Modelagem Estatística para a Determinação de Resultados de Dados Esportivos

Adriano Kamimura Suzuki

Orientador: Prof. Dr. Francisco Louzada Neto

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

São Carlos

Junho de 2007

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S968me

Suzuki, Adriano Kamimura.
Modelagem estatística para a determinação de resultados de dados esportivos / Adriano Kamimura Suzuki. -- São Carlos : UFSCar, 2007.
70 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2007.

1. Distribuição de Poisson bivariada. 2. Inferência Bayesiana. 3. Medida de Definetti. 4. Bootstrap (Estatística). I. Título.

CDD: 519.24 (20ª)

"É melhor tentar e falhar, que preocupar-se a ver a vida passar. É melhor tentar, ainda que em vão, que sentar-se fazendo nada até o final. Eu prefiro na chuva caminhar, que em dias tristes em casa me esconder. Prefiro ser feliz, embora louco, que em conformidade viver."

Agradeço,

à Deus pela força durante todo esses anos de estudos,

aos meus pais, Luis e Ruth, a quem devo mais essa conquista, pelo incentivo, força e por toda a ajuda que recebi no decorrer dessa pesquisa,

à minha irmã Daniela, meu cunhado Jessé, meu sobrinho Heitor, minhas tias Catarina e Luzia, minha avó Fumiko e a todos os meus familiares por acreditarem em mim e pelo incentivo durante a realização deste trabalho,

à minha namorada Meire pelo apoio, compreensão e por seu amor imprescindível durante todo esse processo,

ao professor Dr. Francisco Louzada Neto pela orientação, pelas idéias e principalmente pelo exemplo de dedicação e disciplina ao trabalho,

à CAPES (Coordenação de Aperfeiçoamente de Pessoal de Nível Superior) pelo apoio financeiro,

à todos os professores do Departamento de Estatística da UFSCar, em especial aqueles aos quais tive o privilégio de ser aluno,

à Universidade Federal de São Carlos - UFSCar,

à meus amigos Erlandson, Fausto, Flávia, Luis Ernesto e Olympio pela amizade e apoio durante o período de realização do mestrado,

a todos que, direta ou indiretamente, acompanharam-me durante o desenvolvimento deste trabalho, o meu muito obrigado.

Resumo

O resultado básico de uma partida de futebol é o seu placar final, que pode ser visto como um vetor aleatório bivariado. Teoricamente e baseando-se na literatura existente podemos argumentar que o número de gols marcados por um time em uma dada partida obedeça a uma distribuição (univariada) de Poisson. Assim, são estudadas as distribuições de Poisson Bivariadas, com destaque para a classe "de Holgate" (1964).

Utilizando como informações os resultados recentes dos times, cujo confronto se queira modelar, foram utilizados vários métodos para a estimação de parâmetros da densidade da classe Poisson Bivariada "de Holgate". A idéia é considerar procedimentos que forneçam as probabilidades de ocorrência de placares, para que assim a probabilidade da ocorrência de um determinado resultado (vitória do time mandante, empate ou derrota) possa ser obtido.

Os parâmetros da distribuição de Poisson Bivariada "de Holgate" são assumidos terem dependência de fatores, tais como ataque, defesa e campo, que possivelmente explicam os números de gols feitos.

Palavras-chave: Distribuição de Poisson Bivariada, Inferência Bayesiana, Medida de Definetti, Bootstrap.

Abstract

The basic result of a soccer game is the final scoreboard, which can be seen as a bivariate random vector. Theoretically and based on existent literature we can argue that the number of marked gols by a team in a game obeys a (univariate) Poisson distribution. Thus, the Bivariate Poisson distributions are studied, in special for the class "of Holgate" (1964).

Using as information the recent results of the teams, whose confrontation we want to model, several methods were used for parameters estimation of the Bivariate Poisson class "of Holgate". The idea is to use procedures that supply the probabilities of occurrence of placares, so that thus the probability of the occurrence of a certain result (team home's victory, draw or defeat) can be calculated properly.

The parameters of Bivariate Poisson distribution "of Holgate" are assumed to have a dependence factors, such as attack, defense and field, that possibly explain the numbers of goals.

keywords: Bivariate Poisson Distribution, Bayesian Inference, Measure of Definetti, Bootstrap.

Sumário

1 Introdução **1**

1.1 Estrutura 2

2 Distribuição de Poisson Bivariada **4**

2.1 A Classe Geral de Distribuições de Poisson Bivariadas 4

2.2 A Classe de Holgate 5

2.3 Considerações Finais 6

3 Métodos de Estimação de Parâmetros para Previsão de Resultados de Jogos de Futebol **7**

3.1 Método SD0 8

3.2 Método SD1 12

3.3 Método Chance I 15

3.4 Método Chance II 19

3.5 Considerações Finais 22

4 Abordagem Bayesiana para Estimação de Parâmetros para Previsão de Resultados de Jogos de Futebol **23**

4.1 Assumindo Independência 24

4.1.1 Método 1: Estimador de Bayes com respeito à perda quadrática . . 24

4.1.2 Método 2: Densidade Preditiva 27

4.2 Método 3: Assumindo Dependência 31

4.3 Considerações Finais 41

5 Medida de DeFinetti e Bootstrap **43**

5.1 Medida de DeFinetti 43

5.2 Bootstrap 44

5.3 Aplicação 45

5.4 Considerações Finais 46

6 Aplicações **47**

6.1 Campeonato Brasileiro 47

6.1.1 Metodologia Utilizada 47

6.2 Considerações Finais 48

7 Considerações Finais e Pesquisas Futuras **49**

Apêndices **50**

A Gráficos **51**

B Alguns programas **62**

Referências Bibliográficas **68**

Capítulo 1

Introdução

Considerando a importância do futebol no mundo, estatísticos de países como a Alemanha, os Estados Unidos e o Canadá começaram a construir modelos estatísticos na tentativa de prever o resultado de partidas de futebol. As primeiras publicações foram feitas por Moroney (1956) e, desde então o número de artigos publicados aumentaram significativamente. Isto certamente se deve a crescente popularidade do futebol.

Em uma partida de futebol vários fatores podem influenciar o seu resultado, tais como: jogar em casa; gramado artificial; torcida e estratégia. Além dos fatores citados, podemos mencionar também o horário de realização da partida, a situação das equipes no campeonato, condições climáticas e atmosféricas, árbitros etc.

Pollard (1986) verificou que um time leva vantagem em **jogar em casa** através da porcentagem de vitórias obtida pelo time da casa em todos os jogos de um torneio (em uma competição onde todos os times jogam a mesma quantidade de partidas como mandante e visitante). Em seu trabalho, verificou que há vantagem em jogar em casa entre as diferentes divisões na liga de futebol. O valor dessa vantagem diminui com a importância da liga.

Barnett e Hilditch (1993), a pedido da Liga Inglesa de Futebol durante os anos 1980-1990, verificou se os quatro clubes ingleses que colocaram **gramados artificiais** em seus estádios teriam uma vantagem ao jogarem em casa. Eles mediram os desempenhos destes quatros times por meio de pontos, gols e resultados das partidas, e concluíram que estes times obtiveram vantagem ao jogarem em casa. Assim, a Liga Inglesa de Futebol proibiu gramados artificiais depois da publicação do artigo de Barnett e Hilditch.

Clark e Norman (1995) e Kuk (1995) em seus trabalhos concluíram sobre a importância dos torcedores (**torcida)** em uma partida de futebol. Os torcedores, na maiora das vezes, incentivam o time de diversas formas como ao cantar o hino do time, gritar o nome dos jogadores, vaiar o adversário etc.

Em uma partida, um time que possui jogadores habilidosos pode ter uma nítida vantagem, porém uma boa **estratégia** pode ser crucial para vencer o jogo. O principal problema encontrado para avaliar a infuência das estratégias dos times em um determinado jogo é a quantidade de dados obtidos de todas as movimentações dos jogadores durante a partida. Para reunir esses dados, há alguns métodos desenvolvidos como o de Franks (1988) no Canadá, de Church e Hughes (1987) na Inglaterra e de Paukku (1994) na Finlândia. Este resultado é bastante aceitável quando os times possuem as mesmas habilidades (técnicas), caso contrário, a diferença de qualidade modificaria o resultado. Realmente um time forte freqüentemente ganhará de um fraco independente da vantagem de jogar em casa.

Em uma partida de futebol, o resultado básico é, logicamente, o seu placar final, que pode ser visto como um vetor aleatório bivariado. Teoricamente e baseando-se na literatura existente podemos argumentar que o número de gols marcados por um time em uma dada partida obedeça a uma distribuição (univariada) de Poisson. Assim, são estudadas as distribuições de Poisson Bivariadas, com destaque para a classe "de Holgate" (1964).

1.1 Estrutura

Nesta seção discutimos o que será apresentado nesta dissertação.

No Capítulo 2 apresentamos a classe geral das distribuições de Poisson Bivariada. Discutimos que a classe de Poisson "de Holgate" é bastante adequada à modelagem de resultados de jogos de futebol, pois possui correlação não-negativa e é a única distribuição de Poisson Bivariada infinitamente divisível. De acordo com Dwass e Teicher (1957), um vetor aleatório X é dito infinitamente divisível se, para qualquer inteiro positivo n, X tem a mesma distribuição de uma soma de n vetores aleatórios independentes e identicamente distribuídos.

No Capítulo 3 consideramos quatro métodos para estimação dos parâmetros. Tais métodos podem ser encontrados em Arruda (2000), entretanto são modificados devido a adição de uma covariável de incidência de crise. Dizemos que um time de futebol atravessa por uma crise quando não há uma boa relação entre os jogadores (na linguagem do futebol dizemos que o grupo está dividido), entre jogadores e treinador (neste caso, por exemplo, os jogadores poderão estar fazendo "corpo mole" ou treinador está criticando jogador(es) e vice e versa) ou entre torcedores e o time. Isso ocorre na maioria das vezes quando o time não consegue obter resultados positivos (vitórias) em um determinado campeonato em disputa.

No Capítulo 4 assumindo uma abordagem Bayesiana e distribuição de Poisson Bivariada "de Holgate" para as variáveis aleatórias X e Y que representam, respectivamente, o número de gols marcados pelo time mandante e visitante, apresentamos três métodos para estimação dos parâmetros, sendo que os dois primeiros assumindo independência entre as variáveis aleatórias.

No Capítulo 5 fizemos reamostragens dos jogos do Campeonato Brasileiro de 2006 considerando as r últimas rodadas, $r = 1, ..., n$ e, utilizamos o método SD1 modificado que apresentamos no Capítulo 3 para as previsões da próxima rodada, ou seja, considerando n jogos realizados, calculamos as previsões da $(n+1)^a$ rodada. Para verificação da qualidade da precisão do método utilizamos a medida de DeFinetti. Construímos também um intervalo de confiança para a medida de DeFinetti.

No Capítulo 6, aplicamos os métodos descritos nos capítulos anteriores e apresentamos as previsões do Campeontado Brasileiro 2005 e 2006.

No Capítulo 7 apresentamos algumas conclusões e algumas propostas de continuidade deste trabalho. Um apêndice foi introduzido com o objetivo de apresentar os gráficos discutidos no Capítulo 4 (Apêndice A). Também apresentamos alguns programas desenvolvidos na implementação computacional para os exemplos de aplicação.

Capítulo 2

Distribuição de Poisson Bivariada

Em uma partida de futebol, o seu placar final (resultado da partida) pode ser visto como um vetor aleatório bivariado. Teoricamente podemos argumentar que o número de gols marcados por um time em uma dada partida obedeça a uma distribuição (univariada) de Poisson. Assim, são estudadas as distribuições de Poisson Bivariadas, com destaque para a classe "de Holgate".

2.1 A Classe Geral de Distribuições de Poisson Bivariadas

Seja X o número de gols marcados pelo time mandante e Y o número de gols marcados pelo time visitante. O vetor aleatório (X, Y) com suporte $\mathbb{N}^2$ segue uma distribuição de Poisson Bivariada (Kocherlakota e Kocherlakota, 1992) se

$$\begin{cases} \sum_{x=0}^{\infty}\sum_{y=0}^{\infty} P(X=x, Y=y) = 1; \\ P(X=x, Y=y) \geq 0, \forall x, y \in \{0,1,2,...\}; \\ \sum_{y=0}^{\infty} P(X=x, Y=y) = \frac{e^{-\lambda_x}(\lambda_x)^x}{x!}, para\ algum\ \lambda_x > 0; \\ \sum_{x=0}^{\infty} P(X=x, Y=y) = \frac{e^{-\lambda_y}(\lambda_y)^y}{y!}, para\ algum\ \lambda_y > 0. \end{cases}$$

A classe de todas as distribuições de Poisson é muito ampla. Dentre as diversas classes de distribuições bivariadas, não são todas as que se adequam à modelagem de resultados de partidas de futebol. De acordo com Arruda (2000) para que uma distribuição bivariada seja considerada adequada, devem ser obedecidos os seguintes critérios:

◦ As distribuições marginais (gols marcados por cada equipe) devem ser Poisson;

◦ A distribuição conjunta deve possuir suporte pleno, ao menos perto da origem;

◦ A distribuição conjunta (e as marginais) devem ser infinitamente divisíveis. De acordo com Dwass e Teicher (1957), um vetor aleatório X é dito infinitamente divisível se, para qualquer inteiro positivo n, X tem a mesma distribuição de uma soma de n vetores aleatórios independentes e identicamente distribuídos.

Justifica-se a necessidade da divisibilidade infinita pelo fato dos jogos de futebol serem disputados em um intervalo contínuo de tempo (diferentemente do vôlei, por exemplo, em que o jogo é dividido em *sets* que duram o tempo necessário até que um time atinja um número fixado de pontos), que pode ser dividido em intervalos menores, preservando-se suas características probabilísticas. Em termos práticos, isso significa que o resultado de um jogo de futebol pode ser considerado tanto como o resultado de um jogo de 90 minutos, quanto a soma dos resultados de dois jogos de 45 minutos cada, ou como a soma dos resultados de noventa jogos de 1 minuto cada.

2.2 A Classe de Holgate

A classe de distribuições bivariadas de Poisson construída por Holgate (1964) é definida como a distribuição conjunta das variáveis $X = R_1 + R_{12}$ e $Y = R_2 + R_{12}$, sendo R_1, R_{12} e R_2 três variáveis aleatórias independentes com distribuições de Poisson univariadas.

A distribuição de Poisson Bivariada "de Holgate" pode ser construída a partir de três processos de Poisson independentes, sendo R_1, R_{12} e R_2 os números de ocorrências de cada processo durante um período de duração comum. As variáveis aleatórias R_1, R_{12} e R_2 têm distribuição de Poisson com médias respectivamente iguais a λ_1, λ_{12} e λ_2. Assim, as variáveis $X = R_1 + R_{12}$ e $Y = R_2 + R_{12}$ tem distribuição conjunta Poisson bivariada. Claramente,

$X \backsim Poisson(\lambda_1 + \lambda_{12})$, enquanto $Y \backsim Poisson(\lambda_2 + \lambda_{12})$.

A densidade conjunta de X e Y pode ser construída da seguinte forma

$$
\begin{aligned}
&P(X=x,Y=y) = P(R_1+R_{12}=x, R_2+R_{12}=y) \\
&= \sum_{k=0}^{\infty} P(R_1+R_{12}=x, R_2+R_{12}=y \mid R_{12}=k)P(R_{12}=k) \\
&= \sum_{k=0}^{\min(x,y)} P(R_1=x-k, R_2=y-k)P(R_{12}=k) \\
&= \sum_{k=0}^{\min(x,y)} P(R_1=x-k)P(R_2=y-k)P(R_{12}=k) \\
&= \sum_{k=0}^{\min(x,y)} \frac{e^{-\lambda_1}\lambda_1^{x-k}}{(x-k)!}\frac{e^{-\lambda_2}\lambda_2^{y-k}}{(y-k)!}\frac{e^{-\lambda_{12}}\lambda_{12}^{k}}{k!} = e^{-(\lambda_1+\lambda_2+\lambda_{12})}\sum_{k=0}^{\min(x,y)}\frac{\lambda_1^{x-k}\lambda_2^{y-k}\lambda_{12}^{k}}{(x-k)!(y-k)!k!} \\
&= e^{-(\lambda_1+\lambda_2+\lambda_{12})}\frac{\lambda_1^{x}}{x!}\frac{\lambda_2^{y}}{y!}\sum_{k=0}^{\min(x,y)}\binom{x}{k}\binom{y}{k}k!\left(\frac{\lambda_{12}}{\lambda_1\lambda_2}\right)^{k}.
\end{aligned}
\tag{2.1}
$$

A Classe Poisson Bivariada "de Holgate" é bastante adequada à modelagem de resultados de jogos de futebol, pois além de satisfazer às três condições expostas anteriormente, em particular possui correlação não-negativa. Esta é a única distribuição de Poisson bivariada infinitamente divisível (Dwass e Teicher, 1957). Em outras palavras, a menos que se queira sacrificar as suposições de distribuições marginais Poisson e suporte $\mathbb{N}^2$, a classe "de Holgate" é a única adequada para a modelagem de resultados de jogos de futebol.

Utilizando a distribuição Poisson Bivariada "de Holgate" para modelagem de partidas de futebol, representamos os números de gols marcados pelas equipes Mandante e Visitante por X e Y, respectivamente. Assim, $X \backsim Poisson(\lambda_1+\lambda_{12})$ e $Y \backsim Poisson(\lambda_2+\lambda_{12})$ então $E[X] = \lambda_1 + \lambda_{12}$ e $E[Y] = \lambda_2 + \lambda_{12}$ que são as esperanças marginais dos gols a serem marcados pelas respectivas equipes em um dado jogo.

2.3 Considerações Finais

Neste capítulo comentamos que a distribuição Poisson Bivariada "de Holgate" é bastante adequada à modelagem de partidas de futebol. No Capítulo 3 apresentamos quatro métodos para estimação dos parâmetros desta distribuição.

Para uniformização da terminologia utilizada neste trabalho, os jogos são escritos sempre na forma Mandante $\times$ Visitante, onde "Mandante" é sempre o time que aparecer a esquerda e "Visitante" o que aparecer à direita de $\times$. Mandante é a terminologia usada para um time que jogar em casa (no seu próprio estádio) e visitante é o time adversário que estiver "viajando"para disputar a partida, ou seja, que está jogando fora de seu estádio.

Capítulo 3

Métodos de Estimação de Parâmetros para Previsão de Resultados de Jogos de Futebol

Neste capítulo, utilizando a distribuição Poisson Bivariada "de Holgate" para modelagem de partidas de futebol, apresentamos quatro métodos para estimação dos parâmetros desta distribuição. Os quatro métodos apresentados a seguir podem ser encontrados em Arruda (2000) entretanto, foram modificados pela adição de uma variável de incidência de crise. Dizemos que um time de futebol atravessa por uma crise quando não há uma boa relação entre os jogadores (na linguagem do futebol dizemos que o grupo está dividido), entre jogadores e treinador (neste caso, por exemplo, os jogadores poderão estar fazendo "corpo mole" ou treinador está criticando jogador(es) e vice e versa) ou entre torcedores e o time. Isso ocorre na maioria das vezes quando o time não consegue obter resultados positivos (vitórias) em um determinado campeonato em disputa.

Para exemplificar vamos utilizar um torneio no qual participaram os times Grêmio, Cruzeiro, Flamengo e São Paulo, considerando que o Grêmio esteja atravessando por uma crise. Neste torneio vamos assumir que ocorreram os seguintes resultados:

Grêmio 0x1 Cruzeiro, em Porto Alegre;

São Paulo 2x1 Flamengo, em São Paulo;

Flamengo 2x1 Grêmio, em Brasília;

Cruzeiro 0x2 São Paulo, em Minas Gerais;

Flamengo 1x2 Cruzeiro, no Rio de Janeiro;

Grêmio 0x3 São Paulo, em Brasília.

A partida final desse torneio será realizada entre os times São Paulo e Cruzeiro, na cidade de São Paulo (os dois times melhores colocados). A idéia é tentar prever a possível chance desses times ganharem.

3.1 Método SD0

O método SD0 (Arruda, 2000) pertence à família de Métodos SD (Soma e Diferença) e, admite que a covariância seja nula, ou seja, à admissão de independência entre as quantidades de gols marcados pelas equipes Mandante e Visitante de cada jogo.

Através das propriedades da distribuição Poisson Bivariada "de Holgate", verificamos que

$Cov(X,Y) = Cov(P_1 + P_{12}, P_2 + P_{12}) = Cov(P_1, P_2) + Cov(P_1, P_{12}) + Cov(P_2, P_{12}) + Var(P_{12}) \overset{ind.}{=} 0 + 0 + 0 + \lambda_{12} = \lambda_{12}$;

$E[X-Y] = E[X] - E[Y] = (\lambda_1 + \lambda_{12}) - (\lambda_2 + \lambda_{12}) = \lambda_1 - \lambda_2$ e

$E[X+Y] = E[X] + E[Y] = (\lambda_1 + \lambda_{12}) + (\lambda_2 + \lambda_{12}) = \lambda_1 + \lambda_2 + 2\lambda_{12} \overset{\lambda_{12}=0}{=} \lambda_1 + \lambda_2$.

Assim, pode-se obter a expressão dos valores dos parâmetros de interesse, λ_1 e λ_2, resolvendo o sistema de equações dado por

$$\begin{cases} E[X-Y] = \lambda_1 - \lambda_2 \\ E[X+Y] = \lambda_1 + \lambda_2 \end{cases}, \tag{3.1}$$

sugerindo os estimadores indiretos

$$\begin{cases} \hat{\lambda_1} = \frac{\hat{E}[X-Y] + \hat{E}[X+Y]}{2} \\ \hat{\lambda_2} = \frac{\hat{E}[X+Y] - \hat{E}[X-Y]}{2} \end{cases}. \tag{3.2}$$

Por sua vez, $E[X-Y]$ e $E[X+Y]$ são estimados através dos modelos lineares dados por:

$$(X+Y)_i = S_i\alpha + \varepsilon_{ai} \tag{3.3}$$

e

$$(X - Y)_i = T_i\beta + \varepsilon_{bi}, \tag{3.4}$$

para $i = 1, 2, 3, ..., n$, onde n é o número de jogos no banco de dados; ε_{ai} e ε_{bi} são erros independentes com médias iguais a 0.

No modelo linear (3.3) , $(X+Y)_i$ é o total de gols marcados (por ambas as equipes) no i-ésimo jogo em questão; o vetor α é composto por $N+2$ parâmetros, sendo um parâmetro associado a cada uma das N equipes constantes do banco de dados, um parâmetro associado ao tipo de local onde o jogo se realiza e mais um parâmetro associado se uma das equipes (ou ambas) está em crise. A matriz-linha S_i possui $N + 2$ elementos, sendo N associados ao *status* de cada equipe em relação ao jogo em questão, um componente que indica o tipo de local em que o jogo se realiza e mais um componente que indica se o time está em crise ou não. O *status* de uma equipe é uma variável de incidência que pode assumir os valores 1 se esta participa do i-ésimo jogo ou 0 se não participa. A atribuição comum do valor 1 para as duas equipes envolvidas no jogo se deve ao fato de que o valor de $(X + Y)_i$ não depende da identificação de qual equipe seja mandante e qual seja visitante. Por exemplo, os resultados 2x1, 3x0, 1x2 e 0x3 significam igualmente a ocorrência de três gols ($(X + Y)_i = 3$).

A componente relativa ao local da realização do jogo também é uma variável de incidência que pode assumir os valores 1 se o jogo foi no campo mandante (e estranho ao visitante) ou 0 se foi em campo neutro, seja ele estranho a ambas as equipes. A componente relativa à crise também é uma variável de incidência que pode assumir os valores 1 se uma das equipes está em crise (ou ambas) ou 0, em caso contrário.

Os métodos da família SD se baseiam em modelos lineares sem interceptos, o que pode ser justificado pelo fato de as partidas de futebol começarem sempre em 0x0 (nunca com um placar iniciado em um valor ξ qualquer).

Para o exemplo proposto de aplicação, no primeiro jogo do torneio temos que o número de gols marcados no primeiro jogo, ou seja, o valor de $(X+Y)_1$ é igual a 0+1=1 e, sendo o vetor α dado (por exemplo em ordem alfabética) por $\begin{bmatrix} \alpha_{Cru} & \alpha_{Fla} & \alpha_{Gre} & \alpha_{SP} & \alpha_{Local} & \alpha_{Crise} \end{bmatrix}^t$, a matriz linha S_1 torna-se igual a $\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$.

Considerando conjuntamente todos os jogos do torneio, $X + Y$ passa a ser o vetor de totais de gols e S a matriz de *status*, local e crise (com uma linha referente a cada jogo,

uma coluna referente a cada equipe, uma coluna referente aos locais dos jogos e a última coluna referente à crise). Para os jogos do torneio, o modelo $(X+Y)_i = S_i\alpha + \varepsilon_{ai}$ é dado por

$$\underbrace{\begin{bmatrix} 1 \\ 3 \\ 3 \\ 2 \\ 3 \\ 3 \end{bmatrix}}_{X+Y} = \underbrace{\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}}_{S} . \underbrace{\begin{bmatrix} \alpha_{Cru} \\ \alpha_{Fla} \\ \alpha_{Gre} \\ \alpha_{SP} \\ \alpha_{Local} \\ \alpha_{Crise} \end{bmatrix}}_{\alpha} + \varepsilon_a,$$

para $\varepsilon_a = (\varepsilon_{a1}, \varepsilon_{a2}, ..., \varepsilon_{an})^t$.

Já no modelo linear (3.4), $(X-Y)_i$ é o número de gols marcados pelo mandante menos o número de gols marcados pelo visitante (diferença de gols marcados) no i-ésimo jogo do banco de dados; o vetor β é composto de $N+2$ parâmetros, sendo um parâmetro associado a cada uma das N equipes constantes do banco de dados, um parâmetro associado ao tipo de local onde o jogo se realiza e mais um parâmetro associado se uma das equipes (ou ambas) está em crise. A matriz-linha T_i possui $N+2$ componentes, sendo N associadas ao *status* de cada equipe em relação ao jogo em questão, uma componente que indica o local em que o jogo se realiza e mais um componente que indica se o time está em crise ou não. Neste modelo, o *status* de uma equipe é uma variável sinalizada de incidência que pode assumir os valores 1 se esta equipe é a mandante do jogo, -1 se é a visitante ou 0 se esta não participa do i-ésimo jogo. É necessário a distinção entre mandante e visitante pelo fato do valor de $(X-Y)_i$ depender diretamente da identificação de qual equipe seja mandante e qual seja visitante. Por exemplo, os resultados 3x2 e 2x3 têm significados completamente diferentes pois $(X-Y)_i = 1$ e $(X-Y)_i = -1$, respectivamente.

A componente relativa ao local da realização do jogo também é uma variável de incidência que pode assumir os valores 1 se o jogo foi no campo mandante (e estranho ao visitante) ou 0 se foi em campo neutro, seja ele estranho a ambas as equipes. A componente relativa à crise também é uma variável de incidência que pode assumir os valores 1 se uma das equipes está em crise (ou ambas) ou 0, em caso contrário.

Para o exemplo proposto de aplicação temos que para o quarto jogo o valor $(X-Y)_4$ vale $0-2=-2$ e, sendo o vetor β (por exemplo em ordem alfabética de times) dado por $\left[\begin{array}{cccccc} \beta_{Cru} & \beta_{Fla} & \beta_{Gre} & \beta_{SP} & \beta_{Local} & \beta_{Crise} \end{array}\right]^t$, a matriz linha T_4 torna-se igual a $T_4 = \left[\begin{array}{cccccc} 1 & 0 & 0 & -1 & 1 & 0 \end{array}\right]$.

Considerando conjuntamente todos os jogos do torneio, $X-Y$ passa a ser o vetor de diferenças de gols e T a matriz de *status*, local e crise (com uma linha referente a cada jogo, uma coluna referente a cada equipe, uma coluna referente aos locais dos jogos e a última coluna referente à crise). Para os jogos do torneio, o modelo $(X-Y)_i = T_i\beta + \varepsilon_{bi}$ é dado por

$$\underbrace{\left[\begin{array}{c} -1 \\ 1 \\ 1 \\ -2 \\ -1 \\ -3 \end{array}\right]}_{X-Y} = \underbrace{\left[\begin{array}{cccccc} -1 & 0 & 1 & 0 & 1 & 1 \\ 0 & -1 & 0 & 1 & 1 & 0 \\ 0 & 1 & -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 & 1 & 0 \\ -1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 & 1 \end{array}\right]}_{T} . \underbrace{\left[\begin{array}{c} \beta_{Cru} \\ \beta_{Fla} \\ \beta_{Gre} \\ \beta_{SP} \\ \beta_{Local} \\ \beta_{Crise} \end{array}\right]}_{\beta} + \varepsilon_b,$$

para $\varepsilon_b = (\varepsilon_{b1}, \varepsilon_{b2}, ..., \varepsilon_{bn})^t$.

Dessa forma, no exemplo de aplicação, os estimadores de λ_1 e λ_2 para o jogo final entre São Paulo e Cruzeiro, construídos em (3.2) podem ser calculados a partir de $\hat{E}[X+Y] = S_7.\hat{\alpha}$ e $\hat{E}[X-Y] = T_7.\hat{\beta}$.

Então, para o jogo final entre São Paulo e Cruzeiro, realizado na cidade de São Paulo, os vetores S_7 e B_7 são dados por $S_7 = [\ 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0\ ]$ e $T_7 = [\ -1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0\ ]$, de onde obtêm-se as previsões pontuais:

$\hat{E}[X+Y] = S_7.\hat{\alpha} = \hat{\alpha}_{SP} + \hat{\alpha}_{Cru} + \hat{\alpha}_{Local}$ e $\hat{E}[X-Y] = S_7.\hat{\beta} = \hat{\beta}_{SP} - \hat{\beta}_{Cru} + \hat{\beta}_{Local}$, onde X e Y são, respectivamente, o número de gols marcados pelo São Paulo e Cruzeiro no jogo final realizado em São Paulo.

Da teoria de mínimos quadrados e de matrizes inversas generalizadas de Moore-Penrose (Venables e Ripley, 1999) tem-se, admitindo pesos iguais, as estimativas $\hat{\alpha}$ e $\hat{\beta}$ dadas por

$\hat{\alpha} = (S'S)^{-1}S'(X+Y) = \left[\begin{array}{cccccc} 1.750 & 2.500 & 0.375 & 2.000 & -1.500 & 0.375 \end{array}\right]^t$ e

$\hat{\beta} = (T'T)^{-1}T'(X-Y) = \left[\begin{array}{cccccc} -0.097 & -0.282 & -1.153 & 1.532 & 1.532 & 1.532 \end{array}\right]^t$.

Logo, $\hat{E}[X+Y] = \hat{\alpha}_{SP} + \hat{\alpha}_{Cru} + \hat{\alpha}_{Local} = 2.000 + 1.750 - 1.500 = 2.250$ e
$\hat{E}[X-Y] = \hat{\beta}_{SP} - \hat{\beta}_{Cru} + \hat{\beta}_{Local} = 1.532 + 0.097 - 0.518 = 1.11.$

Assim, obtém-se as estimativas
$\hat{\lambda}_{SP} = \frac{\hat{E}[X+Y]+\hat{E}[X-Y]}{2} = \frac{2.25+1.11}{2} = 1.68$ e
$\hat{\lambda}_{Cru} = \frac{\hat{E}[X+Y]-\hat{E}[X-Y]}{2} = \frac{2.25-1.11}{2} = 0.57.$

Então, o "placar de mínimos quadrados" para o jogo final seria (1,68 x 0,57).

A partir dos valores de $\hat{\lambda}_{SP}$ e $\hat{\lambda}_{Cru}$ pode-se calcular a probabilidade da vitória do São Paulo, fazendo a soma de todas as probabilidades $P(x, y)$ quando $x > y$. Analogamente, para a vitória do Cruzeiro ($x < y$) e empate ($x = y$). Feito isso, chegamos às seguintes probabilidades:

P[vitória do São Paulo] = 0.6449;

P[empate] = 0.2332;

e P[vitória do Cruzeiro] = 0.1219.

3.2 Método SD1

O Método SD1 (Arruda, 2000) pertence a Família de Métodos SD (Soma e Diferença) e, admite-se que a covariância seja não nula.

Temos que $E[(X+Y)^2]-(E[X+Y])^2 = Var[X+Y] = Var[X]+Var[Y]+2Cov[X,Y] = \lambda_1 + \lambda_{12} + \lambda_2 + \lambda_{12} + 2\lambda_{12}$ e, pelas propriedades da distribuição Poisson Bivariada "de Holgate", pode-se obter a expressão dos valores dos parâmetros de interesse λ_1, λ_2 e λ_{12} na forma do sistema de equações

$$\begin{cases} E[X-Y] = \lambda_1 - \lambda_2 \\ E[X+Y] = \lambda_1 + \lambda_2 + 2\lambda_{12} \\ E[(X+Y)^2] - (E[X+Y])^2 = \lambda_1 + \lambda_2 + 4\lambda_{12} \end{cases}, \tag{3.5}$$

sugerindo os estimadores indiretos

$$\begin{cases} \hat{\lambda}_1 = \frac{\hat{E}[X-Y]+2\hat{E}[X+Y]-\{\hat{E}[(X+Y)^2]-(\hat{E}[X+Y])^2\}}{2} \\ \hat{\lambda}_2 = \frac{2\hat{E}[X+Y]-\hat{E}[X-Y]-\{\hat{E}[(X+Y)^2]-(\hat{E}[X+Y])^2\}}{2} \\ \hat{\lambda}_{12} = \frac{\{\hat{E}[(X+Y)^2]-(\hat{E}[X+Y])^2\}-\hat{E}[X+Y]}{2} \end{cases} . \tag{3.6}$$

Neste método $E[X-Y]$ e $E[X+Y]$ são estimados através de modelos lineares com a mesma estrutura do método SD 0 (ver as expressões (3.3) e (3.4)).

Assim, $E[(X+Y)^2]$ é estimado através de modelo linear dado por

$$[(X+Y)^2]_i = S_i\gamma + \varepsilon_{ci}, \tag{3.7}$$

para $i = 1, 2, 3, ..., n$, onde n é o número de jogos no banco de dados; ε_{ci} são erros independentes com médias iguais a 0.

No modelo linear (3.7), $[(X+Y)^2]_i$ é o quadrado do total de gols marcados (por ambas as equipes) no i-ésimo jogo da amostra; o vetor γ é composto por $N+2$ parâmetros, sendo um parâmetro associado a cada uma das N equipes constantes do banco de dados, um parâmetro associado ao tipo de local onde o jogo se realiza e mais um parâmetro associado se uma das equipes (ou ambas) está em crise. A matriz-linha S_i é construída da mesma forma no modelo (3.3). A atribuição do valor 1 para as duas equipes envolvidas no jogo se deve ao fato de que o valor de $[(X+Y)^2]_i$ não depende da identificação de qual equipe seja mandante e qual a visitante. Por exemplo, os resultados 3x0, 2x1,1x2 e 0x3 significam igualmente a ocorrência de três gols e assim, de $[(X+Y)^2]_i = 9$.

Para o exemplo proposto de aplicação temos, por exemplo, no quarto jogo do torneio $[(X+Y)^2]_4 = (0+2)^2 = 4$ e sendo o vetor γ dado por
$\begin{bmatrix} \gamma_{Cru} & \gamma_{Fla} & \gamma_{Gre} & \gamma_{SP} & \gamma_{Local} & \gamma_{Crise} \end{bmatrix}^t$, a matriz linha S_4 torna-se igual a
$S_4 = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$.

Assim, ao considerar conjuntamente todos os jogos do torneio, $(X+Y)^2$ torna o vetor de quadrados totais de gols e S uma matriz de *status,* local e crise (com uma linha referente a cada jogo, uma coluna referente a cada equipe, uma coluna referente aos locais dos jogos e a última coluna referente à crise). Então, para os jogos do torneio, o modelo (3.7) é dado por

$$\underbrace{\begin{bmatrix} 1 \\ 9 \\ 9 \\ 4 \\ 9 \\ 9 \end{bmatrix}}_{(X+Y)^2} = \underbrace{\begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{bmatrix}}_{S} . \underbrace{\begin{bmatrix} \gamma_{Cru} \\ \gamma_{Fla} \\ \gamma_{Gre} \\ \gamma_{SP} \\ \gamma_{Local} \\ \gamma_{Crise} \end{bmatrix}}_{\gamma} + \varepsilon_c,$$

para $\varepsilon_c = (\varepsilon_{c1}, \varepsilon_{c2}, ..., \varepsilon_{cn})^t$.

Os estimadores de λ_1 , λ_2 e λ_{12} para o jogo final entre São Paulo e Cruzeiro, construídos em (3.6) podem ser calculados a partir de $\hat{E}[X+Y] = S_7.\hat{\alpha}$, $\hat{E}[X-Y] = T_7.\hat{\beta}$ e $\hat{E}[(X+Y)^2] = S_7.\hat{\gamma}$.

Então, para o jogo final entre São Paulo e Cruzeiro, realizado na cidade de São Paulo, os vetores S_7 e B_7 são dados por $S_7 = [\ 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0\ ]$ e $T_7 = [\ -1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0\ ]$, de onde obtêm-se as previsões pontuais:

$\hat{E}[X+Y] = S_7.\hat{\alpha} = \hat{\alpha}_{SP} + \hat{\alpha}_{Cru} + \hat{\alpha}_{Local}$, $\hat{E}[X-Y] = T_7.\hat{\beta} = \hat{\beta}_{SP} - \hat{\beta}_{Cru} + \hat{\beta}_{Local}$ e $\hat{E}[(X+Y)^2] = S_7.\hat{\gamma} = \hat{\gamma}_{Cru} + \hat{\gamma}_{SP} + \hat{\gamma}_{Local}$, para X e Y representando, respectivamente, o número de gols marcados pelo São Paulo e Cruzeiro no jogo final realizado em São Paulo.

Da teoria de mínimos quadrados e de matrizes inversas generalizadas de Moore-Penrose (Venables e Ripley, 1999) tem-se, admitindo pesos iguais, as estimativas $\hat{\alpha}$, $\hat{\beta}$ e $\hat{\gamma}$ dadas por

$\hat{\alpha} = (S'S)^{-1}S'(X+Y) = \left[\ 1.750 \quad 2.500 \quad 0.375 \quad 0.375 \quad -1.500 \quad 0.375\ \right]^t$,

$\hat{\beta} = (T'T)^{-1}T'(X-Y) = \left[\ -0.097 \quad -0.282 \quad -1.153 \quad 1.532 \quad -0.518 \quad -0.518\ \right]^t$ e

$\hat{\gamma} = (S'S)^{-1}S'[(X+Y)^2] = \left[\ 4.750 \quad 8.500 \quad 0.875 \quad 6.000 \quad -5.500 \quad 0.875\ \right]^t$.

Logo, $\hat{E}[X+Y] = \hat{\alpha}_{SP} + \hat{\alpha}_{Cru} + \hat{\alpha}_{Local} = 2 + 1.75 - 1.5 = 2.25$,

$\hat{E}[X-Y] = \hat{\beta}_{SP} - \hat{\beta}_{Cru} + \hat{\beta}_{Local} = 1.532 - (-0.097) - 0.518 = 1.111$ e

$\hat{E}[(X+Y)^2] = S_7.\hat{\gamma} = \hat{\gamma}_{Cru} + \hat{\gamma}_{SP} + \hat{\gamma}_{Local} = 4.75 + 6 - 5.5 = 5.25$.

Assim, obtém-se as estimativas

$\hat{\lambda}_{SP} = \frac{\hat{E}[X-Y]+2\hat{E}[X+Y]-\{\hat{E}[(X+Y)^2]-(\hat{E}[X+Y])^2\}}{2} = \frac{1.111+2x2.25-(5.25-2.25^2)}{2} = 2.71175$;

$\hat{\lambda}_{Cru} = \frac{2\hat{E}[X+Y]-\hat{E}[X-Y]-\{\hat{E}[(X+Y)^2]-(\hat{E}[X+Y])^2\}}{2} = \frac{2x2.25-1.111-(5.25-2.25^2)}{2} = 1.60075$;

$\hat{\lambda}_{12} = \frac{\{\hat{E}[(X+Y)^2]-(\hat{E}[X+Y])^2\}-\hat{E}[X+Y]}{2} = \frac{5.25-2.25^2-2.25}{2} = -1.03125$.

Então, o "placar de mínimos quadrados" para o jogo final seria (2.71175x1.60075).

A partir dos valores de $\overset{\wedge}{\lambda}_{SP}$ e $\overset{\wedge}{\lambda}_{Cru}$ pode-se calcular a probabilidade da vitória do São Paulo, fazendo a soma de todas as probabilidades $P(x, y)$ quando $x > y$. Analogamente, para a vitória do Cruzeiro ($x < y$) e empate ($x = y$). Fazendo isso, chegamos às seguintes probabilidades:

P[vitória do São Paulo] = 0.6119;

P[empate] = 0.1750;

e P[vitória do Cruzeiro] = 0.2131.

3.3 Método Chance I

Lee (1997) em seu artigo publicado na revista *Chance,* o que deu o nome ao método, propôs um método para estimar a média de gols marcados pelo time da casa e visitante, em que a média de gols de um time reflete sua "força", a qualidade do seu adversário e a vantagem de jogar em casa.

O método Chance I (Arruda, 2000) admite independência entre os placares X e Y , ou seja, o número de gols marcados pelo time da casa não interfere a distribuição de gols marcados pelo time visitante. Então, pelas propriedades da distribuição Poisson Bivariada "de Holgate", $\lambda_{12} = 0$. Neste método $E[X] = \lambda_1$ e $E[Y] = \lambda_2$ são estimados através do modelo log-linear de Poisson definido da seguinte maneira.

Se definirmos X_{2i-1} e X_{2i} como sendo os números de gols marcados pelas equipes Mandante e Visitante, respectivamente, no i-ésimo jogo (i=1,2,3,...n). Então suas distribuições são dadas por

$$\begin{cases} f(x_{2i-1}) = \frac{e^{-\lambda_{2i-1}}\lambda_{2i-1}^{x_{2i-1}}}{x_{2i-1}!} \\ f(x_{2i}) = \frac{e^{-\lambda_{2i}}\lambda_{2i}^{x_{2i}}}{x_{2i}!}, \end{cases} \tag{3.8}$$

para X_{2i-1} e X_{2i} são independentes.

Os logaritmos de suas funções de verossimilhança podem ser escritos da seguinte forma

$$\begin{cases} l(\lambda_{2i-1}, X_{2i-1}) = -\lambda_{2i-1} + x_{2i-1}\log\lambda_{2i-1} - \log(x_{2i-1})! \\ l(\lambda_{2i}, X_{2i}) = -\lambda_{2i} + x_{2i}\log\lambda_{2i} - \log(x_{2i})! \end{cases} . \tag{3.9}$$

O modelo log-linear de Poisson relaciona a distribuição da variável dependente (gols marcados) às variáveis explicativas através da função (canônica) de ligação dada por

$$\lambda_j = e^{U_j B}, \tag{3.10}$$

para $j = 1, 2, ..., 2n$ onde U_1, U_2, ...,U_{2n} denotam vetores de covariáveis e β é um vetor de parâmetros.

Das expressões (3.9) e (3.10), pode-se então escrever as funções de log-verossimilhança da seguinte forma

$$\begin{cases} l(\lambda_{2i-1}, X_{2i-1}) = -e^{U_{2i-1}\beta} + x_{2i-1}U_{2i-1}\beta - \log(x_{2i-1}!) \\ l(\lambda_{2i}, X_{2i}) = -e^{U_{2i}\beta} + x_{2i}U_{2i}\beta - \log(x_{2i}!) \end{cases} . \tag{3.11}$$

Considerando a realização de n jogos, o modelo pode ser definido por

$$l(\lambda_1, \lambda_2, ..., \lambda_{2n-1}, \lambda_{2n}, X_1, X_2, ..., X_{2n-1}, X_{2n}) = \sum_{j=1}^{2n}(-e^{U_j\beta} + x_j U_j\beta - \log(x_j!). \tag{3.12}$$

Neste método, x_{2i-1} e x_{2i} são os números de gols marcados pelas equipes Mandante e Visitante (respectivamente) no i-ésimo jogo; o vetor β é composto por $2N+3$ parâmetros, sendo um parâmetro o intercepto, dois parâmetros (um relativo ao ataque e outro à defesa) associado a cada uma das N equipes constantes no banco de dados, um parâmetro associado ao tipo de local onde o jogo se realiza e, mais um parâmetro associado à crise. As matrizes-linha U_{2i-1} e U_{2i} possuem $2N + 3$ componentes, sendo a primeira constante e igual a 1 (associada ao intercepto), as N componentes seguintes associadas ao *status* de cada equipe em relação ao ataque, as N subsequentes associadas ao *status* de cada equipe em relação à defesa, a próxima componente que indica o tipo de local em que o jogo se realiza e a última que indica se o time está sofrendo crise. As N componentes da matriz-linha U_{2i-1} relativas ao *status* das equipes em relação ao ataque, assumirão os valores 1, se a equipe correspondente for a Mandante do i-ésimo jogo, ou 0 caso contrário. Analogamente, as N componentes relativas ao *status* das equipes em relação à defesa, assumirão os valores -1, se a equipe correspondente for a Visitante do i-ésimo jogo, ou 0 caso contrário. Da mesma forma, as N componentes da matriz-linha U_{2i} relativas ao *status*

das equipes em relação ao ataque, assumirão os valores 1, se a equipe correspondente for a Visitante do i-ésimo jogo, ou 0 em caso contrário e as N componentes relativas ao *status* das equipes em relação à defesa, assumirão os valores -1, se a equipe correspondente for a Mandante do i-ésimo jogo, ou 0 em caso contrário.

A variável correspondente da matriz-linha U_{2i-1} relativa ao local de realização do jogo é uma variável indicadora que assume os valores 1 se a equipe Mandante do i-ésimo jogo tiver jogando em sua casa, ou 0 em caso contrário. Analogamente, a variável correspondente da matriz-linha U_{2i} relativa ao local de realização do jogo é uma variável inidicadora que assume os valores 1 se a equipe Visitante do i-ésimo jogo tiver jogando em sua casa, ou 0 em caso contrário.

A variável correspondente da matriz-linha U_{2i-1} relativa à crise é uma variável indicadora que assume os valores 1 se a equipe Mandante do i-ésimo jogo tiver jogando contra um adversário que está sofrendo crise, ou 0 em caso contrário. Da mesma forma, a variável correspondente da matriz-linha U_{2i} relativa à crise é uma variável indicadora que assume os valores 1 se a equipe Visitante do i-ésimo jogo tiver jogando contra uma equipe que está sofrendo crise, ou 0 em caso contrário.

Para o exemplo proposto de aplicação, no terceiro jogo do torneio, X_5 e X_6 são respectivamente iguais a 2 e 1 e, o vetor β dado (em ordem alfabética de times e com os parâmetros de ataque colocados antes dos de defesa) por

$$\left[\begin{array}{cccccccccccc} \beta_0 & \beta_{AtCru} & \beta_{AtFla} & \beta_{AtGre} & \beta_{AtSP} & \beta_{DefCru} & \beta_{DefFla} & \beta_{DefGre} & \beta_{DefSP} & \beta_{Local} & \beta_{Crise} \end{array} \right]^T,$$

a matriz-linha U_5 torna-se igual a $\left[\begin{array}{cccccccccccc} 1 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{array} \right]$ e a matriz-linha U_6 torna-se igual a $\left[\begin{array}{cccccccccccc} 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{array} \right]$.

Ao considerar conjuntamente todos os jogos do torneio, U torna-se uma matriz de constantes (relativas ao intercepto), *status*, local e crise (com duas linhas referente a cada jogo, uma coluna referente ao ataque e outra à defesa de cada equipe, uma coluna referente aos locais onde os jogos são realizados e a última coluna referente à crise).

Assim, para os jogos do torneio (exemplo), a matriz U é dada por

$$U = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix} \quad \text{e} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_{AtCru} \\ \beta_{AtFla} \\ \beta_{AtGre} \\ \beta_{AtSP} \\ \beta_{DefCru} \\ \beta_{DefFla} \\ \beta_{DefGre} \\ \beta_{DefSP} \\ \beta_{Local} \\ \beta_{Crise} \end{bmatrix}.$$

Os estimadores de $E[X_{2n+1}] = \lambda_{2n+1}$ e $E[X_{2n+2}] = \lambda_{2n+2}$ podem, então, ser calculados a partir de $\hat{E}[X_{2n+1}] = e^{U_{2n+1}\hat{\beta}}$ e $\hat{E}[X_{2n+2}] = e^{U_{2n+2}\hat{\beta}}$, onde $\hat{\beta}$ é obtido através da estimação por regressão de Poisson, que é a obtenção do vetor $\hat{\beta}$ que maximize a função de log-verossimilhança (3.12). Trata-se de uma estimação de máxima verossimilhança onde não existe uma forma analítica fechada para a expressão de $\hat{\beta}$, é então realizada através de processos numéricos interativos que podem levar a estimatidas $\hat{\beta}$ diferentes, mas que sempre levarão aos mesmos valores das estimativas $\hat{E}[X_j]$. Podemos observar essa característica, por exemplo, considerando sem perda de generalidade o jogo São Paulo e Cruzeiro do torneio hipotético. Sendo a modelagem do número de gols de cada time dada por $\hat{E}[X_{13}] = e^{U_{13}\hat{\beta}} = e^{\hat{B}_0+\hat{\beta}_{AtSP}-\hat{\beta}_{DefCru}+\hat{\beta}_{Local}}$e $\hat{E}[X_{14}] = e^{U_{14}\hat{\beta}} = e^{\hat{B}_0+\hat{\beta}_{AtCru}-\hat{\beta}_{DefSP}}$, podemos notar que as expressões acima podem ser reescritas da seguinte forma (onde K_1 e K_2 são constantes quaisquer):

$\hat{E}[X_{13}] = e^{U_{13}\hat{\beta}} = e^{\hat{B}_0+\hat{\beta}_{AtSP}-\hat{\beta}_{DefCru}+\hat{\beta}_{Local}} = e^{(\hat{B}_0+k_1+k_2)+(\hat{\beta}_{AtSP}-k_1)-(\hat{\beta}_{DefCru}+k_2)+\hat{\beta}_{Local}} = \exp\{\hat{\beta}_0' + \hat{\beta}_{AtSP}' - \hat{\beta}_{DefCru}' + \hat{\beta}_{Local}'\} = \exp\{U_{13}\hat{\beta}'\}$ e $\hat{E}[X_{14}] = e^{U_{14}\hat{\beta}} = e^{\hat{B}_0+\hat{\beta}_{AtCru}-\hat{\beta}_{DefSP}} = e^{(\hat{B}_0+k_1+k_2)+(\hat{\beta}_{AtCru}-k_1)-(\hat{\beta}_{DefSP}+k_2)} = \exp\{\hat{\beta}_0' + \hat{\beta}_{AtCru}' - \hat{\beta}_{DefSP}'\} = \exp\{U_{14}\hat{\beta}'\}$.

Entretanto, pode-se notar que embora o vetor $\hat{\beta}$ não seja único, as expressões acima mostram que as estimativas finais $\hat{E}[X_{13}]$ e $\hat{E}[X_{14}]$ são sempre únicas, devido às constantes K_1 e K_2 que se "cancelam".

Utilizando o *software* R e através do comando **glm** obtém-se a estimativa $\hat{\beta}$ dada por

$$\hat{\beta} = [\ -0.161 \quad -1.143 \quad -0.056 \quad -2.087 \quad 0 \quad -0.958 \quad -2.029 \quad -1.135 \quad 0 \quad -1.112 \quad 0\ ]. \tag{3.13}$$

Para a partida final entre São Paulo e Cruzeiro, obtemos
$\hat{E}[X_{13}] = e^{U_{13}\hat{\beta}} = e^{\hat{B}_0+\hat{\beta}_{AtSP}-\hat{\beta}_{DefCru}+\hat{\beta}_{Local}} = e^{-0.161+0-(-0.958)-1.112} = e^{-0.315} = 0.7298$
e $\hat{E}[X_{14}] = e^{U_{14}\hat{\beta}} = e^{\hat{B}_0+\hat{\beta}_{AtCru}-\hat{\beta}_{DefSP}} = e^{-0.161-1.143-0} = e^{-1.304} = 0.2715.$

Assim, obtém-se as estimativas $\hat{\lambda}_{SP} = 0.7298$ e $\hat{\lambda}_{Cru} = 0.2715$. A partir dos valores de $\hat{\lambda}_{SP}$ e $\hat{\lambda}_{Cru}$ pode-se calcular as probabilidades:

P[vitória do São Paulo] = 0.4303;

P[empate] = 0.4439;

e P[vitória do Cruzeiro] = 0.1258.

3.4 Método Chance II

O método Chance II (Arruda, 2000) tem a mesma estrutura do método Chance I, mas inclui a estimação da covariância entre X e Y. Pelas propriedades da distribuição de Poisson "de Holgate" as esperanças marginais dos números de gols marcados pelos dois adversários são escritas da forma: $E[X] = \lambda_1 + \lambda_{12}$ e $E[Y] = \lambda_2 + \lambda_{12}$, sendo possível decompô-las em uma parcela comum λ_{12} e uma parcela (λ_1 ou λ_2) que se relaciona somente à distribuição marginal (de X ou de Y, respectivamente).

O resultado de uma "linearização" do modelo (3.10) é dado por

$$X_i = \beta_0 + U_i'\beta' + \varepsilon_{bi}, \text{ ou equivalentemente, } E[X_i] = \beta_0 + U_i'\beta'. \tag{3.14}$$

sendo que ε_{bi} são erros independentes com médias iguais a 0 e U' é a mesma matriz U e β' é o mesmo vetor β definidos no método Chance I, ambos alterados apenas com a exclusão da coluna de U e da componente de β relativas ao intercepto.

Pela construção da distribuição "de Holgate" por processos de Poisson é possível estabelecer analogias entre a parcela comum e o processo comum e entre as parcelas específicas e os processos específicos das variáveis X e Y. Dessa forma, estabelecendo um paralelo entre as notações para $E[X] = \lambda_1 + \lambda_{12}$ e o modelo (3.14) temos

$$E[X] = \underbrace{\lambda_{12}}_{\text{parcela comum}} + \underbrace{\lambda_1}_{\text{parcela específica}} = \underbrace{\beta_0}_{\text{parcela comum}} + \underbrace{U_i'\beta'}_{\text{parcela específica}}. \tag{3.15}$$

Então, concluindo-se esse raciocínio baseado em analogias e paralelismo, sugere-se a correspondência $\lambda_{12} = \beta_0$.

Logo, pode-se obter a expressão dos valores dos parâmetros de interesse λ_1, λ_2 e λ_{12} do sistema de equações dado por

$$\begin{cases} E[X] = \lambda_1 + \lambda_{12} \\ E[Y] = \lambda_2 + \lambda_{12} \\ \beta_0 = \lambda_{12} \end{cases}. \tag{3.16}$$

O que sugere os seguintes estimadores indiretos

$$\begin{cases} \hat{\lambda}_1 = \hat{E}[X] - \hat{\beta}_0 \\ \hat{\lambda}_2 = \hat{E}[Y] - \hat{\beta}_0 \\ \hat{\lambda}_{12} = \hat{\beta}_0 \end{cases}. \tag{3.17}$$

Por sua vez, $E[X]$, $E[Y]$ (e β_0) são estimados através do modelo linear definido por

$$X_j = U_j\beta + \varepsilon_{bj}, \tag{3.18}$$

para $j = 1, 2, ..., 2n$, onde ε_{bj} são erros Normais independentes com médias iguais a 0.

No modelo linear (3.18), X_j, β e a matriz-linha U_j são definidas similarmente como no método Chance I. Como X_{2i-1} e X_{2i} são os números de gols marcados pelas equipes Mandante e Visitante no i-ésimo jogo, então considerando conjuntamente todos os jogos do nosso torneio fictício, apresentado no início deste capítulo,o vetor X do modelo (3.18) passa a ser o vetor de gols marcados dado por

$$X = [\ 0 \quad 1 \quad 2 \quad 1 \quad 2 \quad 1 \quad 0 \quad 2 \quad 1 \quad 2 \quad 0 \quad 3\ ]^t.$$

Assim, os estimadores de λ_1, λ_2 e λ_{12} construídos em (3.17) podem então ser calculados a partir de $\hat{E}[X] = U\hat{\beta}$ e $\lambda_{12} = \hat{\beta}_0$.

Da teoria de mínimos quadrados e de matrizes inversas generalizadas de Moore-

Penrose, tem-se que o estimador $\hat{\beta}$ é dado por

$\hat{\beta} = (U'U)^{-}U'X =$

$=[\ 0.9153 \quad -0.0728 \quad 0.479 \quad -0.594 \quad 1.104 \quad -0.146 \quad -0.8015 \quad -0.291 \quad 0.323 \quad -0.587 \quad 0.291\ ]^{t}.$

Para o jogo final entre São Paulo x Cruzeiro, os vetores U_{13} e U_{14} são dados por $U_{13} =$ $[\ 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 1\ ]$ e $U_{14} = [\ 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad -1 \quad 0\ ]$, de onde obtêm-se os estimadores $\hat{E}[X_{13}] = U_{13}\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_{AtSP} - \hat{\beta}_{DefCru} + \hat{\beta}_{Local}$ e $\hat{E}[X_{14}] = U_{14}\hat{\beta}$ $= \hat{\beta}_0 + \hat{\beta}_{AtCru} - \hat{\beta}_{DefSP}$.

Logo, substituindo o valor estimado $\hat{\beta}$ obtemos $\hat{E}[X_{13}] = U_{13}\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_{AtSP} -$ $\hat{\beta}_{DefCru} + \hat{\beta}_{Local} = 0.9153 + 1.104 - (-0.146) - 0.587 = 1.5783$, $\hat{E}[X_{14}] = U_{14}\hat{\beta} =$ $\hat{\beta}_0 + \hat{\beta}_{AtCru} - \hat{\beta}_{DefSP} = 0.9153 - 0.0728 - 0.323 = 0.5195$ e $\hat{\beta}_0 = 0.9153$.

Portanto, substituindo no sistema de equações (3.17) temos

$\hat{\lambda}_{SP} = \hat{E}[X_{13}] - \hat{\beta}_0 = 1.5783 - 0.9153 = 0.663$;

$\hat{\lambda}_{Cru} = \hat{E}[X_{14}] - \hat{\beta}_0 = 0.5195 - 0.9153 = -0.3958$;

$\hat{\lambda}_{12} = \hat{\beta}_0 = 0.9153$.

Arruda (2000) propôs que quando o valor estimado der negativo (como neste caso) utilizar "caminha" ou projeção do ponto estimado ao ponto mais próximo pertencente ao conjunto de estimativas válidas. Algebricamente, isso equivale a igualar a estimativa negativa a zero. Para aplicações práticas, entretanto, pode não ser conveniente igualar a estimativa negativa a zero. Pois, se $\lambda_2 = 0$ então é equivalente a fazer $P_2 \equiv 0$ (pela construção da distribuição "de Holgate") e, consequentemente, $X = P_1 + P_{12}$ e $Y = P_{12}$. Como P_1 e P_{12} são não negativos, X será sempre maior ou igual que Y, o que significa que P(derrota de X)=P(X<Y)=0. No caso do futebol, por maior que seja a diferença entre dois times, nunca se poderá atribuir probabilidade zero à derrota de um deles. Dessa forma, Arruda (2000) sugere uma "caminhada"alternativa que, em vez de igualar a estimativa negativa a zero, iguale-a a um valor ε pré-estabelecido (por exemplo, $\varepsilon = 0.25$).

Com isso, através da adoção de $\hat{\lambda}_{Cru} = 0.25$, obtemos as seguintes probabilidades:

P[vitória do São Paulo] = 0.4060;

P[empate] = 0.4706;

e P[vitória do Cruzeiro] = 0.1234.

3.5 Considerações Finais

Neste capítulo utilizando a distribuição Poisson Bivariada "de Holgate" para modelagem de partidas de futebol, consideramos quatro métodos para estimação dos parâmetros desta distribuição. Nos métodos SD0, SD1 e Chance II pode ocorrer o problema em encontrarmos estimativas negativas de λ_1, λ_2 e λ_{12}. Para contornar este problema usamos "caminhada", igualando o valor negativo estimado à 0.25. Dessa forma, uma possível questão que possa ser levantada é se este valor é adequado?

Os métodos apresentados neste capítulo não incorporam informações subjetivas a respeito de ataque, defesa e campo. No entanto, se considerarmos a metodologia Bayesiana para o problema, pode-se facilmente incorporar tais informações.

No próximo capítulo, considerando a metodologia Bayesiana, apresentamos três métodos para estimação dos parâmetros.

Capítulo 4

Abordagem Bayesiana para Estimação de Parâmetros para Previsão de Resultados de Jogos de Futebol

Neste capítulo, considerando a metodologia Bayesiana, propomos três métodos para calcular a probabilidade de vitória, empate ou derrota de uma determinada partida de futebol. Nos dois primeiros métodos assumimos independência entre X e Y, em que o método 1 consiste em encontrar o valor esperado para λ_1 e λ_2, conseqüentemente calculando essas probabilidades e, o Método 2 através do cálculo das densidades preditivas. No Método 3 assumimos dependência entre X e Y e calculamos as probabilidades de interesse de suas médias da posteriori utilizando uma priori conjunta pertencente à família da distribuição gama multivariada, a qual acomoda correlação entre os parâme- tros. A distribuição a posteriori foi obtida de duas formas: a primeira utilizando uma priori conjunta, sendo o produto de densidades gamas independentes e a segunda uma mistura de densidades gamas condicionalmente independentes.

Para exemplificar os métodos calculamos o possível resultado da partida entre Corinthians e Internacional que foi realizada no dia 20 de novembro de 2005.

Aplicamos os Métodos 1 e 2 no Campeonato Brasileiro de 2005 da mesma forma descrita no Capítulo 3. A aplicação do Método 3 em uma competição não é trivial.

Encontramos algumas dificuldade computacionais para calcular a densidade posteriori $p((\lambda_1, \lambda_2, \lambda_{12}) \mid (x, y))$ devido a sua forma recursiva. Por esse motivo, não o aplicamos no Campeonato Brasileiro de 2005 e 2006.

4.1 Assumindo Independência

4.1.1 Método 1: Estimador de Bayes com respeito à perda quadrática

O objetivo deste método é encontrar o valor esperado do número de gols marcados pelos times mandante e visitante na próxima rodada e assim, calcularmos as probabilidades de vitória, derrota e empate. Para isso, temos o seguinte teorema (Ehlers, 2005):

Teorema 4.1 Seja X uma variável aleatória com distribuição de Poisson com média λ_1, ou seja, $X \mid \lambda_1 \backsim Poisson(\lambda_1)$. Considerando a priori $Gama(\alpha_1, \beta_1)$ para λ_1 então a densidade a posteriori é da forma $(\lambda_1 \mid X_n = x_n) \backsim Gama(\alpha_1 + x_n, \beta_1 + 1)$, onde x_n é o número de gols marcados pelo time mandante na n-ésima rodada.

Prova: Pelo Teorema de Bayes, temos que

$$\Pi(\lambda_1 \mid X_n = x_n) = \frac{f(X_n = x_n \mid \lambda_1)\Pi(\lambda_1)}{\int_0^\infty f(X_n = x_n \mid \lambda_1)\Pi(\lambda_1)d\lambda_1}. \tag{4.1}$$

Por hipótese $X \mid \lambda_1 \backsim Poisson(\lambda_1)$ e $\lambda_1 \backsim Gama(\alpha_1, \beta_1)$. Dessa forma, substituindo as densidades em (4.1) obtemos

$$\Pi(\lambda_1 \mid X_n = x_n) = \frac{\frac{e^{-\lambda_1}\lambda_1^{x_n}\beta_1^{\alpha_1}\lambda_1^{\alpha_1-1}e^{-\beta_1\lambda_1}}{x_n!\Gamma(\alpha_1)}}{\int_0^\infty \frac{e^{-\lambda_1}\lambda_1^{x_n}\beta_1^{\alpha_1}\lambda_1^{\alpha_1-1}e^{-\beta_1\lambda_1}}{x_n!\Gamma(\alpha_1)}d\lambda_1} = \frac{e^{-\lambda_1}\lambda_1^{x_n}\lambda_1^{\alpha_1-1}e^{-\beta_1\lambda_1}}{\int_0^\infty e^{-\lambda_1}\lambda_1^{x_n}\lambda_1^{\alpha_1-1}e^{-\beta_1\lambda_1}d\lambda_1}. \tag{4.2}$$

Como o denominador da expressão (4.2) pode ser reescrito como

$$\int_0^\infty e^{-\lambda_1}\lambda_1^{x_n}\lambda_1^{\alpha_1-1}e^{-\beta_1\lambda_1}d\lambda_1 = \underbrace{\int_0^\infty \frac{(\beta_1+1)^{x_n+\alpha_1}\lambda_1^{x_n+\alpha_1-1}e^{-(\beta_1+1)\lambda_1}}{\Gamma(x_n+\alpha_1)}d\lambda_1}_{=1\ (Gama(x_n+\alpha_1,\beta_1+1))}.\frac{\Gamma(x_n+\alpha_1)}{(\beta_1+1)^{x_n+\alpha_1}} = \frac{\Gamma(x_n+\alpha_1)}{(\beta_1+1)^{x_n+\alpha_1}}. \tag{4.3}$$

De (4.2) e (4.3) obtemos

$$\Pi(\lambda_1 \mid X_n = x_n) = \frac{(\beta_1+1)^{x_n+\alpha_1}\lambda_1^{x_n+\alpha_1-1}e^{-(\beta_1+1)\lambda_1}}{\Gamma(x_n+\alpha_1)}.$$

Portanto, temos que

$$(\lambda_1 \mid X_n = x_n) \backsim Gama(\alpha_1+x_n, \beta_1+1), \tag{4.4}$$

demonstrando assim o teorema. ■

Dessa forma, se considerarmos a priori $Gama(\alpha_2, \beta_2)$ para o λ_2 e, sendo y_n é o número de gols marcados pelo time visitante na n-ésima rodada temos, pelo Teorema 4.1, que

$$(\lambda_2 \mid Y_n = y_n) \backsim Gama(\alpha_2+y_n, \beta_2+1). \tag{4.5}$$

Assim, se $(\lambda_1 \mid X_n = x_n) \backsim Gama(\alpha_1+x_n, \beta_1+1)$ então o valor esperado $E[\lambda_1 \mid X_n = x_n] = \frac{\alpha_1+x_n}{\beta_1+1}$.

Note que $E[\lambda_1 \mid X_n = x_n] = \frac{\alpha_1+x_n}{\beta_1+1} = \frac{\alpha_1}{\beta_1+1}+\frac{x_n}{\beta_1+1} = \frac{\alpha_1}{\beta_1}.\frac{\beta_1}{\beta_1+1}+x_n.\frac{1}{\beta_1+1} = \frac{\alpha_1}{\beta_1}.a+x_n.(1-a)$ onde $a = \frac{\beta_1}{\beta_1+1}$, isto é, o valor esperado é uma combinação convexa entre a média da priori e a média dos resultados (neste caso temos apenas o resultado da última partida).

Da mesma forma encontramos $E[\lambda_2 \mid Y_n = y_n] = \frac{\alpha_2+y_n}{\beta_2+1} = \frac{\alpha_2}{\beta_2+1} + \frac{y_n}{\beta_2+1} = \frac{\alpha_2}{\beta_2}.\frac{\beta_2}{\beta_2+1} + y_n.\frac{1}{\beta_2+1} = \frac{\alpha_2}{\beta_2}.b + y_n.(1-b)$ sendo $b = \frac{\beta_2}{\beta_2+1}$.

Logo, para encontrar esta média precisamos dos valores de α_i e $\beta_i\,, i = 1, 2$. Mas, quais valores atribuiremos?

Para responder esta pergunta podemos escolher estes valores de duas maneiras:

a) Como temos os dados das $(n-1)$-ésimas rodadas já realizadas podemos então fazer a média da priori igual ao estimador de máxima verossimilhança da Poisson, isto é: $\frac{\alpha_1}{\beta_1} = \frac{\sum_{i=1}^{n-1} x_i}{n-1} = \overline{x}_{n-1}$ e, como foi mostrado acima que a média da posteriori é uma combinação

convexa entre a média da priori e o último resultado, podemos fazer $a = \frac{\beta_1}{\beta_1+1} = k$ (*constante arbitrária*), isto é, colocando assim um peso k para a média da priori. Assim, resolvendo o sistema de equações

$$\begin{cases} \frac{\alpha_1}{\beta_1} = \frac{\sum_{i=1}^{n-1} x_i}{n-1} \\ \frac{\beta_1}{\beta_1+1} = k \end{cases},$$

obtemos os valores de α_1 e β_1 .

Analogamente, encontramos os valores de α_2 e β_2 através da solução do sistema

$$\begin{cases} \frac{\alpha_2}{\beta_2} = \frac{\sum_{i=1}^{n-1} y_i}{n-1} \\ \frac{\beta_2}{\beta_2+1} = k \end{cases}.$$

Atribuimos para k os valores: 0.10, 0.25, 0.50, 0.75 e 0.90.

b) Optamos por uma forma quase análoga ao que foi vista em (*a*), mas com a diferença de que igualamos a média da priori à média de gols marcados quando o time for mandante e visitante respectivamente, ou seja,

$$\frac{\alpha_1}{\beta_1} = \frac{\sum \text{número de gols marcados em casa}}{\text{número de jogos realizados em casa}}$$

e

$$\frac{\alpha_2}{\beta_2} = \frac{\sum \text{número de gols marcados fora de casa}}{\text{número de jogos realizados fora de casa}}.$$

A partir do valor esperado dos parâmetros da distribuição pode-se calcular a probabilidade de qualquer resultado específicio (por exemplo, P(1x3)=P(X=1,Y=3)). Consequentemente, pode-se calcular a probabilidade dos eventos relativos ao jogo: P(vitória do mandante) $= \sum_{i>j} P(X = i, Y = j)$; P(empate)=$\sum_{i} P(X = i, Y = i)$ e P(vitória do visitante) $= \sum_{j>i} P(X = i, Y = j)$.

Como exemplo de aplicação, vamos prever o possível resultado da partida entre Corinthians e Internacional que foi realizada no dia 20 de novembro de 2005. Para isso temos os resultados das 32 primeiras rodadas do Campeonato Brasileiro 2005. As Tabelas 4.1 e 4.2 mostram, respectivamente, os resultados obtidos considerando o Método 1 caso (a) e (b).

Tabela 4.1: Método 1 (a)

k	Pr(V)[1]	Pr(E)[2]	Pr(D)[3]
0.10	0.0736	0.3415	0.5849
0.25	0.1669	0.3095	0.5236
0.50	0.2909	0.2675	0.4416
0.75	0.3887	0.2338	0.3775
0.90	0.4383	0.2162	0.3455

Tabela 4.2: Método 1 (b)

k	Pr(V)[1]	Pr(E)[2]	Pr(D)[3]
0.10	0.0798	0.3442	0.5760
0.25	0.1827	0.3131	0.5042
0.50	0.3218	0.2686	0.4096
0.75	0.4319	0.2307	0.3374
0.90	0.4876	0.2106	0.3018

Observa-se que, para esse exemplo de aplicação, as probabilidades de vitória aumentam com o valor de k e mudam de um método (a e b) para outro. As probabilidades de empate permanecem aproximadamente iguais nos dois métodos.

4.1.2 Método 2: Densidade Preditiva

Neste método, calculamos a densidade preditiva de X e Y com as mesmas suposições consideradas no Método 1, isto é, que $\lambda_1 \backsim Gama(\alpha_1, \beta_1)$ e $\lambda_2 \backsim Gama(\alpha_2, \beta_2)$. Para isso, enunciamos os seguintes teoremas:

Teorema 4.2 Sejam X e Y variáveis aleatórias representando o número de gols marcados pelos times mandante e visitante, respectivamente. Supondo X e Y com distribuição de Poisson com média λ_1 e λ_2, respectivamente e, considerando uma priori $Gama(\alpha_1, \beta_1)$ para λ_1 e uma priori $Gama(\alpha_2, \beta_2)$ para λ_2, a probabilidade de ocorrer empate é dada

[1] Pr(V) representa a probabilidade de vitória do time mandante.
[2] Pr(E) representa a probabilidade de empate.
[3] Pr(D) representa a probabilidade de derrota do time mandante.

por

$$P(\text{empate}|\ X_n{=}\ x_n, Y_n{=}\ y_n) = \sum_{k=0}^{\infty} [\frac{(\beta_1+1)^{x_n+\alpha_1}(\beta_2+1)^{y_n+\alpha_2}\Gamma(k+x_n+\alpha_1)\Gamma(k+y_n+\alpha_2)}{k!^2\Gamma(x_n+\alpha_1)\Gamma(y_n+\alpha_2)(\beta_1+2)^{k+x_n+\alpha_1}(\beta_2+2)^{k+y_n+\alpha_2}},$$

onde x_n e y_n representam, respectivamente, o número de gols marcados pelos times mandante e visitante na n-ésima rodada.

Prova: Por definição temos que

$$P(\text{empate} \mid X_n = x_n, Y_n = y_n) = \sum_{k=0}^{\infty} P(X_{n+1} = k, Y_{n+1} = k \mid X_n = x_n, Y_n = y_n). \quad (4.6)$$

Por hipótese X e Y são independentes, dessa forma (4.6) pode ser expressa na forma

$$\begin{aligned} &P(\text{empate}|\ X_n = x_n, Y_n = y_n) = \sum_{k=0}^{\infty} P(X_{n+1} = k \mid X_n = x_n, Y_n = y_n)\times \\ &\times P(Y_{n+1}{=}\ k \mid X_n{=}\ x_n, Y_n{=}\ y_n) = \sum_{k=0}^{\infty} P(X_{n+1}{=}\ k \mid X_n{=}\ x_n).P(Y_{n+1}{=}\ k \mid Y_n{=}\ y_n). \end{aligned} \quad (4.7)$$

Podemos reescrever (4.7) da seguinte maneira

$$\begin{aligned} &P(\text{empate}|\ X_n = x_n, Y_n = y_n) \\ &= \sum_{k=0}^{\infty}[(\int_0^{\infty} P(X_{n+1} = k, \lambda_1 \mid X_n = x_n)d\lambda_1)\times \\ &\quad \times(\int_0^{\infty} P(Y_{n+1} = k, \lambda_2 \mid Y_n = y_n)d\lambda_2)] \\ &= \sum_{k=0}^{\infty}[(\int_0^{\infty} P(X_{n+1} = k \mid \lambda_1, X_n = x_n).P(\lambda_1 \mid X_n = x_n)d\lambda_1)\times \\ &\quad \times(\int_0^{\infty} P(Y_{n+1} = k \mid \lambda_2, Y_n = y_n).P(\lambda_2 \mid Y_n = y_n)d\lambda_2)] \\ &= \sum_{k=0}^{\infty}[(\int_0^{\infty} P(X_{n+1} = k \mid \lambda_1).P(\lambda_1 \mid X_n = x_n)d\lambda_1)\times \\ &\quad \times(\int_0^{\infty} P(Y_{n+1} = k \mid \lambda_2).P(\lambda_2 \mid Y_n = y_n)d\lambda_2)]. \end{aligned} \quad (4.8)$$

Como, por hipótese, $X \backsim Poisson(\lambda_1)$, $Y \backsim Poisson(\lambda_2)$, $\lambda_1 \backsim Gama(\alpha_1, \beta_1)$ e $\lambda_2 \backsim Gama(\alpha_2, \beta_2)$. Segue de (4.8) que

$$
\begin{aligned}
&P(\text{empate}|\ X_n = x_n, Y_n = y_n)\\
&= \sum_{k=0}^{\infty}[(\int_0^{\infty} \frac{e^{-\lambda_1}\lambda_1^k}{k!}.\frac{(\beta_1+1)^{x_n+\alpha_1}\lambda_1^{x_n+\alpha_1-1}e^{-(\beta_1+1)\lambda_1}}{\Gamma(x_n+\alpha_1)}d\lambda_1)\times\\
&\times(\int_0^{\infty} \frac{e^{-\lambda_2}\lambda_2^k}{k!}.\frac{(\beta_2+1)^{y_n+\alpha_2}\lambda_2^{y_n+\alpha_2-1}e^{-(\beta_2+1)\lambda_2}}{\Gamma(y_n+\alpha_2)}d\lambda_2)]\\
&= \sum_{k=0}^{\infty}[\frac{(\beta_1+1)^{x_n+\alpha_1}(\beta_2+1)^{y_n+\alpha_2}}{k!^2\Gamma(x_n+\alpha_1)\Gamma(y_n+\alpha_2)}\times\\
&\times(\int_0^{\infty} e^{-(\beta_1+2)\lambda_1}\lambda_1^{k+x_n+\alpha_1-1}d\lambda_1).(\int_0^{\infty} e^{-(\beta_2+2)\lambda_2}\lambda_2^{k+y_n+\alpha_2-1}d\lambda_2)]\\
&= \sum_{k=0}^{\infty}[\frac{(\beta_1+1)^{x_n+\alpha_1}(\beta_2+1)^{y_n+\alpha_2}\Gamma(k+x_n+\alpha_1)\Gamma(k+y_n+\alpha_2)}{k!^2\Gamma(x_n+\alpha_1)\Gamma(y_n+\alpha_2)(\beta_1+2)^{k+x_n+\alpha_1}(\beta_2+2)^{k+y_n+\alpha_2}}\times\\
&\times(\int_0^{\infty} \frac{(\beta_1+2)^{k+x_n+\alpha_1}}{\Gamma(k+x_n+\alpha_1)}e^{-(\beta_1+2)\lambda_1}\lambda_1^{k+x_n+\alpha_1-1}d\lambda_1)(\int_0^{\infty} \frac{(\beta_2+2)^{k+y_n+\alpha_2}}{\Gamma(k+y_n+\alpha_2)}e^{-(\beta_2+2)\lambda_2}\lambda_2^{k+y_n+\alpha_2-1}d\lambda_2)]\\
&= \sum_{k=0}^{\infty}[\frac{(\beta_1+1)^{x_n+\alpha_1}(\beta_2+1)^{y_n+\alpha_2}\Gamma(k+x_n+\alpha_1)\Gamma(k+y_n+\alpha_2)}{k!^2\Gamma(x_n+\alpha_1)\Gamma(y_n+\alpha_2)(\beta_1+2)^{k+x_n+\alpha_1}(\beta_2+2)^{k+y_n+\alpha_2}}.\blacksquare
\end{aligned}
$$

Teorema 4.3 Sejam X e Y variáveis aleatórias representando o número de gols marcados pelos times Mandante e Visitante, respectivamente. Supondo X e Y com distribuição de Poisson com média λ_1 e λ_2, respectivamente e, considerando uma priori $Gama(\alpha_1, \beta_1)$ para λ_1 e uma priori $Gama(\alpha_2, \beta_2)$ para λ_2, a probabilidade de ocorrer vitória do time Mandante é dada por

$$
\begin{aligned}
&P(\text{Vitória do time mandante}|\ X_n = x_n, Y_n = y_n)\\
&= \sum_{i>j}^{\infty}[\frac{(\beta_1+1)^{x_n+\alpha_1}(\beta_2+1)^{y_n+\alpha_2}\Gamma(i+x_n+\alpha_1)\Gamma(j+y_n+\alpha_2)}{i!j!\Gamma(x_n+\alpha_1)\Gamma(y_n+\alpha_2)(\beta_1+2)^{i+x_n+\alpha_1}(\beta_2+2)^{j+y_n+\alpha_2}}.
\end{aligned}
\quad ,
$$

onde x_n e y_n representam, respectivamente, o número de gols marcados pelo time mandante e visitante na n-ésima rodada.

Prova: Partindo da definição

$$
\begin{aligned}
&P(\text{Vitória do time mandante}|\ X_n = x_n, Y_n = y_n) = P(X_{n+1} > Y_{n+1} \mid X_n = x_n, Y_n = y_n)\\
&= \sum_{i>j} P(X_{n+1} = i, Y_{n+1} = j \mid X_n = x_n, Y_n = y_n),
\end{aligned}
$$

a prova deste Teorema é análoga a do **Teorema 4.2**.

Teorema 4.4 Sejam X e Y variáveis aleatórias representando o número de gols marcados pelos times Mandante e Visitante, respectivamente. Supondo X e Y com distribuição de Poisson com média λ_1 e λ_2, respectivamente e, considerando a priori $Gama(\alpha_1, \beta_1)$

para λ_1 e $Gama(\alpha_2, \beta_2)$ para λ_2, a probabilidade de ocorrer vitória do time Visitante é dada por

$$P(\text{Vitória do time visitante}|\ X_n = x_n, Y_n = y_n) = \sum_{i<j}^{\infty} [\frac{(\beta_1+1)^{x_n+\alpha_1}(\beta_2+1)^{y_n+\alpha_2}\Gamma(i+x_n+\alpha_1)\Gamma(j+y_n+\alpha_2)}{i!j!\Gamma(x_n+\alpha_1)\Gamma(y_n+\alpha_2)(\beta_1+2)^{i+x_n+\alpha_1}(\beta_2+2)^{j+y_n+\alpha_2}}.\ ,$$

onde x_n e y_n representam, respectivamente, o número de gols marcados pelo time mandante e visitante na n-ésima rodada.

Prova: Da mesma forma do teorema anterior, partindo da definição

$$P(\text{Vitória do time visitante}|\ X_n = x_n, Y_n = y_n) = P(X_{n+1} < Y_{n+1} \mid X_n = x_n, Y_n = y_n)$$
$$= \sum_{i<j} P(X_{n+1} = i, Y_{n+1} = j \mid X_n = x_n, Y_n = y_n),$$

a demonstração deste Teorema é análoga a do **Teorema 4.2**.

Assim, utilizando os valores de α_1, β_1, α_2 e β_2 obtidos no Método 1, tanto em (a) quanto no caso (b), calculamos as probabilidades.

Dessa forma, no exemplo de aplicação, para prever o resultado da partida entre Corinthians e Internacional ocorrida em 20 de novembro de 2005, os resultados obtidos pelo Método 2 caso (a) e (b) são mostrados, respectivamente, pelas Tabelas 4.3 e 4.4.

Tabela 4.3: Método 2 (a)

a	Pr(V)[4]	Pr(E)[5]	Pr(D)[6]
0.10	0.4915	0.4276	0.0809
0.25	0.4735	0.3461	0.1804
0.50	0.4329	0.2636	0.3035
0.75	0.3832	0.2232	0.3936
0.90	0.3499	0.2107	0.4394

[4]Pr(V) representa a probabilidade de vitória do time mandante.
[5]Pr(E) representa a probabilidade de empate.
[6]Pr(D) representa a probabilidade de derrota do time mandante.

Tabela 4.4: Método 2 (b)

a	Pr(V)[7]	Pr(E)[8]	Pr(D)[9]
0.10	0.4851	0.4279	0.0870
0.25	0.4585	0.3466	0.1949
0.50	0.4060	0.2632	0.3308
0.75	0.3465	0.2204	0.4331
0.90	0.3078	0.2054	0.4868

4.2 Método 3: Assumindo Dependência

Assumindo a distribuição de Poisson Bivariada "de Holgate" para X e Y, a densidade conjunta pode ser escrita da seguinte forma (ver (2.1))

$$P(X=x,Y=y)=e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^x}{x!}\frac{\lambda_2^y}{y!}\sum_{k=0}^{\min(x,y)}\binom{x}{k}\binom{y}{k}k!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k,$$

onde $\lambda_3=\lambda_{12}$.

Suponhamos os dados $\mathbf{X}_i$ onde cada $\mathbf{X}_i=(x_i,y_i)$ é o par de observações que representam, respectivamente, os gols marcados pelos times mandante e visitante na i-ésima partida $(i=1,...,n)$, a verossimilhança é dada por

$L_n(\boldsymbol{\lambda},\mathbf{X})=\prod_{i=1}^{n}e^{-(\lambda_1+\lambda_2+\lambda_3)}\frac{\lambda_1^{x_i}}{x_i!}\frac{\lambda_2^{y_i}}{y_i!}\sum_{k=0}^{\min(x_i,y_i)}\binom{x_i}{k}\binom{y_i}{k}k!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k$, onde $\boldsymbol{\lambda}=(\lambda_1,\lambda_2,\lambda_3)$. Assim, reescrevemos a verossimilhança através do seguinte Lema encontrado em Karlis e Tsiamyrtiz (2004).

Lema: Defina $v_r^{(n)}=\frac{1}{(x_n-r)!(y_n-r)!r!}$. Dada uma amostra aleatória de tamanho n a verossimilhança pode ser escrita da seguinte forma

$$L_n(\boldsymbol{\lambda},\mathbf{X})=\exp(-n(\lambda_1+\lambda_2+\lambda_3))\lambda_1^{\sum_{i=1}^{n}x_i}\lambda_2^{\sum_{i=1}^{n}y_i}\sum_{k=0}^{S_n}w_k^{(n)}\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k, \tag{4.9}$$

sendo $S_n=\sum_{i=1}^{n}\min(x_i,y_i)$ e $w_k^{(n)}$ são coeficientes que podem ser obtidos recursivamente

[7] Pr(V) representa a probabilidade de vitória do time mandante.
[8] Pr(E) representa a probabilidade de empate.
[9] Pr(D) representa a probabilidade de derrota do time mandante.

através da seguinte expressão

$$w_k^{(n)} = \sum_{r=\max\{0,k-s_n^*\}}^{\min\{k,s_n^*\}} v_r^{(n)} w_{k-r}^{(n-1)}, \tag{4.10}$$

sendo $s_i = \min\{x_i, y_i\}$, $S_k = \sum_{i=1}^{k} s_i$, $s_n^* = \min\{s_n, S_{n-1}\}$ e $w_k^{(1)} = v_k^{(1)}$.

A típica suposição para as distribuições a priori é que estas são densidades gamas independentes, isto é, assumir que

$$\pi(\lambda_1, \lambda_2, \lambda_3) = \frac{\beta_1^{\alpha_1}\beta_2^{\alpha_2}\beta_3^{\alpha_3}}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} \lambda_1^{\alpha_1-1}\lambda_2^{\alpha_2-1}\lambda_3^{\alpha_3-1} \exp\{-\beta_1\lambda_1 - \beta_2\lambda_2 - \beta_3\lambda_3\}. \tag{4.11}$$

Assumimos uma priori conjunta que acomoda correlação entre os parâmetros. Nossa priori conjunta é pertencente à família da distribuição gama multivariada e trata-se de uma mistura de densidades gamas condicionalmente independentes da forma (Karlis e Tsiamyrtiz (2004))

$$\begin{aligned} &\pi(\lambda_1, \lambda_2, \lambda_3) = \sum_{k=0}^{r} p_k G(\alpha_1 - k, \beta_1) G(\alpha_2 - k, \beta_2) G(\alpha_3 + k, \beta_3) \\ &= \sum_{k=0}^{r} w_k (\lambda_1^{\alpha_1-k-1} \exp\{-\lambda_1\beta_1\})(\lambda_2^{\alpha_2-k-1} \exp\{-\lambda_2\beta_2\})(\lambda_3^{\alpha_3+k-1} \exp\{-\lambda_3\beta_3\}). \end{aligned} \tag{4.12}$$

Assumindo que $(X, Y) \mid (\lambda_1, \lambda_2, \lambda_3)$ é uma distribuição de Poisson Bivariada e se fizermos $s = \min\{x, y\}$ temos que a verossimilhança é da forma

$$\begin{aligned} &f((x, y) \mid (\lambda_1, \lambda_2, \lambda_3) = \exp\{-(\lambda_1 + \lambda_2 + \lambda_3)\} \sum_{i=0}^{s} \frac{\lambda_1^{x-i}\lambda_2^{y-i}\lambda_3^i}{(x-i)!(y-i)!i!} \\ &= \sum_{i=0}^{s} v_i (\lambda_1^{x-i} \exp\{-\lambda_1\})(\lambda_2^{y-i} \exp\{-\lambda_2\})(\lambda_3^i \exp\{-\lambda_3\}) \text{ onde } v_i = \frac{1}{(x-i)!(y-i)!i!}. \end{aligned} \tag{4.13}$$

E, a distribuição a priori conjunta para $(\lambda_1, \lambda_2, \lambda_3)$ é dada por

$$\begin{aligned} &\pi(\lambda_1, \lambda_2, \lambda_3) = \sum_{k=0}^{r} p_k G(\alpha_1 - k, \beta_1) G(\alpha_2 - k, \beta_2) G(\alpha_3 + k, \beta_3) \\ &= \sum_{k=0}^{r} w_k (\lambda_1^{\alpha_1-k-1} \exp\{-\lambda_1\beta_1\})(\lambda_2^{\alpha_2-k-1} \exp\{-\lambda_2\beta_2\})(\lambda_3^{\alpha_3+k-1} \exp\{-\lambda_3\beta_3\}), \end{aligned} \tag{4.14}$$

onde $\alpha_1 > r$, $\alpha_2 > r$, $\alpha_3 > 0$, $\beta_i > 0$ ($i = 1, 2, 3$), $\sum_{j=0}^{r} p_j = 1$ e

$$w_k = p_k \frac{(\beta_1)^{\alpha_1 - k}(\beta_2)^{\alpha_2 - k}(\beta_3)^{\alpha_3 + k}}{\Gamma(\alpha_1 - k)\Gamma(\alpha_2 - k)\Gamma(\alpha_3 + k)} \text{ para } k = 0, 1, ..., r. \tag{4.15}$$

Então, de (4.13), (4.14) e (4.15), a distribuição a posteriori é da forma

$$\begin{aligned}
&p((\lambda_1, \lambda_2, \lambda_3) \mid (x, y)) \propto [\sum_{i=0}^{s} v_i(\lambda_1^{x-i} \exp\{-\lambda_1\})(\lambda_2^{y-i} \exp\{-\lambda_2\})\times \\
&\times(\lambda_3^i \exp\{-\lambda_3\})] \times [\sum_{k=0}^{r} w_k(\lambda_1^{\alpha_1-k-1} \exp\{-\lambda_1\beta_1\})(\lambda_2^{\alpha_2-k-1} \exp\{-\lambda_2\beta_2\})(\lambda_3^{\alpha_3+k-1} \exp\{-\lambda_3\beta_3\})] \\
&=\lambda_1^{\alpha_1+x-1}\lambda_2^{\alpha_2+y-1}\lambda_3^{\alpha_3-1} \exp\{-\lambda_1(\beta_1+1) - \lambda_2(\beta_2+1) - \lambda_3(\beta_3+1)\} \times [\sum_{i=0}^{s} v_i(\tfrac{\lambda_3}{\lambda_1\lambda_2})^i][\sum_{j=0}^{r} w_j(\tfrac{\lambda_3}{\lambda_1\lambda_2})^j] \\
&=\sum_{k=0}^{s+r} [\sum_{l=\max\{0,k-s^*\}}^{\min\{k,s^*\}} v_l w_{k-l}](\lambda_1^{\alpha_1+x-k-1}e^{-\lambda_1(\beta_1+1)})(\lambda_2^{\alpha_2+y-k-1}e^{-\lambda_2(\beta_2+1)})(\lambda_3^{\alpha_3+k-1}e^{-\lambda_3(\beta_3+1)}).
\end{aligned} \tag{4.16}$$

Agora, para $k = 0, 1, ..., s + r$ considerarmos

$$\rho_k^* = [\sum_{l=\max\{0,k-s^*\}}^{\min\{k,s^*\}} v_l w_{k-l}]\times\Gamma(\alpha_1+x-k)\Gamma(\alpha_2+y-k)\Gamma(\alpha_3+k)\left(\frac{(\beta_1+1)(\beta_2+1)}{\beta_3+1}\right)^k \tag{4.17}$$

e

$$\rho_k = \frac{\rho_k^*}{\sum_{l=0}^{s+r} \rho_l^*} \text{ para } k = 0, 1, ..., s + r. \tag{4.18}$$

Substituindo (4.17) e (4.18) em (4.16), a distribuição a posteriori terá a seguinte forma (Karlis e Tsiamyrtiz (2004))

$$p((\lambda_1, \lambda_2, \lambda_3) \mid (x, y)) = \sum_{k=0}^{s+r} \rho_k G(\alpha_1 + x - k, \beta_1 + 1)G(\alpha_2 + y - k, \beta_2 + 1)G(\alpha_3 + k, \beta_3 + 1). \tag{4.19}$$

Note que a posteriori é novamente uma mistura de densidades gamas condicionalmente independentes.

No exemplo de aplicação, vamos utilizar no conjunto de dados o resultado (confronto direto entre Internacional e Corinthians) da partida realizada em 10 de agosto de 2005,

em que ocorreu um empate por 0x0 e, os demais resultados para estimar os parâmetros $\alpha_i's$ e $\beta_i's$ $(i = 1, 2, 3)$ através de um sistema de equações para encontrar, primeiramente, um valor médio para $\lambda_i's \geq 0$ obtidos da seguinte forma: igualamos a média de gols marcados pelos times mandantes e visitante à média amostral dos gols quando estes foram, respectivamente, mandante e visitante e, igualamos a correlação a uma constante c. Daí, igualamos a média das prioris a esses valores e as variâncias iguais a uma constante arbitrária d.

Dessa forma, pelas propriedades da distribuição de Poisson Bivariada "de Holgate" temos:

$$\begin{cases} E[X] = \lambda_1 + \lambda_3 = 2.1875 \\ E[Y] = \lambda_2 + \lambda_3 = 1.6667 \\ \rho(x,y) = \dfrac{\lambda_3}{\sqrt{\lambda_1+\lambda_3}\sqrt{\lambda_2+\lambda_3}} = c \end{cases} \Longrightarrow \begin{cases} \lambda_1 = 2.1875 - c\sqrt{2.1875}\sqrt{1.6667} \\ \lambda_2 = 1.6667 - c\sqrt{2.1875}\sqrt{1.6667} \\ \lambda_3 = c\sqrt{2.1875}\sqrt{1.6667} \end{cases} \Longrightarrow$$

$$\Longrightarrow \begin{cases} \lambda_1 = 2.1875 - 1.909426c \\ \lambda_2 = 1.6667 - 1.909426c \\ \lambda_3 = 1.909426c \end{cases}.$$

Note que temos as seguintes restrições: $c\sqrt{\lambda_1+\lambda_3}\sqrt{\lambda_2+\lambda_3} < \lambda_1+\lambda_3$ e $c\sqrt{\lambda_1+\lambda_3}\sqrt{\lambda_2+\lambda_3} < \lambda_2+\lambda_3$ pois $\lambda_1 > 0$ e $\lambda_2 > 0$. E, como a correlação $c = \rho(x,y) \geq 0$ pois $\lambda_i's \geq 0$, atribuimos os seguintes valores para a constante $c = 0.001$, 0.1, 0.2, $0.3, ...,$ 0.8, 0.872. Para valores maiores que 0.872 temos $\lambda_2 < 0$ e, se $c = 0$ então X e Y são independentes.

Dessa maneira, encontramos $\alpha_i's$ e $\beta_i's$ $(i = 1, 2, 3)$ através do seguinte sistema de equações

$$\begin{cases} \frac{\alpha_1}{\beta_1} = 2.1875 - 1.909426c \\ \frac{\alpha_2}{\beta_2} = 1.6667 - 1.909426c \\ \frac{\alpha_3}{\beta_3} = 1.909426c \\ \frac{\alpha_1}{\beta_1^2} = \frac{\alpha_2}{\beta_2^2} = \frac{\alpha_3}{\beta_3^2} = d \end{cases} \Longrightarrow \begin{cases} \alpha_1 = \frac{(2.1875-1.909426c)^2}{d} \\ \beta_1 = \frac{(2.1875-1.909426c)}{d} \\ \alpha_2 = \frac{(1.6667-1.909426c)^2}{d} \\ \beta_2 = \frac{(1.6667-1.909426c)}{d} \\ \alpha_3 = \frac{(1.909426c)^2}{d} \\ \beta_3 = \frac{(1.909426c)}{d} \end{cases}.$$

O valor da constante d é arbitrária e varia de acordo com a incerteza. Atribuimos para d os valores: 0.1, 1, 3, 5, 10.

Simulamos 10000 amostras das posterioris de λ_1, λ_2 e λ_3 e através de suas médias calculamos as probabilidades de vitória, empate e derrota do time mandante. As densidades posterioris foram obtidas de duas maneiras:

1°) Cálculo através de uma priori conjunta sendo densidades gamas independentes, ou seja, $r = 0$. Assim, a posteriori é dada por

$$p((\lambda_1, \lambda_2, \lambda_3) \mid (0,0)) = G(\alpha_1, \beta_1 + 1)G(\alpha_2, \beta_2 + 1)G(\alpha_3, \beta_3 + 1).$$

As Tabelas 4.5 a 4.14 mostram, respectivamente, os resultados obtidos ao assumir a correlação $c = 0.001$, 0.1, 0.2, 0.3, ..., 0.8, 0.872.

Tabela 4.5: Resultados assumindo a correlação c igual a 0.001

c	d	λ_1	λ_2	λ_3	Pr(V)[10]	Pr(E)[11]	Pr(D)[12]
0.001	0.1	2.0925	1.5676	$4.5e^{-07}$	0.4992	0.2105	0.2903
0.001	1	1.5083	1.0344	$3.0e^{-11}$	0.4818	0.2584	0.2598
0.001	3	0.9160	0.5830	$2.9e^{-10}$	0.4173	0.3595	0.2232
0.001	5	0.6686	0.4175	$4.7e^{-86}$	0.3628	0.4385	0.1987
0.001	10	0.3916	0.2434	$1.2e^{-254}$	0.2655	0.5816	0.1529

Tabela 4.6: Resultados assumindo a correlação c igual a 0.1

c	d	λ_1	λ_2	λ_3	Pr(V)[10]	Pr(E)[11]	Pr(D)[12]
0.1	0.1	1.9025	1.3881	0.1271	0.4967	0.2144	0.2889
0.1	1	1.3267	0.8809	0.0307	0.4713	0.2768	0.2519
0.1	3	0.7941	0.4810	0.0138	0.4002	0.3917	0.2081
0.1	5	0.5630	0.3376	0.0068	0.3360	0.4834	0.1806
0.1	10	0.3255	0.1834	0.0038	0.2395	0.6339	0.1266

[10] Pr(V) representa a probabilidade de vitória do time mandante.
[11] Pr(E) representa a probabilidade de empate.
[12] Pr(D) representa a probabilidade de derrota do time mandante.

Tabela 4.7: Resultados assumindo a correlação c igual a 0.2

c	d	λ_1	λ_2	λ_3	Pr(V)[13]	Pr(E)[14]	Pr(D)[15]
0.2	0.1	1.7083	1.1908	0.3056	0.4973	0.2154	0.2873
0.2	1	1.1564	0.7215	0.1031	0.4653	0.2913	0.2434
0.2	3	0.6728	0.3820	0.0444	0.3814	0.4238	0.1948
0.2	5	0.4920	0.2696	0.0253	0.3222	0.5142	0.1636
0.2	10	0.2710	0.1524	0.0142	0.2148	0.6670	0.1182

Tabela 4.8: Resultados assumindo a correlação c igual a 0.3

c	d	λ_1	λ_2	λ_3	Pr(V)[13]	Pr(E)[14]	Pr(D)[15]
0.3	0.1	1.5194	0.9961	0.4869	0.4985	0.2159	0.2856
0.3	1	0.9962	0.5670	0.2065	0.4615	0.3008	0.2377
0.3	3	0.5646	0.2943	0.0903	0.3653	0.4485	0.1862
0.3	5	0.3915	0.2001	0.0584	0.2951	0.5514	0.1535
0.3	10	0.2216	0.1085	0.0326	0.1983	0.6978	0.1039

Tabela 4.9: Resultados assumindo a correlação c igual a 0.4

c	d	λ_1	λ_2	λ_3	Pr(V)[13]	Pr(E)[14]	Pr(D)[15]
0.4	0.1	1.3316	0.8121	0.6746	0.4977	0.2159	0.2864
0.4	1	0.8299	0.4295	0.3310	0.4521	0.3071	0.2408
0.4	3	0.4727	0.2102	0.1591	0.3588	0.4583	0.1829
0.4	5	0.3100	0.1423	0.1018	0.2770	0.5724	0.1506
0.4	10	0.1759	0.0759	0.0549	0.1835	0.7177	0.0988

[13] Pr(V) representa a probabilidade de vitória do time mandante.
[14] Pr(E) representa a probabilidade de empate.
[15] Pr(D) representa a probabilidade de derrota do time mandante.

Tabela 4.10: Resultados assumindo a correlação c igual a 0.5

c	d	λ_1	λ_2	λ_3	Pr(V)[16]	Pr(E)[17]	Pr(D)[18]
0.5	0.1	1.1347	0.6247	0.8639	0.4956	0.2163	0.2881
0.5	1	0.6797	0.2935	0.4775	0.4483	0.3071	0.2446
0.5	3	0.3579	0.1354	0.2341	0.3403	0.4706	0.1891
0.5	5	0.2417	0.0908	0.1526	0.2671	0.5804	0.1525
0.5	10	0.1321	0.0435	0.0788	0.1703	0.7351	0.0946

Tabela 4.11: Resultados assumindo a correlação c igual a 0.6

c	d	λ_1	λ_2	λ_3	Pr(V)[16]	Pr(E)[17]	Pr(D)[18]
0.6	0.1	0.9514	0.4378	1.0511	0.4964	0.2161	0.2875
0.6	1	0.5303	0.1749	0.6106	0.4396	0.3085	0.2519
0.6	3	0.2741	0.0771	0.3195	0.3343	0.4642	0.2015
0.6	5	0.1785	0.0522	0.2212	0.2637	0.5673	0.1690
0.6	10	0.1003	0.0255	0.1135	0.1699	0.7237	0.1064

Tabela 4.12: Resultados assumindo a correlação c igual a 0.7

c	d	λ_1	λ_2	λ_3	Pr(V)[16]	Pr(E)[17]	Pr(D)[18]
0.7	0.1	0.7569	0.2527	1.2462	0.4945	0.2160	0.2895
0.7	1	0.3810	0.0779	0.7545	0.4263	0.3065	0.2672
0.7	3	0.1833	0.0339	0.4145	0.3236	0.4519	0.2245
0.7	5	0.1258	0.0212	0.2878	0.2638	0.5495	0.1867
0.7	10	0.0630	0.0092	0.1610	0.1724	0.7000	0.1276

[16] Pr(V) representa a probabilidade de vitória do time mandante.
[17] Pr(E) representa a probabilidade de empate.
[18] Pr(D) representa a probabilidade de derrota do time mandante.

Tabela 4.13: Resultados assumindo a correlação c igual a 0.8

c	d	λ_1	λ_2	λ_3	Pr(V)[19]	Pr(E)[20]	Pr(D)[21]
0.8	0.1	0.5744	0.0801	1.4297	0.4924	0.2158	0.2918
0.8	1	0.2680	0.0166	0.9208	0.4169	0.2941	0.2890
0.8	3	0.1140	0.0075	0.5068	0.3182	0.4324	0.2494
0.8	5	0.0760	0.0036	0.3675	0.2665	0.5186	0.2149
0.8	10	0.0431	0.0025	0.2030	0.1820	0.6691	0.1489

Tabela 4.14: Resultados assumindo a correlação c igual a 0.872

c	d	λ_1	λ_2	λ_3	Pr(V)[19]	Pr(E)[20]	Pr(D)[21]
0.872	0.1	0.4394	$3.2e-05$	1.5713	0.4811	0.2149	0.3040
0.872	1	0.1836	$3.7e-05$	1.0276	0.4024	0.2866	0.3110
0.872	3	0.0767	$3.8e-135$	0.5977	0.3213	0.4050	0.2737
0.872	5	0.0544	$4.9e-133$	0.4125	0.2696	0.4988	0.2316
0.872	10	0.0259	$2.9e-92$	0.2397	0.1892	0.6422	0.1686

Através dos resultados, esboçamos (ver apêndice A) os gráficos das prioris para o caso c=0.001 (Figuras A.1 a A.15) mostrando o comportamento da priori ao aumentar a variância d. Também, esboçamos os gráficos da correlação *versus* probabilidade (Figuras A.16 a A.30).

2°) Cálculo através da priori com os valores $r = 1$, p_0 e p_1. Dessa forma, a posteriori é dada por

$$p((\lambda_1, \lambda_2, \lambda_3) \mid (0,0)) = \sum_{k=0}^{1} \rho_k G(\alpha_1 - k, \beta_1 + 1) G(\alpha_2 - k, \beta_2 + 1) G(\alpha_3 + k, \beta_3 + 1),$$

onde $\rho_k = \frac{\rho_k^*}{\sum_{l=0}^{1} \rho_l^*}$ e $\rho_k^* = \frac{p_k \beta_1^{\alpha_1 - k} \beta_2^{\alpha_2 - k} \beta_3^{\alpha_3 + k}}{(\beta_1+1)^{\alpha_1 - k} (\beta_2+1)^{\alpha_2 - k} (\beta_3+1)^{\alpha_3 + k}}$, $k = 0, 1$. Agora, temos as seguintes

[19] Pr(V) representa a probabilidade de vitória do time mandante.
[20] Pr(E) representa a probabilidade de empate.
[21] Pr(D) representa a probabilidade de derrota do time mandante.

restrições: $\alpha_1 > r$, $\alpha_2 > r$, $\alpha_3 > 0$, $\beta_i > 0$ ($i = 1, 2, 3$), $\sum_{j=0}^{r} p_j = 1$.

i) Para $p_0 = 0.5$ e $p_1 = 0.5$, obtemos os seguintes resultados mostrados na Tabela 4.15.

Tabela 4.15: Resultados obtidos assumindo $p_0 = 0.5$ e $p_1 = 0.5$

c	d	λ_1	λ_2	λ_3	Pr(V)[22]	Pr(E)[23]	Pr(D)[24]
0.001	0.1	2.0915	1.5662	0.0201	0.4994	0.2093	0.2913
0.001	1	1.4998	1.0342	0.0045	0.4798	0.2587	0.2615
0.1	0.1	1.8818	1.3538	0.2693	0.5001	0.2070	0.2929
0.1	1	1.2275	0.7592	0.2724	0.4803	0.2592	0.2605
0.2	0.1	1.6865	1.1585	0.4023	0.4998	0.2107	0.2895
0.2	1	1.0052	0.5303	0.4254	0.4804	0.2678	0.2518
0.3	0.1	1.4918	0.9598	0.5601	0.5005	0.2130	0.2865
0.3	1	0.7916	0.3173	0.5390	0.4776	0.2815	0.2409
0.4	0.1	1.2944	0.7593	0.7353	0.5011	0.2144	0.2845
0.5	0.1	1.0989	0.5582	0.9149	0.5022	0.2155	0.2823
0.6	0.1	0.9052	0.3472	1.0935	0.5058	0.2165	0.2777
0.7	0.1	0.7022	0.1190	1.2849	0.5111	0.2174	0.2715

ii) Para $p_0 = 0.25$ e $p_1 = 0.75$, obtemos os seguintes resultados mostrados na Tabela 4.16.

[22] Pr(V) representa a probabilidade de vitória do time mandante.
[23] Pr(E) representa a probabilidade de empate.
[24] Pr(D) representa a probabilidade de derrota do time mandante.

Tabela 4.16: Resultados obtidos assumindo $p_0 = 0.25$ e $p_1 = 0.75$

c	d	λ_1	λ_2	λ_3	Pr(V)[25]	Pr(E)[26]	Pr(D)[27]
0.001	0.1	2.0839	1.5658	0.0575	0.4980	0.2075	0.2945
0.001	1	1.5047	1.0336	0.0129	0.4813	0.2572	0.2615
0.1	0.1	1.8713	1.3406	0.3639	0.5009	0.2023	0.2968
0.1	1	1.1419	0.6625	0.4868	0.4852	0.2457	0.2691
0.2	0.1	1.6712	1.1416	0.4582	0.5003	0.2083	0.2914
0.2	1	0.9076	0.4238	0.6116	0.4844	0.2573	0.2583
0.3	0.1	1.4700	0.9394	0.6003	0.5003	0.2118	0.2879
0.3	1	0.6978	0.2027	0.6998	0.4846	0.2726	0.2428
0.4	0.1	1.2815	0.7360	0.7628	0.5033	0.2135	0.2832
0.5	0.1	1.0790	0.5316	0.9385	0.5036	0.2152	0.2812
0.6	0.1	0.8833	0.3108	1.1181	0.5089	0.2160	0.2751
0.7	0.1	0.68065	0.0664	1.2998	0.5179	0.2178	0.2643

iii) Para $p_0 = 0.75$ e $p_1 = 0.25$, obtemos os seguintes resultados mostrados na Tabela 4.17.

[25] Pr(V) representa a probabilidade de vitória do time mandante.
[26] Pr(E) representa a probabilidade de empate.
[27] Pr(D) representa a probabilidade de derrota do time mandante.

Tabela 4.17: Resultados obtidos assumindo $p_0 = 0.75$ e $p_1 = 0.25$

c	d	λ_1	λ_2	λ_3	Pr(V)[28]	Pr(E)[29]	Pr(D)[30]
0.001	0.1	2.0869	1.5657	0.0066	0.4985	0.2109	0.2906
0.001	1	1.5007	1.0468	0.0015	0.4769	0.2586	0.2645
0.1	0.1	1.8930	1.3720	0.1944	0.4984	0.2107	0.2909
0.1	1	1.2882	0.8321	0.1285	0.4754	0.2691	0.2555
0.2	0.1	1.7018	1.1741	0.3511	0.4996	0.2129	0.2875
0.2	1	1.0910	0.6352	0.2520	0.4733	0.2792	0.2475
0.3	0.1	1.5062	0.9820	0.5204	0.4987	0.2145	0.2868
0.3	1	0.8964	0.4444	0.3833	0.4703	0.2886	0.2411
0.4	0.1	1.3175	0.7845	0.7055	0.5006	0.2148	0.2846
0.5	0.1	1.1194	0.5886	0.8902	0.5000	0.2157	0.2843
0.6	0.1	0.9283	0.3925	1.0778	0.5011	0.2159	0.2830
0.7	0.1	0.7303	0.1837	1.2657	0.5033	0.2167	0.2800

As Figuras A.31 a A.36 (ver apêndice A) mostram o gráfico da probabilidade *versus* correlação ao assumir variância 0.1 e 1 para os casos (*i*, *ii* e *iii*).

4.3 Considerações Finais

Neste Capítulo utilizando a distribuição Poisson Bivariada "de Holgate" para modelagem de partidas de futebol, propomos três métodos para estimação dos parâmetros desta distribuição.

Nos dois primeiros métodos assumimos indepêndencia entre X e Y. No caso (*a*) igualamos a média da priori igual ao estimador de máxima verossimilhanca da Poisson, isto é, a média dos gols marcados no campeonato, não levando em consideração se o time é mandante ou visitante. Já no caso (*b*) igualamos a média da priori a média de gols marcados em seu estádio se o time for mandante ou igualamos a média de gols marcados fora de seu estádio se o time for visitante. Uma outra alternativa que poderiamos ter

[28] Pr(V) representa a probabilidade de vitória do time mandante.
[29] Pr(E) representa a probabilidade de empate.
[30] Pr(D) representa a probabilidade de derrota do time mandante.

feito consiste em fazer a média da priori da mesma forma do que foi feita (nos casos (a) e (b)), mas ao invés de supor um valor para a, podemos escolher uma variância para a priori igual a uma constante c, ou seja, e assim encontrar os valores de α_1 e β_1 através da solução do sistema $\begin{cases} \frac{\alpha_1}{\beta_1} = \frac{\sum_{i=1}^{n-1} x_i}{n-1} \\ \frac{\alpha_1}{\beta_1^2} = c \end{cases}$. Da mesma forma fazemos para encontrar os valores de α_2 e β_2 . O valor da constante c é arbitrário e varia de acordo com a incerteza. E, através de painel contendo opniões de especialistas também poderiamos determinar os valores de α_i e β_i , $i = 1, 2$ através da solução do sistema $\begin{cases} \frac{\alpha_i}{\beta_i} = \frac{\sum_{j=1}^{n} x_j}{n} \\ \frac{\alpha_i}{\beta_i^2} = d \end{cases}$ para i =1,2; n: número de especialistas e d constante que varia de acordo com a incerteza.

Para a aplicação do Método 3 em um campeonato deparamos com dificuldades em calcular a densidade posteriori $p((\lambda_1, \lambda_2, \lambda_3) \mid (x, y))$ (ver 4.19) devido a forma recursiva com que esta é obtida. No exemplo de aplicação deste Método, observamos que com estas prioris (i, ii e iii) não podemos obter todas as possíveis variações da correlação e variância vistas no caso $r = 0$, pois para os valores não explicitados nas Tabelas 4.15, 4.16 e 4.17 obtemos resultados que não satisfazem as restrições vistas acima.

Observamos que, através de uma priori composta por densidades gamas independentes e assumindo uma priori informativa, como nos casos $d = 0.1$ e $d = 1$, independente de quão forte for a correlação de X e Y ocorrerá a vitória do time mandante (Corinthians). Se $d = 3$ o Corinthians irá vencer se $c < 0.2$, caso contrário, ocorrerá empate. Para $d = 5$ e $d = 10$ (priori não informativa) o jogo terminará empatado. E, assumindo uma priori composta de uma mistura de densidades gamas condicionalmente independentes com $r = 1$, devido as restrições no valor da variância $d = 0.1$ ou $d = 1$, o resultado da partida em questão será a vitória do time mandante.

Capítulo 5

Medida de DeFinetti e Bootstrap

5.1 Medida de DeFinetti

A medida de DeFinetti (DeFinetti, 1972) consiste na consideração de um simplex contido em $\mathbb{R}^3$ como representação geométrica do conjunto das possíveis previsões probabilísticas. Os vértices desse simplex correspondem às ocorrências dos resultados e os demais pontos a todas as outras possíveis previsões. Formalmente, $S = \{(P_{VIT}, P_{EMP}, P_{DER}) \in \mathbb{R}^3 \mid P_{VIT} + P_{EMP} + P_{DER} = 1,\ P_{VIT} \geq 0,\ P_{EMP} \geq 0,\ P_{DER} \geq 0\}$, onde P_{VIT} denota a probabilidade de vitória do time mandante, P_{EMP} a probabilidade de empate e P_{DER} a probabilidade de derrota do time visitante.

A medida da distância de DeFinetti corresponde à distância euclidiana quadrática entre o ponto correspondente à probabilidade prevista e o vértice correspondente ao resultado efetivamente observado. Para mais de uma previsão, pode-se construir um índice dado pela média aritmética das distâncias de DeFinetti chamado "Medida de DeFinetti".

Para o futebol, associam-se, respectivamente, os vértices (1,0,0), (0,1,0) e (0,0,1) à vitória da equipe mandante, ao empate e à derrota da equipe mandante. Ao vetor de probabilidades atribuídas para uma determinada partida associa-se o ponto $(PVIT, PEMP, PDER) \in S$.

Dessa forma, a distância de DeFinetti será igual a: $(PVIT - 1)^2 + (PEMP - 0)^2 + (PDER - 0)^2$ se ocorrer vitória do time mandante; $(PVIT - 0)^2 + (PEMP - 1)^2 + (PDER - 0)^2$ se ocorrer empate e $(PVIT - 0)^2 + (PEMP - 0)^2 + (PDER - 1)^2$ se ocorrer vitória do time visitante.

Para previsões futebolísticas um padrão comumente utilizado é a atribuição equiprovável de probabilidades ($PVIT = PEMP = PDER = \frac{1}{3}$) que corresponde a um previsor que atribui chances iguais a cada resultado em cada jogo. Para essa atribuição, temos que a medida de DeFinetti é igual a $(\frac{1}{3} - 1)^2 + (\frac{1}{3} - 0)^2 + (\frac{1}{3} - 0)^2 = 0.6667$.

Assim, podem ser considerados métodos de previsões de qualidade minimamente aceitável aqueles que apresentarem medidas de DeFinetti menores que 0.6667 e de má qualidade os que apresentarem medidas de DeFinetti superiores a 0.6667.

5.2 Bootstrap

A técnica Bootstrap foi introduzida por Efron (1979) para o cálculo de intervalos de confiança de parâmetros em situações onde outras técnicas não são aplicáveis, em particular no caso em que o tamanho amostral é reduzido. Posteriormente, seu uso foi generalizado para a resolução de problemas complexos através de técnicas de análise estatística tradicionais.

O método consiste em gerar um grande número de amostras utilizando a função de distribuição empírica dos dados originais. Estas amostras geradas podem ser então utilizadas, por exemplo, para a construção de intervalos de confiança com uma determinada probabilidade de cobertura.

O nome Bootstrap vem do fato de que usar os dados para gerar mais dados lembra o artifício usado pelo fictício Barão de Munchausen que conseguiu sair de um lago puxando ele mesmo pelos laços das suas botas (*bootstrap* em inglês).

Seja $Y_1, Y_2, ..., Y_n$ uma amostra aleatória da distribuição $p(y \mid \theta)$ onde θ é um parâmetro desconhecido. A função de distribuição empírica é definida como

$$\hat{F}_n(y) = \frac{\#\{Y_i \leq y\}}{n},$$

$\forall\ y \in \mathbb{R}$, onde $\#\ A$ significa o número de vezes que o evento A ocorre, $i = 1, ..., n$. Seja $T(\mathbf{Y})$ um estimador de θ.

O procedimento de reamostragem aqui consistirá na seleção com reposição de amostras $Y_1^*,\ Y_2^*, ..., Y_n^*$ da distribuição empírica $\hat{F}_n(y)$. Replicando este procedimento B vezes e calculando $\widehat{T^*}(y^*)$ a cada replicação teremos uma sequência de valores $\widehat{T_1^*},\ \widehat{T_2^*}, ...,\ \widehat{T_B^*}$. É

importante notar que cada Y_i^*, $i = 1, ..., n$, é igual a um dos valores observados e cada valor observado pode aparecer mais de uma vez, já que a amostragem é com reposição. A técnica bootstrap aqui utilizado é o não paramétrico.

5.3 Aplicação

Como aplicação da metodologia desenvolvida podemos, por exemplo, reamostrar os jogos considerando as r últimas rodadas do Campeonato Brasileiro 2006, $r = 1, ..., n$ e, utilizar o método SD0 para as previsões da $(n+1)^a$ rodada. Para isso, reamostramos 3000 jogos e realizamos este procedimento 1000 vezes para a construção do intervalo de confiança bootstrap da medida de Definetti. A cada reamostragem obtivemos uma medida de Definetti. Os resultados são apresentados nas Tabelas 6.1 e 6.2.

Tabela 6.1 Resultados obtidos para a série A na rodada 30

r	$I.C.(95\%)$[1]	def[2]
1	(0.6718;0.6729)	0.6721
3	(0.5274;0.5544)	0.5416
5	(0.5519;0.6115)	0.5778
10	(0.5711;0.6316)	0.6019
15	(0.5651;0.6215)	0.5938
20	(0.5552;0.6220)	0.5887
25	(0.5874;0.6455)	0.6124
29	(0.5611;0.6191)	0.5877

[1] Intervalo de Confiança Bootstrap de 95% para a medida de DeFinetti.
[2] def representa o valor médio obtido da medida de DeFinetti.

Tabela 6.2 Resultados obtidos para a série B na rodada 30

r	$I.C.(95\%)$[3]	def[4]
1	(0.3776;0.4592)	0.4160
3	(0.3691;0.4462)	0.4077
5	(0.3831;0.4520)	0.4185
10	(0.3887;0.4482)	0.4166
15	(0.3834;0.4548)	0.4177
20	(0.3833;0.4494)	0.4155
25	(0.3830;0.4475)	0.4130
29	(0.3733;0.4487)	0.4149

5.4 Considerações Finais

Nesta sessão fizemos uma breve introdução da Medida de DeFinetti e da metodologia Bootstrap. Como aplicação da metodologia desenvolvida, a cada rodada do Campeonato Brasileiro 2006 fizemos 3000 reamostragens dos jogos e através do Método SD0 verificamos sua precisão através da medida de DeFinetti. Exibimos como exemplo de aplicação apenas os resultados obtidos na rodada 30 tanto na série A quanto na série B do Campeonato Brasileiro 2006. Realizamos este procedimento 1000 vezes e construímos o intervalo de confiança bootstrap 95% da medida de DeFinetti.

Como resultado obtivemos uma melhor qualidade de precisão (menor medida de DeFinetti) ao reamostrarmos os jogos considerando apenas as 3 últimas rodadas tanto na série A quanto na série B.

[3] Intervalo de Confiança Bootstrap de 95% para a medida de DeFinetti.
[4] def representa o valor médio obtido da medida de DeFinetti.

Capítulo 6

Aplicações

Neste Capítulo aplicamos os procedimentos apresentados nos Capítulos 3 e 4 para previsão dos jogos do Campeonato Brasileiro 2005 e 2006.

6.1 Campeonato Brasileiro

O Campeonato Brasileiro de futebol é conhecido como Brasileirão. No Brasileirão 2006 o número de clubes participantes passou de 22 para 20 nas séries A e B. Essa quantidade de times torna a competição mais rentável, seletiva e emocionante tanto em cima como embaixo da tabela de classificação. Com a manutenção do regulamento da série A valorizará também a Série B, que será também disputado pelo sistema de pontos corridos, com quatro times subindo para a elite, e não dois como aconteceu nos anos anteriores. Desde o Brasileirão de 2003, o primeiro disputado por pontos corridos, observa-se o aumento das audiência televisiva, tanto na aberta quanto na paga e, do público nos estádios.

6.1.1 Metodologia Utilizada

Para o Campeonato Brasileiro 2005 série A utilizamos os oito métodos apresentados nos Capítulos 3 e 4. Para o Campeonato Brasileiro 2006 estamos utilizando os métodos SD 0 e Método 1 (b), assumindo $a = 0.9$, que apresentaram melhores resultados preditivos nas séries A e B.

Para prever os resultados utilizamos o banco de dados composto apenas por jogos do atual campeonato, ou seja, por exemplo, para a previsão da 26^a rodada utilizamos apenas os resultados dos jogos entre a 1^a e 25^a rodada. Considerando que um método "acerta" o resultado de um determinado jogo quando nos evidencia uma maior probabilidade de ocorrência, isto é, se em um determinado jogo o time mandante vencer e se a previsão para a vitória do time mandante é a maior, consideramos como acerto, caso contrário como erro. Da mesma forma para a ocorrência de empate e de vitória do time visitante.

6.2 Considerações Finais

Conferindo os placares dos jogos do Campeonato Brasileiro 2005, através dos métodos SD0, SD1, Chance I e Chance II obtivemos 54.54%, 52.12%, 34.34% e 30.81% de acertos, respectivamente. O Método 1 (a) e o Método 1 (b) obtiveram melhores resultados preditivo ao assumirem $a = 0.9$. Com essas condições, estes métodos acertaram 53.03% e 62.63%, respectivamente. Já o Método 2 (a) e o Método2 (b) obtiveram melhores resultados preditivo ao assumirem $a = 0.1$. Com essas condições, estes métodos acertaram 43.94% e 41.12%, respectivamente. Para o Campeonato Brasileiro 2006, utilizando o método SD 0 para a previsão das séries A e B obtivemos 56.84% e 54.37%, respectivamente. E, utilizando o Método 1 (b) assumindo $a = 0.9$ obtivemos 57.86% e 64.09%, respectivamente.

Verificamos, através da Medida de Definetti, a precisão dos dois Métodos que apresentaram a maior quantidade de "acerto": Métodos SD 0 e Método 1 (b) assumindo $a = 0.9$. Em ambos os Métodos, a cada rodada do Campeonato Brasileiro 2006 fizemos 3000 reamostragens dos jogos, realizamos este procedimento 1000 vezes e construímos o intervalo de confiança 95% da medida de DeFinetti. Como resultado, tanto no Método SD 0 quanto no Método 1 (b) assumindo $a = 0.9$, obtivemos uma melhor qualidade de precisão (menor medida de DeFinetti) ao reamostrarmos os jogos considerando apenas as 3 últimas rodadas tanto na série A quanto na série B.

Capítulo 7

Considerações Finais e Pesquisas Futuras

Nesta dissertação, utilizamos a distribuição de Poisson Bivariada "de Holgate" para a modelagem de resultados de jogos de futebol. Apresentamos alguns métodos para a estimação dos parâmetros desta distribuição. De posse destes métodos que forneçam as probabilidades de ocorrência de placares, aplicamos essas estimativas para calcular a probabilidade da ocorrência de um determinado resultado, a probabilidade de vitória do time mandante, de empate e de derrota.

Para uma melhor previsão dos resultados do Campeonato Brasileiro poderíamos ter utilizado um banco de dados composto por mais jogos recentes, por exemplo com os jogos dos campeonatos estaduais (Campeonato Paulista, Carioca, Mineiro, etc.).

Arruda (2000) aplicou os métodos em 390 jogos (64 da Copa do mundo de 1998, 297 do Camponato Brasileiro de 1998 e 29 do torneiro Rio-São Paulo de 1999). As menores medidas de DeFinetti obtidas foram 0.6203 e 0.6226 (métodos Chance II e Chance I, respectivamente). Neste trabalho aplicamos em 1191jogos (451 do Campeonato Brasileiro de 2005, 370 do Campeonato Brasileiro de 2006 série A e 370 do Campeonato Brasileiro de 2006 série B). As menores medidas de DeFinetti obtidas foram 0.5421 e 0.5027 (método SD 0 adicionado de uma covariável de incidência de crise e método 1 (b) assumindo a=0.9, respectivamente).

Obtemos uma melhor qualidade de precisão (menor medida de DeFinetti) do Método SD 0 e do Método 1 (b) assumindo $a = 0.9$, ao reamostrarmos os jogos considerando

apenas as 3 últimas rodadas tanto na série A quanto na série B do Campeonato Brasileiro 2006.

Uma linha de pesquisa futura é ampliar para o caso bivariado o modelo dinâmico apresentado em Knorr-Held (2000) e Junior e Gamerman (2004).

Apêndice A

Gráficos

Através dos resultados obtidos no Capítulo 4, esboçamos os seguintes gráficos:

i) Gráficos das prioris, por exemplo, para $c = 0.001$.

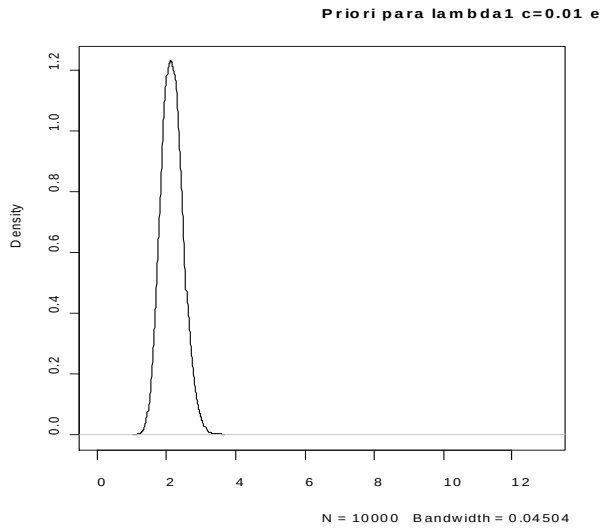

Figura A.1: Gráfico de λ_1 para c=0.01 e d=0.1

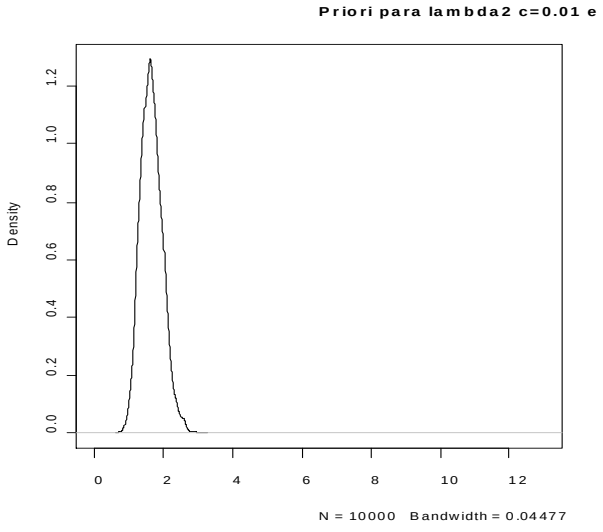

Figura A.2: Gráfico de λ_2 para c=0.01 e d=0.1

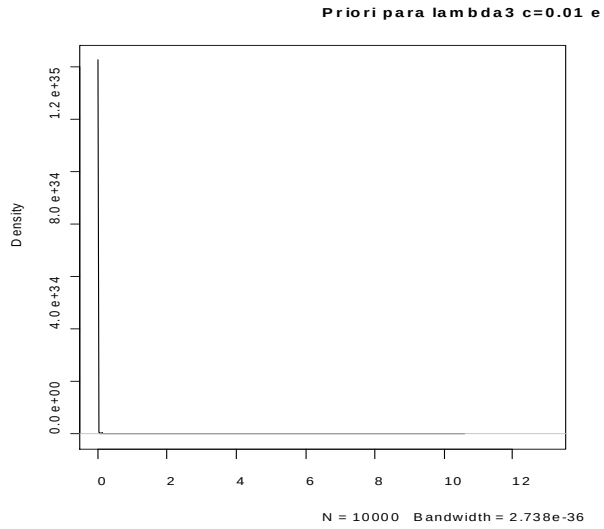

Figura A.3: Gráfico de λ_3 para c=0.01 e d=0.1

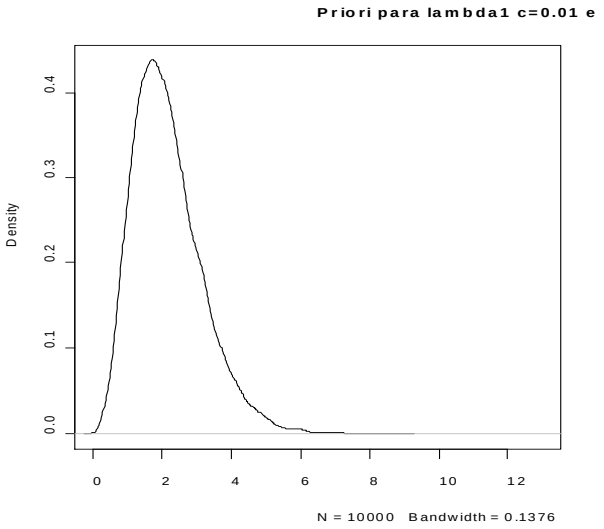

Figura A.4: Gráfico de λ_1 para c=0.01 e d=1

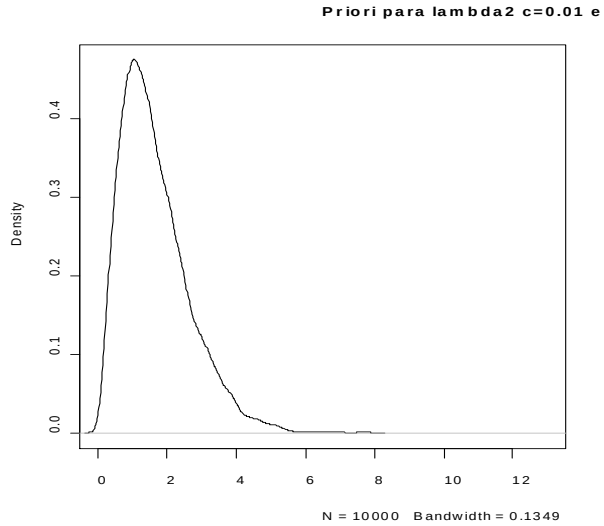

Figura A.5: Gráfico de λ_2 para c=0.01 e d=1

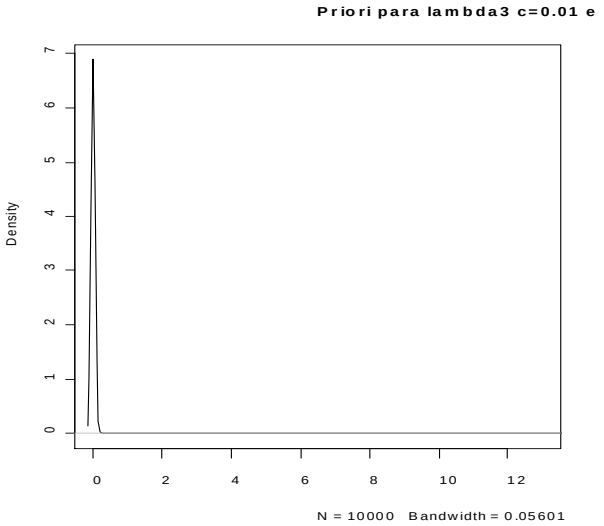

Figura A.6: Gráfico de λ_3 para c=0.01 e d=1

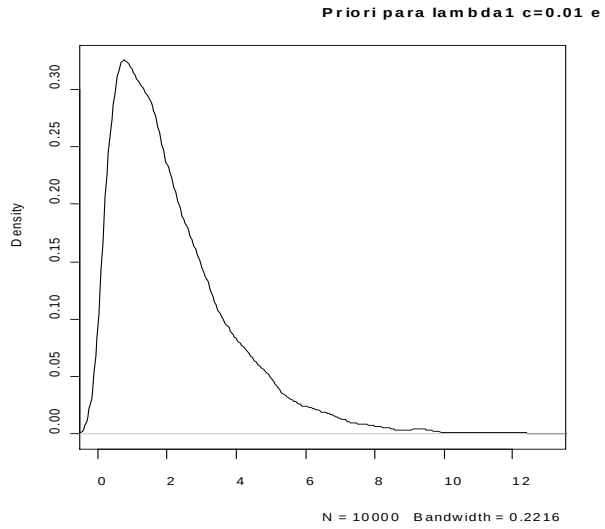

Figura A.7: Gráfico de λ_1 para c=0.01 e d=3

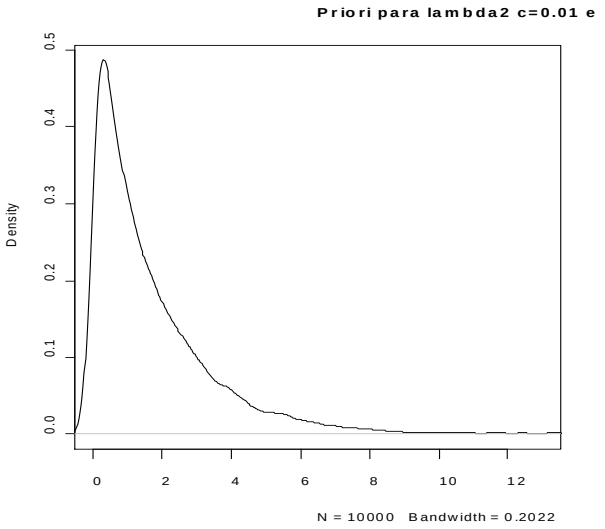

Figura A.8: Gráfico de λ_2 para c=0.01 e d=3

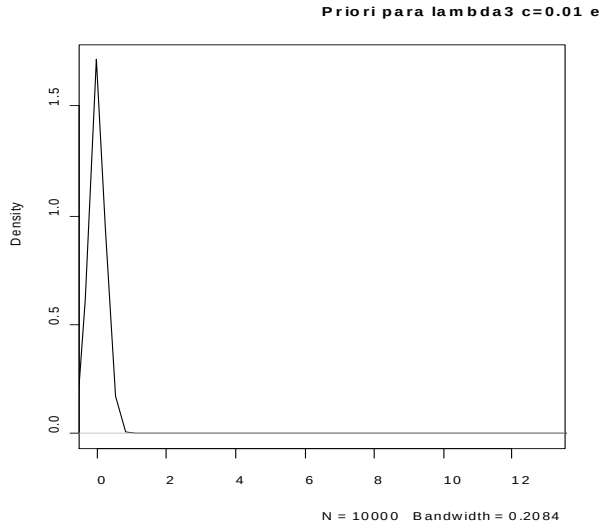

Figura A.9: Gráfico de λ_3 para c=0.01 e d=3

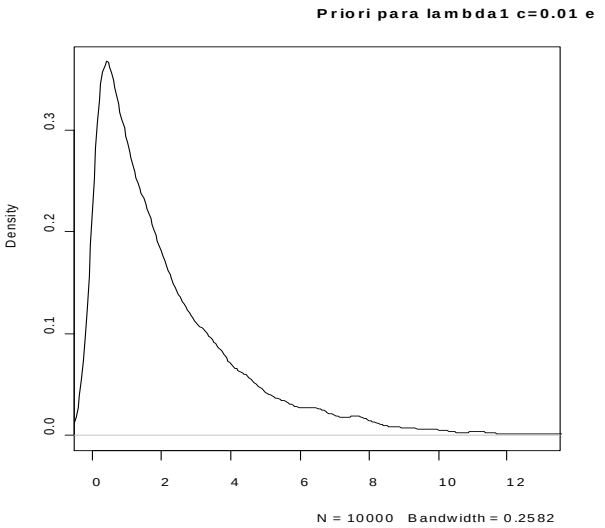

Figura A.10: Gráfico de λ_1 para c=0.01 e d=5

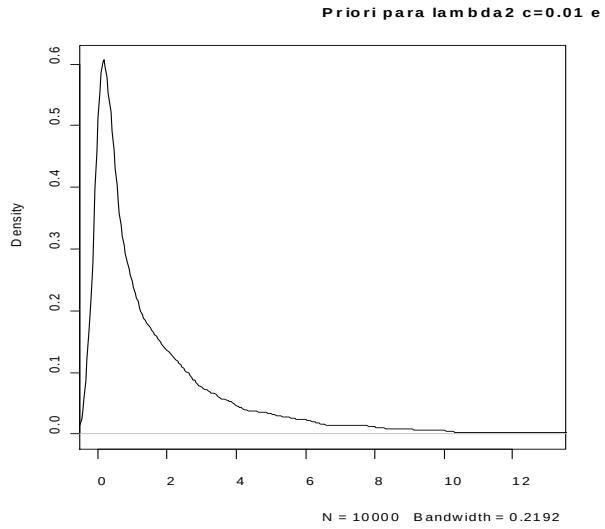

Figura A.11: Gráfico de λ_2 para c=0.01 e d=5

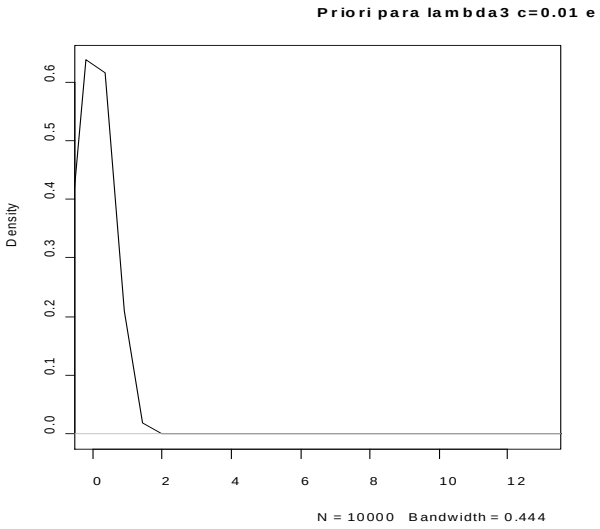

Figura A.12: Gráfico de λ_3 para c=0.01 e d=5

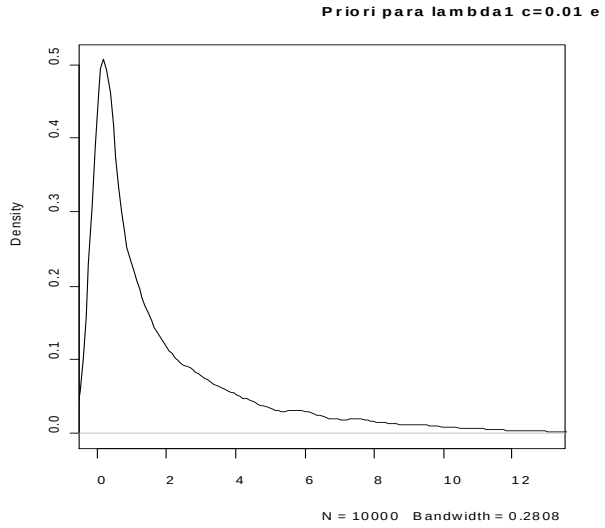

Figura A.13: Gráfico de λ_1 para c=0.01 e d=10

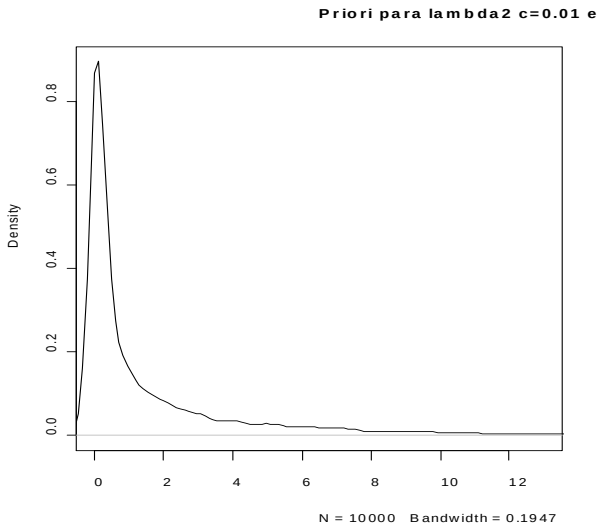

Figura A.14: Gráfico de λ_2 para c=0.01 e d=10

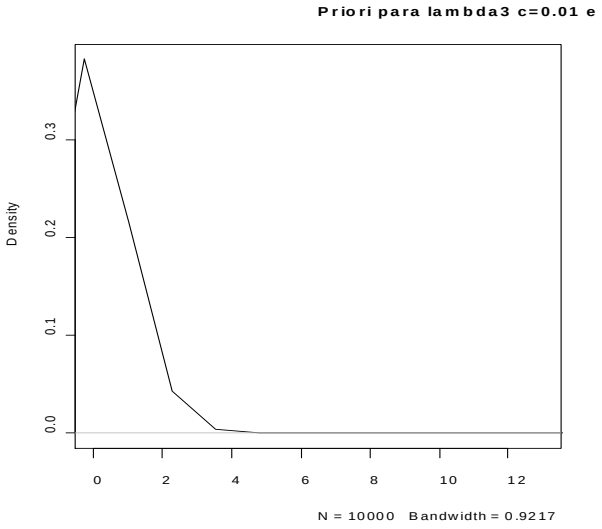

Figura A.15: Gráfico de λ_3 para c=0.01 e d=10

ii) Gráficos da correlação *c versus* probabilidade:

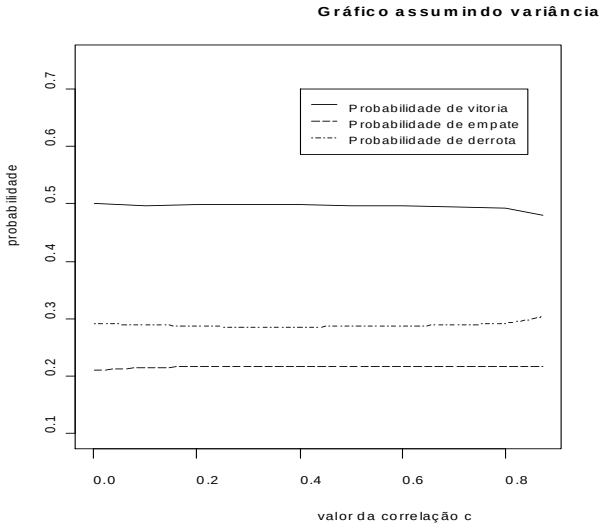

Figura A.16: Gráfico assumindo variância d=0.1

Obs: A legenda dos próximos gráficos será omitida e estas são a mesma do gráfico anterior.

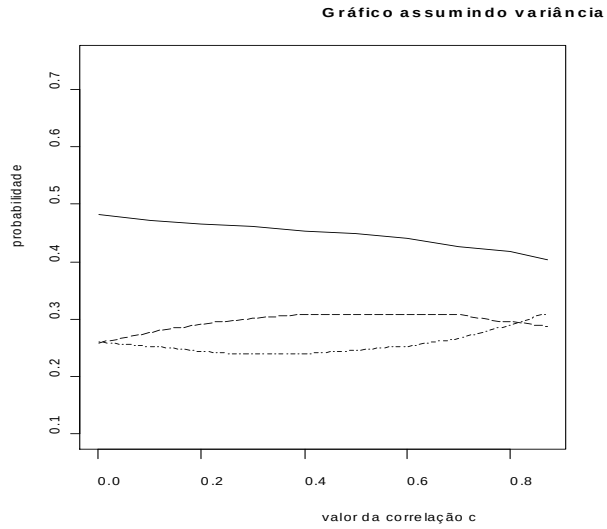

Figura A.17: Gráfico assumindo variância d=1

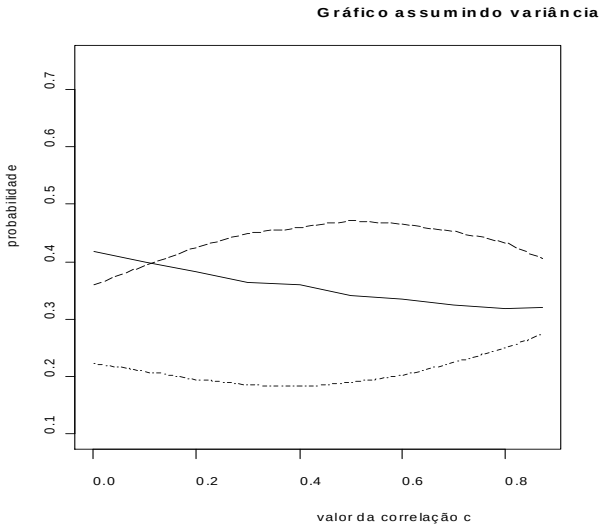

Figura A.18: Gráfico assumindo variância d=3

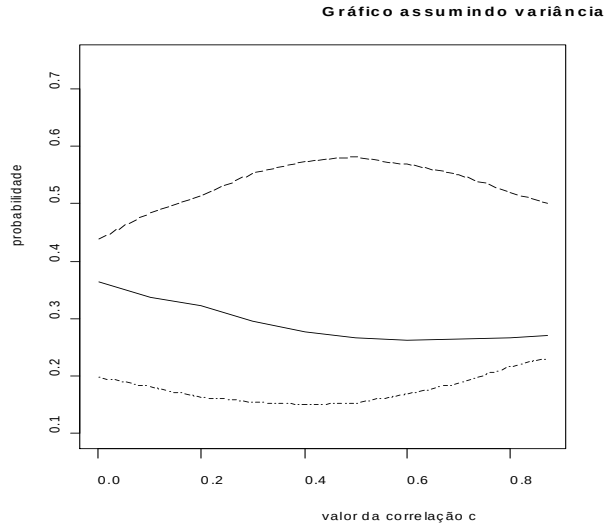

Figura A.19: Gráfico assumindo variância d=5

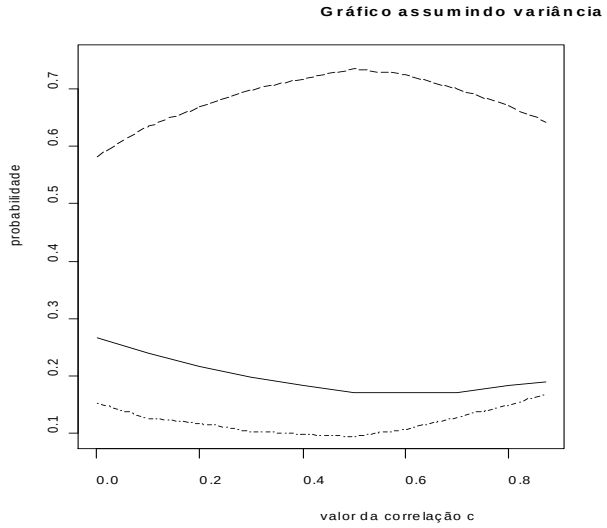

Figura A.20: Gráfico assumindo variância d=10

iii) Gráficos da variância d *versus* probabilidade:

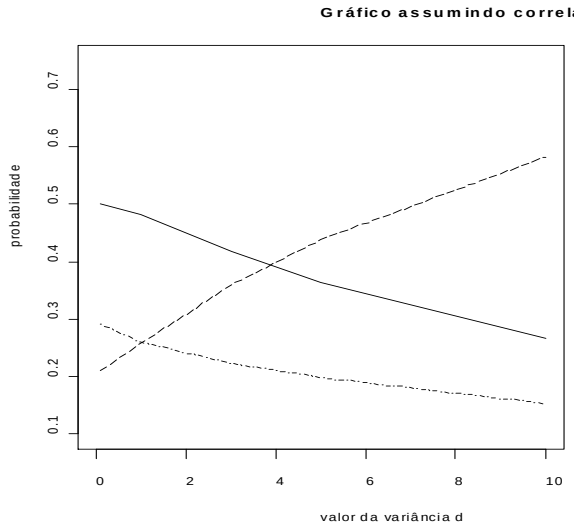

Figura A.21: Gráfico assumindo correlação c=0.01

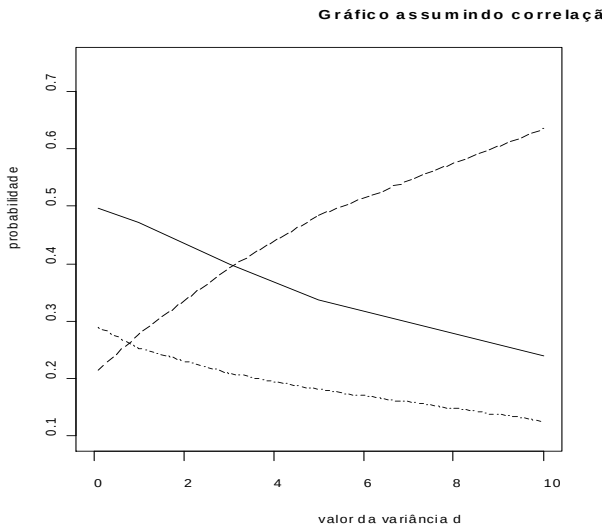

Figura A.22: Gráfico assumindo correlação c=0.1

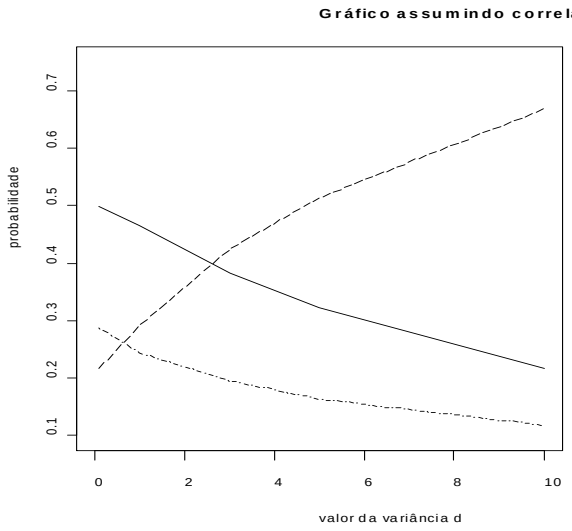

Figura A.23: Gráfico assumindo correlação c=0.2

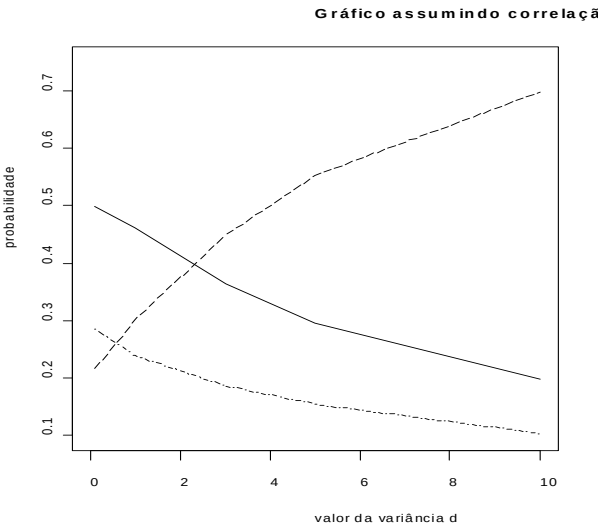

Figura A.24: Gráfico assumindo correlação c=0.3

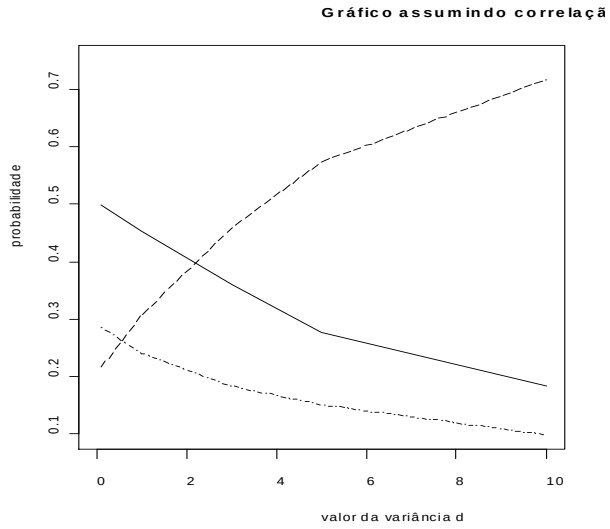

Figura A.25: Gráfico assumindo correlação c=0.4

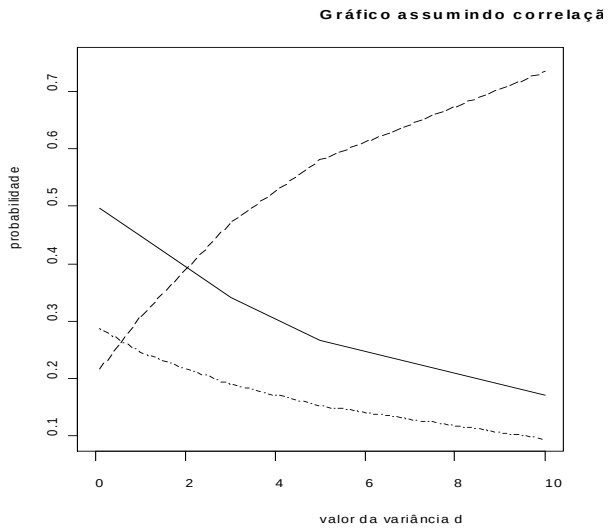

Figura A.26: Gráfico assumindo correlação c=0.5

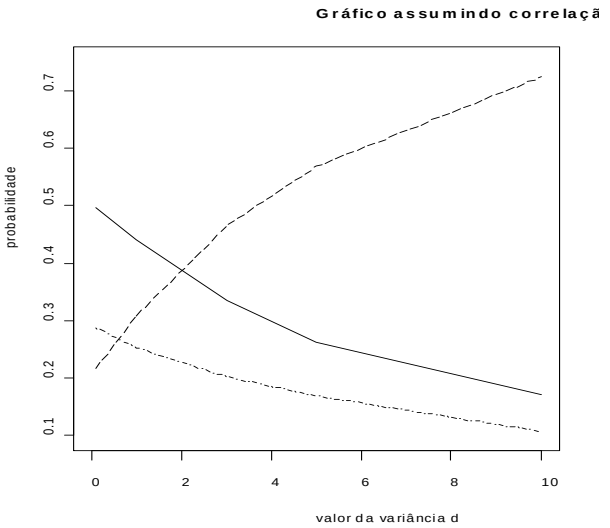

Figura A.27: Gráfico assumindo correlação c=0.6

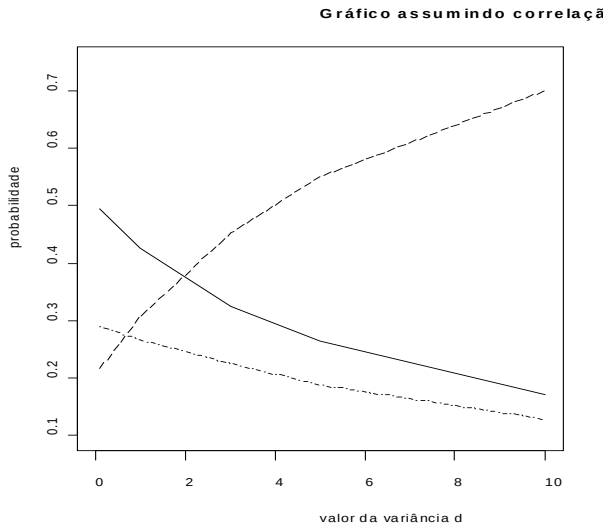

Figura A.28: Gráfico assumindo correlação c=0.7

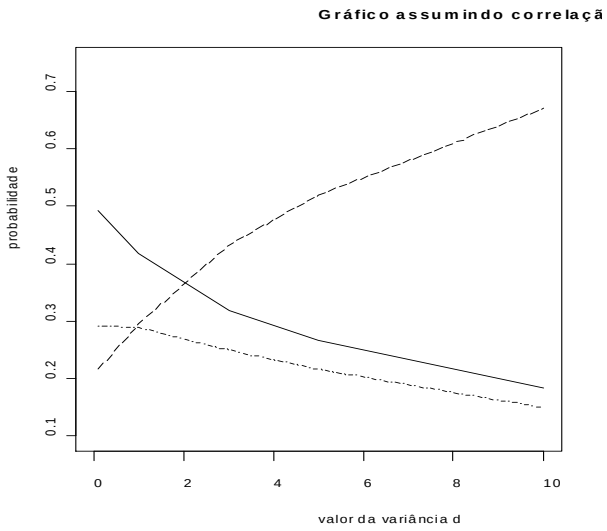

Figura A.29: Gráfico assumindo correlação c=0.8

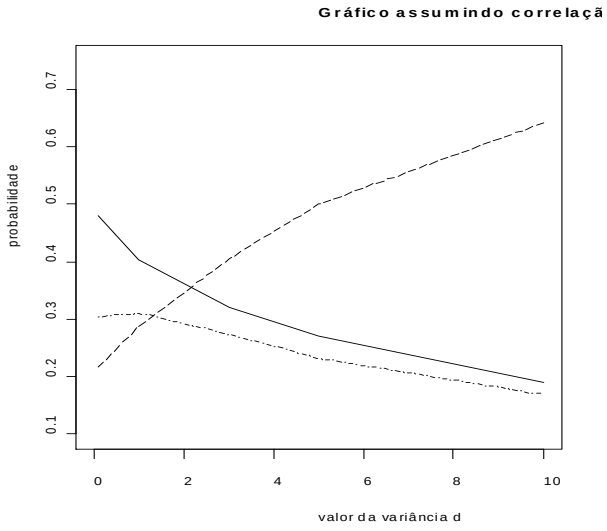

Figura A.30: Gráfico assumindo correlação c=0.872

Gráficos da correlação *c versus* probabilidade:

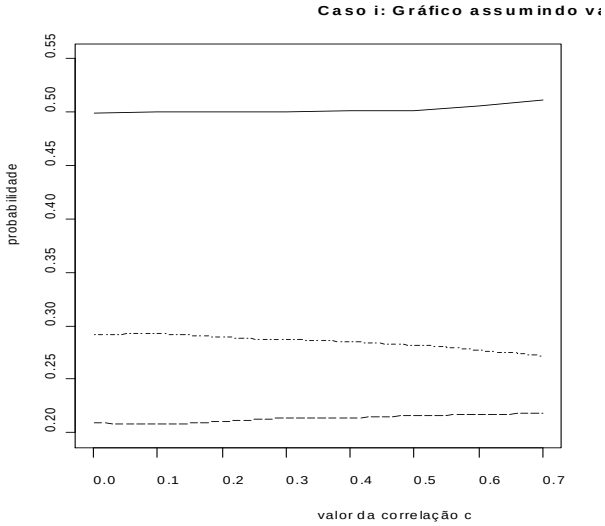

Figura A.31: Caso i assumindo variância d=0.1

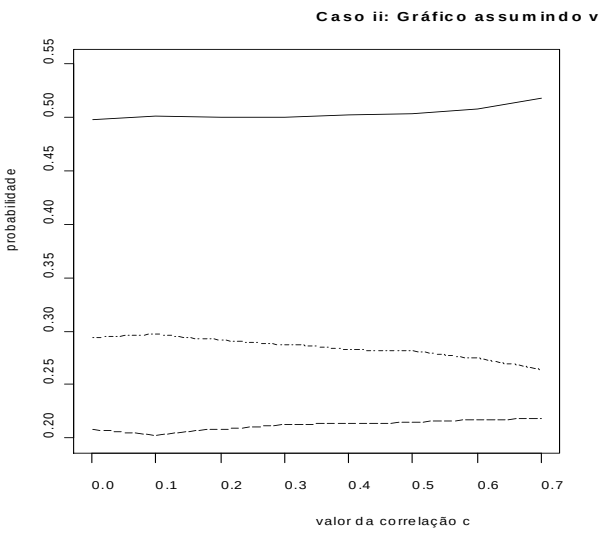

Figura A.32: Caso ii assumindo variância d=0.1

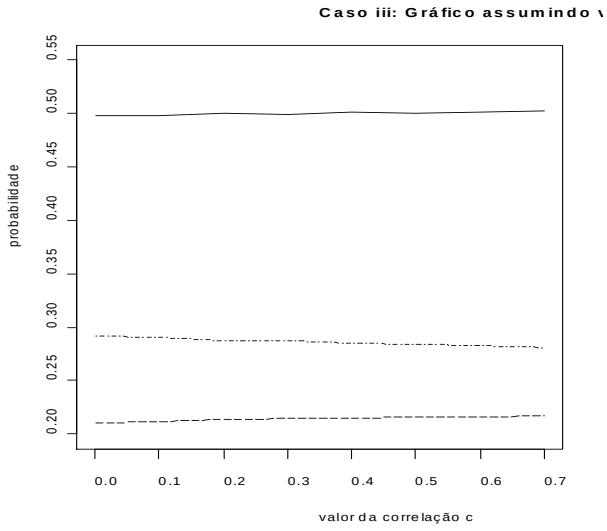

Figura A.33: Caso iii assumindo variância d=0.1

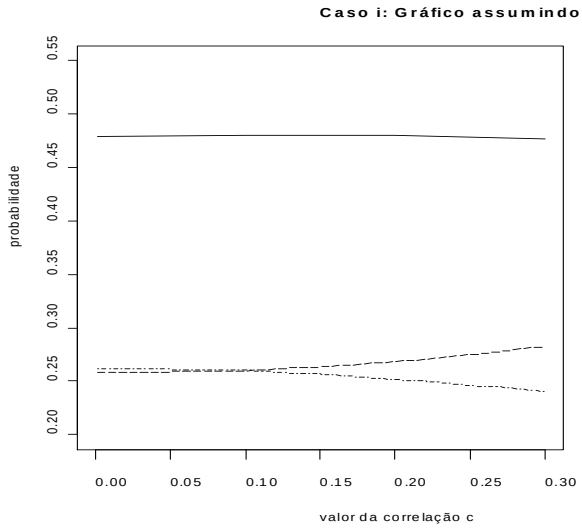

Figura A.34: Caso i assumindo variância d=1

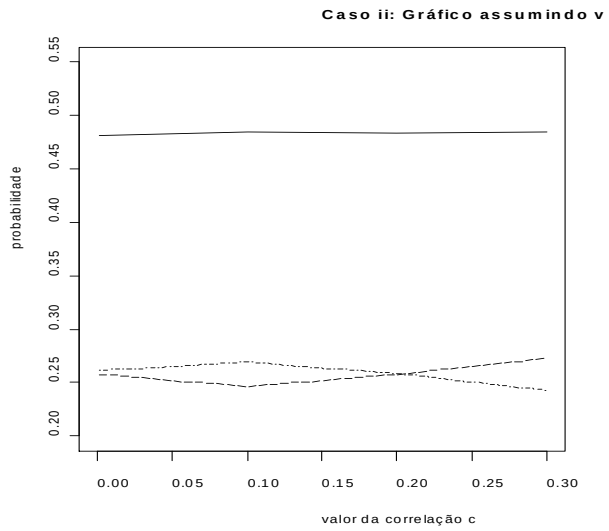

Figura A.35: Caso ii assumindo variância d=1

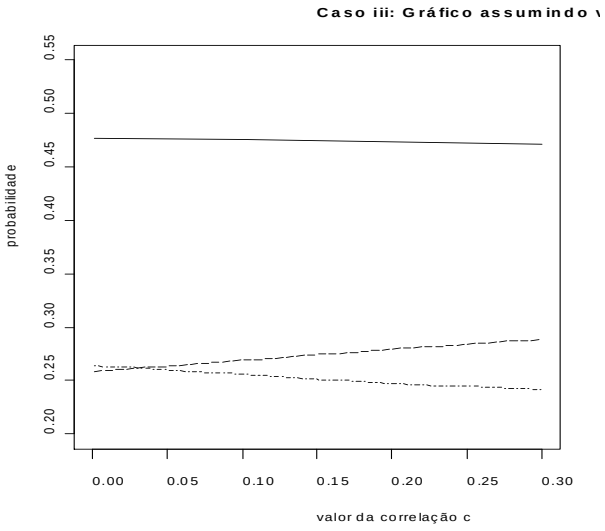

Figura A.36: Caso iii assumindo variância d=1

Apêndice B

Alguns programas

Neste apêndice, apresentamos alguns programas computacionais desenvolvidos no *software* R, utilizados para obter as probabilidades de vitória, empate e derrota.

Para os quatro primeiros programas carregamos o pacote MASS e definimos as matrizes Soma, Dif, S, T, SN, TN, GOLS, U e UN.

```
library(MASS)
Soma<-read.table("c:\\soma.txt")
Dif<-read.table("c:\\dif.txt")
S<-read.table("c:\\s.txt")
T<-read.table("c:\\t.txt")
SN<-read.table("c:\\sn.txt")
TN<-read.table("c:\\tn.txt")
GOLS<-read.table("c:\\gols.txt")
U<-read.table("c:\\u.txt")
UN<-read.table("c:\\un.txt")
```

1. O programa 1 é referente ao Método SD 0 introduzido no Capítulo 3.

Programa 1

```
alpha<-ginv(S)%*%Soma
beta<-ginv(T)%*%Dif
exmy<-SN%*%alpha
exny<-TN%*%beta
```

```
Casa<-(exny+exmy)/2
Visitante<-(exmy-exny)/2
for (i in 1:length(Casa))
                        {
                        if(Casa[i]>0) Casa[i]=Casa[i] else Casa[i]=0.25
                        if(Visitante[i]>0) Visitante[i]=Visitante[i] else Visitante[i]=0.25
                        }
SD0<-function(n)
{
k<-seq(1,n)
y<-seq(0,n)
Vitoria<-as.numeric(0)
Derrota<-as.numeric(0)
Empate<-as.numeric(0)
for (i in 1:length(Visitante))
                              {
                              Vitoria[i]<-sum(dpois(k,Casa[i])*ppois(k-1,Visitante[i]))
                              Derrota[i]<-sum(dpois(k,Visitante[i])*ppois(k-1,Casa[i]))
                              Empate[i]<-sum(dpois(y,Casa[i])*dpois(y,Visitante[i]))
                              }
placar<-cbind(Casa,Visitante,Vitoria,Empate,Derrota)
list(placar=placar)
}
```

2. O programa 2 é referente ao Método SD1 introduzido no Capítulo 3.

Programa 2

```
Somaquadrado=Soma^2
gamma=ginv(S)%*%Somaquadrado
alpha<-ginv(S)%*%Soma
beta<-ginv(T)%*%Dif
exmy<-SN%*%alpha
```

```
exny<-TN%*%beta
exmy<-SN%*%gamma
Casa<-(exny+2*exmy-exmyq+exmy^2)/2
Visitante<-(2*exmy-exny-exmyq+exmy^2)/2
Lambda12<-(exmyq-exmy^2-exmy)/2
for (i in 1:length(Casa))
                        {
                        if(Casa[i]>0) Casa[i]=Casa[i] else Casa[i]=0.25
                        if(Visitante[i]>0) Visitante[i]=Visitante[i] else Visitante[i]=0.25
                        }
SD1<-function(n)
{
k<-seq(1,n)
y<-seq(0,n)
Vitoria<-as.numeric(0)
Derrota<-as.numeric(0)
Empate<-as.numeric(0)
for (i in 1:length(Visitante))
                              {
                              Vitoria[i]<-sum(dpois(k,Casa[i])*ppois(k-1,Visitante[i]))
                              Derrota[i]<-sum(dpois(k,Visitante[i])*ppois(k-1,Casa[i]))
                              Empate[i]<-sum(dpois(y,Casa[i])*dpois(y,Visitante[i]))
                              }
placar<-cbind(Casa,Visitante,Vitoria,Empate,Derrota,Lambda12)
list(placar=placar)
}
```

3. O programa 3 é referente ao Método Chance I introduzido no Capítulo 3.

Programa 3

```
z<-glm(Gols~U-1,family=poisson)
z<-z[1]
betachapeu<-z$coefficients
for (i in 1:length(betachapeu))
                                {
                                if(is.na(betachapeu[i])==FALSE)
                                betachapeu[i]=betachapeu[i] else betachapeu[i]=0
                                }
elevado<-UN%*%betachapeu
Casa<-as.numeric(0)
Visitante<-as.numeric(0)
for (i in 1:length(elevado)/2)
                              {
                              Casa[i]<-exp(elevado[2*i-1])
                              Visitante[i]<-exp(elevado[2*i])
                              }
CHANCE1<-function(n)
{
k<-seq(1,n)
y<-seq(0,n)
Vitoria<-as.numeric(0)
Derrota<-as.numeric(0)
Empate<-as.numeric(0)
for (i in 1:length(Visitante))
                              {
                              Vitoria[i]<-sum(dpois(k,Casa[i])*ppois(k-1,Visitante[i]))
                              Derrota[i]<-sum(dpois(k,Visitante[i])*ppois(k-1,Casa[i]))
                              Empate<-sum(dpois(y,Casa[i])*dpois(y,Visitante[i]))
                              }
```

```
placar<-cbind(Casa,Visitante,Vitoria,Empate,Derrota)
list(placar=placar)
}
```

4. O programa 4 é referente ao Método Chance II introduzido no Capítulo 3.

Programa 4

```
betachance<-ginv(U)%*%Gols
betazero<-betachance[1]
esperanca<-UN%*%betachance
vetorlambda<-esperanca-betazero
for (i in 1:length(vetorlambda))
                                {
                                if(vetorlambda[i]>0) vetorlambda[i]=vetorlambda[i]
                                else vetorlambda[i]=0.25
                                }
Casa<-as.numeric(0)
Visitante<-as.numeric(0)
for (i in 1:length(vetorlambda)/2)
                                  {
                                  Casa[i]<-vetorlambda[2*i-1]
                                  Visitante[i]<-vetorlambda[2*i]
                                  }
CHANCE2<-function(n)
{
k<-seq(1,n)
y<-seq(0,n)
Vitoria<-as.numeric(0)
Derrota<-as.numeric(0)
Empate<-as.numeric(0)
```

```
for (i in 1:length(Visitante))
                                {
                                Vitoria[i]<-sum(dpois(k,Casa[i])*ppois(k-1,Visitante[i]))
                                Derrota[i]<-sum(dpois(k,Visitante[i])*ppois(k-1,Casa[i]))
                                Empate[i]<-sum(dpois(y,Casa[i])*dpois(y,Visitante[i]))
                                }
placar<-cbind(Casa,Visitante,Vitoria,Empate,Derrota,betazero)
list(placar=placar)
}
```

Referências Bibliográficas

[1] Arruda, M.L. (2000), Poisson, Bayes, Futebol e DeFinetti, Tese de Mestrado, IME-USP.

[2] Barnett, V. e Hilditch, S. (1993), The effect of an artificial pitch surface on home team performances in football (soccer). *Journal of the Royal Statistical Society A*, **1956**, 39-50.

[3] Church, S. e Hughes, M. (1987), A computerised approach to soccer notation analysis. *Abstract of the 1st World Congress of Science and Football, Liverpool,* p. 20. Liverpool Polytechnic, Liverpool.

[4] Clark, S. R. e Norman, J. M. (1995), Home ground advantage of individual clubs in English soccer. *The Statistician*, **44**, 509-521.

[5] DeFINETTI, B. (1972), Probability, Induction and Statistics, London: John Wiley.

[6] DeGroot, M.H. (2002), Probability and Statistics – 3a edição, Addison-Wesley.

[7] Dwass, M. e Teicher, H. (1957), On Infinitely Divisible Random Vectors, *Annals of Mathematical Statistics*, **28**, 461-470.

[8] Ehlers, R.S. (2005), Introdução à Inferência Bayesiana. Departamento de Estatística,UFPR. Disponível em http://leg.est.ufpr.br/∼ehlers/notas/bayes2006.

[9] EMONET, B. (2000), Revisiting Statistical Applications in Soccer, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland.

[10] Franks, I. M. (1988), Analysis of association football. *Soccer Journal*, **33**, 35-43.

[11] Gamerman, D. (1997), Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Chapman & Hall, Londres.

[12] Glickman, M. E. (1993), Paired Comparison Models With Time-Varying Parameters. Departament of Statistics, Havard University, Cambridge.

[13] Glickman, M. E. e Stern H. S. (1998), A state-space model for National Football League scores.

[14] Harrison, P. J. e Stevens, C. F. (1996), Bayesian forecasting (com discussão). *Journal of the Royal Statistical Society*, Series B, **38**, 205 – 247

[15] Holgate, P. (1964), Estimation for the Bivariate Poisson Distribution, *Biometrika,* **52**, 241-245.

[16] Johnson, N.L, Kotz, S. e Balakrishnan, N. (1997), Discrete Multivariate Distributions, New York: John Wiley & Sons.

[17] Junior, O. G. S. e Gamerman, D. (2004), Previsão de Partidas de Futebol Usando Modelos Dinâmicos - BPO.

[18] Karlis, D. e Ntzoufras, I. (2003), Analysis of Sports Data by Using Bivariate Poisson Models, *The Statistician,* **52**, Part 3, PP 381-393.

[19] Karlis, D. e Tsiamyrtiz, P. (2004), Exact Bayesian Inference for Bivariate Poisson Data. Athens University of Economics and Business, 76 Patission Str 10434, Athens, Greece.

[20] Keller, J. B. (1994), A Characterization of the Poisson Distribution and the Probability of Winning a Game. *The American Statistician*, **48** (4), 294-298.

[21] Knorr-Held, L. (2000), Journal of the Royal Statistical Society, Series D, *The Statistician*, **49**, 200-225.

[22] Kocherlakota, S. e Kocherlakota, K. (1992), Bivariate Discrete Distributions, New York: Marcel Dekker.

[23] Kuk, A. Y. C. (1995), Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players. T*he Statistician*, **44**, 523-528.

[24] Lee, A. J. (1997), Modeling Scores in the Premier League: Is Manchester United Really the Best?, *Chance*, **10** (1), 15-19.

[25] Moroney, M. J. (1956), Facts from figures. 3rd edition, Penguin, London.

[26] Paukku, T. (1994),And it's another goal. *New Scientist*, **40**, 41-43.

[27] Pole, A., West, M. e Harrison, J. (1994), Applied Bayesian Forecasting and time series analysis. Springer, Nova York.

[28] Pollard, R.(1986), Home advantage in soccer: a retrospective analysis. *Journal of Sports Sciences*, **4**, 237-248.

[29] Tsionas, E.G. (1999), Bayesian Analysis of the Multivariate Poisson Distribution. *Communications in Statistics - Theory and Methods*, **28**, 431-451.

[30] Venables, W. N. e Ripley, B. D. (1999), Modern Applied Statistics with S-PLUS – 3a edição, Springer, p.100.