Estimador de Máxima Verossimilhança - Motivação

Renato Martins Assunção

DCC - UFMG

2013

Origem

- O método de máxima verossimilhança foi criado por Sir Ronald Fisher (1890 - 1962), o maior estatístico que já existiu.
- Ele foi uma espécie de Isaac Newton da estatística, responsável pelos principais conceitos e resultados da inferência estatística, usados até hoje.
- Suas idéias principais em inferência foram publicada de uma só vez, num único artigo publicado em 1922, *On the mathematical foundations of theoretical statistics*.
- Alguns dos principais conceitos (verossimilhança, suficiência e eficiência, por exemplo) e resultados que serão estudados no curso apareceram neste artigo espetacular, publicado quando ele tinha 32 anos de idade.

Sir Ronald A. Fisher

Figura: Sir Ronald A. Fisher.

Glioblastoma multiforme IV

- O mais agressivo tipo de câncer do cérebro, o tempo de vida após o diagnóstico é curto.
- Suponha que, usando o tratamento cirúrgico e terapêutico padrão nestes casos, o tempo médio de sobrevida seja de 12 meses.
- Uma inovação médica parece promissora mas é muito mais cara.
- Como não existe a certeza de que o novo tratamento seja realmente melhor que o anterior, existem também restrições éticas quanto a sua adoção indiscriminada.
- Tanto as seguradoras de saúde quanto os pacientes e médicos envolvidos precisam tomar uma decisão mais bem informada sobre a adoção do novo tratamento em substituição ao antigo.

Questão de interesse

- Suponha que $X_1, \ldots, X_n$ sejam os tempos de vida de n indivíduos após o novo tratamento cirúrgico.
- Suponha também que elas sejam variáveis aleatórias i.i.d. com distribuição contínua.
- O interesse é em fazer inferência sobre o valor esperado de X_i.
- Isto é, fazer inferência sobre $E(X_i) = \mu$
- Se μ for maior que 12 meses, o novo procedimento deveria ser considerado atentamente.

Inferência

- Para decidir se μ é maior que 12, vamos estimar μ a paryir dos dados da amostra.
- Se a amostra é grande $\overline{X} \approx \mathbb{E}(X_i) = \mu$.
- Isto é garantido por um teorema chamado de Lei dos Grandes Números:
- Qual a natureza de $\overline{X}$?
- É uma constante?
- É uma função matemática?
- É uma v.a.?

Lei dos Grandes Números

- $X_1, X_2, \ldots, X_n$ são v.a.'s iid com $\mathbb{E}(X_i) = \mu$ e $\mathbb{V}(X_i) = \sigma^2$.
- $\overline{X}$ é v.a.
- Em cada amostra particular, ela fica instanciada num n¸emero específico.
- Alguns números são mais prováveis que outros.
- Qual é a sua esperança? $\mathbb{E}(\overline{X})$? É μ.
- Qual é $\mathbb{V}(\overline{X})$? É σ^2/n
- LGN: Se $X_1, X_2, \ldots, X_n$ são v.a.'s iid com $\mathbb{E}(X_i) = \mu$ então $\overline{X} \to \mu$

Inferência

- Assim, checar o valor de $\overline{X}$ dá uma boa base para uma tomada de decisão acerca do valor de μ, principalmnete se a amostra é grande.
- Para obter $\overline{X}$ é necessário esperar que todos os indivíduos da amostra faleçam e isto pode demorar um longo tempo.
- Se o novo tratamento não for melhor nem pior que o tratamento padrão, podemos ter, por exemplo, $\mathbb{P}(X_i > 36) = 0.10$
- Numa amostra de 100 indivíduos, 3 anos após o início dos estudos ainda teríamos aproximadamente 10 pacientes ainda vivos.

Esperar ou não?

- Nem sempre é possível esperar tanto tempo.
- Os diversos interessados na decisão (pacientes, familiares, seguradoras e médicos) precisam tomar decisões, mesmo que sujeitas a revisões posteriores.
- As decisões precisam ser bem informadas mas não podem esperar tanto tempo pela coleta dos dados.
- É sempre possível rever decisões errôneas mas, num dado momento, alguma decisão deve ser tomada.
- E de preferência, ela deve uma decisão baseada nas evidências disponíveis no momento.

Censura

- Uma solução muito comum nos experimentos bioestatísticos é tomar uma *amostra censurada*.
- Observamos os pacientes até um tempo limite. Digamos, 18 meses.
- Para aqueles que viverem mais que o tempo limite, simplesmente anotamos que este evento ocorreu.
- Isto é, a amostra é composta das variáveis aleatórias $Y_1, \ldots, Y_n$ onde Y_i é igual ao tempo de vida X_i se $X_i < 18$.
- Caso ocorra o evento $X_i > 18$, então anota-se $Y_i = 18$.
- Em notação matemática:

$$Y_i = \min\{X_i, 18\} = \begin{cases} X_i, & \text{se } X_i < 18 \\ 18, & \text{se } X_i \geq 18 \end{cases}$$

Ilustração

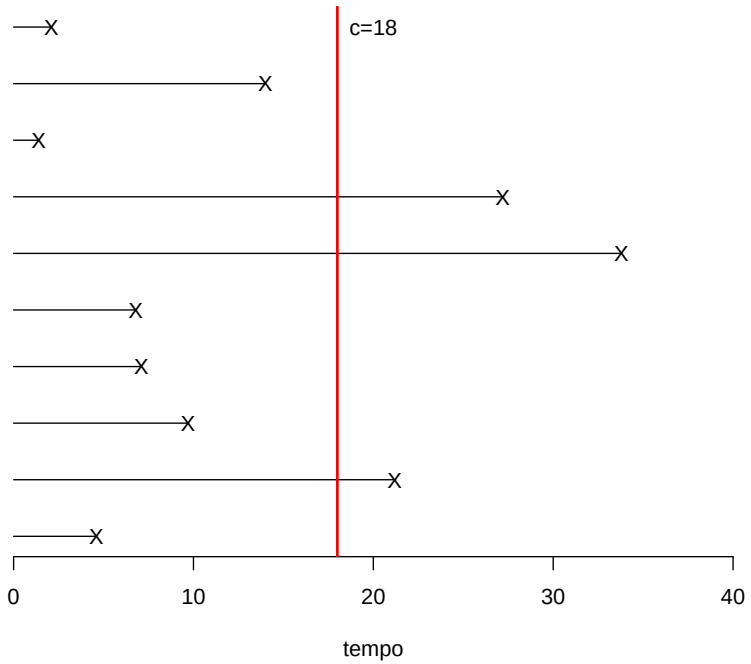

Figura: Uma amostra particular de $n = 10$ tempos de vida x_i. Cinco valores são maiores que $c = 18$ e portanto não serão observados até o fim. Sabe-se apenas que, nestes casos, $x_i \geq 18$.

Os dados

Tabela: Dados de tempos de sobrevida x_i de uma amostra de 10 indivíduos e os dados censurados y_i que seriam realmente registrados. O tempo de censura é 18 meses.

i	1	2	3	4	5	6	7	8	9	10
x_i	4.6	21.2	9.7	7.1	6.8	33.8	27.2	1.4	14.0	2.1
y_i	4.6	18	9.7	7.1	6.8	18	18	1.4	14.0	2.1

Como estimar o tempo esperado μ de sobrevida neste problema de dados censurados?

Como estimar?

- Quando os dados não eram censurados, simplesmente tomávamos a média aritmética das variáveis aleatórias X_i.
- O que nós temos agora são as variáveis Y_i que nunca superam o tempo 18.
- Os maiores tempos de sobrevida (os três valores acima de 18 na Tabela) são substituídos pelo tempo de censura (18 meses).
- Portanto, a média aritmética dos tempos censurados y_i será menor que a média aritmética dos tempos não-censurados x_i.
- Ela tende a subestimar o verdadeiro tempo esperado de sobrevida e por isto ela não é um estimador razoável para μ.

Como estimar?

- Outra opção: ignorar os tempos que foram de fato censurados.
- Tomar a média aritmética apenas dos tempos $Y_i = X_i$ restantes.
- Também não é uma boa idéia.
- Ela daria um estimador até pior que a média de todos os Y_i.
- Teríamos apenas os tempos de sobrevida de quem faleceu muito rapidamente.

A idéia de Fisher

- Fisher fez um raciocínio muito engenhoso que produz um candidato a estimador que parece razoável neste problema.
- Não só neste problema mas em quase qualquer outro modelo estatístico o mesmo raciocínio pode ser aplicado.
- Mais surpreendente ainda, este raciocínio gera estimadores imbatíveis num certo sentido. NENHUM estimador pode ser melhor que o que vamos obter aplicando o método criado por Fisher!!
- Este conceito mudou completamente a história da estatística.
- Ele transformou o que era um conjunto de idéias e técnicas desconectadas numa ciência.

Um modelo para os dados

- Para o método de Fisher, precisamos de um MODELO de probabilidade para os dados
- Tempso de vida são i.i.d
- Na prática, queremos a distribuição de X_i dependente de idade, sexo, estágio do tumor no diagnóstico, talvez via modelo de regressão.
- NESTE MOMENTO, vamos assumir que os pacientes são idênticos com relação a todas estas características que poderiam afetar a distribuição de X_i
- Isto é suponha que eles tenham a mesma idade, mesmo sexo, mesmo estágio, etc.
- Então as v.a.'s são i.i.d.

Um modelo para X_i

- Para explicar o método, assuma que $X_i \sim \exp(\lambda)$.
- Veremos o método de Fisher neste caso particular mas ele funciona do mesmo modo para QUALQUER modelo que você assuma.
- Assim, quqremos estimar $\mu = \mathbb{E}(X_i)$.
- μ está associado com λ pois $\mu = 1/\lambda$.
- Então estimar μ é o mesmo que estimar λ.

Quais valores de λ são verossímeis?

- Considere os 10 dados $y_1, \ldots, y_{10}$ registrados na amostra censurada:

$$4.6, 18^c, 9.7, 7.1, 6.8, 18^c, 18^c, 1.4, 14.0, 2.1$$

- Alguns valores de λ não eram compatíveis com os dados observados.
- Por exemplo, não parece plausível que $\lambda = 100$ (e portanto $E(X_i) = 0.01$) pois os dados observados são muito maiores que a esperança $1/\lambda = 0.01$.
- Do mesmo modo, os dados observados não dão suporte à afirmação de que $\lambda = 0.01$ (e portanto de que a esperança seja $1/0.01 = 100$).
- Estes valores extremos para λ são facilmente descartados.
- Fisher procurou pensar no princípio lógico que nós usamos para descartar esses valores extremos.

$\mathbb{P}(Y_1 \in (4.6 \pm \Delta/2))$

- Fixado algum valor para o parâmetro desconhecido λ, é possível calcular a chance de observar uma amostra tal como aquela realmente registrada.
- Por exemplo, o primeiro elemento da amostra foi igual a $y_1 = 4.6$.
- Considere um pequeno intervalo
  $(4.6 - \Delta/2,\ 4.6 + \Delta/2) = (4.6 \pm \Delta/2)$
- Como 4.6 está longe da região de censura, temos $Y_i = X_i$ e a probabilidade é igual a

$$\mathbb{P}\left(Y_1 \in (4.6 \pm \Delta/2)\right) = \mathbb{P}\left(X_1 \in (4.6 \pm \Delta/2)\right) \approx \lambda \exp(-4.6\lambda)\Delta\,,$$

- onde aproximamos a probabilidade pela área do retângulo de base Δ e altura igual a $f_\lambda(4.6)$, a densidade da distribuição exponencial no ponto 4.6.
- De maneira análoga, calculamos a probabilidade para todos os outros elementos da amostra em que o valor registrado foi de fato o tempo de sobrevida.

A probabilidade de observar uma censura

- Passemos agora aos três elementos da amostra que tiveram o valor registrado $y_i = 18^c$.
- Nós só registramos $y_i = 18^c$ se, e somente se, o valor correspondente x_i tiver sido maior que 18.
- Isto é, o tempo de sobrevida foi superior ao tempo de censura de 18 meses.
- Neste caso, a probabilidade de registrarmos $y_i = 18$ é dada por

$$\mathbb{P}(Y_i = 18^c) = \mathbb{P}(X_i \geq 18) = \exp(-18\lambda)$$

A probabilidade do que foi visto

- A probabilidade conjunta de extrairmos uma amostra aproximadamente igual a que realmente obtivemos é dada por

$$\mathbb{P}\left(Y_1 \in (4.6 \pm \Delta/2), Y_2 = 18^c, \ldots, Y_{10} \in (2.1 \pm \Delta/2)\right)$$

- Como as variáveis $Y_1, \ldots, Y_{10}$ são i.i.d., isto é igual ao produto das probabilidades marginais:

$$\mathbb{P}\left(Y_1 \in (4.6 \pm \Delta/2)\right)\ \mathbb{P}\left(Y_2 = 18^c\right)\ \ldots\ \mathbb{P}\left(Y_{10} \in (2.1 \pm \Delta/2)\right)$$

- Por sua vez, esta é igual a

$$\begin{aligned} \lambda \exp(-4.6\lambda)\Delta\ \exp(-18\lambda)\ \ldots\ \lambda \exp(-2.1\lambda)\Delta &= \lambda^7 \exp(-\lambda(4.6 + 18 + \ldots + 2.1))\Delta^7 \\ &= \lambda^7 \exp(-99.7\lambda)\Delta^7 \end{aligned}$$

- Note que o expoente da exponencial multiplica λ pela soma 99.7 dos 10 valores registrados de y_i, somando tanto os 3 valores censurados quanto os 7 outros não censurados.

$L(\lambda)$: Likelihood function

- Temos

$$\mathbb{P}(Y_1 \approx 4.6, Y_2 = 18^c, \ldots, Y_{10} \approx 2.1) = \underbrace{\lambda^7 \exp(-99.7\lambda)}_{L(\lambda)} \Delta^7$$

- Considerando os dados da amostra como números fixos, a expressão $L(\lambda)$ é função apenas de λ.
- O valor de Δ é completamente arbitrário e é escolhido pelo usuário sem relação com o verdadeiro valor do parâmetro λ ou com os valores dos dados na amostra.
- Note que se variarmos λ, o valor de Δ não se altera. Assim, com respeito a λ, o valor de Δ é uma constante.

A função $L(\lambda)$

- Para diferentes valores de λ teremos valores diferentes da probabilidade aproximada $L(\lambda)\Delta^7$ de obter uma amostra tal como a que realmente obtivemos.
- Para valores tais como $\lambda > 0.15$, a probabilidade de obter a amostra é praticamente zero.

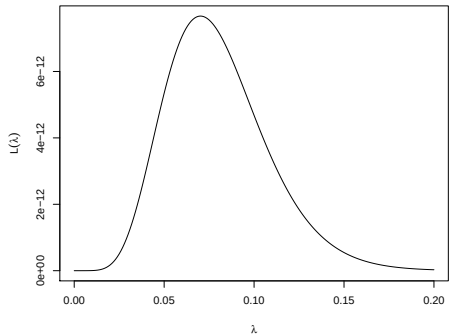

Figura: Gráfico da função $L(\lambda) = \lambda^7 \exp(-99.7\lambda)$ versus λ.

Verossimilhança - Likelihood

- Fisher dizia que estes valores tão extremos para λ não são *verossímeis*.
- *vero:* verdadeiro, real, autêntico; *símil:* semelhante, similar.
- algo é verossímil se parece verdadeiro, se não repugna à verdade, se é semelhante à verdade, se é coerente o suficiente para se passar por verdade.
- Portanto, ao dizer que algo é verossímil, não dizemos que é verdadeiro mas que parece verdadeiro pois está de acordo com todas as evidências disponíveis
- A notação $L(\lambda)$ é devido à palavra *likelihood*.

Verossimilhança relativa

- Compare dois valores de λ, $\lambda = 0.06$ e $\lambda = 0.15$, quanto a sua verossimilhança, quanto a sua suposta veracidade, levando em conta os dados que foram observados.
- $$\frac{L(0.06)}{L(0.15)} = \frac{0.06^7 \exp(-99.7 * 0.06)}{0.15^7 \exp(-99.7 * 0.15)} = 12.92$$
- Quando $\lambda = 0.06$, a probabilidade de obter uma amostra como a que realmente obtivemos é quase 13 vezes maior que a mesma probabilidade quando $\lambda = 0.15$.

Verossimilhança relativa

- Neste sentido, o valor $\lambda = 0.06$ é mais verossímil que o valor $\lambda = 0.15$.
- Ambos podem ser considerados como candidatos para λ mas os dados que observamos na amostra podem ocorrer com probabilidade muito maior quando $\lambda = 0.06$ do que quando $\lambda = 0.15$.
- Se temos que inferir sobre o verdadeiro valor de λ com base nesta amostra, porquê alguém iria preferir $\lambda = 0.15$ a $\lambda = 0.06$?

EMV

- A idéia então é acompanhar os valores $L(\lambda)$ à medida em que os valores de λ varrem o espaço paramétrico.
- Quanto maior o valor de $L(\lambda)$, mais verossímil o valor de λ correspondente.
- O valor de λ que leva ao valor máximo de $L(\lambda)$ é chamado de estimativa de máxima verossimilhança.
- Maximum Likelihood Estimator: MLE
- Estimador Máxima Verossimilhança: EMV

Obtendo o EMV: visualmente

Pela figura, o valor de λ que vai maximizar $L(\lambda)$ é aproximadamente 0.075.

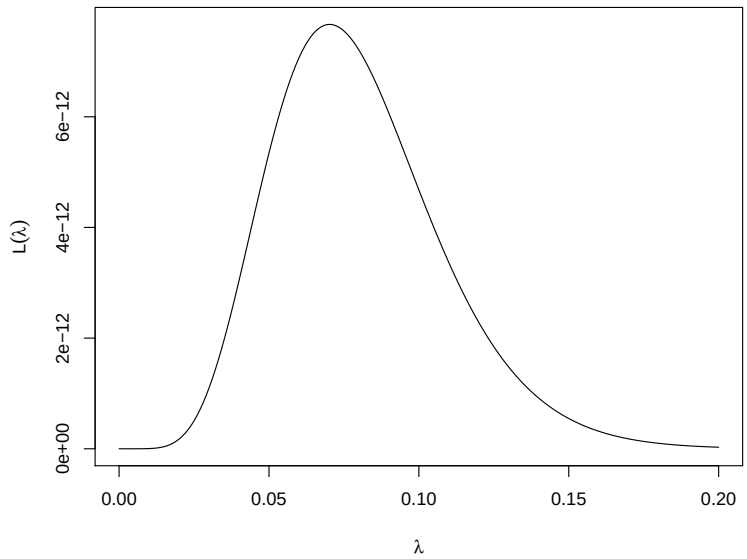

Figura: Gráfico da função $L(\lambda) = \lambda^7 \exp(-99.7\lambda)$ versus λ.

Obtendo o EMV: analiticamente

- Basta derivar $L(\lambda)$ com respeito a λ e igualar a zero:

$$\frac{dL(\lambda)}{d\lambda} = 7\lambda^6 e^{-99.7\lambda} - 99.7\lambda^7 e^{-99.7\lambda} = 0$$

  o que implica em

$$7 - 99.7\lambda = 0\,,$$

- cuja solução é $\hat{\lambda} = 7/99.7$. Portanto, uma estimativa para o tempo médio de sobrevida é

$$\frac{1}{\hat{\lambda}} = \frac{99.7}{7} = \frac{10}{7}\,\frac{99.7}{10} = \frac{10}{7}\overline{Y}\,.$$

Obtendo o EMV: analiticamente

- Primeiro: calcule a média artimética $\overline{Y}$ de todos os valores da amostra, censurados e não-censurados, obtendo $99.7/10$.
- Esta estimativa $\overline{Y}$ tende a ser menor que o valor verdadeiro e deveríamos aumentá-la.
- Este é o papel do fator $10/7 > 1$ que, multiplicando a média $\overline{Y}$, vai trazer a estimativa mais para perto do valor verdadeiro.

O caso geral

- $k =$ número de observações censuradas
- $\sum_i Y_i$ é a soma de todos os n valores registrados (k censurados e $n - k$ não censurados),
- Então

$$\widehat{\mu} = \frac{1}{\widehat{\lambda}} = \frac{n}{n-k} \frac{\sum_i Y_i}{n} = \frac{n}{n-k} \overline{Y} \tag{1}$$

- Se $k = 0$, o EMV é a media aritmética simples $\sum_i Y_i/n$.
- Se $k > 0$, a fração $n/(n-k)$ será maior que 1 e seu efeito é dilatar a subestimativa $\overline{Y}$, possivelmente trazendo-a mais para perto do verdadeiro valor que queremos estimar.
- A fração $n/(n-k)$ aumenta com o número de observações censuradas.
- Se todas as observações forem censuradas (isto é, se $k = n$), o estimador de máxima verossimilhança não está definido.

Generalidade

- O método de máxima versossimilhança pode ser aplicado em praticamente toda situação de inferência em que os dados aleatórios sigam um modelo estatístico paramétrico $\mathcal{P}_{\boldsymbol{\theta}} = \{f(\boldsymbol{y}; \boldsymbol{\theta})\}$.
- Isto é, os dados possuem uma distribuição de probabilidade que depende de um parâmetro desconhecido $\boldsymbol{\theta}$.
- Para modelos com um único parâmetro θ, o método pode ser resumido de maneira informal da seguinte maneira:
- Suponha que $y_1, \ldots, y_n$ são os dados da amostra.
- Usando o modelo estatístico $\mathcal{P}_\theta$, calcule o valor aproximado da probabilidade de observar os dados da amostra e obtenha a *função de verossimilhança* $L(\theta)$ onde apenas θ pode variar.
- Obtenha o valor $\hat{\theta}$ que maximiza $L(\theta)$. Este valor é a estimativa de máxima verossimilhança.

Resumo

- Em resumo, o método de máxima verossimilhança encontra o valor $\hat{\theta}$ de θ que é o mais verossímil tendo em vista os dados à mão.
- O valor $\hat{\theta}$ é aquele em que, aproximadamente, a probabilidade de observar os dados realmente observados é máxima.
- Note que NÃO ESTAMOS encontrando o θ mais provável.
- Estamos encontrando o θ tal que seja máxima a probabilidade de gerar OS DADOS que realmente temos em mãos.

Por quê usar o método de máxima verossimilhança?

- generalidade: o método é muito geral e pode ser usado quando a intuição não conseguir sugerir bons estimadores para θ.
- É fácil obter $L(\theta)$ e basta maximizá-la em θ.
- **Fisher:** se a amostra cresce então a estimativa $\hat{\theta}$ converge para θ QUALQUER QUE SEJA O PROBLEMA ESTATÍSTICO.
- **Fisher:** se a amostra cresce então a estimativa $\hat{\theta}$ é aproximadamente não-viciada para θ.
- OBS: Um estimador é não-viciado se as estimativas que fazemos com ele tendem a oscilar em torno do verdadeiro valor desconhecido de θ (veremos isto mais a frente).

Por quê usar o método de máxima verossimilhança?

- outra razão para usar a estimativa de máxima verossimilhança.
- Esta razão també é de **Fisher**, e o resultado é sensacional: qualquer estimador não-viciado ou aproximadamente não-viciado terá um erro médio de estimação maior que o estimador de máxima verossimilhança. E isto é válido para praticamente *qualquer* modelo estatístico.
- **Fisher** de novo: o estimador de máxima verossimilhança possui distribuição aproximadamente normal, não importa quão complicada seja a sua fórmula. Este é um fato fundamental para intervalos de confiança e testes de hipóteses.