3 Estimação por Intervalos

3.1 Introdução

Situamo-nos no âmbito da inferência paramétrica:

- População X tem distribuição na família $\mathcal{F} = \{f(\cdot \mid \theta) : \theta \in \Theta\}$
- Espaço do parâmetro $\Theta \subset \mathbb{R}^k$
- Define-se processo de amostragem; em geral $(X_1, \ldots, X_n)$ amostra casual de tamanho n proveniente de X
- Espaço amostral $\mathcal{X}$

Problema: na situação em que $\Theta \subset \mathbb{R}$,

- estimar θ por um intervalo, ou seja, escolher uma aplicação $\mathbf{x} \in \mathcal{X} \mapsto (T_1(\mathbf{x}), T_2(\mathbf{x}))$, com $T_1 < T_2$, que a cada amostra observada faz corresponder um intervalo. Esse intervalo deve conter θ com um determinado *grau de confiança*
- Podemos por vezes estar interessados em estimar por um intervalo uma função de θ, $\tau(\theta)$

Intervalo aleatório e intervalo de confiança

Definição 3.1 Intervalo aleatório para θ: *Se $T_1(X_1, \ldots, X_n)$ e $T_2(X_1, \ldots, X_n)$, com $T_1 < T_2$, são duas estatísticas verificando*

$$P(T_1 < \theta < T_2) = 1 - \alpha \ , \quad \forall \theta \in \Theta$$

para algum $\alpha \in (0, 1)$ (α não depende de θ), então dizemos que (T_1, T_2) é um intervalo aleatório para θ de probabilidade $1 - \alpha$. ■

Definição 3.2 Intervalo de confiança para θ: *Seja (T_1, T_2) um intervalo aleatório para θ de probabilidade $1 - \alpha$, e seja $(x_1, \ldots, x_n)$ uma amostra observada, $t_1 = T_1(x_1, \ldots, x_n)$ e $t_2 = T_2(x_1, \ldots, x_n)$. Então, ao intervalo (t_1, t_2) dá-se o nome de intervalo de confiança a $(1 - \alpha) \times 100\%$ para θ.* ■

Observações:

- um intervalo de confiança é a concretização de um intervalo aleatório, tal como uma estimativa é uma concretização de um estimador
- $P(T_1 < \theta < T_2) = 1 - \alpha$, mas $P(t_1 < \theta < t_2) = I_{(t_1,t_2)}(\theta)$, ou seja, ou $\theta \in (t_1, t_2)$, ou $\theta \notin (t_1, t_2)$; mas como θ é desconhecido não sabemos em que situação nos encontramos
- Interpretação frequencista da probabilidade: se recolhermos um número muito grande de amostras casuais de dimensão n, e para cada uma delas calcularmos o valor observado do intervalo (T_1, T_2), então em aproximadamente $(1 - \alpha) \times 100\%$ dos casos o intervalo observado incluirá o verdadeiro valor de θ
- É este o significado de *confiança*: temos uma confiança de $(1 - \alpha) \times 100\%$ que o intervalo que produzimos esteja entre os que efectivamente contêm θ!

Exemplo 3.1 *Seja $X_1, \ldots, X_n$ uma amostra casual de tamanho $n = 10$ proveniente de uma população $N(\mu, 1)$. É fácil de verificar que*

$$(T_1, T_2) = (\bar{X} - 1.96/\sqrt{10}, \bar{X} + 1.96/\sqrt{10})$$

é um intervalo aleatório para μ de probabilidade 0.95: $P(T_1 < \mu < T_2) = 0.95$.

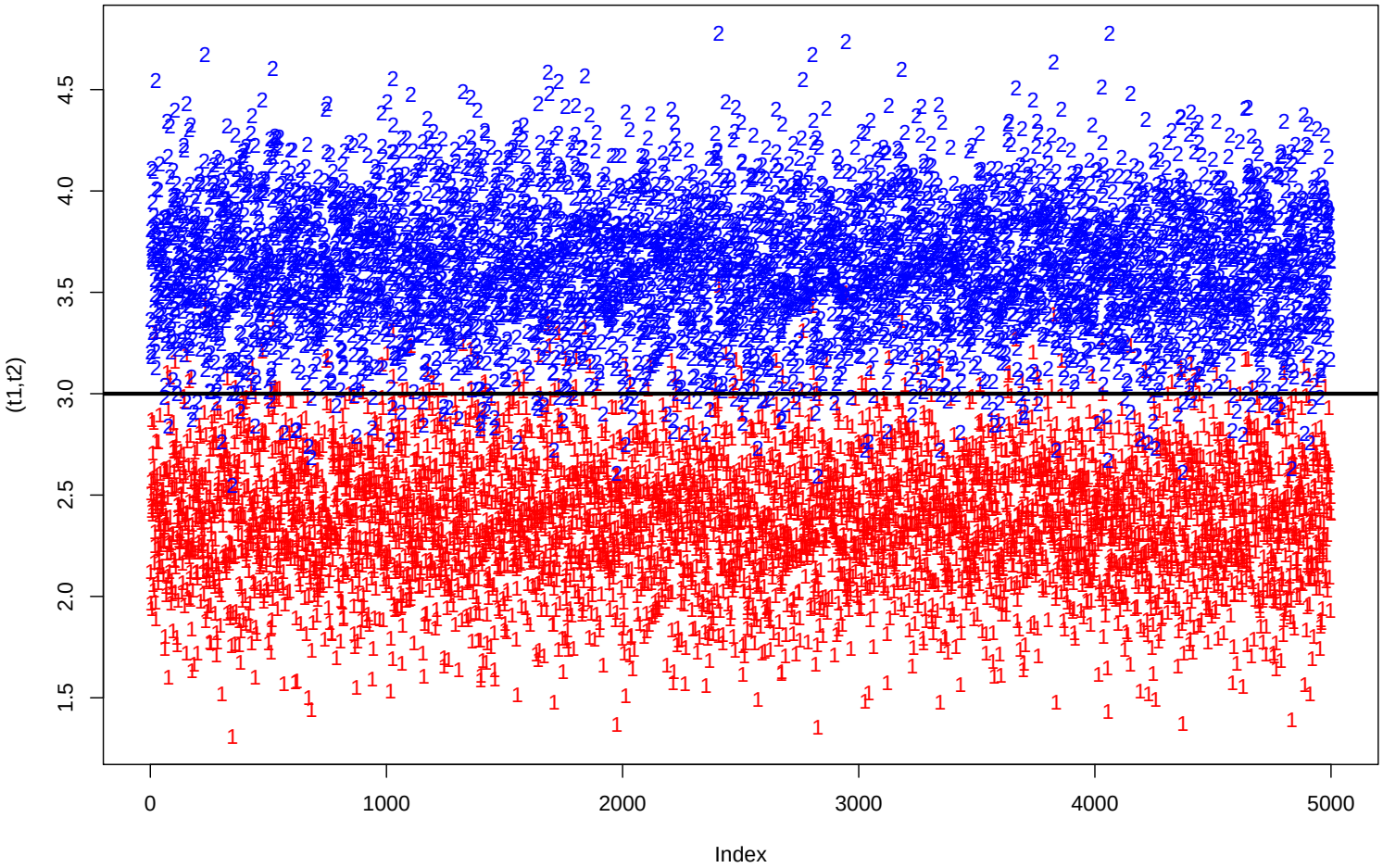
Suponhamos que $\mu = 3$. Simulamos $M = 5000$ amostras de dimensão 10 provenientes de uma população $N(3, 1)$, e para cada uma delas calculamos o valor observado de (T_1, T_2). Qual a proporção de intervalos que contêm o número 3?

```
n <- 10
M <- 5000
data <- matrix(rnorm(n*M,mean=3),ncol=n)
datameans <- apply(data,1,mean)

t1 <- datameans-1.96/sqrt(n)
t2 <- datameans+1.96/sqrt(n)

plot(t1,ylim=c(min(t1),max(t2)),pch="1",col=2,ylab="(t1,t2)")
points(t2,pch="2",col=4)
abline(h=3,lwd=3)

sum((t1<3)*(t2>3))/M
```


(t1,t2)
4.5
4.0
3.5
3.0
2.5
2.0
1.5
0
1000
2000
3000
4000
5000
Index

3.2 Método da variável fulcral

Definição 3.3 Variável fulcral *Seja $Z(X_1, \ldots, X_n, \theta)$ uma função da amostra casual e de θ. Diz-se que Z é uma variável fulcral para θ se a sua distribuição de probabilidade não depender de parâmetros desconhecidos.* ■

Exemplo 3.2 *Se $X_1, \ldots, X_n$ é uma amostra casual proveniente de uma população $N(\mu, 1)$, então*

$$Z = \sqrt{n}(\bar{X} - \mu)$$

é uma variável fulcral para μ já que a sua distribuição de probabilidades é $N(0, 1)$ qualquer que seja μ. ■

Observações:

- Uma variável fulcral não é uma estatística já que depende do parâmetro desconhecido θ; não pode depender de mais nenhum parâmetro desconhecido
- Designações alternativas: pivot ou quantidade pivotal para θ

Método:

1. Encontrar variável fulcral adequada ao problema
2. Fixar grau de confiança desejado $1-\alpha$
3. Encontrar dois números no suporte de Z, $z_1(\alpha)$ e $z_2(\alpha)$ verificando

$$P(z_1(\alpha) < Z < z_2(\alpha)) = 1-\alpha \quad \forall\theta \in \Theta$$

   Existem na maior parte dos casos uma infinidade de escolhas possíveis para z_1, z_2 — idealmente tenta-se encontrar um par que minimize a amplitude (esperada) do intervalo resultante. A escolha verificando $P(Z < z_1(\alpha)) = P(Z > z_2(\alpha)) = \alpha/2$ resulta num intervalo aleatório denominado de central.

4. Manipular a desigualdade $z_1(\alpha) < Z < z_2(\alpha)$ de forma a obter a desigualdade equivalente $T_1(X_1, \ldots, X_n) < \theta < T_2(X_1, \ldots, X_n)$
5. Por construção, $(T_1(X_1, \ldots, X_n), T_2(X_1, \ldots, X_n))$ é um intervalo aleatório de probabilidade $1-\alpha$ para θ
6. Um intervalo de confiança a $(1-\alpha)\times 100\%$ para θ é

$$(T_1(x_1, \ldots, x_n), T_2(x_1, \ldots, x_n))$$

Exemplo 3.3 *$X_1, \ldots, X_n$ amostra de dimensão n proveniente de população $N(\mu, 1)$. Obter intervalo de confiança a $(1-\alpha)\times 100\%$ para μ*

1. Variável fulcral:
$$Z = \sqrt{n}\,\frac{\bar{X}-\mu}{\sigma} = \sqrt{n}\,(\bar{X}-\mu) \sim N(0,1)$$
2. Escolhendo $z_2(\alpha) = -z_1(\alpha)$, tem-se $z_2 = \Phi^{-1}(1-\alpha/2) \equiv z_{\alpha/2}$
3. $$-z_{\alpha/2} < \sqrt{n}\,(\bar{X}-\mu) < z_{\alpha/2} \Leftrightarrow \bar{X} - z_{\alpha/2}/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2}/\sqrt{n}$$
4. Por construção,
$$(\bar{X} - z_{\alpha/2}/\sqrt{n}, \bar{X} + z_{\alpha/2}/\sqrt{n})$$
é um intervalo aleatório para μ de probabilidade $1-\alpha$
5. Para $n = 16$ e $\alpha = 0.05$, vem que $(\bar{X} - 0.49, \bar{X} + 0.49)$ é um intervalo aleatório de probabilidade 0.95 para μ, sendo o correspondente intervalo de confiança dado por $(\bar{x} - 0.49, \bar{x} + 0.49)$

3.3 Aplicação a populações normais

Seja $X_1, \ldots, X_n$ amostra de dimensão n proveniente de população $\mathrm{N}(\mu, \sigma^2)$. Obtenção de intervalo de confianca a $(1-\alpha) \times 100\%$...

3.3.1 para μ com σ^2 conhecido

- Variável fulcral
$$Z = \sqrt{n}\,\frac{\bar{X} - \mu}{\sigma} \sim N(0,1)$$
- $z_1 < Z < z_2 \Leftrightarrow \bar{X} - z_2\sigma/\sqrt{n} < \mu < \bar{X} - z_1\sigma/\sqrt{n}$
- a amplitude do intervalo aleatório é $(z_2 - z_1)\sigma/\sqrt{n}$ que não é aleatória. A simetria e unimodalidade da normal permitem concluir que esta amplitude é minimizada quando $z_2 = -z_1$.

3.3.2 para μ com σ^2 desconhecido

- Variável fulcral

$$Z = \sqrt{n}\,\frac{\bar{X} - \mu}{S'} \sim t_{(n-1)}$$

- $t_1 < Z < t_2 \Leftrightarrow \bar{X} - t_2 S'/\sqrt{n} < \mu < \bar{X} - t_1 S'/\sqrt{n}$
- a amplitude do intervalo aleatório é $(t_2 - t_1)S'/\sqrt{n}$ que é aleatória. Minimiza-se então a amplitude esperada, que é dada por $(t_2 - t_1)c(n)\sigma$. A simetria e unimodalidade da $t_{(n-1)}$ permitem concluir que esta amplitude é minimizada quando $t_2 = -t_1$.

3.3.3 para σ^2

- Variável fulcral

$$Z = \frac{nS^2}{\sigma^2} \sim \chi^2(n-1)$$

- $q_1 < Z < q_2 \Leftrightarrow \frac{nS^2}{q_2} < \sigma^2 < \frac{nS^2}{q_1}$
- a amplitude do intervalo aleatório é $(t_1 - t_2)S'/\sqrt{n}$ que é aleatória. Minimizar a amplitude esperada, que é dada por $(n-1)\sigma^2/(1/q_1 - 1/q_2)$, equivale a resolver o problema $\min(1/q_1 - 1/q_2)$ sujeito a $\int_{q_1}^{q_2} f_{\chi^2(n-1)}(x)dx = 1 - \alpha$. Opta-se então pelo intervalo central: $P(Z > q_2) = P(Z < q_1) = \alpha/2$.

Observação: Procede-se similarmente para obter intervalos de confiança para a diferença de médias de duas populações normais e para o quociente entre as variâncias de duas populações normais.

3.4 Caso de grandes amostras

3.4.1 Populações Bernoulli

- Variável fulcral

$$Z = \sqrt{n}\frac{\bar{X} - \theta}{\sqrt{\theta(1-\theta)}} \overset{a}{\sim} \mathrm{N}(0,1)$$

- Se $z_{\alpha/2} : P(Z > z_{\alpha/2}) = \alpha/2$, então

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = P\left(\bar{X} - z_{\alpha/2}\sqrt{\theta(1-\theta)}/\sqrt{n} < \theta < \bar{X} + z_{\alpha/2}\sqrt{\theta(1-\theta)}/\sqrt{n}\right) \approx 1 - \alpha$$

- O intervalo acima não é um intervalo aleatório para θ porque os extremos do intervalo involvem θ. Para resolver este problema existem duas possibilidades:
  - Substituir θ por $\bar{X}$ no denominador de Z. A distribuição aproximada mantem-se, dado que $\bar{X}$ é um estimador consistente de θ. O intervalo $1 - \alpha$ resultante é

$$\left(\bar{X} - z_{\alpha/2}\sqrt{\bar{X}(1-\bar{X})}/\sqrt{n}, \bar{X} + z_{\alpha/2}\sqrt{\bar{X}(1-\bar{X})}/\sqrt{n}\right)$$

- Alternativamente, podemos resolver a dupla desigualdade explicitamente em ordem a θ:

$$|Z| < z_{\alpha/2} \Leftrightarrow (n + z^2_{\alpha/2})\theta^2 - (2n\bar{X} + z^2_{\alpha/2})\theta + n\bar{X}^2 < 0$$

Calculados os zeros deste polinómio, o correspondente intervalo vem

$$\frac{(2n\bar{X} + z^2_{\alpha/2}) \pm z_{\alpha/2}\sqrt{z^2_{\alpha/2} + 4n\bar{X}(1 - \bar{X})}}{2(n + z^2_{\alpha/2})}$$

mas, como n é grande, tem-se que $n + z^2_{\alpha/2} \approx n$, recuperando-se então o intervalo anterior.

- Esta estratégia aplica-se também ao problema de construir um intervalo aleatório para a diferença de duas proporções, se bem que apenas a segunda solução seja prática
- Variável fulcral

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\theta_1 - \theta_2)}{\sqrt{\frac{\bar{X}_1(1-\bar{X}_1)}{m} + \frac{\bar{X}_2(1-\bar{X}_2)}{n}}} \overset{a}{\sim} \mathrm{N}(0,1)$$

  Note-se que θ_1 e θ_2 foram substituídos no denominador pelos seus estimadores consistentes $\bar{X}_1$ e $\bar{X}_2$, respectivamente
- O intervalo aleatório resultante é

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\bar{X}_1(1-\bar{X}_1)}{m} + \frac{\bar{X}_2(1-\bar{X}_2)}{n}}$$

3.4.2 População Poisson

- A estratégia anterior aplica-se em princípio a qualquer população com variância finita, dado que resulta da utilização do Teorema do Limite Central
- No caso de uma população $\mathrm{Po}(\lambda)$, a primeira alternativa origina

$$\left(\bar{X} - z_{\alpha/2}\sqrt{\bar{X}}/\sqrt{n}, \bar{X} + z_{\alpha/2}\sqrt{\bar{X}}/\sqrt{n}\right)$$

enquanto a segunda produz

$$\frac{(2n\bar{X} + z^2_{\alpha/2}) \pm z_{\alpha/2}\sqrt{-z^2_{\alpha/2} + 4n\bar{X}}}{2n}$$

que, dado que n é grande, estará muito próximo do intervalo obtido seguindo a primeira alternativa

3.4.3 Estimador de Máxima Verosimilhança

- A distribuição assintótica do estimador de máxima verosimilhança válida em condições bastante gerais,

$$\sqrt{\mathcal{I}_{(X_1,\ldots,X_n)}(\theta)}(\hat{\theta}-\theta) \overset{a}{\sim} \mathrm{N}(0,1)$$

  pode em alguns casos ser manipulada para produzir intervalos aleatórios (aproximados) para θ, muitas vezes após $\mathcal{I}_{(X_1,\ldots,X_n)}(\theta)$ ser substituído pelo seu estimador consistente $\mathcal{I}_{(X_1,\ldots,X_n)}(\hat{\theta})$

Exemplo 3.4 *Considere-se uma população com função densidade dada por* $f(x \mid \theta) = \theta x^{\theta-1}$, $0 < x < 1$, $\theta > 0$. *É fácil determinar que* $\hat{\theta} = -n/\sum \ln X_i$ *é o estimador de máxima verosimilhança de* θ, *e que* $\mathcal{I}_X(\theta) = \theta^{-2}$. *Assim, segue-se que*

$$Z = \sqrt{n\theta^{-2}}\left(-n/\sum \ln X_i - \theta\right) = -\sqrt{n}\left(\frac{n}{\sum \ln X_i}\frac{1}{\theta}+1\right) \overset{a}{\sim} N(0,1)$$

facto que permite determinar intervalos aleatórios (aproximados) para θ. ■