**ANÁLISE BAYESIANA DE
DADOS CITOGENÉTICOS**

Maria Regina Madruga

DISSERTAÇÃO APRESENTADA
AO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DA
UNIVERSIDADE DE SÃO PAULO
PARA
OBTENÇÃO DO GRAU DE MESTRE
EM
ESTATÍSTICA

Área de Concentração: Estatística

Orientador: Prof. Dr. Carlos Alberto de Bragança Pereira

-São Paulo, Março de 1994-

ANÁLISE BAYESIANA DE DADOS CITOGENÉTICOS

Este exemplar corresponde à redação
final da dissertação devidamente corrigida
e defendida por Maria Regina Madruga
e aprovada pela comissão julgadora.

São Paulo, 8 de Março de 1994.

Banca examinadora:

- Prof. Dr. Carlos Alberto de Bragança Pereira (orientador) - IME-USP
- Prof. Dr. Sérgio Wechsler - IME-USP
- Profa. Dra. M. Nazareth Rabello-Gay - ILSI

*À memória
de minha mãe,
um exemplo
de mulher*

AGRADECIMENTOS

- Ao Carlinhos pela orientação dedicada na elaboração deste trabalho.
- À minha família pela confiança em mim depositada, em especial, ao meu pai e à Beth.
- Ao Héilton pelo carinho e companheirismo, sempre presentes durante a redação deste trabalho.
- Aos amigos que me apoiaram e ajudaram durante esta etapa da minha vida, entre outros, Cleonice, Giovani, Isaura, Rosângela e Ricardo.
- Aos colegas do Depto. de Matemática e Estatística da Universidade Federal da Paraíba (campus II) pelo incentivo durante a elaboração desta dissertação, em especial, aos professores: Antônio José, Chiquinho e Claudia.
- Ao Cláudio, à Mara e ao Pinheiro, pela convivência alegre no CRUSP.
- À Tânia H. Ochi-Lohmann, pesquisadora do IPEN, por ter cedido os dados utilizados no Capítulo 3.
- À CAPES pelo apoio financeiro no início do programa de mestrado.

RESUMO

Problemas estatísticos em citogenética normalmente envolvem dados categorizados relacionados a danos ocorridos em células. Esses danos necessariamente são eventos raros quando relacionados com o total de células. Isto produz os já conhecidos problemas de análise estatística de dados provenientes de eventos raros. Soluções Bayesianas para comparação de proporções são apresentadas e contempladas com as alternativas clássicas, que normalmente sofrem de imprecisões devido às conclusões condicionais obtidas. Uma calibração Bayesiana é apresentada para solucionar problemas de ajuste de modelo em experimentos de dosimetria. Esta parece ser a primeira calibração feita sem aproximações assintóticas em dados multinomiais.

ABSTRACT

Statistical problems of cytogenetics experiments usually involve categorical data obtained by counting of cells with aberration. The consequence is the well known problem of analysing statistical data obtained from the observation of rare events. Bayesian solutions to compare proportions are presented and confronted with the classical alternatives that usually suffer from conditional conclusions. A Bayesian calibration is introduced to solve adjustment problems in dosimetry experiments. This seems to be the first calibration obtained without the use of asymptotic arguments.

Conteúdo

1	INTRODUÇÃO	3
2	TESTE DE IGUALDADE DE PROPORÇÕES	6
2.1	Considerações iniciais	6
2.2	Teste de Significância Condicional	12
2.3	Teste χ^2 de homogeneidade	19
2.4	Teste Bayesiano	26
2.4.1	Informação <i>a priori</i>	26
2.4.2	Procedimento do Teste	27
3	DOSIMETRIA BAYESIANA	36
3.1	Considerações iniciais	36
3.2	Caso Binomial	38
3.3	Caso Multinomial	58
3.4	Outras aplicações	65
3.4.1	Região de Credibilidade	65
3.4.2	Comparação entre as frequências de anomalias de diversos grupos de indivíduos	67

4 CONSIDERAÇÕES FINAIS	73
A APROXIMAÇÃO NORMAL BIVARIADA	74
BIBLIOGRAFIA	81

Capítulo 1

INTRODUÇÃO

O desenvolvimento industrial tem como consequência inevitável a exposição do homem a um número crescente de agentes que, se por um lado lhe aumentam o conforto e a riqueza, constituem, por outro, um risco para sua saúde. As aberrações cromossômicas constituem uma fração relevante do dano causado por agentes físicos, químicos e biológicos ao material genético. Muitos compostos capazes de causar mutações gênicas também causam mutações cromossômicas. Mutação é toda alteração do material genético de uma célula que não resulta de segregação ou recombinação. Assim, é imprescindível a inclusão de testes citogenéticos na avaliação do potencial mutagênico de um composto. (Para mais neste tópico, ver Rabello-Gay, Rodrigues & Monteleone-Neto, 1991). Estes testes produzem resultados numéricos que precisam ser avaliados estatisticamente. O objetivo principal desta dissertação é apresentar técnicas estatísticas Bayesianas para avaliação de alguns resultados citogenéticos. Estes estudos contribuirão para o estabelecimento de normas de uso destes agentes.

Em experimentos citogenéticos as unidades observacionais, em geral, são células, e o grande interesse é estudar o efeito destes agentes em células de indivíduos expostos ao mesmo. Um dos problemas que surgem é a comparação das proporções de células com aberrações cromossômicas entre indivíduos que sofreram exposição a uma alta dosagem e indivíduos que sofrem uma exposição natural (no sentido de que todo indivíduo está sujeito a uma exposição diária de agentes suspeitos). No Capítulo 2 discutimos este problema, apresentando algumas soluções indicadas na literatura estatística Clássica e apresentando a solução Bayesiana. Na seção 2.2 descrevemos o Teste de Significância Condicional, que baseia-se no cálculo de um nível de significância descritivo, utilizando a teoria de Inferência Parcial (ver Pereira & Lindley, 1987). A distribuição de probabilidades condicional utilizada neste teste envolve uma redução do espaço amostral, obtida pelo Método da Redução. A validade deste método é questionável, uma vez que não temos uma regra geral para encontrar a reparametrização adequada. Ainda, procedimentos baseados em níveis descritivos violam o Princípio da Verossimilhança, que enuncia o seguinte: toda inferência deve ser baseada apenas no ponto amostral observado, não em outros pontos (mais extremos) que poderiam ter sido observados e não foram. Na seção 2.3 apresentamos o Teste χ^2 de homogeneidade, que é utilizado nos casos em que o número de aberrações observadas é relativamente grande. Na seção 2.4 descrevemos o Teste Bayesiano, que assume conhecimentos precedentes para as chances das hipóteses em teste e utiliza a visão subjetivista de probabilidade.

No Capítulo 3, tratamos da questão da dosimetria. O problema estatístico fundamental em Toxicologia é obter uma relação funcional entre dose e res-

posta, onde dose refere-se à concentração do agente suspeito ao qual o indivíduo é exposto e resposta é o fenômeno observado no indivíduo após a exposição. A solução que apresentaremos envolve técnicas Bayesianas como, por exemplo, em Bender et al. (1988) e Groër & Pereira (1987). Nestes artigos considerou-se a distribuição de Poisson e modelos lineares, pois o interesse era em níveis de dose baixos (e, consequentemente, respostas baixas). Os dados que analisaremos envolvem valores baixos e altos da dose-resposta, e as soluções apresentadas nestes artigos não são adequadas. Na seção 3.2 apresentamos os resultados relativos ao caso univariado, isto é, a resposta é uma variável unidimensional. Na seção 3.3, apresentamos o caso em que a resposta é um vetor de três componentes. Na seção 3.4 apresentamos algumas aplicações envolvendo estas técnicas.

Capítulo 2

TESTE DE IGUALDADE DE PROPORÇÕES

2.1 Considerações iniciais

Em estudos citogenéticos, o teste de igualdade de proporções é realizado com o objetivo de detectar efeitos mutagênicos (ou anomalias), de algum agente suspeito, em indivíduos que sofreram exposição ao mesmo. Os dados são obtidos, para cada uma das unidades experimentais, registrando-se a ocorrência ou não de anomalias. As unidades experimentais, em geral, são células e estes efeitos mutagênicos podem ser aberrações cromossômicas. O experimento consiste na análise de células, obtidas em culturas de sangue, de indivíduos que foram expostos ao produto (grupo dos expostos) e de indivíduos que não foram expostos (grupo dos controles). A exposição ao produto pode ter sido "in vitro" ou "in vivo", ou seja, as células dos

indivíduos sofrem exposição direta em laboratório ou os indivíduos sofrem a exposição e, através da cultura de sangue, suas células são analisadas.

O principal objetivo neste capítulo é a comparação entre as frequências de anomalias dos dois grupos, pois deseja-se saber se há um efeito significativo do produto nos indivíduos expostos. O modelo naturalmente associado a este tipo de experimento é o Binomial, pois trata-se de um fenômeno dicotômico com resultados independentes: cada unidade experimental, pertencente a cada grupo, é classificada como portadora ou não de anomalias, e a ocorrência ou não de anomalias em uma determinada célula não influencia o efeito observado em outras células. Logo, o modelo Binomial é o adequado neste experimento, uma vez que a independência e a dicotomia o caracterizam.

O método mais utilizado para testar igualdade de proporções é o teste χ^2 de homogeneidade. Contudo, este teste é eficiente nas situações em que as amostras são grandes e a probabilidade do evento não se afasta "muito" de 0.5. Os dados aqui discutidos estão associados a um evento considerado raro, pois necessita-se de um número razoavelmente grande de células para que seja possível detectar uma quantidade significativa de ocorrências de uma determinada anomalia. Ou seja, a primeira condição para a eficiência do teste χ^2 é satisfeita, visto que um número grande de células é observado, mas a segunda condição não é satisfeita, pois a probabilidade de ocorrência de anomalias é bastante pequena. Observe que é necessário assumir uma exposição a uma baixa dosagem do produto, visto que a agressão pode ser fatal ao indivíduo. A dosagem baixa é a causa pela qual o evento se torna raro. No capítulo 3 será discutida a relação entre a dosagem do produto a que as células foram expostas e a frequência observada de anomalias. E-

xistem correções que são utilizadas em situações de frequências baixas, mas estas correções são teoricamente indicadas para os casos de amostras não muito grandes. Uma alternativa para as situações em que as amostras são grandes e as frequências pequenas é um teste de significância condicional, (apresentado na seção 2.2), semelhante ao teste exato de Fisher. Na seção 2.4, apresentaremos um método que utiliza as técnicas Bayesianas para testes de hipóteses.

O exemplo a seguir ilustra o tipo de problema citogenético.

Exemplo 2.1.1 *Com o objetivo de detectar o possível efeito citogenético de certo inseticida observou-se 5 operários que lidam com a droga (expostos) e 5 outros, da mesma fábrica, que não têm contato com o inseticida (controles). Prepararam-se lâminas de culturas de linfócitos de cada operário e analisaram-se 1000 células de cada indivíduo, registrando-se o número de células com anomalias cromossômicas. As frequências observadas são mostradas na tabela 2.1.*

Tabela 2.1: Frequências de células com anomalias

Operários	Frequências		Total
	com anomalias	normais	
Expostos	6	4994	5000
Controles	1	4999	5000
Total	7	9993	10000

Observando a tabela 2.1, percebe-se que o exemplo acima se encaixa na situação de amostra grande (10000 células) e frequência baixa (7 células). A

hipótese de interesse, H_1 , é a de que a proporção de células com anomalias nos operários expostos é diferente dos operários controles. A hipótese nula, H_0 , é a de que as duas proporções são iguais.

Antes de dar início a um procedimento de teste, alguns conceitos a serem utilizados posteriormente serão enunciados:

Definição 2.1.1 *Nível de significância de um teste é um limite superior, que se fixa, para a probabilidade de rejeitar a hipótese nula quando ela é verdadeira.*

Definição 2.1.2 *Nível descritivo de um teste é a probabilidade, sob a hipótese nula, do evento formado pelo que se observou e por todos os casos mais extremos que possam ser observados. Mais extremos no sentido de que contrariam a hipótese nula, em favor da hipótese alternativa, mais do que os dados efetivamente observados.*

Definição 2.1.3 *Uma estatística T é ancilar com respeito a um parâmetro θ se a verossimilhança pode ser escrita como $p(x|\theta) = p(x|t, \theta)p(t)$, ou seja, se a distribuição marginal de T não depende de θ .*

Definição 2.1.4 *Uma estatística T é suficiente com respeito a um parâmetro θ se a verossimilhança pode ser escrita como $p(x|\theta) = p(x|t)p(t|\theta)$, ou seja, se a distribuição condicional de X , dado T , não depende de θ .*

Como mencionado anteriormente, o modelo aqui considerado é o Binomial. A função de probabilidade associada a uma variável aleatória X com distribuição Binomial com parâmetro (n, p) , ou seja, observa-se n unidades

amostrais com probabilidade de ocorrência de anomalias em cada uma das unidades igual a p , é dada por

$$P[X = x] = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}; \quad x = 0, 1, \dots, n \quad (2.1)$$

Observe que na situação de exposição a baixa dosagem do produto (como no exemplo 2.1.1), tem-se um contexto de evento raro, ou seja, n é "grande" e p "pequeno". Neste caso, a dificuldade na utilização da expressão (2.1) é grande (no exemplo 2.1.1 seria necessário calcular o fatorial de 5000). Para simplificar os cálculos recomenda-se o uso da distribuição Poisson, que é uma excelente aproximação para a distribuição Binomial, no caso de eventos raros. Portanto, pode-se aproximar a expressão (2.1) por

$$P[X = x] = \frac{(np)^x e^{-np}}{x!}; \quad x = 0, 1, \dots \quad (2.2)$$

Note que considerar o número de células com anomalias, X , tendo distribuição Binomial com parâmetro (n, p) , implica em se afirmar que células entre indivíduos podem ser permutadas. Outra implicação deste modelo é o fato de esquecermos o indivíduo como elemento de observação e considerarmos apenas a célula como unidade observacional, uma vez que supomos o mesmo parâmetro p para as distribuições Binomiais associadas a diferentes indivíduos. Apesar deste modelo ser utilizado com frequência na literatura, objeções ao seu uso podem ser facilmente apresentadas. Um modelo mais realístico, porém muito mais complexo, seria o da mistura de Binomiais. Isto é, podemos imaginar um gerador aleatório de valores para p . Para a situação do exemplo 2.1.1 observamos as variáveis X_i e Y_j ($i, j = 1, 2, 3, 4, 5$),

que representam o número de células com aberrações no i -ésimo indivíduo do grupo dos expostos e no j -ésimo indivíduo do grupo dos controles, respectivamente. A cada uma destas variáveis, X_i ou Y_j , está associada a respectiva proporção, p_i ou q_j . No modelo da mistura podemos supor, por exemplo, que p_1, p_2, p_3, p_4, p_5 são observações independentes de uma distribuição Beta com parâmetro (α_1, β_1) e q_1, q_2, q_3, q_4, q_5 são observações independentes de uma distribuição Beta com parâmetro (α_2, β_2) . Desta forma, para o grupo dos expostos, temos o seguinte modelo:

$$X_i | [P = p_i] \sim B(n_i, p_i) \quad i = 1, 2, 3, 4, 5$$

E, para o grupo dos controles, temos

$$Y_j | [Q = q_j] \sim B(m_j, q_j) \quad j = 1, 2, 3, 4, 5$$

Logo, $X_i | \alpha_1, \beta_1$ terá distribuição Beta-binomial com parâmetro $(n_i; \alpha_1, \beta_1)$ e $Y_j | \alpha_2, \beta_2$ terá distribuição Beta-binomial com parâmetro $(m_j; \alpha_2, \beta_2)$. Ver Basu & Pereira (1982) para um completo estudo da distribuição Beta-binomial. Assim, a comparação entre controles e expostos se resumiria ao teste da hipótese $H : \alpha_1 = \alpha_2, \beta_1 = \beta_2$.

O nosso interesse neste capítulo é apresentar uma alternativa Bayesiana ao teste frequentemente utilizado. O tratamento estatístico deste modelo mais complexo fugirá aos nossos objetivos, mas pretendemos em um trabalho posterior nos dedicarmos à solução Bayesiana do teste comparativo entre expostos e controles neste modelo de mistura.

2.2 Teste de Significância Condicional

Observando os dados da tabela 2.1, a intuição nos leva a considerar que o total de células anômalas (7 células) nada deve informar sobre a relação entre as frequências de ocorrência de anomalias nos dois grupos de operários. Ou seja, suponha que a tabela 2.1 não estivesse disponível e a única informação acessível com relação ao que se observou fosse este total de 7 células anômalas. Neste caso, não seria possível conhecer a relação informativa de 1 para 6 observada entre as proporções. Esta relação é informativa no sentido de que ela nos traz conhecimento a respeito das quantidades de interesse p_1 e p_2 , as probabilidades de ocorrência de anomalias nos operários expostos e controles, respectivamente. Como as duas amostras são de mesmo tamanho, 5000, sob a hipótese de igualdade nos dois grupos, espera-se que a frequência de células anômalas obtida nos expostos seja equivalente à frequência de "caras" observada em 7 lançamentos de uma moeda honesta. A probabilidade condicional de obter 6 células anômalas no grupo dos operários expostos, dado o total de 7 células anômalas, sob a hipótese de igualdade das proporções, é dada por

$$\frac{7!}{6!(7-6)!}(0.5)^6(1-0.5)^{7-6} = 0.0547.$$

Um resultado mais extremo, no sentido de que favoreceria a hipótese alternativa H_1 mais do que o resultado obtido, seria a obtenção de 7 células anômalas para os expostos e nenhuma para os controles. A probabilidade para este resultado mais extremo, sob a hipótese nula, é dada por

$$\frac{7!}{7!(7-7)!}(0.5)^7(1-0.5)^{7-7} = 0.0078.$$

Segundo a Definição 2.1.2, o nível descritivo é dado pela soma destas duas probabilidades, ou seja, $0.0547 + 0.0078 = 0.0625$. Logo, a hipótese nula seria rejeitada para todo nível de significância maior que 0.0625. A interpretação para o nível descritivo é a seguinte: se seu valor for pequeno, conclui-se que os dados obtidos têm uma forte evidência contra a hipótese nula; ou seja, esta hipótese é incompatível com os dados; caso contrário, conclui-se que a hipótese nula é compatível com os dados obtidos.

Considere uma situação geral, Tabela 2.2, em que se retira uma amostra de n_1 células da população de indivíduos expostos a uma determinada droga e uma amostra de n_2 células da população de indivíduos controles.

Tabela 2.2: Tabela de contingência 2x2 (caso geral)

Grupos	Frequências		Total
	com anomalias	normais	
Expostos	x_1	$n_1 - x_1$	n_1
Controles	x_2	$n_2 - x_2$	n_2
Total	y	$n - y$	n

Os valores x_1 e x_2 na tabela 2.2 são, respectivamente, os valores observados das variáveis aleatórias X_1 e X_2 , que representam as frequências de células com anomalias nos dois grupos de indivíduos analisados. O planejamento descrito para a situação geral mostrada nesta tabela nos leva a concluir que as variáveis aleatórias X_1 e X_2 são independentes, tendo distribuição Bi-

nomial com parâmetros (n_1, p_1) e (n_2, p_2) , respectivamente. As hipóteses de interesse a serem confrontadas são $H_0 : p_1 = p_2$ e $H_1 : p_1 \neq p_2$. A função de probabilidade conjunta para (X_1, X_2) , a verossimilhança, é dada por

$$L(x_1, x_2 | p_1, p_2) = \frac{n_1!}{x_1!(n_1 - x_1)!} p_1^{x_1} (1 - p_1)^{n_1 - x_1} \frac{n_2!}{x_2!(n_2 - x_2)!} p_2^{x_2} (1 - p_2)^{n_2 - x_2} \quad (2.3)$$

O teste condicional, que será discutido nesta seção, é indicado para testar a igualdade de proporções associadas a eventos raros. Neste caso, como p_1 e p_2 são próximos de zero e produzem valores de x_1 e x_2 pequenos, em relação a n_1 e n_2 , deve-se utilizar as distribuições aproximadas para X_1 e X_2 , ou seja, a função de verossimilhança (2.3) é aproximada por

$$L(x_1, x_2 | p_1, p_2) = \frac{e^{-\lambda_1} (\lambda_1)^{x_1}}{x_1!} \frac{e^{-\lambda_2} (\lambda_2)^{x_2}}{x_2!} \quad (2.4)$$

onde $\lambda_1 = n_1 p_1$ e $\lambda_2 = n_2 p_2$

Este teste não utiliza a distribuição incondicional de (X_1, X_2) expressa em (2.4). O teste baseia sua decisão de favorecer ou não H_0 na distribuição condicional de X_1 dada a estatística $Y = X_1 + X_2$. O argumento usado no teste de significância condicional é que toda informação relacionada as hipóteses acima está incluída nesta distribuição condicional. Por outro lado, supõe-se que a distribuição marginal de Y é não informativa com relação às hipóteses. Observe que podemos fazer a seguinte reparametrização

$$\theta = \frac{\lambda_1}{\lambda_1 + \lambda_2} \quad e \quad \phi = \lambda_1 + \lambda_2 \quad (2.5)$$

e, de forma análoga, testar as seguintes hipóteses $H_0 : \theta = 1/2$ e $H_1 : \theta \neq 1/2$ (ou seja, comparar as médias das duas distribuições Poisson).

A função de verossimilhança associada a esta reparametrização, e substituindo x_2 por $y - x_1$, é dada por

$$L(x_1, y | \theta, \phi) = \frac{e^{-\phi} \phi^y}{y!} \frac{y!}{x_1! (y - x_1)!} \theta^{x_1} (1 - \theta)^{y - x_1} \quad (2.6)$$

Note que a função de verossimilhança pode ser escrita como o produto das seguintes funções de probabilidade

$$L(x_1, y | \theta, \phi) = l_1(y | \phi) \times l_2(x_1 | y, \theta), \quad (2.7)$$

ou seja, a distribuição marginal de Y é Poisson com parâmetro ϕ e a distribuição condicional de X_1 dado Y é Binomial com parâmetro (y, θ) . O teste de significância condicional baseia-se no seguinte fato: a relação entre os dados observados (x_1, x_2) e o parâmetro de interesse é descrita unicamente pela função de verossimilhança. Desta forma, o valor da função de verossimilhança deveria alterar-se quando houvesse variação no valor do parâmetro, caso contrário, não seria possível obter alguma informação relevante sobre o parâmetro através destes dados. O argumento para o uso do teste é que na fatoração de $L(x_1, y | \theta, \phi)$, expressão (2.7), a função $l_1(y | \phi)$ não se altera ao variarmos o valor do parâmetro de interesse θ , portanto, pode ser abandonada. Consequentemente, deve-se considerar apenas aqueles fatores que são funções de θ . Observe que a estatística Y é suficiente com respeito a ϕ e é ancilar com respeito a θ . O argumento acima pode ser traduzido da seguinte forma: se Y é ancilar nenhuma informação sobre θ pode ser obtida

de Y , portanto, as inferências sobre θ podem ser baseadas apenas na função $l_2(x_1|y, \theta)$.

Sob a hipótese $H_0 : p_1 = p_2 = p$, a frequência de células com anomalias no grupo dos expostos, X_1 , dado o total Y , tem distribuição Binomial com parâmetro $(y, n_1/n)$, pois

$$\theta = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{n_1 p}{n_1 p + n_2 p} = \frac{n_1 p}{n p} = \frac{n_1}{n} \quad (2.8)$$

Ou seja, sob H_0 , a função de probabilidade condicional de x_1 dado y é dada por

$$l_2(x_1|y) = \frac{y!}{x_1!(y-x_1)!} \left(\frac{n_1}{n}\right)^{x_1} \left(\frac{n_2}{n}\right)^{y-x_1}; \quad x_1 = 0, 1, \dots, y. \quad (2.9)$$

O procedimento, para o teste condicional de H_0 , consiste em encontrar os pontos amostrais mais extremos do que o ponto observado x_1 e calcular para cada um desses pontos sua probabilidade condicional (dado y). A soma dessas probabilidades é o nível de significância descritivo do teste, α_d . A decisão de favorecer ou não H_0 é tomada através da comparação entre este nível e o nível de significância pré-fixado α , ou seja, se

$$\alpha_d = l_2(x_1|y) + \sum_{x \in S_1} l_2(x|y) > \alpha$$

aceita-se H_0 , caso contrário, rejeita-se H_0 , onde $S_1 = \{x: x_1 \text{ favorece a hipótese nula mais do que } x\}$. O nível descritivo é dado por

$$\alpha_d = \sum_{r=x_1}^y l_2(r|y) = \sum_{r=x_1}^y \frac{y!}{r!(y-r)!} \left(\frac{n_1}{n}\right)^r \left(\frac{n_2}{n}\right)^{y-r}.$$

Para o exemplo 2.1.1 tem-se $x_1 = 6$, $y = 7$, $n_1/n = 5000/10000 = 0.5$ e, o único ponto mais extremo que $x_1 = 6$ é o ponto $x = 7$, ou seja, $S_1 = \{7\}$. Portanto, o nível de significância descritivo do teste é $\alpha_d = 0.0625$, como obtido anteriormente. O exemplo apresentado a seguir refere-se a dois grupos com tamanhos de amostras diferentes, $n_1 \neq n_2$. Consequentemente, a distribuição condicional de X_1 , dada a estatística ancilar y , tem o parâmetro p diferente de 0.5. Isto significa que a frequência de células com anomalias, x_1 , não é semelhante à frequência de "caras" no lançamento de uma moeda honesta.

Exemplo 2.2.1 *Amostras de sangue coletadas de alguns indivíduos foram expostas à radiação em laboratório e a frequência de células com micronúcleos (MN) foi observada. Com o objetivo de estudar o efeito da radiação, esta frequência também foi observada em amostras de sangue que não foram expostas à radiação (controles). As frequências observadas são mostradas na tabela 2.3.*

Tabela 2.3: Frequências de células com MN

Indivíduos	Frequências de células		Total
	com MN	sem MN	
Expostos	2	998	1000
Controles	1	4999	5000
Total	3	5997	6000

Analisar se há ou não efeito da radiação, consiste em decidir sobre a veracidade de uma das seguintes hipóteses:

H_0 : a proporção de células com MN nos indivíduos expostos (p_1) é igual à proporção nos indivíduos controles (p_2).

H_1 : a proporção de células com MN nos indivíduos expostos (p_1) é diferente da proporção nos indivíduos controles (p_2).

Os valores observados nas tabelas 2.1 e 2.3 nos levam a desconfiar que haja mais evidências contra a hipótese de igualdade das proporções na tabela 2.3 do que na tabela 2.1, pois a proporção observada 2/1000 é dez vezes maior que 1/5000 (na tabela 2.3), enquanto que a proporção 6/5000 é apenas seis vezes maior que 1/5000 (na tabela 2.1). Contudo, o cálculo do nível descritivo para o exemplo 2.2.1 mostrará que seu valor é um pouco maior do que o valor calculado no exemplo 2.1.1.

Analisando a tabela 2.3, observa-se que o ponto $x = 3$ seria o único ponto que favoreceria menos a hipótese de igualdade das proporções do que o ponto observado $x_1 = 2$. O nível de significância descritivo, baseado na distribuição condicional das frequências, é dado por

$$\frac{3!}{2!1!} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^1 + \frac{3!}{3!0!} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^0 = 0.0694 + 0.0046 = 0.0740$$

Neste exemplo, o nível de significância é um pouco maior do que o obtido no exemplo 2.1.1. Ou seja, a intensidade da evidência contra H_0 é menor para estas observações. Isto contradiz as suspeitas iniciais, baseadas na comparação das frequências obtidas nas duas tabelas (2.1 e 2.3), de que os dados observados na tabela 2.3 seriam mais significantes do que aqueles na tabela 2.1. Mas, esta contradição aparente é consequência do tamanho da amostra,

menor na tabela 2.3, isto é, o teste da tabela 2.3 deve ser menos poderoso que o teste da tabela 2.1.

2.3 Teste χ^2 de homogeneidade

Nesta seção, consideramos casos onde os tamanhos de amostras são bem superiores àqueles da seção anterior. Mesmo estando trabalhando com eventos raros, e assim com frequências amostrais relativas baixas, a frequência de anomalias observada é considerável. Para o teste de igualdade de proporções em situações em que observa-se uma frequência de anomalias relativamente alta, o clássico teste χ^2 de homogeneidade pode ser utilizado. O procedimento para o teste χ^2 consiste no cálculo de uma estatística de teste, que leva em conta os desvios entre as frequências observadas e as frequências esperadas, sob a veracidade da hipótese de igualdade das proporções. Essas frequências esperadas são obtidas com o cálculo da esperança da distribuição condicional da variável aleatória X_1 , dado o total y . Na situação geral, mostrada na tabela 2.2, vimos que esta distribuição condicional sob H_0 é Binomial com parâmetro $(y, n_1/n)$, portanto, sua esperança é dada por

$$E = y \times \frac{n_1}{n}.$$

Deveríamos esperar que a frequência de células anômalas para o grupo dos expostos fosse em torno de E , caso houvesse igualdade de proporções nos dois grupos.

Na seção anterior, calculava-se o nível descritivo usando esta distribuição condicional, e seu valor media a evidência trazida pelos dados a favor da hipótese nula. Nesta seção, como y é um valor razoavelmente grande, podemos calcular uma estatística cuja distribuição aproximada é χ^2 com l graus de liberdade. A estatística para o teste será

$$T = \sum \frac{(O - E)^2}{E} \quad (2.10)$$

onde

E =frequência esperada de células com anomalias;

O =frequência observada de células com anomalias.

O valor de T expressará a evidência a favor da hipótese nula, ou seja, se seu valor for "grande" a evidência é contra H_0 , caso contrário, a evidência é a favor de H_0 . O valor de l é calculado de forma padrão. O exemplo a seguir ilustra o uso do teste χ^2 .

Exemplo 2.3.1 *Pretende-se estudar o efeito de um composto no índice mitótico de células da medula óssea. Para isso, observou-se índices mitóticos de células da medula óssea de camundongos, divididos em dois grupos distintos, tratados (expostos ao composto) e controles (não expostos ao composto). Cada grupo foi dividido em 3 subgrupos. No primeiro, os camundongos foram sacrificados após 24 horas de exposição; no segundo, após 48 horas e no terceiro, após 72 horas. As frequências observadas de células em metáfase são mostradas na tabela 2.4.*

Inicialmente a comparação das proporções será feita conjuntamente, ou seja, verificando se as proporções são iguais entre as caselas. A tabela 2.5

Tabela 2.4: Estudo do índice mitótico

Grupos	Exposição	Número de animais	Frequências de células	
			Em metáfase	Total
Tratado	24 h	6	358	18000
Controle	24 h	5	424	15000
Tratado	48 h	6	256	18000
Controle	48 h	5	378	15000
Tratado	72 h	6	342	18000
Controle	72 h	4	257	12000

mostra as frequências observadas e esperadas para cada casela. A tabela 2.6 mostra os desvios entre observados e esperados e o valor da estatística T . Observa-se que o valor da estatística, $T = 94.32$, é bastante alto quando comparado com o valor tabelado de uma variável χ^2 com 6 graus de liberdade. Logo, não podemos aceitar a hipótese de igualdade entre todas as proporções.

Tabela 2.5: Frequências observadas (esperadas) de células em metáfase

Grupos	Exposição			Total
	24h	48h	72h	
Tratado	358 (378)	256 (378)	342 (378)	956
Controle	424 (315)	378 (315)	257 (251)	1059
Total	782	634	599	2015

Com o objetivo de descobrir onde há diferenças entre as proporções, deve-se analisar estes dados em duas etapas: a primeira etapa consistirá da com-

Tabela 2.6: Desvios $\frac{(o-e)^2}{e}$

Grupos	Exposição			Total
	24h	48h	72h	
Tratado	1.06	39.38	3.43	43.87
Controle	37.72	12.60	0.14	50.46
Total	38.78	51.98	3.57	94.33

paração entre tratados e controles dentro de cada exposição, e a segunda da comparação das diferentes exposições dentro de cada grupo. Para a comparação dos dois grupos em cada exposição, inicialmente fixaremos um determinado tempo de exposição, por exemplo 24 horas. Observe que novamente temos observações associadas a eventos raros, pois, $x_1 = 358$ e $x_2 = 424$ são valores pequenos quando comparados com $n_1 = 18000$ e $n_2 = 15000$. A diferença com relação à situação descrita na seção anterior é o total y , que agora é um valor razoavelmente grande, $y = x_1 + x_2 = 782$. Como foi visto anteriormente, se não existe efeito do composto no índice mitótico das células, a distribuição condicional de X_1 , dado y , é Binomial com parâmetro $(782, 18/33)$. A tabela 2.7 mostra o valor obtido da estatística T para a exposição de 24 h.

Observamos que o valor da estatística, $T = 24.56$, é alto, expressando uma forte evidência a favor da existência de efeito do composto na exposição de 24 h. De maneira análoga, obtém-se o valor de T para a exposição de 48 h, os resultados são mostrados na tabela 2.8. Note que a distribuição condicional para esta exposição é Binomial com parâmetro $(634, 18/33)$.

Tabela 2.7: T para a exposição de 24 h

Grupos	Frequências		$\frac{(o-e)^2}{e}$
	observadas	esperadas	
tratados	358	427	11.15
controles	424	355	13.41
total	782	782	24.56

Tabela 2.8: T para a exposição de 48 h

Grupos	Frequências		$\frac{(o-e)^2}{e}$
	observadas	esperadas	
tratados	256	346	23.41
controles	378	288	28.12
total	634	634	51.53

Tabela 2.9: T para a exposição de 72 h

Grupos	Frequências		$\frac{(o-e)^2}{e}$
	observadas	esperadas	
tratados	342	359	0.81
controles	257	240	1.20
total	599	599	2.01

O valor obtido da estatística, $T = 51.53$, também expressa uma forte evidência a favor da existência de efeito do composto para a exposição de 48 h. Para a exposição de 72 h, a distribuição condicional é Binomial com parâmetro $(599, 18/30)$. Os valores observados e esperados, juntamente com o respectivo valor da estatística T , são mostrados na tabela 2.9.

Neste caso, o valor obtido da estatística T é baixo, e expressa uma forte evidência contra a existência de efeito do composto quando os camundongos são sacrificados após 72 h de exposição.

Nesta primeira etapa da análise, concluímos que há evidências da existência de diferenças entre tratados e controles dentro de cada nível de exposição, portanto, é necessário estudar o efeito do composto sobre o índice mitótico segundo os diferentes tempos de exposição. A segunda etapa da análise compreende a comparação das diferentes exposições em cada grupo. Como o fator de comparação possui três níveis (24 h, 48 h, 72 h), a distribuição condicional do número de células em metáfase para um determinado grupo, dado o total, será uma Trinomial com parâmetro $(t, n_1/n, n_2/n, n_3/n)$, onde t é o total de células em metáfase observado neste grupo, n_1, n_2 e n_3 são, respectivamente, os totais de células observados neste grupo para as exposições de

24 h, 48 h e 72 h, e $n = n_1 + n_2 + n_3$. No grupo dos tratados temos $t = 956$, $n_1/n = n_2/n = n_3/n = 18000/54000 = 1/3$ e no grupo controle $t = 1059$, $n_1/n = n_2/n = 15000/42000 = 5/14$ e $n_3/n = 12000/42000 = 4/14$. Neste caso, a estatística do teste tem, aproximadamente, distribuição χ^2 com dois graus de liberdade. Os valores da estatística T para os grupos tratados e controles são mostrados, respectivamente, nas tabelas 2.10 e 2.11.

Tabela 2.10: T para o grupo dos tratados

Exposição	Frequências		$\frac{(o-e)^2}{e}$
	observadas	esperadas	
24 h	358	318	5.03
48 h	256	319	12.44
72 h	342	319	1.66
Total	956	956	19.13

Tabela 2.11: T para o grupo dos controles

Exposição	Frequências		$\frac{(o-e)^2}{e}$
	observadas	esperadas	
24 h	424	378	5.60
48 h	378	378	0.00
72 h	257	303	6.98
Total	1059	1059	12.58

Nos dois grupos o valor observado de T é grande, isto nos dá evidências contra a hipótese de igualdade das proporções de células em metáfase nos

diferentes tempos de exposição. Ou seja, em ambos os grupos a proporção de células em metáfase é diferente para cada tempo de exposição.

2.4 Teste Bayesiano

2.4.1 Informação *a priori*

Experimentos são realizados com o objetivo de obter-se informações a respeito de um estado da natureza, representado por uma quantidade desconhecida θ , e a finalidade da Inferência Estatística é extrair toda informação a respeito de θ contida nos dados experimentais x . O interesse pelo parâmetro θ é despertado, em geral, por algum conhecimento precedente sobre o mesmo. Este conhecimento precedente pode ser a informação de que θ existe e assume valores em um conjunto bem definido, isto já é suficiente para se atribuir uma distribuição de probabilidades para θ . Quando o verdadeiro valor de θ é conhecido a informação é total, e a consulta a um estatístico é desnecessária. Se θ é desconhecido, o grande interesse é a redução desse desconhecimento. A intensidade da incerteza sobre o valor de θ pode assumir diferentes graus e o segmento Bayesiano da estatística representa estes diferentes graus de incerteza por modelos probabilísticos distintos. Considerar diferentes graus de incerteza significa admitir que esta incerteza pode variar de indivíduo para indivíduo, ou seja, o Seu conhecimento inicial sobre θ depende apenas das informações que Você dispõe inicialmente. Observe que esta é uma visão subjetivista de probabilidade, que considera a probabilidade de um determinado evento como a medida de incerteza sobre a ocorrência ou não do mesmo.

As técnicas Bayesianas para solução de problemas estatísticos utilizam a informação contida nos dados para calibrar nosso conhecimento inicial sobre o fenômeno em estudo. Este conhecimento inicial é representado pela distribuição de probabilidades *a priori* e o novo conhecimento (após a calibração) pela distribuição de probabilidades *a posteriori*. O processo de calibração, chamado de operação Bayesiana de passagem da *priori* para a *posteriori*, ocorre através da utilização da fórmula de Bayes. Estas técnicas estendem-se para o caso do teste das hipóteses H_0 e H_1 , onde admite-se conhecimentos precedentes sobre o fenômeno em estudo atribuindo-se chances *a priori* para as duas hipóteses. Após a realização do experimento, calibra-se esta informação precedente e obtém-se novas chances para as hipóteses a serem testadas. Este processo de calibração só é possível através da função de verossimilhança, que representa a relação entre o experimento realizado e a quantidade em estudo presente nas hipóteses. O critério de decisão do teste é baseado no Escore a favor de H_0 *a posteriori*, que é obtido multiplicando-se o Escore a favor de H_0 *a priori* pela razão das Funções Preditivas sob H_0 e H_1 . O Escore *a priori* é a razão entre as chances iniciais atribuídas às hipóteses e as Funções Preditivas são médias ponderadas da função de verossimilhança condicionadas às hipóteses.

2.4.2 Procedimento do Teste

Voltando ao caso geral descrito na tabela 2.2, tem-se as seguintes hipóteses de interesse

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2$$

Considere o seguinte parâmetro

$$\delta = \begin{cases} 0 & \text{se } (p_1, p_2) \in \Omega_0 \\ 1 & \text{se } (p_1, p_2) \in \Omega_1 \end{cases}$$

onde Ω_0 e Ω_1 são os respectivos espaços paramétricos associados às hipóteses H_0 e H_1 :

$$\Omega_0 = \{(p_1, p_2); \quad 0 \leq p_1 \leq 1, p_1 = p_2\}$$

$$\Omega_1 = \{(p_1, p_2); \quad 0 \leq p_1 \leq 1, 0 \leq p_2 \leq 1, p_1 \neq p_2\}$$

O primeiro passo é representar a incerteza inicial sobre o par (p_1, p_2) por uma função densidade conjunta, $\pi(p_1, p_2)$, em seguida atribui-se chances iniciais aos valores de δ , ou seja,

$$P[\delta = 0] = \eta$$

$$P[\delta = 1] = 1 - \eta.$$

Se existe um desconhecimento total, ou seja, não há preferências com relação às hipóteses, os valores de η e $1 - \eta$ devem ser iguais, $\eta = 1 - \eta = 1/2$. Caso seja necessário decidir-se sobre uma das hipóteses antes da realização do experimento, deve-se analisar a razão entre estas chances, $\varepsilon = \eta/(1 - \eta)$, denominada *Escore a Favor de H_0 a priori*. A decisão será em favor da hipótese cuja chance for maior. Após a realização do experimento, calibra-se este escore inicial utilizando-se os dados experimentais (x_1, x_2) . Este novo escore é dado por

$$\varepsilon(x_1, x_2) = \frac{P[\delta = 0 \mid x_1, x_2]}{P[\delta = 1 \mid x_1, x_2]}$$

e denomina-se *Escore a Favor de H_0 a posteriori*.

Observe que a relação entre δ e os dados (x_1, x_2) se dá através de (p_1, p_2) e da função de verossimilhança $L(x_1, x_2 | p_1, p_2)$. Desta forma, o cálculo da distribuição condicional conjunta $\delta, p_1, p_2 | (x_1, x_2)$ é mais fácil que o cálculo da distribuição $\delta | (x_1, x_2)$. Mas, esta última distribuição pode ser obtida através da integração da primeira nos parâmetros p_1 e p_2 . Ou seja, a partir das probabilidades $P[\delta = 0, p_1, p_2 | x_1, x_2]$ e $P[\delta = 1, p_1, p_2 | x_1, x_2]$ obtém-se, respectivamente, as probabilidades $P[\delta = 0 | x_1, x_2]$ e $P[\delta = 1 | x_1, x_2]$. Com a utilização da fórmula de Bayes, temos que

$$\begin{aligned} P[\delta = 0, p_1, p_2 | x_1, x_2] &= \frac{L(\delta = 0, x_1, x_2 | p_1, p_2)\pi(p_1, p_2 | \delta = 0)P[\delta = 0]}{f(x_1, x_2)} \\ &= \frac{\eta L(\delta = 0, x_1, x_2 | p_1, p_2)\pi(p_1, p_2 | \delta = 0)}{f(x_1, x_2)} \end{aligned}$$

logo,

$$\begin{aligned} P[\delta = 0 | x_1, x_2] &= \int_A P[\delta = 0, p_1, p_2 | x_1, x_2] dp_1 dp_2 \\ &= \frac{\eta}{f(x_1, x_2)} \int_A L(\delta = 0, x_1, x_2 | p_1, p_2)\pi(p_1, p_2 | \delta = 0) dp_1 dp_2 \end{aligned}$$

onde A é o conjunto dos possíveis valores do par (p_1, p_2) .

Note que estamos trabalhando com os valores de (p_1, p_2) pertencentes ao espaço paramétrico Ω_0 , pois as funções envolvidas estão condicionadas à hipótese nula, ou seja, ao valor $\delta = 0$. A função $L(\delta = 0, x_1, x_2 | p_1, p_2)$ é a função de verossimilhança, $L(x_1, x_2 | p_1, p_2)$, definida no conjunto Ω_0 . A função $\pi(p_1, p_2 | \delta = 0)$ é a função densidade conjunta para (p_1, p_2) definida no conjunto Ω_0 , que precisa ser normalizada para continuar sendo uma função

densidade de probabilidade. Isto é,

$$\pi(p_1, p_2 \mid \delta = 0) = \begin{cases} \frac{\pi(p_1, p_2)}{\int_A \pi(p_1, p_2) dp_1 dp_2} & \text{se } (p_1, p_2) \in \Omega_0 \\ 0 & \text{caso contrário} \end{cases}$$

A integral $\int_A L(\delta = 0, x_1, x_2 \mid p_1, p_2) \pi(p_1, p_2 \mid \delta = 0) dp_1 dp_2$ é uma média ponderada da verossimilhança condicionada à hipótese nula. Este condicionamento se dá pelo fato do par (p_1, p_2) pertencer à Ω_0 e a ponderação é feita pela distribuição conjunta *a priori* de (p_1, p_2) . Esta média ponderada denomina-se *Função Preditiva sob H_0* e é dada, através de Integrais de Contorno (ver Pereira, 1985), por

$$h_0(x_1, x_2) = h(x_1, x_2 \mid H_0) = \frac{\int_{\Omega_0} \pi(p_1, p_2) L(x_1, x_2 \mid p_1, p_2) ds}{\int_{\Omega_0} \pi(p_1, p_2) ds} \quad (2.11)$$

onde s é o comprimento do arco definido pela função identidade, $p_1 = p_2$, e o conjunto Ω_0 , ou seja, é o comprimento da reta $p_1 = p_2$ definida no quadrado unitário (ver figura 2.1).

Assim, podemos escrever

$$P[\delta = 0 \mid x_1, x_2] = \frac{\eta}{f(x_1, x_2)} h_0(x_1, x_2) \quad (2.12)$$

Para o cálculo de $h_0(x_1, x_2)$ são necessárias a verossimilhança, expressa em (2.3) e a função densidade conjunta *a priori*, $\pi(p_1, p_2)$. Para facilitar o desenvolvimento analítico, o estatístico Bayesiano expressa sua opinião inicial sobre (p_1, p_2) através da família de Distribuições Conjugadas *a priori* em relação à verossimilhança considerada. Uma família de ditribuições de probabilidade é chamada Conjugada *a priori* em relação à verossimilhança

Figura 2.1: Quadrado unitário

L se o fato da distribuição *a priori* π pertencer a esta família implicar a permanência da distribuição *a posteriori* nesta família. A verossimilhança considerada é a distribuição Binomial e sua família de distribuições conjugadas *a priori* é a família de distribuições Beta. Ou seja, assumindo que p_1 e p_2 são independentes e possuem distribuição *a priori* $Beta(a_1, b_1)$ e $Beta(a_2, b_2)$ respectivamente, temos

$$\pi(p_1, p_2) = \pi_1(p_1)\pi_2(p_2) \quad (2.13)$$

onde

$$\pi_i(p_i) = \frac{1}{B(a_i, b_i)} p_i^{a_i-1} (1 - p_i)^{b_i-1} \quad , i = 1, 2 \quad (2.14)$$

$$B(a_i, b_i) = \Gamma(a_i)\Gamma(b_i)/\Gamma(a_i + b_i)$$

e $\Gamma(\cdot)$ representa a função Gamma.

Consideraremos um caso particular da Beta, $a_i = b_i = 1$, que é a distribuição Uniforme $[0,1]$, pois ela representa falta total de informação sobre o par (p_1, p_2) . Neste caso, temos

$$\pi(p_1, p_2) = \begin{cases} 1 & , 0 \leq p_1, p_2 \leq 1 \\ 0 & , \text{caso contrário} \end{cases}$$

Agora, podemos proceder o cálculo de $h_0(x_1, x_2)$. O numerador é dado por

$$\begin{aligned} \int_{\Omega_0} \pi(p_1, p_2) L(x_1, x_2 | p_1, p_2) ds &= \int_0^1 \pi(p_1, p_1) L(x_1, x_2 | p_1, p_1) \sqrt{2} dp_1 = \\ &= \int_0^1 \frac{n_1!}{x_1!(n_1 - x_1)!} \frac{n_2!}{x_2!(n_2 - x_2)!} p_1^{x_1+x_2} (1 - p_1)^{n_1+n_2-(x_1+x_2)} \sqrt{2} dp_1 = \\ &= \sqrt{2} \frac{n_1!}{x_1!(n_1 - x_1)!} \frac{n_2!}{x_2!(n_2 - x_2)!} \int_0^1 p_1^y (1 - p_1)^{n-y} dp_1 = \\ &= \frac{\sqrt{2} n_1! n_2!}{x_1! x_2! (n_1 - x_1)! (n_2 - x_2)!} B(y+1, n-y+1) \end{aligned}$$

e o denominador por

$$\int_{\Omega_0} \pi(p_1, p_2) ds = \int_0^1 \pi(p_1, p_1) \sqrt{2} dp_1 = \sqrt{2}$$

onde $y = x_1 + x_2$ e $n = n_1 + n_2$.

Portanto,

$$h_0(x_1, x_2) = \frac{n_1!}{x_1!(n_1 - x_1)!} \frac{n_2!}{x_2!(n_2 - x_2)!} B(y+1, n-y+1)$$

Ou seja,

$$h_0(x_1, x_2) = \frac{n_1!}{x_1!(n_1 - x_1)!} \frac{n_2!}{x_2!(n_2 - x_2)!} \frac{y!(n - y)!}{(n + 1)!} \quad (2.15)$$

pois $\Gamma(p + 1) = p!$ se p é um número inteiro. Analogamente, obtém-se $P[\delta = 1 \mid x_1, x_2]$:

$$\begin{aligned} P[\delta = 1, p_1, p_2 \mid x_1, x_2] &= \frac{L(\delta = 1, x_1, x_2 \mid p_1, p_2) \pi(p_1, p_2 \mid \delta = 1) P[\delta = 1]}{f(x_1, x_2)} = \\ &= \frac{(1 - \eta) L(\delta = 1, x_1, x_2 \mid p_1, p_2) \pi(p_1, p_2 \mid \delta = 1)}{f(x_1, x_2)} \end{aligned}$$

logo,

$$P[\delta = 1 \mid x_1, x_2] = \frac{(1 - \eta)}{f(x_1, x_2)} \int_A L(\delta = 1, x_1, x_2 \mid p_1, p_2) \pi(p_1, p_2 \mid \delta = 1) dp_1 dp_2$$

Neste caso, tem-se o par (p_1, p_2) pertencente ao espaço paramétrico Ω_1 . De forma semelhante define-se a *Função Preditiva sob H_1* :

$$h_1(x_1, x_2) = h(x_1, x_2 \mid H_1) = \frac{\int_{\Omega_1} \pi(p_1, p_2) L(x_1, x_2 \mid p_1, p_2) dp_1 dp_2}{\int_{\Omega_1} \pi(p_1, p_2) dp_1 dp_2} \quad (2.16)$$

Assim, podemos escrever

$$P[\delta = 1 \mid x_1, x_2] = \frac{(1 - \eta)}{f(x_1, x_2)} h_1(x_1, x_2) \quad (2.17)$$

O numerador de $h_1(x_1, x_2)$ é dado por

$$\int_{\Omega_1} \pi(p_1, p_2) L(x_1, x_2 \mid p_1, p_2) dp_1 dp_2 =$$

$$= \int_0^1 \int_0^1 \frac{n_1!}{x_1!(n_1 - x_1)!} \frac{n_2!}{x_2!(n_2 - x_2)!} p_1^{x_1} (1 - p_1)^{n_1 - x_1} \times \\ p_2^{x_2} (1 - p_2)^{n_2 - x_2} dp_2 dp_1 = \frac{1}{(n_1 + 1)(n_2 + 1)}$$

Para o denominador de $h_1(x_1, x_2)$, temos

$$\int_{\Omega_1} \pi(p_1, p_2) dp_1 dp_2 = \int_0^1 \int_0^1 dp_2 dp_1 = 1$$

Portanto, a função preditiva sob H_1 é dada por

$$h_1(x_1, x_2) = \frac{1}{(n_1 + 1)(n_2 + 1)}$$

Usando as expressões (2.12) e (2.17) obtemos a seguinte relação

$$\varepsilon(x_1, x_2) = \frac{\eta}{1 - \eta} \frac{h_0(x_1, x_2)}{h_1(x_1, x_2)} = \varepsilon \times \frac{h_0(x_1, x_2)}{h_1(x_1, x_2)}$$

Substituindo os valores de $h_0(x_1, x_2)$ e $h_1(x_1, x_2)$, vem

$$\varepsilon(x_1, x_2) = \varepsilon \times \frac{(n_1 + 1)!(n_2 + 1)!y!(n - y)!}{x_1!(n_1 - x_1)!x_2!(n_2 - x_2)!(n + 1)!} \quad (2.18)$$

O critério de decisão do teste é baseado neste escore *a posteriori*:

- Se $\varepsilon(x_1, x_2) < k$, rejeita-se H_0 ;
- Se $\varepsilon(x_1, x_2) \geq k$, não rejeita-se H_0 .

Assim, dada uma constante k , obtém-se o conjunto dos pares (x_1, x_2) que levam à rejeição de H_0 . Este conjunto é a *Região Crítica do teste*, e é dado por

$$R.C. = \{(x_1, x_2) : \varepsilon(x_1, x_2) < k\}$$

Esta constante k indica quantas vezes H_0 deverá ser mais provável que H_1 , após a realização do experimento, para que a decisão seja favorável à H_0 . A

determinação do valor de k envolve perdas associadas à tomada incorreta de decisões, ou seja, perdas que podem ser sofridas caso um dos seguintes erros seja cometido: rejeitar H_0 quando H_0 é verdadeira ou aceitar H_0 quando H_0 é falsa.

Consideraremos $k = 1$, isto é, basta H_0 ser mais provável que H_1 uma única vez para que a decisão seja a seu favor. Isto significa que os dois tipos de erro terão a mesma importância, no sentido de que as perdas sofridas serão as mesmas qualquer que seja o erro cometido.

O escore *a posteriori*, $\varepsilon(x_1, x_2)$, para os dados do exemplo 2.1.1, é dado por

$$\varepsilon(x_1, x_2) = \varepsilon \times \frac{5001!5001!7!9993!}{6!4994!1!4999!10001!}$$

Suponha que as chances iniciais dadas às hipóteses H_0 e H_1 são as mesmas, isto é, $\varepsilon = 1$. Logo,

$$\varepsilon(x_1, x_2) = 136.64$$

Como $\varepsilon(x_1, x_2) > 1$, não rejeitamos a hipótese de igualdade das proporções nos dois grupos. Para os dados do exemplo 2.2.1, também considerando $\varepsilon = 1$, temos

$$\varepsilon(x_1, x_2) = \frac{1001!5001!3!5997!}{2!998!1!4999!6001!} = 57.90$$

Logo, não rejeitamos a hipótese de igualdade das proporções no exemplo 2.2.1.

Capítulo 3

DOSIMETRIA BAYESIANA

3.1 Considerações iniciais

No capítulo anterior consideramos que os indivíduos sofriam exposição a um único nível de dose. Neste capítulo, consideraremos que ocorre exposição a diferentes níveis de doses de um determinado produto, e nosso principal interesse é estabelecer um modelo que explique adequadamente as modificações causadas pela mudança de nível na dose, ou seja, obter uma relação funcional entre dose e resposta. Com o estabelecimento desta relação funcional pode-se descobrir a dose de exposição cada vez que a resposta do indivíduo é medida. Em nossas análises iremos considerar a radiação como o produto suspeito, e a resposta será a frequência de células com aberrações citogenéticas.

A distribuição que vem sendo usada na literatura para análise da frequência de células com anomalias é a distribuição de Poisson. Esta distribuição, quando associada à contagem de eventos, só pode ser usada quando consi-

deramos eventos raros, uma vez que ela é uma aproximação da distribuição Binomial. O problema com esta aproximação é que, em doses muito altas, não podemos mais considerar o evento como sendo raro, visto que é muito provável a ocorrência de anomalias nas células. Portanto, em doses altas, a distribuição de Poisson não deve ser usada, pois se afastaria muito da distribuição Binomial. A relação funcional geralmente indicada na literatura é um modelo linear - simples (Fenech & Morley, 1985) ou quadrático (Prosser et al., 1988). Contudo, alguns autores (Kormos & Köteles, 1988; Balasem & Ali, 1991) recomendam estes modelos apenas para doses baixas, quando usamos a distribuição de Poisson. A utilização da distribuição Binomial acarreta a imposição de limites bem determinados ao número de células defeituosas, estes limites tendo probabilidades de ocorrência consideráveis. Assim, um modelo que exploda e faça com que os valores estimados ultrapassem esses limites estabelecidos não pode ser considerado para explicar a relação entre a frequência de células defeituosas e os níveis de dose. Consequentemente, os modelos lineares devem ser eliminados do conjunto dos possíveis modelos considerados. O próprio fenômeno em estudo, ocorrência de anomalias em células expostas à radiação, exige limites na frequência de células anômalas, pois a partir de um determinado nível de dose há uma pulverização da célula e não se consegue mais observar as anomalias. Uma possível solução para resolver este problema seria considerar transformações adequadas das frequências, de forma que os valores possíveis dessas transformações não sejam limitados.

Como trabalharemos com níveis de dose baixos e altos, propomos um modelo que utiliza uma aproximação Normal para uma transformação da frequência relativa populacional. Logo, não há restrições de domínio ao mo-

delo proposto. O ajuste deste modelo será feito utilizando-se técnicas Bayesianas.

3.2 Caso Binomial

Suponha que estamos interessados em estabelecer um modelo para as proporções de células anômalas de indivíduos expostos a k diferentes doses de algum tipo de radiação. Admita que amostras de n_i células são expostas a níveis de doses D_i , $i = 1, \dots, k$, e observa-se as frequências de células com anomalias em cada nível, representando-as pelas variáveis aleatórias Y_i 's. A distribuição de Y_i é Binomial com parâmetro (n_i, p_i) , onde p_i é a proporção populacional de células com anomalias no i -ésimo nível de dose. Observe que as Y_i 's, $i = 1, \dots, k$, são variáveis aleatórias independentes entre si, pois a frequência de células com anomalias observada em um determinado nível de dose independe da frequência observada em outro nível. Portanto, a função de verossimilhança é

$$\begin{aligned} L(y_1, \dots, y_k | p_1, \dots, p_k) &= P[Y_1 = y_1, \dots, Y_k = y_k | p_1, \dots, p_k] \\ &= \prod_{i=1}^k \frac{n_i!}{y_i!(n_i - y_i)!} p_i^{y_i} (1 - p_i)^{n_i - y_i} \end{aligned}$$

Utilizando a respectiva família conjugada, temos que a distribuição *a priori* para p_i é uma distribuição Beta com parâmetros a_i e b_i . Um raciocínio análogo ao anteriormente feito com os Y_i 's nos leva a concluir que os p_i 's também são variáveis aleatórias independentes. Logo, sua função densidade de probabilidade conjunta *a priori* é dada por

$$\pi(p_1, \dots, p_k) = \prod_{i=1}^k \frac{1}{B(a_i, b_i)} p_i^{a_i-1} (1-p_i)^{b_i-1} \quad (3.1)$$

Utilizando o Teorema de Bayes, tem-se cada p_i com uma distribuição *a posteriori* Beta com parâmetros A_i e B_i , onde $A_i = a_i + y_i$ e $B_i = b_i + n_i - y_i$. Para obter uma estimação pontual de p_i , consideramos a função perda quadrática. Logo, o estimador de Bayes de p_i , a média posterior, é dado por

$$\hat{p}_i = \frac{A_i}{A_i + B_i} \quad (3.2)$$

e a variância posterior é

$$\text{Var}(p_i | y_i) = \frac{A_i B_i}{(A_i + B_i + 1)(A_i + B_i)^2} \quad (3.3)$$

Como sugerido anteriormente, pode-se trabalhar com uma transformação adequada das frequências. A literatura estatística indica uma transformação bem conhecida, denominada pelo nome de *log-odds*, e expressa por

$$\theta_i = \ln \frac{p_i}{1 - p_i} \quad i = 1, \dots, k \quad (3.4)$$

A transformação inversa é conhecida por transformação logística, e é dada por

$$p_i = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} \quad i = 1, \dots, k \quad (3.5)$$

Observe que os valores possíveis desta transformação podem assumir qualquer valor na reta, tanto positivo quanto negativo. A grande vantagem desta transformação é que θ_i tem uma distribuição aproximadamente Normal.

Isto é proveniente da relação existente entre as distribuições Beta e Fisher-Snedecor, que será enunciada nos dois resultados a seguir.

Lema 3.2.1 *Seja P uma variável aleatória com distribuição $Beta(a, b)$. A variável aleatória $Y = \frac{bP}{a(1-P)}$ tem distribuição Fisher-Snedecor com $2a$ e $2b$ graus de liberdade.*

Demonstração:

Sejam X_1 e X_2 variáveis aleatórias independentes, cada uma tendo distribuição Gama com os seguintes parâmetros: $X_1 \sim Gama(a; \lambda)$ e $X_2 \sim Gama(b; \lambda)$.

Portanto,

$$P = \frac{X_1}{X_1 + X_2} \sim Beta(a, b)$$

$$1 - P = \frac{X_2}{X_1 + X_2} \sim Beta(b, a)$$

Observe que as distribuições de P e $1 - P$ não dependem de λ . Logo, a razão

$$\frac{P}{1 - P} = \frac{X_1}{X_2}$$

também não depende de λ , ou seja, é válida para qualquer valor de λ . Em particular, é válida para $\lambda = 1/2$. Neste caso, X_1 e X_2 têm distribuições Qui-quadrado com $2a$ e $2b$ graus de liberdade, respectivamente. Podemos escrever esta razão da seguinte maneira

$$\frac{P}{1 - P} = \frac{X_1}{X_2} = \frac{2a \frac{X_1}{2a}}{2b \frac{X_2}{2b}} = \frac{2a}{2b} Y$$

com $Y = \frac{X_1/2a}{X_2/2b}$.

Mas, razão de variáveis aleatórias independentes com distribuições Qui-quadrado, divididas por seus respectivos graus de liberdade, tem distribuição Fisher-Snedecor. Logo, a variável aleatória

$$Y = \frac{b}{a} \frac{P}{1 - P}$$

tem distribuição Fisher-Snedecor com $2a$ e $2b$ graus de liberdade.

□

Teorema 3.2.1 *Seja Y uma variável aleatória com distribuição de Fisher com v_1 e v_2 graus de liberdade. Então, para v_1 e v_2 grandes, a variável aleatória $Z = \ln Y$ tem distribuição aproximadamente Normal.*

Demonstração:

A função densidade de probabilidade de Y é dada por

$$f(y) = \frac{1}{B(v_1/2, v_2/2)} \frac{v_1^{v_1/2} v_2^{v_2/2} y^{(v_1-2)/2}}{(v_1 y + v_2)^{(v_1+v_2)/2}}$$

A variável aleatória $Z = \ln Y$ tem a seguinte função característica

$$\Phi_Z(t) = E[e^{itZ}] = E[e^{it \ln Y}] = E[Y^{it}],$$

ou seja, é o momento de ordem r da variável Y , $r = it$. Logo,

$$\Phi_Z(r) = \int_0^\infty \frac{1}{B(v_1/2, v_2/2)} \frac{v_1^{v_1/2} v_2^{v_2/2} y^{(v_1-2)/2}}{(v_1 y + v_2)^{(v_1+v_2)/2}} y^r dy$$

$$= \left(\frac{v_2}{v_1}\right)^r \frac{\Gamma(v_1/2 + r)\Gamma(v_2/2 - r)}{\Gamma(v_1/2)\Gamma(v_2/2)}$$

A função geradora de cumulantes de Z é dada por

$$\ln \Phi_Z(r) = r \ln\left(\frac{v_2}{v_1}\right) + \ln \Gamma\left(\frac{v_1}{2} + r\right) + \ln \Gamma\left(\frac{v_2}{2} - r\right) - \ln \Gamma(v_1/2) - \ln \Gamma(v_2/2)$$

Através da expansão de Stirling para $\ln \Gamma(x)$ (ver Kendall, Stuart & Ord, 1987), obtém-se

$$\ln \Phi_Z(r) = r\left(\frac{1}{v_2} - \frac{1}{v_1}\right) + \frac{r^2}{2}\left(\frac{1}{v_1} + \frac{1}{v_2}\right) + o\left(\frac{1}{v}\right) \dots$$

ou, em termos de t ,

$$\ln \Phi_Z(t) = it\left(\frac{1}{v_2} - \frac{1}{v_1}\right) - \frac{t^2}{2}\left(\frac{1}{v_1} + \frac{1}{v_2}\right) + o\left(\frac{1}{v}\right) \dots$$

onde v é uma função de v_1 e v_2 . Portanto, para v_1 e v_2 grandes, Z terá a função geradora de cumulantes de uma variável aleatória com distribuição Normal.

□

Usando os dois resultados anteriores concluímos que $\theta_i = \ln \frac{P_i}{1-P_i}$ tem distribuição aproximadamente Normal quando A_i e B_i são grandes. Como foi discutido no capítulo anterior, coleta-se uma amostra grande de células a fim de obter-se uma frequência considerável de células anômalas, consequentemente, os valores de A_i e B_i são grandes ($i = 1, \dots, k$). As expressões exatas para a esperança e a variância de θ_i são enunciadas no próximo resultado.

Lema 3.2.2 *Seja P uma variável aleatória com distribuição Beta(a, b). A esperança e variância da variável aleatória $\theta = \ln \frac{P}{1-P}$ são dadas por*

$$E[\theta] = \psi(a) - \psi(b)$$

$$Var[\theta] = \psi'(a) + \psi'(b)$$

onde $\psi(x)$ é a função Digamma e $\psi'(x)$ é a função Trigamma.

Demonstração:

Como $P \sim Beta(a, b)$, vale a seguinte identidade

$$\int_0^1 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} dp = 1 \quad (3.6)$$

i) Derivando com relação a a ambos os membros de (3.6), e trocando a ordem dos operadores integração e diferenciação, obtemos

$$\frac{\partial \ln \Gamma(a+b)}{\partial a} + E[\ln P] - \frac{\partial \ln \Gamma(a)}{\partial a} = 0$$

ou, seja,

$$E[\ln P] = \psi(a) - \psi(a+b) \quad (3.7)$$

ii) Derivando com relação a b ambos os membros de (3.6), e trocando a ordem dos operadores integração e diferenciação, obtemos

$$\frac{\partial \ln \Gamma(a+b)}{\partial b} + E[\ln(1-P)] - \frac{\partial \ln \Gamma(b)}{\partial b} = 0$$

ou, seja,

$$E[\ln(1-P)] = \psi(b) - \psi(a+b) \quad (3.8)$$

Assim, de (3.7) e (3.8), temos

$$E[\theta] = E\left[\ln\left(\frac{P}{1-P}\right)\right] = E[\ln P - \ln(1-P)] = \psi(a) - \psi(b)$$

De (3.7), temos que

$$\int_0^1 \ln p \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1} dp = \psi(a) - \psi(a+b) \quad (3.9)$$

iii) Derivando com relação a a ambos os membros de (3.9), e trocando a ordem dos operadores integração e diferenciação, obtemos

$$E[\ln^2 P] = \psi'(a) - \psi'(a+b) + E^2[\ln P]$$

Fazendo operações equivalentes em (3.8), temos que

$$E[\ln^2(1-P)] = \psi'(b) - \psi'(a+b) + E^2[\ln(1-P)]$$

Portanto,

$$Var[\ln P] = E[\ln^2 P] - E^2[\ln P] = \psi'(a) - \psi'(a+b)$$

$$Var[\ln(1-P)] = E[\ln^2(1-P)] - E^2[\ln(1-P)] = \psi'(b) - \psi'(a+b)$$

iv) Derivando duas vezes com relação a a e a b ambos os membros de (3.6), e trocando a ordem dos operadores integração e diferenciação, obtemos

$$\begin{aligned} E[\ln P \ln(1-P)] &= \psi(a+b) [\psi(a+b) - \psi(a) - \psi(b)] + \\ &\quad \psi(a)\psi(b) - \psi'(a+b) \end{aligned}$$

Logo,

$$\begin{aligned}
 Var[\theta] &= Var\left[\ln \frac{P}{1-P}\right] = Var[\ln P - \ln(1-P)] \\
 &= Var[\ln P] + Var[\ln(1-P)] - 2Cov[\ln P, \ln(1-P)] \\
 &= \psi'(a) + \psi'(b)
 \end{aligned}$$

□

Usando estes resultados, concluímos que

$$\mu_i = E[\theta_i] = \psi(A_i) - \psi(B_i) \quad (3.10)$$

$$\sigma_i^2 = Var[\theta_i] = \psi'(A_i) + \psi'(B_i) \quad (3.11)$$

Portanto, θ_i tem distribuição aproximadamente Normal com média μ_i e variância σ_i^2 , $\forall i = 1, \dots, k$. Outra maneira de mostrar a aproximação Normal para a distribuição de θ_i é esboçando o gráfico de sua função densidade de probabilidade e comparando-o ao gráfico da função densidade de uma Normal com mesma média e variância. As figuras 3.1, 3.2, 3.3, 3.4, e 3.5 representam estes gráficos para diferentes valores de A_i e B_i . Estes gráficos indicam que a aproximação é muito boa.

O próximo passo é propor uma relação funcional entre os θ_i 's e os níveis de doses. Observe que o valor de θ_i , $i = 1, \dots, k$, é desconhecido, mas conhecemos sua média, μ_i , que é o estimador de Bayes quando usamos a função perda quadrática. Portanto, podemos estabelecer uma relação entre μ_i e D_i .

Figura 3.1: Funções densidades de θ_i , para $A_i = 5$ e $B_i = 10$, e da Normal com mesma média e variância

Figura 3.2: Funções densidades de θ_i , para $A_i = 50$ e $B_i = 30$, e da Normal com mesma média e variância

Figura 3.3: Funções densidades de θ_i , para $A_i = 100$ e $B_i = 153$, e da Normal com mesma média e variância

Figura 3.4: Funções densidades de θ_i , para $A_i = 300$ e $B_i = 350$, e da Normal com mesma média e variância

Figura 3.5: Funções densidades de θ_i , para $A_i = 1000$ e $B_i = 600$, e da Normal com mesma média e variância

A Tabela 3.1 apresenta a frequência de células com Micronúcleos (MN) em 4 indivíduos sadios expostos a 8 diferentes doses de radiação Gama. A presença de MN foi observada em linfócitos mono, bi e multinucleados.

Tabela 3.1: Frequência de MN induzidos por diferentes doses de radiação Gama em linfócitos de indivíduos sadios

i	Dose (cGy)	Mononucleados		Binucleados		Multinucleados	
		y_i	n_i	y_i	n_i	y_i	n_i
0	0	35	51272	32	2373	2	84
1	20	109	24000	51	2662	7	150
2	50	204	26892	142	1991	20	183
3	100	521	26437	236	2047	26	164
4	200	1067	24549	407	2611	21	112
5	300	821	9344	708	2442	56	148
6	400	1003	10811	777	2398	50	154
7	500	1130	8814	741	1746	30	59

A Tabela 3.2 apresenta os parâmetros das distribuições Beta *a priori* (*a posteriori*). Os parâmetros da *priori* refletem a opinião dos biólogos de que os p_i 's crescem com os níveis de dose. A Tabela 3.3 apresenta as estimativas de Bayes dos p_i 's e suas variâncias posteriores multiplicadas por 10^5 .

Observe que as estimativas das proporções de células com MN crescem com a dose. Isto é bastante razoável, uma vez que a chance de ocorrência de MN aumenta à medida que aumentamos a dose de radiação. A Tabela 3.4 apresenta os parâmetros da distribuição *a posteriori* dos θ_i 's.

Tabela 3.2: Parâmetros das distribuições *a priori* (*a posteriori*) dos p_i 's

i	Mononucleados		Binucleados		Multinucleados	
	$a_i(A_i)$	$b_i(B_i)$	$a_i(A_i)$	$b_i(B_i)$	$a_i(A_i)$	$b_i(B_i)$
0	1 (36)	9 (51246)	1 (33)	9 (2350)	1 (3)	9 (91)
1	1 (110)	8 (23899)	1 (52)	8 (2619)	1 (8)	8 (151)
2	1 (205)	6 (26694)	1 (143)	6 (1855)	1 (21)	6 (169)
3	1 (522)	5 (25921)	1 (237)	5 (1816)	1 (27)	5 (143)
4	1 (1068)	4 (23486)	1 (408)	4 (2208)	1 (22)	4 (95)
5	1 (822)	3 (8526)	1 (709)	3 (1737)	1 (57)	3 (95)
6	1 (1004)	2 (9810)	1 (778)	2 (1623)	1 (51)	2 (106)
7	1 (1131)	1 (7685)	1 (742)	1 (1006)	1 (31)	1 (30)

Tabela 3.3: Estimativas e variâncias posteriores ($\times 10^5$) dos p_i 's

i	Mononucleados		Binucleados		Multinucleados	
	$\hat{p}_i$	$V(p_i)$	$\hat{p}_i$	$V(p_i)$	$\hat{p}_i$	$V(p_i)$
0	0.0007	0.0014	0.0138	0.5728	0.0319	32.5224
1	0.0046	0.0190	0.0195	0.7144	0.0503	29.8643
2	0.0076	0.0281	0.0716	3.3241	0.1105	51.4713
3	0.0197	0.0732	0.1154	4.9715	0.1588	78.1278
4	0.0435	0.1694	0.1560	5.0301	0.1880	129.3876
5	0.0879	0.8578	0.2899	8.4120	0.3750	153.1863
6	0.0928	0.7787	0.3240	9.1189	0.3248	138.8096
7	0.1283	1.2683	0.4245	13.9678	0.5082	403.1174

Tabela 3.4: Parâmetros da distribuição *a posteriori* dos θ_i 's

i	Mononucleados		Binucleados		Multinucleados	
	μ_i	σ_i^2	μ_i	σ_i^2	μ_i	σ_i^2
0	-7.27	0.02819	-4.28	0.03119	-3.58	0.40598
1	-5.38	0.00917	-3.93	0.01980	-3.00	0.13978
2	-4.87	0.00493	-2.57	0.00756	-2.11	0.05470
3	-3.91	0.00196	-2.04	0.00478	-1.68	0.04475
4	-3.09	0.00098	-1.69	0.00290	-1.48	0.05708
5	-2.34	0.00134	-0.90	0.00198	-0.51	0.02828
6	-2.28	0.00110	-0.73	0.00191	-0.74	0.02928
7	-1.92	0.00101	-0.30	0.00234	0.03	0.06667

O ajuste do modelo dose-resposta será feito para cada grupo de linfócitos. No caso dos linfócitos mononucleados, o modelo proposto é o seguinte:

$$-\mu_i = \frac{\beta}{\gamma + D_i} \quad (3.12)$$

onde β e γ são números reais. Com os dados apresentados na Tabela 3.4, e usando o método de regressão linear, obtemos

$$\beta = 1369.70 \quad (3.13)$$

$$\gamma = 230.91 \quad (3.14)$$

Para os linfócitos binucleados, o modelo proposto é o seguinte:

$$-\mu_i = \alpha + \frac{\beta}{\gamma + D_i} \quad (3.15)$$

onde α é um número real. Com os dados da Tabela 3.4, e usando o procedimento de regressão não-linear, NLR, do software SPSS, obtemos

$$\alpha = -0.40 \quad (3.16)$$

$$\beta = 549.81 \quad (3.17)$$

$$\gamma = 115.98 \quad (3.18)$$

Para os linfócitos multinucleados, o modelo proposto é o seguinte:

$$-\mu_i = \frac{\beta}{\gamma + D_i} \quad (3.19)$$

Com os dados da Tabela 3.4 (sem a dose 500), e usando o método de regressão linear, obtemos

$$\beta = 286.03 \quad (3.20)$$

$$\gamma = 78.82 \quad (3.21)$$

A Tabela 3.5 apresenta os valores ajustados, $\hat{\mu}_i$'s, e seus respectivos erros. Para obter uma função que relaciona os níveis de doses e os p_i 's, usamos a expressão (3.5), substituindo θ_i por μ_i , e usamos as funções ajustadas para os μ_i 's, expressões (3.12), (3.15) e (3.19) para os linfócitos mono, bi e multinucleados, respectivamente. Logo, obtemos a seguinte relação funcional entre as doses e os p_i 's:

- Mononucleados e multinucleados:

$$p_i = \frac{\exp\left\{-\left(\frac{\beta}{\gamma + D_i}\right)\right\}}{1 + \exp\left\{-\left(\frac{\beta}{\gamma + D_i}\right)\right\}} \quad (3.22)$$

- Binucleados:

$$p_i = \frac{\exp\{-(\alpha + \frac{\beta}{\gamma + D_i})\}}{1 + \exp\{-(\alpha + \frac{\beta}{\gamma + D_i})\}} \quad (3.23)$$

A Tabela 3.6 apresenta os valores ajustados, $\tilde{p}_i$'s, e os correspondentes erros observados, $\tilde{\varepsilon}_i$'s. As Figuras 3.6, 3.7 e 3.8 ilustram o ajuste para as proporções, p_i 's, nos linfócitos mono, bi e multinucleados, respectivamente.

Tabela 3.5: Os μ_i 's ajustados e seus respectivos erros

i	Mononucleados		Binucleados		Multinucleados	
	$\hat{\mu}_i$	ε_i	$\hat{\mu}_i$	ε_i	$\hat{\mu}_i$	ε_i
0	-5.93	-1.34	-4.34	0.06	-3.63	0.05
1	-5.46	0.08	-3.64	-0.29	-2.89	-0.11
2	-4.87	0.00	-2.91	0.34	-2.22	0.11
3	-4.14	0.23	-2.14	0.10	-1.60	-0.08
4	-3.18	0.09	-1.34	-0.35	-1.02	-0.46
5	-2.58	0.24	-0.92	0.02	-0.75	0.24
6	-2.17	-0.11	-0.66	-0.07	-0.60	-0.14
7	-1.87	-0.05	-0.49	0.19	-	-

Tabela 3.6: Os p_i 's ajustados e seus respectivos erros

i	Mononucleados		Binucleados		Multinucleados	
	$\tilde{p}_i$	$\tilde{\epsilon}_i$	$\tilde{p}_i$	$\tilde{\epsilon}_i$	$\tilde{p}_i$	$\tilde{\epsilon}_i$
0	0.0026	-0.0019	0.0129	0.0009	0.0258	0.0061
1	0.0042	0.0004	0.0256	-0.0061	0.0526	-0.0023
2	0.0076	0.0000	0.0517	0.0199	0.0980	0.0125
3	0.0157	0.0040	0.1053	0.0101	0.1680	-0.0092
4	0.0399	0.0036	0.2075	-0.0515	0.2650	-0.0770
5	0.0704	0.0175	0.2849	0.0050	0.3208	0.0542
6	0.1025	-0.0097	0.3407	-0.0167	0.3543	-0.0295
7	0.1335	-0.0052	0.3799	0.0446	-	-

Figura 3.6: Proporção de células com MN nos linfócitos mononucleados

Figura 3.7: Proporção de células com MN nos linfócitos binucleados

Figura 3.8: Proporção de células com MN nos linfócitos multinucleados

3.3 Caso Multinomial

Nesta seção, observaremos a variável aleatória Y_{ij} , a frequência de células com j anomalias no i -ésimo nível de dose, $i = 1, 2, \dots, k$; $j = 1, 2, \dots, m$. A fim de facilitar o desenvolvimento analítico, vamos considerar a análise da ocorrência de MN (Tabela 3.1) em 3 categorias: células com zero MN, células com um MN e células com dois ou mais MN. Os dados observados, para uma particular dose i ($i = 0, 1, \dots, 7$), consistem de um vetor de 3 componentes, $\mathbf{y}_i = (y_{i0}, y_{i1}, y_{i2})$, correspondentes às frequências das 3 categorias já mencionadas. A distribuição considerada para qualquer dose i é a distribuição Multinomial com parâmetro $(n_i; p_{i0}, p_{i1}, p_{i2})$, onde $n_i = y_{i0} + y_{i1} + y_{i2}$, $p_{i0} + p_{i1} + p_{i2} = 1$. Portanto, a verossimilhança é dada por

$$L_i(\mathbf{y}_i | \mathbf{p}_i) = \frac{n_i!}{y_{i0}! y_{i1}! y_{i2}!} p_{i0}^{y_{i0}} p_{i1}^{y_{i1}} p_{i2}^{y_{i2}} \quad (3.24)$$

onde $\mathbf{p}_i = (p_{i0}, p_{i1}, p_{i2})$ é o parâmetro de interesse para a dose i , e p_{i0} , p_{i1} e p_{i2} são, respectivamente, as proporções de células com zero MN, um MN e dois ou mais MN.

A verossimilhança para os dados completos é o produto das funções de verossimilhanças das 8 doses, pois temos experimentos independentes entre doses. Isto é,

$$L(\mathbf{y} | \mathbf{p}) = \prod_{i=0}^7 L_i(\mathbf{y}_i | \mathbf{p}_i) \quad (3.25)$$

onde $\mathbf{y} = (\mathbf{y}_0, \dots, \mathbf{y}_7)$ e $\mathbf{p} = (\mathbf{p}_0, \dots, \mathbf{p}_7)$.

A classe conjugada de distribuições *a priori* da Multinomial é a família Dirichlet (ver Basu & Pereira, 1982). Suponha que a distribuição *a priori*

para $\mathbf{p}_i$ é uma Dirichlet de ordem 3 com parâmetro $\mathbf{a}_i = (a_{i0}, a_{i1}, a_{i2})$, onde $a_{ij} \geq 0$, $j = 0, 1, 2$, ou seja, $\mathbf{p}_i \sim D_3(\mathbf{a}_i)$. A distribuição *a posteriori* de $\mathbf{p}_i$ é $D_3(\mathbf{A}_i)$, onde $\mathbf{A}_i = \mathbf{a}_i + \mathbf{y}_i = (A_{i0}, A_{i1}, A_{i2})$. Por simplicidade, denotaremos $a_{i.} = a_{i0} + a_{i1} + a_{i2}$ e $A_{i.} = A_{i0} + A_{i1} + A_{i2}$. A função densidade *a posteriori* de $\mathbf{p}_i$ é dada por

$$f_i(\mathbf{p}_i | \mathbf{y}_i) = \frac{\Gamma(A_{i.})}{\Gamma(A_{i0})\Gamma(A_{i1})\Gamma(A_{i2})} p_{i0}^{A_{i0}} p_{i1}^{A_{i1}} p_{i2}^{A_{i2}}. \quad (3.26)$$

O estimador de Bayes de $\mathbf{p}_i$, considerando a função perda quadrática, é a média posterior, dada por

$$\hat{\mathbf{p}}_i = (\hat{p}_{i0}, \hat{p}_{i1}, \hat{p}_{i2}) = \frac{1}{A_{i.}}(A_{i0}, A_{i1}, A_{i2}) \quad (3.27)$$

A matriz de variância-covariância posterior é dada por

$$\frac{1}{A_{i.} + 1} \begin{pmatrix} \hat{p}_{i0}(1 - \hat{p}_{i0}) & -\hat{p}_{i0}\hat{p}_{i1} & -\hat{p}_{i0}\hat{p}_{i2} \\ -\hat{p}_{i0}\hat{p}_{i1} & \hat{p}_{i1}(1 - \hat{p}_{i1}) & -\hat{p}_{i1}\hat{p}_{i2} \\ -\hat{p}_{i0}\hat{p}_{i2} & -\hat{p}_{i1}\hat{p}_{i2} & \hat{p}_{i2}(1 - \hat{p}_{i2}) \end{pmatrix}.$$

Em consequência da independência *a priori* dos $\mathbf{p}_i$'s e da independência estatística dos $\mathbf{y}_i$'s, a densidade conjunta *a posteriori* dos $\mathbf{p}_i$'s é dada por

$$f(\mathbf{p} | \mathbf{y}) = \prod_{i=0}^7 f_i(\mathbf{p}_i | \mathbf{y}_i). \quad (3.28)$$

Analisaremos os dados da Tabela 3.1 correspondentes aos linfócitos binucleados, para os outros dois grupos, mono e multinucleados, o procedimento é semelhante. A Tabela 3.7 apresenta estes dados e os parâmetros das distribuições Dirichlet *a priori* (*a posteriori*). A Tabela 3.8 apresenta as

Tabela 3.7: Dados e parâmetros das distribuições *a priori* (*a posteriori*)

i	Dose (cGy)	Dados				Parâmetros <i>a priori</i> (<i>a posteriori</i>)			
		y_{i0}	y_{i1}	y_{i2}	n_i	$a_{i0}(A_{i0})$	$a_{i1}(A_{i1})$	$a_{i2}(A_{i2})$	$a_{i.}(A_{i.})$
0	0	2341	31	1	2373	10 (2351)	2 (33)	1 (2)	13 (2386)
1	20	2611	45	6	2662	9 (2620)	2 (47)	1 (7)	12 (2674)
2	50	1849	117	25	1991	8 (1857)	2 (119)	1 (26)	11 (2002)
3	100	1811	189	47	2047	7 (1818)	2 (191)	1 (48)	10 (2057)
4	200	2204	325	82	2611	6 (2210)	2 (327)	1 (83)	9 (2620)
5	300	1734	501	207	2442	5 (1739)	2 (503)	1 (208)	8 (2450)
6	400	1621	523	254	2398	4 (1625)	2 (525)	1 (255)	7 (2405)
7	500	1005	456	285	1746	3 (1008)	2 (458)	1 (286)	6 (1752)

estimativas de Bayes e as variâncias posteriores (multiplicadas por 10^5) dos p_i 's.

O próximo passo é estabelecer uma relação dose-resposta. A transformação *log-odds* é dada por

$$\theta_{ik} = \ln \frac{p_{ik}}{p_{i0}} \quad k = 1, 2 \quad (3.29)$$

e a transformação inversa por

$$p_{ik} = \frac{\exp(\theta_{ik})}{1 + \exp(\theta_{i1}) + \exp(\theta_{i2})}. \quad (3.30)$$

Novamente tem-se uma propriedade bastante interessante: $\theta_i = (\theta_{i1}, \theta_{i2})$ tem, aproximadamente, distribuição Normal bivariada quando p_i é Dirichlet (Aitchison & Shen, 1980). Seguindo etapas semelhantes àquelas da demons-

Tabela 3.8: Estimativas e variâncias posteriores ($\times 10^5$) dos p_{ij} 's

i	$\hat{p}_{i0}$	$\hat{p}_{i1}$	$\hat{p}_{i2}$	$V(p_{i0})$	$V(p_{i1})$	$V(p_{i2})$
0	0.9853	0.0138	0.0008	0.6068	0.5701	0.0335
1	0.9798	0.0176	0.0026	0.7399	0.6464	0.0969
2	0.9276	0.0594	0.0130	3.3529	2.7894	0.6406
3	0.8838	0.0929	0.0233	4.9902	4.0947	1.1058
4	0.8435	0.1248	0.0317	5.0365	4.1673	1.1711
5	0.7098	0.2053	0.0849	8.4041	6.6565	3.1698
6	0.6757	0.2183	0.1060	9.1076	7.0925	3.9386
7	0.5753	0.2614	0.1632	13.9378	11.0137	7.7904

tração do Lema 3.2.2, pode-se mostrar que os parâmetros da distribuição *a posteriori* de θ_i são dados por

$$\mu_{ik} = E[\theta_{ik}] = \psi(A_{ik}) - \psi(A_{i0}) \quad (3.31)$$

$$\sigma_{ik}^2 = Var[\theta_{ik}] = \psi'(A_{ik}) + \psi'(A_{i0}) \quad (3.32)$$

$$c_i = Cov[\theta_{i1}, \theta_{i2}] = \psi'(A_{i0}) \quad (3.33)$$

No Apêndice A esboçamos os gráficos da função densidade de θ_i , para diferentes valores do vetor (A_{i0}, A_{i1}, A_{i2}) , e da função densidade de uma variável com distribuição Normal bivariada com os mesmos parâmetros. Estes gráficos indicam que a aproximação é muito boa. A Tabela 3.9 apresenta os valores destes parâmetros posteriores. A relação funcional entre os níveis de doses e os θ_i 's é semelhante àquela proposta na seção anterior (expressão (3.15)):

Tabela 3.9: Parâmetros da distribuição *a posteriori* de $(\theta_{i1}, \theta_{i2})$

i	μ_{i1}	μ_{i2}	σ_{i1}^2	σ_{i2}^2	c_i
0	-4.28	-7.34	0.03119	0.64535	0.00042
1	-4.03	-6.00	0.02188	0.15392	0.00038
2	-2.75	-4.29	0.00898	0.03975	0.00054
3	-2.25	-3.64	0.00580	0.02160	0.00055
4	-1.91	-3.29	0.00351	0.01257	0.00045
5	-1.24	-2.12	0.00256	0.00539	0.00057
6	-1.13	-1.85	0.00252	0.00454	0.00061
7	-0.79	-1.26	0.00317	0.00449	0.00099

$$-\mu_{ik} = \alpha_k + \frac{\beta_k}{\gamma_k + D_i}. \quad (3.34)$$

Utilizando o procedimento NLR do SPSS, obtemos

$$\alpha_1 = 0.18 \quad \alpha_2 = 0.78 \quad (3.35)$$

$$\beta_1 = 449.01 \quad \beta_2 = 506.63 \quad (3.36)$$

$$\gamma_1 = 107.05 \quad \gamma_2 = 78.20 \quad (3.37)$$

A Tabela 3.10 apresenta os valores ajustados, $\hat{\mu}_{ik}$'s, e os correspondentes erros. Usando a expressão (3.30), com $\hat{\mu}_{ik}$ substituindo θ_{ik} , obtem-se a relação funcional estimada entre as doses e os p_{ik} 's. A Tabela 3.11 apresenta os valores ajustados, $\tilde{p}_{ik}$'s, e os respectivos erros, $\tilde{\epsilon}_{ik}$'s. Note que $\tilde{p}_{i0} = 1 - \tilde{p}_{i1} - \tilde{p}_{i2}$. As Figuras 3.9, 3.10 e 3.11 ilustram este ajuste.

Tabela 3.10: Valores ajustados, $\hat{\mu}_{ik}$'s, e seus respectivos erros

i	$\hat{\mu}_{i1}$	ϵ_{i1}	$\hat{\mu}_{i2}$	ϵ_{i2}
0	-4.37	0.09	-7.26	-0.08
1	-3.71	-0.32	-5.94	-0.06
2	-3.04	0.29	-4.73	0.44
3	-2.35	0.10	-3.62	-0.02
4	-1.64	-0.27	-2.60	-0.69
5	-1.28	0.04	-2.12	0.00
6	-1.06	-0.07	-1.84	-0.01
7	-0.92	0.13	-1.66	0.40

Tabela 3.11: Valores ajustados, $\tilde{p}_{ik}$'s, e seus respectivos erros

i	$\tilde{p}_{i0}$	$\tilde{\epsilon}_{i0}$	$\tilde{p}_{i1}$	$\tilde{\epsilon}_{i1}$	$\tilde{p}_{i2}$	$\tilde{\epsilon}_{i2}$
0	0.9868	-0.0015	0.0125	0.0013	0.0007	0.0001
1	0.9736	0.0062	0.0238	-0.0062	0.0026	0.0000
2	0.9464	-0.0188	0.0453	0.0141	0.0083	0.0047
3	0.8911	-0.0073	0.0850	0.0079	0.0239	-0.0006
4	0.7885	0.0550	0.1529	-0.0281	0.0586	-0.0269
5	0.7153	-0.0055	0.1989	0.0064	0.0858	-0.0009
6	0.6643	0.0114	0.2302	-0.0119	0.1055	0.0005
7	0.6295	-0.0542	0.2508	0.0106	0.1197	0.0435

Figura 3.9: Proporção de células sem MN

Figura 3.10: Proporção de células com um MN

Figura 3.11: Proporção de células com dois ou mais MN

3.4 Outras aplicações

3.4.1 Região de Credibilidade

Para exemplificar uma situação prática onde nosso método poderia ser adequado, considere um acidente nuclear fictício em que um trabalhador sofreu exposição a uma dose de radiação desconhecida, ϑ . Com o objetivo de estimar a dose de radiação que este indivíduo sofreu, planeja-se um experimento de calibração: culturas de sangue de outros trabalhadores de mesma atividade, mas que não sofreram o mesmo acidente, são expostas a diferentes níveis de radiação e os valores das respectivas respostas, $\mathbf{y}$'s, são obtidos. Assuma que os dados do experimento de calibração são aqueles da Tabela 3.7 e os valores observados para o indivíduo exposto foram $\mathbf{y}_{\vartheta} = (1620, 500, 236)$. Supo-

nha que o parâmetro posterior da distribuição Dirichlet para a frequência de células com zero MN, um MN e dois ou mais MN é $A_{\partial} = (1625, 503, 237)$. As constantes estimadas do modelo dose-resposta são aquelas em (3.35), (3.36) e (3.37).

Nosso interesse é a obtenção de uma região de credibilidade para a dose desconhecida, ∂ . Mas, para obtê-la diretamente seria necessário conhecer a distribuição de probabilidades de ∂ . Esta distribuição não é conhecida, mas sabemos que a variável $\theta_{\partial} = (\theta_{\partial 1}, \theta_{\partial 2})$ tem distribuição Normal bivariada. Logo, podemos obter uma região de credibilidade para θ_{∂} e, usando as expressões (3.29), (3.30), (3.35), (3.36) e (3.37), obter uma região de credibilidade para ∂ . Para isto, precisamos conhecer os pontos $(\theta_{\partial 1}, \theta_{\partial 2})$ pertencentes a região de credibilidade de θ_{∂} . O procedimento que utilizamos foi o seguinte:

- 1) Gerar vários pares $(\theta_{\partial 1}, \theta_{\partial 2})$ e os respectivos valores da sua função densidade;
- 2) Ordenar os pares gerados pelo valor da função densidade;
- 3) Selecionar 95% dos pares que têm maiores valores da função densidade;

Observe que estes pares selecionados formam uma região com 95% de credibilidade para θ_{∂} . O próximo passo é o cálculo dos correspondentes valores da dose para obtenção de sua região de credibilidade:

- 4) Obter os valores de $\tilde{p}_{\partial} = (\tilde{p}_{\partial 0}, \tilde{p}_{\partial 1}, \tilde{p}_{\partial 2})$ usando (3.30);
- 5) Obter os valores das doses, (∂_1, ∂_2) , usando os valores obtidos na etapa 4 e as expressões (3.29), (3.35), (3.36) e (3.37), ou seja,

$$\begin{aligned}\partial_1 &= \frac{449.01}{\ln(\tilde{p}_{\partial 0}/\tilde{p}_{\partial 1}) - 0.18} - 107.05 \\ \partial_2 &= \frac{506.63}{\ln(\tilde{p}_{\partial 0}/\tilde{p}_{\partial 2}) - 0.78} - 78.20\end{aligned}$$

- 6) Encontrar os valores mínimos e máximos de ∂_1 e ∂_2 ;
- 7) Selecionar o maior valor entre os mínimos, ∂_{min} , e o menor valor entre os máximos, ∂_{max} ;
- 8) Somar os valores da função densidade para as doses no intervalo $(\partial_{min}, \partial_{max})$;
- 9) Dividir a soma calculada no item anterior pela soma total da função densidade correspondente aos pares selecionados na etapa 3.

Esta razão indica a credibilidade neste intervalo, $(\partial_{min}, \partial_{max})$. Se quisermos uma credibilidade maior (menor), aumentamos (diminuimos) o intervalo e repetimos as etapas 8 e 9. Os resultados obtidos para o nosso exemplo foram os seguintes:

$$\partial_{min} = 278.58 \quad \partial_{max} = 432.30$$

A razão entre as somas da função densidade para o intervalo (278.58, 432.30) é aproximadamente um. Diminuindo o intervalo para (310, 401), o valor da razão é 0.9730. Ou seja, temos 97,30% de credibilidade de que a dose de radiação sofrida pelo trabalhador é um número no intervalo (310, 401).

3.4.2 Comparação entre as frequências de anomalias de diversos grupos de indivíduos

Nesta seção, estamos interessados em comparar as frequências de anomalias de grupos de indivíduos expostos a diferentes doses de radiação. Uma solução intuitiva, envolvendo as técnicas desenvolvidas neste capítulo, é a comparação dos modelos dose-resposta de cada grupo de indivíduo. Para

Tabela 3.12: Dados e parâmetros das distribuições *a priori* (*a posteriori*) dos indivíduos do Grupo 2

i	Dose (cGy)	Dados				Parâmetros <i>a priori</i> (<i>a posteriori</i>)			
		y_{i0}	y_{i1}	y_{i2}	n_i	$a_{i0}(A_{i0})$	$a_{i1}(A_{i1})$	$a_{i2}(A_{i2})$	$a_{i.}(A_{i.})$
0	0	1010	23	10	1043	8 (1018)	2 (25)	1 (11)	11 (1054)
1	180	769	202	27	998	7 (776)	2 (204)	1 (28)	10 (1008)
2	360	924	350	73	1047	6 (930)	2 (352)	1 (74)	9 (1056)
3	540	540	434	87	1061	5 (545)	2 (436)	1 (88)	8 (1069)
4	3150	690	430	121	1241	4 (694)	2 (432)	1 (122)	7 (1248)
5	6300	591	445	148	1184	3 (594)	2 (447)	1 (149)	6 (1190)

ilustrar, analisaremos dados referentes às frequências de células com anomalias observadas em linfócitos binucleados de 3 grupos de indivíduos, que chamaremos Grupo 1, Grupo 2 e Grupo 3. Os dados referentes aos indivíduos do Grupo 1 são aqueles apresentados na Tabela 3.7. As Tabelas 3.12 e 3.13 apresentam as frequências observadas e os parâmetros das distribuições Dirichlet *a priori* (*a posteriori*) dos indivíduos dos Grupos 2 e 3, respectivamente.

De (3.34), (3.35), (3.36) e (3.37), obtemos o modelo ajustado para os indivíduos do Grupo 1. Seguindo as mesmas etapas desenvolvidas anteriormente, obtemos os ajustes para os Grupos 2 e 3. Estes ajustes são dados por

Grupo 1 (G1):

$$-\hat{\mu}_{i1} = 0.18 + \frac{449.01}{107.05 + D_i} \quad (3.38)$$

Tabela 3.13: Dados e parâmetros das distribuições *a priori* (*a posteriori*) dos indivíduos do Grupo 3

i	Dose (cGy)	Dados				Parâmetros <i>a priori</i> (<i>a posteriori</i>)			
		y_{i0}	y_{i1}	y_{i2}	n_i	$a_{i0}(A_{i0})$	$a_{i1}(A_{i1})$	$a_{i2}(A_{i2})$	$a_{i.}(A_{i.})$
0	0	920	31	2	953	10 (930)	2 (33)	1 (3)	13 (966)
1	20	989	41	8	1038	9 (998)	2 (43)	1 (9)	12 (1050)
2	50	933	56	14	1003	8 (941)	2 (58)	1 (15)	11 (1014)
3	100	939	114	32	1085	7 (946)	2 (116)	1 (33)	10 (1095)
4	200	794	176	67	1037	6 (800)	2 (178)	1 (68)	9 (1046)
5	300	683	209	59	951	5 (688)	2 (211)	1 (60)	8 (959)
6	400	742	256	107	1105	4 (746)	2 (258)	1 (108)	7 (1112)
7	500	771	327	143	1241	3 (774)	2 (329)	1 (144)	6 (1247)

$$-\hat{\mu}_{i2} = 0.78 + \frac{506.63}{78.20 + D_i} \quad (3.39)$$

Grupo 2 (G2):

$$-\hat{\mu}_{i1} = 0.22 + \frac{259.45}{73.82 + D_i} \quad (3.40)$$

$$-\hat{\mu}_{i2} = 1.32 + \frac{677.89}{205.78 + D_i} \quad (3.41)$$

Grupo 3 (G3):

$$-\hat{\mu}_{i1} = -0.08 + \frac{625.81}{177.66 + D_i} \quad (3.42)$$

$$-\hat{\mu}_{i2} = 1.13 + \frac{411.63}{87.74 + D_i} \quad (3.43)$$

Para cada grupo, o modelo dose-resposta é dado por

$$\tilde{p}_{i0} = \frac{1}{1 + \exp(\hat{\mu}_{i1}) + \exp(\hat{\mu}_{i2})} \quad (3.44)$$

$$\tilde{p}_{i1} = \frac{\exp(\hat{\mu}_{i1})}{1 + \exp(\hat{\mu}_{i1}) + \exp(\hat{\mu}_{i2})} \quad (3.45)$$

$$\tilde{p}_{i2} = \frac{\exp(\hat{\mu}_{i2})}{1 + \exp(\hat{\mu}_{i1}) + \exp(\hat{\mu}_{i2})} \quad (3.46)$$

As Figuras 3.12, 3.13 e 3.14 esboçam, respectivamente, os modelos dose-resposta dos grupos para as proporções de células com zero MN, um MN e dois ou mais MN. Estas figuras nos indicam que não parece haver diferenças entre os indivíduos dos Grupos 1 e 3, enquanto que os indivíduos do Grupo 2 parecem ter um comportamento diferente.

Figura 3.12: Proporção de células com zero MN

Figura 3.13: Proporção de células com um MN

Figura 3.14: Proporção de células com dois ou mais MN

Capítulo 4

CONSIDERAÇÕES FINAIS

Observe que não nos detivemos nos conceitos biológicos associados aos dados analisados. O nosso maior interesse era apresentar os detalhes das técnicas estatísticas Bayesianas utilizadas. Pretendemos, entretanto, em um próximo trabalho, apresentar esses conceitos em maiores detalhes.

No Capítulo 3, propositadamente não foram feitas conclusões com relação aos resultados obtidos, visto que decisões a serem tomadas exigem análise biológica elaborada. As técnicas que utilizamos para solução do problema de ajuste em experimentos de dosimetria são inéditas e, esperamos ter contribuído com as pesquisas que estão sendo desenvolvidas na análise deste tipo de experimento.

Apêndice A

APROXIMAÇÃO NORMAL BIVARIADA

Com o objetivo de analisar a aproximação Normal bivariada para a distribuição da transformação *log-odds*, esboçamos alguns gráficos da função densidade de θ_i e da função densidade de uma variável com distribuição Normal bivariada. Os parâmetros da distribuição de θ_i foram escolhidos entre aqueles da Tabela 3.7. Cada par de gráficos representa uma comparação entre estas densidades para um determinado vetor de parâmetros. Por exemplo, as Figuras A.1 e A.2 representam, respectivamente, a função densidade de θ_i , para $A_i = (2351, 33, 2)$, e a função densidade de uma variável com distribuição Normal bivariada com o mesmo vetor de médias e a mesma matriz de variância-covariância.

Figura A.1: Função densidade de $(\theta_{i1}, \theta_{i2})$ para $A_i = (2351, 33, 2)$

Normal

Figura A.2: Função densidade da Normal bivariada correspondente ao parâmetro $A_i = (2351, 33, 2)$

Figura A.3: Função densidade de $(\theta_{:1}, \theta_{:2})$ para $A_i = (1857, 119, 26)$

Normal

Figura A.4: Função densidade da Normal bivariada correspondente ao parâmetro $A_i = (1857, 119, 26)$

Figura A.5: Função densidade de $(\theta_{i,1}, \theta_{i,2})$ para $A_i = (1739, 503, 208)$

Normal

Figura A.6: Função densidade da Normal bivariada correspondente ao parâmetro $A_i = (1739, 503, 208)$

Bibliografia

- [1] Aitchison, J. & Shen, S.M. 1980. Logistic-normal distributions: some properties and uses. *Biometrika* 67: 261-72.
- [2] Balasem, A.N. & Ali, A.S.K. 1991. Establishment of dose-response relationships between doses of Cs-137 γ -rays and frequencies of micronuclei in human peripheral blood lymphocytes. *Mutation Research* 259: 133-8.
- [3] Basu, D. & Pereira, C.A. de B. 1982. On the Bayesian analysis of categorical data: the problem of nonresponse. *Journal of Statistical Planning and Inference* 6: 345-62.
- [4] Bender, M.A.; Awa, A.A.; Brooks, A.L.; Evans, H.J.; Groër, P.G.; Littlefield, L.G.; Pereira, C.A. de B.; Preston, R.J. & Wachholz, B.W. 1988. Current status of cytogenetic procedures to detect and quantify previous exposures to radiation. *Mutation Research* 196: 103-59.
- [5] Fenech, M. & Morley, A.A. 1985. Measurement of micronuclei in lymphocytes. *Mutation Research* 147: 29-36.
- [6] Groër, P.G. & Pereira, C.A. de B. 1987. Calibration of a radiation detector: chromosome dosimetry for neutrons. In: R. Viertl(ed.) *Probability*

and *Bayesian Statistics*, Plenum Publ. Corp., pp. 225-32.

- [7] Kendall, M.; Stuart, A. & Ord, J.K. 1987. *Kendall's Theory of Statistics. London, Charles Griffin & Company Limited. V.1.*
- [8] Kormos, C. & Kóteles, G.J. 1988. Micronuclei in X-irradiated human lymphocytes. *Mutation Research* 199: 31-5.
- [9] Pereira, C.A. de B. 1985. Teste de hipóteses definidas em espaços de diferentes dimensões: visão Bayesiana e interpretação clássica. *Tese de Livre-Docência. IME-USP.*
- [10] Pereira, C.A. de B. & Lindley, D.V. 1987. Examples questioning the use of partial likelihood. *The Statistician* 36: 15-20.
- [11] Prosser, J.S.; Moquet, J.E.; Lloyd, D.C. & Edwards, A.A. 1988. Radiation induction of micronuclei in human lymphocytes. *Mutation Research* 199: 37-45.
- [12] Rabello-Gay, M.N.; Rodrigues, M.A.L.R. & Moontealeone-Neto, R. 1991. Mutagênese, Teratogênese e Carcinogênese - Métodos e Critérios de Avaliação. *Sociedade Brasileira de Genética.*