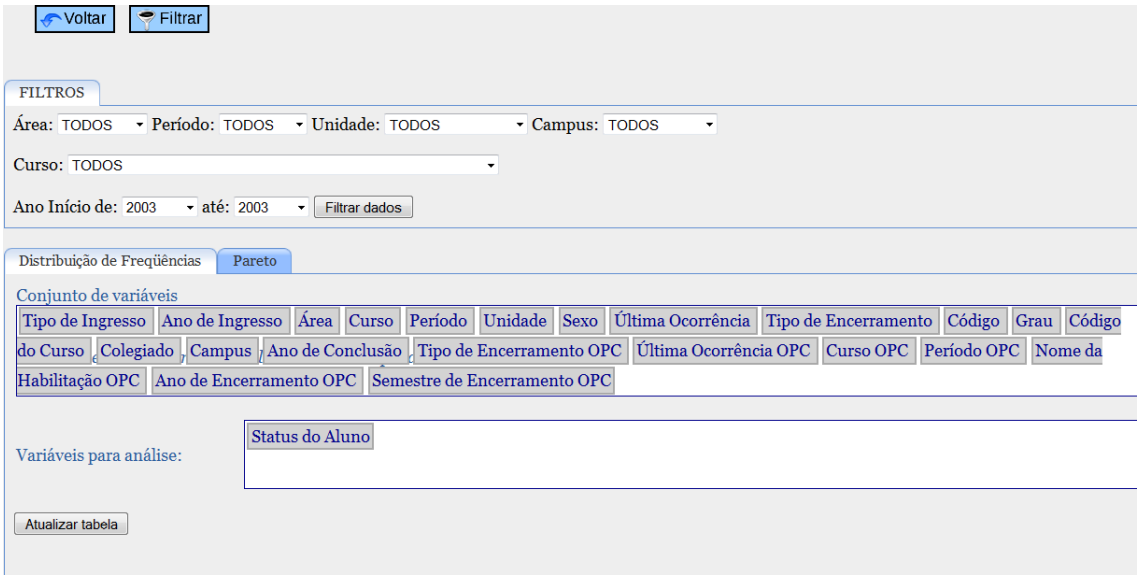

Figura 7.15: *Análise de distribuição de freqüência e filtros possíveis de sereme feitos.*

Linha	Fator	Freqüência	Porcentagem
1	Ativos	48165	55.03
2	Concluíram	25983	29.69
3	Evadiram	10296	11.76
4	Faleceram	59	0.07
5	Migraram	2964	3.39
6	NULL	54	0.06
7	Total	87521	100

Figura 7.16: *Resultado obtido na distribuição de freqüência.*

7.4 Conclusão

Analisando os dados para alunos ingressantes entr 2000 e 2004, verificamos que, de forma geral, a evasão vem diminuindo significativamente ao longo dos anos. Concluímos que existe o fator período do curso influência no comportamento da evasão que é maior nos períodos noturno e vespertino (não há diferença entre esse dois), seguido dos períodos matutino, diurno e integral. Comparando os sexos, a evasão é sempre maior entre os homens. A área do curso

também influência na evasão, e é maior para humanas, seguido de exatas e biológicas.

Capítulo 8

Conclusão e Trabalhos Futuros

Neste trabalho propomos um modelo que é uma extensão do modelo de intensidade multiplicativo desenvolvido por Aalen (1978) e apresentado em Fleming & Harrington (2005), para variáveis aleatórias multinomiais, com as seguintes características:

- Categorias nulas;
- Razão de chances não proporcionais; e
- Dados na presença de censura.

O modelo sugerido é também adequado para análise de dados multinomiais que atendem às hipóteses necessárias no modelo de regressão multinomial clássico.

Estimamos os parâmetros do modelo utilizando o critério de martingale. Demonstramos que os estimadores da função intensidade, da função intensidade acumulada e da distribuição de probabilidade são estimadores consistentes. Verificamos esse resultados realizando um estudo de simulação.

Propomos testes de hipóteses para comparação dos parâmetros para uma e duas populações. A estatística de teste para comparação de duas populações com dados censurados, apresentada neste trabalho, foi proposta por Mantel (1966) e é baseada na estatística de Mantel-Haenszel para tabelas de contigências estratificadas. Esta estatística ficou conhecida como estatística de *logrank* e é provavelmente a mais utilizada como estatística de teste para comparação de duas populações com dados censurados. Realizamos estudos de simulação e verificamos que estes

testes forneceram taxas de não rejeição da hipótese nula bem próximas ao nível de confiança de 95%, para tamanhos amostrais a partir de 80.

Finalmente, aplicamos o modelo proposto para analisar o comportamento da evasão dos alunos de graduação da Universidade de São Paulo. Analisando os dados para alunos ingressantes entr 2000 e 2004, verificamos que, de forma geral, a evasão vem diminuindo significativamente ao longo dos anos. Concluímos que o fator período do curso influencia no comportamento da evasão e que é maior nos períodos noturno e vespertino (não há diferença entre esse dois), seguido dos períodos matutino, diurno e integral. Comparando os sexos, a evasão é sempre maior entre os homens. A área do curso também influência na evasão e é maior para humanas, seguido de exatas e biológicas.

Como proposta de trabalho futuros, listamos os seguintes itens:

- Extensão da teoria assintótica de integrais estocásticas para diversas populações.
- Extensão da teoria assintótica de teste de hipótese para diversas populações.
- Análise da variância assintótica associada à estatística de *logrank*.
- Comparação das estatísticas de *logrank* e Gehan-Wilcoxon sugerida em Fleming & Harrington (2005) para comparação de diversas populações.

Apêndice A

Espaço de Cantor

Algumas demonstrações associadas à construção do espaço de probabilidade para acomodar uma variável aleatória multinomial.

Lema A.1. *A classe $\mathcal{H}$ separa pontos em S^∞, ou seja, dados x e y dois pontos distintos pertencentes a S^∞ existe um conjunto $\Lambda \in \mathcal{H}$ tal que*

$$\mathbb{1}_{\{\Lambda\}}(x) \neq \mathbb{1}_{\{\Lambda\}}(y).$$

Demonstração: Tomemos dois elementos distintos, x e y, em S^∞. Temos que esses elementos são da seguinte forma

$$\begin{aligned} x &= (\alpha_1, \alpha_2, \ldots, \alpha_k, \ldots), \quad \alpha_i = 0 \text{ ou } 1 \\ y &= (\beta_1, \beta_2, \ldots, \beta_k, \ldots), \quad \beta_i = 0 \text{ ou } 1, \quad i = 1, 2, \ldots, k, \ldots. \end{aligned}$$

Como x e y são distintos, então x e y diferem em pelo menos uma coordenada, digamos na j-ésima coordenada. Assim, $\alpha_j \neq \beta_j$ e, conseqüentemente,

$$\pi_j^{-1}(\{\alpha_j\}) \neq \pi_j^{-1}(\{\beta_j\}).$$

Pelo resultado 1,

$$\pi_j^{-1}(\{\alpha_j\}) \cap \pi_j^{-1}(\{\beta_j\}) = \emptyset.$$

Desta forma, basta tomarmos $\Lambda = \pi_j^{-1}(\{\alpha_j\})$ para termos

$$\mathbb{1}_{\{\Lambda\}}(x) = 1 \neq \mathbb{1}_{\{\Lambda\}}(y) = 0.$$

$\mathbf{\Lambda}$

Lema A.2. *A classe $\mathcal{H}$ de subconjuntos de S^∞ é compacta, isto é, para toda seqüência $\{C_n\} \in \mathcal{H}$ tal que $\bigcap_{n=1}^{\infty} C_n = \emptyset$, existe $n_0 \in \mathbb{N}$ tal que*

$$\bigcap_{n=1}^{n_0} C_i = \emptyset.$$

Demonstração: Consideremos a seqüência $\{C_n\}_n \in \mathcal{H}$ tal que

$$\bigcap_{n=1}^{\infty} C_n = 0.$$

Assim, existem pelo menos dois elementos da seqüência $\{C_n\}_n$ que diferem em pelo menos uma coordenada. Ou seja, existem $r,\ s \in \mathbb{N}$ tais que

$$C_r = \pi_k^{-1}(\{1\}) \text{ e } C_s = \pi_k^{-1}(\{0\})$$

para algum $k \in \mathbb{N}$. Tomemos $n_0 = \max\{r, s\}$. Assim,

$$\bigcap_{n=1}^{n_0} C_n = \emptyset.$$

$\mathbf{\Lambda}$

Lema A.3. *A classe Δ separa pontos em S^∞.*

Demonstração: Sejam A e B elementos distintos de Δ. Ou seja,

$$A = \pi_{v_1^A}^{-1}(\{\omega_{v_1^A}\}) \cap \ldots \cap \pi_{v_n^A}^{-1}(\{\omega_{v_n^A}\})$$

e

$$B = \pi_{v_1^B}^{-1}(\{\omega_{v_1^B}\}) \cap \ldots \cap \pi_{v_n^B}^{-1}(\{\omega_{v_n^B}\})$$

com $\pi_{v_k^A}^{-1}(\{\omega_{v_k^A}\}) \neq \pi_{v_k^B}^{-1}(\{\omega_{v_k^B}\})$ para algum $k \in \mathbb{N}$. Como já vimos, $\mathcal{H}$ separa pontos. Logo, existe $\Lambda \in \mathcal{H}$ tal que

$$\pi_{v_k^A}^{-1}(\{\omega_{v_k^A}\}) \in \Lambda \text{ e } \pi_{v_k^B}^{-1}(\{\omega_{v_k^B}\}) \in ??\Lambda.$$

Assim, $A \in \Lambda$ e $B \in ??\Lambda$. Mais do que isso, existe $\Lambda_1 \subset \Delta$ ($\Lambda_1 \subset \Lambda$) tal que $A \in \Lambda_1$, mas $B \in ???\Lambda_1$. Portanto, Δ separa pontos.

$\mathbf{\Lambda}$

Lema A.4. Δ *é uma classe compacta.*

Demonstração: Seja $\{\Delta_n\}_n \subset \Delta$ uma seqüência de conjuntos tais que

$$\bigcap_{n=1}^{\infty} \Delta_n = \emptyset,$$

sendo

$$\Delta_n = \pi_{v_1^n}^{-1}(\{\omega_{v_1^n}\}) \cap \ldots \cap \pi_{v_{k_n}^n}^{-1}(\{\omega_{v_{k_n}^n}\}), \quad \pi_{v_i^n}^{-1}(\{\omega_{v_i^n}\}) \in \mathcal{H}, \quad i = 1, \ldots, k_n, \quad n \in \mathbb{N}.$$

Assim,

$$\bigcap_{n=1}^{\infty} \Delta_n = \bigcap_{n=1}^{\infty} \left(\bigcap_{i=1}^{k_n} \pi_{v_i^n}^{-1}(\{\omega_{v_i^n}\}) \right).$$

Com isso, $\bigcap_{n=1}^{\infty} \Delta_n$ é interseção enumerável de elementos de $\mathcal{H}$. Portanto, podemos escrever esta interseção como

$$\bigcap_{n=1}^{\infty} \Delta_n = \bigcap_{n=1}^{\infty} C_n, \quad C_n \in \mathcal{H}.$$

Como $\mathcal{H}$ é classe compacta, existe $n_0 \in \mathbb{N}$ tal que

$$\bigcap_{n=1}^{n_0} \Delta_n = \bigcap_{n=1}^{n_0} C_n = \emptyset.$$

$\mathbf{\Lambda}$

Proposição A.1. *A classe* $\mathcal{C}$ *é compacta.*

Demonstração: Consideremos em $\mathcal{C}$ uma seqüência de conjuntos $\{C_n\}_{n\in\mathbb{N}}$ com a p.f.i. Para provar que $\mathcal{C}$ é uma classe compacta devemos verificar que $\bigcap_{i=1}^{\infty} C_i \neq \emptyset$. Agora, pela definição da classe $\mathcal{C}$ temos que

$$C_j = A_{j1} \cup A_{j2} \cup \ldots \cup A_{jk_j}, \quad \forall\, j \quad \text{e} \quad A_{j_s} \in \Delta, \quad s = 1, \ldots, k_j.$$

Como Δ é classe compacta, basta obter uma seqüênia $A_{1i_1}, A_{2i_2}, \ldots$ com a p.i.f., porque $\bigcap_{j=1}^{\infty} C_{ji_j} \subseteq \bigcap_{j=1}^{\infty} A_j$. A partir do Lema **??** obtemos esta seqüência, construindo-a indutivamente:

$$C_{1i_1}, C_{2i_2}, \ldots, C_{ni_n}, A_{n+1}, A_{n+2}.$$

$\square$

Referências Bibliográficas

Aalen, O. O. (1978). Nonparametric Inference for a Family of Counting Processes. *Annals of Statistics*, **6**, 701–726.

Begg, C. & Gray, R. (1984). Calculations of polychotomous logistic regression estimates using individualized regressions. *Biometrika*, **71**, 1–18.

Fleming, T. & Harrington, D. (2005). *Counting Processes and Survival Analysis*. Wiley, New Jersey.

Helland, I. S. (1982). Central Limit Theorems for Martingales with Discrete or Continuous Time. *Scand J Statist,*, **9**, 79–94.

Hosmer, D. & Lemeshow, S. (1989). *Applied Logistic Regression*. Wiley, New york, first edition.

Leao, D., Fragoso, M., Val, J. & Andrade, M. (2005). Separable Hausdorff Measurable Randon Spaces. *Statistics e Probability Letters,*, **71**, 347–359.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother*, **50**, 163–70.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

Rom, M. & Cohen, A. (1995). Estimation in the polytomous logistic regression model. *Jornal of Statistical Planning and Inference.*, **43**, 341–353.

Shiryayev, A. (1984). *Probability*. Springer-Verlag, New York.

Stout, W. (1970). *Almost sure Convergence*. Academic Press, Inc.(London) Ltd., New York.