Estimadores Bayesianos Empíricos Espaciais de Taxas

Renato Assunção - Depto de Estatística, UFMG

1 Introdução

A análise de dispersão espacial do risco de uma doença é feita principalmente através de mapas de taxas de incidência. Estes mapas são instrumentos valiosos para apontar associações entre fontes potenciais de contaminação e áreas de risco elevado, para sugerir determinantes locais de doenças e fatores etiológicos desconhecidos e para visualização da distribuição espacial da doença. Mason (1995) apresenta os vários estudos de campo realizados como conseqüência das questões ligadas aos determinantes ambientais do câncer levantadas após análise dos diversos Atlas editados pelo National Cancer Intitute americano. Estes estudos abrangem câncer oral (Winn *et al.*, 1981), câncer do intestino (Pickle *et al.*, 1981), câncer de pulmão (Ziegle *et al.*, 1984), câncer de bexiga (Hoover and Strasser, 1980). Glass *et al.* (1995) produziram um mapa de risco de doenças de Lyme a partir de dados epidemiológicos e um sistema de informação geográfica.

A maior parte dos mapas são constituídos por mapas temáticos ou coropléticos onde um conjunto de áreas são sombreadas de acordo com seus valores em certa variável de interesse. Por exemplo, a Figura 1 mostra um mapa da mortalidade infantil no estado de São Paulo no ano de 1998. O mapa está subdividido nos 644 municípios existentes em 1998 e cada município recebeu uma tonalidade de cinza de acordo com o valor do índice de mortalidade infantil (MI). O valor de MI é definido como a razão entre o número de crianças abaixo de um ano de idade mortas durante um ano e o número de crianças nascidas vivas naquele mesmo ano multiplicado por 1000.

O mapa fornece uma descrição da distribuição da MI na região. Existe um gradiente do Sudeste do estado, onde o risco é mais elevado, para o Nordeste, onde o risco declina consideravelmente. A intensidade deste acréscimo pode também ser apreciada com uma rápida consulta à legenda. Fica evidente a presença de autocorrelação espacial entre as taxas: municípios vizinhos tendem a ter taxas mais similares do que dois municípios escolhidos ao acaso dentre os 644 existentes.

Por outro lado, o mapa também apresenta características que revelam alguns dos problemas. Nota-se a presença de alguns municípios com valores muito discrepantes de seus vizinhos. Na verdade, alguns dos valores mais extremos das taxas são observados nestes casos de discrepância espacial.

Flutuações extremas nas taxas levam algumas vezes à decisão de não se divulgar taxas ao nível de municípios ou quando as áreas são muito pequenas. No entanto, isto entra em conflito com um dos principais objetivos de se fazer os mapas. Para alcançar plenamente estes objetivos, os mapas devem possuir resolução geográfica adequada. Isto implica que a maior utilidade dos mapas ocorre quando eles utilizam pequenas regiões geográficas como unidades de análise. Várias destas pequenas áreas possuirão pequenas populações de risco o que acarretará taxas de incidência com muita

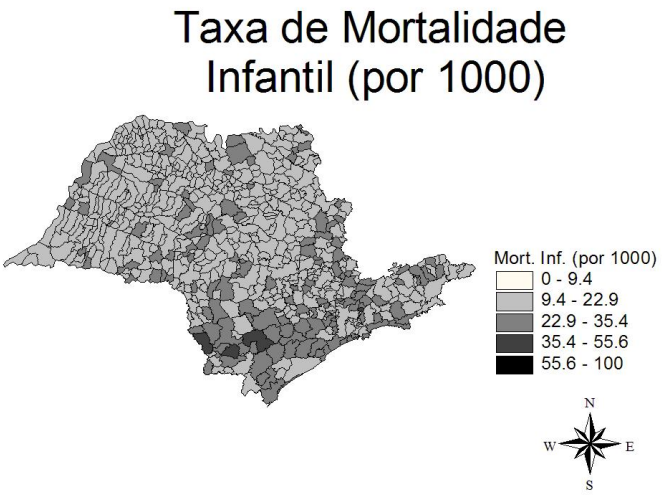

Figure 1: Mapa temático de mortalidade infantil nos municípios de São Paulo no ano de 1998.

instabilidade. Isto é, o acréscimo ou decréscimo de um caso nestas áreas poderá causar mudanças drásticas nas taxas. Em termos estatísticos, as taxas das diversas áreas não são comparáveis já que possuem variâncias muito diferentes.

Neste capítulo, nosso objetivo é apresentar os principais métodos usados atualmente para mapear taxas de doenças. Os métodos são apresentados no contexto epidemiológico mas eles podem ser adaptados para mapas de variáveis sociais ou econômicas de outros tipos tais como taxas de desempregou ou de furtos. Vamos considerar inicialmente o modelo clássico de riscos relativos e a alternativa simples de fazer mapas de taxas padronizadas, incluindo a descrição das várias formas de padronizá-las. A seguir, apresentamos de forma mais detalhada a estrutura aleatória das variáveis observadas e dos parâmetros. Finalmente, consideramos a solução proposta por Choynowski (1959) e um método baseado numa abordagem bayesiana empírica.

2 Modelo Clássico de Riscos Relativos

Os dados usuais para estudar a variação espacial de riscos à saúde são constituídos pelas contagens de eventos em cada área de um mapa e da população em risco de sofrer estes eventos. Atualmente, existe uma disponibilidade enorme de dados para que interessados façam estudos epidemiológicos. O endereço http://www.datasus.gov.br, por exemplo, disponibiliza muitos dados de saúde georeferenciados por municípios ou áreas maiores (estados, microregiões, regiões, etc). Encontra-se também disponível neste endereço o software TABWIN que permite fazer mapas destas informações. O IBGE distribui pro um preço muito acessível (R$15,00) um CD com a malha de todos os municípios brasileiros e dos setores censitários dos municípios com população superior a

25 mil habitantes.

O mapa permite fazer comparações entre as áreas. É claro que não faz sentido fazer mapas e comparações das contagens brutas já que elas dependem das populações das áreas. Para permitir comparações entre diferentes populações no espaço ou no tempo, as contagens devem ser padronizadas para gerar taxas que tiram o efeito das diferentes populações. O mais comum é padronizar as contagens considerando as diferenças em tamanho, estrutura etária e por sexo das populações de risco de cada área. A padronização pode ser também por área, por tempo de exposição ou outras características.

Vamos considerar inicialmente as diferenças de tamanhos entre as populações das áreas. Considere um mapa particionado em n áreas indexadas por i com $i = 1, \ldots, n$. Seja y_i a contagem de eventos na área num dado período de tempo e N_i a sua população em risco de sofrer o evento. A taxa per capita na área é definida como $r_i = y_i/N_i$. É comum usar também $r_i = (y_i/N_i)\,100000$, a taxa por 100 mil, a qual é interpretada como sendo o número de eventos que seriam observados na área i caso ela tivesse uma população total de 100000 indivíduos. Esta forma simples padroniza as contagens pois elas agora referem-se a uma mesma população hipotética de tamanho constate em casa área.

Suponha que deseja-se fazer comparações entre dois mapas de uma mesma região, sendo os mapas de doenças diferentes e possuindo incidências muito diferentes. Ou então mapas referentes a uma mesma doença mas com dados de períodos de tempo muitos distantes. Para situações como estas, é comum trabalharmos com medidas de risco relativas ao nível médio na região. Esta medida é a razão (de mortalidade, morbidade, etc) padronizada, ou *standardized (mortality, morbidity, etc) ratio*, ou *SMR*, em inglês. Em cada área , calculamos o número esperado de eventos caso o risco na área seja igual ao risco na região total na região é dado pela taxa (per capita) global $r = \sum_i y_i / \sum_i N_i$. Assim, $E_i = rN_i$ e a *SMR* da área i é obtida comparando-se o número observado de eventos em i com o número esperado nesta área: $SMR_i = y_i/E_i$. Desta forma, um valor de $SMR_i = 1$ indica que a área i teve tantos casos observados quanto o que seria esperado caso seu risco fosse idêntico ao da região toda. Se $SMR_i = 1$ (ou $= 0.5$), então a área i teve um número de casos duas vezes maior (ou menor) que o esperado se seu risco fosse idêntico ao da região toda. É comum ter a *SMR* multiplicada por 100 sendo então 100 o nível de referência da área global.

Para começar a introduzir o modelo estatístico para mapas de doenças, nós vamos considerar o número observado y_i de eventos na área i como ma variável aleatória. A justificativa para isto é baseada no entendimento de que o valor realizado na área i poderia ser diferente caso o tempo pudesse transcorrer novamente pois vários eventos foram frutos de decisões ou ações que podem ser vistas como aleatórias. O fato é que supor as variáveis y_i como aleatórias implica que possui distribuição de probabilidade, valor esperado, variância, etc. A hipótese mais comum quando os eventos são raros (relativamente ao tamanho da população) é que y_i possui distribuição de Poisson com valor esperado μ_i específico por área. Além de farta evidência empírica que dá suporte a esta hipótese, existem também resultados teóricos para justificá-la. Usando hipótese bastante realistas e fracas, Brillinger (1986) derivou esta distribuição de Poisson para dados de contagens tratando as taxas vitais como estatísticas obtidas a partir de processos de Poisson gerando eventos no diagrama de Lexis, uma técnica gráfica básica de análise demográfica. Se o risco é constante na região, então $\mu_i = E_i = rN_i$.

Quando o risco varia com a idade ou o sexo, não é suficiente fazer apenas a padronização pelo tamanho da população. A maioria dos fenômenos de interesse serão deste tipo tais como mortes violentas, que atingem mais os homens jovens, a maior parte dos cânceres e doenças cardíacas, que atingem os mais velhos, AIDS, etc. Para tirar o efeito das diferenças de distribuição etária e por sexo das populações (além de seu tamanho), realizamos padronização indireta. Seja i o índice da área e j o índice da classe de idade-sexo. Por exemplo, $j = 1$ indica masculino de 0 a 4 anos de idade, $j = 2$ indica masculino de 5 a 9 anos de idade, etc.

Vamos fixar a atenção inicialmente numa classe j. Por exemplo, $j = 5$, que representa a classe de homens de 20 a 24 anos de idade. Seja y_{ij} o número de eventos que ocorreram entre pessoas da classe j na área i e N_{ij} o número de pessoas da classe j na área i. A taxa global em todo mapa referente apenas à classe de idade-sexo j é dada por $r_j = \sum_i y_{ij} / \sum_i N_{ij}$. Então, $E_{ij} = r_j N_{ij}$ é o número esperado de eventos na classe j e na área i se o risco na classe j fosse constante no espaço.

O número total de eventos esperados na área i se o risco de cada uma das classes j de idade-sexo é constante no espaço é dado por $E_i = \sum_j E_{ij}$, a soma dos números esperados na área i nas diferentes classes de idade-sexo. A *SMR* é então calculada como a razão entre o número observado de eventos e número esperado caso o risco fosse constante no espaço: $SMR_i = y_i/E_i$. Assim, na hipótese de que o risco é constante no espaço em cada classe de idade-sexo, temos $y_i \sim$Poisson(E_i)onde E_i é calculado como foi explicado acima.

Na verdade, ao fazer um mapa para estudar a variação espacial, já estamos assumindo que o risco não é constante no espaço. Em geral, a região de estudo será grande o suficiente para que o risco não seja considerado constante. Situações em que a hipótese de risco constante é de interesse são comuns na situação de testes de conglomerados onde as regiões de estudo são pequenas e homogêneas. Este assunto será tratado em outro capítulo deste livro.

2.1 Estimadores e parâmetros

O modelo mais realista é aquele em que os riscos relativos variam no espaço. Isto é, y_i possui distribuição de Poisson com valor esperado $\mu_i = \theta_i E_i$. O parâmetro θ_i varia de área para área e quantifica a diferença do risco da área i relativamente à média da região global ou relativamente ao valor esperado E_i. Um valor $\theta_i = 1.5$, por exemplo, indicaria que a área i possui um risco relativo 50% maior do que a média do mapa. Um valor menor que 1, como por exemplo $\theta_i = 0.75$, indica um risco 25% menor relativamente à média do mapa da região global analisada.

É claro que a estatística $SMR_i = y_i/E_i$ é uma estimativa do parâmetro θ_i. De fato, o estimador de máxima verossimilhança $\widehat{\theta_i}$ do risco relativo θ_i é a SMR_i se assumimos que as contagens y_i são variáveis aleatórias independentes com distribuição Poisson$(\theta_i E_i)$ (*Mostre isto como exercício*). Sob estas condições, o estimador $\widehat{\theta_i} = SMR_i = y_i/E_i$ possui propriedades ótimas pois é o estimador não viciado uniformemente de mínima variância (*Mostre isto como exercício*).

Caso o modelo use as taxas subjacentes como parâmetros, temos resultados inteiramente análogos. Isto é, se as contagens y_i são variáveis aleatórias independentes com distribuição Poisson$(\theta_i n_i)$ onde n_i é a população sob risco na área i, então $\widehat{\theta_i} = y_i/n_i$ é o estimador de máxima verossimilhança de θ_i e é também o estimador não viciado uniformemente de mínima variância.

Considerando novamente o modelo de riscos relativos com $y_i \sim$Poisson$(\theta_i E_i)$, o estimador $\widehat{\theta_i} = y_i/E_i$ possui variância inversamente proporcional ao número esperado de eventos (*Mostre*

isto como exercício) . Assim, quando este número esperado de eventos for pequeno, a variabilidade do estimador do risco relativo pode ser muito grande. Esta situação ocorre com bastante freqüência com dados epidemiológicos e demográficos quando as unidades espaciais de análise são municípios ou unidades ainda menores. Neste caso, as populações destas pequenas áreas serão, na maioria dos casos, bastante pequena e vai gerar um valor de E_i bastante pequeno. A conseqüência é que os valores extremos de $\widehat{\theta_i}$ tendem a ocorrer nas áreas com pequenas populações. Isto é, o que mais chama a atenção num mapa, os seus valores extremos, é o menos precisamente estimado. As maiores oscilações do risco relativo não estarão, em geral, associadas com variações no risco subjacente que as populações sofrem mas serão apenas flutuação aleatória casual.

Este problema afeta também uma análise baseada nas taxas empíricas $r_i = y_i/E_i$, ao invés das estimativas $\widehat{\theta_i}$ do risco relativo. Ao trabalharmos com as taxas simples, assumimos um modelo simples em que a padronização por idade e sexo não é efetuada e assim y_i possui distribuição com valor esperado $\theta_i N_i$ onde θ_i é a taxa subjacente da área i estimada por r_i. Ela possui variância inversamente proporcional ao tamanho da população N_i. Assim, locais com pequenas populações terão grande variabilidade em suas taxas. Uma pequena diferença no número de casos observados poderá levar a grande variação nos valores das taxas.

2.2 Exemplos

Num artigo clássico e pioneiro, Choynowski (1959) considera a ocorrência de tumores cerebrais em condados poloneses. Ele apontou o problema nos seguintes termos: "Tendo construído um mapa que mostrava as taxas de tumores no cérebro em sessenta condados (poviat) do sul da Polônia, ... verifiquei que algumas das taxas eram muito altas ou muito baixas em comparação com a média da área total, igual a 5,17 por 100 mil habitantes. Como essas grandes irregularidades geográficas eram bastante surpreendentes, estudei os dados detidamente e notei que os condados que se desviavam (muito da média) sem qualquer explicação aparente (tais como qualidade de cuidado médico, diferenças na composição etária, etc) tinham populações pequenas. Conseqüentemente, mesmo uma pequena diferença em freqüências absolutas criava uma diferença substancial nas taxas. Isto é, os desvios poderiam muito bem ser atribuídos a variações amostrais."

Considerando os dados, notamos que a maior taxa, correspondente ao condado de Lesko, está associada à menor população (17.000 habitantes). A ocorrência de dois casos nesse condado é responsável pela taxa de 11,77 por 100 mil. Caso tivesse ocorrido apenas um caso, essa taxa baixaria para 5,9 por 100 mil, um valor consistente com as taxas dos outros condados. Por outro lado, a ocorrência de três casos faria a taxa pular para 17,7 por 100 mil.

Esse efeito drástico que a adição ou subtração de um ou dois casos acarreta na taxa não se verifica naqueles condados com grandes populações. Por exemplo, considere Gorlice que possui a segunda maior taxa, igual a 10,8 por 100 mil. Se um caso é subtraído ou adicionado, a taxa muda para 9,6 ou 12,0 por 100 mil, respectivamente. Populações moderadas teriam feito intermediário como, por exemplo, Przeworsk que teria taxas iguais a 1,8 e 5,3 no caso da subtração ou adição de um caso adicional.

É importante enfatizar que a caracterização do que constitui uma população grande ou pequena depende do risco subjacente. Uma mesma população pode ser considerada pequena para calcular uma taxa associada com um risco muito baixo e considerada grande para um risco comum. Suponha,

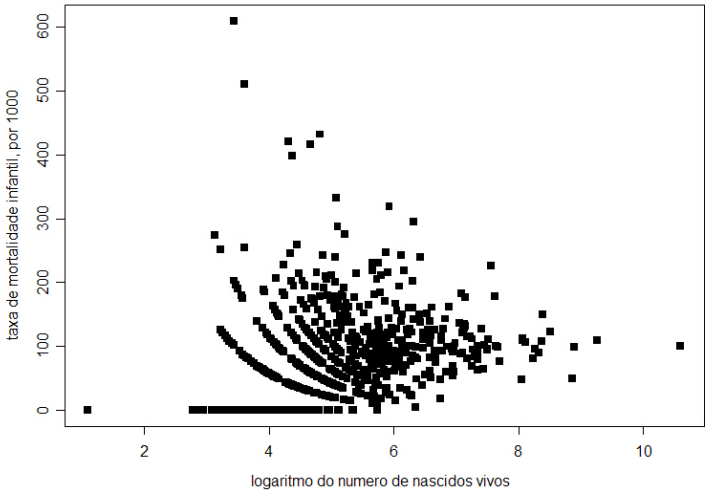

Figure 2: Taxa de mortalidade infantil (por 1000) versus o logaritmo do número de nascidos vivos em municípios mineiros em 1994. Cada ponto representa um município.

por exemplo, que o interesse reside na razão de sexos ou, equivalentemente, na proporção de pessoas do sexo masculino numa população. Uma pessoa escolhida ao acaso possui uma chance ("risco") aproximadamente igual a 0,5 de ser homem. Uma taxa igual a proporção de homens na população vai variar muito pouco se o tamanho da população for 500 ou mais. No entanto, essa população de 500 é claramente inadequada para calcular a taxa de mortalidade por câncer de pulmão, devido à grande variabilidade da taxa resultante.

O problema portanto é que a variação das taxas é muito grande quando a população é pequena para o risco que está sob consideração. Esse problema está presente no exemplo de mortalidade infantil em Minas Gerais apresentado a seguir.

A Figura 2 mostra as taxas de mortalidade infantil (por 1000) em municípios de Minas Gerais em 1994. A mortalidade infantil é uma taxa cujo numerador é o número de crianças que morreram com menos de um ano de idade num certo período dividido pelo número de crianças nascidas vivas no mesmo período. Em MG, encontramos uma taxa global de 31,8 por 1000. No gráfico, o eixo horizontal é o logaritmo do número de nascidos vivos e casa ponto representa um município. Observa-se a nítida forma de um funil com a variação das taxas sendo muito maior nos municípios menores. Embora parte desta variação possa estar associada com riscos muito diferentes em municípios pequenos, nenhum epidemiológo acreditaria em taxas neste período tão altas quanto 200 mortes a casa 1000. A enorme variação das taxas quando o número de nascidos vivos é pequeno é devido a pura flutuação aleatória. Para se convencer disto, considere os 15 municípios que tiveram 0 mortes e menos que 30 nascidos vivos. Se uma única morte tivesse sido registrada nestes municípios, as taxas passariam de 0 (o menor valor possível) para valores entre 37 e 63, todos, exceto

um, acima do terceiro quartil.

Para solucionar este problema, algumas opções podem ser consideradas. A primeira é agregar áreas para formar áreas maiores para análise. A desvantagem disto é perder a informação localizada. Outra opção foi sugerida por Choynowski (1959) e é baseada em mapas de probabilidades, como descreveremos na próxima seção. Esta solução vem caindo em desuso pois apresenta certas inconveniências e porque vem sendo superada por outras soluções. O restante do capítulo é dedicado a uma estratégia melhor, a de procurar estimar melhor risco localizado de cada área. Pode-se obter grande redução do problema da variabilidade de pequenas áreas utilizando abordagens bayesianas.

3 Mapas de Probabilidades (leitura opcional)

Para evitar o problema da grande variabilidade de taxas baseadas em pequenas populações, Choynowski (1959) propõe fazer mapas substituindo as taxas por probabilidades similares ao P-valor de um teste. Um mapa temático baseado nessa proposta é chamado de mapa de probabilidade. Ao invés de mapear a taxa, faze-se um mapa da probabilidade de obter uma contagem que é mais extrema que aquela de fato observado, sob a hipótese de que o risco é constante na região.

Suponha que o número observado de casos y_i possui distribuição de Poisson com valor esperado E_i significando um risco constante no espaço em cada classe de idade-sexo. Seja X uma variável aleatória com a mesma distribuição que y_i. Isto é, $X \sim$Poisson(E_i). Defina então:

$$\rho_i = \begin{cases} P(X \geq y_i) \text{ se } y_i \geq E_i \\ P(X \leq y_i) \text{ se } y_i < E_i \end{cases}$$

Assim, uma valor de ρ_i muito próximo de zero indica que a taxa é muito alta ou muito baixa (relativamente ao valor esperado). A vantagem deste método é que ele substitui a taxa por uma medida que leva em conta a natureza estocástica da contagem considerando sua variabilidade. Deste modo, ele evita atribuir significado a taxas muito extremas mas baseadas em pequenos excessos aleatórios casuais em pequenas populações e que são compatíveis com a hipótese de que o risco naquela área é igual ao esperado sob hipótese de risco constante no espaço.

Entretanto, esta medida sofre do problema contrário ao apresentado anteriormente. Seus valores serão muito próximos de zero caso a população seja muito grande. De fato, do mesmo modo que um valor p-valor, ρ_i é influenciado pelo tamanho da população de risco. Se E_i é grande, vamos ter pequenas diferenças entre y_i e E_i causando um valor $\rho_i \approx 0$. Por exemplo, com $E_i = 1000$, se $y_i > 1052$ então $\rho_i < 0.05$. No entanto, a diferença entre esses valores é de apenas 5,2% do valor esperado. De forma um pouco mais geral, suponha que $y_i - E_i = 1,05E_i$. Usando o teorema Central do Limite, temos

$$\rho_i = P\left(X \geq y_i\right) = P\left(\frac{X - E_i}{\sqrt{E_i}} > \frac{y_i - E_i}{\sqrt{E_i}}\right) \approx P\left(N(0,1) > \frac{1.05E_i}{\sqrt{E_i}}\right) \longrightarrow 0$$

quando $E_i \longrightarrow \infty$.

Embora a idéia dos mapas de probabilidades seja permitir comparações por meio da padronização das taxas em uma escala de probabilidade, isto não é possível se algumas populações de risco são

grande pois então teremos valores extremos da probabilidade como conseqüência do poder do teste para detectar pequenos afastamentos de hipóteses feitas para o cálculo das probabilidades ρ_i.

No entanto, a principal desvantagem deste método é supor que as taxas de cada local deveriam ser iguais a uma taxa global constante em todo o mapa. Isto está implícito no cálculo dos ρ_i's quando assume-se que a distribuição de X possui média E_i. Isto não será razoável num mapa de uma área extensa ou com grande variação no risco tais como mapas de estados brasileiros divididos em municípios ou mesmo mapas de metrópoles divididas em bairros.

Um importante problema adicional é que eles ignoram o valor da taxa de modo que, num mapa, duas áreas com a mesma podem ter ρ_i's muito diferentes e as áreas mais extremas podem ser simplesmente aquelas com as maiores populações.

Outra desvantagem desses mapas de probabilidade é que eles não levam em conta as esperadas similaridades entre áreas contíguas. Estas similaridades são devido a usual variação suave do risco sobre área sendo mapeada. Incorporar esta informação nas estimativas de risco pode levar a mapas com menos flutuação aleatória e assim a uma diferenciação mais precisa entre o que é de fato risco elevado e o que é flutuação aleatória causada por pequenas populações ou grande potência de detectar diferenças substantivamente desprezíveis.

O melhor é não fazer os mapas de probabilidades baseados nas medidas ρ_i. No entanto, estas medidas são facilmente calculáveis. Assim, para amenizar estas desvantagens e ainda usar as medidas ρ_i, pode-se proceder em duas etapas ao fazer mapas de estimativas de riscos relativos ou taxas. Primeiramente, considerando-se apenas o conhecimento do problema e dos valores típicos ou toleráveis de uma doença, responde-se à questão: A taxa r_i (ou SMR_i) é muito alta/baixa ou pode ser consideradas típica/aceitável? Se for típica, nada a fazer. Se alta (ou baixa), responde-se: O valor ρ_i é próximo de 0? Se a resposta for não, nada a fazer. Se for sim, então esta taxa r_i é alta (ou baixa) e possui pequena variância podendo ser considerada uma área de risco elevado (ou mais baixo que o típico da doença).

FIGURA 3.3 ?????

Para utilizar este procedimento de duas etapas no exemplo da mortalidade infantil em Minas Gerais, considere o gráfico da Figura 2 que mostra a taxa de mortalidade no eixo horizontal e a probabilidade ρ_i no eixo vertical. Valores de mortalidade infantil maiores que 50 mil seriam motivo de preocupação, do ponto de vista da saúde pública. Municípios com estes valores são aqueles a direita da linha vertical. Por outro lado, apenas municípios com ρ_i pequenos, digamos abaixo de 0,05, deveriam ser motivo de preocupação e estes são aqueles abaixo da linha horizontal. Vemos que quase todos os valores de taxas muito elevadas continuam aparecem abaixo da linha 0,05. No entanto, veremos mais tarde com os métodos bayesianos que este quadro não permanece assim. Com dissemos antes, um dos principais defeitos da metodologia baseada nos ρ_i é supor um risco constante no espaço, o que não é verdade no caso de mortalidade infantil em MG.

Para superar estas dificuldades, métodos Bayesianos empíricos ou inteiramente Bayesianos têm sido propostos na literatura. Estes métodos, ao estimar o risco de uma pequena área, têm como idéia central o uso de informação das outras áreas que compõem a região de estudo. Deste modo, diminui-se o erro quadrático médio total da estimação dos riscos. O método bayesiano empírico é apresentado neste capítulo enquanto que o método inteiramente bayesiano é apresentado no próximo capítulo. Atualmente, o método inteiramente bayesiano é preferível por levar em conta

toca a variabilidade presente nos dados e por permitir inferência muito mais ricas em modelos mais sofisticados.

4 Taxas e riscos relativos como variáveis aleatórias

Considere uma região particionada em n áreas indexadas por $i = 1, \ldots, n$. Suponha que a taxa anual de eventos (desconhecida) em cada área é denotada por θ_i e que $r_i = y_i/n_i$ é a taxa observada, onde y_i é o número de eventos na área i e n_i é o número de pessoas em risco. É comum supor que, dada a taxa desconhecida θ_i, a contagem y_i do número de casos possui distribuição de probabilidade de Poisson com esperança (condicional em θ_i) igual a $E(y_i|\theta_i) = n_i\theta_i$ e variância condicional $Var(y_i|\theta_i) = n_i\theta_i$. Num contexto não-bayesiano, quando y_i possui distribuição Poisson ou binomial, a taxa r_i é a melhor estimativa de θ_i no sentido de minimizar o erro médio de estimação $E\left(\theta_i - \widehat{\theta}_i\right)^2$ dentre os estimadores não viciados de θ_i. Esta esperança é tomada com respeito à densidade de probabilidade das observações y_i. Pouco pode ser feito para obter estimativas melhores num contexto como este. Já vimos nos exemplos anteriores que estes estimadores podem até ser ótimos mas não são bons o suficiente se as áreas possume populações pequenas. A única saída é impor mais estrutura ao problema e com isto possivelmente obter estimadores melhores.

Parece razoável supor que as diferenças entre as taxas desconhecidas θ_i possuem uma regularidade ou homogeneidade derivada dos processos sociais e ambientais subjacentes que afetam a região em estudo. Essas regularidade nos permite modelar as variações nos θ_i através de uma distribuição de probabilidade. Este modelo para as taxas desconhecidas θ_i reconhece que elas não são números totalmente arbitrários, completamente desconectados, mas que guardam certa relação entre si. Por exemplo, para um dado fenômeno numa região, as taxas de cada sub-área provavelmente vão estar concentradas numa faixa de valores com alguns poucos valores possivelmente mais afastados, muito superiores ou muito inferiores em relação ao comportamento típico das taxas subjacentes e desconhecidas θ_i .

Uma abordagem bayesiana é a mais adequada e flexível para incorporar esta idéia intuitivamente simples e facilmente aceitável. Esta abordagem, na sua versão orientada para análise de dados, assume que nosso conhecimento ou falta de conhecimento (incerteza) acerca das taxas fixas e desconhecidas $\theta_i, i = 1, \ldots, n$, pode ser representado por uma distribuição de probabilidade. Mais especificamente, os valores desconhecidos e fixos $\theta_i, i = 1, \ldots, n$, seriam realizações de variáveis aleatórias $\Theta_i, i = 1, \ldots, n$, com certa distribuição conjunta. O objetivo é atualizar nosso conhecimento acerca destas quantidades desconhecidas após observar os dados. É comum usar uma notação um pouco mais ambígua denotando ambas, tanto a variável aleatória Θ_i quanto o seu valor observado θ_i, por um único símbolo, que será θ_i neste livro.

O caso mais simples supõe que $\theta_i, i = 1 \ldots, n$, são independentes e possuem uma distribuição de probabilidade comum com valor esperado μ_θ e variância σ^2_θ. É útil imaginar cada área i escolhendo sua taxa de forma aleatória a partir de uma distribuição comum. Desse modo, pode-se interpretar μ_θ como sendo o valor esperado da taxa de uma área escolhida ao acaso na região.

A distribuição comum das taxas poderia ser bastante complicada com várias modas e sem uma expressão matemática conhecida. No entanto, na prática, escolhe-se uma distribuição simples.

Por exemplo, as taxas θ_i poderiam estar sendo geradas a partir de uma distribuição gama com parâmetros α e β, denotada por $\Gamma(\alpha, \beta)$. Na Figura 3, o gráfico A à esquerda mostra a densidade de probabilidade de uma gama com parâmetros $\alpha = 2$ e $\beta = 10$. A distribuição gama $\Gamma(\alpha, \beta)$ possui valor esperado $\mu_\theta = \alpha/\beta$ e variância $\sigma^2_\theta = \alpha/\beta^2$. Para o exemplo, temos $\mu_\theta = 0,2$ e variância $\sigma^2_\theta = 0,02$ ou $\sigma_\theta = 0,14$.

No caso dos estados de Minas Gerais, dificilmente esperaríamos taxas θ_i iguais a 0,2 per capita (ou 200 por mil) em alguns municípios. Isto implicaria que, nesses municípios, uma criança em cada cinco morreria antes de completar um ano de vida. Estas taxas são inaceitavelmente altas e dificilmente verossímisseis no quadro de saúde pública brasileira nos tempos atuais. Uma mortalidade infantil de 20% ou próxima disso seria uma verdadeira calamidade pública. Como a distribuição da Figura 3A atribui massa de probabilidade igual a 0,406 para valores acima de 0,2, ela seria uma escolha muito grosseira para representar nosso conhecimento da mortalidade infantil em Minas Gerais. Após um pouco de tentativa e erro na escolha dos parâmetros da distribuição Gama, fazendo gráficos de densidade de probabilidades, calculando probabilidades diversas e auxiliado por discussões com epidemiológos, somos levados a uma distribuição muito mais realista, a de uma gama de parâmetros $\alpha = 1,2$ e $\beta = 24$. Esta distribuição possui valor esperado 0,05 e desvio padrão 0,046 com densidade representada no gráfico da Figura 3B. Nesta nova distribuição, a chance de observarmos um valor tão extremo quanto 0,2 é apenas 0,013, não impossível mas razoavelmente improvável. Vamos voltar a discutir várias vezes a questão da escolha de uma priori apropriada.

5 Inferência Bayesiana

Vamos fazer uma breve revisão dos principais aspectos de inferência bayesiana necessários para prosseguir com o estudo de mapas de taxas e proporções. Nosso texto é muito resumido e parcial. Recomendamos os excelentes livros de Gelman et al e Carlin e Louis para maiores detalhes. Na inferência Bayesiana, assim como na inferência clássica, num dado problema específico, os possíveis valores dos parâmetros $\theta_i, i = 1, \ldots, n$, pertencem a um certo conjunto do espaço euclidiano, chamado *espaço paramétrico*. Por exemplo, no caso da moratlidade infantil em MG no ano de 1994, os verdadeiros valores das taxas $\theta_i, i = 1, \ldots, n$, de mortalidade infantil (per capita) são desconhecidas e pertecem e ao intervalo $[0, \infty)$. Entretanto, embora todo valor em $[0, \infty)$ seja *possível*, eles não são igualmente *prováveis*. Na verdade, a chance de que algum dos θ_i pertença ao intervalo $[0.5, \infty)$ é praticamente desprezível. Mesmo no intervalo $[0, 0.5)$, existe conhecimento suficiente para acreditarmos firmemente que há muita diferença entre seus subintervalos. Este é um conhecimento que possuimos antes de observar qualquer dado no problema específico que iremos analisar. Ele vem de conhecimentos teóricos sobre o problema, da observação empírica de problemas semelhantes no passado na mesma região ou em regiões diferentes. De qualquer modo, é um conhecimento *prévio* à observação dos dados específicos do problema a ser analisado e ele é expresso numa *distribuição de probabilidade* para os possíveis valores de θ_i. Esta distribuição é chamada de *distribuição (de probabilidade) a priori* sobre os possíveis valores de θ_i e a densidade de probabilidades é denotada por $p(\theta_1, \ldots, \theta_n)$. Caso a distribuição seja discreta, teremos a função de probabilidade ao invés de uma densidade. Usaremos a mesma notação $p(\theta_1, \ldots, \theta_n)$ para representar as duas coisas, uma densidade ou uma função de probabilidade.

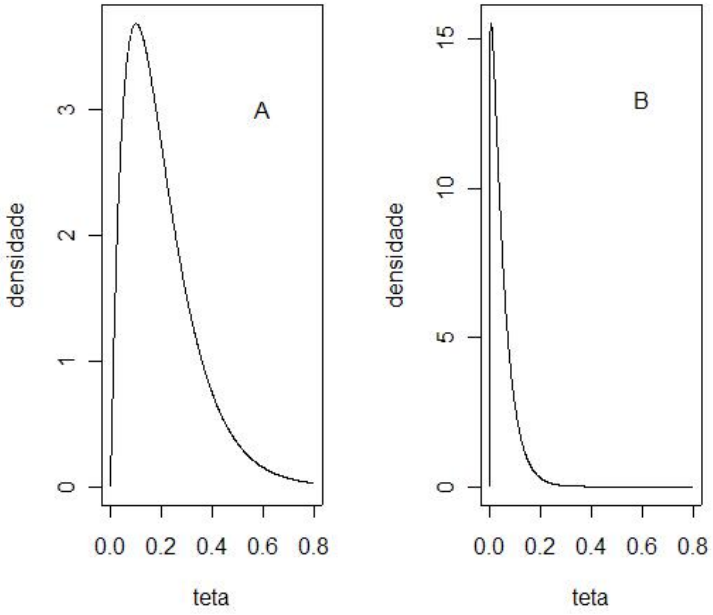

Figure 3: A: Densidade de probabilidade de uma distribuição Gama (2,10). B: Densidade de probabilidade uma distribuição Gama (1.2,24).

Observe que, quando mais de um parâmetro está presente no problema, esta distribuição a priori é uma distribuição *multivariada* para todos parâmetros $\theta_1, \ldots, \theta_n$. Por exemplo, ela poderia ser uma distribuição multivariada normal n-dimensional. Outra possibilidade que simplifica muito vários problemas (mas nem sempre é muito realista) é assumir que, a priori, os parâmetros $\theta_1, \ldots, \theta_n$ são i.i.d. Neste caso, as distribuições marginais são idênticas e a distribuição conjunta é o produto das marginais:

$$p(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} p(\theta_i)$$

Assim, basta declarar qual é a (única) distribuição de probablidade marginal $p(\theta)$.

Por exemplo, no caso da mortalidade infantil em Minas Gerais, podemos assumir que, a priori, as taxas $\theta_1, \ldots, \theta_n$ são independentes e identicamente distribuídas com distriubuição comum gama com parâmetros 1.2 e 24. Alternativamente, poderíamos expressar um conhecimento mais refinado assumindo que certas regiões de Minas possuem taxas de mortalidade infantil mais altas que outras. Isto poderia ser feito, por exemplo, com uma distribuição de probabilidade conjunta composta por variáveis aleatórias θ_i que não são identicamente distribuídas ou que não são independentes entre si. Veremos mais sobre este tipo de priori no próximo capítulo.

As observações ou dados estocásticos $y_1, \ldots, y_n$ possuem uma distribuição de probabilidade que depende dos parametros desconhecidos. Dados os valores dos parâmetros $\theta_1, \ldots, \theta_n$, as observações $y_1, \ldots, y_n$ possuem distribuição $p(y_1, \ldots, y_n|\theta_1, \ldots, \theta_n)$. Vista como função dos parâmetros esta função é conhecida como a *função de verossimilhança* dos parâmetros. É comum assumir que as observações são independentes condicionalmente nos parâmetros e que a distribuição de y_i depende apenas de θ_i e não dos demais parâmetros. Isto é, será comum assumir que $p(y_1, \ldots, y_n|\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} p(y_i|\theta_i)$.

No caso da mortalidade infantil em Minas Gerais, podemos assumir que, dados os parâmetros $\theta_1, \ldots, \theta_n$, as contagens $y_1, \ldots, y_n$ são variáveis aleatórias independentes de Poisson com valor esperado $E(y_i|\theta_i) = E_i\theta_i$.

A inferência bayesiana é baseada na atualização da distribuição a priori dos parâmetros após obseravarmos os dados. Esta atualização é feita através da Regra de Bayes e a distribuição resultante é chamada de distribuição *a posteriori*. Como recordação de probabilidade elementar, a probabilidade de ocorrer o evento A dado que ocorreu o evento B é dada por $P(B|A) = P(B \cap A)/P(A)$, quando $P(B) > 0$. Suponha agora que o espaço amostral é particionado em eventos $B_1, \ldots, B_k$ tais que $B_i \cap B_j = \emptyset$ e a união $B_1 \cup B_2 \cup \ldots \cup B_k$ é o espaço amostral. Além disso, são conhecidas a priori as probabilidades $P(B_i)$ para todo i. Também são conhecidas as probabilidades $P(A|B_i)$ para todo i. Após observar que o evento A ocorreu, deseja-se atualizar a probabilidade de que o evento B_i possa ocorrer para cada um dos $B_i, i = 1, \ldots k$. Para isto, usa-se a regra de Bayes:

$$P(B_i|A) = \frac{P(B_i \cap A)}{P(A)} = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)}$$

Apesar do que parece indicar a expressão do denominador na última fração acima, observe que o denominador é apenas $P(A)$ e não depende do evento específico B_i.

A regra de Bayes vale também para manipular densidades de probabilidade (ou funções de probabilidade no caso de variáveis aleatórias discretas), e não apenas eventos. Assim, podemos escrever que a distribuição a posteriori de $\theta_1, \ldots, \theta_n$, dadas as observações $y_1, \ldots, y_n$, é igual a

$$\begin{aligned} p(\theta_1, \ldots, \theta_n | y_1, \ldots, y_n) &= \frac{p(y_1, \ldots, y_n | \theta_1, \ldots, \theta_n) p(\theta_1, \ldots, \theta_n)}{p(y_1, \ldots, y_n)} \\ &= \frac{p(y_1, \ldots, y_n | \theta_1, \ldots, \theta_n) p(\theta_1, \ldots, \theta_n)}{\int \ldots \int p(y_1, \ldots, y_n | \theta_1, \ldots, \theta_n) p(\theta_1, \ldots, \theta_n) d\theta_1 d\theta_2 \ldots d\theta_n} \end{aligned}$$

Como antes, o denominador não depende dos valores específicos $\theta_1, \ldots, \theta_n$ nos quais a densidade a posteriori está sendo avaliada. Estes pontos ficarão mais claros no exemplo a seguir.

Suponha que existam apenas $n = 5$ áreas e que $\theta_1, \ldots, \theta_5$ são parâmetros fixos e desconhecidos. Para representar nosso conhecimento sobre esses parâmetros nós adotamos um modelo onde $\theta_1, \ldots, \theta_5$ são realizações de variáveis aleatórias i.i.d. com distribuição Gama com parâmetros $\alpha = 3$ e $\beta = 100$. A Figura 3 mostra a função de densidade de probabilidade dessa distribuição comum dos valores desconhecidos $\theta_1, \ldots, \theta_5$. Condicionalmente aos valores de θ, os dados y possuem distribuição Poisson. Mais especificamente, dados os valores $\theta_1, \ldots, \theta_5$, temos $y_i|\theta_i \sim \text{Poisson}(n_i\theta_i)$ onde n_i á a população de risco da área i. Assim, θ_i é o risco (per capita) de ocorrer o evento de interesse na área i. Para sermos mais específicos ainda, vamos considerar $n_1 = 24$, $n_2 = 78$, $n_3 = 143$, $n_4 = 323$, $n_5 = 39591$. Estes números correspondem ao número de crianças nascidas vivas em 1995 no municípios mineiros de Fama, Ijaci, Fronteira, Paraguaçu e Belo Horizonte, respectivamente. Neste mesmo ano e nestes mesmos municípios, foram observadas as seguintes contagens para o número de crianças que morreram antes de completar um ano de idade: $y_1 = 0, y_2 = 1, y_3 = 0, y_4 = 6, y_5 = 1262$.

Um cálculo simples de probabilidade usando o Teorema de Bayes mostra que, após observar as contagens acima, a distribuição para os parâmetros desconhecidos $\theta_1, \ldots, \theta_5$ torna-se a de variáveis aleatórias independentes (dadas as observações) com distribuição $(\theta_i|y_i) \sim Gama\,(\alpha + y_i, \beta + n_i)$. Para verificar isto, vamos considerar inicialmente apenas o segundo município com o parâmetro desconhecido $\theta_2 \sim Gama(3, 100)$ e $y_2|\theta_2 \sim \text{Poisson}(78\ \theta_2)$. O valor realmente observado de y_2 foi 1. Utilizando a regra de Bayes, a distribuição a posteriori de θ_2 é igual a

$$\begin{aligned} p(\theta_2|y_2 &= 0) = \frac{p(y_2 = 1|\theta_2)p(\theta_2)}{p(y_2 = 1)} = \left(\frac{(78\theta_2)^1 \exp(-78\theta_1)}{1!} \frac{100^3}{\Gamma(3)} \theta_2^{3-1} \exp(-100\theta_2) \right) \div p(y_2 = 1) \\ &= 78(100)^3/\Gamma(3)\ \ \theta_2^{3+1-1} \exp(-178\theta_2) \div p(y_2 = 1) \end{aligned}$$

O numerador é sempre simples de ser calculado pois é apenas o produto da verossimilhança pela distribuição a priori. O denominador é um pouco mais difícil de ser obtido:

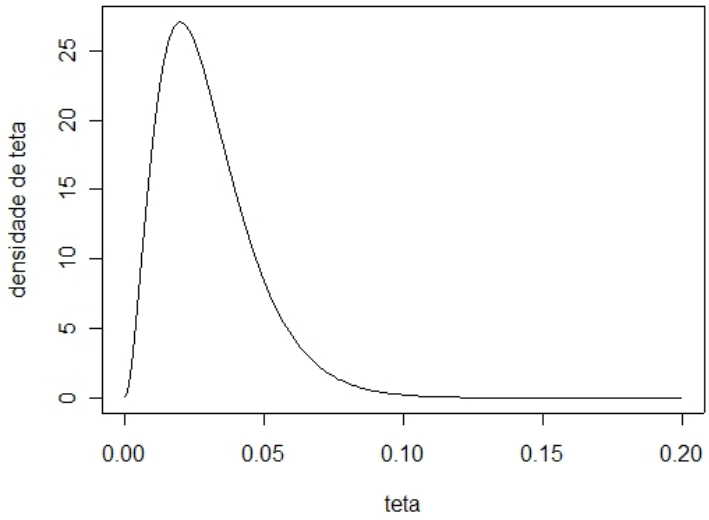

Figure 4: Função densidade de probabilidade de uma distribuição gama com parâmetros $\alpha = 3$ e $\beta = 100$.

$$
\begin{aligned}
p(y_2 &= 1) = \int p(y_2 = 1|\theta_2)p(\theta_2)d\theta_2 = \int_0^\infty \frac{(78\theta_2)^1 \exp(-78\theta_2)}{1!}\frac{100^3}{\Gamma(3)}\theta^{3-1}\exp(-100\theta_2)d\theta_2 \\
&= \frac{78\times 100^3}{\Gamma(3)}\int_0^\infty \theta_2^{4-1}\exp(-178\theta_2)d\theta_2 \\
&= \frac{78\times 100^3}{\Gamma(3)}\frac{\Gamma(4)}{178^4}\int_0^\infty \frac{178^4}{\Gamma(4)}\theta_2^{4-1}\exp(-178\theta_2)d\theta_2 \\
&= \frac{78\times 100^3}{\Gamma(3)}\frac{\Gamma(4)}{178^4}\times 1 \approx 0.233
\end{aligned}
$$

sendo que a integral da penúltima linha é igual a 1 por ser a densidade de uma gama. Observe que este denominador depende apenas do valor observado de y_2, ele é constante em θ_2 ou, equivalentemente, não depende de θ_2. Assim, a densidade a posteriori fica igual a

$$
p(\theta_2|y_2 = 1) = \left(78(100)^3/(0.233\Gamma(3))\right)\theta_2^{4-1}\exp(-178\theta_2)
$$

Esta é a densidade de uma distribuição gama com parâmetros 4 e 178. Uma forma mais simples de reconhecer isto (e *quase sempre mais fácil de ser usada*)é verificar que o numerador é proporcional a $\theta^{4-1}\exp(-178\theta_1)$. A menos de uma constante, isto é proporcional à expressão da densidade de uma distribuição Gama com parâmetros $\alpha + y_2 = 3 + 1$ e $\beta + n_2 = 100 + 78$. Como a cosntante de integração de uma densidade é obtida fazendo-se com que a integral da densidade na reta seja igual a 1, isto implica que a constante da densidade a posteriori tem de ser a mesma constante de integração de uma gama com parâmetros 4 e 178. Assim, a regra simples para obter a porteriori é: olhe para o numerador da regra de Bayes ignorando todas as constantes. Se este numerador for igual (a menos de constnates) à expressão de uma densidade conhecida (normal, beta, gama, Pareto, etc.) então a densidade a posteriori tem de ser esta densidade reconhecida. Infelizmente, em modelos mais complicados, não conseguiremos aplicar esta regra pois, em geral, não conseguiremos reconhecer uma densidade a partir do numerador da regra de Bayes. Outras técnicas numéricas serão necessárias nesses casos. De qualquer modo, é sempre muito útil reconhecer que o denominador na regra de Bayes não envolve o parâmetro descohecido e assim a função de distribuição a porteriori pode sempre ser considerada como proporcional ao produto da verossimilhança pela priori:

$$
p(\theta_2|y_2 = 1) = \frac{p(y_2 = 1|\theta_2)p(\theta_2)}{p(y_2 = 1)} = Cp(y_2 = 1|\theta_2)p(\theta_2) \propto p(y_2 = 1|\theta_2)p(\theta_2)
$$

O mesmo cálculo feito para o município acima pode ser feito de forma geral: se $\theta \sim Gama(\alpha, \beta)$ e $(y|\theta) \sim Poisson(n\theta)$ então $(\theta|y = k) \sim Gama(\alpha + k, \beta + n)$. De fato, nós temos

$$
\begin{aligned}
p(\theta|y &= k) = \frac{p(y = k|\theta)p(\theta)}{p(y = k)} \\
&\propto p(y = k|\theta)p(\theta) = \frac{(n\theta)^k \exp(-n\theta)}{k!}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta) \\
&= C\theta^{\alpha+k-1}\exp(-(\beta + n)\theta)
\end{aligned}
$$

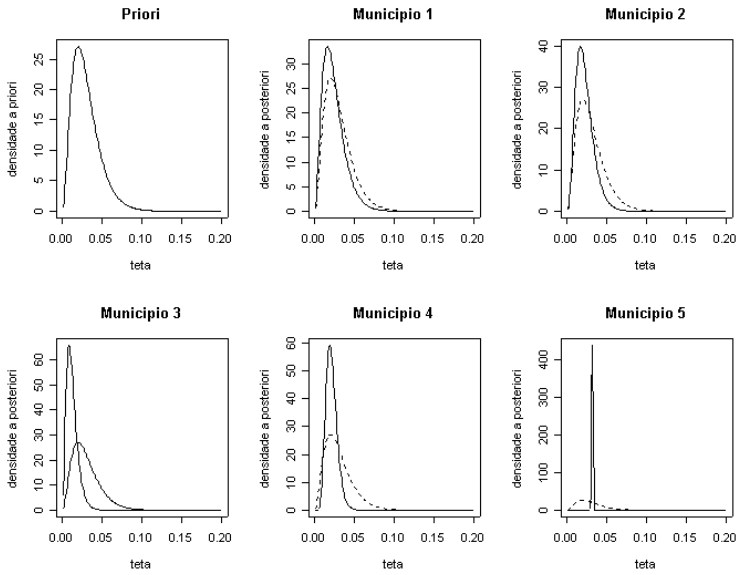

Figure 5: Gráfico com as densidade a priori da mortalidade infantil de um município em Minas Gerais e densidades a posteriori de cada um dos 5 municípios discutidos no texto.

É comum escrevermos apenas $(\theta|y) \sim Gama(\alpha+y, \beta+n)$ ao invés de $(\theta|y=k) \sim Gama(\alpha+k, \beta+n)$. Fica subentendido que y deve ser substituído pelo seu valor realmente observado. Assim, por exemplo, no caso do primeiro município, temos $(\theta|y=0) \sim Gama(3, 124)$.

O caso multivariado também é uma consequência imediata desses cálculos pois:

$$\begin{aligned} p(\theta_1, \ldots, \theta_5|y_1 &= 0, \ldots, y_5 = 1262) \propto p(y_1 = 0, \ldots, y_5 = 1262|\theta_1, \ldots, \theta_5)p(\theta_1, \ldots, \theta_5) \\ &= \prod_{i=1}^{5} \frac{(n_i\theta_i)^{y_i} \exp(-n_i\theta_i)}{y_i!} \prod_{i=1}^{5} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_i^{\alpha-1} \exp(-\beta\theta_i) \\ &\propto \prod_{i=1}^{5} \theta_i^{\alpha+y_i-1} \exp(-(\beta+n_i)\theta_i) \end{aligned}$$

e este último produto é proporcional á densidade conjunta de 5 variáveis aleatórias *independentes* com distribuição gama e parâmetros $\alpha + y_i$ e $\beta + n_i$.

A Figura 5 mostra a densidade a priori Gama(3,100) e as densidades a posteriori de cada um dos 5 municípios já discutidos. Todas são distribuições gama com parâmetros iguais a $3 + y_i$ e $100 + n_i, i = 1 \ldots, 5$.

A inferência bayesiana é toda baseada na distribuição a posteriori. A partir dela, podemos obter estimativa pontuais ou intervalares para os parâmetros ou outras informações mais complicadas. Por exemplo, quando estivermos trabalhando com todos os municípios de Minas Gerais, é provável

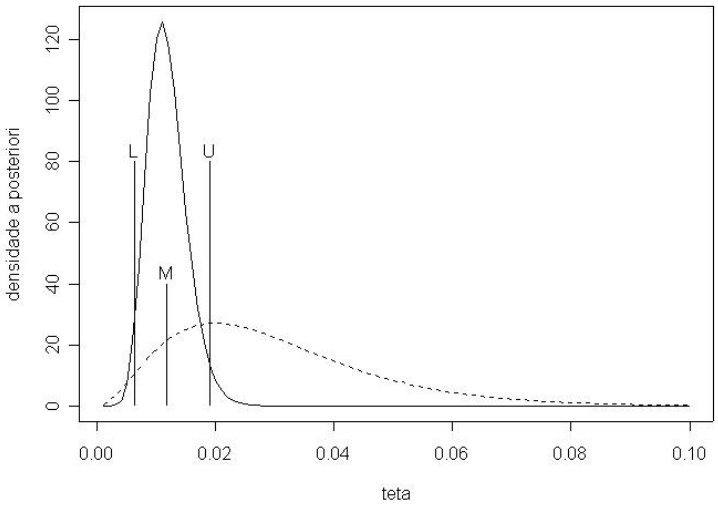

Figure 6: Densidade a priori (linha pontinlhada) Gama(3,100) e densidade a posteriori Gama(13, 1100) com linhas verticais indicando a média a posteriori (M) e os limites inferior (L) e superior (U) de um intervalo de credibilidade de 95%.

que haja interesse em resumir toda a informação representada na posteriori de cada município em apenas um único número. Um candidato natural para este número-resumo no município i é a esperança (ou média) da distribuição a posteriori $p(\theta_i|y_i)$. Outras opções poderiam ser a mediana ou a moda dessa distribuição a posteriori. Um intervalo para os prováveis valores de θ_i é o chamado *intervalo de credibilidade de 95%*: trata-se de um intervalo $[a, b]$ tal que, a posteriori, o parâmetro $\theta_i \in [a, b]$ com probabilidade 0.95. A Figura 6 mostra a densidade a priori Gama(3,100) em linha tracejada e a densidade a posteriori Gama(13, 1100) com linhas verticais indicando a média a posteriori (M) e os limites inferior (L) e superior (U) de um intervalo de credibilidade de 95%.

6 Estimador Bayesiano Ótimo

Suponha que desejamos obter um estimador pontual para cada o valor desconhecido θ_i. Qual é o melhor estimador possível? Lembre-se que um estimador de um parâmetro é simplesmente uma função dos dados observados e de constnates conhecidas (como o tamanho da amostra). Assim, qualquer estimador $\widehat{\theta_i}$ do parâmetro θ_i pode sempre ser escrito como $\widehat{\theta_i} = \widehat{\theta_i}(y_1, \ldots, y_n) = g(y_1, \ldots, y_n)$ onde g é uma função real arbitrária.

A solução para o problema de encontrar o melhor estimador de θ_i depende de qual é o critério para dizer que um estimador é bom, ruim ou ótimo. A abordagem bayesiana ou de teoria da decisão coloca este problema nos seguintes termos. Defina uma função que reflete o custo de estimar com

$\widehat{\theta}$ o verdadeiro valor desconhecido de θ. Esta função é chamada de *função de perda*. Uma das escolhas mais comuns para esta função é a função de perda quadrática dada por $\left(\widehat{\theta}-\theta\right)^2$. Na verdade, como temos vários parâmetros, queremos que o erro de estimação seja pequeno para todos eles ao mesmo tempo. A forma mais comum de definir a função de perda é então como sendo igual a $\sum_{i=1}^{n}\left(\theta_i-\widehat{\theta}_i\right)^2$. Assim, queremos um estimador que torne pequena a soma dos erros quadráticos de estimação. Como o erro de estimação (quadrático) $\sum_{i=1}^{n}\left(\theta_i-\widehat{\theta}_i\right)^2$ é uma variável aleatória, então procuramos a solução $\widehat{\theta}_i=\widehat{\theta}_i(y_1,\ldots,y_n)$ que minimize o *valor esperado* de $\sum_{i=1}^{n}\left(\theta_i-\widehat{\theta}_i\right)^2$. Isto é, queremos estimadores $\widehat{\theta_1},\ldots,\widehat{\theta_n}$ que minimizem $E\left[\sum_{i=1}^{n}\left(\theta_i-\widehat{\theta}_i\right)^2\right]$.

Na estatística clássica, buscamos estimadores que torne mínima esta esperança considerando os parâmetros θ_i como fixos e desconhecidos. Sobre estes parâmetros sabemos apenas que eles pertencem a certo espaço paramétrico mas não temos nenhuma idéia a priori sobre onde, dentro desse espaço paramétrico, poderiam estar os verdadeiros (e desconhecidos) valores de θ_i. Tipicamente, impomos algumas restrições na classe de estimadores possíveis (por exemplo, que ele seja não-viciado) e procuramos então o estimador que minimize o erro de estimação. Uma escolha que é quese sempre feita é a de usar o estimadores de máxima verossimilhança pois, sob condições de regularidade e se a amostra é grande, eles são aproximadamente não-viciados e com o menor erro de estimação possível para um estimador não viciado. Este resultado assintótico é o que justifica a imensa popularidade dos estimadores de máxima verossimilhança. Sendo assim, o que mais podemos esperar? Podemos fazer algo ainda obter melhores estimadores que os de máxima verossmilhança?

A resposta é que não poderemos fazer nada melhor se nada mais for assumido sobre os valores dos parâmetros. Na inferência bayesiana, mesmo antes de observar os dados, nós estamos dispostos a dizer que certas regiões do espaço paramétrico possume mais probabilidade de conter os valores verdadeiros dos parâmetros que outras regiões. Suponha que este conhecimento a priori é expresso na forma de uma distribuição de probabilidade a priori para os verdadeiros valores dos parâmetros.

Então nós buscamos a solução $\widehat{\theta_1},\ldots,\widehat{\theta_n}$, que minimiza $E\left[\sum_{i=1}^{n}\left(\theta_i-\widehat{\theta}_i\right)^2\right]$ onde cada $\widehat{\theta}_i$ é uma função apenas dos dados $y_1,\ldots,y_n$. A diferença com a situação anterior é que agora a esperança é calculada considerando como variáveis aleatórias aos dois elementos, tanto as observações $y_1,\ldots,y_n$ quanto os parâmetros $\theta_1,\ldots,\theta_n$. Intuitivamente, ao invés de encontrar estimadores $\widehat{\theta_1},\ldots,\widehat{\theta_n}$ que minimizam $E\left[\sum_{i=1}^{n}\left(\theta_i-\widehat{\theta}_i\right)^2\right]$ para todo e qualquer conjunto de valores possíveis de $\theta_1,\ldots,\theta_n$, nós reconhecemos que alguns valores $\theta_1,\ldots,\theta_n$ são possíveis mas altamante improváveis. Assim, não vale a pena obrigar um estimador a ser bom mesmo nessas situações altamente improváveis. Esta é a razõa para procurar estimadores que sejam bons para todos os valores possíveis, *mas ponderados pela priori*, de $\theta_1,\ldots,\theta_n$.

Mais especificamente, queremos encontrar $\widehat{\theta_1}, \ldots, \widehat{\theta_n}$ que minimizem

$$E\left[\sum_{i=1}^{n}\left(\theta_i - \widehat{\theta_i}\right)^2\right] = \sum_{i=1}^{n} E\left(\theta_i - \widehat{\theta_i}\right)^2 = \sum_{i=1}^{n} E_y\left[E_\theta\left(\left(\theta_i - \widehat{\theta_i}\right)^2 |y\right)\right]$$

Vamos considerar uma realização arbitrária mas específica y. Vamos encontrar a solução do problema acima para esta realização arbitrária. Isto será obtido mostrando que a função $g(y) = E_\theta\left(\left(\theta_i - \widehat{\theta_i}\right)^2 |y\right)$ é minimizada se tomamos $\widehat{\theta_i} = E_\theta(\theta_i|y)$. Como esta função é mínimizada para todo valor possível de y, o valor aleatório $g(Y)$ será o estimador que buscamos. Devido ao papel importante dessa esperança condicional nas contas a seguir, vamos denotar por μ_y o valor esperado de θ_i condicionado nas observações y. Isto é, $\mu_y = E_\theta(\theta_i|y)$. Assim, podemos escrever

$$\begin{aligned} E_\theta\left(\left(\theta_i - \widehat{\theta_i}\right)^2 |y\right) &= E_\theta\left(\left(\theta_i - \mu_y + \mu_y - \widehat{\theta_i}\right)^2 |y\right) = E_\theta\left((\theta_i - \mu_y)^2 + \left(\mu_y - \widehat{\theta_i}\right)^2 + 2(\theta_i - \mu_y)\left(\mu_y - \widehat{\theta_i}\right)|y\right. \\ &= Var_\theta(\theta_i|y) + E_\theta\left(\mu_y - \widehat{\theta_i}\right)^2 + 2\left(\mu_y - \widehat{\theta_i}\right)E_\theta(\theta_i - \mu_y) \\ &= Var_\theta(\theta_i|y) + E_\theta\left(\mu_y - \widehat{\theta_i}\right)^2 \geq Var_\theta(\theta_i|y) \end{aligned}$$

Por outro lado, $E_\theta\left((\theta_i - \mu_y)^2 |y\right) = Var_\theta(\theta_i|y)$. Assim, $\widehat{\theta_i} = \mu_y = E_\theta(\theta_i|y)$ é o estimador que minimiza $E\left[\sum_{i=1}^{n}\left(\theta_i - \widehat{\theta_i}\right)^2\right]$.

6.1 Exemplo numérico

Vamos apresentar um exemplo numérico detalhado mas que será útil tam,b´me para que o leitor melhor o resultado acima. Vamos continuar considerando o exemplo das $n = 5$ áreas com $\theta_1, \ldots, \theta_5$ i.i.d. com distribuição Gama com parâmetros $\alpha = 3$ e $\beta = 100$. Além disso, dados os valores $\theta_1, \ldots, \theta_5$ dos parâmetros, temos $y_i|\theta_i \sim$ Poisson$(n_i\theta_i)$ onde n_i á a população de risco da área i correspondendo ao número de crianças nascidas vivas em 1995. Para os cinco municípios mineiros de Fama, Ijaci, Fronteira, Paraguaçu e Belo Horizonte, temos $n_1 = 24$, $n_2 = 78$, $n_3 = 143$, $n_4 = 323$, $n_5 = 39591$, respectivamente. As contagens para o número de crianças que morreram antes de completar um ano de idade foram: $y_1 = 0, y_2 = 1, y_3 = 0, y_4 = 6, y_5 = 1262$.

Já mostramos que, nestas condições, a distribuição a posteriori dos θ_i é composta por variáveis aleatórias independentes com $(\theta_i|y_i) \sim Gama(\alpha + y_i, \beta + n_i) = Gama(3 + y_i, 100 + n_i)$. Perceba que, após observar os dados, os parâmetros dessa distribuição a posetriori ficam completamente determinados. O estimador ótimo de Bayes para a perda quadrática é dado por $\widehat{\theta_i} = \mu_y = E_\theta(\theta_i|y) = (\alpha + y_i)/(\beta + n_i) = (3 + y_i)/(100 + n_i)$. Observe como este estimador bayesiano depende apenas dos dados observados e de constnates conhecidas (o tamanho n_i da população e os parâmetros conhecidos da distribuição a priori).

É instrutivo refazer para este exemplo a demosntração dada acima para o caso geral. COMPLETAR NO FUTURO ????

A Tabela ??? mostra os valores das estimativas de cada uma das cinco cidades consideradas, bem como o valor do estimador de máxima verossilhança dado por y_i/n_i. Cabe notar que o estimador ótimo de Bayes $\widehat{\theta}_i = E_\theta(\theta_i|y) = (\alpha + y_i)/(\beta + n_i)$ é um valor intermediário entre a média da distribuição a priori (α/β) e o estimador de máxima verossilhança (y_i/n_i). De fato, suponha que $y_i/n_i < \alpha/\beta$. Então $y_i/n_i < (\alpha + y_i)/(\beta + n_i) < \alpha/\beta$. Por outro lado, se $y_i/n_i > \alpha/\beta$, então $y_i/n_i > (\alpha + y_i)/(\beta + n_i) > \alpha/\beta$. Dessa forma, podemos imaginar que o estimador Bayesiano faz uma contração do estimador de máxima verossilhança y_i/n_i em direção à média α/β da distribuição a priori. A quantidade de contração está associada com o tamanho da população n_i, com contrações menores se a população é grande e contrações maiores se a população é pequena.

	Fama	Ijaci	Fronteira	Paraguaçu	Belo Horizonte
Máx. Ver.	0.000	0.013	0.000	0.019	0.032
Bayes ótimo	0.024	0.022	0.012	0.021	0.032

O gráfico à esquerda na Figura 6 mostra o resultado desse estimador bayesiano para os 756 municípios mineiros em 1994. No eixo horizontal temos os estimadores de máxima verossimilhança e no eixo vertical, os estimadores Bayesiano ótimos usando o modelo acima. Os eixos possuem a mesma escala, a reta diagonal é a linha $y = x$ e a reta horizontal marca a posição da média 0.03 da distribuição a priori. Cada ponto representa um município em Minas Gerais. Podemos observar o claro efeito de contração no eixo vertical: a nuvem de pontos está ligeiramente deitada para a direita. Veja que estimativas de máxima verossimilhança extremas são trazidas para próximo da média. Assim, estimativas de máxima verossimilhança acima de 0.10 (1 em cada 10 crianças morrendo) sâo contraídas em direção a 0.03, ficando todas abaixo de 0.10. O gráfico à direita na Figura 6 mostra que este efeito de contração é maior nos municípios menores. No eixo horizontal, temos o logaritmo do número de nascidos no município e no eixo vertical, a diferença entre as estimativas de máxima verossimilhança e as estimativas ótimas de BAyes em cada município.

7 Estimador Bayesiano Empírico

O estimador ótimo de Bayes depende de sermos capazes de obter a esperança a posteriori $\mu_y = E_\theta(\theta_i|y) = \widehat{\theta}_i$. Isto só é possível se a distribuição a priori for dada explicitamente com todos os seus parâmetros conhecidos. No caso particular da mortalidade infantil em Minas, vimos que isto se traduz em dizer quais são os valores dos parâmetros α e β da distribuição gama do parâmetro θ_i. Nós usamos nos cálculos anteriores os valores $\alpha = 3$ e $\beta = 100$, o que implicava em uma média de 0.03. Vimos também que o estimador ótimo de Bayes representa uma contração do estimador de máxima verossimilhança em direção à esta média a priori de 0.03. Em várias situações, pode ser difícil dar um valor para α e β, por várias razões. A primeira é que um mesmo pesquisador pode não estar seguro sobre sua escolha e achar que outros valores de α e β diferentes de 3 e 100 deveriam também ser considerados. Outra razão é que pode não haver conhecimento suficiente para fixar com alguma segurança quaisquer valores para α e β. Finalmente, poderá haver discordância entre pesquisadores distintos quanto aos valores que deveriam ser adotados para α e β.

Este problema levou ao surgimento do método de Bayes empírico que propõe utilizar os dados observados para estimar os parâmetros da priori. No caso da mortalidade infantil, isto implicaria em usar as observações $y_1, \ldots, y_n$ para estimar α e β, os parâmetros da priori $Gama(\alpha, \beta)$, ao invés

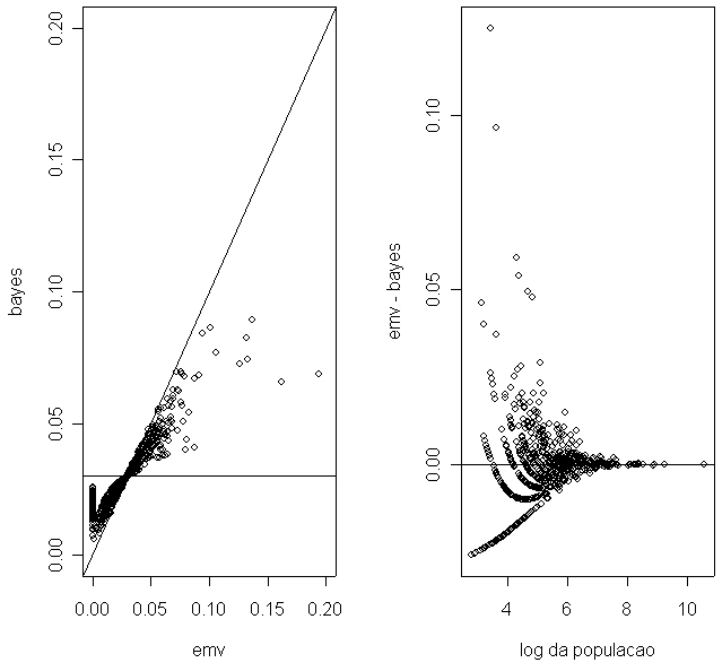

Figure 7: Direita: Diagrama de dispersão com cada ponto representando um município em MInas Gerais em 1994. No eixo horizontal, temos as estimativas de máxima verossimilhança de mortalidade infantil e no eixo vertical, temos as estimativas ótima de Bayes usando uma Gama(3,100) como priori. Esquerda: efeito de tamanho de população. NO eixo horizontal, temos o logaritmo do número de nascidos vivos e, no eixo horizontal, a diferença entre as estimativas de máxima verossimilhança e as estimativas ótima de Bayes.

de fixar estes parâmetros como $\alpha = 3$ e $\beta = 100$, como fizemos.

Para diferenciar os parâmetros nos quais temos interesse ($\theta_1, \ldots, \theta_n$, em nosso exemplo) dos parâmetros da priori (α e β, em nosso exemplo), chamamos a estes últimos de *hiperparâmetros*.

Para estimar os hiperparâmetros, em geral, o método Bayesiano empírico calcula a distribuição das observações $y_1, \ldots, y_n$ não condicionada aos parâmetros $\theta_1, \ldots, \theta_n$. Esta distribuição marginal dos dados vai depender dos hiperparâmetros. Assim, pode-se usar o método de momentos ou de máxima verossimilhança para obter estimativas para os hiperparâmetros a partir dos dados observados. Estas estimativas são então substituídas nas expressões dos estimadores ótimos de Bayes. Vamos ver como é o funcionamento do método no nosso exemplo de mortalidade infantil.

Considere apenas uma única observação inicialmente. Então, a distribuição marginal (ou não condicional nos parâmetros) da observação y é dada por

$$\begin{aligned} p(y &= k) = \int p(y=k|\theta)p(\theta)d\theta = \int_0^\infty \frac{(n\theta)^k \exp(-n\theta)}{k!}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}\exp(-\beta\theta)d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{n^k}{k!}\frac{\Gamma(\alpha+k)}{(\beta+n)^{\alpha+k}}\int_0^\infty \frac{(\beta+n)^{\alpha+k}}{\Gamma(\alpha+k)}\theta^{\alpha+k-1}\exp(-(\beta+n)\theta)d\theta \\ &= \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{n^k}{k!}\frac{\Gamma(\alpha+k)}{(\beta+n)^{\alpha+k}} \times 1 = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!}\left(\frac{\beta}{\beta+n}\right)^\alpha\left(\frac{n}{\beta+n}\right)^k \end{aligned}$$

É fácil obter a esperança e a variância de y sem precisar fazer este cálculo bastando usar a regra de esperanças iteradas. De fato, nós temos:

$$\begin{aligned} E(y) &= E\left[E\left(y|\theta\right)\right] = E\left[n\theta\right] = n\alpha/\beta \\ Var(y) &= Var\left[E(y|\theta)\right] + E\left[Var\left(y|\theta\right)\right] = Var\left[n\theta\right] + E\left[n\theta\right] \\ &= n^2\alpha/\beta^2 + n\alpha/\beta = \alpha n(\beta+n)/\beta^2 \end{aligned}$$

No caso particular de termos α como um inteiro positivo, teremos $\Gamma(\alpha) = (\alpha-1)!$ a probabilidade $p(y=k)$ reduz-se a

$$p(y=k) = \left(\begin{array}{c} \alpha+k-1 \\ k \end{array}\right)\left(\frac{\beta}{\beta+n}\right)^\alpha\left(\frac{n}{\beta+n}\right)^k$$

Esta distribuição mais simples é conhecida como *binomial negativa* e pode ser obtida como o número y de fracassos antes de observar o α-ésimo sucesso numa sucessão de ensaios de Bernoulli independentes com probabilidade de sucesso $\beta/(\beta+n)$.

O caso multivariado é simples:

$$
\begin{aligned}
p(y_1 &= k_1, \ldots, y_n = k_n) = \int_{\theta_1,\ldots,\theta_n} p(y_1 = k_1, \ldots, y_n = k_n|\theta_1, \ldots, \theta_n)p(\theta_1, \ldots, \theta_n)d\theta_1 \ldots d\theta_n \\
&= \int_{\theta_1,\ldots,\theta_n} \left(\prod_{i=1}^{n} p(y_i = k_i|\theta_i)p(\theta_i)\right) d\theta_1 \ldots d\theta_n = \prod_{i=1}^{n} \int_{\theta_i} p(y_i = k_i|\theta_i)p(\theta_i)d\theta_i \\
&= \prod_{i=1}^{n} p(y_i = k_i) = \prod_{i=1}^{n} \frac{\Gamma(\alpha + k_i)}{\Gamma(\alpha)k_i!} \left(\frac{\beta}{\beta + n_i}\right)^{\alpha} \left(\frac{n_i}{\beta + n_i}\right)^{k_i}
\end{aligned}
$$

e portanto $y_1, \ldots, y_n$ são independentes com a distribuição dada acima no caso univariado.

Portanto, a distribuição marginal do vetor $y_1, \ldots, y_n$ (ou distribuição não condicionada nos valores de θ) depende de α e β. Assim, os valores observados de $y_1, \ldots, y_n$ podem dar informação sobre quais seriam os valores de α e β. Podemos usar um método tradicional de estimação (momentos ou máxima verossimilhança) para obter estimadores dos hiperparâmetros α e β. Vamos considerar estas duas possibilidades de estimação para nosso exemplo de mortalidade infantil.

- **Método de momentos:**Vamos definir $r_i = y_i/n_i$. Então $E(r_i) = \alpha/\beta$. Qualquer média ponderada $\sum_i w_i r_i$ dos valores r_i ainda terá média α/β por causa da linearidade da esperança. Assim, vamos tomar a média ponderada pelo tamanho da população $m = \sum_i n_i r_i / \sum_i n_i = \sum_i y_i / \sum_i n_i$ como estimativa de α/β. Isto é, $\widehat{\alpha}/\widehat{\beta} = m$. Temos que $Var(r_i) = \alpha(\beta + n_i)/(\beta^2 n_i)$. Considere o desvio padrão ponderado dos r_i dado por $s^2 = \sum_i n_i(r_i - m)^2 / \sum_i n_i$. Ignorando o erro de assumir que m seja exatamente igual à média α/β de r_i, encontramos

$$
E(s^2) = \sum_i n_i E(r_i - m)^2 / \sum_i n_i \approx \sum_i n_i \left(\alpha(\beta + n_i)/(\beta^2 n_i)\right) / \sum_i n_i = \frac{\alpha}{\beta^2} + \frac{\alpha}{\beta \overline{n}}
$$

onde $\overline{n} = \sum_i n_i/n$ é o número médio de indivíduos por município. Igualando s^2 ao seu valor esperado, e usando que $\widehat{\alpha}/\widehat{\beta} = m$, obtemos a estimativa de momentos de β:

$$
s^2 = \frac{m}{\widehat{\beta}} + \frac{m}{\overline{n}} \Rightarrow \widehat{\beta} = \frac{m\overline{n}}{\overline{n}s^2 - m}
$$

Finalmente, $\widehat{\alpha} = m\widehat{\beta}$. Os dois valores α e β que aparecem na expressão do estimador ótimo de Bayes $\widehat{\theta}_i = E_\theta(\theta_i|y) = (\alpha + y_i)/(\beta + n_i)$ são substituídos pelos valores estimados $\widehat{\alpha}$ e $\widehat{\beta}$ fornecendo os estimadores empíricos de Bayes $\widehat{\widehat{\theta}}_i = E_\theta \widehat{(\theta_i|y)} = (\widehat{\alpha} + y_i)/(\widehat{\beta} + n_i)$

Como exemplo, considere todos os 756 municípios de Minas Gerais e os dados de mortalidade infantil. Então $m = 0.0318$. Além disso,

- **Método de máxima verossimilhança:** Completar no futuro ??

7.1 Estimador de Bayes Linear Ótimo

Como vimos anteriormente, o estimador $r_i = y_i/n_i$ não é uma boa escolha e os estimadores de Bayes ótimos com os parâmetros da priori estimados empiricamente a partir da distribuição marginal dos dados são uma boa solução alternativa. Entretanto, apenas em situações simples é possível calcular estes estimador bayesianos empíricos. Assim, estimadores *lineares* de Bayes passaram a ser utilizados pois o estimador ótimo nesta classe depende apenas dos dois primeiros momentos da distribuição a priori. Muitas vezes estes dois momentos a priori podem ser estimados a partir dos dados usando alguma relação entre os momentos a priori dos parâmetros e a distribuição marginal dos dados.

Como antes no caso da mortalidade infantil, suponha que, dadas as taxas subjacentes $\theta_i, i = 1, \ldots, n$, as contagens possuem distribuições de Poisson independentes e com média $n_i\theta_i$ e que as taxas θ_i, $i = 1, \ldots, n$, são independentes mas não assumimos nada acerca desta distribuição a priori a não ser que ela possui média e variância dadas por μ_θ e σ^2_θ, respectivamente.

O estimador linear ótimo de Bayes é dado por

$$\widehat{\theta_i} = w_i r_i + (1 - w_i)\mu_\theta$$

onde $w_i = \sigma^2_\theta / Var\,(r_i)$ está entre 0 e 1 (Griffin and Krutchkoff, 1971). Isto é, o estimador ??? é uma média ponderada entre r_i e μ_θ.

Para entender um pouco melhor essa média, observe que $Var(r_i) = Var\,(E\,(r_i|\theta_i)) + E\,(Var\,(r_i|\theta_i))$. No entanto, r_i é não-viciado para estimar a taxa desconhecida θ_i correspondente. Então, $Var\,(E\,(r_i|\theta_i)) = Var(\theta_i) = \sigma^2_\theta$. Por outro lado, $E\,(Var\,(r_i|\theta_i)) = E\,(\theta_i/n_i) = \mu_\theta/n_i$ e, desse modo,

$$w_i = \frac{\sigma^2_\theta}{\sigma^2_\theta + \mu_\theta/n_i}$$

Fica claro que o efeito do estimador de Bayes 3.1 é contrair o estimador usual r_i em direção ao valor esperado μ_θ dos θ_i de acordo com o valor de μ_θ/n_i. Uma pequena população n_i vai produzir um valor relativamente grande de μ_θ/n_i, o que indica que r_i é um estimador instável e, como conseqüência, w_i será próximo de zero produzindo uma grade contração. Se a população n_i for grande, então, r_i tem pequena variância, w_i é próximo de 1 e a contração é também pequena.

Em resumo, a estimativa $\widehat{\theta_i}$ em 3.1 é um valor intermediário entre a taxa r_i da área e o valor esperado μ_θ das taxas desconhecidas. O peso de cada uma na estimativa depende da incerteza da estimativa r_i: se essa for uma taxa com pequena variabilidade, r_i praticamente não é a alterada. Se, por outro lado, r_i tiver grande variância, então pouco peso é atribuído ao instável r_i tomando $\widehat{\theta_i}$ mais próximo do valor esperado μ_θ de uma área escolhida ao acaso.

8 Estimação Bayesiana Empírica

Uma das fontes de grande discussão científica no passado era, que como esta distribuição expressa a incerteza de um observador sobre o fenômeno (a taxa subjacente), a inferência passava a fazer sentido apenas para ele e para outros que compartilhavam da mesma expressão da incerteza. Por outro lado, a determinação de forma prévia à observação dos dados dos parâmetros α e β (ou

dos outros parâmetros determinantes de alguma outra distribuição que pudesse ser escolhida ao invés da distribuição gama) é fundamental para os resultados finais e, dado seu caráter subjetivo, o método bayesiano ficava necessariamente numa posição defensiva para convencer observadores céticos. Por causa, disto, o método bayesiano empírico apareceu como uma solução que evita a especificação prévia deste parâmetros da distribuição das taxas θ_i's.

Um dos resultados mais provocantes de teoria estatística nas últimas décadas é o fenômeno chamado paradoxo de Stein ou estimadores de James-Stein ou ainda estimadores de contração (*shrinkage estimators*). Na sua forma mais simples, a situação é a seguinte: uma coleção de medições independentes $X_1, X_2, \ldots, X_k$ é feita e $X_i \sim N(\theta_i, 1)$ onde os θ_i's são constantes desconhecidas e fixas. Os parâmetros $\theta_1, \theta_2, \ldots, \theta_k$ não precisam ser relacionados, podendo referir-se a problemas completamente distintos como, por exemplo, o preço médio do quilo de feijão em Belo Horizonte e a altura média dos japoneses adultos que vivem em Kyoto. Deseja-se estimar os θ_i's *simultaneamente* com a função de perda composta

$$L\left(\theta, \widehat{\theta}\right) = \sum_{i=1}^{k} \left(\theta_i - \widehat{\theta}_i\right)^2$$

onde $\theta = (\theta_1, \ldots, \theta_k)$ e $\widehat{\theta} = \left(\widehat{\theta_1}, \ldots, \widehat{\theta_k}\right)$. O desempenho do estimador $\widehat{\theta}$ é medido pela função de risco

$$R\left(\theta, \widehat{\theta}\right) = EL\left(\theta, \widehat{\theta}\right)$$

onde a esperança é tomada com relação à distribuição conjunta das variáveis $X_1, X_2, \ldots, X_k$.

A descoberta de Stein (1959), refinada em James e Stein (1961), é que, se $k \geq 3$, médias ponderadas dos X_i's são estimadores melhores dos θ_i's do que o estimador óbvio $\widehat{\theta}_i^o = X_i$. Efron e Morris (1973) mostraram que estes estimadores podiam ser vistos como estimadores bayesianos empíricos motivando então uma série de trabalhos utilizando esta metodologia. Um conjunto desses trabalhos apareceu na área de mapeamento de doenças, como descrevemos a seguir.

8.1 Estimador Ótimo Linear Bayesiano Empírico

Marshall (1991) propôs um método bayesiano empírico extremamente simples de ser implementado e que não supõe nenhuma distribuição específica para os . Ele é baseado no método dos momentos aplicado a distribuição marginal das contagens.

9 Estimação de Modelos Bayesianos Empíricos

Considere uma região particionada em N áreas indexadas por $i, i = 1, \ldots, N$. Suponha que a taxa anual de eventos (desconhecida) em cada área é denotada por θ_i e que $r_i = y_i/n_i$ é a taxa observada, onde y_i é o número de eventos na área i e n_i é o número de pessoas em risco. É comum supor que, dada a taxa desconhecida θ_i, a contagem y_i do número de casos possui distribuição de probabilidade de Poisson com esperança condicional $E(y_i|\theta_i) = n_i\ \theta_i$ e variância condicional $\mathrm{Var}(y_i|\theta_i) = n_i\ \theta_i$

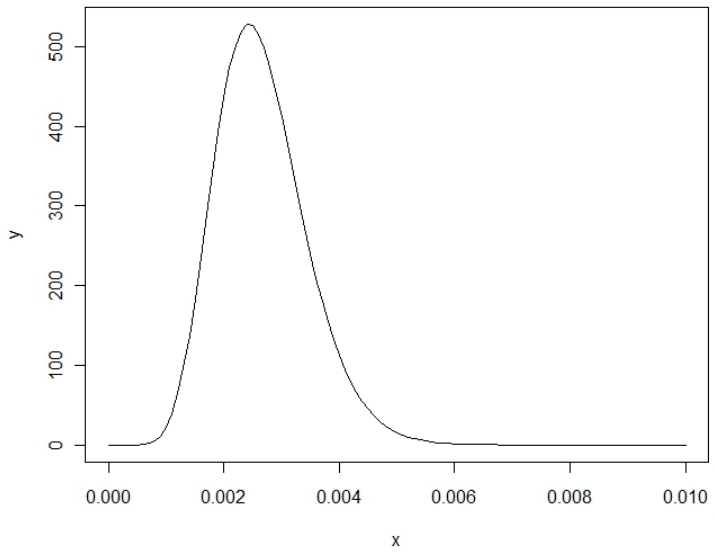

Figure 8: Densidade de uma distribuição gamma com $\alpha = 11.55$ e $\beta = 4.336 \times 10^3$.

Num contexto não-bayesiano, quando y_i possui distribuição Poisson ou binomial, a taxa r_i é a melhor estimativa de θ_i no sentido de minimizar o erro médio de estimação $E(\theta_i - \hat{\theta}_i)^2$.

Uma abordagem bayesiana supõe que as taxas desconhecidas $\theta_i, i = 1, \ldots, n$, são independentes e possuem uma distribuição de probabilidade comum com valor esperado μ_θ e variância σ_θ^2. É útil imaginar cada área i escolhendo sua taxa θ_i de forma aleatória a partir de uma distribuição comum. Desse modo, pode-se interpretar mu_θ como sendo o valor esperado da taxa de uma área escolhida ao acaso na região. A distribuição comum das taxas poderia ser bastante complicada com várias modas e sem uma expressão matemá tica conhecida. No entanto, na prática, escolhe-se uma distribuição simples. Por exemplo, as taxas θ_i poderiam estar sendo geradas a partir de uma distribuição gama com parâmetros α e β, denotada por $\Gamma(\alpha, \beta)$. A figura **??** mostra a densidade de probabilidade de uma gamma com parâmetros $\alpha = 11.55$ e $beta = 4.336 \times 10^3$.

Parece razoável supor que as diferenças entre as taxas desconhecidas θ_i possuem uma regularidade ou homogeneidade derivada dos processos sociais e ambientais subjacentes que afetam a região em estudo. Essa regularidade nos permite modelar as variações nos θ_i através de uma distrbuição de probabilidade. Este modelo para as taxas desconhecidas θ_i reconhece que elas não são números totalmente arbitrários, completamente desconectados, mas que guardam certa relação entre si. Por exemplo, para um dado fenômeno numa região, as taxas de cada sub-área provavelmente vão estar concentradas numa faixa de valores com alguns poucos valores possivelmente mais afastados.

A distrbuição gama $\Gamma(\alpha, \beta)$ possui valor esperado $\mu_\theta = \alpha/\beta$ e variância $\sigma_\theta^2 = \alpha/\beta^2$. Para o exemplo da figura **??**, temos $\mu_\theta = 0.00266$ e $\sigma_\theta^2 = 7.28 \times 10^{-6}$.

O interesse reside em estimar as taxas desconhecidas com base nos dados observados. Como existe interesse em *todas* as taxas de nosso mapa, e não numa taxa de uma área particular, adota-se como critério de escolha para um estimador aquele que minimize a *soma* dos erros quadráticos de estimação das áreas $\sum_i E(\theta_i - \hat{\theta}_i)^2$. O estimador linear ótimo de Bayes é dado por

$$\hat{\theta}_i = w_i\, r_i + (1 - w_i)\mu_\theta \tag{1}$$

onde $w_i = \sigma^2_\theta / \mathrm{Var}(r_i)$ está entre 0 e 1 (Griffin and Krutchkoff, 1971). Isto é, o estimador 1 é uma média ponderada entre r_i e μ_θ.

Para entender um pouco melhor essa média, observe que $\mathrm{Var}(r_i) = \mathrm{Var}(E(r_i|\theta_i)) + E(\mathrm{Var}(r_i|\theta_i))$. No entanto, r_i é não-viciado para estimar a taxa desconhecida θ_i correspondente. Então, $\mathrm{Var}E(r_i\theta_i) = \mathrm{Var}(\theta_i) = \sigma^2_\theta$ e, desse modo, $w_i = \sigma^2_\theta / (\sigma^2_\theta + E(\mathrm{Var}(r_i|\theta_i))$.

Fica claro que o efeito do estimador 1 de Bayes é contrair o estimador usual r_i em direção à média μ_θ dos θ_i de acordo com o valor de $E(\mathrm{Var}(r_i|\theta_i))$. Um grande valor indica que, em média, r_i é um estimador instável e, como consequência, o estimador terá w_i próximo de zero produzindo uma grande contração. Se $E(\mathrm{Var}(r_i|\theta_i))$ é pequeno então, em média, r_i tem pequena variância, w_i épróximo de 1 e a contração é também pequena.

Em resumo, a estimativa $\hat{\theta}_i$ em 1 é um valor intermediá rio entre a taxa r_i da área e a média μ_θ das taxas desconhecidas. O peso de cada uma na estimativa depende da incerteza da estimativa r_i: se essa for uma taxa com pequena variabilidade, r_i praticamente não é alterada. Se, por outro lado, r_i tiver grande variância em média, então pouco peso é atribuído ao instável r_i tomando $\hat{\theta}_i$ mais próximo do valor esperado μ_θ de uma área escolhida ao acaso.

Em princípio, o cálculo de w_i é simples se a distribuição dos θ_i e r_i são conhecidas. No caso em que $\theta_i \sim Gamma(\alpha, \beta)$ e $(r_i|\theta_i) \sim Poisson(n_i\, \theta_i)$, encontramos $\sigma^2_\theta = \alpha/beta^2$ e

$$E(\mathrm{Var}(r_i|\theta_i)) = E(\theta_i/n_i) = \mu_\theta/n_i = \alpha/(n_i\beta)$$

. Assim,

$$w_i = \frac{\alpha/\beta^2}{\alpha/\beta^2 + \alpha/(n_i\beta)}$$

e, após alguma manipulação algébrica, encontra-se que 1 reduz-se a:

$$\hat{\theta}_i = \frac{y_i + \alpha}{n_i + \beta}\,. \tag{2}$$

O problema com a fórmula acima é que, em geral, α e β não são conhecidos e existe pouca disposição em chutar algum núemro. No caso geral, esse problema aparec no desconhecimento de μ_θ e σ^2_θ. Uma saída é estimar esses parâmetros a partir dos próprios dados. Essa é a principal característica do método *bayesiano empírico*. A estimativa pode ser *paramétrica* ou *não-paramétrica*.

9.1 Estimativa bayesiana empírica paramétrica

9.2 Estimativa bayesiana empírica não-paramétrica: Método de Momentos de Marshall

Marshall (1991) propos um método extremamente simples de ser implementado e que não supõe nenhuma distribuição específica para os θ_i.

As fórmulas bá sicas são:

$$\hat{\theta}_i = C_i\, r_i + (1 - C_i)\, \hat{m}$$

onde $\hat{m} = \sum_i y_i / \sum_i n_i$, a taxa global, e

$$C_i = \frac{s^2 - \hat{m}/\bar{n}}{s^2 - \hat{m}/\bar{n} + \hat{m}/n_i}$$

com $\bar{n} = \sum_i /N$, o número médio de pessoas em risco, e $s^2 = \sum_i n_i\,(r_i - \hat{m})^2/n$, onde $n = \sum_i n_i$.

Pode acontecer de $s^2 < \hat{m}/\bar{n}$. Nesse caso, $\hat{\theta}_i = \hat{m}$.

9.3 Verossimilhança Penalizada

9.4 Resumo

As fórmulas básicas são:

???

Onde ???, a taxa global, e ???

com ??? , o número médio de pessoas em risco, e ???, onde ???. Pode acontecer de ???. Neste caso, ???.

No caso da mortalidade infantil em São Paulo, aplicando este método proposto por Marshall produz o mapa da Figura 3.6 na mesma escala em que foi feito o mapa das taxas brutas na Figura 3.1. Observe a enorme contração das taxas estimadas pelo método bayesiano mostrado que boa parte da variabilidade presente nas taxas brutas pode ser creditada a fatores aleatórios não associados com os riscos subjacentes.

A Figura 3.7 mostra o mesmo mapa mas utilizando uma escala diferente da taxa bruta, mas apropriada para a pequena variação das taxas brutas bayesianas empíricas.

Figura 3.6 : Mapa das taxas bayesianas empíricas de mortalidade infantil em São Paulo em 1998. Escala é a mesma de taxas brutas.

9.5 Estimador Inteiramente Bayesiano

O estimador bayesiano empírico de Marshall (1991) é muito fácil de ser utilizado e produz boas estimativas pontuais, comparáveis às dos métodos inteiramente bayesianos. Entretanto, ele possui duas desvantagens não compartilhadas por um método puramente bayesiano: ele ignora a variabilidade introduzida na estimação dos parâmetros da priori e não lida bem com a variabilidade dos estimadores produzidos, além disto, ele não pode ser generalizado facilmente para as situações bem mais complexas, como de interação entre espaço e tempo ou a introdução de covariáveis que começam a ganhar a atenção dos pesquisadores. O assunto do próximo capítulo é uma introdução aos métodos inteiramente bayesianos para a produção de mapas de doenças.

Figura 3.7: Mapa das taxas bayesianas empíricas de mortalidade infantil em São Paulo em 1998.