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 parâmetro 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 $\theta > 0$,

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

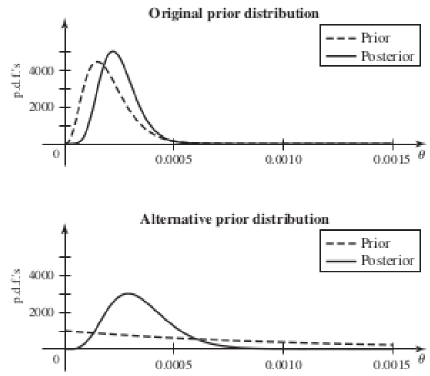
Precisamos calcular $g_n(\mathbf{x})$, que é a integral de (7.2.8) sobre todo θ:

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

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 $i = 1,\dots,n$, seja $X_i = 1$ se o i-ésimo item for defeituoso, e seja $X_i = 0$ caso contrário. Então $X_1,\dots,X_n$ 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 $X_1,\dots,X_n$ pode ser escrita na seguinte forma para $x_i = 0$ ou 1 $(i = 1,\dots,n)$:

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

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 $X_1 = x_1, \dots, X_{n-1} = x_{n-1}$ de acordo com a versão condicional da lei da probabilidade total (3.7.16). Por exemplo, se θ tem uma distribuição contínua,

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

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

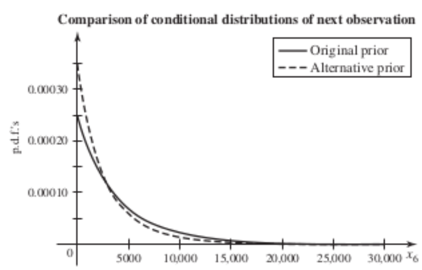
$$\Pr(X_6 > 3000|\mathbf{x}) = \int_{3000}^\infty \frac{9.555 \times 10^{41}}{9 \times 39.178^9}dx_6 = \frac{9.555 \times 10^{41}}{9 \times 39.178^9} = 0.4882.$$

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.