


Bioinformática

Métodos de Estimação

Manuela Neves

ISA/ULisboa

4 e 6 de Maio 2020

- 1 Estimação: métodos clássicos
- 2 Propriedades dos estimadores
- 3 Métodos de Estimação
 - O método dos momentos
 - O Método da Máxima Verosimilhança
 - Propriedades dos estimadores de Máxima Verosimilhança
 - A máxima verosimilhança no 

Estimação: métodos clássicos

No **Bloco 1** das aulas de Métodos Estatísticos debruçamo-nos sobre Testes de Hipóteses.

Os testes têm como objectivo “validar” uma suposição formulada sobre um ou mais **parâmetros desconhecidos** da população. Para isso, o papel fulcral reside na construção de **uma variável aleatória** que, sob a hipótese nula (i.e. uma vez concretizado(s) o(s) valor(es) dos parâmetros) tem, pelo menos aproximadamente uma distribuição conhecida - estas hipóteses tinham a designação de **hipóteses paramétricas**.

Aquela variável tinha o nome de **Estatística do Teste**

Falámos de testes especiais: os testes do qui-quadrado em tabelas de contingência.

Antes, porém, de falarmos em Testes de Hipóteses ou Intervalos de Confiança (que não tratámos aqui), deve falar-se em procedimentos que utilizam a amostra para dela retirarem da “melhor” forma a informação necessária para estimar quantidades desconhecidas.

Vamos, por isso, discutir critérios usuais que permitem dizer quando é que um estimador (cuja definição vamos relembrar já a seguir) de um parâmetro desconhecido se pode considerar “bom” – **seguiremos o capítulo 8 de Ewens & Grant**

Introdução:

A Teoria da Estimação utiliza **procedimentos** sobre uma amostra para obter informação sobre:

- o valor desconhecido de um parâmetro, que designamos genericamente por θ , usando o conhecimento, mesmo aproximado, da distribuição da população (**estimação paramétrica**);
- a função distribuição desconhecida da variável em estudo, F , ou parâmetros, sem o pressuposto do conhecimento de um modelo para a população subjacente (**estimação não paramétrica**).

Estimação pontual: estimador e estimativa

O problema da **estimação paramétrica** consiste em utilizar a informação dada pela amostra para “adivinhar” o valor desconhecido de um parâmetro, i.e., encontrar um número que possa estar próximo do parâmetro.

Seja dada uma amostra aleatória (X_1, X_2, \dots, X_n)

Definição

Chama-se **estimador** de θ a uma função da amostra aleatória que não envolve parâmetros desconhecidos.

Quando essa **amostra é observada** podemos então fazer-lhe corresponder um valor que estima θ , a que chamamos uma **estimativa de θ** .

Sendo θ um parâmetro desconhecido

Vamos representar

$$\begin{aligned}\Theta^* &\equiv \Theta^*(X_1, X_2, \dots, X_n) && \text{um estimador de } \theta \\ \theta^* &\equiv \theta^*(x_1, x_2, \dots, x_n) && \text{uma estimativa } \theta\end{aligned}$$

Não confundir

- estimador – variável aleatória
- estimativa – valor aproximado do parâmetro, obtido pelo estimador usando uma amostra particular

Exemplo: São estimadores

$$\text{i) } \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{e} \quad \text{ii) } S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$

Observada, por exemplo, a amostra **(1, 2, 0, 3, 1, 5)**

As estimativas associadas àqueles estimadores são:

$$\text{i) } \bar{x} = \frac{1}{6} \sum_{i=1}^6 x_i = 2 \quad \text{e} \quad \text{ii) } s^2 = \frac{1}{5} \sum_{i=1}^6 (x_i - \bar{x})^2 = 3.2$$

Como se podem definir vários estimadores de um parâmetro, põe-se o problema de escolher, se possível “**o melhor**”.

Há então que considerar certas **propriedades que um estimador deve verificar**.

Vejamos algumas:

1. Consistência


Um estimador Θ^* diz-se convergente ou consistente para o parâmetro θ se $\Theta^* \xrightarrow{P} \theta$.

Prova-se que é condição suficiente de convergência de um estimador Θ^* para θ que

$$\text{Var}(\Theta^*) \rightarrow 0 \quad \text{e} \quad E(\Theta^*) \rightarrow \theta$$

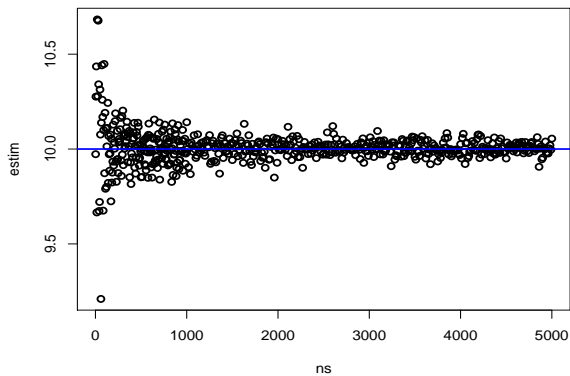
Vejamos o seguinte exemplo ilustrando a convergência de um estimador.

Exemplo

Consideremos uma população $\mathcal{N}(10, 2)$ e vejamos, utilizando o , o comportamento da **média amostral** para várias dimensões de amostras, $n = 2, 5, 10, 15, 20, \dots, 1000, 1010, 1020, \dots, 5000$

```
> ns <- c(2, seq(5, 1000, by = 5), seq(1010, 5000, by = 10))
> estim <- numeric(length(ns)) # tb podia fazer-se estim<-c()
> for (i in 1:length(ns)) {
+ amostra <- rnorm(ns[i], 10, 2)
+ estim[i] <- mean(amostra)
+ }
> plot(ns, estim, lwd=2)
> abline(h = 10, lwd=2, col=4)
```

Ilustrando a convergência



2. Não enviesamento

O estimador Θ^* diz-se centrado ou não enviesado se $E(\Theta^*) = \theta$.

Na verdade é uma propriedade que muitos estimadores não possuem e daí ter mais interesse considerar a **diferença entre o valor esperado do estimador e o parâmetro** a estimar.

É o chamado **viés** (ou em inglês '**bias**') que representaremos por $b_{\theta}(\Theta^*)$ e se define como

$$b_{\theta}(\Theta^*) = E(\Theta^*) - \theta$$

3. Erro quadrático médio

Chama-se **erro quadrático médio** do estimador Θ^* a **$EQM(\Theta^*) = E[(\Theta^* - \theta)^2]$** .

Esta propriedade é um dos critérios mais usados para comparar estimadores.

É muito fácil mostrar que

$$EQM(\Theta^*) = Var[\Theta^*] + [b_\theta(\Theta^*)]^2$$

Se Θ^* é um **estimador centrado** de um parâmetro então o **erro quadrático médio** \equiv **variância**, i.e.

$$E[(\Theta^* - \theta)^2] = Var(\Theta^*)$$

Ilustração da precisão e da exactidão

O termo **exactidão (accuracy)** refere-se à proximidade de uma medição ou estimativa ao verdadeiro valor.

O termo **precisão ou variância (precision)** refere-se ao “grau de concordância numa sucessão de medições”.



Accuracy & Precision



Accurate
but , not precise



Precise
but , not accurate



Accurate
and **Precise**

Até aqui falámos em **estimadores** e nas propriedades que devem possuir. Interessa ter procedimentos que construam estimadores com boas propriedades.

Vamos então falar dos **principais métodos de estimação paramétrica**.

Dos **métodos de estimação paramétrica** vamos referir:

- Método dos momentos e
- Método da Máxima verosimilhança

O método dos momentos

Introduzido por Karl Pearson no início do século XX, foi o primeiro método de estimação a ser apresentado e que tem uma filosofia muito simples.

O método consiste em:

– considerar como estimadores dos parâmetros desconhecidos as soluções das equações que se obtêm igualando os [momentos teóricos](#) aos [momentos empíricos](#).

É um método de aplicação geral, tendo como única condição que a distribuição tenha um número suficiente de momentos (teóricos).

O método dos momentos

Sejam $\theta_1, \dots, \theta_l$ parâmetros desconhecidos de uma v.a. X .
O método dos momentos consiste em igualar momentos teóricos e momentos empíricos.

Dada uma v.a. X , chama-se **momento (ordinal) teórico de ordem k** e representa-se por $E[X^k] = \mu'_k$ a:

$$E[X^k] = \sum_i x_i^k p_i \quad X \text{ é v.a. discreta com distribuição } (x_i, p_i)$$

$$E[X^k] = \int_{-\infty}^{+\infty} x^k f(x) dx \quad X \text{ é v.a. contínua com f.d.p. } f(x)$$

Nota: Os momentos só existem se “aquela soma existir” ou o integral for absolutamente convergente.

O método dos momentos

Vamos recordar que

$E[X] = \mu'_1$ é o conhecido valor médio de X , μ
 $E[X^2] = \mu'_2$ surgiu na fórmula de cálculo da variância, i.e.,

$$\sigma^2 = E[X^2] - \mu^2$$

Dada uma amostra aleatória (X_1, X_2, \dots, X_n) , retirada de uma população X , chama-se **momento (ordinal) empírico de ordem k** e representa-se por M'_k a

$$M'_k = \sum_i^n \frac{X_i^k}{n}$$

O método dos momentos

O método dos momentos consiste em considerar o seguinte sistema de equações

$$\begin{array}{lll} E[X] = M'_1 & \text{onde, como se sabe} & M'_1 = \frac{1}{n} \sum_{i=1}^n X_i \\ E[X^2] = M'_2 & \text{onde, como se sabe} & M'_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \\ \vdots & & \\ E[X^k] = M'_k & \text{onde, como se sabe} & M'_k = \frac{1}{n} \sum_{i=1}^n X_i^k \end{array}$$

Quando temos a amostra observada, (x_1, \dots, x_n) , temos

$$m'_k = \frac{1}{n} \sum_{i=1}^n x_i^k.$$

Da resolução daquele sistema de igualdades temos **os estimadores** ou **as estimativas**, se passamos a trabalhar com a amostra concreta com as expressões correspondentes.

Exemplo

Consideremos $X \sim \mathcal{N}(\mu, \sigma)$. Quais os **estimadores** obtidos pelo método dos momentos de μ e de σ^2 ?

Tem-se :

$$\begin{array}{ll} E[X] = \mu & \text{e} \quad M'_1 = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \\ E[X^2] = \sigma^2 + \mu^2 & \text{e} \quad M'_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{array}$$

então

$$\left\{ \begin{array}{l} \mu = \bar{X} \\ \sigma^2 + \mu^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \end{array} \right. \iff \left\{ \begin{array}{l} \mu^* = \bar{X} \\ (\sigma^2)^* = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 \end{array} \right.$$

Exercício

Considere $X \sim \text{Exp}(\lambda)$. Determine o **estimador** de momentos de λ .

Os cálculos não são complicados, mas no cálculo dos momentos empíricos aparecem potências de expoente elevado quando há muitos parâmetros, conduzindo a estimativas instáveis.

Por isso, **como regra prática deve evitar-se recorrer ao método dos momentos para mais de quatro parâmetros.**

Nota: Os estimadores obtidos pelo método dos momentos são menos eficientes do que os estimadores de máxima verosimilhança, que passamos já a apresentar.

A Verosimilhança

Seja X uma v.a. cuja distribuição depende de um **parâmetro** θ , desconhecido, e (X_1, \dots, X_n) uma amostra aleatória.

Seja (x_1, \dots, x_n) a amostra observada.

Definição


Chama-se **verosimilhança da amostra** e representa-se por $\mathcal{L}(\theta | \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ a

$$f(x_1, \dots, x_n | \theta) = \prod_{i=1}^n f(x_i | \theta) \quad \text{caso contínuo}$$

$$P(X_1 = x_1, \dots, X_n = x_n | \theta) = \prod_{i=1}^n P(X_i = x_i | \theta) \quad \text{caso discreto}$$

O Método da Máxima Verosimilhança (MMV)

O **método da máxima verosimilhança**, proposto por Fisher em 1922 e desenvolvido em 1925 consiste em escolher como **estimativa de θ** o valor que **maximiza a verosimilhança** $\mathcal{L}(\theta | \mathbf{x}_1, \dots, \mathbf{x}_n)$, face a uma amostra observada (x_1, \dots, x_n) .

Vejamos uma ilustração no 

Exemplo muito simples

Num pequeno lago que temos no jardim há peixes. Eu gostava de saber quantos peixes terá o lago. Como posso “estimar”?

Um procedimento comum é o chamado método de captura-recaptura, de que já falámos quando demos o modelo de probabilidade hipergeométrico. Como se descreve o método?

- Apanhámos $M = 7$ peixes, marcámos cada um e libertámo-lo de seguida.
- Aguarda-se uns dias para voltarem “aos seus hábitos” e vamos pescar uns quantos de novo.
- Seja X a v.a. que conta o número de capturados-marcados, i.e. são os recapturados.
- Suponhamos que da segunda vez apanhámos $n = 4$, dos quais 3 estão marcados, então $x = 3$.

O Método da Máxima Verosimilhança (MMV)

Ora pretendemos saber o número N de peixes que existem no lago

$$N = M + (N - M);$$

X designa o número de peixes recapturados.

$$X \sim \mathcal{H}(N, n = 4, M = 7) \quad \text{então} \quad P[X = x] = \frac{\binom{7}{x} \binom{N-7}{4-x}}{\binom{N}{4}}$$

Se **encontrámos 3 marcados**, qual o número de peixes que com maior probabilidade lá existirá?

Sabemos que $N \geq 7$, mas poderá ser $N = 8, 9, 10, 11, 12, 13, 14 \dots$?

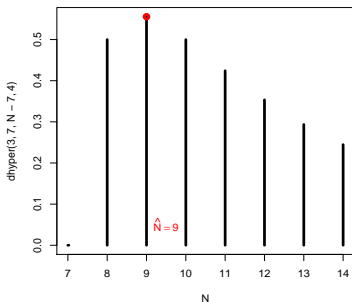
Foi Sir Ronald Fisher fez a pergunta: “*What is the value of N which has the highest likelihood?*”

Ou seja, para os valores possíveis para N qual o que torna máxima a probabilidade acima?

Exemplo muito simples

Resolver no , calculando as probabilidades e fazendo um gráfico

```
N<-seq(7,14);N  
plot(N,dhyper(3,7,N-7,4),type="h",lwd=4)  
points(9,dhyper(3,7,2,4),type="p",col=2,lwd=4)  
text(9.5,0.05,expression(hat(N)==9),col=2)
```



O Método da Máxima Verosimilhança (MMV)

Em muitas situações as funções de verosimilhança satisfazem condições que permitem que o valor para o qual a verosimilhança é máxima seja obtido tal como aprenderam na Análise Matemática, i.e., por derivação.

Porém, e como a função logarítmica é monótona, regra geral é mais cómodo trabalhar com a função **log-verosimilhança**,

$$\log \mathcal{L}(\theta|\underline{x})$$

O Método da Máxima Verosimilhança (MMV)

Então, no caso de existirem derivadas (e para um único parâmetro θ), o valor do maximizante é obtido de modo que:

$$\frac{d \log \mathcal{L}}{d\theta} = 0 \quad \text{e} \quad \frac{d^2 \log \mathcal{L}}{d\theta^2} < 0$$

Note-se que
$$\frac{d \log \mathcal{L}}{d\theta} = \sum_{i=1}^n \frac{d \log f(x_i|\theta)}{d\theta}$$

A solução, $\hat{\theta}(\mathbf{x}_1, \dots, \mathbf{x}_n)$ é **a estimativa de máxima verosimilhança**, que é uma realização da v.a. $\hat{\Theta} = \hat{\Theta}(X_1, \dots, X_n)$.

Exercício (muito simples)


Determine o estimador de máxima verosimilhança do parâmetro λ do modelo exponencial.

Propriedades dos estimadores de Máxima Verosimilhança

- Se $\hat{\theta}$ é estimador de máxima verosimilhança de θ e se $g(\theta)$ é uma função biunívoca então $g(\hat{\theta})$ é estimador de máxima verosimilhança de $g(\theta)$.
- Se existir estimador eficiente de um parâmetro, o estimador de máxima verosimilhança coincide com esse estimador.
- Regra geral o estimador de máxima verosimilhança não é centrado, mas é **assintoticamente normal** com **valor médio θ** e **variância assintótica** dada por

$$E \left[\left(\frac{\partial \log \mathcal{L}}{\partial \theta} \right)^2 \right] = \frac{1}{nE \left[\left(\frac{\partial \log f}{\partial \theta} \right)^2 \right]} = \frac{1}{-E \left[\frac{\partial^2 \log \mathcal{L}}{\partial \theta^2} \right]}$$

A máxima verosimilhança no R

No  existem funções para obter as **estimativas de máxima verosimilhança**, usando procedimentos iterativos.

Temos, por exemplo, as funções


`mle()` do *package* **stats4**

`fitdistr()` do *package* **MASS**

`optimize ()` – adequado quando há só um parâmetro

`optim ()` – adequado quando há dois ou mais parâmetros

Exemplo de utilização do R

Consideremos o ficheiro de dados do  `PlantGrowth`.

Vamos usar a amostra de valores observados da variável “weight”.

Vamos supor que as observações são a concretização de uma a.a.

(X_1, X_2, \dots, X_n) , i.i.d. $\mathcal{N}(\mu, \sigma)$

Primeiro necessitamos de escrever a função de log-verosimilhança

Observação importante a função de optimização, por omissão calcula o mínimo da função log-verosimilhança, que deverá por isso ser negativa

```
logL_negativa <- function(mu, sigma){  
  -sum(dnorm(x, mean = mu, sd = sigma, log = TRUE))}
```

Para obter a estimativa de máxima verosimilhança, o algoritmo de optimização requer valores iniciais para os parâmetros a estimar.


Resolução do Exemplo

Consideremos como **valores iniciais** a média e o desvio padrão observados

```
> media<-mean(x);media
> desvio<-sd(x);desvio
> MLest1 <- mle(logL_negativa, start = list(mu = media, sigma= desvio))
> summary(MLest1)
Maximum likelihood estimation
Call:
mle(minuslogl = logL_negativa, start = list(mu = media, sigma = desvio))

Coefficients:
      Estimate Std. Error
mu      5.0730000 0.12586801
sigma  0.6894075 0.08900152
-2 log L: 62.82084
```

RESUMO de algumas distribuições no R

Nome da distribuição no 	Função	Argumentos
Beta	beta	shape1, shape2
Binomial	binom	size, prob
Cauchy	cauchy	location, scale
Chisquare	chisq	df
Exponential	exp	rate
FDist	f	df1, df2
GammaDist	gamma	shape, scale
Geometric	geom	prob
Hypergeometric	hyper	m, n, k
Lognormal	lnorm	meanlog, sdlog
Logistic	logis	location, scale
NegBinomial	nbinom	size, prob
Normal	norm	mean, sd
Poisson	pois	lambda
TDist	t	df
Uniform	unif	min,max
Weibull	weibull	shape, scale