Inferência para CS
Tópico 10 - Princípios de Estimação Pontual

Renato Martins Assunção

DCC - UFMG

2013

Critérios para escolher estimadores

- Para escolher bons estimadores, precisamos de uma TEORIA que nos guie.
- Nesta teoria, é FUNDAMENTAL ver os estimadores como variáveis (ou vetores) aleatórias.
- O que é uma variável ou vetor aleatório?
- Duas coisas...

μ e $\overline{Y}$

- Suponha que $Y_1, Y_2, \ldots, Y_n$ sejam i.i.d. $N(\mu, 1)$.
- Queremos estimar μ.
- Usamos $\overline{Y} = (Y_1 + \ldots + Y_n)/n$
- Qual a diferença entre μ e $\overline{Y}$?
- Exemplo

μ e $\overline{Y}$

- Gerei no R cinco v.a.s i.i.d. $N(0,1)$.
- Assim, $\mu = 0$.
- Resultado: `-0.962 -0.293 0.259 -1.152 0.196`
- Tivemos a estimativa $\overline{y} = -0.390$
- Nova simulação: `0.030 0.085 1.117 -1.219 1.267`
- Nova estimativa: $\overline{y} = 0.256$
- μ não muda de valor quando nova amostra é retirada. Temos sempre $\mu = 0$ aqui.
- $\overline{y}$ muda de valor de amostra para amostra. Isto indica que μ e $\overline{y}$ não podem ser as mesmas coisas.

Estimador e estimativa

- Experimento: retirar duas amostras de tamanho 5 de $N(0, 1)$.
- Você vai repetir o experimento com nova semente.
- Qual das duas amostras, a 1a. ou a 2a., será a melhor?
- $\overline{y}$ é a estimativa: a instância específica que se materializa numa amostra, um número.
- $\overline{Y}$ é a VARIÁVEL ALEATÓRIA: duas coisas...
- Por exemplo, no caso de n v.a.'s i.i.d. $N(\mu, \sigma^2)$ temos: $\overline{Y} \sim N(\mu, \sigma^2/n)$

Um caso muito simples

- Suponha que $Y_1, Y_2, \ldots, Y_5$ sejam i.i.d. $N(\mu, 1)$.
- Queremos estimar μ.
- Podemos usar a VARIÁVEL ALEATÓRIA $\overline{Y} = (Y_1 + \ldots + Y_5)/5$
- Algumas vezes (em algumas amostras) teremos a estimativa com $|\overline{y} - \mu| \approx 0$ mas algumas vezes teremos $|\overline{y} - \mu| >> 0$.
- Ou podemos usar a VARIÁVEL ALEATÓRIA mediana:
  - ordene a amostra: $Y_{(1)} = \min\{Y_1, \ldots, Y_5\}$,
  - $Y_{(2)}$ é o segundo menor, etc,
  - $Y_{(5)} = \max\{Y_1, \ldots, Y_5\}$
  - Pegue $M = Y_{(3)}$ como estimador de μ
- Quem é melhor para estimar μ: a VARIÁVEL ALEATÓRIA M ou a VARIÁVEL ALEATÓRIA $\overline{Y}$?

Verifique...

- Com as duas amostras de uma $N(0,1)$ (isto é, $\mu = 0$) tivemos:
  - Primeira amostra: $\overline{y} = -0.390$ e $m = -0.292$
  - Segunda amostra: $\overline{y} = 0.256$ e $m = 0.085$
- Nas duas amostras, a v.a. M esteve mais próxima de μ que $\overline{Y}$.
- Isto talvez seja um indicativo de que M tem um erro de estimação SEMPRE menor que $\overline{Y}$
- FALSO: numa 3a. amostra temos os dados `-0.745 -1.131 -0.716 0.253 0.152` com $\overline{y} = -0.437$ e $m = -0.716$.
- As vezes, teremos $|\overline{y} - \mu| < |m - \mu|$ mas as vezes teremos o contrário.
- O que acontece EM GERAL? Qual o comportamento ESTATÍSTICO das v.a.'s M e $\overline{Y}$?

M e $\overline{Y}$

- Simulando 1000 amostras de tamanho 5.

```
mat = matrix(rnorm(5*1000), ncol=5)
media = apply(mat, 1, mean)
med = apply(mat, 1, median)
aux = range(c(med, media))
plot(media, med, asp=1); abline(0,1)
plot(abs(media), abs(med), asp=1)
sum(abs(media - 0) > abs(med - 0))
```

- Podemos concluir que $\overline{Y}$ é melhor que M sempre? Para todo tamanho de amostra n? Para todo valor de μ? Para todo valor de σ^2? Como conlcuir de forma geral e definitiva?

Estatísticas

- Temos amostra aleatória $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ de v.a.'s
- Distribuição de $\mathbf{Y}$ pertence a uma família (ou modelo) paramétrico $F(\mathbf{y}, \boldsymbol{\theta}) = F(y_1, y_2, \ldots, y_n; \boldsymbol{\theta})$.
- Parâmetro $\boldsymbol{\theta} \in \Theta$ (espaço paramétrico)
- Deseja-se inferir sobre $q(\boldsymbol{\theta})$.
- **Definição:** Uma *estatística* é uma função matemática $g(\mathbf{Y})$ que tenha como argumento $\mathbf{Y}$ e que tome valores em $\mathbb{R}^h$
- Uma estatística não pode envolver os parâmetros desconhecidos $\boldsymbol{\theta}$.
- **Definição:** Um *estimador pontual* de $q(\boldsymbol{\theta})$ será qualquer estatística $\mathbf{T} = g(\mathbf{Y})$.
- A única diferença entre uma estatística e um estimador é que ao definir um estimador precisamos declarar o quê ele está estimando (declarar $q(\boldsymbol{\theta})$).

Exemplo de estimador pontual

- Amostra $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ de v.a.'s iid com $\mathbb{E}(Y_i) = \mu$ e $\text{Var}(Y_i) = \mathbb{E}(Y_i - \mu)^2 = \sigma^2$.
- Seja $\overline{Y} = \frac{1}{n}\sum_i Y_i$, a média aritmética das v.a.'s
- $\overline{Y}$ é um estimador de μ.
- $S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$ é a variância amostral
- S^2 é um estimador para σ^2.

Estimador pontual como vetor

- Amostra $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ de v.a.'s iid com $\mathbb{E}(Y_i) = \mu$ e $\text{Var}(Y_i) = \mathbb{E}(Y_i - \mu)^2 = \sigma^2$.
- Seja que $\boldsymbol{\theta} = (\mu, \sigma^2)$
- É comum usarmos o vetor bi-dimensional

$$\mathbf{T} = g(\mathbf{X}) = \left(\overline{Y}, \frac{1}{n} \sum_{i=1}^{n} (Y_i - \overline{Y})^2 \right)$$

  para fazer inferência sobre $\boldsymbol{\theta}$.
- $\mathbf{T}$ é uma estatística bi-dimensional já que cada entrada de $\mathbf{T}$ é uma função dos dados $\mathbf{Y}$.
- A primeira entrada do vetor $\mathbf{T}$ é a média aritmética $\sum_i Y_i/n$ e ela é usada para inferir sobre o valor desconhecido de μ.
- A segunda entrada é uma medida empírica da variação dos dados em torno de $\overline{Y}_n$ e ela é usada para inferir sobre o valor de $\sigma^2 = \mathbb{E}(Y - \mu)^2$.

Quando algo é um estimador?

- Definição de estimador permite que QUALQUER estatística $g(\mathbf{Y})$ seja estimador de $\boldsymbol{\theta}$.
- Mas então podemos usar $\overline{Y}$ ou $W = \max\{Y_1, \ldots, Y_n\}$ como estimadores de $\sigma^2 = \mathbb{V}(Y_i)$?
- Podemos mas não devemos.
- Vamos ver que $\overline{Y}$ e W tem propriedades muito ruins como estimadores de σ^2.
- Podemos facilmente encontrar estimadores de σ^2, tais como $S^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, que são muito melhores que $\overline{Y}$ ou $W = \max\{Y_1, \ldots, Y_n\}$.

Outros exemplos de estimadores pontuais

- $\mathbf{T} = g(\mathbf{Y}) = (\overline{Y}, (Y_{(n)} - Y_{(1)})/2)$, a média e metade da variação total (range) da amostra
- $\mathbf{T} = g(\mathbf{Y}) = \widehat{F}_n(3.2) = \#\{Y_i; \quad Y_i \leq 3.2\}/n$ onde $\#A$ é o número de elementos (ou cardinalidade) do conjunto A.
- Isto é, $\widehat{F}_n(3.2)$ é a proporção de elementos da amostra que são menores ou iguais a 3.2.
- Poderíamos substituir o ponto $x = 3.2$ por qualquer outro no exemplo acima.

Estimadores não-intuitivos

- Alguns estimadores possuem fórmulas matemáticas complicadas e não-intuitivas.
- Por exemplo, para variáveis aleatórias Y_i positivas (isto é, com $P(Y_i > 0) = 1$) podemos definir a estatística $\mathbf{T} = g(\mathbf{Y}) = \frac{n}{\sum_{i=1}^{n} \log Y_i}$.
- Esta estatística estranha é o MLE de um parâmetro θ em um certo modelo estatístico (v.a.'s i.i.d. com distribuição Pareto ou power-law).

Estimadores não-intuitivos

- Em outro problema, o método de máxima verossimilhança leva ao seguinte estimador:
- $T = T(\mathbf{Y})$ é o valor de λ que satisfaz à seguinte restrição:

$$\frac{\lambda}{1 - e^{-\lambda}} = \bar{Y}$$

- A solução desta equação não-linear deve ser encontrada numericamente.
- A solução é função dos dados através de $\bar{Y}$ no lado direito.

Estimadores não-intuitivos

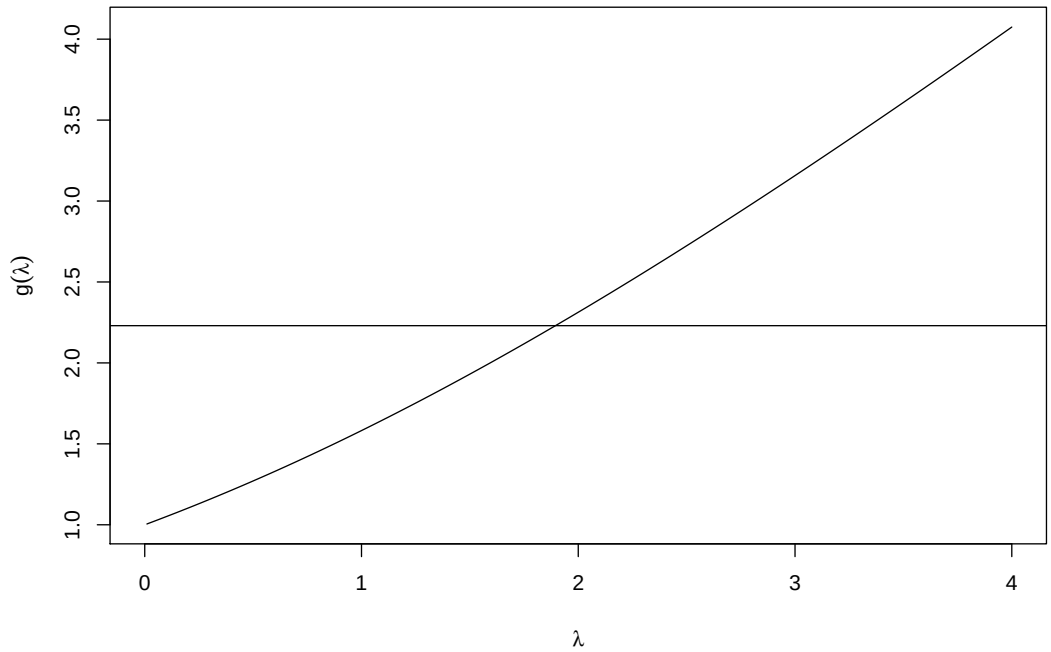

Figura: Gráfico da função $g(\lambda) = \lambda/(1 - e^{-\lambda})$. A linha horizontal corresponde à média aritmética $\bar{y} = 2.23$ dos dados amostrais. O estimador de λ é o valor $\hat{\lambda}$ tal que $g(\lambda) = \bar{y}$. Podemos ver que $\hat{\lambda} \approx 1.95$.

Estimadores pontuais em regressão linear

- No modelo de regressão linear múltipla, temos v.a.'s $Y_1, \ldots, Y_n$ que são independentes mas não são i.d.
- Para a i-ésima observação, temos o modelo:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i = \mathbf{x}_i' \ \boldsymbol{\beta} + \varepsilon_i$$

- $\mathbf{x} = (1, x_{i1}, \ldots, x_{ik})^t$ é o vetor $(k+1) \times 1$ com as covariáveis (ou features) associadas com a observação i.
- Os erros $\varepsilon_1, \ldots, \varepsilon_n$ são i.i.d. seguindo uma gaussiana $N(0, \sigma^2)$.
- Os erros ε_i NÃO SÃO observados diretamente. Observamos apenas Y_i e as covariáveis em $\mathbf{x}_i$.
- O modelo implica que $Y_i \sim N(\mathbf{x}_i' \ \boldsymbol{\beta}, \sigma^2)$ e as v.a.'s são independentes.

Estimadores pontuais em regressão linear

- A versão matricial do modelo de regressão linear é

$$\mathbf{Y} = \mathbf{X}\ \boldsymbol{\beta} + \varepsilon$$

- onde $\mathbf{Y}$ é vetor $n \times 1$, a matriz de desenho $\mathbf{X}$ é de dimensão $n \times (k+1)$ com k covariáveis e a colunas de 1's.
- O vetor $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ de dimensão $n \times 1$ é composto de v.a.'s i.i.d. $N(0, \sigma^2)$.
- O parâmetro $\boldsymbol{\theta}$ é $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2) = (\beta_0, \ldots, \beta_p, \sigma^2)$
- Queremos estimar $\boldsymbol{\beta}$ e σ^2.

MLE de $\boldsymbol{\beta}$

- O MLE $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ coincide com o estimador de mínimos quadrados.
- $\widehat{\boldsymbol{\beta}}$ é uma função do vetor de dados aleatórios $\mathbf{Y}$ e da matriz de regressores $\mathbf{X}$ (que é considerada uma matriz de constantes conhecida).
- Temos o vetor $(k+1) \times 1$

$$\widehat{\boldsymbol{\beta}} = (\widehat{\beta_0}, \ldots, \widehat{\beta_p})' = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y}$$

- Se escrevermos a matriz $k \times n$ dada por $(\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t$ por A podemos ver que $\widehat{\boldsymbol{\beta}} = A\,\mathbf{Y}$.
- Assim, cada elemento de $\widehat{\boldsymbol{\beta}}$ é uma combinação linear dos elementos do vetor $\mathbf{Y}$.

MLE de σ^2

- O modelo prediz ou estima o valor de Y_i usando $\widehat{\beta}$.
- O vetor $n \times 1$ com os valores preditos pelo modelo para os valores realmente observados $\mathbf{Y}$ são dados por

$$\widehat{\mathbf{Y}} = \mathbf{X}\ \beta = \mathbf{X}\ (\mathbf{X}^t\,\mathbf{X})^{-1}\mathbf{X}^t\,\mathbf{Y}$$

- A diferença entre $\mathbf{Y}$ e a predição $\widehat{\mathbf{Y}}$ forma o vetor de resíduos ou vetor de erros de predição $\mathbf{Y} - \widehat{\mathbf{Y}}$.
- O MLE de σ^2 é dado pela média dos resíduos ao quadrado:

$$\widehat{\sigma^2} = \frac{1}{n}||\,\mathbf{Y} - \widehat{\mathbf{Y}}||^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{x}_i'\,\widehat{\beta})^2$$

Estimadores pontuais em regressão logística

- Na regressão logística, onde $p_i = p_i(\theta) = 1/(1 + \exp(-\mathbf{x}_i^t\,\theta))$, o MLE é a solução $\hat{\theta}$ do sistema de equações não-lineares

$$\mathbf{X}\,\mathbf{Y} = \mathbf{X}\,\mathbf{p}(\theta)$$

- onde $\mathbf{p}(\theta) = (p_1(\theta), p_2(\theta), \ldots, p_n(\theta))$
- Embora não exista uma expressão analítica, uma fórmula, para o MLE, podemos ver que a solução vai depender apenas de $\mathbf{X}$ e de $\mathbf{Y}$.
- Assim, como em regressão múltipla, $\hat{\theta}$ é função dos dados aleatórios binários $\mathbf{Y}$ e da matriz de regressores (ou constantes) $\mathbf{X}$ (embora não possamos escrever explicitamente esta função).

$T(\mathbf{Y})$ é v.a.

- Vamos escrever $T(\mathbf{Y})$ ou simplesmente T.
- Conceito CRUCIAL: $T(\mathbf{Y})$ é uma v.a.
- Exemplo: $Y_1, \ldots, Y_n$ i.i.d. $N(\mu, \sigma^2)$.
- Considere $\overline{Y} = (Y_1 + \ldots + Y_n)/n$ é um estimador natural para μ.
- $\overline{Y}$ é uma v.a!!!
- Até sabemos qual é a sua distribuição de probabilidade a partir das propriedades de combinação linear de uma normal multivariada:

$$\overline{Y} = \frac{1}{n} (1, 1, \ldots 1)^t \mathbf{Y} \sim N(\mu, \sigma^2/n)$$

onde $\mathbf{Y} = (Y_1, \ldots, Y_n)$ é normal multivariada com vetor esperado $(\mu, \mu, \ldots, \mu)$ e matriz de covariância $\sigma^2 I_n$ onde I_n é a matriz identidade.

$T(\mathbf{Y})$ e $T(\mathbf{y})$

- Devemos distinguir entre a estatística ou estimador $T(\mathbf{Y})$ (que é uma v.a.) e o seu valor observado num conjunto específico de instâncias (que é um número específico).
- Uma amostra de tamanho 4: $y_1 = 1.6$, $y_2 = 1.8$, $y_3 = 1.5$, $y_4 = 1.8$.
- Estes são os valores observados das v.a.'s Y_1, Y_2, Y_3, Y_4 nesta amostra particular.
- O valor observado da v.a. $\overline{Y}$ é o valor $\overline{y} = 1.675$.
- Note que $\overline{y} = 1.675$ não é uma v.a.: o número 1.675 não possui uma lista de valores possíveis e probabilidades associadas.
- $\overline{y} = 1.675$ é um dos valores possíveis da v.a. $\overline{Y}$, um valor especial: aquele que calhou de ocorrer na amostra que temos à mão.

$T(\mathbf{Y})$ e $T(\mathbf{y})$

- $\mathbf{Y} = (Y_1, \ldots, Y_{11})$ é vetor com 11 variáveis aleatórias i.i.d com distribuição comum $U[0, \theta]$.
- Suponha que o verdadeiro valor de θ é 1 mas *isto é desconhecido pelo usuário*. Deseja-se estimar θ.
- Estimador de θ: $T = 12 \max\{Y_1, \ldots, Y_{11}\}/11 = 12 Y_{(11)}/11$.
- Explicação: $\max\{Y_1, \ldots, Y_{11}\}$ deve estar próximo, mas abaixo, do maior valor possível, que é θ. Uma maneira de obter um estimador de θ seria incrementar um pouco o máximo multiplicando por alguma constante maior que 1.
- A fração 12/11 faz com que T seja ligeiramente maior que o máximo $Y_{(n)}$.
- Veremos que isto torna o estimador T não-viciado para estimar θ (lista de exercícios).

$T(\mathbf{Y})$ e $T(\mathbf{y})$

- Uma amostra particular $\mathbf{y}_1$: obtem $t_1 = g(\mathbf{y}_1) = 1.0437$.
- Com outra amostra particular (um segundo dia) $\mathbf{y}_2$ obtemos $t_2 = g(\mathbf{x}_2) = 0.9845$.
- Repetindo independentemente 1000 vezes.
- Geramos mil vetores $\mathbf{y}_1, \ldots, \mathbf{y}_{1000}$, cada um deles de tamanho $n = 11$.
- Em cada uma destas 1000 amostras, calculamos 1000 valores $t_1, t_2, \ldots, t_{1000}$.
- Isto é, obtemos uma amostra de 1000 valores i.i.d. de T.

$T(\mathbf{Y})$ e $T(\mathbf{y})$

- Com esta amostra de 1000 valores do ESTIMADOR T podemos ter uma idéia da distribuição de probabilidade da v.a. T fazendo um histograma:

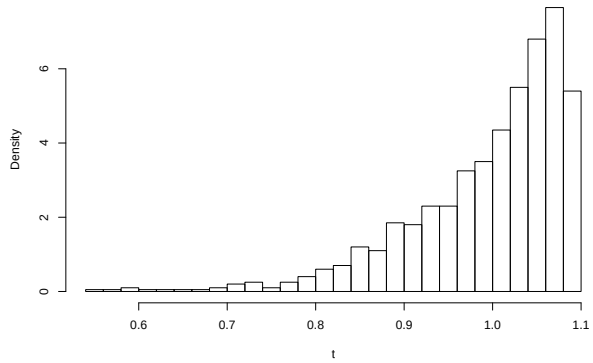

Figura: Histograma dos 1000 valores de $T = 12/11 Y_{(11)}$

- Aproximadamente 1% dos valores observados de T caíram abaixo de 0.70: bad days.

O drama da realidade

- O drama do usuário é que, na prática, ele provavelmente fará apenas uma única estimação, num único dia específico.
- Ele fará isto usando uma única amostra de 11 dados nos quais deve se basear para estimar o valor desconhecido de θ.
- Por isto, ele nunca saberá, nesse dia da estimação, qual o tamanho do erro de estimação que ele está cometendo.

Candidatos a estimar θ

- Seja $\mathbb{E}(Y) = \mu$ e uma amostra $Y_1, \ldots, Y_n$
- Por que não considerar o estimador mediana amostral? Ordene os dados: $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. Então

$$M = \begin{cases} X_{(k+1)}, & \text{caso } n \text{ seja ímpar, } n = 2k+1 \\ (X_{(k)} + X_{(k+1)})/2, & \text{caso } n \text{ seja par, } n = 2k \end{cases}$$

- Ou que sabe a média do primeiro e do terceiro quartil?
- Como escolher um deles? qual o critério de escolha?

Mais candidatos a estimar θ

- Ou então uma média ponderada entre a média aritmética $\overline{X}_n$ e a mediana:

$$T = wM + (1 - w)\overline{X}_n$$

  onde $0 \leq w \leq 1$.
- Como w varia continuamente entre 0 e 1, teremos infinitos possíveis estimadores deste tipo.
- Ou quem sabe uma média ponderada entre $\overline{X}_n$, a mediana e a média dos quartis?
- Como escolher um deles? qual o critério de escolha?
- Obviamente algum critério que faça referência ao tamanho dos erros de estimação, mas qual é esse critério?

Erro de estimação

- O erro de estimação que vamos cometer é a *variável aleatória* $T(\mathbf{Y}) - \theta$.
- Como v.a., ela possui duas coisas: uma "lista" de valores possíveis e uma "lista" de probabilidades associadas.
- Na prática, como não conhecemos o valor de θ, nunca saberemos o valor do erro de estimação que cometemos em cada caso particular.
- Apesar disso, podemos conhecer as propriedades estatísticas do erro de estimação.
- Podemos saber como erraremos em geral, embora não possamos saber como erramos em cada caso particular.

Vício

- $T(\mathbf{Y})$ é um estimador vetorial de dimensão k de uma característica vetorial $\boldsymbol{\theta}$ de dimensão k da população.
- Definição: A diferença vetorial $\mathbb{E}(T(\mathbf{Y})) - \boldsymbol{\theta}$ é chamada de vício do estimador (para estimar $\boldsymbol{\theta}$) e é denotada por $b(\boldsymbol{\theta})$.
- Quando $b(\boldsymbol{\theta}) = \mathbf{0}$, dizemos que o estimador é não viciado para estimar o parâmetro $\boldsymbol{\theta}$.
- Um estimador não-viciado tem sua distribuição centrada em torno do parâmetro $\boldsymbol{\theta}$ que desejamos estimar.
- Ocasionalmente ele vai subestimar ou superestimar θ.
- Mas ele nem subestima nem superestima *sistematicamente*.
- Um estimador não-viciado é um estimador *acurado*.

Vício:exemplos

- Seja $Y_1, \ldots, Y_n$ uma amostra de variáveis i.i.d. com *qualquer* distribuição.
- Suponha que $\mathbb{E}(Y_i) = \mu$.
- Então $\overline{Y}_n$ é um estimador não-viciado para estimar μ.
- Se os n dados tiverem distribuição i.i.d. normal, a mediana amostral M é não-viciada para estimar μ se n é ímpar e é ligeiramente viciada se n é par.
- Se $w \in (0, 1)$, então $w\overline{Y} + (1 - w)M$ é não-viciado para estimar μ

Vício:exemplos

- Seja $Y_1, \ldots, Y_n$ uma amostra de variáveis i.i.d. com *qualquer* distribuição.
- Suponha que $\text{Var}(Y_i) = \sigma^2$.
- Então S^2 é um estimador não-viciado para estimar σ^2 onde

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}_n\right)^2$$

- A prova é simples e fica para a lista de exercícios.
- Esta é a razão para ter o denominador $n-1$ no estimador acima: torná-lo não-viciado para estimar σ^2

Vício

- O estimador

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2$$

  é viciado
- Veja que $\widehat{\sigma}^2 = (n-1)/nS^2$
- O vício de $\widehat{\sigma}^2$ é dado por

$$b(\sigma^2) = \mathbb{E}(\widehat{\sigma}^2) - \sigma^2 = \left(\frac{n-1}{n} - 1\right)\sigma^2 = -\frac{\sigma^2}{n}$$

- Vício $b(\sigma^2)$ tende a zero quando a amostra cresce.

Vício: exemplo

- Tempo de vida $T_1, \ldots, T_n$ são iid com distribuição exp(λ).
- Observar tempos de vida até um tempo máximo C.
- Quando $T_i > C$ anotamos simplesmente C.
- O verdadeiro valor de T_i nestes casos não é observado
- Dizemos que eles são censurados.
- Estimar $\mathbb{E}(T_i) = \mu$ usando a média aritmética dos T_i's **registrados** é um estimador viciado (subestima μ).

Vício: exemplo

- $\mathbf{Y} = (Y_1, \ldots, Y_n)$ são n variáveis aleatórias i.i.d com distribuição comum $U[0, \theta]$.
- Suponha que o verdadeiro valor de θ é desconhecido e que desejamos estimar θ.
- Estimador: $T = ((n+1)/n) \times \max\{Y_1, \ldots, Y_n\}$.
- T é não-viciado para estimar θ pois $\mathbb{E}(T) = \theta$.
- A constante $(n+1)/n$ é a quantidade perfeita para incrementar o máximo $\max\{Y_1, \ldots, Y_n\}$ e "empurrá-lo" para ficar em torno de θ.

MSE

- $T = T(\mathbf{Y})$ é estimador de θ
- *Definição:* $MSE = \mathbb{E}(T - \theta)^2$ é chamado de erro quadrático médio de estimação (*mean squared error*, em inglês).
- *Definição:* $\mathbb{E}(|T - \theta|)$ é chamado de erro absoluto médio de estimação.
- O erro absoluto é mais intuitivo mas temos mais resultados usando o MSE.

Decomposição do MSE

- *Teorema:* Se $T = T(\mathbf{Y})$ é estimador de θ então

$$\mathbb{E}(T-\theta)^2 = \text{Var}(T) + b(\theta)^2 \tag{1}$$

- *Prova:* Seja $\mathbb{E}(T) = \mu$.
- Some e subtraia μ:

$$\begin{aligned}
\mathbb{E}(T-\theta)^2 &= \mathbb{E}(T-\mu+\mu-\theta)^2 \\
&= \mathbb{E}\left[(T-\mu)^2 + 2(T-\mu)(\mu-\theta) + (\mu-\theta)^2\right]^2 \\
&= \mathbb{E}\left[(T-\mu)^2\right] + 2\mathbb{E}\left[(T-\mu)(\mu-\theta)\right] + \mathbb{E}\left[(\mu-\theta)^2\right] \\
&= \text{Var}(T) + 2(\mu-\theta)\mathbb{E}\left(T-\mu\right) + (\mu-\theta)^2
\end{aligned}$$

- Para terminar: $\mathbb{E}\left(T-\mu\right) = 0$ pois $\mathbb{E}(T) = \mu$ e assim

$$\mathbb{E}(T-\theta)^2 = \text{Var}(T) + (\mu-\theta)^2 = \text{Var}(T) + b^2(\theta)\,.$$

Decomposição do MSE

- Esta decomposição é de grande utilidade por duas razões.
- 1) ela permite quebrar o problema de encontrar um bom estimador (um com erro de estimação tipicamente pequeno) em dois sub-problemas.
- Devemos encontrar um estimador que possua um vício pequeno e, ao mesmo tempo, variância pequena.
- 2) Por outro lado, ele fornece um critério simples para encontrar bons estimadores.
- Se o vício de dois estimadores é zero, o *MSE* deles é igual à suas variâncias e assim basta escolher aquele estimador que possui a menor variância.
- Este estimador não-viciado e de menor variância terá o menor *MSE*.

Ilustração

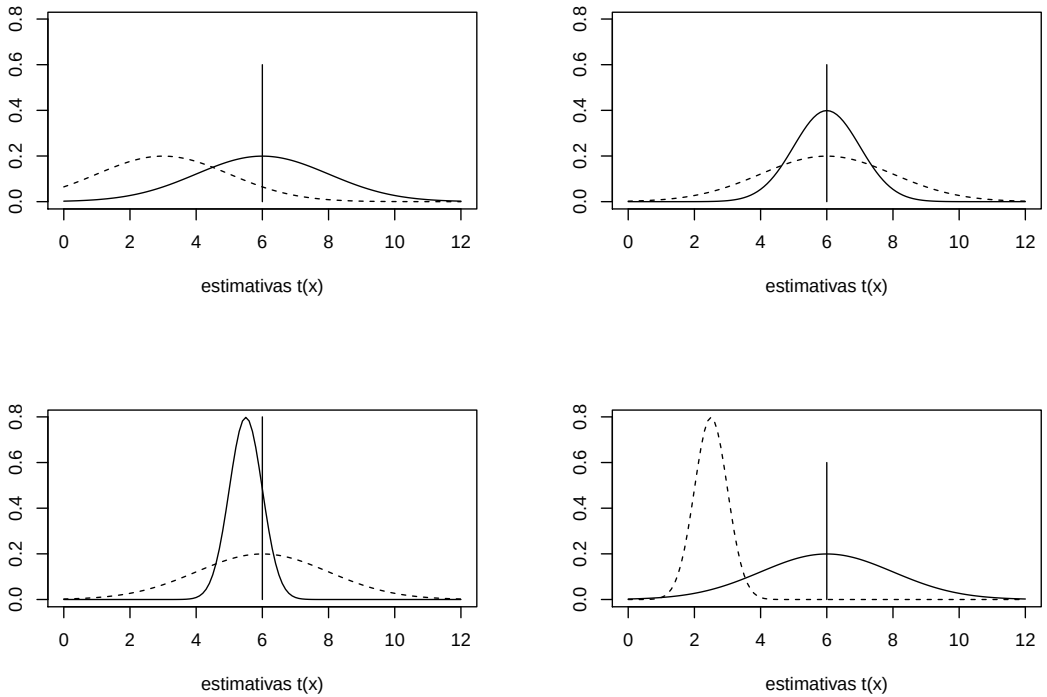

Figura: Densidades de dois estimadores T_1 e T_2 de um mesmo parâmetro θ. Nestes exemplos, o valor de θ é 6.