FUNDAÇÃO GETULIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA
Inferência Estatística

Tradução de Probability and Statistics by DeGroot & Schervish

VICTOR GABRIEL HARUO IWAMOTO

Rio de Janeiro – RJ
August 2025

7.1 Inferência Estatística

Lembre-se de nossos vários exemplos de ensaios clínicos. O que poderíamos dizer sobre a probabilidade de um futuro paciente responder com sucesso ao tratamento depois de observarmos os resultados de uma coleção de outros pacientes? Essa é a questão que a inferência estatística se destina a abordar. Em geral, a inferência estatística consiste em fazer declarações probabilísticas sobre quantidades desconhecidas. Por exemplo, podemos calcular médias, variâncias, quantis e algumas outras quantidades que ainda serão introduzidas sobre variáveis aleatórias não observadas e parâmetros desconhecidos de distribuições. Nosso objetivo será dizer o que aprendemos sobre as quantidades desconhecidas após observar alguns dados que acreditamos conter informações relevantes. Aqui estão alguns outros exemplos de questões que a inferência estatística pode tentar responder. O que podemos dizer sobre se uma máquina está funcionando corretamente após observarmos parte de sua produção? Em um processo cível, o que podemos dizer sobre se houve discriminação após observar como diferentes grupos étnicos foram tratados? Os métodos de inferência estatística, que desenvolveremos para abordar essas questões, são construídos sobre a teoria da probabilidade abordada nos capítulos anteriores deste texto.

Probabilidade e Modelos Estatísticos

Nos capítulos anteriores deste livro, discutimos a teoria e os métodos da probabilidade. À medida que novos conceitos em probabilidade eram introduzidos, também introduzimos exemplos do uso desses conceitos em problemas que agora reconheceremos como *inferência estatística*. Antes de discutir a inferência estatística formalmente, é útil nos recordarmos daqueles conceitos de probabilidade que fundamentarão a inferência.

Exemplo 7.1.1 Tempo de Vida de Componentes Eletrônicos. Uma empresa vende componentes eletrônicos e está interessada em saber por quanto tempo cada componente provavelmente durará. Eles podem coletar dados sobre componentes que foram usados sob condições típicas. Eles optam por usar a família de distribuições exponenciais para modelar o tempo (em anos) desde o momento em que um componente é colocado em serviço até sua falha. Eles gostariam de modelar os componentes como tendo todos a mesma taxa de falha θ, mas há incerteza sobre o valor numérico específico de θ. Para ser mais preciso, seja $X_1, X_2, \ldots$ uma sequência de tempos de vida de componentes em anos. A empresa acredita que, se soubessem a taxa de falha θ, então $X_1, X_2, \ldots$ seriam variáveis aleatórias i.i.d. (independentes e identicamente distribuídas) com distribuição exponencial com parâmetro θ. (Ver Seção 5.7 para a definição de distribuições exponenciais. Estamos usando o símbolo θ para o parâmetro de nossas distribuições exponenciais em vez de β para corresponder ao restante da notação neste capítulo.) Suponha que os dados que a empresa irá observar

consistam nos valores de $X_1, \ldots, X_m$, mas que eles ainda estejam interessados em $X_{m+1}, X_{m+2}, \ldots$. Eles também estão interessados em θ porque está relacionado ao tempo de vida médio. Como vimos na Eq. (5.7.17), a média de uma variável aleatória exponencial com parâmetro θ é $1/\theta$, razão pela qual a empresa pensa em θ como a taxa de falha.

Imaginamos um experimento cujos resultados são sequências de tempos de vida como descrito acima. Como mencionado, se soubéssemos o valor de θ, então $X_1, X_2, \ldots$ seriam variáveis aleatórias i.i.d. Nesse caso, a lei dos grandes números (Teorema 6.2.4) diz que a média $\frac{1}{n}\sum_{i=1}^{n} X_i$ converge em probabilidade para a média $1/\theta$. E o Teorema 6.2.5 diz que $n/\sum_{i=1}^{n} X_i$ converge em probabilidade para θ. Como θ é uma função da sequência de tempos de vida que constituem cada resultado experimental, ele pode ser tratado como uma variável aleatória. Suponha que, antes de observar os dados, a empresa acredite que a taxa de falha é provavelmente em torno de 0,5/ano, mas há uma certa incerteza sobre isso. Eles modelam θ como uma variável aleatória com distribuição gama com parâmetros 1 e 2. Parafraseando o que foi dito anteriormente, eles também modelam $X_1, X_2, \ldots$ como variáveis aleatórias exponenciais condicionalmente i.i.d. com parâmetro θ dado θ. Eles esperam aprender mais sobre θ examinando os dados da amostra $X_1, \ldots, X_m$. Eles nunca podem aprender θ precisamente, pois isso exigiria observar toda a sequência infinita $X_1, X_2, \ldots$. Por essa razão, θ é apenas hipoteticamente observável.

O Exemplo 7.1.1 ilustra várias características que serão comuns à maioria dos problemas de inferência estatística e que constituem o que chamamos de modelo estatístico.

Definição 7.1.1 Modelo Estatístico. Um *modelo estatístico* consiste em uma identificação das variáveis aleatórias de interesse (tanto observáveis quanto apenas hipoteticamente observáveis), uma especificação de uma distribuição de probabilidade conjunta ou de uma família de possíveis distribuições conjuntas para as variáveis aleatórias observáveis, a identificação de quaisquer parâmetros dessas distribuições que se assumem desconhecidos e possivelmente hipoteticamente observáveis, e (se desejado) uma especificação de uma distribuição conjunta para os parâmetros desconhecidos. Quando tratamos os parâmetros desconhecidos θ como aleatórios, então a distribuição de probabilidade conjunta das variáveis aleatórias observáveis indexada por θ é entendida como a distribuição condicional das variáveis aleatórias observáveis dado θ.

No Exemplo 7.1.1, as variáveis aleatórias observáveis de interesse formam a sequência $X_1, X_2, \ldots$, enquanto a taxa de falha θ é hipoteticamente observável. A família de possíveis distribuições conjuntas de $X_1, X_2, \ldots$ é indexada

pelo parâmetro θ. A distribuição de probabilidade conjunta das observáveis correspondente ao valor θ é aquela em que $X_1, X_2, \ldots$ são variáveis aleatórias i.i.d. cada uma com distribuição exponencial com parâmetro θ. Esta também é a distribuição condicional de $X_1, X_2, \ldots$ dado θ porque estamos tratando θ como uma variável aleatória. A distribuição de θ é a distribuição gama com parâmetros 1 e 2.

Nota: Redefinindo Ideias Antigas. O leitor notará que um modelo estatístico nada mais é do que uma formalização de muitas características que temos usado em vários exemplos ao longo dos capítulos anteriores deste livro. Alguns exemplos precisam apenas de algumas das características que compõem a especificação completa de um modelo estatístico, enquanto outros exemplos usam a especificação completa. Nas Seções 7.1–7.4, nós vamos introduzir uma quantidade considerável de terminologia, grande parte da qual é a formalização de conceitos que foram introduzidos e usados em vários lugares anteriormente no livro. O propósito de toda essa formalidade é nos ajudar a manter os conceitos organizados para que possamos dizer quando estamos aplicando as mesmas ideias de novas maneiras e quando estamos introduzindo novas ideias.

Estamos agora prontos para introduzir formalmente a inferência estatística.

Definição 7.1.2 Inferência Estatística. Uma *inferência estatística* é um procedimento que produz uma declaração probabilística sobre alguns ou todas as partes de um modelo estatístico.

Por uma "declaração probabilística", queremos dizer uma declaração que faz uso de qualquer um dos conceitos de teoria da probabilidade que foram discutidos no texto ou que ainda serão discutidos mais tarde. Por exemplo, eles incluem uma média, uma média condicional, um quantil, uma variância, uma distribuição condicional de uma variável aleatória dada outra, a probabilidade de um evento, uma probabilidade condicional de um evento dado algum outro, e assim por diante. No Exemplo 7.1.1, aqui estão alguns exemplos de inferências estatísticas que alguém poderia querer fazer:

- Produzir uma variável aleatória Y (uma função de $X_1, \ldots, X_m$) tal que $\Pr(Y \geq \theta|\theta) = 0.9$.
- Produzir uma variável aleatória Y que se espera que esteja próxima de θ.
- Computar quão provável é que a média dos próximos 10 tempos de vida, $\frac{1}{10}\sum_{i=m+1}^{m+10} X_i$, seja pelo menos 2.
- Dizer algo sobre quão confiantes estamos de que $\theta \leq 0.4$ após observar $X_1, \ldots, X_m$.

Todos esses tipos de inferência e outros serão discutidos com mais detalhes neste livro.

Na Definição 7.1.1, distinguimos entre variáveis aleatórias observáveis e hipoteticamente observáveis. Reservamos o nome *observável* para uma variável aleatória que temos certeza de que poderíamos observar se dedicássemos o esforço necessário para observá-la. O nome *hipoteticamente observável* foi usado para uma variável aleatória que exigiria recursos infinitos para ser observada, como o limite (quando $n \to \infty$) das médias amostrais das primeiras n observáveis. Neste texto, tal variável aleatória hipoteticamente observável corresponderá aos parâmetros da distribuição conjunta dos observáveis como no Exemplo 7.1.1. Como esses parâmetros figuram de forma proeminente em muitos dos tipos de problemas de inferência que veremos, vale a pena formalizar o conceito de parâmetro.

Definição 7.1.3 Parâmetro/Espaço de parâmetros. Em um problema de inferência estatística, uma característica ou combinação de características que determina a distribuição conjunta para as variáveis aleatórias de interesse é chamada de *parâmetro* da distribuição. O conjunto Ω de todos os valores possíveis de um parâmetro θ ou de um vetor de parâmetros $(\theta_1, \ldots, \theta_k)$ é chamado de *espaço de parâmetros*.

Todas as famílias de distribuições introduzidas anteriormente (e a serem introduzidas mais tarde) neste livro têm parâmetros que estão incluídos nos nomes dos membros individuais da família. Por exemplo, a família de distribuições binomiais tem parâmetros que chamamos de n e p, a família de distribuições normais é parametrizada pela média μ e variância σ^2 de cada distribuição, a família de distribuições uniformes em intervalos é parametrizada pelos extremos dos intervalos, a família de distribuições exponenciais é parametrizada pela taxa de parâmetro θ, e assim por diante. No Exemplo 7.1.1, o parâmetro θ (a taxa de falha) deve ser positivo. Portanto, a menos que certos valores positivos de θ possam ser explicitamente descartados como valores possíveis de θ, o espaço de parâmetros Ω será o conjunto de todos os números positivos. Como outro exemplo, suponha que a distribuição das alturas dos indivíduos em uma certa população seja assumida como a distribuição normal com média μ e variância σ^2, mas que os valores exatos de μ e σ^2 sejam desconhecidos. A média μ e a variância σ^2 determinam a distribuição normal particular para as alturas dos indivíduos. Assim, (μ, σ^2) pode ser considerado um par de parâmetros. Neste exemplo de alturas, tanto μ quanto σ^2 devem ser positivos. Portanto, o espaço de parâmetros Ω pode ser considerado como o conjunto de todos os pares (μ, σ^2) tais que $\mu > 0$ e $\sigma^2 > 0$. Se a distribuição normal neste exemplo representa a distribuição das alturas em polegadas dos indivíduos nesta população particular, podemos ter certeza de que $30 < \mu < 100$ e $\sigma^2 < 50$. Nesse caso, o espaço de parâmetros Ω poderia ser considerado como o conjunto menor de todos os

pares (μ, σ^2) tais que $30 < \mu < 100$ e $0 < \sigma^2 < 50$.

A característica importante do espaço de parâmetros Ω é que ele deve conter todos os valores possíveis dos parâmetros em um dado problema, para que possamos ter certeza de que o valor real do vetor de parâmetros é um ponto em Ω.

Exemplo 7.1.2 Um Ensaio Clínico. Suponha que 40 pacientes receberão um tratamento para uma condição e que observaremos para cada paciente se ele se recupera ou não da condição. Podemos também estar interessados em uma grande coleção de pacientes adicionais que receberão o mesmo tratamento. Para ser específico, para cada paciente $i = 1, 2, \ldots$, seja $X_i = 1$ se o paciente i se recuperar, e seja $X_i = 0$ se não. Como uma coleção de possíveis distribuições para $X_1, X_2, \ldots$, poderíamos escolher dizer que os X_i são i.i.d. tendo a distribuição de Bernoulli com parâmetro p para $0 \leq p \leq 1$. Neste caso, o parâmetro p é conhecido por estar no intervalo $[0, 1]$, e este intervalo poderia ser considerado o espaço de parâmetros. Note também que a lei dos grandes números (Teorema 6.2.4) diz que p é o limite, quando n tende ao infinito, da proporção dos primeiros n pacientes que se recuperam.

Na maioria dos problemas, existe uma interpretação natural para o parâmetro como uma característica de possíveis distribuições de nossos dados. No Exemplo 7.1.2, o parâmetro p tem uma interpretação natural como a proporção de nossa população de pacientes que se recupera do tratamento. No Exemplo 7.1.1, o parâmetro θ tem uma interpretação natural como uma taxa de falha, ou seja, um sobre o tempo de vida médio de uma grande população de tempos de vida. Tais casos, inferência sobre parâmetros pode ser interpretada como inferência sobre as características que o parâmetro representa. Neste texto, todos os parâmetros terão tais interpretações naturais. Em exemplos que se encontram fora de um curso introdutório, as interpretações podem não ser tão diretas.

Exemplos de Inferência Estatística

Aqui estão alguns dos exemplos de modelos estatísticos e inferências que foram introduzidos anteriormente no texto.

Exemplo 7.1.3 Um Ensaio Clínico. O ensaio clínico introduzido no Exemplo 2.1.4 estava preocupado com a probabilidade de os pacientes evitarem uma recaída enquanto recebiam vários tratamentos. Para cada i, seja $X_i = 1$ se o paciente i no tratamento com imipramina evitar a recaída e $X_i = 0$ caso contrário. Seja P a proporção de pacientes que evitam a recaída em um grande grupo recebendo tratamento com imipramina. Se P for desconhecido, podemos modelar $X_1, X_2, \ldots$ como i.i.d. variáveis aleatórias de Bernoulli com parâme-

tro p condicional a $P = p$. Os pacientes na coluna da imipramina da Tabela 2.1 devem nos fornecer alguma informação que mude nossa incerteza sobre P. Uma inferência estatística consistiria em fazer uma declaração de probabilidade sobre os dados e/ou P, e o que os dados e P nos dizem um sobre o outro. Por exemplo, no Exemplo 4.7.8, assumimos que P tinha a distribuição uniforme no intervalo $[0, 1]$, e encontramos a distribuição condicional de P dados os resultados observados do estudo. Também calculamos a média condicional de P dados os resultados do estudo, bem como o E.M.Q. (Erro Médio Quadrático) para prever P tanto antes quanto depois de observar os resultados do estudo.

Exemplo 7.1.4 Partículas Radioativas. No Exemplo 5.7.8, partículas radioativas atingem um alvo de acordo com um processo de Poisson com taxa desconhecida β. No Exercício 22 da Seção 5.7, foi solicitado que você encontrasse a distribuição condicional de β após observar o processo de Poisson por um certo período de tempo.

Exemplo 7.1.5 Antropometria de Besouros Pulga. No Exemplo 5.10.2, plotamos duas medidas físicas de uma amostra de 31 besouros pulga juntamente com contornos de uma distribuição normal bivariada. A família de distribuições normais bivariadas é parametrizada por cinco quantidades: as duas médias, as duas variâncias e a correlação. A escolha de qual conjunto desses cinco parâmetros usar para os dados ajustados é uma forma de inferência estatística conhecida como *estimação*.

Exemplo 7.1.6 Intervalo para a Média. Suponha que as alturas dos homens em uma certa população sigam a distribuição normal com média μ e variância 9, como no Exemplo 5.6.7. Desta vez, assuma que não conhecemos o valor da média μ, mas desejamos aprender sobre ela amostrando da população. Suponha que decidamos amostrar $n = 36$ homens e seja $\bar{X}_n$ a média de suas alturas. Então o intervalo $(\bar{X}_n - 0.98, \bar{X}_n + 0.98)$ calculado no Exemplo 5.6.8 tem a propriedade de que conterá o valor de μ com probabilidade 0.95.

Exemplo 7.1.7 Discriminação na Seleção do Júri. No Exemplo 5.8.4, estávamos interessados em saber se havia evidência de discriminação contra Mexicano-Americanos na seleção do júri. A Figura 5.8 mostra como pessoas que entraram no caso com diferentes opiniões sobre a extensão da discriminação (se houver) poderiam alterar suas opiniões à luz do aprendizado da evidência numérica apresentada no caso.

Exemplo 7.1.8 Tempos de Serviço em uma Fila. Suponha que clientes em uma fila devam esperar por serviço e que estamos interessados em observar os tempos de serviço de vários clientes. Suponha que estejamos interessados na taxa em que os clientes são atendidos. Seja Z a taxa de serviço, e no Exemplo 5.7.4, mostramos como encontrar a distribuição condicional de Z dados vários tempos de serviço observados.

Classes Gerais de Problemas de Inferência

Previsão. Uma forma de inferência é tentar prever variáveis aleatórias que ainda não foram observadas. No Exemplo 7.1.1, podemos estar interessados na média dos próximos 10 tempos de vida, $\frac{1}{10}\sum_{i=m+1}^{m+10} X_i$. No exemplo do ensaio clínico (Exemplo 7.1.3), podemos estar interessados em quantos pacientes no grupo da imipramina terão sucesso. Em praticamente todo problema de inferência estatística em que não observamos todos os dados relevantes, a previsão é possível. Quando a quantidade não observada a ser prevista é um parâmetro, a previsão é geralmente chamada de *estimação*, como no Exemplo 7.1.5.

Problemas de Decisão Estatística. Em muitos problemas de inferência estatística, após a análise de dados experimentais, devemos escolher entre várias decisões de classes com a propriedade de que as consequências de cada decisão disponível dependem do valor desconhecido de algum parâmetro. Por exemplo, podemos ter que estimar a taxa de falha θ de nossos componentes eletrônicos quando as consequências dependem de quão próxima nossa estimativa de θ está do valor correto. Como outro exemplo, podemos ter que decidir se a proporção desconhecida P de pacientes no exemplo da imipramina (Exemplo 7.1.3) é maior ou menor que uma constante especificada quando as consequências dependem de onde P se encontra em relação à constante. Este último tipo de inferência está intimamente relacionado a *testes de hipóteses*, o assunto do Capítulo 9.

Delineamento Experimental. Em alguns problemas de inferência estatística, temos algum controle sobre o tipo ou a quantidade de dados experimentais que serão coletados. Por exemplo, considere um experimento para determinar a resistência média à tração de um certo tipo de liga como uma função da pressão e temperatura em que a liga é produzida. Dentro dos limites de certos orçamentos e restrições de tempo, pode ser possível para o experimentador escolher os níveis de pressão e temperatura nos quais os espécimes experimentais da liga serão produzidos, e também especificar o número de espécimes a serem produzidos em cada um desses níveis. Tal problema, no qual o experimentador pode escolher (pelo menos até certo ponto) o delineamento experimental particular a ser realizado, é chamado de problema de *delineamento experimental*. Obviamente, o delineamento de um experimento e a análise estatística dos dados experimentais estão intimamente relacionados. Não se pode projetar um experimento eficaz sem considerar a análise estatística subsequente que será realizada nos dados que serão obtidos. E não se pode realizar uma análise estatística significativa

de dados experimentais sem considerar o delineamento experimental particular do qual os dados foram derivados.

Outras Inferências. As classes gerais de problemas descritas acima, bem como os exemplos mais específicos que apareceram anteriormente, pretendem ser ilustrações de tipos de inferências estatísticas que poderemos realizar com a teoria e métodos introduzidos neste texto. A gama de possíveis modelos, inferências e métodos que podem surgir quando os dados são observados em problemas de pesquisa reais excede em muito o que podemos introduzir aqui. Espera-se que, ao obter uma compreensão dos problemas que cobrimos aqui, o leitor terá uma apreciação do que precisa ser feito quando um problema estatístico mais desafiador surge.

Definição de uma Estatística

Exemplo 7.1.9 Tempos de Falha de Rolamentos de Esferas. No Exemplo 5.6.9, tínhamos uma amostra dos números de milhões de revoluções antes da falha para 23 rolamentos de esferas. Modelamos os tempos de vida como uma amostra aleatória de uma distribuição lognormal. Podemos supor que os parâmetros μ e σ^2 dessa distribuição lognormal são desconhecidos e que talvez queiramos fazer alguma inferência sobre eles. Gostaríamos de fazer uso dos 23 valores observados para fazer qualquer inferência. Mas precisamos acompanhar todos os 23 valores ou existem alguns resumos dos dados sobre os quais nossa inferência será baseada?

Cada inferência estatística que aprenderemos a realizar neste livro será baseada em um ou alguns resumos dos dados disponíveis. Tais resumos de dados surgem com tanta frequência e são tão fundamentais para a inferência que recebem um nome especial.

Definição 7.1.4 Estatística. Suponha que as variáveis aleatórias observáveis de interesse sejam $X_1, \ldots, X_n$. Seja r uma função de valor real arbitrária de n variáveis reais. Então a variável aleatória $T = r(X_1, \ldots, X_n)$ é chamada de *estatística*.

Três exemplos de estatísticas são a média amostral $\bar{X}_n$, o máximo Y_n dos valores de $X_1, \ldots, X_n$, e a função $r(X_1, \ldots, X_n)$, que tem o valor constante 3 para todos os valores de $X_1, \ldots, X_n$.

Exemplo 7.1.10 Tempos de Falha de Rolamentos de Esferas. No

Exemplo 7.1.9, suponha que estivéssemos interessados em fazer uma declaração sobre quão longe μ está de 40. Então, poderíamos querer usar a estatística

$$T = \left| \frac{1}{36} \sum_{i=1}^{36} \log(X_i) - 4 \right|$$

em nosso procedimento de inferência. Neste caso, T é uma medida ingênua de quão longe os dados sugerem que μ está de 40.

Exemplo 7.1.11 Intervalo para a Média. No Exemplo 7.1.6, construímos um intervalo que tem probabilidade 0.95 de conter μ. Os extremos do intervalo, a saber, $\bar{X}_n - 0.98$ e $\bar{X}_n + 0.98$, são estatísticas.

Muitas inferências podem prosseguir sem construir estatísticas explicitamente como um passo preliminar. No entanto, a maioria das inferências envolverá o uso de estatísticas que poderiam ser identificadas antecipadamente. E saber quais estatísticas são úteis em quais circunstâncias pode simplificar muito a inferência. Expressar uma inferência em termos de uma estatística também pode nos ajudar a decidir quão bem a inferência atende às nossas necessidades. Por exemplo, no Exemplo 7.1.10, se estimamos $\mu - 40$ por T, podemos usar a distribuição de T para nos ajudar a determinar quão provavelmente é que T difira de $|\mu - 40|$ por uma grande quantidade. À medida que construímos inferências específicas mais tarde neste livro, chamaremos a atenção para aquelas estatísticas que desempenham papéis importantes na inferência.

Parâmetros como Variáveis Aleatórias

Há alguma controvérsia sobre se os parâmetros devem ser tratados como variáveis aleatórias ou meramente como números que indexam uma distribuição. Por exemplo, no Exemplo 7.1.3, seja P a proporção de pacientes que evitam a recaída em um grande grupo que recebe imipramina. Então, dizemos que $X_1, X_2, \ldots$ são i.i.d. variáveis aleatórias de Bernoulli com parâmetro p condicional a $P = p$. Estamos explicitamente pensando em P como uma variável aleatória, e damos a ele uma distribuição. Uma alternativa seria dizer que $X_1, X_2, \ldots$ são i.i.d. variáveis aleatórias de Bernoulli com parâmetro p onde p é desconhecido e deixar por isso mesmo. Se realmente queremos calcular algo como a probabilidade condicional de que a proporção de pacientes seja maior que 0.5 dados os resultados dos primeiros 40 pacientes, então devemos tratar P como uma variável aleatória. Por outro lado, se estamos apenas interessados em fazer declarações de probabilidade que são indexadas pelo valor de p, então não precisamos pensar em p como uma variável aleatória. Por exemplo, podemos desejar encontrar duas variáveis aleatórias Y_1 e Y_2 (funções de $X_1, \ldots, X_{40}$) tais que, não importa qual p seja, a probabilidade de que $Y_1 \leq p \leq Y_2$ seja de

pelo menos 0.9. Algumas das inferências que discutiremos mais adiante neste livro são do primeiro tipo que requerem o tratamento de P como uma variável aleatória, e algumas são do último tipo em que p é meramente um índice para uma distribuição.

Alguns estatísticos acreditam que é possível e útil tratar parâmetros como variáveis aleatórias em todos os problemas de inferência estatística. Eles acreditam que a distribuição de um parâmetro é uma probabilidade subjetiva que representa as crenças subjetivas e informadas de um experimentador individual sobre onde o valor verdadeiro do parâmetro provavelmente está. Uma vez que eles atribuem uma distribuição a um parâmetro, essa distribuição não é diferente de qualquer outra distribuição de probabilidade usada no campo da estatística, e todas as regras da teoria da probabilidade se aplicam a cada distribuição. De fato, em todos os casos descritos neste livro, os parâmetros podem realmente ser identificados como limites de funções de grandes coleções de observações potenciais. Aqui está um exemplo típico.

Exemplo 7.1.12 Parâmetro como um Limite de Variáveis Aleatórias. No Exemplo 7.1.3, o parâmetro P pode ser entendido da seguinte forma: Imagine uma sequência infinita de pacientes potenciais recebendo tratamento com imipramina. Suponha que, para cada inteiro n, os resultados de cada subconjunto ordenado de n pacientes dessa sequência infinita tenham a mesma distribuição conjunta que os resultados de qualquer outro subconjunto ordenado de n pacientes. Em outras palavras, suponha que a ordem em que os pacientes aparecem na sequência seja irrelevante para o resultado do tratamento. Seja P_n a proporção de pacientes que não recaem entre os primeiros n pacientes. Pode-se mostrar que a probabilidade é 1 de que P_n convirja para algo quando $n \to \infty$. Essa algo pode ser pensado como P, o que tem sido chamado de proporção de sucessos em uma população muito grande. Nesse sentido, P é uma variável aleatória porque é uma função de outros modelos de variáveis aleatórias. Um argumento semelhante pode ser feito em todos os modelos estatísticos deste livro, envolvendo parâmetros, mas a matemática necessária para tornar esses argumentos precisos é muito avançada para ser apresentada aqui (o Capítulo 12 de Schervish (1995) contém os detalhes necessários). Estatísticos que argumentam desta forma são ditos aderir à filosofia Bayesiana de estatística e são chamados de *Bayesianos*.

Há outra linha de raciocínio que leva naturalmente a tratar P como uma variável aleatória no Exemplo 7.1.12 sem depender de uma sequência infinita de pacientes potenciais. Suponha que o número de pacientes potenciais, embora grande, seja finito, digamos N. Então podemos fazer a aproximação na Seção 5.3.4 aplicável. Então P é apenas a proporção de sucessos entre a grande população de N pacientes. Condicional a $P = p$, o número de sucessos em uma amostra de n pacientes será aproximadamente uma variável aleatória binomial

com parâmetros n e p de acordo com o Teorema 5.3.4. Se os resultados dos pacientes na amostra são variáveis aleatórias, entre outras coisas, então a proporção de sucessos entre eles também é uma variável aleatória. Há outro grupo de estatísticos que acredita que em muitos problemas não é apropriado atribuir uma distribuição a um parâmetro, mas em vez disso, afirma que o valor verdadeiro do parâmetro é um certo número fixo cujo valor por acaso é desconhecido para o experimentador. Esses estatísticos atribuem uma distribuição a um parâmetro apenas quando há extensa informação prévia sobre as frequências relativas com que parâmetros similares tomaram cada um de seus valores possíveis em experimentos passados. Se dois cientistas diferentes pudessem concordar sobre quais experimentos passados eram similares ao experimento atual, então eles poderiam concordar sobre uma distribuição a ser atribuída ao parâmetro. Por exemplo, suponha que a proporção θ de itens defeituosos em um grande lote manufaturado seja desconhecida. Suponha também que o mesmo fabricante produziu muitos desses lotes de itens no passado e que registros detalhados foram mantidos sobre as proporções de itens defeituosos em lotes passados. As frequências relativas para lotes passados poderiam então ser usadas para construir uma distribuição para θ. Estatísticos que argumentariam desta forma são ditos aderir à filosofia frequentista de estatística e são chamados de *frequentistas*.

Os frequentistas baseiam-se na suposição de que existem sequências infinitas de variáveis aleatórias para dar sentido à maioria de suas declarações de probabilidade. Uma vez que se assume a existência de tal sequência infinita, descobre-se que os parâmetros das distribuições que estão sendo usadas são limites de funções das sequências infinitas, assim como fazem os Bayesianos descritos acima. Desta forma, os parâmetros são variáveis aleatórias porque são funções de outras variáveis aleatórias. O ponto de desacordo entre os dois grupos é se é útil ou mesmo possível atribuir uma distribuição a tais parâmetros.

Tanto Bayesianos quanto frequentistas concordam sobre a utilidade de famílias de distribuições para observações indexadas por parâmetros. Os Bayesianos referem-se à distribuição indexada pelo valor do parâmetro θ como a distribuição condicional das observações dado que o parâmetro é igual a θ. Os frequentistas referem-se à distribuição indexada por θ como a distribuição das observações quando θ é o valor verdadeiro do parâmetro. Os dois grupos concordam que sempre que uma distribuição pode ser atribuída a um parâmetro, a teoria e os métodos a serem descritos neste capítulo são aplicáveis e úteis. Nas Seções 7.2–7.4, nós explicitamente assumiremos que cada parâmetro é uma variável aleatória e atribuiremos a ele uma distribuição que representa as probabilidades de que o parâmetro esteja em vários subconjuntos do espaço de parâmetros. A partir da Seção 7.5, consideraremos técnicas de estimação que não se baseiam na atribuição de distribuições a parâmetros.

7.2 Distribuições a Priori e a Posteriori

A distribuição de um parâmetro antes da observação de quaisquer dados é chamada de distribuição *a priori*. A distribuição condicional do parâmetro, dados os valores observados, é chamada de distribuição *a posteriori*. Se inserirmos os valores observados dos dados na f.d.p. (função de densidade de probabilidade) ou f.p. (função de probabilidade) condicional, e considerarmos o resultado como uma função apenas do parâmetro, o resultado é chamado de função de *verossimilhança*.

A Distribuição a Priori

Exemplo 7.2.1 Tempo de Vida de Componentes Eletrônicos. No Exemplo 7.1.1, os tempos de vida $X_1, X_2, \ldots$ de componentes eletrônicos foram modelados como variáveis aleatórias i.i.d. exponenciais com parâmetro θ condicional a θ, e θ foi interpretado como a taxa de falha dos componentes. Notamos que $n/\sum_{i=1}^{n} X_i$ deveria convergir em probabilidade para θ quando $n \to \infty$. Dissemos então que θ tinha a distribuição gama com parâmetros 1 e 2.

A distribuição de θ mencionada no final do Exemplo 7.2.1 foi atribuída antes de se observar a vida útil de qualquer componente. Por essa razão, chamamos isso de *distribuição a priori*.

Definição 7.2.1 Distribuição a Priori/f.p./f.d.p. Suponha que se tenha um modelo estatístico com parâmetro θ. Se trata θ como uma variável aleatória, então a distribuição que se atribui a θ antes de observar quaisquer outras variáveis aleatórias de interesse é chamada de sua *distribuição a priori*. Se o espaço de parâmetros for no máximo contável, então a distribuição a priori é discreta e sua f.p. é chamada de *f.p. a priori* de θ. Se a distribuição a priori for contínua, então sua f.d.p. é chamada de *f.d.p. a priori* de θ. Usaremos comumente o símbolo $\xi(\theta)$ para denotar a f.p. ou f.d.p. a priori de θ.

Quando se trata um parâmetro como uma variável aleatória, o nome "distribuição a priori" é meramente outro nome para a distribuição marginal do parâmetro.

Exemplo 7.2.2 Moeda Justa ou de Duas Caras. Seja θ a probabilidade de obter uma cara quando uma certa moeda é lançada, e suponha que se saiba que a moeda é justa ou tem cara em ambos os lados. Portanto, os únicos valores possíveis de θ são $\theta = 1/2$ e $\theta = 1$. Se a probabilidade a priori de que a moeda é justa for 0.8, então a f.p. a priori de θ é $\xi(1/2) = 0.8$ e $\xi(1) = 0.2$.

Exemplo 7.2.3 Proporção de Itens Defeituosos. Suponha que a proporção θ de itens defeituosos em um grande lote manufaturado seja desconhecida e que a distribuição a priori atribuída a θ seja a distribuição uniforme no intervalo $[0, 1]$. Então a f.d.p. a priori de θ é

$$\xi(\theta) = \begin{cases} 1 & \text{para } 0 < \theta < 1, \\ 0 & \text{caso contrário.} \end{cases} \tag{7.2.1}$$

A distribuição a priori de um parâmetro θ deve ser uma distribuição de probabilidade sobre o espaço de parâmetros Ω. Assumimos que o experimentador ou estatístico será capaz de resumir seu conhecimento prévio e crenças sobre onde o valor de θ provavelmente se encontra em Ω na forma de uma distribuição a priori para θ. Ou seja, antes que os dados experimentais tenham sido coletados ou observados, a experiência e o conhecimento passados do experimentador o levarão a acreditar que θ tem maior probabilidade de estar em certas regiões de Ω do que em outras. Assumiremos que as verossimilhanças relativas das diferentes regiões podem ser expressas em termos de uma distribuição de probabilidade em Ω, ou seja, a distribuição a priori de θ.

Exemplo 7.2.4 Tempo de Vida de Lâmpadas Fluorescentes. Suponha que os tempos de vida (em horas) de lâmpadas fluorescentes de um certo tipo devam ser observados e que o tempo de vida de qualquer lâmpada em particular tenha a distribuição exponencial com parâmetro θ. Suponha também que o valor exato de θ seja desconhecido, e com base na experiência prévia, a distribuição a priori de θ seja considerada a distribuição gama para a qual a média é 0.0002 e o desvio padrão é 0.0001. Determinaremos a f.d.p. a priori de θ. Suponha que a distribuição a priori de θ seja a distribuição gama com parâmetros α_0 e β_0. Foi mostrado no Teorema 5.7.4 que a média desta distribuição é α_0/β_0 e a variância é α_0/β_0^2. Portanto, $\alpha_0/\beta_0 = 0.0002$ e $\alpha_0/\beta_0^2 = (0.0001)^2$. Essas duas equações resultam em $\alpha_0 = 4$ e $\beta_0 = 20.000$. Segue-se de Eq. (5.7.13) que a f.d.p. a priori de θ para $\theta > 0$ é a seguinte:

$$\xi(\theta) = \frac{(20.000)^4}{3!}\theta^3 e^{-20.000\theta}. \tag{7.2.2}$$

Além disso, $\xi(\theta) = 0$ para $\theta \leq 0$.

No restante desta seção e nas Seções 7.3 e 7.4, focaremos em problemas de inferência estatística nos quais o parâmetro θ é uma variável aleatória e, portanto, precisa ter uma distribuição atribuída. Referir-nos-emos à distribuição indexada por θ para as outras variáveis aleatórias de interesse como a distribuição condicional para essas variáveis aleatórias dado θ. Esta é precisamente a linguagem usada no Exemplo 7.2.1 onde o parâmetro é θ, a taxa de falha. Referindo-se à f.p. ou f.d.p. condicional de variáveis aleatórias condicionais e

suas f.p.s e f.d.p.s. não condicionais, usaremos a notação da Seção 7.2.1. Por exemplo, se seja $\mathbf{X} = (X_1, \ldots, X_m)$ no Exemplo 7.2.1, a f.d.p. condicional de $\mathbf{X}$ dado θ é

$$f_m(\mathbf{x}|\theta) = \begin{cases} \theta^m \exp(-\theta[x_1 + \cdots + x_m]) & \text{para todos } x_i > 0, \\ 0 & \text{caso contrário.} \end{cases} \tag{7.2.3}$$

Em muitos problemas, como no Exemplo 7.2.1, os dados observáveis $X_1, X_2, \ldots$ são modelados como uma amostra aleatória de uma distribuição univariada indexada por θ. Nestes casos, seja $f(x|\theta)$ a f.p. ou f.d.p. de uma única variável aleatória sob a distribuição indexada por θ. Em tal caso, usando a notação acima,

$$f_m(\mathbf{x}|\theta) = f(x_1|\theta) \cdots f(x_m|\theta).$$

Quando tratamos θ como uma variável aleatória, $f(x_i|\theta)$ é a f.p. ou f.d.p. condicional de cada observação X_i dado θ, e as observações são condicionalmente i.i.d. dado θ. Em resumo, as duas expressões a seguir devem ser entendidas como equivalentes:

- $X_1, \ldots, X_n$ formam uma amostra aleatória com f.p. ou f.d.p. $f(x|\theta)$.
- $X_1, \ldots, X_n$ são condicionalmente i.i.d. dado θ com f.p. ou f.d.p. condicional $f(x|\theta)$.

Embora geralmente usemos a primeira expressão por simplicidade, é frequente que a segunda expressão seja útil para lembrar que as duas expressões são equivalentes quando tratamos θ como uma variável aleatória.

Análise de Sensibilidade e Prioris Impróprias

No Exemplo 2.3.8 na página 84, vimos uma situação em que dois conjuntos muito diferentes de probabilidades a priori foram usados para uma coleção de eventos. Após a observação dos dados, no entanto, as probabilidades a posteriori eram bastante semelhantes. No Exemplo 5.8.4 na página 330, usamos uma grande coleção de distribuições a priori para a probabilidade de um parâmetro a fim de ver o quanto o impacto de uma distribuição a priori sobre a probabilidade a posteriori de um único evento importante. É uma prática comum comparar as distribuições a posteriori que surgem de várias distribuições a priori diferentes para ver o quanto o efeito da distribuição a priori tem sobre as respostas a questões importantes. Tais comparações são chamadas de *análise de sensibilidade*.

É muito comum o caso de que diferentes distribuições a priori não fazem muita diferença depois que os dados foram observados. Isso é especialmente verdadeiro se houver muitos dados ou se as distribuições a priori que estão sendo comparadas são muito dispersas. Essa observação tem duas implicações importantes. Primeiro, o fato de que diferentes experimentadores podem não concordar sobre uma distribuição a priori torna-se menos importante se houver muitos dados. Segundo, os experimentadores podem estar menos inclinados

a gastar tempo especificando uma distribuição a priori se não for fazer muita diferença qual deles é especificado. Infelizmente, se não se especifica alguma distribuição a priori, não há como calcular uma distribuição condicional do parâmetro dados os dados.

Como um expediente, existem alguns cálculos disponíveis que tentam capturar a ideia de que os dados contêm muito mais informações do que as disponíveis a priori. Geralmente, esses cálculos envolvem o uso de uma função $\xi(\theta)$ como se fosse uma f.d.p. a priori para o parâmetro θ, mas tal que $\int \xi(\theta)d\theta = \infty$, o que viola claramente a definição de f.d.p. Tais prioris são chamadas de *impróprias*. Discutiremos prioris impróprias mais detalhadamente na Seção 7.3.

A Distribuição a Posteriori

Exemplo 7.2.5 Tempo de Vida de Lâmpadas Fluorescentes. No Exemplo 7.2.4, construímos uma distribuição a priori para o parâmetro θ que especifica a distribuição exponencial para uma coleção de tempos de vida de lâmpadas fluorescentes. Suponha que observemos uma coleção de n tais tempos de vida. Como mudaríamos a distribuição de θ para levar em conta os dados observados?

Definição 7.2.2 Distribuição/f.p./f.d.p. a Posteriori. Considere um problema de inferência estatística com parâmetro θ e variáveis aleatórias $X_1, \ldots, X_n$, a serem observadas. A distribuição condicional de θ dados $X_1, \ldots, X_n$ é chamada de *distribuição a posteriori* de θ. A f.p. ou f.d.p. condicional de θ dados $X_1 = x_1, \ldots, X_n = x_n$ é chamada de *f.p. a posteriori* ou *f.d.p. a posteriori* de θ e é tipicamente denotada por $\xi(\theta|x_1, \ldots, x_n)$.

Quando se trata o parâmetro como uma variável aleatória, o nome "distribuição a posteriori" é meramente outro nome para a distribuição condicional do parâmetro dados os dados. O teorema de Bayes para variáveis aleatórias (3.6.13) e para vetores aleatórios (3.7.15) nos diz como derivar a f.p. ou f.d.p. a posteriori de θ após observar os dados. Reafirmaremos o teorema de Bayes aqui usando a notação específica de distribuições e parâmetros a priori.

Teorema 7.2.1 Suponha que as n variáveis aleatórias $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição para a qual a f.d.p. ou a f.p. é $f(x|\theta)$. Suponha também que o valor do parâmetro θ seja desconhecido e a f.p. ou f.d.p. a priori de θ seja $\xi(\theta)$. Então a f.d.p. ou f.p. a posteriori de θ é

$$\xi(\theta|\mathbf{x}) = \frac{f(x_1|\theta)\cdots f(x_n|\theta)\xi(\theta)}{g_n(\mathbf{x})} \quad \text{para } \theta \in \Omega,$$

onde g_n é a f.d.p. ou f.p. conjunta marginal de $X_1, \ldots, X_n$.

Prova Por simplicidade, assumiremos que o espaço de parâmetros Ω é um intervalo da reta real ou a reta real inteira e que $\xi(\theta)$ é uma f.d.p. a priori, em vez de uma f.p. a priori. No entanto, a prova que será dada aqui pode ser facilmente adaptada a um problema em que $\xi(\theta)$ é uma f.p. Uma vez que as variáveis aleatórias $X_1, \ldots, X_n$ formam uma amostra aleatória da distribuição para a qual a f.d.p. é $f(x|\theta)$, segue-se que sua f.d.p. ou f.p. conjunta condicional dado θ é

$$f_n(x_1, \ldots, x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta). \tag{7.2.4}$$

Se usarmos a notação vetorial $\mathbf{x} = (x_1, \ldots, x_n)$, então a f.d.p. conjunta em Eq. (7.2.4) pode ser escrita mais compactamente como $f_n(\mathbf{x}|\theta)$. Eq. (7.2.4) expressa meramente o fato de que $X_1, \ldots, X_n$ são condicionalmente independentes e identicamente distribuídas dado θ, cada uma tendo f.d.p. ou f.p. $f(x|\theta)$. Se multiplicarmos a f.d.p. ou f.p. conjunta de θ por a f.d.p. de $\xi(\theta)$, obtemos a f.d.p. ou f.p. conjunta $(n+1)$-dimensional de $X_1, \ldots, X_n$ e θ na forma

$$f(\mathbf{x}, \theta) = f_n(\mathbf{x}|\theta)\xi(\theta). \tag{7.2.5}$$

A f.d.p. ou f.p. conjunta marginal de $X_1, \ldots, X_n$ pode agora ser obtida integrando o lado direito da Eq. (7.2.5) sobre todos os valores de θ. Portanto, a f.d.p. ou f.p. conjunta marginal n-dimensional de $X_1, \ldots, X_n$ pode ser escrita na forma

$$g_n(\mathbf{x}) = \int_\Omega f_n(\mathbf{x}|\theta)\xi(\theta)d\theta. \tag{7.2.6}$$

Eq. (7.2.6) é apenas uma instância da lei da probabilidade total para variáveis aleatórias (3.7.14). Ademais, a f.d.p. condicional de θ dado que $X_1 = x_1, \ldots, X_n = x_n$, a saber, $\xi(\theta|\mathbf{x})$, deve ser igual a $f(\mathbf{x}, \theta)$ dividido por $g_n(\mathbf{x})$. Assim, temos

$$\xi(\theta|\mathbf{x}) = \frac{f_n(\mathbf{x}|\theta)\xi(\theta)}{g_n(\mathbf{x})} \quad \text{para } \theta \in \Omega, \tag{7.2.7}$$

que é o teorema de Bayes reafirmado para parâmetros e amostras aleatórias. Se $\xi(\theta)$ é uma f.p., de modo que a distribuição a priori é discreta, basta substituir a integral em (7.2.6) pela soma sobre todos os valores possíveis de θ. ■

Exemplo 7.2.6 Tempo de Vida de Lâmpadas Fluorescentes. Suponha novamente, como nos Exemplos 7.2.4 e 7.2.5, que a distribuição dos tempos de vida de lâmpadas fluorescentes de um certo tipo seja a distribuição exponencial com parâmetro θ, e a distribuição a priori de θ seja uma distribuição gama particular para a qual a f.d.p. $\xi(\theta)$ é dada por Eq. (7.2.2). Suponha também que os tempos de vida de n lâmpadas deste tipo sejam observados. Determinaremos a f.d.p. a posteriori de θ dado que $X_1 = x_1, \ldots, X_n = x_n$. Pela Eq. (5.7.16), a f.d.p. de cada observação X_i é

$$f(x|\theta) = \begin{cases} \theta e^{-\theta x} & \text{para } x > 0, \\ 0 & \text{caso contrário.} \end{cases}$$

A f.d.p. conjunta de $X_1, \ldots, X_n$ pode ser escrita na seguinte forma, para $x_i > 0\,(i = 1, \ldots, n)$:

$$f_n(\mathbf{x}|\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i} = \theta^n e^{-\theta y},$$

onde $y = \sum_{i=1}^{n} x_i$. Como $f_n(\mathbf{x}|\theta)$ será usado na construção da distribuição a posteriori de θ, é agora aparente que a estatística $Y = \sum_{i=1}^{n} X_i$ será usada em qualquer inferência que faça uso da distribuição a posteriori.

Uma vez que a f.d.p. a priori $\xi(\theta)$ é dada por Eq. (7.2.2), segue-se que para $\theta > 0$,

$$f_n(\mathbf{x}|\theta)\xi(\theta) = \theta^n e^{-\theta y} \frac{(20.000)^4}{3!} \theta^3 e^{-20.000\theta} = \frac{(20.000)^4}{3!} \theta^{n+3} e^{-(y+20.000)\theta}. \tag{7.2.8}$$

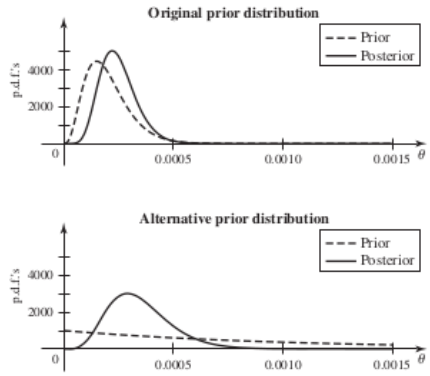
Precisamos calcular $g_n(\mathbf{x})$, que é a integral de (7.2.8) sobre todo θ:

$$g_n(\mathbf{x}) = \int_0^\infty \frac{(20.000)^4}{3!} \theta^{n+3} e^{-(y+20.000)\theta} d\theta = \frac{(20.000)^4}{3!} \frac{\Gamma(n+4)}{(y+20.000)^{n+4}}$$

onde a última igualdade segue do Teorema 5.7.3. Portanto,

$$\begin{aligned} \xi(\theta|\mathbf{x}) &= \frac{\frac{(20.000)^4}{3!} \theta^{n+3} e^{-(y+20.000)\theta}}{\frac{(20.000)^4}{3!} \frac{\Gamma(n+4)}{(y+20.000)^{n+4}}} \\ &= \frac{(y+20.000)^{n+4}}{\Gamma(n+4)} \theta^{n+3} e^{-(y+20.000)\theta}, \end{aligned} \tag{7.2.9}$$

para $\theta > 0$. Quando comparamos esta expressão com Eq. (5.7.13), podemos ver que é a f.d.p. da distribuição gama com parâmetros $n+4$ e $y+20.000$. Portanto, esta distribuição gama é a distribuição a posteriori de θ. Como um exemplo específico, suponha que observamos os seguintes $n = 5$ tempos de vida em horas: 2911, 4403, 3237, 5509 e 3118. Então $y = 16.178$, e a distribuição a posteriori de θ é a distribuição gama com parâmetros 9 e 36.178. O painel superior da Fig. 7.1 exibe tanto a f.d.p. a priori quanto a posteriori neste exemplo. Fica claro a partir dos dados que os dados fizeram com que a distribuição de θ mudasse um pouco da priori para a posteriori. Neste ponto, pode ser apropriado realizar uma análise de sensibilidade. Por exemplo, como a distribuição a posteriori mudaria se tivéssemos escolhido uma distribuição a priori diferente? Para ser específico, considere a priori gama com parâmetros 1 e 1000. Esta priori tem o mesmo desvio padrão da priori original, mas a média é cinco vezes maior. A distribuição a posteriori seria então a distribuição gama com parâmetros 6 e 17.178. As f.d.p.s desta priori e posteriori estão no painel inferior da Fig. 7.1. Pode-se ver que tanto a priori quanto a posteriori no painel inferior estão mais espalhadas do que suas contrapartes no painel superior.



É claro que a escolha da priori fará diferença com este pequeno conjunto de dados. Os nomes "a priori" e "a posteriori" derivam das palavras latinas para "anterior" e "posterior". A distribuição a priori é a distribuição de θ que vem antes da observação dos dados, e a distribuição a posteriori vem depois da observação dos dados.

A Função de Verossimilhança

O denominador no lado direito da Eq. (7.2.7) é simplesmente a integral do numerador sobre todos os valores possíveis de θ. Embora o valor desta integral dependa dos valores observados $x_1, \ldots, x_n$, ele não depende de θ e pode ser tratado como uma constante quando o lado direito da Eq. (7.2.7) é considerado como uma f.d.p. de θ. Podemos, portanto, substituir a Eq. (7.2.7) pela seguinte relação:

$$\xi(\theta|\mathbf{x}) \propto f_n(\mathbf{x}|\theta)\xi(\theta). \tag{7.2.10}$$

O símbolo de proporcionalidade $\propto$ é usado aqui para indicar que o lado esquerdo é igual ao lado direito, exceto possivelmente por um fator constante, cujo valor pode depender dos valores observados $x_1, \ldots, x_n$, mas não depende de θ. A constante apropriada que estabelecerá a igualdade dos dois lados na relação (7.2.10) pode ser determinada a qualquer momento usando o fato de que $\int_\Omega \xi(\theta|\mathbf{x})d\theta = 1$, porque $\xi(\theta|\mathbf{x})$ é uma f.d.p. de θ. Uma das duas funções no lado direito da Eq. (7.2.10) é a f.d.p. a priori de θ. A outra função também tem um nome especial.

Definição 7.2.3 Função de Verossimilhança. Quando a f.p. ou f.d.p. conjunta $f_n(\mathbf{x}|\theta)$ das observações em uma amostra aleatória é considerada como uma função de θ para valores dados de $x_1, \ldots, x_n$, ela é chamada de *função de verossimilhança*.

A relação (7.2.10) afirma que a f.d.p. a posteriori de θ é proporcional ao produto da função de verossimilhança e da f.d.p. a priori de θ. Usando a relação de proporcionalidade (7.2.10), muitas vezes é possível determinar a f.d.p. a posteriori de θ sem realizar explicitamente a integração em Eq. (7.2.6). Se pudermos reconhecer o lado direito da relação (7.2.10) como sendo, exceto por uma das f.d.p.s padrão introduzidas no Capítulo 5 ou em outro lugar neste livro, exceto possivelmente por um fator constante, então podemos facilmente determinar o fator apropriado que converterá o lado direito de (7.2.10) em uma f.d.p. adequada de θ. Ilustraremos essas ideias considerando novamente o Exemplo 7.2.3.

Exemplo 7.2.7 Proporção de Itens Defeituosos. Suponha novamente, como no Exemplo 7.2.3, que a proporção θ de itens defeituosos em um grande lote manufaturado seja desconhecida e que a distribuição a priori de θ seja uma distribuição uniforme no intervalo $[0, 1]$. Suponha também que uma amostra aleatória de n itens seja retirada do lote, e para $i = 1, \ldots, n$, seja $X_i = 1$ se o i-ésimo item for defeituoso, e seja $X_i = 0$ caso contrário. Então $X_1, \ldots, X_n$ formam n ensaios de Bernoulli com parâmetro θ. Determinaremos a f.d.p. a posteriori de θ. Segue-se da Eq. (5.2.2) que a f.p. de cada observação X_i é

$$f(x|\theta) = \begin{cases} \theta^x(1-\theta)^{1-x} & \text{para } x = 0, 1, \\ 0 & \text{caso contrário.} \end{cases}$$

Portanto, se seja $y = \sum_{i=1}^{n} x_i$, então a f.p. conjunta de $X_1, \ldots, X_n$ pode ser escrita na seguinte forma para $x_i = 0$ ou $1\,(i = 1, \ldots, n)$:

$$f_n(\mathbf{x}|\theta) = \theta^y(1-\theta)^{n-y}. \tag{7.2.11}$$

Uma vez que a f.d.p. a priori $\xi(\theta)$ é dada por Eq. (7.2.1), segue-se que para $0 < \theta < 1$,

$$f_n(\mathbf{x}|\theta)\xi(\theta) = \theta^y(1-\theta)^{n-y}. \tag{7.2.12}$$

Quando comparamos esta expressão com Eq. (5.8.3), podemos ver que, exceto por um fator constante, é a f.d.p. da distribuição beta com parâmetros $\alpha = y+1$ e $\beta = n-y+1$. Uma vez que a f.d.p. a posteriori $\xi(\theta|\mathbf{x})$ é proporcional ao lado direito da Eq. (7.2.12), segue-se que $\xi(\theta|\mathbf{x})$ deve ser a f.d.p. da distribuição beta com parâmetros $a = y + 1$ e $b = n - y + 1$. Portanto, para $0 < \theta < 1$,

$$\xi(\theta|\mathbf{x}) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y}. \tag{7.2.13}$$

Neste exemplo, a estatística $Y = \sum_{i=1}^{n} X_i$ está sendo usada para construir a distribuição a posteriori, e, portanto, será usada em qualquer inferência que se baseie na distribuição a posteriori.

Nota: Constante de Normalização para f.d.p. a Posteriori. Os passos que nos levaram de (7.2.12) para (7.2.13) são um exemplo de uma técnica muito comum para determinar uma f.d.p. a posteriori. Pode-se extrair qualquer fator constante inconveniente da f.p. ou f.d.p. a priori e da função de verossimilhança antes de multiplicá-los juntos como em (7.2.10). Então, olhamos para o produto resultante, chame-o de $g(\theta)$, para ver se o reconhecemos como se parecendo com parte de uma f.d.p. que já vimos. Se, de fato, encontrarmos uma f.d.p. nomeada com a qual estamos familiarizados que seja igual a $cg(\theta)$, então nossa f.d.p. a posteriori também é $cg(\theta)$, e nossa distribuição a posteriori tem o nome correspondente, assim como no Exemplo 7.2.7.

Observações Sequenciais e Previsão

Em muitos experimentos, as observações $X_1, \ldots, X_n$, que formam a amostra aleatória, devem ser obtidas sequencialmente, ou seja, uma de cada vez. Em tal experimento, o valor de X_1 é observado primeiro, o valor de X_2 é observado em seguida, o valor de X_3 é então observado, e assim por diante. Suponha que a f.d.p. a posteriori do parâmetro θ após o valor de x_1 ter sido observado, possa ser calculada da maneira usual a partir da relação

$$\xi(\theta|x_1) \propto f(x_1|\theta)\xi(\theta). \tag{7.2.14}$$

Como X_1 e X_2 são condicionalmente independentes dado θ, a f.p. ou f.d.p. condicional de X_2 dado θ e $X_1 = x_1$ é a mesma que dado θ apenas, a saber, $f(x_2|\theta)$. Portanto, a f.d.p. a posteriori de θ na Eq. (7.2.14) serve como a f.d.p. a priori de θ quando o valor de X_2 está para ser observado. Assim, após o valor de x_2 ter sido observado, a f.d.p. a posteriori de θ pode ser calculada a partir da relação

$$\xi(\theta|x_1, x_2) \propto f(x_2|\theta)\xi(\theta|x_1). \tag{7.2.15}$$

Podemos continuar desta forma, calculando uma f.d.p. a posteriori atualizada de θ após cada observação e usando essa f.d.p. como a f.d.p. a priori de θ para a próxima observação. A f.d.p. a posteriori $\xi(\theta|x_1, \ldots, x_{n-1})$ após os valores $x_1, \ldots, x_{n-1}$ terem sido observados, será em última análise a f.d.p. a priori de θ para o valor final observado x_n. A f.d.p. a posteriori após todos os n valores $x_1, \ldots, x_n$ terem sido observados, será, portanto, especificada pela relação

$$\xi(\theta|x_1, \ldots, x_n) \propto f(x_n|\theta)\xi(\theta|x_1, \ldots, x_{n-1}). \tag{7.2.16}$$

Alternativamente, depois de todos os n valores $x_1, \ldots, x_n$ terem sido observados, poderíamos calcular a f.d.p. a posteriori de θ da maneira usual, combinando a f.d.p. ou f.p. conjunta $f_n(\mathbf{x}|\theta)$ com a f.d.p. a priori original $\xi(\theta)$, como indicado em Eq. (7.2.7). Pode ser mostrado (ver Exercício 8) que a f.d.p. a posteriori $\xi(\theta|\mathbf{x})$ será a mesma, independentemente de ser calculada diretamente usando Eq. (7.2.7) ou sequencialmente usando Eqs. (7.2.14), (7.2.15) e (7.2.16). Esta propriedade foi ilustrada na Seção 2.3 (ver página 80) para uma moeda que se sabe ser justa ou ter cara em ambos os lados. Após cada lançamento da moeda,

a probabilidade a posteriori de a moeda ser justa é atualizada. As constantes de proporcionalidade nas Eqs. (7.2.14)-(7.2.16) têm uma interpretação útil. Por exemplo, em (7.2.16) a constante de proporcionalidade é 1 sobre a integral do lado direito com respeito a θ. Mas esta integral é a f.d.p. ou f.p. condicional de X_n dado $X_1 = x_1, \ldots, X_{n-1} = x_{n-1}$ de acordo com a versão condicional da lei da probabilidade total (3.7.16). Por exemplo, se θ tem uma distribuição contínua,

$$f(x_n|x_1, \ldots, x_{n-1}) = \int f(x_n|\theta)\xi(\theta|x_1, \ldots, x_{n-1})d\theta. \tag{7.2.17}$$

Se estivermos interessados em prever a n-ésima observação após observar as primeiras $n-1$, podemos usar (7.2.17), que também é 1 sobre a constante de proporcionalidade em Eq. (7.2.16), como a f.d.p. ou f.p. condicional de X_n dados os primeiros $n-1$ observáveis.

Exemplo 7.2.8 Tempo de Vida de Lâmpadas Fluorescentes. No Exemplo 7.2.6, condicional a θ, os tempos de vida de lâmpadas fluorescentes são variáveis aleatórias exponenciais independentes com parâmetro θ. Também observamos os tempos de vida de cinco lâmpadas, e a distribuição a posteriori de θ foi encontrada como sendo a distribuição gama com parâmetros 9 e 36.178. Suponha que queiramos prever o tempo de vida X_6 da próxima lâmpada. A f.d.p. condicional de X_6, o tempo de vida da próxima lâmpada, dadas as primeiras cinco vidas, integra o produto de $\xi(\theta|\mathbf{x})$ e $f(x_6|\theta)$ em relação a θ. A f.d.p. a posteriori de θ é $\xi(\theta|\mathbf{x}) = 2.633 \times 10^{36}\theta^8 e^{-36.178\theta}$ para $\theta > 0$. Então, para $x_6 > 0$

$$\begin{aligned} f(x_6|\mathbf{x}) &= \int_0^\infty 2.633 \times 10^{36}\theta^8 e^{-36.178\theta}\theta e^{-x_6\theta}d\theta \\ &= 2.633 \times 10^{36}\int_0^\infty \theta^9 e^{-(x_6+36.178)\theta}d\theta \\ &= 2.633 \times 10^{36}\frac{\Gamma(10)}{(x_6+36.178)^{10}} = \frac{9.555 \times 10^{41}}{(x_6+36.178)^{10}}. \end{aligned} \tag{7.2.18}$$

Podemos usar esta f.d.p. para realizar qualquer cálculo que desejarmos sobre a distribuição de X_6 dados os tempos de vida observados. Por exemplo, a probabilidade de que a sexta lâmpada dure mais de 3000 horas é igual a

$$\Pr(X_6 > 3000|\mathbf{x}) = \int_{3000}^\infty \frac{9.555 \times 10^{41}}{9 \times 39.178^9}dx_6 = \frac{9.555 \times 10^{41}}{9 \times 39.178^9} = 0.4882.$$

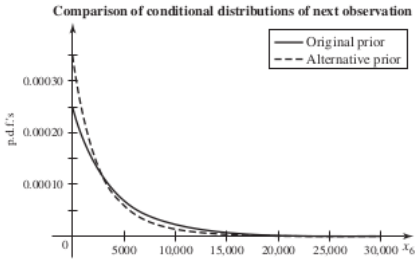
Podemos continuar a análise de sensibilidade que foi iniciada no Exemplo 7.2.6. É importante saber a probabilidade de que o próximo tempo de vida seja de pelo menos 3000, podemos ver quanta influência a escolha da distribuição a priori teve nesta computação. Usando a segunda distribuição a priori (gama com parâmetros 1 e 1000), descobrimos que a distribuição a posteriori de θ era

a gama com parâmetros 6 e 17.178. Poderíamos calcular a f.d.p. condicional de X_6 dados os dados observados da mesma forma que fizemos com a priori original, e seria

$$f(x_6|\mathbf{x}) = \frac{1.542 \times 10^{26}}{(x_6 + 17.178)^7}, \quad \text{para } x_6 > 0. \tag{7.2.19}$$

Com esta f.d.p., a probabilidade de que $X_6 > 3000$ é

$$\Pr(X_6 > 3000|\mathbf{x}) = \int_{3000}^{\infty} \frac{1.542 \times 10^{26}}{(x_6 + 17.178)^7} dx_6 = \frac{1.542 \times 10^{26}}{6 \times 20.178^6} = 0.3807.$$



Como notamos no final do Exemplo 7.2.6, as diferentes prioris fazem uma diferença considerável nas inferências que podemos fazer. É importante ter um valor preciso de $\Pr(X_6 > 3000|\mathbf{x})$, precisamos de uma amostra maior. Os dois f.d.p.s diferentes de X_6 podem ser comparados na Fig. 7.2. A f.d.p. de Eq. (7.2.18) é maior para valores intermediários de x_6, enquanto a de Eq. (7.2.19) é maior para os valores extremos de x_6.

Resumo

A distribuição a priori de um parâmetro descreve nossa incerteza sobre o parâmetro antes de observar quaisquer dados. A função de verossimilhança é a f.d.p. ou f.p. condicional dos dados, considerada como uma função do parâmetro dados os dados. A verossimilhança nos diz o quanto os dados alteram nossa incerteza. Valores grandes da verossimilhança corresponderão a valores de parâmetro onde a posteriori será maior do que a priori. Valores baixos da verossimilhança ocorrerão em valores de parâmetro onde a posteriori será menor do que a priori. A distribuição a posteriori do parâmetro é a distribuição condicional do parâmetro dados os dados. Ela é obtida usando o teorema de Bayes para variáveis aleatórias, que vimos pela primeira vez na página 148. Podemos

prever observações futuras que são condicionalmente independentes dos dados observados dado θ usando a versão condicional da lei da probabilidade total que vimos na página 163.

Exercícios

1. Considere novamente a situação descrita no Exemplo 7.2.8. Desta vez, suponha que o experimentador acredite que a distribuição a priori de θ é a distribuição gama com parâmetros 1 e 5000. Que valor o experimentador calcularia para $\Pr(X_6 > 3000|\mathbf{x})$?

2. Suponha que a proporção θ de itens defeituosos em um grande lote manufaturado seja 0,1 ou 0,2, e que a f.p. (função de probabilidade) a priori de θ seja a seguinte:

$$\xi(0.1) = 0.7 \quad \text{e} \quad \xi(0.2) = 0.3.$$

   Suponha também que, quando oito itens são selecionados aleatoriamente do lote, descobre-se que exatamente dois deles são defeituosos. Determine a f.p. a posteriori de θ.

3. Suponha que o número de defeitos em um rolo de fita de gravação magnética tenha uma distribuição de Poisson para a qual a média λ é 1,0 ou 1,5, e a f.p. a priori de λ é a seguinte:

$$\xi(1.0) = 0.4 \quad \text{e} \quad \xi(1.5) = 0.6.$$

   Se um rolo de fita selecionado aleatoriamente apresentar três defeitos, qual é a f.p. a posteriori de λ?

4. Suponha que a distribuição a priori de algum parâmetro θ seja uma distribuição gama para a qual a média é 10 e a variância é 5. Determine a f.d.p. (função de densidade de probabilidade) a priori de θ.

5. Suponha que a distribuição a priori de algum parâmetro θ seja uma distribuição beta para a qual a média é 1/3 e a variância é 1/45. Determine a f.d.p. a priori de θ.

6. Suponha que a proporção θ de itens defeituosos em um grande lote manufaturado seja desconhecida, e a distribuição a priori de θ seja a distribuição uniforme no intervalo $[0, 1]$. Quando oito itens são selecionados aleatoriamente do lote, descobre-se que exatamente três deles são defeituosos. Determine a distribuição a posteriori de θ.

7. Considere novamente o problema descrito no Exercício 6, mas suponha agora que a f.d.p. a priori de θ seja a seguinte:

$$\xi(\theta) = \begin{cases} 2(1-\theta) & \text{para } 0 < \theta < 1, \\ 0 & \text{caso contrário.} \end{cases}$$

Como no Exercício 6, suponha que em uma amostra aleatória de oito itens, exatamente três sejam defeituosos. Determine a distribuição a posteriori de θ.

8. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição para a qual a f.d.p. é $f(x|\theta)$, o valor de θ é desconhecido, e a f.d.p. a priori de θ é $\xi(\theta)$. Mostre que a f.d.p. a posteriori $\xi(\theta|\mathbf{x})$ é a mesma, quer seja calculada diretamente usando a Eq. (7.2.7) ou sequencialmente usando as Eqs. (7.2.14), (7.2.15) e (7.2.16).

9. Considere novamente o problema descrito no Exercício 6, e assuma a mesma distribuição a priori de θ. Suponha, no entanto, que em vez de selecionar uma amostra aleatória de oito itens do lote, realizemos o seguinte experimento: os itens do lote são selecionados aleatoriamente um a um até que exatamente três defeituosos tenham sido encontrados. Se descobrirmos que devemos selecionar um total de oito itens neste processo, qual é a distribuição a posteriori de θ no final do experimento?

10. Suponha que uma única observação X deva ser retirada da distribuição uniforme no intervalo $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, o valor de θ é desconhecido, e a distribuição a priori de θ é a distribuição uniforme no intervalo $[10, 20]$. Se o valor observado de X for 12, qual é a distribuição a posteriori de θ?

11. Considere novamente as condições do Exercício 10, e assuma a mesma distribuição a priori de θ. Suponha, no entanto, que seis observações sejam selecionadas aleatoriamente da distribuição uniforme no intervalo $[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, e seus valores sejam 11.0, 11.5, 11.7, 11.1, 11.4 e 10.9. Determine a distribuição a posteriori de θ.

7.3 Distribuições a Priori Conjugadas

Para cada um dos modelos estatísticos mais populares, existe uma família de distribuições para o parâmetro com uma propriedade muito especial. Se a distribuição a priori for escolhida como um membro dessa família, então a distribuição a posteriori também será um membro da mesma família. Tal família de distribuições é chamada de *família conjugada*. A escolha de uma distribuição a priori de uma família conjugada tornará o cálculo da distribuição a posteriori particularmente simples.

Amostragem de uma Distribuição de Bernoulli

Exemplo 7.3.1 Um Ensaio Clínico. No Exemplo 5.8.5 (página 330), estávamos observando pacientes em um ensaio clínico. A proporção P de resultados bem-sucedidos entre todos os pacientes possíveis era uma variável aleatória para a qual escolhemos uma distribuição da família de distribuições beta. Essa escolha tornou o cálculo da distribuição condicional de P dados os dados observados

muito simples no final daquele exemplo. De fato, a distribuição condicional de P dados os dados era outro membro da família beta.

O resultado do Exemplo 7.3.1 ocorre em geral e é o assunto do próximo teorema.

Teorema 7.3.1 Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória da distribuição de Bernoulli com parâmetro θ, que é desconhecido ($0 < \theta < 1$). Suponha também que a distribuição a priori de θ seja a distribuição beta com parâmetros $\alpha > 0$ e $\beta > 0$. Então a distribuição a posteriori de θ dado $X_i = x_i\,(i = 1, \ldots, n)$ é a distribuição beta com parâmetros $\alpha + \sum_{i=1}^{n} x_i$ e $\beta + n - \sum_{i=1}^{n} x_i$.

O Teorema 7.3.1 é apenas uma reafirmação do Teorema 5.8.2 (página 329), e sua prova é essencialmente o cálculo no Exemplo 5.8.3.

Atualizando a Distribuição a Posteriori

Uma implicação do Teorema 7.3.1 é a seguinte: Suponha que a proporção θ de itens defeituosos em um grande lote seja desconhecida, a distribuição a priori de θ seja a distribuição beta com parâmetros α e β, e os itens sejam selecionados um de cada vez aleatoriamente do lote. Suponha que o primeiro item inspecionado seja defeituoso, a distribuição a posteriori de θ será a distribuição beta com parâmetros $\alpha + 1$ e β. Se o primeiro item não for defeituoso, a distribuição a posteriori será a distribuição beta com parâmetros α e $\beta + 1$. O processo pode ser continuado da seguinte maneira: a cada item inspecionado, a distribuição a posteriori atual de θ é trocada por uma nova distribuição beta na qual o valor de ou o parâmetro α ou o parâmetro β é aumentado em uma unidade, a cada vez que um item defeituoso é encontrado, e o valor do parâmetro β é aumentado em uma unidade a cada vez que um item não defeituoso é encontrado.

Definição 7.3.1 Família Conjugada/Hiperparâmetros. Seja $\mathcal{F}$ uma família de possíveis distribuições sobre um espaço de parâmetros Ω. Suponha que, não importa qual observação $\mathbf{X} = (X_1, \ldots, X_n)$ observarmos, nem qual distribuição a priori $\xi \in \mathcal{F}$ escolhermos, a distribuição a posteriori $\xi(\theta|\mathbf{x})$ é um membro de $\mathcal{F}$. Então $\mathcal{F}$ é chamada de *família conjugada* de distribuições para amostras das distribuições $f(x|\theta)$. Suponha também que a família $\mathcal{F}$ seja parametrizada por outros parâmetros, então os parâmetros associados para a distribuição a priori são chamados de *hiperparâmetros*, e os parâmetros associados da distribuição a posteriori são chamados de *hiperparâmetros a posteriori*.

O Teorema 7.3.1 diz que a família de distribuições beta é uma família conjugada de distribuições a priori para amostras de uma distribuição de Bernoulli. Se uma distribuição a priori é uma distribuição beta, então a distribuição a

posteriori em cada estágio de amostragem também será uma distribuição beta, independentemente dos valores da amostra observados. Os parâmetros α e β na distribuição a priori do Teorema 7.3.1 são os hiperparâmetros. Os correspondentes hiperparâmetros a posteriori de θ são $\alpha + \sum_{i=1}^{n} x_i$ e $\beta + n - \sum_{i=1}^{n} x_i$. A estatística $\sum_{i=1}^{n} X_i$ é necessária para calcular a distribuição a posteriori, portanto, ela estará presente em qualquer inferência baseada na distribuição a posteriori. Exercícios 23 e 24 introduzem uma coleção geral de f.p.s e f.d.p.s $f(x|\theta)$ para as quais as famílias de distribuições conjugadas existem. A maioria das famílias de distribuições nomeadas abordadas por esses exercícios são notáveis exceções.

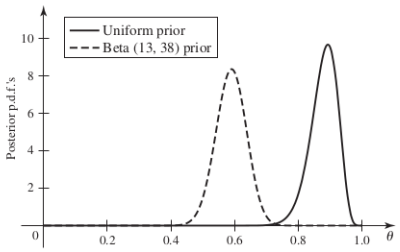
Exemplo 7.3.2 A Variância da Distribuição Beta a Posteriori. Suponha que a proporção θ de itens defeituosos em um grande lote seja desconhecida, a distribuição a priori de θ seja a distribuição uniforme no intervalo $[0, 1]$, e os itens devam ser selecionados aleatoriamente do lote e inspecionados até que a variância da distribuição a posteriori de θ tenha sido reduzida a 0.01 ou menos. Devemos determinar o número total de itens defeituosos e não defeituosos que devem ser obtidos antes que o processo de amostragem seja interrompido. Conforme afirmado na Seção 5.8, a distribuição uniforme no intervalo $[0, 1]$ é a distribuição beta com parâmetros 1 e 1. Portanto, depois que y itens defeituosos e z itens não defeituosos tiverem sido obtidos, a distribuição a posteriori de θ será a distribuição beta com $\alpha = y + 1$ e $\beta = z + 1$. Foi mostrado no Teorema 5.8.3 que a variância da distribuição beta com parâmetros α e β é $\alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]$. Portanto, a variância V da distribuição a posteriori de θ será

$$V = \frac{(y + 1)(z + 1)}{(y + z + 2)^2(y + z + 3)}.$$

A amostragem deve parar assim que o número de defeituosos y e o número de não defeituosos z que foram obtidos sejam tais que $V \leq 0.01$. Pode ser mostrado (ver Exercício 2) que não será necessário selecionar mais de 22 itens, mas é necessário selecionar pelo menos sete itens.

Exemplo 7.3.3 Uso de Luvas por Enfermeiras. Friedland et al. (1992) estudaram 23 enfermeiras em um hospital do centro da cidade antes e depois de um programa educacional sobre a importância de usar luvas. Eles registraram se as enfermeiras usavam ou não luvas em procedimentos nos quais poderiam entrar em contato com fluidos corporais. Antes do programa educacional, as enfermeiras foram observadas durante 51 procedimentos, e usaram luvas em apenas 13 deles. Seja θ a probabilidade de uma enfermeira usar luvas dois meses após o programa educacional. Podemos estar interessados em como θ se compara a 13/51, a proporção observada antes do programa. Vamos considerar duas distribuições a priori diferentes para θ para ver quão sensível a distribuição a posteriori de θ é à escolha da distribuição a priori. A primeira distribuição a

priori será uniforme no intervalo $[0, 1]$, que também é a distribuição beta com parâmetros 1 e 1. A segunda distribuição a priori será a distribuição beta com parâmetros 13 e 38. Esta segunda distribuição a priori tem uma variância muito menor que a primeira e tem sua média em $13/51$. Alguém que sustenta a segunda priori acredita firmemente que o programa educacional não terá efeito perceptível. Dois meses após o programa educacional, 56 procedimentos foram observados, com as enfermeiras usando luvas em 50 deles. A distribuição a posteriori de θ, com base na primeira priori, seria a distribuição beta com parâmetros $1 + 50 = 51$ e $1 + 6 = 7$. Em particular, a média a posteriori de θ é $51/(51 + 7) = 0.88$, e a probabilidade a posteriori de que $\theta > 13/51$ é essencialmente 1. Com base na segunda priori, a distribuição a posteriori de θ seria a distribuição beta com parâmetros $13 + 50 = 63$ e $38 + 6 = 44$. A média a posteriori seria $63/(63 + 44) = 0.59$, e a probabilidade a posteriori de que $\theta > 13/51$ é 0.95. Assim, mesmo alguém que era inicialmente cético sobre o programa educacional parece ter sido convencido. A probabilidade é bastante alta de que as enfermeiras tenham pelo menos o dobro de probabilidade de usar luvas após o programa como antes. A Figura 7.3 mostra as f.d.p.s de ambas as distribuições a posteriori calculadas acima. As distribuições são claramente muito diferentes. Por exemplo, a primeira dá uma probabilidade maior que 0.99 de que $\theta > 0.7$, enquanto a segunda dá uma probabilidade menor que 0.001 de que $\theta > 0.7$. No entanto, uma vez que estamos apenas interessados na probabilidade de que $\theta > 13/51 = 0.5098$, vemos que ambas as prioris concordam que essa probabilidade é bastante grande.



Amostragem de uma Distribuição de Poisson

Exemplo 7.3.4 Chegadas de Clientes. O dono de uma loja modela as chegadas de clientes como um processo de Poisson com uma taxa desconhecida de θ por hora. Ele atribui a θ uma distribuição gama com parâmetros 3 e 2.

Seja X o número de clientes que chegam em um período específico de uma hora. Se $X = 3$ for observado, o dono da loja quer atualizar a distribuição de θ.

Quando as amostras são retiradas de uma distribuição de Poisson, a família de distribuições gama é uma família conjugada de distribuições a priori. Essa relação é mostrada no próximo teorema.

Teorema 7.3.2 Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória da distribuição de Poisson com média $\theta > 0$, e θ seja desconhecido. Suponha também que a distribuição a priori de θ seja a distribuição gama com parâmetros $\alpha > 0$ e $\beta > 0$. Então a distribuição a posteriori de θ, dado que $X_i = x_i$ $(i = 1, \ldots, n)$, é a distribuição gama com parâmetros $\alpha + \sum_{i=1}^{n} x_i$ e $\beta + n$.

Prova Seja $y = \sum_{i=1}^{n} x_i$. Então a função de verossimilhança $f_n(\mathbf{x}|\theta)$ satisfaz a relação

$$f_n(\mathbf{x}|\theta) \propto e^{-n\theta}\theta^y.$$

Nesta relação, um fator que envolve $\mathbf{x}$ mas não depende de θ foi descartado do lado direito. Além disso, a f.d.p. a priori de θ tem a forma

$$\xi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta} \quad \text{para } \theta > 0.$$

Como a f.d.p. a posteriori $\xi(\theta|\mathbf{x})$ é proporcional a $f_n(\mathbf{x}|\theta)\xi(\theta)$, segue-se que

$$\xi(\theta|\mathbf{x}) \propto \theta^{\alpha+y-1}e^{-(\beta+n)\theta} \quad \text{para } \theta > 0.$$

O lado direito desta relação pode ser reconhecido como sendo, exceto por um fator constante, a f.d.p. da distribuição gama com parâmetros $\alpha + y$ e $\beta + n$. Portanto, a distribuição a posteriori de θ é como especificado no teorema. ■

No Teorema 7.3.2, os números α e β são os hiperparâmetros a priori. Note que os hiperparâmetros a posteriori são $\alpha + \sum_{i=1}^{n} x_i$ e $\beta + n$. Note que a estatística $Y = \sum_{i=1}^{n} X_i$ é usada para calcular a distribuição a posteriori de θ, e, portanto, fará parte de qualquer inferência baseada na posteriori.

Exemplo 7.3.5 Chegadas de Clientes. No Exemplo 7.3.4, podemos aplicar o Teorema 7.3.2 com $n = 1, \alpha = 3, \beta = 2$, e $x_1 = 3$. A distribuição a posteriori de θ dado $X = 3$ é a distribuição gama com parâmetros 6 e 3.

Exemplo 7.3.6 A Variância da Distribuição Gama a Posteriori. Con-

sidere uma distribuição de Poisson para a qual a média θ é desconhecida, e suponha que a f.d.p. a priori de θ seja a seguinte:

$$\xi(\theta) = \begin{cases} 2e^{-2\theta} & \text{para } \theta > 0, \\ 0 & \text{para } \theta \leq 0. \end{cases}$$

Suponha também que as observações devam ser retiradas da distribuição de Poisson até que a variância da distribuição a posteriori de θ tenha sido reduzida a 0.01 ou menos. Devemos determinar o número de observações que devem ser feitas antes que o processo de amostragem seja interrompido. A f.d.p. a priori dada é a da distribuição gama com hiperparâmetros $\alpha = 1$ e $\beta = 2$. Portanto, depois de termos obtido valores observados $x_1, \ldots, x_n$, a soma dos quais é $y = \sum_{i=1}^{n} x_i$, a distribuição a posteriori de θ será a distribuição gama com hiperparâmetros $y+1$ e $n+2$. Foi mostrado no Teorema 5.4.2 que a variância da distribuição gama com parâmetros α e β é α/β^2. Portanto, a variância V da distribuição a posteriori de θ será

$$V = \frac{y+1}{(n+2)^2}.$$

A amostragem deve parar assim que os valores observados $x_1, \ldots, x_n$ forem tais que $V \leq 0.01$. Ao contrário do Exemplo 7.3.2, não há limite uniforme sobre quão grande n precisa ser porque y pode ser arbitrariamente grande, não importa qual seja n. Claramente, são necessárias pelo menos $n = 8$ observações antes que $V \leq 0.01$.

Amostragem de uma Distribuição Normal

Exemplo 7.3.7 Emissões Automotivas. Considere novamente o exemplo de emissões automotivas em Exemplo 5.6.1 na página 302. Antes de observar os dados, suponha que um engenheiro acreditasse que cada medida de emissão tinha a distribuição normal com média θ e desvio padrão 0.5, mas que θ era desconhecido. O engenheiro também acredita que θ tem uma distribuição normal com média 2.0 e desvio padrão 1.0. Depois de ver os dados na Fig. 5.1, como este engenheiro descreveria sua incerteza sobre θ?

Quando as amostras são retiradas de uma distribuição normal para a qual o valor da média θ é desconhecido mas o valor da variância σ^2 é conhecido, a família de distribuições normais é uma família conjugada de distribuições a priori, como mostrado no próximo teorema.

Teorema 7.3.3 Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição normal para a qual o valor da média θ é desconhecido e o valor da variância $\sigma^2 > 0$ é conhecido. Suponha também que a distribuição a

priori de θ seja uma distribuição normal com média μ_0 e variância ν_0^2. Então a distribuição a posteriori de θ dado que $X_i = x_i\,(i = 1, \ldots, n)$ é a distribuição normal com média μ_1 e variância ν_1^2, onde

$$\mu_1 = \frac{\sigma^2\mu_0 + n\nu_0^2\bar{x}_n}{\sigma^2 + n\nu_0^2} \tag{7.3.1}$$

e

$$\nu_1^2 = \frac{\sigma^2\nu_0^2}{\sigma^2 + n\nu_0^2}. \tag{7.3.2}$$

Prova A função de verossimilhança, $f_n(\mathbf{x}|\theta)$ tem a forma

$$f_n(\mathbf{x}|\theta) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta)^2\right].$$

Aqui, um fator constante foi descartado do lado direito. O método de completar o quadrado (ver Exercício 24 na Seção 5.6) nos diz que

$$\sum_{i=1}^{n}(x_i - \theta)^2 = n(\theta - \bar{x}_n)^2 + \sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

Omitindo um fator que envolve $x_1, \ldots, x_n$ mas não depende de θ, podemos reescrever $f_n(\mathbf{x}|\theta)$ na seguinte forma:

$$f_n(\mathbf{x}|\theta) \propto \exp\left[-\frac{n}{2\sigma^2}(\theta - \bar{x}_n)^2\right].$$

Como a f.d.p. a priori $\xi(\theta)$ tem a forma

$$\xi(\theta) \propto \exp\left[-\frac{1}{2\nu_0^2}(\theta - \mu_0)^2\right],$$

segue-se que a f.d.p. a posteriori $\xi(\theta|\mathbf{x})$ satisfaz a relação

$$\xi(\theta|\mathbf{x}) \propto \exp\left\{-\frac{1}{2}\left[\frac{n}{\sigma^2}(\theta - \bar{x}_n)^2 + \frac{1}{\nu_0^2}(\theta - \mu_0)^2\right]\right\}.$$

Se μ_1 e ν_1^2 são como especificado nas Eqs. (7.3.1) e (7.3.2), completar o quadrado novamente estabelece a seguinte identidade:

$$\frac{n}{\sigma^2}(\theta - \bar{x}_n)^2 + \frac{1}{\nu_0^2}(\theta - \mu_0)^2 = \frac{1}{\nu_1^2}(\theta - \mu_1)^2 + \frac{n}{\sigma^2 + n\nu_0^2}(\bar{x}_n - \mu_0)^2.$$

Como o termo final no lado direito desta equação não envolve θ, ele pode ser absorvido na constante de proporcionalidade, e obtemos a relação

$$\xi(\theta|\mathbf{x}) \propto \exp\left[-\frac{1}{2\nu_1^2}(\theta - \mu_1)^2\right].$$

O lado direito desta relação pode ser reconhecido como sendo, exceto por um fator constante, a f.d.p. da distribuição normal com média μ_1 e variância ν_1^2. Portanto, a distribuição a posteriori de θ é como especificado no teorema. ■

No Teorema 7.3.3, os números μ_0 e ν_0^2 são os hiperparâmetros a priori, enquanto μ_1 e ν_1^2 são os hiperparâmetros a posteriori. Note que a estatística $\bar{X}_n$ é usada na construção da distribuição a posteriori, e, portanto, desempenhará um papel em qualquer inferência baseada na posteriori.

Exemplo 7.3.8 Emissões Automotivas. Podemos aplicar o Teorema 7.3.3 para responder à pergunta no final do Exemplo 7.3.7. Na notação do teorema, temos $n = 46, \sigma^2 = 0.5^2 = 0.25, \mu_0 = 2$ e $\nu_0^2 = 1.0$. A média das 46 medições é $\bar{x}_n = 1.329$. A distribuição a posteriori de θ é então a distribuição normal com média e variância dadas por

$$\mu_1 = \frac{0.25 \times 2 + 46 \times 1 \times 1.329}{0.25 + 46 \times 1} = 1.333,$$

$$\nu_1^2 = \frac{0.25 \times 1}{0.25 + 46 \times 1} = 0.0054.$$

A média μ_1 da distribuição a posteriori de θ, como dada na Eq. (7.3.1), pode ser reescrita da seguinte forma:

$$\mu_1 = \frac{\sigma^2}{\sigma^2 + n\nu_0^2}\mu_0 + \frac{n\nu_0^2}{\sigma^2 + n\nu_0^2}\bar{x}_n. \tag{7.3.3}$$

Pode-se ver da Eq. (7.3.3) que μ_1 é uma média ponderada da média a priori μ_0 e da média amostral $\bar{x}_n$. Além disso, pode-se ver que o peso relativo dado a $\bar{x}_n$ satisfaz as três propriedades a seguir: (1) Para valores fixos de ν_0^2 e σ^2, quanto maior o tamanho da amostra n, maior será o peso relativo dado a $\bar{x}_n$. (2) Para valores fixos de ν_0^2 e n, quanto maior a variância σ^2 de cada observação na amostra, menor será o peso relativo dado a $\bar{x}_n$. (3) Para valores fixos de σ^2 e n, quanto maior a variância ν_0^2 da distribuição a priori, maior será o peso relativo dado a $\bar{x}_n$. Além disso, pode-se ver da Eq. (7.3.2) que a variância ν_1^2 da distribuição a posteriori de θ depende do número n de observações que foram feitas, mas não depende das magnitudes dos valores observados. Portanto, para uma amostra aleatória de n observações a ser retirada de uma distribuição normal para a qual o valor da média θ é desconhecido, o valor da variância é conhecido, e a distribuição a priori de θ é uma distribuição normal especificada. Então, antes de qualquer observação ter sido feita, podemos usar a Eq. (7.3.2) para calcular o valor da variância ν_1^2 da distribuição a posteriori. No entanto, o valor da média μ_1 da distribuição a posteriori dependerá dos valores observados que são obtidos na amostra. O fato de a variância a posteriori não depender dos valores observados deve-se à suposição de que a variância σ^2 das observações individuais é conhecida. Na Seção 8.6, relaxaremos essa suposição.

Exemplo 7.3.9 A Variância da Distribuição Normal a Posteriori. Suponha que as observações devam ser retiradas de uma distribuição normal com média θ e variância 1, e que o valor de θ seja desconhecido. Suponha também que a distribuição a priori de θ seja uma distribuição normal com variância 4. Além disso, as observações devem ser feitas até que a variância da distribuição a posteriori de θ tenha sido reduzida a 0.01 ou menos. Devemos determinar o número de observações que devem ser feitas antes que o processo de amostragem seja interrompido. Segue-se da Eq. (7.3.2) que, após n observações terem sido feitas, a variância ν_1^2 da distribuição a posteriori será

$$\nu_1^2 = \frac{4}{4n+1}.$$

Portanto, a relação $\nu_1^2 \leq 0.01$ será satisfeita se e somente se $n \geq 99.75$. Portanto, a variância a posteriori será de 0.01 ou menos depois de 100 observações terem sido feitas e não antes.

Exemplo 7.3.10 Contagem de Calorias em Alimentos Preparados. Allison, Heshka, Sepulveda e Heymsfield (1993) amostraram 20 alimentos preparados nacionalmente e compararam o conteúdo de calorias declarado com o conteúdo de calorias determinado no laboratório. A Figura 7.4 é um histograma das diferenças percentuais entre as medições observadas em laboratório e o conteúdo de calorias anunciado nos rótulos dos alimentos. Suponha que modelamos a distribuição condicional das diferenças dado θ como a distribuição normal com média θ e variância 100. (Nesta seção, assumimos que a variância é conhecida. Na Seção 8.6, seremos capazes de lidar com o caso em que tanto a média quanto a variância são tratadas como variáveis aleatórias com uma distribuição conjunta.) Usaremos uma distribuição a priori para θ que é a distribuição normal com média 0 e uma variância de 60. Os dados $\mathbf{X}$ compreendem as 20 diferenças na Fig. 7.4, cuja média é 0.125. A distribuição a posteriori de θ seria então a distribuição normal com média

$$\mu_1 = \frac{100 \times 0 + 20 \times 60 \times 0.125}{100 + 20 \times 60} = 0.1154,$$

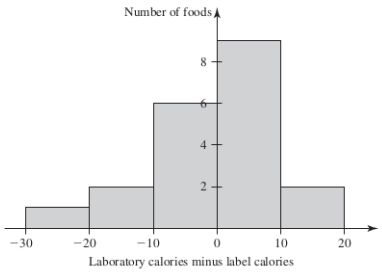
e variância

$$\nu_1^2 = \frac{100 \times 60}{100 + 20 \times 60} = 4.62.$$

Por exemplo, podemos estar interessados em saber se os empacotadores estão subestimando sistematicamente as calorias em seus alimentos em pelo menos 1 por cento. Isso corresponderia a $\theta > 1$. Usando o Teorema 5.6.6, podemos encontrar

$$\Pr(\theta > 1|\mathbf{x}) = 1 - \Phi\left(\frac{1 - 0.1154}{\sqrt{4.62}}\right) = 1 - \Phi(1.12) = 0.3403.$$

Há uma chance não desprezível, mas não esmagadora, de que os empacotadores estejam omitindo um por cento ou mais de suas calorias dos rótulos.



Amostragem de uma Distribuição Exponencial

Exemplo 7.3.11 Tempo de Vida de Componentes Eletrônicos. No Exemplo 7.2.1, suponha que observemos os tempos de vida de três componentes, $X_1 = 3, X_2 = 1.5, X_3 = 2.1$. Eles foram modelados como i.i.d. variáveis exponenciais, dado θ. Nossa distribuição a priori para θ foi a distribuição gama com parâmetros 1 e 2. Qual é a distribuição a posteriori de θ dados esses tempos de vida observados?

Ao amostrar de uma distribuição exponencial para a qual o valor do parâmetro θ é desconhecido, a família de distribuições gama serve como uma família conjugada de distribuições a priori, como mostrado no próximo teorema.

Teorema 7.3.4 Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição exponencial com parâmetro $\theta > 0$ que é desconhecido. Suponha também que a distribuição a priori de θ seja a distribuição gama com parâmetros $\alpha > 0$ e $\beta > 0$. Então a distribuição a posteriori de θ dado que $X_i = x_i \, (i = 1, \ldots, n)$ é a distribuição gama com parâmetros $\alpha+n$ e $\beta+\sum_{i=1}^{n} x_i$.

Prova Novamente, seja $y = \sum_{i=1}^{n} x_i$. Então a função de verossimilhança $f_n(\mathbf{x}|\theta)$ tem a forma

$$f_n(\mathbf{x}|\theta) = \theta^n e^{-\theta y}.$$

Além disso, a f.d.p. a priori $\xi(\theta)$ tem a forma

$$\xi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} \quad \text{para } \theta > 0.$$

Portanto, segue-se que a f.d.p. a posteriori $\xi(\theta|\mathbf{x})$ tem a forma

$$\xi(\theta|\mathbf{x}) \propto \theta^{\alpha+n-1} e^{-(\beta+y)\theta} \quad \text{para } \theta > 0.$$

O lado direito desta relação pode ser reconhecido como sendo, exceto por um fator constante, a f.d.p. da distribuição gama com parâmetros $\alpha + n$ e $\beta + y$. Portanto, a distribuição a posteriori de θ é como especificado no teorema. ■

A distribuição a posteriori de θ no Teorema 7.3.4 depende do valor observado da estatística $Y = \sum_{i=1}^{n} X_i$; portanto, toda inferência sobre θ baseada na distribuição a posteriori dependerá do valor observado de Y.

Exemplo 7.3.12 Tempo de Vida de Componentes Eletrônicos. No Exemplo 7.3.11, podemos aplicar o Teorema 7.3.4 para encontrar a distribuição a posteriori. Na notação do teorema e sua prova, temos $n = 3, \alpha = 1, \beta = 2$, e

$$y = \sum_{i=1}^{n} x_i = 3 + 1.5 + 2.1 = 6.6.$$

A distribuição a posteriori de θ é então a distribuição gama com parâmetros $\alpha + n = 1 + 3 = 4$ e $\beta + y = 2 + 6.6 = 8.6$. O leitor deve notar que o Teorema 7.3.4 teria encurtado muito a derivação da posterior no Exemplo 7.2.6.

Distribuições a Priori Impróprias

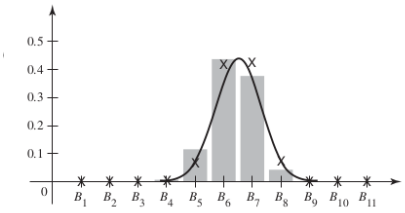
Na Seção 7.2, mencionamos prioris impróprias como expedientes que tentam capturar a ideia de que há muito mais informação nos dados do que a contida em nossa distribuição a priori. Cada uma das famílias conjugadas que vimos nesta seção tem uma priori imprópria como um caso limite.

Exemplo 7.3.13 Um Ensaio Clínico. O que ilustramos aqui será aplicado a todos os exemplos em que os dados compreendem uma amostra i.i.d. (dado θ) da distribuição de Bernoulli com parâmetro θ. Considere os sujeitos no grupo da imipramina no Exemplo 2.1.4. A proporção de sucessos entre todos os pacientes que poderiam receber imipramina foi chamada de P em exemplos anteriores, mas vamos chamá-la de θ desta vez para manter a notação geral deste capítulo. Suponha que θ tenha a distribuição beta com parâmetros α e β, e um conjugado geral a priori. Existem $n = 40$ pacientes no grupo da imipramina, e 22 deles são sucessos. A distribuição a posteriori de θ é a distribuição beta com parâmetros $\alpha + 22$ e $\beta + 18$, como vimos no Teorema 7.3.1. A média da distribuição a posteriori é $(\alpha + 22)/(\alpha + \beta + 40)$. Se α e β são pequenos, então a média a posteriori está próxima de $22/40$, que é a proporção observada de sucessos. De fato, se $\beta = 0$, o que não corresponde a uma distribuição beta real, então a média a posteriori é exatamente $22/40$. No entanto, olhe o que acontece com α e β quando eles se aproximam de 0. A f.d.p. beta (ignorando a constante) é $\theta^{\alpha-1}(1-\theta)^{\beta-1}$. Podemos definir $\alpha = \beta = 0$ e fingir que $\xi(\theta) \propto \theta^{-1}(1-\theta)^{-1}$ é a

f.d.p. a priori de θ. A função de verossimilhança é $f_{40}(\mathbf{x}|\theta) = \binom{40}{22}\theta^{22}(1-\theta)^{18}$. Podemos ignorar o fator constante $\binom{40}{22}$ e obter o produto

$$\xi(\theta|\mathbf{x}) \propto \theta^{21}(1-\theta)^{17}, \quad \text{para } 0 < \theta < 1.$$

Isso é facilmente reconhecido como sendo a mesma coisa que a f.d.p. da distribuição beta com parâmetros 22 e 18, exceto por um fator constante. Assim, se usarmos a "priori imprópria"beta com hiperparâmetros 0 e 0, obtemos a distribuição a posteriori beta para θ com hiperparâmetros 22 e 18. Note que o Teorema 7.3.1 produz a distribuição a posteriori correta mesmo neste caso de priori imprópria. A Figura 7.5 adiciona a f.d.p. da distribuição beta a posteriori calculada aqui à Fig. 2.4, que representava as probabilidades a posteriori para duas distribuições a priori discretas diferentes. Todas as três posteriores são muito próximas.



Definição 7.3.2 Priori Imprópria. Seja $\xi(\theta)$ uma função não negativa cujo domínio inclui o espaço de parâmetros de um modelo estatístico. Suponha que $\int \xi(\theta)d\theta = \infty$. Se pretendemos usar $\xi(\theta)$ como a f.d.p. a priori de θ, então estamos usando uma *priori imprópria* para θ.

A Definição 7.3.2 não é muito útil para determinar uma priori imprópria a ser usada em uma aplicação particular. Existem muitos métodos para escolher uma priori imprópria, e a esperança é que todos eles levem a distribuições a posteriori semelhantes, de modo que não importe muito qual deles se escolhe. O método mais direto para escolher uma priori imprópria é começar com a família de distribuições conjugadas, se tal família existir. Na maioria dos casos, se a parametrização da família conjugada (hiperparâmetros) for escolhida com cuidado, a posteriori terá a mesma forma que a priori conjugada mais uma estatística. Alguém então substituiria cada um desses hiperparâmetros por 0 para obter o que parece ser a fórmula para a f.d.p. a posteriori. Isso é o que fizemos no Exemplo 7.3.13; cada um dos hiperparâmetros a posteriori era igual aos hiperparâmetros a priori correspondentes mais alguma estatística. Nesse exemplo, substituímos ambos os hiperparâmetros a priori por 0 para obter a priori imprópria. O método descrito anteriormente precisa ser modificado se

alguém escolher uma parametrização "inconveniente"da priori conjugada, como no Exemplo 7.3.15 abaixo.

Exemplo 7.3.14 Mortes no Exército Prussiano. Bortkiewicz (1898) contou o número de soldados prussianos mortos por coices de cavalo em várias unidades do exército durante o século dezenove (um problema mais sério naquela época do que hoje). Havia 14 corpos do exército para cada um dos 20 anos, para um total de 280 contagens. As 280 contagens tiveram os seguintes valores: 144 contagens são 0, 91 contagens são 1, 32 contagens são 2, 11 contagens são 3, e 2 contagens são 4. Nenhuma unidade sofreu mais de quatro mortes por coices de cavalo em um único ano. (Esses dados foram reportados e analisados por Winsor, 1947.) Suponha que vamos modelar as 280 contagens como uma amostra aleatória de variáveis aleatórias de Poisson $X_1, \ldots, X_{280}$ com média θ condicional ao parâmetro θ. Uma priori conjugada seria um membro da família de distribuições gama com hiperparâmetros α e β. O Teorema 7.3.2 diz que a distribuição a posteriori de θ seria a distribuição gama com hiperparâmetros a posteriori 196 e 280, uma vez que a soma das 280 contagens é 196. A menos que α ou β seja muito grande, a distribuição gama a posteriori é quase a mesma que a distribuição gama com hiperparâmetros posteriores 196 e 280. Este problema de distribuição a posteriori parece ser o resultado do uso de uma priori conjugada com hiperparâmetros a priori 0 e 0. Ignorando o fator constante, a f.d.p. da distribuição gama com parâmetros α e β é $\theta^{\alpha-1}e^{-\beta\theta}$ para $\theta > 0$. Se definirmos $\alpha = 0$ e $\beta = 0$ nesta fórmula, obtemos a "f.d.p. a priori imprópria"θ^{-1} para $\theta > 0$. Fingindo que isso realmente era uma f.d.p. a priori e aplicando o teorema de Bayes para variáveis aleatórias (Teorema 3.6.4) resultaria em

$$\xi(\theta|\mathbf{x}) \propto \theta^{195}e^{-280\theta}, \quad \text{para } \theta > 0.$$

Isso é facilmente reconhecido como sendo a f.d.p. da distribuição gama com parâmetros 196 e 280, exceto por um fator constante. Este exemplo de resultado se aplica a todos os casos em que modelamos dados com distribuições de Poisson. A "distribuição gama"imprópria com hiperparâmetros a priori 0 e 0 pode ser usada no Teorema 7.3.2, e a conclusão ainda se manterá.

Exemplo 7.3.15 Tempos de Falha de Rolamentos de Esferas. Suponha que modelamos os 23 logaritmos de tempos de falha dos rolamentos de esferas do Exemplo 5.6.9 como variáveis aleatórias normais $X_1, \ldots, X_{23}$ com média θ e variância 0.25. Uma priori conjugada para θ seria a distribuição normal com média μ_0 e variância ν_0^2. Para os log-tempos de falha, $\bar{x}_{23} = 4.15$, então a distribuição a posteriori de θ seria a distribuição normal com média $\mu_1 = (0.25\mu_0 + 23 \times 4.15\nu_0^2)/(0.25 + 23\nu_0^2)$ e variância $\nu_1^2 = (0.25\nu_0^2)/(0.25 + 23\nu_0^2)$. Se deixarmos $\nu_0^2 \to \infty$ nas fórmulas para μ_1 e ν_1^2, obtemos $\mu_1 \to 4.15$ e $\nu_1^2 \to 0.25/23$. Ter variância infinita para a distribuição a priori de θ é como dizer que θ é igualmente provável de estar em qualquer lugar na reta real. Isso

acontece em todos os exemplos em que modelamos dados $X_1, \ldots, X_n$ como uma amostra aleatória da distribuição normal com média θ e variância conhecida σ^2 condicional a θ. Se usarmos uma "distribuição normal"imprópria com variância infinita (a priori não precisa de média), o cálculo no Teorema 7.3.3 resultaria na distribuição normal com média $\bar{x}_n$ e variância σ^2/n. Neste caso, $\xi(\theta)$ seria igual a uma constante. Este exemplo poderia ser uma aplicação do método descrito após a Definição 7.3.2 se tivéssemos os hiperparâmetros "mais convenientes": 1 sobre a variância $1/\nu_0^2$ e a média sobre a variância μ_0/ν_0^2. Em termos desses hiperparâmetros, a distribuição posterior tem 1 sobre sua variância igual a $1/\nu_1^2 = 1/\nu_0^2 + n/\sigma^2 = 1/\nu_0^2 + 23/0.25$ e média sobre variância igual a $\mu_1/\nu_1^2 = \mu_0/\nu_0^2 + 23 \times 4.15/0.25$. Cada um dos hiperparâmetros a posteriori tem a forma do hiperparâmetro a priori correspondente mais uma estatística. A priori imprópria com a média sobre variância e 1 sobre variância iguais a 0 também tem $\xi(\theta)$ igual a uma constante.

Existem outros exemplos de prioris impróprias para outros modelos de amostragem. O leitor pode verificar (no Exercício 21) que a "distribuição gama"com parâmetros 0 e 0 leva a resultados semelhantes aos de uma amostra aleatória de uma distribuição exponencial. Os Exercícios 23 e 24 introduzem uma coleção geral de f.d.p.s $f(x|\theta)$ para as quais é fácil construir prioris impróprias. Prioris impróprias foram introduzidas para casos em que os dados observados continham muito mais informação do que nossa distribuição a priori. Implicitamente, estamos assumindo que os dados são informativos. Quando os dados não contêm muita informação, as prioris impróprias podem ser altamente inapropriadas.

Exemplo 7.3.16 Eventos Muito Raros. No Exemplo 5.4.7, estávamos discutindo um contaminante de água potável conhecido como criptosporídio que geralmente ocorre em concentrações muito baixas. Suponha que uma autoridade de água modele os oocistos de criptosporídio no abastecimento de água como um processo de Poisson com taxa de θ oocistos por litro. Eles decidem amostrar 25 litros de água para aprender sobre θ. Suponha que eles usem a priori gama imprópria com "f.d.p."θ^{-1}. (Esta é a mesma priori imprópria usada no Exemplo 7.3.14.) Se a amostra de 25 litros não contiver oocistos, a autoridade de água teria uma distribuição a posteriori para θ que é a distribuição gama com parâmetros 0 e 5, o que não é uma distribuição real. Não importa quantos litros eles amostrem, a distribuição a posteriori não será uma distribuição real até que pelo menos um oocisto seja observado. Ao amostrar eventos raros, pode-se ser forçado a quantificar a informação a priori na forma de uma distribuição a priori adequada para fazer inferências baseadas na distribuição a posteriori.

Resumo

Para cada uma das várias famílias diferentes de modelos estatísticos, para dados de uma determinada família, encontramos uma família conjugada de distribuições para o parâmetro. Essas famílias têm a propriedade de que, se a distribuição a priori for escolhida da família, então a distribuição a posteriori é um membro da família. Para dados com distribuições relacionadas à Bernoulli, como binomial, geométrica e binomial negativa, a família conjugada para o parâmetro de sucesso é a família de distribuições beta. Para dados com distribuições relacionadas ao processo de Poisson, como Poisson e gama (com primeiro parâmetro conhecido), e exponencial, a família conjugada para o parâmetro de taxa é a família de distribuições gama. Para dados com uma distribuição normal com variância conhecida, a família conjugada para a média é a família normal. Também descrevemos o uso de prioris impróprias. Prioris impróprias não são distribuições de probabilidade, mas se fingirmos que são, podemos calcular distribuições a posteriori que se aproximam daquelas que teríamos obtido usando prioris conjugadas apropriadas com valores extremos dos hiperparâmetros a priori.

Exercícios

1. Considere novamente a situação descrita no Exemplo 7.3.10. Mais uma vez, suponha que a distribuição a priori de θ seja uma distribuição normal com média 0, mas desta vez, seja a variância a priori $\nu^2 > 0$. Se a média a posteriori de θ for 0.12, que valor de ν^2 foi usado?

2. Mostre que no Exemplo 7.3.2 deve ser verdade que $V \leq 0.01$ depois que 22 itens tiverem sido selecionados. Mostre também que $V > 0.01$ até que pelo menos sete itens tenham sido selecionados.

3. Suponha que a proporção θ de itens defeituosos em um grande lote seja desconhecida e que a distribuição a priori de θ seja a distribuição beta com parâmetros 2 e 200. Se 100 itens forem selecionados aleatoriamente do lote e três desses itens forem encontrados como defeituosos, qual é a distribuição a posteriori de θ?

4. Considere novamente as condições do Exercício 3. Suponha que, depois que um certo estatístico observou que havia três itens defeituosos entre os 100 itens selecionados aleatoriamente, a distribuição a posteriori que ele atribui a θ seja uma distribuição beta para a qual a média é $2/51$ e a variância é $98/[(51)^2(103)]$. Qual distribuição a priori o estatístico atribuiu a θ?

5. Suponha que o número de defeitos em um rolo de 1200 pés de fita de gravação magnética tenha uma distribuição de Poisson para a qual o valor da média θ é desconhecido e que a distribuição a priori de θ seja a distribuição gama com parâmetros $\alpha = 3$ e $\beta = 1$. Quando cinco rolos

desta fita são selecionados aleatoriamente e inspecionados, o número de defeitos encontrados nos rolos são 2, 2, 6, 0 e 3. Determine a distribuição a posteriori de θ.

6. Seja θ o número médio de defeitos por 100 pés de um certo tipo de fita magnética. Suponha que o valor de θ seja desconhecido e que a distribuição a priori de θ seja a distribuição gama com parâmetros $\alpha = 2$ e $\beta = 10$. Quando um rolo de 1200 pés desta fita é inspecionado, exatamente quatro defeitos são encontrados. Determine a distribuição a posteriori de θ.

7. Suponha que as alturas dos indivíduos em uma certa população tenham uma distribuição normal para a qual o valor da média θ é desconhecido e o desvio padrão é 2 polegadas. Suponha também que a distribuição a priori de θ seja uma distribuição normal para a qual a média é 68 polegadas e o desvio padrão é 1 polegada. Se 10 pessoas forem selecionadas aleatoriamente da população, e sua altura média for de 69.5 polegadas, qual é a distribuição a posteriori de θ?

8. Considere novamente o problema descrito no Exercício 7.

   (a) Qual intervalo de 1 polegada de comprimento tinha a maior probabilidade a priori de conter o valor de θ?

   (b) Qual intervalo de 1 polegada de comprimento tem a maior probabilidade a posteriori de conter o valor de θ?

   (c) Encontre os valores das probabilidades nas partes (a) e (b).

9. Suponha que uma amostra aleatória de 20 observações seja retirada de uma distribuição normal para a qual o valor da média θ é desconhecido e a variância é 1. Após os valores da amostra terem sido observados, descobre-se que $\bar{X}_n = 10$, e que a distribuição a posteriori de θ é uma distribuição normal para a qual a média é 8 e a variância é 1/25. Qual foi a distribuição a priori de θ?

10. Suponha que uma amostra aleatória deva ser retirada de uma distribuição normal para a qual o valor da média θ é desconhecido e o desvio padrão é 2, e a distribuição a priori de θ é uma distribuição normal para a qual o desvio padrão é 1. Qual é o menor número de observações que devem ser incluídas na amostra para reduzir o desvio padrão da distribuição a posteriori de θ para o valor 0.1?

11. Suponha que uma amostra aleatória de 100 observações deva ser retirada de uma distribuição normal para a qual o valor da média θ é desconhecido e o desvio padrão é 2, e a distribuição a priori de θ seja uma distribuição normal. Mostre que, não importa quão grande seja o desvio padrão da distribuição a priori, o desvio padrão da distribuição a posteriori será menor que 1/5.

12. Suponha que o tempo em minutos necessário para servir um cliente em uma determinada instalação tenha uma distribuição exponencial para a qual o valor do parâmetro θ é desconhecido e que a distribuição a priori de θ seja uma distribuição gama para a qual a média é 0.2 e o desvio padrão é 1. Se o tempo médio necessário para servir uma amostra aleatória de 20 clientes for observado como 3.8 minutos, qual é a distribuição a posteriori de θ?

13. Para uma distribuição com média $\mu \neq 0$ e desvio padrão $\sigma > 0$, o *coeficiente de variação* da distribuição é definido como $\sigma/|\mu|$. Considere novamente o problema descrito no Exercício 12, e suponha que o coeficiente de variação da distribuição gama a priori de θ seja 2. Qual é o menor número de clientes que devem ser observados para reduzir o coeficiente de variação da distribuição a posteriori para 0.1?

14. Mostre que a família de distribuições beta é uma família conjugada de distribuições a priori para amostras de uma distribuição binomial negativa com um valor conhecido do parâmetro r e um valor desconhecido do parâmetro p $(0 < p < 1)$.

15. Seja $\xi(\theta)$ uma f.d.p. que é definida da seguinte forma para constantes $\alpha > 0$ e $\beta > 0$:

$$\xi(\theta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta} & \text{para } \theta > 0, \\ 0 & \text{para } \theta \leq 0. \end{cases}$$

    Uma distribuição com esta f.d.p. é chamada de *distribuição gama inversa*.

    (a) Verifique se $\xi(\theta)$ é realmente uma f.d.p. verificando que $\int_0^\infty \xi(\theta) d\theta = 1$.

    (b) Considere a família de distribuições de probabilidade que pode ser representada por uma f.d.p. $\xi(\theta)$ tendo a forma dada para todos os pares de constantes $\alpha > 0$ e $\beta > 0$ possíveis. Mostre que esta família é uma família conjugada de distribuições a priori para amostras de uma distribuição normal com um valor conhecido da média μ e um valor desconhecido da variância θ.

16. Suponha que no Exercício 15 o parâmetro seja considerado o desvio padrão da distribuição normal, em vez da variância. Determine uma família conjugada de distribuições a priori para amostras de uma distribuição normal com um valor conhecido da média μ e um valor desconhecido do desvio padrão σ.

17. Suponha que o número de minutos que uma pessoa deve esperar por um ônibus a cada manhã tenha a distribuição uniforme no intervalo $[0, \theta]$, onde o valor do ponto final θ é desconhecido. Suponha também que a

f.d.p. a priori de θ seja a seguinte:

$$\xi(\theta) = \begin{cases} \frac{192}{\theta^4} & \text{para } \theta \geq 4, \\ 0 & \text{caso contrário.} \end{cases}$$

Se os tempos de espera observados em três manhãs sucessivas forem 5, 3 e 8 minutos, qual é a f.d.p. a posteriori de θ?

18. A distribuição de Pareto com parâmetros $x_0 > 0$ e $\alpha > 0$ é definida no Exercício 16 da Seção 5.7. Mostre que a família de distribuições de Pareto é uma família conjugada de distribuições a priori para amostras de uma distribuição uniforme no intervalo $[0, \theta]$, onde o valor do ponto final θ é desconhecido.

19. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição para a qual a f.d.p. $f(x|\theta)$ é a seguinte:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{para } 0 < x < 1, \\ 0 & \text{caso contrário.} \end{cases}$$

Suponha também que o valor do parâmetro θ seja desconhecido ($\theta > 0$), e a distribuição a priori de θ seja a distribuição gama com parâmetros $\alpha > 0$ e $\beta > 0$. Determine a média e a variância da distribuição a posteriori de θ.

20. Suponha que modelamos os tempos de vida (em meses) de componentes eletrônicos como variáveis aleatórias exponenciais independentes com parâmetro desconhecido β. Modelamos β como tendo a distribuição gama com parâmetros a e b. Acreditamos que o tempo de vida médio é de quatro meses. Se fôssemos observar 10 componentes com uma vida útil média observada de seis meses, então reivindicaríamos que o tempo de vida médio é de cinco meses. Determine a e b. *Dica:* Use o Exercício 21 da Seção 5.7.

21. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória da distribuição exponencial com parâmetro θ. Seja a distribuição a priori de θ imprópria com "f.d.p."$1/\theta$ para $\theta > 0$. Encontre a distribuição a posteriori de θ e mostre que a média a posteriori de θ é $1/\bar{x}_n$.

22. Considere os dados no Exemplo 7.3.10. Desta vez, suponha que usemos a priori imprópria com "f.d.p."$\xi(\theta) = 1$ (para todo θ). Encontre a distribuição a posteriori de θ e a probabilidade a posteriori de que $\theta > 1$.

23. Considere uma distribuição para a qual a f.d.p. ou f.p. é $f(x|\theta)$, onde θ pertence a algum espaço de parâmetros Ω. É dito que a família de distribuições obtidas deixando θ variar sobre todos os valores em Ω é uma

família exponencial, ou uma *família de Koopman-Darmois*, se $f(x|\theta)$ pode ser escrito da seguinte forma para $\theta \in \Omega$ e todos os valores de x:

$$f(x|\theta) = a(\theta)b(x)\exp[c(\theta)d(x)].$$

Aqui $a(\theta)$ e $c(\theta)$ são funções arbitrárias de θ, e $b(x)$ e $d(x)$ são funções arbitrárias de x. Seja

$$H = \left\{(\alpha, \beta) : \int_\Omega a(\theta)^\alpha \exp[c(\theta)\beta]d\theta < \infty\right\}.$$

Para cada $(\alpha, \beta) \in H$, seja

$$\xi_{\alpha,\beta}(\theta) = \frac{a(\theta)^\alpha \exp[c(\theta)\beta]}{\int_\Omega a(\eta)^\alpha \exp[c(\eta)\beta]d\eta},$$

e seja Ψ o conjunto de todas as distribuições de probabilidade que têm f.d.p.s da forma $\xi_{\alpha,\beta}(\theta)$ para alguns $(\alpha, \beta) \in H$.

(a) Mostre que Ψ é uma família conjugada de distribuições a priori para amostras de $f(x|\theta)$.

(b) Suponha que observemos uma amostra aleatória de tamanho n da distribuição com f.d.p. $f(x|\theta)$. Se a f.d.p. a priori de θ for ξ_{α_0,β_0}, mostre que os hiperparâmetros a posteriori são

$$\alpha_1 = \alpha_0 + n, \quad \beta_1 = \beta_0 + \sum_{i=1}^{n} d(x_i).$$

24. Mostre que cada uma das seguintes famílias de distribuições é uma família exponencial, como definido no Exercício 23:

(a) A família de distribuições de Bernoulli com um parâmetro desconhecido p.

(b) A família de distribuições de Poisson com uma média desconhecida.

(c) A família de distribuições binomiais negativas para as quais o valor de r é conhecido e o valor de p é desconhecido.

(d) A família de distribuições normais com uma média desconhecida e uma variância conhecida.

(e) A família de distribuições normais com uma variância desconhecida e uma média conhecida.

(f) A família de distribuições gama para as quais o valor de α é desconhecido e o valor de β é conhecido.

(g) A família de distribuições gama para as quais o valor de α é conhecido e o valor de β é desconhecido.

(h) A família de distribuições beta para as quais o valor de α é desconhecido e o valor de β é conhecido.

(i) A família de distribuições beta para as quais o valor de α é conhecido e o valor de β é desconhecido.

25. Mostre que a família de distribuições uniformes nos intervalos $[0, \theta]$ para $\theta > 0$ não é uma família exponencial como definido no Exercício 23. *Dica:* Olhe para o suporte de cada distribuição uniforme.

26. Mostre que a família de distribuições uniformes discretas nos conjuntos de inteiros $\{0, 1, \ldots, \theta\}$ para θ um inteiro não negativo não é uma família exponencial como definido no Exercício 23.

7.4 Estimadores de Bayes

Um estimador de um parâmetro é alguma função dos dados que esperamos que seja próxima ao parâmetro. Um estimador de Bayes é um estimador que é escolhido para minimizar a média a posteriori de alguma medida de quão longe um estimador está do parâmetro, como o erro quadrático ou o erro absoluto.

Natureza de um Problema de Estimação

Exemplo 7.4.1 Contagem de Calorias em Rótulos de Alimentos. No Exemplo 7.3.10, encontramos a distribuição a posteriori de θ, a diferença percentual média entre as calorias medidas e as anunciadas. Um grupo de consumidores pode desejar relatar um único número como uma estimativa de θ sem especificar a distribuição inteira para θ. Como escolher tal estimativa de número único é o assunto desta seção.

Começamos com uma definição apropriada para um valor de variável real, como no Exemplo 7.4.1. Uma definição mais geral se seguirá depois que nos familiarizarmos mais com o conceito de estimação.

Definição 7.4.1 Estimador/Estimativa. Seja $X_1, \ldots, X_n$ observáveis, cuja distribuição conjunta é indexada por um parâmetro θ assumindo valores em um subconjunto Ω da reta real. Um *estimador* do parâmetro θ é uma função de valor real $\delta(X_1, \ldots, X_n)$. Se $X_1 = x_1, \ldots, X_n = x_n$ são observados, então $\delta(x_1, \ldots, x_n)$ é chamada de *estimativa* de θ.

Note que todo estimador é, por natureza de ser uma função dos dados, uma estatística no sentido da Definição 7.1.4. Como o valor de θ deve pertencer a Ω, pode parecer razoável exigir que todo valor possível de um estimador $\delta(X_1, \ldots, X_n)$ também deva pertencer a Ω. Não exigiremos essa restrição, no entanto. Se um estimador pode assumir valores fora do espaço de parâmetros Ω, o experimentador precisará decidir no problema específico se isso parece apropriado ou se tem propriedades menos desejáveis. Na Definição 7.1.4, distinguimos

entre os termos *estatística* e *estimativa*. Como um estimador $\delta(X_1, \ldots, X_n)$ é uma função dos dados, o próprio estimador é uma variável aleatória, e sua distribuição de probabilidade pode ser derivada da distribuição conjunta de $X_1, \ldots, X_n$ se esta for conhecida. Por outro lado, uma *estimativa* é o valor específico $\delta(x_1, \ldots, x_n)$ do estimador que é determinado usando valores observados específicos $x_1, \ldots, x_n$. Se usarmos a notação vetorial $\mathbf{X} = (X_1, \ldots, X_n)$ e $\mathbf{x} = (x_1, \ldots, x_n)$, então um estimador é uma função $\delta(\mathbf{X})$ do vetor aleatório $\mathbf{X}$, e uma estimativa é um valor específico $\delta(\mathbf{x})$. Muitas vezes será conveniente denotar um estimador $\delta(\mathbf{X})$ simplesmente pelo símbolo δ.

Funções de Perda

Exemplo 7.4.2 Contagem de Calorias em Rótulos de Alimentos. No Exemplo 7.4.1, o grupo de consumidores pode sentir que, quanto mais distante sua estimativa $\delta(\mathbf{x})$ estiver da verdadeira diferença média θ, mais constrangimento e possíveis ações legais eles enfrentarão. Idealmente, eles gostariam de quantificar a magnitude das repercussões negativas como uma função de θ e da estimativa $\delta(\mathbf{x})$, e então poderiam ter alguma ideia da probabilidade de encontrar vários níveis de transtorno como resultado de sua estimação. O requisito mais fundamental de um bom estimador δ é que ele produza uma estimativa de θ que esteja próxima do valor real de θ. Em outras palavras, um bom estimador é aquele para o qual é altamente provável que o erro $\delta(\mathbf{X}) - \theta$ esteja próximo de 0. Assumiremos que para cada valor possível de $\theta \in \Omega$ e cada estimativa possível a, existe um número $L(\theta, a)$ que mede a perda ou o custo para o estatístico quando o verdadeiro valor do parâmetro é θ e sua estimativa é a. Normalmente, quanto maior a distância entre a e θ, maior será o valor de $L(\theta, a)$.

Definição 7.4.2 Função de perda. Uma *função de perda* é uma função de valor real de duas variáveis, $L(\theta, a)$, onde $\theta \in \Omega$ e a é um número real. A interpretação é que um estatístico perde $L(\theta, a)$ se o parâmetro for igual a θ e a estimativa for igual a a.

Como antes, seja $\xi(\theta)$ a f.d.p. a priori de θ no conjunto Ω, e considere um problema no qual o estatístico deve estimar o valor de θ sem poder observar os valores em uma amostra aleatória. Se o estatístico escolher uma estimativa particular a, então sua perda esperada será

$$E[L(\theta, a)] = \int_{\Omega} L(\theta, a)\xi(\theta)d\theta. \tag{7.4.1}$$

Assumiremos que o estatístico deseja escolher uma estimativa a para a qual a perda esperada em Eq. (7.4.1) é um mínimo.

Definição de um Estimador de Bayes

Suponha agora que o estatístico possa observar o valor $\mathbf{x}$ do vetor aleatório $\mathbf{X}$ antes de estimar θ, e seja $\xi(\theta|\mathbf{x})$ a f.d.p. a posteriori de θ no conjunto Ω. (O caso de um parâmetro discreto pode ser tratado de maneira semelhante.) Para cada estimativa a que o estatístico possa usar, sua perda esperada neste caso será

$$E[L(\theta, a)|\mathbf{x}] = \int_{\Omega} L(\theta, a)\xi(\theta|\mathbf{x})d\theta. \tag{7.4.2}$$

Portanto, o estatístico deve agora escolher uma estimativa a para a qual a esperança em Eq. (7.4.2) é um mínimo. Para cada valor possível $\mathbf{x}$ do vetor aleatório $\mathbf{X}$, seja $\delta^*(\mathbf{x})$ o valor de uma estimativa a para a qual a perda esperada em Eq. (7.4.2) é um mínimo. Então o estimador $\delta^*(X)$ para o qual os valores são especificados desta forma será um estimador de θ.

Definição 7.4.3 Estimador/Estimativa de Bayes. Seja $L(\theta, a)$ uma função de perda. Para cada valor possível $\mathbf{x}$ de $\mathbf{X}$, seja $\delta^*(\mathbf{x})$ um valor de a tal que $E[L(\theta, a)|\mathbf{x}]$ é minimizado. Então δ^* é chamado de *Estimador de Bayes* de θ. Uma vez que $\mathbf{X} = \mathbf{x}$ é observado, $\delta^*(\mathbf{x})$ é chamado de *Estimativa de Bayes* de θ.

Outra maneira de descrever uma estimativa de Bayes $\delta^*(\mathbf{x})$ é notar que, para cada valor possível $\mathbf{x}$ de $\mathbf{X}$, o valor $\delta^*(\mathbf{x})$ é escolhido de modo que

$$E[L(\theta, \delta^*(\mathbf{x}))|\mathbf{x}] = \min_a E[L(\theta, a)|\mathbf{x}]. \tag{7.4.3}$$

Em resumo, consideramos um problema de estimação no qual uma amostra aleatória $\mathbf{X} = (X_1, \ldots, X_n)$ deve ser retirada de uma distribuição envolvendo um parâmetro θ que tem um valor desconhecido em algum conjunto especificado Ω. Para cada função de perda $L(\theta, a)$ e cada f.d.p. a priori $\xi(\theta)$ dadas, o estimador de Bayes de θ é o estimador $\delta^*(\mathbf{X})$ para o qual a Eq. (7.4.3) é satisfeita para cada valor possível $\mathbf{x}$ de $\mathbf{X}$. Deve ser enfatizado que a forma do estimador de Bayes dependerá tanto da função de perda que é usada no problema quanto da distribuição a priori que é atribuída a θ. Nos problemas descritos neste texto, os estimadores de Bayes existirão. No entanto, existem situações mais complicadas nas quais nenhuma função δ^* satisfaz (7.4.3).

Diferentes Funções de Perda

De longe, a função de perda mais comumente usada em problemas de estimação é a função de perda de erro quadrático.

Definição 7.4.4 Função de Perda de Erro Quadrático. A função de

perda

$$L(\theta, a) = (\theta - a)^2 \tag{7.4.4}$$

é chamada de *perda de erro quadrático.*

Quando a função de perda de erro quadrático é usada, a estimativa de Bayes $\delta^*(\mathbf{x})$ para cada valor observado de $\mathbf{x}$ será o valor de a para o qual a esperança $E[(\theta-a)^2|\mathbf{x}]$ é um mínimo. O Teorema 4.7.3 afirma que a esperança de $(\theta-a)^2$ é calculada com respeito a uma certa distribuição de θ, esta esperança será um mínimo quando a for escolhido como igual à média $E(\theta|\mathbf{x})$ da distribuição a posteriori, se a média a posteriori for finita. Se a média a posteriori não for finita, então a perda esperada é infinita para cada valor possível de a. Portanto, temos o seguinte corolário do Teorema 4.7.5.

Corolário 7.4.1 Seja θ um parâmetro de valor real. Suponha que a função de perda de erro quadrático (7.4.4) seja usada e que a média a posteriori, $E(\theta|\mathbf{X})$, seja finita. Então, um estimador de Bayes de θ é $\delta^*(\mathbf{X}) = E(\theta|\mathbf{X})$.

Exemplo 7.4.3 Estimando o Parâmetro de uma Distribuição de Bernoulli. Seja a amostra aleatória $X_1, \ldots, X_n$ da distribuição de Bernoulli com parâmetro θ, que é desconhecido e deve ser estimado. Seja a distribuição a priori de θ a distribuição beta com parâmetros $\alpha > 0$ e $\beta > 0$. Suponha que a função de perda de erro quadrático seja usada, como especificado em Eq. (7.4.4), para $0 < \theta < 1$ e $0 < a < 1$. Determinaremos a estimativa de Bayes de θ. Para valores observados $x_1, \ldots, x_n$, seja $y = \sum_{i=1}^n x_i$. Segue-se do Teorema 7.3.1 que a distribuição a posteriori de θ será a distribuição beta com parâmetros $\alpha_1 = \alpha + y$ e $\beta_1 = \beta + n - y$. Uma vez que a média da distribuição beta com parâmetros α_1 e β_1 é $\alpha_1/(\alpha_1+\beta_1)$, a média da distribuição a posteriori de θ será $(\alpha + y)/(\alpha + \beta + n)$. A estimativa de Bayes $\delta^*(\mathbf{x})$ será igual a este valor para cada vetor observado $\mathbf{x}$. Portanto, o estimador de Bayes $\delta^*(\mathbf{X})$ é especificado da seguinte forma:

$$\delta^*(\mathbf{X}) = \frac{\alpha + \sum_{i=1}^n X_i}{\alpha + \beta + n}. \tag{7.4.5}$$

Exemplo 7.4.4 Estimando a Média de uma Distribuição Normal. Seja a amostra aleatória $X_1, \ldots, X_n$ de uma distribuição normal para a qual o valor da média θ é desconhecido e o valor da variância σ^2 é conhecido. Suponha também que a distribuição a priori de θ seja uma distribuição normal com média μ_0 e variância ν_0^2. Finalmente, suponha que a função de perda de erro quadrático seja usada, como especificado em Eq. (7.4.4), para $-\infty < \theta < \infty$ e $-\infty < a < \infty$. Determinaremos o estimador de Bayes de θ. Segue-se do Teorema 7.3.3 que para todos os valores observados $x_1, \ldots, x_n$, a distribuição

a posteriori de θ será a distribuição normal com média μ_1 especificada em Eq. (7.3.1). Portanto, o estimador de Bayes $\delta^*(\mathbf{X})$ é especificado da seguinte forma:

$$\delta^*(\mathbf{X}) = \frac{\sigma^2\mu_0 + n\nu_0^2\bar{X}_n}{\sigma^2 + n\nu_0^2}. \tag{7.4.6}$$

A variância a posteriori de θ não entra neste cálculo.

Outra função de perda comumente usada em problemas de estimação é a função de perda de erro absoluto.

Definição 7.4.5 Função de Perda de Erro Absoluto. A função de perda

$$L(\theta, a) = |\theta - a| \tag{7.4.7}$$

é chamada de *perda de erro absoluto.*

Para cada valor observado de $\mathbf{x}$, a estimativa de Bayes $\delta^*(\mathbf{x})$ será agora o valor de a para o qual a esperança $E(|\theta - a||\mathbf{x})$ é um mínimo. Foi mostrado no Teorema 4.5.3 que para cada distribuição de probabilidade dada de θ, a esperança de $|\theta - a|$ será um mínimo quando a for escolhido como a mediana da distribuição de θ. Portanto, quando a esperança de $|\theta - a|$ é calculada com respeito à distribuição a posteriori de θ, esta esperança será um mínimo quando a for escolhido como a mediana da distribuição a posteriori de θ.

Corolário 7.4.2 Quando a função de perda de erro absoluto (7.4.7) é usada, um estimador de Bayes de um parâmetro de valor real θ é $\delta^*(\mathbf{X})$ igual a uma mediana da distribuição a posteriori de θ.

Vamos agora reconsiderar os Exemplos 7.4.3 e 7.4.4, mas usaremos a função de perda de erro absoluto em vez da função de perda de erro quadrático.

Exemplo 7.4.5 Estimando o Parâmetro de uma Distribuição de Bernoulli. Considere novamente as condições do Exemplo 7.4.3, mas suponha que a função de perda de erro absoluto seja usada, como especificado na Eq. (7.4.7). Para todos os valores observados $x_1, \ldots, x_n$, a estimativa de Bayes $\delta^*(\mathbf{x})$ será igual à mediana da distribuição a posteriori de θ, que é a distribuição beta com parâmetros $\alpha + y$ e $\beta + n - y$. Não há expressão simples para a mediana desta distribuição. Ela deve ser determinada por aproximações numéricas para cada conjunto de valores observados. A maioria dos softwares de computador estatísticos pode calcular a mediana de uma distribuição beta arbitrária. Como

um exemplo específico, considere a situação descrita no Exemplo 7.3.13 no qual uma priori imprópria foi usada. A distribuição a posteriori de θ nesse exemplo foi a distribuição beta com parâmetros 22 e 18. A média desta distribuição beta é $22/40 = 0.55$. A mediana é 0.5508.

Exemplo 7.4.6 Estimando a Média de uma Distribuição Normal. Considere novamente as condições do Exemplo 7.4.4, mas suponha agora que a função de perda de erro absoluto seja usada, como especificado na Eq. (7.4.7). Para todos os valores observados, a estimativa de Bayes $\delta^*(\mathbf{x})$ será igual à mediana da distribuição normal a posteriori de θ. No entanto, uma vez que a média e a mediana de cada distribuição normal são iguais, $\delta^*(\mathbf{x})$ é igual à média da distribuição a posteriori. Portanto, o estimador de Bayes com respeito à função de perda de erro absoluto é o mesmo que o estimador de Bayes com respeito à função de perda de erro quadrático, e é novamente dado pela Eq. (7.4.6).

Outras Funções de Perda

Embora a função de perda de erro quadrático e, em menor grau, a função de perda de erro absoluto sejam as funções de perda mais comumente usadas em problemas de estimação, nenhuma delas pode ser apropriada em um problema de estimação particular. Em alguns problemas, pode ser apropriado usar uma função de perda com a forma $L(\theta, a) = |\theta - a|^k$, onde k é algum número positivo diferente de 1 ou 2. Em outros problemas, o fato de a estimativa a ter uma magnitude grande pode depender do valor real de θ. Em tal problema, pode ser apropriado usar uma função de perda com a forma $L(\theta, a) = \lambda_1(\theta)|\theta - a|$ ou $L(\theta, a) = \lambda_2(\theta)(\theta - a)^2$, onde $\lambda(\theta)$ é uma função positiva de θ. Em outros problemas ainda, pode ser mais custoso superestimar θ por uma certa quantidade do que subestimá-lo pela mesma quantidade. Uma função de perda específica que reflete essa propriedade é a seguinte:

$$L(\theta, a) = \begin{cases} 3(\theta - a)^2 & \text{para } \theta \leq a, \\ (\theta - a)^2 & \text{para } \theta > a. \end{cases}$$

Vários outros tipos de funções de perda podem ser relevantes em problemas de estimação específicos. No entanto, neste livro, daremos a maior parte de nossa atenção às funções de perda de erro quadrático e de erro absoluto.

A Estimativa de Bayes para Amostras Grandes

Efeito de Diferentes Distribuições a Priori. Suponha que a proporção θ de itens defeituosos em um grande lote seja desconhecida e que a distribuição a priori de θ seja a distribuição uniforme no intervalo $[0, 1]$. Suponha também que o valor de θ deva ser estimado, e que a função de perda de erro quadrático seja usada. Suponha, finalmente, que em uma amostra aleatória de 100 itens do lote, 10 itens sejam defeituosos. Como a distribuição uniforme é a distribuição

beta com parâmetros $\alpha = 1$ e $\beta = 1$, e como $n = 100$ e $y = 10$, segue-se de Eq. (7.4.5) que a estimativa de Bayes é $\delta^*(\mathbf{x}) = 11/102 = 0.108$. Agora, suponha que a f.d.p. a priori de θ tenha a forma $\xi(\theta) = 2(1-\theta)$ para $0 < \theta < 1$. Em vez de uma distribuição uniforme, e que em uma amostra aleatória de 100 itens, exatamente 10 itens sejam defeituosos. Como $\xi(\theta)$ é a f.d.p. da distribuição beta com parâmetros $\alpha = 1$ e $\beta = 2$, segue-se de Eq. (7.4.5) que neste caso a estimativa de Bayes de θ é $\delta(\mathbf{x}) = 11/103 = 0.107$. As duas distribuições a priori consideradas aqui são bastante diferentes. A média da distribuição uniforme é $1/2$, e a média da outra distribuição a priori é $1/3$. No entanto, como o número de observações na amostra é grande ($n = 100$), as estimativas de Bayes com respeito às duas distribuições a priori diferentes são quase as mesmas. Além disso, os valores de ambas as estimativas são muito próximos da proporção observada de itens defeituosos na amostra, $\bar{x}_n = 0.1$.

Exemplo 7.4.7 Medidas de Peito. Quetelet (1846) relatou (com alguns erros) dados sobre as medidas de peito (em polegadas) de 5732 milicianos escoceses. Esses dados apareceram anteriormente em um artigo de 1817 de um médico militar e foram discutidos por Stigler (1986). A Fig. 7.6 mostra um histograma dos dados. Suponha que modelamos as medidas individuais do peito como uma amostra aleatória (dado θ) de variáveis aleatórias normais com média θ e variância 4. A média do peito da amostra é $\bar{x}_n = 39.85$. Se tivéssemos a distribuição normal a priori com média μ_0 e variância ν_0^2, então usando a Eq. (7.3.1) a distribuição a posteriori de θ seria normal com média

$$\mu_1 = \frac{4\mu_0 + 5732 \times \nu_0^2 \times 39.85}{4 + 5732 \times \nu_0^2}$$

e variância

$$\nu_1^2 = \frac{4\nu_0^2}{4 + 5732 \times \nu_0^2}.$$

A estimativa de Bayes será então $\delta(\mathbf{x}) = \mu_1$. Note que, a menos que μ_0 seja incrivelmente grande ou ν_0^2 seja muito pequeno, teremos μ_1 quase igual a 39.85 e ν_1^2 quase igual a $4/5732$. De fato, se a f.d.p. a priori de θ for qualquer função contínua que seja positiva em torno de $\theta = 39.85$ e não seja extremamente grande longe de 39.85, então a f.d.p. a posteriori de θ com uma f.d.p. normal muito próxima com média 39.85 e variância $4/5732$. A média e a mediana da distribuição a posteriori estão próximas de 39.85, independentemente da distribuição a priori.

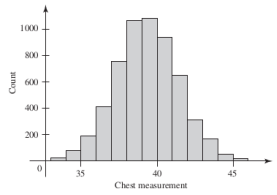


Figure 7.6 Histogram of chest measurements of Scottish militiamen in Example 7.4.7.

Consistência do Estimador de Bayes. Seja $X_1, \ldots, X_n$ uma amostra aleatória (dado θ) da distribuição de Bernoulli com parâmetro θ. Suponha que usamos uma priori conjugada para θ. Como θ é a média da distribuição da qual a amostra está sendo retirada, segue-se das leis dos grandes números discutidas na Seção 6.2 que $\bar{X}_n$ converge em probabilidade para θ quando $n \to \infty$. Como a diferença entre o estimador de Bayes $\delta^*(\mathbf{X})$ e $\bar{X}_n$ converge em probabilidade para 0 quando $n \to \infty$, também pode ser concluído que $\delta^*(\mathbf{X})$ converge em probabilidade para o valor desconhecido de θ quando $n \to \infty$.

Definição 7.4.6 Sequência Consistente de Estimadores. Uma sequência de estimadores que converge em probabilidade para o valor desconhecido do parâmetro que está sendo estimado, quando $n \to \infty$, é chamada de *sequência consistente de estimadores*.

Assim, mostramos que os estimadores de Bayes $\delta^*(\mathbf{X})$ formam uma sequência consistente de estimadores no problema considerado aqui. A interpretação prática deste resultado é a seguinte: Quando um grande número de observações é feito, há alta probabilidade de que o estimador de Bayes esteja muito próximo do valor desconhecido de θ. Os resultados que acabamos de apresentar para estimar o parâmetro de uma distribuição de Bernoulli também são verdadeiros para outros problemas de estimação. Sob condições gerais razoáveis e para uma ampla classe de funções de perda, os estimadores de Bayes de alguns parâmetros θ formarão uma sequência consistente de estimadores à medida que o tamanho da amostra $n \to \infty$. Em particular, para amostras aleatórias de qualquer uma das várias famílias de distribuições discutidas na Seção 7.3, se uma distribuição a priori conjugada for atribuída aos parâmetros e a função de perda de erro quadrático for usada, os estimadores de Bayes formarão sequências consistentes de estimadores. Por exemplo, considere novamente as condições do Exemplo 7.4.4. Nesse exemplo, uma amostra aleatória é retirada de uma distribuição normal para a qual o valor da média θ é desconhecido, e o estimador de Bayes $\delta^*(\mathbf{X})$ é especificado por Eq. (7.4.6). Como $\bar{X}_n$ convergirá para o valor desconhecido de θ quando $n \to \infty$, pode-se ver a partir da Eq. (7.4.6) que $\delta^*(\mathbf{X})$ também convergirá para θ quando $n \to \infty$. Assim, os estimadores de Bayes novamente

formam uma sequência consistente de estimadores. Outros exemplos são dados nos Exercícios 7 e 11 no final desta seção.

Parâmetros e Estimadores Mais Gerais

Até agora nesta seção, consideramos apenas parâmetros de valor real e estimadores desses parâmetros. Existem duas generalizações muito comuns desta situação que são fáceis de lidar com as mesmas técnicas descritas acima. A primeira generalização é para parâmetros multidimensionais, como o parâmetro bidimensional de uma distribuição normal com média e variância desconhecidas. A segunda generalização é para funções do parâmetro em vez do próprio parâmetro. Por exemplo, se θ é a taxa de falha no Exemplo 7.1.1, podemos estar interessados em estimar $1/\theta$, o tempo médio até a falha. Como outro exemplo, se nossos dados provêm de uma distribuição normal com média e variância desconhecidas, podemos desejar estimar apenas a média em vez do parâmetro inteiro. As mudanças necessárias na Definição 7.4.1 para lidar com ambas as generalizações que acabamos de mencionar são dadas na Definição 7.4.7.

Definição 7.4.7 Estimador/Estimativa. Seja $X_1, \ldots, X_n$ observáveis, cuja distribuição conjunta é indexada por um parâmetro θ assumindo valores em um subconjunto Ω de um espaço k-dimensional. Seja h uma função de Ω para um espaço d-dimensional. Defina $\psi = h(\theta)$. Um *estimador* de ψ é uma função $\delta(X_1, \ldots, X_n)$ que assume valores no espaço d-dimensional. Se $X_1 = x_1, \ldots, X_n = x_n$ são observados, então $\delta(x_1, \ldots, x_n)$ é chamada de *estimativa* de ψ.

Quando h na Definição 7.4.7 é a função identidade $h(\theta) = \theta$, então $\psi = \theta$ e estamos estimando o parâmetro original θ. Quando $h(\theta)$ é uma coordenada de θ, então ψ é essa coordenada, e estamos estimando apenas essa coordenada. Haverá uma série de parâmetros multidimensionais em seções posteriores e capítulos deste livro. Aqui está um exemplo de estimar uma função de um parâmetro.

Exemplo 7.4.8 Tempo de Vida de Componentes Eletrônicos. No Exemplo 7.3.12, suponha que queiramos estimar $\psi = 1/\theta$, o tempo médio até a falha dos componentes eletrônicos. A distribuição a posteriori de θ é a distribuição gama com parâmetros 4 e 8.6. Se usarmos a função de perda de erro quadrático $L(\theta, a) = (\psi - a)^2$, o Teorema 4.7.3 diz que a estimativa de Bayes é

a média da distribuição a posteriori de ψ. Que é,

$$\begin{aligned}\delta^*(\mathbf{x}) = E(\psi|\mathbf{x}) &= E\left(\frac{1}{\theta}\middle|\mathbf{x}\right)\\ &= \int_0^\infty \frac{1}{\theta}\xi(\theta|\mathbf{x})d\theta\\ &= \int_0^\infty \frac{1}{\theta}\frac{8.6^4}{6}\theta^3 e^{-8.6\theta}d\theta\\ &= \frac{8.6^4}{6}\int_0^\infty \theta^2 e^{-8.6\theta}d\theta\\ &= \frac{8.6^4}{6}\frac{2}{8.6^3} = \frac{8.6}{3} = 2.867.\end{aligned}$$

onde a igualdade final segue do Teorema 5.7.3. A média de $1/\theta$ é ligeiramente maior que $1/E(\theta|\mathbf{x}) = 8.6/4 = 2.15$.

Nota: Funções de Perda e Utilidade. Na Seção 4.8, introduzimos o conceito de utilidade para medir os valores para um tomador de decisão de vários resultados aleatórios. O conceito de função de perda está intimamente relacionado ao de utilidade. Em um sentido, uma função de perda é como o negativo de uma utilidade. De fato, o Exemplo 4.8.8 mostra como converter perda de erro absoluto em uma utilidade. Nesse exemplo, o papel do parâmetro θ e do estimador $d(W)$ desempenham os papéis do estimador. De maneira semelhante, pode-se converter outras funções de perda em utilidades. Portanto, não é surpreendente que o objetivo de maximizar a utilidade esperada na Seção 4.8 tenha sido substituído pelo objetivo de minimizar a perda esperada na seção atual.

Limitações dos Estimadores de Bayes

A teoria dos estimadores de Bayes, como descrita nesta seção, fornece uma teoria satisfatória e coerente para a estimação de parâmetros. De fato, para os estatísticos que aderem à filosofia Bayesiana, ela fornece a única teoria de estimação que pode ser desenvolvida. No entanto, existem certas limitações para a aplicabilidade da teoria em problemas estatísticos práticos. Para aplicar a teoria, é necessário especificar uma função de perda particular, como o erro quadrático ou o erro absoluto, e também uma distribuição a priori para o parâmetro. Especificações significativas podem existir, em princípio, mas pode ser muito difícil e demorado para um estatístico determiná-las. Em alguns problemas, o estatístico deve determinar as especificações que seriam apropriadas para clientes ou empregadores que não estão disponíveis ou não conseguem comunicar suas preferências e conhecimento. Em outros problemas, pode ser necessário que uma estimativa seja feita por membros de um grupo ou comitê, e pode ser difícil para os membros do grupo chegarem a um acordo sobre uma função de perda apropriada e distribuição a priori. Outra dificuldade possível é que

em um problema particular o parâmetro θ pode ser na verdade um vetor de parâmetros de valor real para os quais todos os valores são desconhecidos. A teoria dos estimadores de Bayes, que foi desenvolvida na seção anterior, pode ser facilmente generalizada para incluir a estimação de um parâmetro vetorial θ. No entanto, para aplicar esta teoria em tal problema é necessário especificar uma priori multivariada para este vetor e também especificar uma função de perda $L(\theta, \mathbf{a})$ para um vetor $\mathbf{a}$ multivariado. Mesmo que o estatístico possa estar interessado em estimar apenas um ou dois componentes do vetor θ em um determinado problema, ele ainda deve atribuir uma priori multivariada para todo o vetor θ. Em muitos problemas estatísticos, alguns dos quais serão discutidos mais adiante neste livro, θ pode ter um grande número de componentes. Em tal problema, é especialmente difícil especificar uma distribuição a priori significativa na parametrização multidimensional. Deve ser enfatizado que não há uma maneira simples de resolver essas dificuldades. Outros métodos de estimação que não se baseiam em distribuições a priori e funções de perda normalmente têm limitações práticas, e esses outros métodos também costumam ter sérios defeitos em sua estrutura teórica.

Resumo

Um estimador de um parâmetro θ é uma função δ dos dados $\mathbf{X}$. Se $\mathbf{X} = \mathbf{x}$ for observado, o valor $\delta(\mathbf{x})$ é chamado de nossa estimativa, o valor observado do estimador $\delta(\mathbf{X})$.

Exercícios

1. Em um ensaio clínico, seja θ a probabilidade de um resultado bem-sucedido. Suponha que θ tenha uma distribuição a priori que é a distribuição uniforme no intervalo $[0, 1]$, que também é a distribuição beta com parâmetros 1 e 1. Suponha que o primeiro paciente tenha um resultado bem-sucedido. Encontre as estimativas de Bayes de θ que seriam obtidas para as funções de perda de erro quadrático e de erro absoluto.

2. Suponha que a proporção θ de itens defeituosos em um grande lote seja desconhecida, e a distribuição a priori de θ seja a distribuição beta para a qual os parâmetros são $\alpha = 5$ e $\beta = 10$. Suponha também que 20 itens sejam selecionados aleatoriamente do lote, e que exatamente um desses itens seja encontrado como defeituoso. Se a função de perda de erro quadrático for usada, qual é a estimativa de Bayes de θ?

3. Considere novamente as condições do Exercício 2. Suponha que a distribuição a priori de θ seja como dada no Exercício 2, e suponha novamente que 20 itens sejam selecionados aleatoriamente do lote.

   (a) Para qual número de itens defeituosos na amostra o erro quadrático médio da estimativa de Bayes será máximo?

(b) Para qual número o erro quadrático médio da estimativa de Bayes será mínimo?

4. Suponha que uma amostra aleatória de tamanho n seja retirada da distribuição de Bernoulli com parâmetro θ, que é desconhecido, mas para a qual a distribuição a priori de θ é uma distribuição beta para a qual a média é μ_0. Mostre que a média da distribuição a posteriori de θ será uma média ponderada com a forma $\gamma_n \bar{X}_n + (1-\gamma_n)\mu_0$, e mostre que $\gamma_n \to 1$ quando $n \to \infty$.

5. Suponha que o número de defeitos em um rolo de 1200 pés de fita de gravação magnética tenha uma distribuição de Poisson para a qual o valor da média θ é desconhecido, e a distribuição a priori de θ seja a distribuição gama para a qual os parâmetros são $\alpha = 3$ e $\beta = 1$. Quando cinco rolos desta fita são selecionados aleatoriamente e inspecionados, os números de defeitos encontrados nos rolos são 2, 2, 6, 0 e 3. Se a função de perda de erro quadrático for usada, qual é a estimativa de Bayes de θ? (Ver Exercício 5 da Seção 7.3.)

6. Suponha que uma amostra aleatória de tamanho n seja retirada de uma distribuição de Poisson para a qual a média θ é desconhecida, e a distribuição a priori de θ seja uma distribuição gama para a qual a média é μ_0. Mostre que a média da distribuição a posteriori de θ será uma média ponderada com a forma $\gamma_n \bar{X}_n + (1-\gamma_n)\mu_0$, e mostre que $\gamma_n \to 1$ quando $n \to \infty$.

7. Considere novamente as condições do Exercício 6, e suponha que o valor de θ deva ser estimado usando a função de perda de erro quadrático. Mostre que os estimadores de Bayes, para $n = 1, 2, \ldots$, formam uma sequência consistente de estimadores de θ.

8. Suponha que as alturas dos indivíduos em uma certa população tenham uma distribuição normal para a qual o valor da média θ é desconhecido e o desvio padrão é 2 polegadas. Suponha também que a distribuição a priori de θ seja uma distribuição normal para a qual a média é 68 polegadas e o desvio padrão é 1 polegada. Suponha finalmente que 10 pessoas sejam selecionadas aleatoriamente da população, e sua altura média seja de 69.5 polegadas.

   (a) Se a função de perda de erro quadrático for usada, qual é a estimativa de Bayes de θ?

   (b) Se a função de perda de erro absoluto for usada, qual é a estimativa de Bayes de θ? (Ver Exercício 7 da Seção 7.3.)

9. Suponha que uma amostra aleatória deva ser retirada de uma distribuição normal para a qual o valor da média θ é desconhecido e o desvio padrão é 2, a distribuição a priori de θ é uma distribuição normal para a qual o desvio padrão é 1, e o valor de θ deva ser estimado usando a função de

perda de erro quadrático. Qual é o menor tamanho da amostra aleatória que deve ser tomado para que o erro quadrático médio do estimador de Bayes de θ seja 0.01 ou menos? (Ver Exercício 10 da Seção 7.3.)

10. Suponha que o tempo em minutos necessário para servir um cliente em um determinado caixa de banco tenha uma distribuição exponencial para a qual o valor do parâmetro θ é desconhecido, a distribuição a priori de θ é uma distribuição gama para a qual a média é 0.2 e o desvio padrão é 1, e o tempo médio necessário para servir uma amostra aleatória de 20 clientes seja observado como 3.8 minutos. Se a função de perda de erro quadrático for usada, qual é a estimativa de Bayes de θ? (Ver Exercício 12 da Seção 7.3.)

11. Suponha que uma amostra aleatória de tamanho n seja retirada de uma distribuição exponencial para a qual o valor do parâmetro θ é desconhecido, a distribuição a priori de θ é uma distribuição gama especificada, e o valor de θ deva ser estimado usando a função de perda de erro quadrático. Mostre que os estimadores de Bayes, para $n = 1, 2, \ldots$, formam uma sequência consistente de estimadores de θ.

12. Seja θ a proporção de eleitores registrados em uma grande cidade que são a favor de uma certa proposição. Suponha que o valor de θ seja desconhecido, e dois estatísticos, A e B, atribuam a θ as seguintes f.d.p.s a priori diferentes, $\xi_A(\theta)$ e $\xi_B(\theta)$, respectivamente:

$$\xi_A(\theta) = 2\theta \quad \text{para } 0 < \theta < 1,$$

$$\xi_B(\theta) = 4\theta^3 \quad \text{para } 0 < \theta < 1.$$

Em uma amostra aleatória de 1000 eleitores registrados da cidade, descobre-se que 710 são a favor da proposição.

(a) Encontre a distribuição a posteriori que cada estatístico atribui a θ.

(b) Encontre a estimativa de Bayes para cada estatístico com base na função de perda de erro quadrático.

(c) Mostre que, após as opiniões dos 1000 eleitores na amostra aleatória terem sido obtidas, as estimativas de Bayes para os dois estatísticos não poderiam diferir por mais de 0.002, independentemente do número na amostra que eram a favor da proposição.

13. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória da distribuição uniforme no intervalo $[0, \theta]$, onde o valor do parâmetro θ é desconhecido. Suponha também que a distribuição a priori de θ seja a distribuição de Pareto com parâmetros x_0 e α ($x_0 > 0$ e $\alpha > 0$), como definido no Exercício 16 da Seção 5.7. Se o valor de θ deve ser estimado usando a função de perda de erro quadrático, qual é o estimador de Bayes de θ? (Ver Exercício 18 da Seção 7.3.)

14. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição exponencial para a qual o valor do parâmetro θ é desconhecido ($\theta > 0$). Seja $\xi(\theta)$ a f.d.p. a priori de θ, e seja $\hat{\theta}$ o estimador de Bayes de θ com respeito à f.d.p. a priori $\xi(\theta)$ quando a função de perda de erro quadrático é usada. Seja $\psi = \theta^2$, e suponha que, em vez de estimar θ, seja desejado estimar o valor de ψ sujeito à seguinte função de perda de erro quadrático:
$$L(\psi, a) = (\psi - a)^2 \quad \text{para } \psi > 0 \text{ e } a > 0.$$
Seja $\hat{\psi}$ o estimador de Bayes de ψ. Explique por que $\hat{\psi} > \hat{\theta}^2$. *Dica:* Olhe o Exercício 4 na Seção 4.4.

15. Seja $c > 0$ e considere a função de perda
$$L(\theta, a) = \begin{cases} c|\theta - a| & \text{se } \theta < a, \\ |\theta - a| & \text{se } \theta \geq a. \end{cases}$$
Assuma que θ tem uma distribuição contínua. Prove que um estimador de Bayes de θ será qualquer quantil $1/(1 + c)$ da distribuição a posteriori de θ. *Dica:* A prova é muito parecida com a prova do Teorema 4.5.3. O resultado se mantém mesmo que θ não tenha uma distribuição contínua, mas a prova é mais complicada.

7.5 Estimadores de Máxima Verossimilhança

A estimação de máxima verossimilhança é um método para escolher estimadores de parâmetros que evita o uso de distribuições a priori e funções de perda. Ela escolhe como a estimativa de θ o valor de θ que fornece o maior valor da função de verossimilhança.

Introdução

Exemplo 7.5.1: Tempo de Vida de Componentes Eletrônicos. Suponha que observamos os dados no Exemplo 7.3.11 consistindo nos tempos de vida de três componentes eletrônicos. Existe um método para estimar a taxa de falha θ sem primeiro construir uma distribuição a priori e uma função de perda?

Nesta seção, desenvolveremos um método relativamente simples de construir um estimador sem ter que especificar uma função de perda e uma distribuição a priori. É chamado de método de *máxima verossimilhança*, e foi introduzido por R. A. Fisher em 1912. A estimação de máxima verossimilhança pode ser aplicada na maioria dos problemas, tem um forte apelo intuitivo, e frequentemente produzirá um estimador razoável de θ. Além disso, se a amostra for grande, o método normalmente produzirá um excelente estimador de θ. Por essas razões, o método de máxima verossimilhança é provavelmente o método de estimação mais amplamente utilizado em estatística.

Nota: Terminologia. Como a estimação de máxima verossimilhança, assim como muitos outros procedimentos a serem introduzidos posteriormente no texto, não envolve a especificação de uma distribuição a priori do parâmetro, uma terminologia um pouco diferente é frequentemente usada na descrição dos modelos estatísticos aos quais esses procedimentos são aplicados. Em vez de dizer que $X_1, \ldots, X_n$ são i.i.d. com f.p. ou f.d.p. $f(x|\theta)$ condicional a θ, podemos dizer que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição com f.p. ou f.d.p. $f(x|\theta)$ onde θ é desconhecido. Mais especificamente, no Exemplo 7.5.1, poderíamos dizer que os tempos de vida formam uma amostra aleatória da distribuição exponencial com um parâmetro de taxa de falha desconhecido θ.

Definição de um Estimador de Máxima Verossimilhança

Sejam as variáveis aleatórias $X_1, \ldots, X_n$ formando uma amostra aleatória de uma distribuição discreta ou uma distribuição contínua para a qual a f.p. ou a f.d.p. é $f(x|\theta)$, onde o parâmetro θ pertence a algum espaço de parâmetros Ω. Aqui, θ pode ser um valor real ou um vetor. Para cada vetor observado $\mathbf{x} = (x_1, \ldots, x_n)$ na amostra, o valor da f.p. conjunta ou f.d.p. conjunta será, como de costume, denotado por $f_n(\mathbf{x}|\theta)$.

Definição 7.5.1: Função de Verossimilhança. Quando a f.d.p. conjunta ou a f.p. conjunta $f_n(\mathbf{x}|\theta)$ das observações em uma amostra aleatória é considerada como uma função de θ para valores dados de $x_1, \ldots, x_n$, ela é chamada de *função de verossimilhança*.

Considere, primeiro, o caso em que o vetor observado $\mathbf{x}$ veio de uma distribuição discreta. Se uma estimativa de θ deve ser selecionada, nós certamente não selecionaríamos qualquer valor de $\theta \in \Omega$ para o qual seria impossível obter o vetor $\mathbf{x}$ que foi realmente observado. Além disso, suponha que a probabilidade $f_n(\mathbf{x}|\theta)$ de obter o vetor observado real $\mathbf{x}$ é muito alta quando θ tem um valor particular, digamos, $\theta = \theta_0$, e é muito pequena para todo outro valor de $\theta \in \Omega$. Nós então naturalmente estimaríamos o valor de θ como sendo θ_0 (a menos que tivéssemos forte informação prévia que superasse a evidência da amostra e apontasse para algum outro valor). Quando a amostra vem de uma distribuição contínua, seria novamente natural tentar encontrar um valor de θ para o qual a densidade de probabilidade $f_n(\mathbf{x}|\theta)$ é grande e usar este valor como uma estimativa de θ. Para cada vetor observado $\mathbf{x}$, somos levados por este raciocínio a considerar um valor de θ para o qual a função de verossimilhança $f_n(\mathbf{x}|\theta)$ é máxima e usar este valor como uma estimativa de θ. Este conceito é formalizado na seguinte definição.

Definição 7.5.2: Estimador/Estimativa de Máxima Verossimilhança. Para cada vetor observado possível $\mathbf{x}$, seja $\delta(\mathbf{x}) \in \Omega$ um valor de $\theta \in \Omega$ para o qual a função de verossimilhança $f_n(\mathbf{x}|\theta)$ é máxima, e seja $\delta(\mathbf{X})$ o estimador de θ definido desta forma. O estimador $\delta(\mathbf{X})$ é chamado de *estimador de máxima*

verossimilhança de θ. Depois que $\mathbf{X} = \mathbf{x}$ é observado, o valor $\delta(\mathbf{x})$ é chamado de *estimativa de máxima verossimilhança* de θ.

As expressões *estimador de máxima verossimilhança* e *estimativa de máxima verossimilhança* são abreviadas como M.L.E. (do inglês, *Maximum Likelihood Estimator/Estimate*). Deve-se confiar no contexto para determinar se a abreviação se refere a um estimador ou a uma estimativa. Note que o M.L.E. deve ser um elemento do espaço de parâmetros Ω, ao contrário dos estimadores/estimativas gerais para os quais não existe tal requisito.

Exemplos de Estimadores de Máxima Verossimilhança

Exemplo 7.5.2: Tempo de Vida de Componentes Eletrônicos. No Exemplo 7.3.11, os dados observados foram $X_1 = 3$, $X_2 = 1.5$ e $X_3 = 2.1$. As variáveis aleatórias foram modeladas como uma amostra aleatória de tamanho 3 da distribuição exponencial com parâmetro θ. A função de verossimilhança é, para $\theta > 0$,

$$f_3(\mathbf{x}|\theta) = \theta^3 \exp(-6.6\theta),$$

onde $\mathbf{x} = (2, 1.5, 2.1)$. O valor de θ que maximiza a função de verossimilhança $f_3(\mathbf{x}|\theta)$ será o mesmo que o valor de θ que maximiza $\log f_3(\mathbf{x}|\theta)$, uma vez que o logaritmo é uma função crescente. Portanto, será conveniente determinar o M.L.E. encontrando o valor de θ que maximiza

$$L(\theta) = \log f_3(\mathbf{x}|\theta) = 3\log(\theta) - 6.6\theta.$$

Tomando a derivada $dL(\theta)/d\theta$, igualando a derivada a 0, e resolvendo para θ resulta em $\theta = 3/6.6 = 0.455$. A segunda derivada é negativa neste valor de θ, então ele fornece um máximo. A estimativa de máxima verossimilhança é então 0.455.

Deve ser notado que em alguns problemas, para certos vetores observados $\mathbf{x}$, o valor máximo de $f_n(\mathbf{x}|\theta)$ pode não ser de fato alcançado para nenhum ponto $\theta \in \Omega$. Em tal caso, um M.L.E. de θ não existe. Para certos outros vetores observados $\mathbf{x}$, o valor máximo de $f_n(\mathbf{x}|\theta)$ pode na verdade ser alcançado em mais de um ponto no espaço Ω. Em tal caso, o M.L.E. não é unicamente definido, e qualquer um desses pontos pode ser escolhido como o valor do estimador $\hat{\theta}$. Em muitos problemas práticos, no entanto, o M.L.E. existe e é unicamente definido.

Ilustraremos agora o método de máxima verossimilhança e essas várias possibilidades, considerando vários exemplos. Em cada exemplo, tentaremos determinar um M.L.E.

Exemplo 7.5.3: Teste para uma Doença. Suponha que você está andando na rua e percebe que o Departamento de Saúde Pública está oferecendo um teste médico gratuito para uma certa doença. O teste é 90 por cento confiável no seguinte sentido: Se uma pessoa tem a doença, há uma probabilidade de 0.9 de que o teste dará uma resposta positiva; enquanto que, se uma pessoa

não tem a doença, há uma probabilidade de apenas 0.1 de que o teste dará uma resposta positiva. Seja X o resultado do teste, onde $X = 1$ significa que o teste é positivo e $X = 0$ significa que o teste é negativo. Seja o espaço de parâmetros $\Omega = \{0.1, 0.9\}$, onde $\theta = 0.1$ significa que a pessoa testada não tem a doença, e $\theta = 0.9$ significa que a pessoa tem a doença. Dado θ, X tem a distribuição de Bernoulli com parâmetro θ. A função de verossimilhança é

$$f(x|\theta) = \theta^x (1-\theta)^{1-x}.$$

Se $x = 0$ é observado, então

$$f(0|\theta) = \begin{cases} 0.9 & \text{se } \theta = 0.1, \\ 0.1 & \text{se } \theta = 0.9. \end{cases}$$

Claramente, $\theta = 0.1$ maximiza a verossimilhança quando $x = 0$ é observado. Se $x = 1$ é observado, então

$$f(1|\theta) = \begin{cases} 0.1 & \text{se } \theta = 0.1, \\ 0.9 & \text{se } \theta = 0.9. \end{cases}$$

Claramente, $\theta = 0.9$ maximiza a verossimilhança quando $x = 1$ é observado. Portanto, temos que o M.L.E. é

$$\hat{\theta} = \begin{cases} 0.1 & \text{se } X = 0, \\ 0.9 & \text{se } X = 1. \end{cases}$$

Exemplo 7.5.4: Amostragem de uma Distribuição de Bernoulli. Suponha que as variáveis aleatórias $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição de Bernoulli com parâmetro θ, que é desconhecido ($0 \leq \theta \leq 1$). Para todos os valores observados $x_1, \ldots, x_n$, onde cada x_i é 0 ou 1, a função de verossimilhança é

$$f_n(\mathbf{x}|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}. \quad (7.5.1)$$

Em vez de maximizar a função de verossimilhança $f_n(\mathbf{x}|\theta)$ diretamente, é novamente mais fácil maximizar $\log f_n(\mathbf{x}|\theta)$:

$$\begin{aligned} L(\theta) = \log f_n(\mathbf{x}|\theta) &= \sum_{i=1}^{n} [x_i \log\theta + (1-x_i)\log(1-\theta)] \\ &= \left(\sum_{i=1}^{n} x_i\right) \log\theta + \left(n - \sum_{i=1}^{n} x_i\right) \log(1-\theta). \end{aligned}$$

Agora calcule a derivada $dL(\theta)/d\theta$, iguale esta derivada a 0, e resolva a equação resultante para θ. Se $\sum_{i=1}^{n} x_i \notin \{0, n\}$, encontramos que a derivada é 0 em

$\theta = \bar{x}_n$, e pode ser verificado (por exemplo, examinando a segunda derivada) que este valor de fato maximiza $L(\theta)$ e a função de verossimilhança definida pela Eq. (7.5.1). Se $\sum_{i=1}^{n} x_i = 0$, então $L(\theta)$ é uma função decrescente de θ para todo θ, e portanto L atinge seu máximo em $\theta = 0$. Similarmente, se $\sum_{i=1}^{n} x_i = n$, L é uma função crescente, e atinge seu máximo em $\theta = 1$. Nestes dois últimos casos, note que o máximo da verossimilhança ocorre em $\theta = \bar{x}_n$. Portanto, segue-se que o M.L.E. de θ é $\hat{\theta} = \bar{X}_n$.

Segue-se do Exemplo 7.5.4 que se $X_1, \ldots, X_n$ são considerados como n ensaios de Bernoulli e se o espaço de parâmetros é $\Omega = [0, 1]$, então o M.L.E. da probabilidade desconhecida de sucesso é simplesmente a proporção de sucessos observados nos n ensaios. No Exemplo 7.5.3, temos $n = 1$ ensaio de Bernoulli, mas o espaço de parâmetros é $\Omega = \{0.1, 0.9\}$ em vez de $[0, 1]$, e o M.L.E. difere da proporção de sucessos.

Exemplo 7.5.5: Amostragem de uma Distribuição Normal com Média Desconhecida. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição normal para a qual a média μ é desconhecida e a variância σ^2 é conhecida. Para todos os valores observados $x_1, \ldots, x_n$, a função de verossimilhança $f_n(\mathbf{x}|\mu)$ será

$$f_n(\mathbf{x}|\mu) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]. \quad (7.5.2)$$

Pode ser visto da Eq. (7.5.2) que $f_n(\mathbf{x}|\mu)$ será maximizada pelo valor de μ que minimiza

$$Q(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2.$$

Vemos que Q é uma quadrática em μ com coeficiente positivo em μ^2. Segue-se que Q será minimizado onde sua derivada é 0. Se agora calcularmos a derivada $dQ(\mu)/d\mu$, igualarmos esta derivada a 0, e resolvermos a equação resultante para μ, encontramos que $\mu = \bar{x}_n$. Segue-se, portanto, que o M.L.E. de μ é $\hat{\mu} = \bar{X}_n$.

Pode ser visto no Exemplo 7.5.5 que o estimador $\hat{\mu}$ não é afetado pelo valor da variância σ^2, que assumimos ser conhecida. O M.L.E. da média desconhecida μ é simplesmente a média amostral $\bar{X}_n$, independentemente do valor de σ^2. Veremos isso novamente no próximo exemplo, no qual tanto μ quanto σ^2 devem ser estimados.

Exemplo 7.5.6: Amostragem de uma Distribuição Normal com Média e Variância Desconhecidas. Suponha novamente que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição normal, mas suponha agora que tanto a média μ quanto a variância σ^2 são desconhecidas. O parâmetro é então $\theta = (\mu, \sigma^2)$. Para todos os valores observados $x_1, \ldots, x_n$, a função de verossimilhança $f_n(\mathbf{x}|\mu, \sigma^2)$ será novamente dada pelo lado direito da Eq.

(7.5.2). Esta função deve agora ser maximizada sobre todos os valores possíveis de μ e σ^2, onde $-\infty < \mu < \infty$ e $\sigma^2 > 0$. Em vez de maximizar a função de verossimilhança $f_n(\mathbf{x}|\mu, \sigma^2)$ diretamente, é novamente mais fácil maximizar $\log f_n(\mathbf{x}|\mu, \sigma^2)$. Temos

$$L(\theta) = \log f_n(\mathbf{x}|\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \quad (7.5.3)$$

Nós encontraremos o valor de $\theta = (\mu, \sigma^2)$ para o qual $L(\theta)$ é máximo em três estágios. Primeiro, para cada σ^2 fixado, encontraremos o valor $\hat{\mu}(\sigma^2)$ que maximiza o lado direito de (7.5.3). Segundo, encontraremos o valor $\hat{\sigma}^2$ de σ^2 que maximiza $L(\theta')$ quando $\theta' = (\hat{\mu}(\sigma^2), \sigma^2)$. Finalmente, o M.L.E. de θ será o vetor aleatório cujo valor é $(\hat{\mu}(\hat{\sigma}^2), \hat{\sigma}^2)$. O primeiro estágio já foi resolvido no Exemplo 7.5.5. Lá, obtivemos $\hat{\mu}(\sigma^2) = \bar{x}_n$. Para o segundo estágio, definimos $\theta' = (\bar{x}_n, \sigma^2)$ e maximizamos

$$L(\theta') = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2. \quad (7.5.4)$$

Isso pode ser maximizado definindo sua derivada em relação a σ^2 igual a 0 e resolvendo para σ^2. A derivada é

$$\frac{d}{d\sigma^2}L(\theta') = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

Igualar isso a 0 resulta em

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2. \quad (7.5.5)$$

A segunda derivada de (7.5.4) é negativa no valor de σ^2 em (7.5.5), então encontramos o máximo. Portanto, o M.L.E. de $\theta = (\mu, \sigma^2)$ é

$$\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2) = \left(\bar{X}_n, \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right). \quad (7.5.6)$$

Note que a primeira coordenada do M.L.E. na Eq. (7.5.6) é chamada de *média amostral* dos dados. Da mesma forma, chamamos a segunda coordenada deste M.L.E. de *variância amostral*. Não é difícil ver que o valor observado da variância amostral é a variância de uma distribuição que atribui probabilidade $1/n$ a cada um dos n valores observados $x_1, \ldots, x_n$ na amostra.

Exemplo 7.5.7: Amostragem de uma Distribuição Uniforme. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória da distribuição uniforme

no intervalo $[0, \theta]$, onde o valor do parâmetro θ é desconhecido ($\theta > 0$). A f.d.p. $f(x|\theta)$ de cada observação tem a seguinte forma:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{para } 0 \leq x \leq \theta, \\ 0 & \text{caso contrário.} \end{cases} \quad (7.5.7)$$

Portanto, a f.d.p. conjunta $f_n(\mathbf{x}|\theta)$ de $X_1, \ldots, X_n$ tem a forma

$$f_n(\mathbf{x}|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{para } 0 \leq x_i \leq \theta \text{ (para } i = 1, \ldots, n), \\ 0 & \text{caso contrário.} \end{cases} \quad (7.5.8)$$

Pode ser visto da Eq. (7.5.8) que o M.L.E. de θ deve ser um valor de θ para o qual $\theta \geq x_i$ para $i = 1, \ldots, n$ e que maximiza $1/\theta^n$ entre todos esses valores. Como $1/\theta^n$ é uma função decrescente de θ, a estimativa será o menor valor de θ tal que $\theta \geq x_i$ para $i = 1, \ldots, n$. Como este valor é $\theta = \max\{x_1, \ldots, x_n\}$, o M.L.E. de θ é $\hat{\theta} = \max\{X_1, \ldots, X_n\}$.

Limitações da Estimação de Máxima Verossimilhança

Apesar de seu apelo intuitivo, o método de máxima verossimilhança não é necessariamente apropriado em todos os problemas. Por exemplo, no Exemplo 7.5.7, o M.L.E. $\hat{\theta}$ não parece ser um estimador adequado de θ. Como $\max\{X_1, \ldots, X_n\} \leq \theta$ com probabilidade 1, segue-se que $\hat{\theta}$ certamente subestima o valor de θ. De fato, se qualquer distribuição a priori for atribuída a θ, então o estimador de Bayes de θ será seguramente maior que $\hat{\theta}$. O valor real pelo qual o estimador de Bayes excederá $\hat{\theta}$ irá, é claro, depender da distribuição a priori particular que é usada e dos valores observados de $X_1, \ldots, X_n$. O Exemplo 7.5.7 também levanta outra dificuldade com a máxima verossimilhança, como ilustramos no Exemplo 7.5.8.

Exemplo 7.5.8: Inexistência de um M.L.E. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória da distribuição uniforme no intervalo $[0, \theta]$. No entanto, suponha agora que, em vez de escrevermos a f.d.p. $f(x|\theta)$ da distribuição uniforme na forma dada na Eq. (7.5.7), nós a escrevemos na seguinte forma:

$$f(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{para } 0 < x < \theta, \\ 0 & \text{caso contrário.} \end{cases} \quad (7.5.9)$$

A única diferença entre a Eq. (7.5.7) e a Eq. (7.5.9) é que o valor da f.d.p. em cada um dos pontos extremos 0 e θ foi alterado, substituindo as desigualdades fracas na Eq. (7.5.7) por desigualdades estritas na Eq. (7.5.9). Portanto, qualquer uma das equações poderia ser usada como a f.d.p. da distribuição uniforme. No entanto, se a Eq. (7.5.9) for usada como f.d.p., então um M.L.E. de θ será um valor de θ para o qual $\theta > x_i$ para $i = 1, \ldots, n$ e que maximiza $1/\theta^n$ entre todos esses valores. Deve-se notar que os valores possíveis de θ não incluem mais o valor $\theta = \max\{x_1, \ldots, x_n\}$, porque θ deve ser *estritamente*

maior que cada valor observado x_i $(i = 1, \ldots, n)$. Como θ pode ser escolhido arbitrariamente próximo ao valor $\max\{x_1, \ldots, x_n\}$, mas não pode ser igual a este valor, segue-se que o M.L.E. de θ não existe.

Em todas as nossas discussões anteriores sobre f.d.p.'s, enfatizamos o fato de que é irrelevante se a f.d.p. da distribuição uniforme é escolhida para ser igual a $1/\theta$ no intervalo aberto $0 < x < \theta$ ou no intervalo fechado $0 \leq x \leq \theta$. Agora, no entanto, vemos que a existência de um M.L.E. depende dessa escolha irrelevante e sem importância. Essa dificuldade é facilmente evitada no Exemplo 7.5.8 usando a f.d.p. dada pela Eq. (7.5.7) em vez daquela dada pela Eq. (7.5.9). Em muitos outros problemas também, uma dificuldade deste tipo pode ser evitada simplesmente escolhendo uma versão particular apropriada da f.d.p. para representar a distribuição. No entanto, como veremos no Exemplo 7.5.10, a dificuldade nem sempre pode ser evitada.

Exemplo 7.5.9: Não-unicidade de um M.L.E. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória da distribuição uniforme no intervalo $[\theta, \theta + 1]$, onde o valor do parâmetro θ é desconhecido $(-\infty < \theta < \infty)$. Neste exemplo, a f.d.p. conjunta $f_n(\mathbf{x}|\theta)$ tem a forma

$$f_n(\mathbf{x}|\theta) = \begin{cases} 1 & \text{para } \theta \leq x_i \leq \theta + 1 \text{ (para } i = 1, \ldots, n), \\ 0 & \text{caso contrário.} \end{cases} \quad (7.5.10)$$

A condição de que $\theta \leq x_i$ para $i = 1, \ldots, n$ é equivalente à condição de que $\theta \leq \min\{x_1, \ldots, x_n\}$. Similarmente, a condição de que $x_i \leq \theta + 1$ para $i = 1, \ldots, n$ é equivalente à condição de que $\theta \geq \max\{x_1, \ldots, x_n\} - 1$. Portanto, em vez de escrever $f_n(\mathbf{x}|\theta)$ na forma dada na Eq. (7.5.10), podemos usar a seguinte forma:

$$f_n(\mathbf{x}|\theta) = \begin{cases} 1 & \text{para } \max\{x_1, \ldots, x_n\} - 1 \leq \theta \leq \min\{x_1, \ldots, x_n\}, \\ 0 & \text{caso contrário.} \end{cases} \quad (7.5.11)$$

Assim, é possível selecionar como um M.L.E. qualquer valor de θ no intervalo

$$\max\{x_1, \ldots, x_n\} - 1 \leq \theta \leq \min\{x_1, \ldots, x_n\}. \quad (7.5.12)$$

Neste exemplo, o M.L.E. não é unicamente especificado. De fato, o método de máxima verossimilhança fornece muito pouca ajuda na escolha de uma estimativa de θ. A verossimilhança de cada valor de θ fora do intervalo (7.5.12) é na verdade 0. Nenhum valor θ fora deste intervalo seria estimado, e todos os valores dentro do intervalo são M.L.E.'s.

Exemplo 7.5.10: Amostragem de uma Mistura de Duas Distribuições. Considere uma variável aleatória X que pode vir com igual probabilidade da distribuição normal com média 0 e variância 1 ou de outra distribuição normal com média μ e variância σ^2, onde tanto μ quanto σ^2 são desconhecidos. Sob essas condições, a f.d.p. de X será a média das f.d.p.'s das duas distribuições

normais. Assim, a f.d.p. $f(x|\mu,\sigma^2)$ de X será

$$f(x|\mu,\sigma^2) = \frac{1}{2}\frac{1}{(2\pi)^{1/2}}\exp\left(-\frac{x^2}{2}\right) + \frac{1}{2}\frac{1}{(2\pi)^{1/2}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \quad (7.5.13)$$

Suponha agora que $X_1,\ldots,X_n$ formam uma amostra aleatória da distribuição para a qual a f.d.p. é dada pela Eq. (7.5.13). Como de costume, a função de verossimilhança $f_n(\mathbf{x}|\mu,\sigma^2)$ tem a forma

$$f_n(\mathbf{x}|\mu,\sigma^2) = \prod_{i=1}^{n} f(x_i|\mu,\sigma^2). \quad (7.5.14)$$

Para encontrar o M.L.E. de $\theta = (\mu,\sigma^2)$, devemos encontrar os valores de μ e σ^2 para os quais $f_n(\mathbf{x}|\mu,\sigma^2)$ é maximizada. Seja x_k um dos valores observados $x_1,\ldots,x_n$. Se fizermos $\mu = x_k$ e deixarmos $\sigma^2 \to 0$, então o fator $f(x_k|\mu,\sigma^2)$ no lado direito da Eq. (7.5.14) crescerá sem limite, enquanto cada fator $f(x_i|\mu,\sigma^2)$ para $x_i \neq x_k$ se aproximará do valor

$$\frac{1}{2(2\pi)^{1/2}}\exp\left(-\frac{x_i^2}{2}\right).$$

Portanto, quando $\mu = x_k$ e $\sigma^2 \to 0$, descobrimos que $f_n(\mathbf{x}|\mu,\sigma^2) \to \infty$. O valor 0 não é uma estimativa permissível de σ^2, porque sabemos de antemão que $\sigma^2 > 0$. Como a função de verossimilhança pode ser tornada arbitrariamente grande escolhendo $\mu = x_k$ e escolhendo σ^2 arbitrariamente próximo de 0, segue-se que o M.L.E. não existe.

Se tentarmos corrigir essa dificuldade permitindo que o valor 0 seja uma estimativa permissível de σ^2, então descobrimos que existem n M.L.E.'s diferentes de μ e σ^2; a saber,

$$\hat{\theta}_k = (\hat{\mu},\hat{\sigma}^2) = (X_k, 0) \text{ para } k = 1,\ldots,n.$$

Nenhum desses estimadores parece apropriado. Considere novamente a descrição, dada no início deste exemplo, das duas distribuições normais das quais cada observação pode ter vindo. Suponha, por exemplo, que $n = 1000$ e usamos o estimador $\hat{\theta}_3 = (X_3, 0)$. Então, estaríamos estimando o valor da variância desconhecida como sendo 0; também, estaríamos efetivamente nos comportando como se exatamente uma das X_i's (a saber, X_3) viesse da distribuição normal desconhecida, enquanto todas as outras 999 observações viessem da distribuição normal com média 0 e variância 1. De fato, no entanto, como cada observação tinha a mesma probabilidade de vir de qualquer uma das duas distribuições, é muito mais provável que centenas de observações, em vez de apenas uma, tenham vindo da distribuição normal desconhecida. Neste exemplo, o método de máxima verossimilhança é obviamente insatisfatório. Uma solução Bayesiana para este problema é delineada no Exercício 10 da Seção 12.5.

Finalmente, devemos mencionar um ponto referente à interpretação do M.L.E. O M.L.E. é o valor de θ que maximiza a f.p. ou f.d.p. condicional dos dados

X dado θ. Portanto, a estimativa de máxima verossimilhança é o valor de θ que atribuiu a maior probabilidade de ver os dados observados. Não é necessariamente o valor do parâmetro que parece ser o mais provável, dados os dados. Para dizer quão prováveis são os diferentes valores do parâmetro, seria necessária uma distribuição de probabilidade para o parâmetro. É claro que a distribuição a posteriori do parâmetro (Seção 7.2) serviria a esse propósito, mas nenhuma distribuição a posteriori está envolvida no cálculo do M.L.E. Portanto, não é legítimo interpretar o M.L.E. como o valor mais provável do parâmetro depois de ver os dados.

Por exemplo, considere uma situação coberta pelo Exemplo 7.5.4. Suponha que vamos lançar uma moeda algumas vezes, e estamos preocupados se ela tem um leve viés para cara ou para coroa. Seja $X_i = 1$ se o i-ésimo lançamento for cara e $X_i = 0$ se não. Se obtivermos quatro caras e uma coroa nos primeiros cinco lançamentos, o valor observado do M.L.E. será 0.8. Mas seria difícil imaginar uma situação em que sentiríamos que o valor mais provável de θ, a probabilidade de caras, é tão grande quanto 0.8 com base em apenas cinco lançamentos do que parecia a priori ser uma moeda típica. Tratar o M.L.E. como se fosse o valor mais provável do parâmetro é muito parecido com ignorar a informação prévia sobre a doença rara no teste médico dos Exemplos 2.3.1 e 2.3.3. Se o teste é positivo nesses exemplos, descobrimos (no Exemplo 7.5.3) que o M.L.E. assume o valor $\hat{\theta} = 0.9$, que corresponde a ter a doença. No entanto, se a probabilidade a priori de você ter a doença é tão pequena quanto no Exemplo 2.3.1, a probabilidade a posteriori de que você tenha a doença ($\theta = 0.9$) ainda é pequena mesmo após o resultado positivo do teste. O teste não é preciso o suficiente para superar completamente a informação prévia. O mesmo acontece com o lançamento da moeda; cinco lançamentos não são informação suficiente para superar as crenças anteriores sobre a moeda ser típica. Somente quando os dados contêm muito mais informação do que está disponível a priori será aproximadamente correto pensar no M.L.E. como o valor próximo do qual acreditamos que o parâmetro tem maior probabilidade de estar. Isso pode acontecer quando o M.L.E. é baseado em muitos dados ou quando há muito pouca informação a priori.

Resumo

A estimativa de máxima verossimilhança de um parâmetro θ é aquele valor de θ que fornece o maior valor da função de verossimilhança $f_n(\mathbf{x}|\theta)$ para um dado $\mathbf{x}$ fixo. Se $\delta(\mathbf{x})$ denota a estimativa de máxima verossimilhança, então $\hat{\theta} = \delta(\mathbf{X})$ é o estimador de máxima verossimilhança (M.L.E.). Calculamos o M.L.E. quando os dados compreendem uma amostra aleatória de uma distribuição de Bernoulli, uma distribuição normal com variância conhecida, uma distribuição normal com ambos os parâmetros desconhecidos, ou a distribuição uniforme no intervalo $[0, \theta]$ ou no intervalo $[\theta, \theta + 1]$.

Exercícios

1. Sejam $x_1, \ldots, x_n$ números distintos. Seja Y uma variável aleatória discreta com a seguinte f.p.:

$$f(y) = \begin{cases} \frac{1}{n} & \text{se } y \in \{x_1, \ldots, x_n\}, \\ 0 & \text{caso contrário.} \end{cases}$$

   Prove que $\text{Var}(Y)$ é dada pela Eq. (7.5.5).

2. Não se sabe qual proporção p das compras de uma certa marca de cereal matinal é feita por mulheres e qual proporção é feita por homens. Em uma amostra aleatória de 70 compras deste cereal, verificou-se que 58 foram feitas por mulheres e 12 foram feitas por homens. Encontre o M.L.E. de p.

3. Considere as condições no Exercício 2, mas suponha que se saiba que $\frac{1}{2} \leq p \leq \frac{2}{3}$. Se as observações na amostra aleatória de 70 compras são como as dadas no Exercício 2, qual é o M.L.E. de p?

4. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória da distribuição de Bernoulli com parâmetro θ, que é desconhecido, mas sabe-se que θ está no intervalo aberto $0 < \theta < 1$. Mostre que o M.L.E. de θ não existe se cada valor observado for 0 ou se cada valor observado for 1.

5. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição de Poisson para a qual a média θ é desconhecida ($\theta > 0$).

   (a) Determine o M.L.E. de θ, assumindo que pelo menos um dos valores observados é diferente de 0.

   (b) Mostre que o M.L.E. de θ não existe se cada valor observado for 0.

6. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição normal para a qual a média μ é conhecida, mas a variância σ^2 é desconhecida. Encontre o M.L.E. de σ^2.

7. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição exponencial para a qual o valor do parâmetro β é desconhecido ($\beta > 0$). Encontre o M.L.E. de β.

8. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição para a qual a f.d.p. $f(x|\theta)$ é a seguinte:

$$f(x|\theta) = \begin{cases} e^{\theta - x} & \text{para } x > \theta, \\ 0 & \text{para } x \leq \theta. \end{cases}$$

   Suponha também que o valor de θ é desconhecido ($-\infty < \theta < \infty$).

   (a) Mostre que o M.L.E. de θ não existe.

(b) Determine outra versão da f.d.p. desta mesma distribuição para a qual o M.L.E. de θ existirá, e encontre este estimador.

9. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição para a qual a f.d.p. $f(x|\theta)$ é a seguinte:

$$f(x|\theta) = \begin{cases} \theta x^{\theta-1} & \text{para } 0 < x < 1, \\ 0 & \text{caso contrário.} \end{cases}$$

Suponha também que o valor de θ é desconhecido ($\theta > 0$). Encontre o M.L.E. de θ.

10. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição para a qual a f.d.p. $f(x|\theta)$ é a seguinte:

$$f(x|\theta) = \frac{1}{2} e^{-|x-\theta|} \quad \text{para } -\infty < x < \infty.$$

Suponha também que o valor de θ é desconhecido ($-\infty < \theta < \infty$). Encontre o M.L.E. de θ. *Dica: Compare isso com o problema de minimizar o E.M.A (Erro Médio Absoluto) como no Teorema 4.5.3.*

11. Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória da distribuição uniforme no intervalo $[\theta_1, \theta_2]$, onde tanto θ_1 quanto θ_2 são desconhecidos ($-\infty < \theta_1 < \theta_2 < \infty$). Encontre os M.L.E.'s de θ_1 e θ_2.

12. Suponha que uma certa população grande contém k tipos diferentes de indivíduos ($k \geq 2$), e seja θ_i a proporção de indivíduos do tipo i, para $i = 1, \ldots, k$. Aqui, $0 \leq \theta_i \leq 1$ e $\theta_1 + \cdots + \theta_k = 1$. Suponha também que em uma amostra aleatória de n indivíduos desta população, exatamente n_i indivíduos são do tipo i, onde $n_1 + \cdots + n_k = n$. Encontre os M.L.E.'s de $\theta_1, \ldots, \theta_k$.

13. Suponha que os vetores bidimensionais $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ formam uma amostra aleatória de uma distribuição normal bivariada para a qual as médias de X e Y são desconhecidas, mas as variâncias de X e Y e a correlação entre X e Y são conhecidas. Encontre os M.L.E.'s das médias.

7.6 Propriedades dos Estimadores de Máxima Verossimilhança

Nesta seção, exploramos várias propriedades dos E.M.V.'s (Estimadores de Máxima Verossimilhança), incluindo:

- *A relação entre o E.M.V. de um parâmetro e o E.M.V. de uma função daquele parâmetro*

- *A necessidade de algoritmos computacionais*
- *O comportamento do E.M.V. à medida que o tamanho da amostra aumenta*
- *A falta de dependência do E.M.V. no plano de amostragem*

Também introduzimos um método alternativo popular de estimação (método dos momentos) que às vezes concorda com a máxima verossimilhança, mas pode ser computacionalmente mais simples.

Invariância

Exemplo 7.6.1 Tempos de Vida de Componentes Eletrônicos. No Exemplo 7.1.1, o parâmetro θ foi interpretado como a taxa de falha de componentes eletrônicos. No Exemplo 7.4.8, encontramos uma estimativa de Bayes de $\psi = 1/\theta$, o tempo de vida médio. Existe um método correspondente para calcular o E.M.V. de ψ?

Suponha que $X_1, \ldots, X_n$ formam uma amostra aleatória de uma distribuição para a qual a f.p. (função de probabilidade) ou a f.d.p. (função densidade de probabilidade) é $f(x|\theta)$, onde o valor do parâmetro θ é desconhecido. O parâmetro pode ser unidimensional ou um vetor de parâmetros. Seja $\hat{\theta}$ o E.M.V. de θ. Assim, para todos os valores observados $x_1, \ldots, x_n$, a função de verossimilhança $f_n(\mathbf{x}|\theta)$ é maximizada quando $\theta = \hat{\theta}$.

Suponha agora que mudamos o parâmetro na distribuição da seguinte forma: Em vez de expressar a f.p. ou a f.d.p. $f(x|\theta)$ em termos do parâmetro θ, vamos expressá-la em termos de um novo parâmetro $\psi = g(\theta)$, onde g é uma função um-para-um de θ. Existe uma relação entre o E.M.V. de θ e o E.M.V. de ψ?

Teorema 7.6.1 Propriedade de Invariância dos E.M.V.'s. Se $\hat{\theta}$ é o estimador de máxima verossimilhança de θ e se g é uma função um-para-um, então $g(\hat{\theta})$ é o estimador de máxima verossimilhança de $g(\theta)$.

Prova O novo espaço de parâmetros é Γ, a imagem de Ω sob a função g. Deixaremos $\theta = h(\psi)$ denotar a função inversa. Então, expressa em termos do novo parâmetro ψ, a f.p. ou f.d.p. de cada valor observado será $f[x|h(\psi)]$, e a função de verossimilhança será $f_n[\mathbf{x}|h(\psi)]$. O E.M.V. $\hat{\psi}$ de ψ será igual ao valor de ψ para o qual $f_n[\mathbf{x}|h(\psi)]$ é maximizado. Como $f_n(\mathbf{x}|\theta)$ é maximizado quando $\theta = \hat{\theta}$, segue-se que $f_n[\mathbf{x}|h(\psi)]$ é maximizado quando $h(\psi) = \hat{\theta}$. Portanto, o E.M.V. $\hat{\psi}$ deve satisfazer a relação $h(\hat{\psi}) = \hat{\theta}$ ou, equivalentemente, $\hat{\psi} = g(\hat{\theta})$. ■

Exemplo 7.6.2 Tempos de Vida de Componentes Eletrônicos. De acordo com o Teorema 7.6.1, o E.M.V. de ψ é um sobre o E.M.V. de θ. No Exemplo 7.5.2, calculamos o valor observado de $\hat{\theta} = 0.455$. O valor observado de $\hat{\psi}$ seria então $1/0.455 = 2.2$. Isso é um pouco menor do que a estimativa de Bayes usando a perda de erro quadrático de 2.867 encontrada no Exemplo 7.4.8. ▲

A propriedade de invariância pode ser estendida para funções que não são um-para-um. Por exemplo, suponha que desejamos estimar a média μ de uma distribuição normal quando tanto a média quanto a variância são desconhecidas. Então μ não é uma função um-para-um do parâmetro $\theta = (\mu, \sigma^2)$. Nesse caso,

a função que desejamos estimar é $g(\theta) = \mu$. Existe uma maneira de definir o E.M.V. de uma função de θ que não é necessariamente um-para-um. Uma maneira popular é a seguinte.

Definição 7.6.1 E.M.V. de uma função. Seja $g(\theta)$ uma função arbitrária do parâmetro, e seja G a imagem de Ω sob a função g. Para cada $t \in G$, defina $G_t = \{\theta : g(\theta) = t\}$ e defina

$$L^*(t) = \max_{\theta \in G_t} \log f_n(\mathbf{x}|\theta).$$

Finalmente, defina o E.M.V. de $g(\theta)$ como sendo $\hat{t}$ onde

$$L^*(\hat{t}) = \max_{t \in G} L^*(t). \quad (7.6.1)$$

O resultado a seguir mostra como encontrar o E.M.V. de $g(\theta)$ com base na Definição 7.6.1.

Teorema 7.6.2 Seja $\hat{\theta}$ um E.M.V. de θ, e seja $g(\theta)$ uma função de θ. Então um E.M.V. de $g(\theta)$ é $g(\hat{\theta})$.

Prova Provaremos que $\hat{t} = g(\hat{\theta})$ satisfaz (7.6.1). Como $L^*(t)$ é o máximo de $\log f_n(\mathbf{x}|\theta)$ sobre θ em um subconjunto de Ω, e como $\log f_n(\mathbf{x}|\hat{\theta})$ é o máximo sobre todo θ, sabemos que $L^*(t) \leq \log f_n(\mathbf{x}|\hat{\theta})$ para todo $t \in G$. Seja $\hat{t} = g(\hat{\theta})$. Terminamos se pudermos mostrar que $L^*(\hat{t}) = \log f_n(\mathbf{x}|\hat{\theta})$. Note que $\hat{\theta} \in G_{\hat{t}}$. Como $\hat{\theta}$ maximiza $f_n(\mathbf{x}|\theta)$ sobre todo θ, ele também maximiza $f_n(\mathbf{x}|\theta)$ sobre $\theta \in G_{\hat{t}}$. Portanto, $L^*(\hat{t}) = \log f_n(\mathbf{x}|\hat{\theta})$ e $\hat{t} = g(\hat{\theta})$ é um E.M.V. de $g(\theta)$. ■

Exemplo 7.6.3 Estimando o Desvio Padrão e o Segundo Momento. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição normal para a qual tanto a média μ quanto a variância σ^2 são desconhecidas. Determinaremos o E.M.V. do desvio padrão σ e o E.M.V. do segundo momento da distribuição normal $E(X^2)$. Foi encontrado no Exemplo 7.5.6 que o E.M.V. de $\theta = (\mu, \sigma^2)$ é $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$. A partir da propriedade de invariância, podemos concluir que o E.M.V. $\hat{\sigma}$ do desvio padrão é simplesmente a raiz quadrada da variância amostral. Em símbolos, $\hat{\sigma} = (\hat{\sigma}^2)^{1/2}$. Além disso, como $E(X^2) = \sigma^2 + \mu^2$, o E.M.V. de $E(X^2)$ será $\hat{\sigma}^2 + \hat{\mu}^2$. ▲

Consistência

Considere um problema de estimação no qual uma amostra aleatória deve ser retirada de uma distribuição envolvendo um parâmetro θ. Suponha que para todo tamanho de amostra n suficientemente grande, isto é, para todo valor de n maior que um certo número mínimo, exista um único E.M.V. de θ. Então, sob certas condições, que são tipicamente satisfeitas em problemas práticos, a sequência de E.M.V.'s é uma sequência consistente de estimadores de θ. Em outras palavras, em tais problemas, a sequência de E.M.V.'s converge em probabilidade para o valor desconhecido de θ quando $n \to \infty$.

Observamos na Seção 7.4 que, sob certas condições gerais, a sequência de estimadores de Bayes de um parâmetro θ também é uma sequência consistente

de estimadores. Portanto, para uma dada distribuição a priori e um tamanho de amostra n suficientemente grande, o estimador de Bayes e o E.M.V. de θ serão tipicamente muito próximos um do outro, e ambos estarão muito próximos do valor desconhecido de θ.

Não apresentaremos quaisquer detalhes formais das condições necessárias para provar este resultado. (Detalhes podem ser encontrados no capítulo 7 de Schervish, 1995.) Iremos, no entanto, ilustrar o resultado considerando novamente uma amostra aleatória $X_1, \ldots, X_n$ da distribuição de Bernoulli com parâmetro θ, que é desconhecido ($0 \leq \theta \leq 1$). Foi mostrado na Seção 7.4 que se a distribuição a priori de θ for uma distribuição beta, então a diferença entre o estimador de Bayes de θ e a média amostral $\bar{X}_n$ converge para 0 quando $n \to \infty$. Além disso, foi mostrado no Exemplo 7.5.4 que o E.M.V. de θ é $\bar{X}_n$. Assim, quando $n \to \infty$, a diferença entre o estimador de Bayes e o E.M.V. convergirá para 0. Finalmente, a lei dos grandes números (Teorema 6.2.4) diz que a média amostral $\bar{X}_n$ converge em probabilidade para θ quando $n \to \infty$. Portanto, tanto a sequência de estimadores de Bayes quanto a sequência de E.M.V.'s são sequências consistentes.

Cálculo Numérico

Em muitos problemas, existe um E.M.V. (Estimador de Máxima Verossimilhança) único $\hat{\theta}$ de um dado parâmetro θ, mas este E.M.V. não pode ser expresso em forma fechada como uma função das observações na amostra. Em tal problema, para um dado conjunto de valores observados, é necessário determinar o valor de $\hat{\theta}$ por cálculo numérico. Ilustraremos esta situação com dois exemplos.

Exemplo 7.6.4 Amostragem de uma Distribuição Gama. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória da distribuição gama para a qual a f.d.p. (função densidade de probabilidade) é a seguinte:

$$f(x|\alpha) = \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} \quad \text{para } x > 0. \quad (7.6.2)$$

Suponha também que o valor de α é desconhecido ($\alpha > 0$) e deve ser estimado. A função de verossimilhança é

$$f_n(\mathbf{x}|\alpha) = \frac{1}{\Gamma^n(\alpha)} \left(\prod_{i=1}^{n} x_i \right)^{\alpha-1} \exp\left(-\sum_{i=1}^{n} x_i \right). \quad (7.6.3)$$

O E.M.V. de α será o valor de α que satisfaz a equação

$$\frac{\partial \log f_n(\mathbf{x}|\alpha)}{\partial \alpha} = 0. \quad (7.6.4)$$

Quando aplicamos a Eq. (7.6.4) neste exemplo, obtemos a seguinte equação:

$$\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \frac{1}{n} \sum_{i=1}^{n} \log x_i. \quad (7.6.5)$$

Tabelas da função $\Gamma'(\alpha)/\Gamma(\alpha)$, que é chamada de *função digama*, estão incluídas em várias coleções publicadas de tabelas matemáticas. A função digama também está disponível em diversos pacotes de software matemático. Para todos os valores dados de $x_1, \ldots, x_n$, o valor único de α que satisfaz a Eq. (7.6.5) deve ser determinado ou consultando essas tabelas ou realizando uma análise numérica da função digama. Este valor será o E.M.V. de α. ▲

Exemplo 7.6.5 Amostragem de uma Distribuição de Cauchy. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição de Cauchy centrada em um ponto desconhecido θ $(-\infty < \theta < \infty)$, para a qual a f.d.p. é a seguinte:

$$f(x|\theta) = \frac{1}{\pi[1 + (x - \theta)^2]} \quad \text{para} \quad -\infty < x < \infty. \quad (7.6.6)$$

Suponha também que o valor de θ deve ser estimado. A função de verossimilhança é

$$f_n(\mathbf{x}|\theta) = \frac{1}{\pi^n \prod_{i=1}^{n}[1 + (x_i - \theta)^2]}. \quad (7.6.7)$$

Portanto, o E.M.V. de θ será o valor que minimiza

$$\prod_{i=1}^{n}[1 + (x_i - \theta)^2]. \quad (7.6.8)$$

Para a maioria dos valores de $x_1, \ldots, x_n$, o valor de θ que minimiza a expressão (7.6.8) deve ser determinado por um cálculo numérico. ▲

Uma alternativa para a solução exata da Eq. (7.6.4) é começar com um estimador heurístico de α e então aplicar o método de Newton.

Definição 7.6.2 Método de Newton. Seja $f(\theta)$ uma função de valor real de uma variável real, e suponha que desejamos resolver a equação $f(\theta) = 0$. Seja θ_0 uma estimativa inicial da solução. O *método de Newton* substitui a estimativa inicial pela estimativa atualizada

$$\theta_1 = \theta_0 - \frac{f(\theta_0)}{f'(\theta_0)}.$$

A lógica por trás do método de Newton é ilustrada na Fig. 7.7. A função $f(\theta)$ é a curva sólida. O método de Newton aproxima a curva por uma reta tangente à curva, ou seja, a linha tracejada que passa pelo ponto $(\theta_0, f(\theta_0))$, indicado pelo círculo. A reta de aproximação cruza o eixo horizontal na estimativa revisada θ_1. Tipicamente, substitui-se a estimativa inicial pela estimativa revisada e itera-se o método de Newton até que os resultados se estabilizem.

Figura 7.7 Método de Newton para aproximar a solução de $f(\theta) = 0$. A estimativa inicial é θ_0, e a estimativa revisada é θ_1.

Exemplo 7.6.6 Amostragem de uma Distribuição Gama. No Exemplo 7.6.4, suponha que observemos $n = 20$ variáveis aleatórias gama $X_1, \ldots, X_{20}$ com parâmetros α e 1. Suponha que os valores observados sejam tais que $\frac{1}{20}\sum_{i=1}^{20} \log(x_i) = 1.220$ e $\frac{1}{20}\sum_{i=1}^{20} x_i = 3.679$. Desejamos usar o método de

Newton para aproximar o E.M.V. Uma estimativa inicial razoável baseia-se no fato de que $E(X_i) = \alpha$. Isso sugere usar $\alpha_0 = 3.679$, a média amostral. A função $f(\alpha)$ é $\psi(\alpha) - 1.220$, onde ψ é a função digama. A derivada $f'(\alpha)$ é $\psi'(\alpha)$, que é conhecida como a função trigama. O método de Newton atualiza a estimativa inicial α_0 para

$$\alpha_1 = \alpha_0 - \frac{\psi(\alpha_0) - 1.220}{\psi'(\alpha_0)} = 3.679 - \frac{1.1607 - 1.220}{0.3120} = 3.871.$$

Aqui, usamos um software estatístico que calcula tanto a função digama quanto a trigama. Após mais duas iterações, a aproximação se estabiliza em 3.876. ▲

O método de Newton pode falhar terrivelmente se $f'(\theta)/f(\theta)$ se aproximar de 0 entre θ_0 e a solução real de $f(\theta) = 0$. Existe uma versão multidimensional do método de Newton, que não apresentaremos aqui. Existem também muitos outros métodos numéricos para maximizar funções. Qualquer texto sobre otimização numérica, como Nocedal e Wright (2006), descreverá alguns deles.

Método dos Momentos

Exemplo 7.6.7 Amostragem de uma Distribuição Gama. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória da distribuição gama com parâmetros α e β. No Exemplo 7.6.4, explicamos como se poderia encontrar o E.M.V. (Estimador de Máxima Verossimilhança) de α se β fosse conhecido. O método envolvia a função digama, que não é familiar para muitas pessoas. Uma estimativa de Bayes também seria difícil de encontrar neste exemplo, porque teríamos que integrar uma função que inclui um fator de $1/\Gamma(\alpha)^n$. Não há outra maneira de estimar o parâmetro vetorial θ neste exemplo? ▲

O método dos momentos é um método intuitivo para estimar parâmetros quando outros métodos, mais atraentes, podem ser muito difíceis. Ele também pode ser usado para obter uma estimativa inicial para aplicar o método de Newton.

Definição 7.6.3 Método dos Momentos. Assuma que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição indexada por um parâmetro k-dimensional θ e que tenha pelo menos k momentos finitos. Para $j = 1, \ldots, k$, seja $\mu_j(\theta) = E(X_i^j|\theta)$. Suponha que a função $\mu(\theta) = (\mu_1(\theta), \ldots, \mu_k(\theta))$ é uma função um-para-um de θ. Seja $M(\mu_1, \ldots, \mu_k)$ a função inversa, ou seja, para todo θ,

$$\theta = M(\mu_1(\theta), \ldots, \mu_k(\theta)).$$

Defina os *momentos amostrais* por $m_j = \frac{1}{n}\sum_{i=1}^n X_i^j$ para $j = 1, \ldots, k$. O *estimador pelo método dos momentos* de θ é $M(m_1, \ldots, m_j)$.

A maneira usual de implementar o método dos momentos é montar as k equações $m_j = \mu_j(\theta)$ e então resolver para θ.

Exemplo 7.6.8 Amostragem de uma Distribuição Gama. No Exemplo 7.6.4, consideramos uma amostra de tamanho n da distribuição gama com parâmetros α e 1. A média de cada uma dessas variáveis aleatórias é $\mu_1(\alpha) = \alpha$. O estimador pelo método dos momentos é então $\hat{\alpha} = m_1$, a média amostral. Essa

foi a estimativa inicial usada para iniciar o método de Newton no Exemplo 7.6.6. ▲

Exemplo 7.6.9 Amostragem de uma Distribuição Gama com Ambos os Parâmetros Desconhecidos. O Teorema 5.7.5 nos diz que os dois primeiros momentos da distribuição gama com parâmetros α e β são

$$\mu_1(\theta) = \frac{\alpha}{\beta},$$

$$\mu_2(\theta) = \frac{\alpha(\alpha+1)}{\beta^2}.$$

O método dos momentos diz para igualar os momentos populacionais aos momentos amostrais e então resolver para α e β. Neste caso, obtemos

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2},$$

$$\hat{\beta} = \frac{m_1}{m_2 - m_1^2},$$

como os estimadores pelo método dos momentos. Note que $m_2 - m_1^2$ é apenas a variância amostral. ▲

Teorema 7.6.3 Suponha que $X_1, X_2, \ldots$ são i.i.d. (independentes e identicamente distribuídas) com uma distribuição indexada by um vetor de parâmetros k-dimensional θ. Suponha que os primeiros k momentos dessa distribuição existem e são finitos para todo θ. Suponha também que a função inversa M na Definição 7.6.3 é contínua. Então, a sequência de estimadores pelo método dos momentos baseada em $X_1, \ldots, X_n$ é uma sequência consistente de estimadores de θ.

Prova A lei dos grandes números diz que os momentos amostrais convergem em probabilidade para os momentos $\mu_1(\theta), \ldots, \mu_k(\theta)$. A generalização do Teorema 6.2.5 para funções de k variáveis implica que M avaliado nos momentos amostrais (ou seja, o estimador pelo método dos momentos) converge em probabilidade para θ. ■

Exemplo 7.6.10 Amostragem de uma Distribuição Uniforme. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória da distribuição uniforme no intervalo $[\theta, \theta + 1]$. Nesse exemplo, descobrimos que o E.M.V. não é único e há um intervalo de E.M.V.'s

$$\max\{x_1, \ldots, x_n\} - 1 \leq \theta \leq \min\{x_1, \ldots, x_n\}. \quad (7.6.9)$$

Este intervalo contém todos os valores possíveis de θ que são consistentes com os dados observados. Aplicaremos agora o método dos momentos, que produzirá um único estimador. A média de cada X_i é $\theta + 1/2$, então o estimador pelo método dos momentos é $\bar{X}_n - 1/2$. Tipicamente, seria de se esperar que o valor observado do estimador pelo método dos momentos fosse um número no intervalo (7.6.9). No entanto, nem sempre é o caso. Por exemplo, se $n = 3$ e $X_1 = 0.2, X_2 = 0.99, X_3 = 0.01$ são observados, então (7.6.9) é o intervalo

$[-0.01, 0.01]$, enquanto $\bar{X}_3 = 0.4$. A estimativa pelo método dos momentos é então -0.1, que não poderia ser o valor verdadeiro de θ. ▲

Existem vários exemplos em que os estimadores pelo método dos momentos também são E.M.V.'s. Alguns destes são temas de exercícios no final desta seção.

Apesar de problemas ocasionais como o do Exemplo 7.6.10, os estimadores pelo método dos momentos serão tipicamente consistentes no sentido da Definição 7.4.6.

E.M.V.'s e Estimadores de Bayes

Estimadores de Bayes e E.M.V.'s (Estimadores de Máxima Verossimilhança) dependem dos dados unicamente através da função de verossimilhança. Eles usam a função de verossimilhança de maneiras diferentes, mas em muitos problemas eles serão muito semelhantes. Quando a função $f(\mathbf{x}|\theta)$ satisfaz certas condições de suavidade (como uma função de θ), pode ser mostrado que a função de verossimilhança tenderá a parecer cada vez mais com uma f.d.p. normal à medida que o tamanho da amostra aumenta. Mais especificamente, à medida que n aumenta, a função de verossimilhança começa a parecer uma constante (não dependendo de θ, mas possivelmente dependendo dos dados) vezes

$$\exp\left[-\frac{1}{2V_n(\theta)/n}(\theta - \hat{\theta})^2\right], \quad (7.6.10)$$

onde $\hat{\theta}$ é o E.M.V. e $V_n(\theta)$ é uma sequência de variáveis aleatórias que tipicamente converge quando $n \to \infty$ para um limite que chamaremos de $v_\infty(\theta)$. Quando n é grande, a função em (7.6.10) sobe rapidamente para seu pico à medida que θ se aproxima de $\hat{\theta}$ e então cai igualmente rápido à medida que θ se afasta de $\hat{\theta}$. Sob essas condições, desde que a f.d.p. a priori de θ seja relativamente plana em comparação com a função de verossimilhança acentuadamente pontiaguda, a f.d.p. a posteriori se parecerá muito com a verossimilhança multiplicada pela constante necessária para transformá-la em uma f.d.p. A média posterior de θ será então aproximadamente $\hat{\theta}$. De fato, a distribuição a posteriori de θ será aproximadamente a distribuição normal com média $\hat{\theta}$ e variância $V_n(\hat{\theta})/n$. De maneira similar, a distribuição do estimador de máxima verossimilhança (dado θ) será aproximadamente a distribuição normal com média θ e variância $v_\infty(\theta)/n$. As condições e provas para tornar essas afirmações precisas estão além do escopo deste texto, mas podem ser encontradas no capítulo 7 de Schervish (1995).

Exemplo 7.6.11 Amostragem de uma Distribuição Exponencial. Suponha que $X_1, X_2, \ldots$ são i.i.d. (independentes e identicamente distribuídas) tendo a distribuição exponencial com parâmetro θ. Seja $T_n = \sum_{i=1}^n X_i$. Então o E.M.V. de θ é $\hat{\theta}_n = n/T_n$. (Isso foi encontrado no Exercício 7 da Seção 7.5.) Como $1/\hat{\theta}_n$ é uma média de variáveis aleatórias i.i.d. com variância finita, o teorema central do limite nos diz que a distribuição de $1/\hat{\theta}_n$ é aproximadamente normal. A média e a variância, neste caso, dessa distribuição normal aproximada

são, respectivamente, $1/\theta$ e $1/(n\theta^2)$. O método delta (Teorema 6.3.2) diz que $\hat{\theta}$ tem então aproximadamente a distribuição normal com média θ e variância θ^2/n. Na notação acima, temos $V_n(\theta) = \theta^2$. A seguir, seja a distribuição a priori de θ a distribuição gama com parâmetros α e β. O Teorema 7.3.4 diz que a distribuição a posteriori de θ será a distribuição gama com parâmetros $\alpha+n$ e $\beta + t_n$. Concluímos mostrando que esta distribuição gama é aproximadamente uma distribuição normal. Assuma por simplicidade que α é um inteiro. Então a distribuição a posteriori de θ é a mesma que a distribuição da soma de $\alpha + n$ variáveis aleatórias exponenciais i.i.d. com parâmetro $\beta + t_n$. Tal soma tem aproximadamente a distribuição normal com média $(\alpha+n)/(\beta+t_n)$ e variância $(\alpha + n)/(\beta + t_n)^2$. Se α e β são pequenos, a média aproximada é então quase $n/t_n = \hat{\theta}$, e a variância aproximada é então quase $n^2/t_n^2 = \hat{\theta}^2/n = V_n(\hat{\theta})/n$. ▲

Exemplo 7.6.12 Mortes no Exército Prussiano. No Exemplo 7.3.14, encontramos a distribuição a posteriori de θ, o número médio de mortes por ano por coice de cavalo em unidades do exército prussiano, com base em uma amostra de 280 observações. A distribuição posterior encontrada foi a distribuição gama com parâmetros 196 e 280. Pelo mesmo argumento usado no Exemplo 7.6.11, esta distribuição gama é aproximadamente a distribuição da soma de 196 i.i.d. variáveis aleatórias exponenciais com parâmetro 280. A distribuição desta soma é aproximadamente a distribuição normal com média $196/280$ e variância $196/280^2$. Usando os mesmos dados do Exemplo 7.3.14, podemos encontrar o E.M.V. de θ, que é a média das 280 observações (de acordo com o Exercício 5 da Seção 7.5). A distribuição da média de 280 i.i.d. variáveis aleatórias de Poisson com média θ é aproximadamente a distribuição normal com média θ e variância $\theta/280$ de acordo com o teorema central do limite. Temos então $V_n(\theta) = \theta$ na notação anterior. O estimador de máxima verossimilhança com os dados observados é $\hat{\theta} = 196/280$, a média da distribuição posterior. A variância da distribuição posterior também é $V_n(\hat{\theta})/n = \hat{\theta}/280$. ▲

Figura 7.8 F.d.p. a posteriori junto com a f.d.p. do E.M.V. e a f.d.p. normal aproximada no Exemplo 7.6.13. Para a f.d.p. do E.M.V., o valor de $\theta = 3/6.6$ é usado para tornar as f.d.p.'s tão semelhantes quanto possível.

Existem duas situações comuns nas quais as distribuições a posteriori e as distribuições de E.M.V.'s não são semelhantes às distribuições normais como na discussão precedente. Uma é quando o tamanho da amostra não é muito grande, e a outra é quando a função de verossimilhança não é suave. Um exemplo com tamanho de amostra pequeno é o nosso exemplo de componentes eletrônicos.

Exemplo 7.6.13 Tempos de Vida de Componentes Eletrônicos. No Exemplo 7.3.12, temos uma amostra de $n = 3$ variáveis aleatórias exponenciais com parâmetro θ. A distribuição a posteriori encontrada lá foi a distribuição gama com parâmetros 4 e 8.6. O E.M.V. é $\hat{\theta} = 3/(X_1 + X_2 + X_3)$, que tem a distribuição de 1 sobre uma variável aleatória gama com parâmetros 3 e 3θ. A Figura 7.8 mostra a f.d.p. a posteriori junto com a f.d.p. do E.M.V. assumindo que $\theta = 3/6.6$, o valor observado do E.M.V. As duas f.d.p.'s, embora semelhantes, ainda são diferentes. Além disso, ambas as f.d.p.'s são semelhantes, mas ainda diferentes, da f.d.p. normal com a mesma média e variância que a

posterior, que também aparece no gráfico. ▲

Um exemplo de uma função de verossimilhança não suave envolve a distribuição uniforme no intervalo $[0, \theta]$.

Exemplo 7.6.14 Amostragem de uma Distribuição Uniforme. No Exemplo 7.5.7, encontramos o E.M.V. de θ com base em uma amostra de tamanho n da distribuição uniforme no intervalo $[0, \theta]$. O E.M.V. é $\hat{\theta} = \max\{X_1, \ldots, X_n\}$. Podemos encontrar a distribuição exata de $\hat{\theta}$ usando o resultado do Exemplo 3.9.6. A f.d.p. de $Y = \hat{\theta}$ é

$$g_n(y|\theta) = n[F(y|\theta)]^{n-1} f(y|\theta), \quad (7.6.11)$$

onde $f(\cdot|\theta)$ é a f.d.p. da distribuição uniforme em $[0, \theta]$ e $F(\cdot|\theta)$ é a c.d.f. (função de distribuição acumulada) correspondente. Substituindo essas funções bem conhecidas na Eq. (7.6.11) resulta na f.d.p. de $Y = \hat{\theta}$:

$$g_n(y|\theta) = n\left[\frac{y}{\theta}\right]^{n-1} \frac{1}{\theta} = \frac{n y^{n-1}}{\theta^n},$$

para $0 < y < \theta$. Esta f.d.p. não é nem um pouco parecida com uma f.d.p. normal. É muito assimétrica e tem seu máximo no maior valor possível do E.M.V. De fato, pode-se calcular a média e a variância de $\hat{\theta}$, respectivamente, como

$$E(\hat{\theta}) = \frac{n}{n+1}\theta,$$

$$Var(\hat{\theta}) = \frac{n}{(n+1)^2(n+2)}\theta^2.$$

A variância diminui como $1/n^2$ em vez de como $1/n$ nos exemplos aproximadamente normais que vimos anteriormente. Se n é grande, a distribuição a posteriori de θ terá uma f.d.p. que é aproximadamente a função de verossimilhança vezes a constante necessária para torná-la uma f.d.p. A verossimilhança está na Eq. (7.5.8). Integrar essa função sobre θ para obter a constante necessária leva à seguinte f.d.p. a posteriori aproximada de θ:

$$\xi(\theta|\mathbf{x}) \approx \begin{cases} \frac{(n-1)\hat{\theta}^{n-1}}{\theta^n} & \text{para } \theta > \hat{\theta}, \\ 0 & \text{caso contrário.} \end{cases}$$

A média e a variância desta distribuição posterior aproximada são, respectivamente, $(n-1)\hat{\theta}/(n-2)$ e $(n-1)\hat{\theta}^2/[(n-2)^2(n-3)]$. A média posterior ainda é quase igual ao E.M.V. (mas um pouco maior), e a variância posterior diminui a uma taxa como $1/n^2$, assim como a variância do E.M.V. Mas a distribuição posterior não é nem um pouco normal, já que a f.d.p. tem seu máximo no menor valor possível de θ e decresce a partir daí. ▲

O Algoritmo EM

Há uma série de situações complicadas nas quais é difícil calcular o E.M.V. (Estimador de Máxima Verossimilhança). Muitas dessas situações envolvem formas

de dados faltantes. O termo ”dados faltantes”pode se referir a vários tipos diferentes de informação. O mais óbvio seriam as observações que planejamos ou esperávamos observar, mas que não foram observadas. Por exemplo, imagine que planejamos coletar as alturas e os pesos de uma amostra de atletas. Por razões que podem estar além do nosso controle, é possível que tenhamos observado tanto as alturas quanto os pesos para a maioria dos atletas, mas apenas as alturas para um subconjunto de atletas e apenas os pesos para outro subconjunto. Se modelarmos as alturas e os pesos como tendo uma distribuição normal bivariada, podemos querer calcular o E.M.V. dos parâmetros dessa distribuição. Para uma coleção completa de pares, o Exercício 24 desta seção fornece fórmulas para o E.M.V. Não é difícil ver o quão mais complicado seria calcular o E.M.V. na situação descrita acima com dados faltantes.

O *algoritmo EM* é um método iterativo para aproximar E.M.V.'s quando dados faltantes estão dificultando encontrar o E.M.V. em forma fechada. Começa-se (como na maioria dos procedimentos iterativos) no estágio 0 com um vetor de parâmetros inicial $\theta^{(0)}$. Para passar do estágio j para o estágio $j+1$, primeiro escreve-se a *log-verossimilhança de dados completos*, que é o logaritmo da função de verossimilhança que teríamos se tivéssemos observado os dados faltantes. Os valores dos dados faltantes aparecem na log-verossimilhança de dados completos como variáveis aleatórias em vez de valores observados. O passo ”E”do algoritmo EM é o seguinte: Calcule a distribuição condicional dos dados faltantes, dados os dados observados se o parâmetro θ fosse igual a $\theta^{(j)}$, e então calcule a média condicional da log-verossimilhança de dados completos tratando θ como uma constante e os dados faltantes como variáveis aleatórias. O passo E se livra das variáveis aleatórias não observadas da log-verossimilhança de dados completos e deixa θ onde estava. Para o passo ”M”, escolha $\theta^{(j+1)}$ para maximizar o valor esperado da log-verossimilhança de dados completos que você acabou de calcular. O passo M leva você para o estágio $j+1$. Idealmente, o passo de maximização não é mais difícil do que seria se os dados faltantes tivessem sido realmente observados.

Exemplo 7.6.15 Alturas e Pesos. Suponha que tentemos observar $n = 6$ pares de alturas e pesos, mas obtemos apenas três vetores completos, mais um peso e duas alturas. Modelamos os pares como vetores aleatórios normais bivariados e queremos encontrar o E.M.V. do vetor de parâmetros $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$. (Este exemplo é apenas para fins ilustrativos. Não se pode obter uma boa estimativa de um vetor de parâmetros de cinco dimensões com apenas nove observações e nenhuma informação a priori.) Os dados estão na Tabela 7.1. Os dados faltantes são a altura $X_{4,1}$, o peso $X_{5,2}$ e a altura $X_{6,1}$. A log-verossimilhança de dados completos é a soma dos logaritmos de seis expressões da forma da Eq. (5.10.2) com cada uma das linhas da Tabela 7.1 substituída pelas variáveis fictícias (x_1, x_2). Por exemplo, o termo correspondente à quarta linha da Tabela 7.1 é

$$
\begin{aligned}
&-\log(2\pi\sigma_1\sigma_2) - \frac{1}{2}\log(1-\rho^2) \\
&\quad - \frac{1}{2(1-\rho^2)}\left[\left(\frac{68-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{68-\mu_1}{\sigma_1}\right)\left(\frac{X_{4,2}-\mu_2}{\sigma_2}\right)\right. \\
&\quad \left. + \left(\frac{X_{4,2}-\mu_2}{\sigma_2}\right)^2\right].
\end{aligned} \tag{7.6.12}
$$

Como um vetor de parâmetros inicial, escolhemos uma estimativa ingênua calculada a partir dos dados observados:

$$\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \sigma_1^{2(0)}, \sigma_2^{2(0)}, \rho^{(0)}) = (69.60, 194.75, 2.87, 14.82, 0.1764).$$

Isso consiste nos E.M.V.'s baseados nas distribuições marginais das duas coordenadas, juntamente com a correlação amostral calculada a partir das três observações completas.

Tabela 1: Alturas e pesos para o Exemplo 7.6.15. Os valores faltantes recebem nomes de variáveis aleatórias.

Altura	**Peso**
72	197
70	204
73	208
68	$X_{4,2}$
65	$X_{5,2}$
$X_{6,1}$	170

O passo E finge que $\theta = \theta^{(0)}$ e calcula a média condicional da log-verossimilhança de dados completos, dados os dados observados. Para a quarta linha da Tabela 7.1, a distribuição condicional de $X_{4,2}$ dados os dados observados e $\theta = \theta^{(0)}$ pode ser encontrada a partir do Teorema 5.10.4 como sendo a distribuição normal com média

$$194.75 + 0.1764 \times (14.82)^{1/2}\left(\frac{68-69.60}{2.87^{1/2}}\right) = 193.3$$

e variância $(1-0.1764^2)14.82^2 = 212.8$. A média condicional de $(X_{4,2}-\mu_2)^2$ seria então $212.8 + (193.3-\mu_2)^2$. A média condicional da expressão em (7.6.12) seria então

$$
\begin{aligned}
&-\log(2\pi\sigma_1\sigma_2) - \frac{1}{2}\log(1-\rho^2) \\
&\quad - \frac{1}{2(1-\rho^2)}\left[\left(\frac{68-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{68-\mu_1}{\sigma_1}\right)\left(\frac{193.3-\mu_2}{\sigma_2}\right)\right. \\
&\quad \left. + \frac{(193.3-\mu_2)^2}{\sigma_2^2} + \frac{212.8}{\sigma_2^2}\right].
\end{aligned} \tag{1}
$$

O ponto a notar sobre esta última expressão é que, exceto pelo último termo $212.8/\sigma_2^2$, é exatamente a contribuição para a log-verossimilhança que teríamos obtido se $X_{4,2}$ tivesse sido igual a 193.3, sua média condicional. Cálculos semelhantes podem ser feitos para as outras duas observações com coordenadas faltantes. Cada uma produzirá uma contribuição para a log-verossimilhança que é a variância condicional da coordenada faltante dividida por sua variância mais o que a log-verossimilhança teria sido se o valor faltante tivesse sido igual à sua média condicional. Isso torna o passo M quase idêntico a encontrar o E.M.V. para um conjunto de dados completamente observado. A única diferença das fórmulas no Exercício 24 é a seguinte: Para cada observação que está faltando X, adicione a variância condicional de X dado Y a $\sum_{i=1}^n (X_i - \bar{X}_n)^2$ tanto na fórmula para $\hat{\sigma}_1^2$ quanto na para $\hat{\rho}$. Da mesma forma, para cada observação que está faltando Y, adicione a variância condicional de Y dado X a $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ na fórmula para $\hat{\sigma}_2^2$ e $\hat{\rho}$.

Agora ilustramos a primeira iteração do algoritmo EM com os dados deste exemplo. Já temos $\theta^{(0)}$, e podemos calcular a função de log-verossimilhança a partir dos dados observados em $\theta^{(0)}$ como sendo -31.359. Para iniciar o algoritmo, já calculamos a média condicional e a variância da segunda coordenada faltante da quarta linha da Tabela 7.1. As médias e variâncias condicionais correspondentes para a quinta e sexta linhas são 190.6 e 212.8 para a quinta linha e 68.76 e 7.98 para a sexta linha. Para o passo E, substituímos os dados faltantes por suas médias condicionais e adicionamos as variâncias condicionais às somas dos desvios quadrados. Para o passo M, inserimos os valores recém-calculados nas fórmulas do Exercício 24, conforme descrito acima. O novo vetor é

$$\theta^{(1)} = (69.46, 193.81, 2.88, 14.83, 0.3742),$$

e a log-verossimilhança é -31.03. Após 32 iterações, a estimativa e a log-verossimilhança param de mudar. A estimativa final é

$$\theta^{(32)} = (68.86, 189.71, 3.15, 15.03, 0.8965),$$

com log-verossimilhança -29.66. ▲

Exemplo 7.6.16 Mistura de Distribuições Normais. Um uso muito popular do algoritmo EM é no ajuste de distribuições de mistura. Sejam $X_1, \ldots, X_n$ variáveis aleatórias tais que cada uma é amostrada ou da distribuição normal com média μ_1 e variância σ^2 (com probabilidade p) ou da distribuição normal com média μ_2 e variância σ^2 (com probabilidade $1-p$), onde $\mu_1 < \mu_2$. A restrição $\mu_1 < \mu_2$ é feita para tornar o modelo identificável no seguinte sentido. Se $\mu_1 = \mu_2$ for permitido, então todo valor de p leva à mesma distribuição conjunta dos dados observáveis. Além disso, se nenhuma das médias for restrita a estar abaixo da outra, então trocar as duas médias e mudar p para $1-p$ produzirá a mesma distribuição conjunta para os dados observáveis. A restrição $\mu_1 < \mu_2$ garante que cada vetor de parâmetro distinto produza uma distribuição conjunta diferente para os dados observáveis. Os dados na Fig. 7.4 têm a aparência típica de uma distribuição que é uma mistura de duas normais com médias não muito distantes. Como assumimos que as variâncias das duas

distribuições são as mesmas, não teremos o problema que surgiu no Exemplo 7.5.10. A função de verossimilhança das observações $X_1 = x_1, \ldots, X_n = x_n$ é

$$\prod_{i=1}^{n}\left[\frac{p}{(2\pi)^{1/2}\sigma}\exp\left(\frac{-(x_i-\mu_1)^2}{2\sigma^2}\right)+\frac{1-p}{(2\pi)^{1/2}\sigma}\exp\left(\frac{-(x_i-\mu_2)^2}{2\sigma^2}\right)\right]. \quad (7.6.13)$$

O vetor de parâmetros é $\theta = (\mu_1, \mu_2, \sigma^2, p)$, e maximizar a verossimilhança como está é um desafio. No entanto, podemos introduzir observações faltantes $Y_1, \ldots, Y_n$ onde $Y_i = 1$ se X_i foi amostrado da distribuição com média μ_1 e $Y_i = 0$ se X_i foi amostrado da distribuição com média μ_2. A log-verossimilhança de dados completos pode ser escrita como a soma do logaritmo da f.p. marginal dos dados Y ausentes mais o logaritmo da f.d.p. condicional dos dados X observados, dados os dados Y. Ou seja,

$$\sum_{i=1}^{n} Y_i \log(p) + \left(n - \sum_{i=1}^{n} Y_i\right)\log(1-p) - \frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[Y_i(x_i-\mu_1)^2 + (1-Y_i)(x_i-\mu_2)^2\right]. \quad (7.6.14)$$

No estágio j com estimativa $\theta^{(j)}$ de θ, o passo E primeiro encontra a distribuição condicional de $Y_1, \ldots, Y_n$ dados os dados observados e $\theta = \theta^{(j)}$. Como $(X_1, Y_1), \ldots, (X_n, Y_n)$ são pares independentes, podemos encontrar a distribuição condicional separadamente para cada par. A distribuição conjunta de (X_i, Y_i) é uma distribuição mista com f.p./f.d.p.

$$f(x_i, y_i|\theta^{(j)}) = \frac{p^{y_i}(1-p)^{1-y_i}}{(2\pi)^{1/2}\sigma} \times \exp\left(-\frac{1}{\sigma^{2(j)}}\left[y_i(x_i-\mu_1^{(j)})^2 + (1-y_i)(x_i-\mu_2^{(j)})^2\right]\right). \quad (2)$$

A f.d.p. marginal de X_i é o i-ésimo fator em (7.6.13). É direto determinar que a distribuição condicional de Y_i dados os dados observados é a distribuição de Bernoulli com parâmetro

$$q_i^{(j)} = \frac{p^{(j)}\exp\left(-\frac{(x_i-\mu_1^{(j)})^2}{2\sigma^{2(j)}}\right)}{p^{(j)}\exp\left(-\frac{(x_i-\mu_1^{(j)})^2}{2\sigma^{2(j)}}\right) + (1-p^{(j)})\exp\left(-\frac{(x_i-\mu_2^{(j)})^2}{2\sigma^{2(j)}}\right)}. \quad (7.6.15)$$

Como a log-verossimilhança de dados completos é uma função linear dos Y_i's, o passo E simplesmente substitui cada Y_i em (7.6.14) por $q_i^{(j)}$. O resultado é

$$\sum_{i=1}^{n} q_i^{(j)} \log(p) + \left(n - \sum_{i=1}^{n} q_i^{(j)}\right)\log(1-p) - \frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[q_i^{(j)}(x_i-\mu_1)^2 + (1-q_i^{(j)})(x_i-\mu_2)^2\right]. \quad (7.6.16)$$

Maximizar (7.6.16) é direto. Como p aparece apenas nos dois primeiros termos, vemos que $p^{(j+1)}$ é apenas a média dos $q_i^{(j)}$'s. Além disso, $\mu_1^{(j+1)}$ é a média ponderada dos x_i's com pesos $q_i^{(j)}$. Da mesma forma, $\mu_2^{(j+1)}$ é a média ponderada dos x_i's com pesos $1 - q_i^{(j)}$. Finalmente,

$$\sigma^{2(j+1)} = \frac{1}{n}\sum_{i=1}^{n}\left[q_i^{(j)}(x_i - \mu_1^{(j+1)})^2 + (1 - q_i^{(j)})(x_i - \mu_2^{(j+1)})^2\right]. \quad (7.6.17)$$

Vamos ilustrar os primeiros passos E e M usando os dados do Exemplo 7.3.10. Para o vetor de parâmetros inicial $\theta^{(0)}$, vamos deixar $\mu_1^{(0)}$ ser a média das 10 menores observações e $\mu_2^{(0)}$ ser a média das 10 maiores observações. Definimos $p^{(0)} = 1/2$, e $\sigma^{2(0)}$ é a média da variância amostral das 10 menores observações e da variância amostral das 10 maiores observações. Isso resulta em

$$\theta^{(0)} = (\mu_1^{(0)}, \mu_2^{(0)}, \sigma^{2(0)}, p^{(0)}) = (-7.65, 7.36, 46.28, 0.5).$$

Para cada uma das 20 observações x_i, calculamos $q_i^{(0)}$. Por exemplo, $x_{10} = -4.0$. De acordo com (7.6.15),

$$q_{10}^{(0)} = \frac{0.5\exp\left(-\frac{(-4.0+7.65)^2}{2\times 46.28}\right)}{0.5\exp\left(-\frac{(-4.0+7.65)^2}{2\times 46.28}\right) + 0.5\exp\left(-\frac{(-4.0-7.36)^2}{2\times 46.28}\right)} = 0.7774. \quad (3)$$

Um cálculo similar para $x_8 = 9.0$ resulta em $q_8^{(0)} = 0.0489$. A log-verossimilhança inicial, calculada como o logaritmo de (7.6.13), é -75.98. A média dos 20 valores de $q_i^{(0)}$ é $p^{(1)} = 0.4402$. A média ponderada dos valores dos dados usando os $q_i^{(0)}$'s como pesos é $\mu_1^{(1)} = -7.736$, e a média ponderada usando os $1 - q_i^{(0)}$'s é $\mu_2^{(1)} = 6.3068$. Usando (7.6.17), obtemos $\sigma^{2(1)} = 56.5491$. A log-verossimilhança sobe para -75.19. Após 25 iterações, os resultados se estabilizam em $\theta^{(25)} = (-21.9715, 2.6802, 48.6864, 0.1037)$ com uma log-verossimilhança final de -72.84. O histograma da Fig. 7.4 é reproduzido na Fig. 7.9 juntamente com a f.d.p. de uma observação da distribuição de mistura ajustada, a saber,

$$f(x) = \frac{0.1037}{(2\pi\times 48.6864)^{1/2}}\exp\left(-\frac{(x+21.9715)^2}{2\times 48.6864}\right) + \frac{1-0.1037}{(2\pi\times 48.6864)^{1/2}}\exp\left(-\frac{(x-2.6802)^2}{2\times 48.6864}\right). \quad (4)$$

Além disso, a f.d.p. ajustada baseada em uma única distribuição normal também é mostrada na Fig. 7.9. A média e a variância dessa única distribuição normal são 0.1250 e 110.6809, respectivamente. ▲

Figura 7.9 Histograma dos dados do Exemplo 7.3.10 juntamente com a f.d.p. ajustada do Exemplo 7.6.16 (curva sólida). A f.d.p. foi escalada para corresponder ao fato de que o histograma fornece contagens em vez de uma f.d.p. estimada. Além disso, a curva tracejada fornece a f.d.p. estimada para uma única distribuição normal.

Pode-se provar que a log-verossimilhança aumenta a cada iteração do algoritmo EM e que o algoritmo converge para um máximo local da função de verossimilhança. Assim como com outras rotinas de maximização numérica, é difícil garantir a convergência para um máximo global.

Planos de Amostragem

Suponha que um experimentador deseje fazer observações de uma distribuição para a qual a f.p. (função de probabilidade) ou a f.d.p. (função densidade de probabilidade) é $f(x|\theta)$ a fim de obter informações sobre o valor do parâmetro θ. O experimentador poderia simplesmente coletar uma amostra aleatória de um tamanho predeterminado. Em vez disso, no entanto, ele pode começar observando alguns valores aleatórios da distribuição e anotando o custo e o tempo gastos para fazer essas observações. Ele pode decidir observar mais alguns valores aleatórios da distribuição e estudar todos os valores obtidos até então. Em algum momento, o experimentador decidirá parar de fazer observações e estimará o valor de θ a partir de todos os valores observados que foram obtidos até aquele ponto. Ele pode decidir parar porque sente que tem informações suficientes para fazer uma boa estimativa de θ ou porque não pode mais arcar com o dinheiro ou o tempo para a amostragem.

Neste experimento, o número n de observações na amostra não é fixado de antemão. É uma variável aleatória cujo valor pode muito bem depender das magnitudes das observações à medida que são obtidas.

Suponha que um experimentador contemple o uso de um plano de amostragem no qual, para cada n, a decisão de parar ou não a amostragem após a coleta de n observações seja uma função das n observações vistas até então. Independentemente de o experimentador escolher tal plano de amostragem ou decidir fixar o valor de n antes que quaisquer observações sejam feitas, pode-se mostrar que a função de verossimilhança baseada nos valores observados é proporcional (como uma função de θ) a

$$f(x_1|\theta) \cdots f(x_n|\theta).$$

Em tal situação, o E.M.V. (Estimador de Máxima Verossimilhança) de θ dependerá apenas da função de verossimilhança e não do tipo de plano de amostragem utilizado. Em outras palavras, o valor de $\hat{\theta}$ depende apenas dos valores $x_1, \ldots, x_n$ que são realmente observados e não depende do plano (se houver algum) que foi usado pelo experimentador para decidir quando parar a amostragem.

Para ilustrar essa propriedade, suponha que os intervalos de tempo, em minutos, entre as chegadas de clientes sucessivos em uma determinada instalação

de serviço sejam variáveis aleatórias i.i.d. (independentes e identicamente distribuídas). Suponha também que cada intervalo tenha a distribuição exponencial com o parâmetro θ, e que um conjunto de intervalos observados $X_1, \ldots, X_n$ forme uma amostra aleatória dessa distribuição. Segue-se do Exercício 7 da Seção 7.5 que o E.M.V. de θ será $\hat{\theta} = 1/\bar{X}_n$. Além disso, como a média μ da distribuição exponencial é $1/\theta$, segue-se da propriedade de invariância do E.M.V. que $\hat{\mu} = \bar{X}_n$. Em outras palavras, o E.M.V. da média é a média das observações na amostra.

Considere agora os três seguintes planos de amostragem:

1. Um experimentador decide de antemão coletar exatamente 20 observações, e a média dessas 20 observações acaba sendo 6. Então o E.M.V. de μ é $\hat{\mu} = 6$.

2. Um experimentador decide coletar observações $X_1, X_2, \ldots$ até que ela obtenha um valor maior que 10. Ela descobre que $X_i < 10$ para $i = 1, \ldots, 19$ e que $X_{20} > 10$. Portanto, a amostragem termina após 20 observações. Se a média dessas 20 observações for 6, então o E.M.V. é novamente $\hat{\mu} = 6$.

3. Um experimentador coleta observações uma de cada vez, sem nenhum plano em particular, até que seja forçada a parar de amostrar ou se canse de amostrar. Ela tem certeza de que nenhuma dessas causas (ser forçada a parar ou se cansar) depende de forma alguma de μ. Se, por qualquer uma dessas razões, ela parar assim que tiver coletado 20 observações e se a média dessas 20 observações for 6, então o E.M.V. é novamente $\hat{\mu} = 6$.

Às vezes, um experimento deste tipo deve ser encerrado durante um intervalo em que o experimentador está esperando o próximo cliente chegar. Se uma certa quantidade de tempo passou desde a chegada do último cliente, esse tempo deve ser omitido da amostra de dados, embora o intervalo completo até a chegada do próximo cliente não tenha sido observado. Suponha, por exemplo, que a média das primeiras 20 observações seja 6, o experimentador espere mais 15 minutos, mas nenhum outro cliente chegue, e então ela encerre o experimento. Neste caso, sabemos que o E.M.V. de μ teria que ser maior que 6, já que o valor da 21ª observação deve ser maior que 15, mesmo que seu valor exato seja desconhecido. O novo E.M.V. pode ser obtido multiplicando a função de verossimilhança para as primeiras 20 observações pela probabilidade de a 21ª observação ser maior que 15, a saber, $\exp(-15\theta)$, e encontrando o valor de θ que maximiza esta nova função de verossimilhança (ver Exercício 15).

Lembre-se que o E.M.V. é determinado pela função de verossimilhança. A única maneira pela qual o E.M.V. pode depender do plano de amostragem é através da função de verossimilhança. Se a decisão sobre quando parar de observar os dados for baseada unicamente nas observações vistas até o momento, então essa informação já foi incluída na função de verossimilhança. Se a decisão de parar for baseada em outra coisa, é preciso avaliar a probabilidade de que "outra coisa"ocorra, dado cada valor possível de θ, e incluir essa probabilidade na verossimilhança.

Outras propriedades dos E.M.V.'s serão discutidas posteriormente neste capítulo e no Capítulo 8.

1. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição com a f.d.p. (função densidade de probabilidade) dada no Exercício 10 da Seção 7.5. Encontre o E.M.V. (Estimador de Máxima Verossimilhança) de $e^{-1/\theta}$.

2. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição de Poisson para a qual a média é desconhecida. Determine o E.M.V. do desvio padrão da distribuição.

3. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição exponencial para a qual o valor do parâmetro β é desconhecido. Determine o E.M.V. da mediana da distribuição.

4. Suponha que o tempo de vida de um certo tipo de lâmpada tenha uma distribuição exponencial para a qual o valor do parâmetro β é desconhecido. Uma amostra aleatória de n lâmpadas deste tipo é testada por um período de T horas e o número X de lâmpadas que falham durante este período é observado, mas os momentos em que as falhas ocorreram não são anotados. Determine o E.M.V. de β com base no valor observado de X.

5. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória da distribuição uniforme no intervalo $[a, b]$, onde ambos os extremos a e b são desconhecidos. Encontre o E.M.V. da média da distribuição.

6. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição normal para a qual tanto a média quanto a variância são desconhecidas. Encontre o E.M.V. do quantil 0.95 da distribuição, ou seja, do ponto θ tal que $\Pr(X < \theta) = 0.95$.

7. Para as condições do Exercício 6, encontre o E.M.V. de $v = \Pr(X > 2)$.

8. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição gama para a qual a f.d.p. é dada pela Eq. (7.6.2). Encontre o E.M.V. de $\Gamma'(\alpha)/\Gamma(\alpha)$.

9. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição gama para a qual ambos os parâmetros α e β são desconhecidos. Encontre o E.M.V. de α/β.

10. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição beta para a qual ambos os parâmetros α e β são desconhecidos. Mostre que os E.M.V.'s de α e β satisfazem a seguinte equação:

$$\frac{\Gamma'(\hat{\alpha}) - \Gamma'(\hat{\beta})}{\Gamma(\hat{\alpha})\Gamma(\hat{\beta})} = \frac{1}{n}\sum_{i=1}^{n} \log \frac{X_i}{1 - X_i}.$$

11. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de tamanho n da distribuição uniforme no intervalo $[0, \theta]$, onde o valor de θ é desconhecido. Mostre que a sequência de E.M.V.'s de θ é uma sequência consistente.

12. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição exponencial para a qual o valor do parâmetro β é desconhecido. Mostre que a sequência de E.M.V.'s de β é uma sequência consistente.

13. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória de uma distribuição para a qual a f.d.p. é como especificado no Exercício 9 da Seção 7.5. Mostre que a sequência de E.M.V.'s de θ é uma sequência consistente.

14. Suponha que um cientista deseje estimar a proporção p de borboletas-monarca que possuem um tipo especial de marcação em suas asas.

    (a) Suponha que ele capture borboletas-monarca uma de cada vez até encontrar cinco que tenham essa marcação especial. Se ele precisar capturar um total de 43 borboletas, qual é o E.M.V. de p?

    (b) Suponha que, ao final de um dia, o cientista tenha capturado 58 borboletas-monarca e tenha encontrado apenas três com a marcação especial. Qual é o E.M.V. de p?

15. Suponha que 21 observações sejam coletadas aleatoriamente de uma distribuição exponencial para a qual a média μ é desconhecida ($\mu > 0$). A média de 20 dessas observações é 6 e, embora o valor exato da outra observação não pôde ser determinado, sabia-se que era maior que 15. Determine o E.M.V. de μ.

16. Suponha que cada um de dois estatísticos, A e B, deva estimar um certo parâmetro θ cujo valor é desconhecido ($\theta > 0$). O estatístico A pode observar o valor de uma variável aleatória X, que tem a distribuição gama com parâmetros α e β, onde $\alpha = 3$ e $\beta = \theta$; o estatístico B pode observar o valor de uma variável aleatória Y, que tem a distribuição de Poisson com média 2θ. Suponha que o valor observado pelo estatístico A é $X = 2$ e o valor observado pelo estatístico B é $Y = 3$. Mostre que as funções de verossimilhança determinadas por esses valores observados são proporcionais e encontre o valor comum do E.M.V. de θ obtido por cada estatístico.

17. Suponha que cada um de dois estatísticos, A e B, deva estimar um certo parâmetro p cujo valor é desconhecido ($0 < p < 1$). O estatístico A pode observar o valor de uma variável aleatória X, que tem a distribuição binomial com parâmetros $n = 10$ e p; o estatístico B pode observar o valor de uma variável aleatória Y, que tem a distribuição binomial negativa com parâmetros $r = 4$ e p. Suponha que o valor observado pelo estatístico A é $X = 4$ e o valor observado pelo estatístico B é $Y = 6$. Mostre que as funções de verossimilhança determinadas por esses valores observados são proporcionais e encontre o valor comum do E.M.V. de p obtido por cada estatístico.

18. Prove que o estimador pelo método dos momentos para o parâmetro de uma distribuição de Bernoulli é o E.M.V.

19. Prove que o estimador pelo método dos momentos para o parâmetro de uma distribuição exponencial é o E.M.V.

20. Prove que o estimador pelo método dos momentos da média de uma distribuição de Poisson é o E.M.V.

21. Prove que os estimadores pelo método dos momentos da média e da variância de uma distribuição normal também são os E.M.V.'s.

22. Seja $X_1, \ldots, X_n$ uma amostra aleatória da distribuição uniforme no intervalo $[0, \theta]$.

    (a) Encontre o estimador pelo método dos momentos de θ.

    (b) Mostre que o estimador pelo método dos momentos não é o E.M.V.

23. Suponha que $X_1, \ldots, X_n$ formem uma amostra aleatória da distribuição beta com parâmetros α e β. Seja $\theta = (\alpha, \beta)$ o vetor de parâmetros.

    (a) Encontre o estimador pelo método dos momentos para θ.

    (b) Mostre que o estimador pelo método dos momentos não é o E.M.V.

24. Suponha que os vetores bidimensionais $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ formem uma amostra aleatória de uma distribuição normal bivariada para a qual as médias de X e Y, as variâncias de X e Y, e a correlação entre X e Y são desconhecidas. Mostre que os E.M.V.'s desses cinco parâmetros são os seguintes:

$$\hat{\mu}_1 = \bar{X}_n \quad \text{e} \quad \hat{\mu}_2 = \bar{Y}_n,$$

$$\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{e} \quad \hat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2,$$

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\left[\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right]^{1/2}\left[\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2\right]^{1/2}}.$$

    Dica: Primeiro, reescreva a f.d.p. conjunta de cada par (X_i, Y_i) como o produto da f.d.p. marginal de X_i e da f.d.p. condicional de Y_i dado X_i. Segundo, transforme os parâmetros para μ_1, σ_1^2 e

$$\alpha = \mu_2 - \frac{\rho\sigma_2}{\sigma_1}\mu_1,$$

$$\beta = \frac{\rho\sigma_2}{\sigma_1},$$

$$\sigma_{2.1}^2 = (1 - \rho^2)\sigma_2^2.$$

    Terceiro, maximize a função de verossimilhança como uma função dos novos parâmetros. Finalmente, aplique a propriedade de invariância dos E.M.V.'s para encontrar os E.M.V.'s dos parâmetros originais. A transformação acima simplifica muito a maximização da verossimilhança.

25. Considere novamente a situação descrita no Exercício 24. Desta vez, suponha que, por razões não relacionadas aos valores dos parâmetros, não podemos observar os valores de $Y_{n-k+1}, \ldots, Y_n$. Ou seja, seremos capazes de observar todos os $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_{n-k}$, mas não os últimos k valores de Y. Usando a dica do Exercício 24, encontre os E.M.V.'s de $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ e ρ.