Verossimilhança e Máxima Verossimilhança[1]

JOÃO LUÍS F. BATISTA[2]
Maio - 2009

1 Introdução

Tradicionalmente a inferência estatística sobre a média de uma população se apoia no Teorema Central do Limite para construir Intervalos de Confiança ou testar hipóteses sobre o valor do parâmetro. Esta abordagem da estatística tradicional pode ser extendida para inferências a respeito de qualquer parâmetro, não só a média. Da mesma forma que no caso da média populacional se usa a distribuição t de Student ou a distribuição Normal Padronizada, no caso de outros parâmetros se utiliza outras *distribuições amostrais*. Essas distribuições são chamadas *amostrais* porque representam o comportamento das estimativas baseado na repetição incontável do processo de amostragem.

Na prática científica, no entanto, sempre se realiza uma única amostragem, o que resulta em uma única amostra. Assim, o conceito de distribuição amostral é até certo ponto artificial, pois em pesquisa científica não raciocinamos em termos de repetições incontáveis de experimentos ou processos de observação. O resultado disto é que o conceito de teste estatístico de hipótese e de intervalo de

[1] Material de estudo que acompanha aula sobre o tema

[2] Centro de Métodos Quantitativos (http://cmq.esalq.usp.br/), Departamento de Ciências Florestais, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Campus Piracicaba.

confiança são frequentemente mal compreendidos.

O desenvolvimento da inferência estatística a partir do conceito de verossimilhança tem sido utilizado como uma alternativa à abordagem estatística frequentista e, segundo alguns autores (como por exemplo Royall, 1997), é mais coerente com a prática científica.

2 Lei da Verossimilhança

Considere uma variável aleatória X, cujo comportamento pode ser explicado por duas hipóteses (hipóteses A e B) que se deseja comparar. Foi realizado um estudo e se obteve uma observação de X, cujo valor foi x.

O que as hipótese dizem a respeito dessa observação?

- A hipótese A implica que $X = x$ seria observado com probabilidade $p_A(x)$, enquanto
- A hipótese B implica que $X = x$ seria observado com probabilidade $p_B(x)$.

No processo de investigação científica, no entanto, o que interessa é a pergunta: "O que a observação de $X = x$ diz a respeito das hipóteses A e B?" A **Lei da Verossimilhança** afirma que a observação $X = x$ é uma evidência que favorece a hipótese A sobre a hipótese B se e somente se

$$p_A(x) > p_B(x).$$

Mais ainda, a Lei da Verossimilhança implica que a **Razão de Verossimilhança**

$$\frac{p_A(x)}{p_B(x)}$$

mede a ***força de evidência*** em favor da hipótese A sobre a hipótese B.

2.1 Exemplo de Regeneração Natural

Deseja-se saber o número médio de plântulas numa parcela de regeneração em floresta nativa. Para isso se utilizou uma parcela circular de 3 m de raio (28.3 m^2). Há duas hipóteses competindo:

- Hipótese A: o número médio de plântulas na parcela é 16 (5700 ind/ha);

- Hipótese B: o número médio de plântulas na parcela é 35 (12500 ind/ha);

Foi medida uma parcela de regeneração e observou-se 24 plântulas (8470 ind/ha).

Neste exemplo, a variável aleatória X é o número de plântulas por parcela e a observação tomada foi de $X = 24$. Faz-se necessário um modelo (distribuição estatística) para se calcular as probabilidades de se observar $X = 24$ sob as duas hipóteses. Como se trata de dados de contagem, a distribuição *Poisson* é uma candidata "natural". Se a variável aleatória X tem distribuição Poisson, sua função de densidade é

$$P(X = x) = \frac{\mathrm{e}^{-\mu}\mu^x}{x!}$$

onde μ (o parâmetro) é o número médio de plântulas.

Portanto as probabilidades de acordo com as hipóteses são:

- Hipótese A: $\mu = 16$

$$p_A(24) = \frac{\mathrm{e}^{-16}16^{24}}{24!} = 0.01437018$$

- Hipótese B: $\mu = 35$

$$p_B(24) = \frac{\mathrm{e}^{-35}35^{24}}{24!} = 0.01160434$$

Logo, a observação $X = 24$ favorece a hipótese A ($\mu = 16$) sobre a hipótese B ($\mu = 35$). A *força de evidência* em favor de A sobre B é:

$$\frac{p_A(24)}{p_B(24)} = \frac{0.01437018}{0.01160434} = 1.238345.$$

Ou seja, pode se dizer que a observação $X = 24$ é evidência que a hipótese A é aproximadamente 1.3 vêzes mais *verossímil* que a hipótese B.

2.2 Função de Verossimilhança

Existe uma diferença sutil entre probabilidade e verossimilhança. Note que para se comparar as hipóteses no exemplo acima utilizou-se a *função de densidade* da distribuição Poisson:

$$P(X = x) = \frac{\mathrm{e}^{-\mu}\mu^x}{x!}$$

onde x é o valor de uma observação da variável aleatória Poisson X e μ é o seu parâmetro. Ao se utilizar essa função a observação $X = x$ ($X = 24$) foi dada e, portanto, o valor de x é conhecido e fixo. A função portanto ficaria:

$$P(X = 24) = \frac{\mathrm{e}^{-\mu}\mu^{24}}{24!}$$

Note que essa expressão não é mais uma função do valor da observação x, mas uma função do valor do parâmetro μ, que varia da hipótese A para hipótese B.

Quando numa função de densidade, a observação é fixa e o parâmetro variável não se tem mais uma função de densidade e sim uma **função de verossimilhança**. A função de verossimilhança indica a verossimilhança de uma dada hipótese, por exemplo hipótese A ($\mu = 16$), dado que se obteve uma observação $X = x$ ($X = 24$). Para tornar mais claro este conceito se utiliza uma notação diferente para a função de verossimilhança:

$$\mathcal{L}\{\text{ hipótese } | \text{ dados }\} \quad \text{ou} \quad \mathcal{L}\{A|X = x\} \quad \text{ou (ainda mais curto)} \quad \mathcal{L}\{\mu|X\}$$

No exemplo acima, a expressão matemática está condicionada à observação $X = 24$ sendo, portanto, a própria função de verossimilhança:

$$\mathcal{L}\{\mu|X = 24\} = \frac{\mathrm{e}^{-\mu}\mu^{24}}{24!}.$$

A figura 1 apresenta o gráfico dessa função para valores de μ entre 10 e 50.

É importante notar que o valor da observação obtida influencia fortemente o comportamento da função de verossimilhança, o que pode ser observado na figura 2.

2.3 Múltiplas Observações

Nos exemplos apresentados até agora, tinha-se uma única observação ($X = x$), o que é uma situação bastante peculiar pois quando fazemos um estudo tomamos uma série de observações para compor uma amostra.

Em geral, a amostra é composta de observações *independentes*. Assumindo-se como verdadeira uma certa hipótese A, a probabilidade de se opter uma amostra $\mathbf{X}_n$ ($\mathbf{X}_n = \{x_1, x_2, \ldots, x_n\}$), composta de n observações independentes de X ($x_1, x_2, \ldots, x_n$), é:

$$P(\mathbf{X}_n|A) = P(X = x_1|A) \cdot P(X = x_2|A) \cdot \ldots \cdot P(X = x_n|A)$$

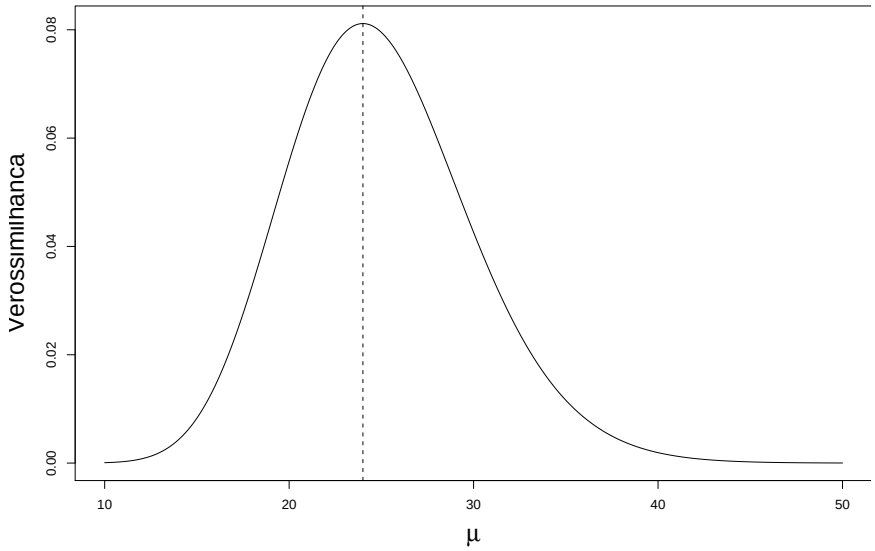

Figura 1: Função de verossimilhança da distribuição Poisson, quando se obtem uma observação $X = 24$.

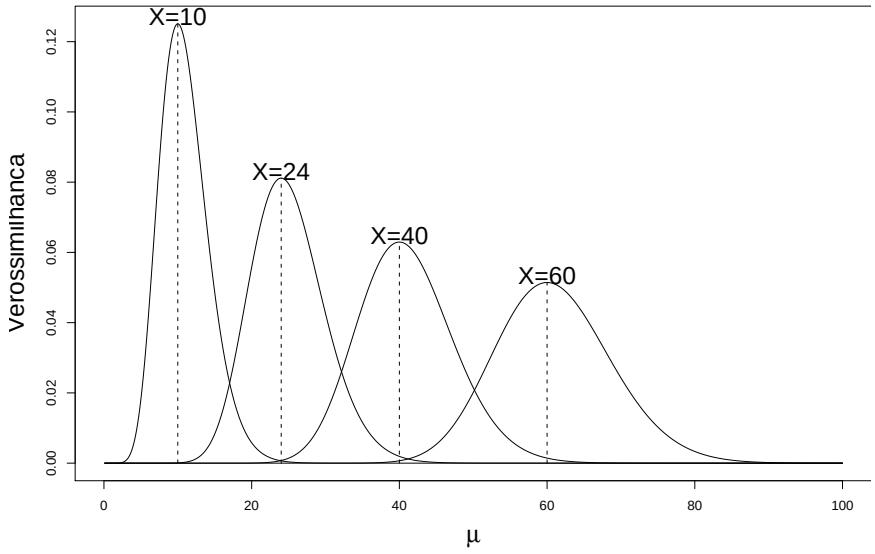

Figura 2: Funções de verossimilhança da distribuição Poisson para diferentes valores observados de X.

Ou seja, a probabilidade de se obter a amostra, dado a hipótese A, é igual ao produto das probabilidades das observações individuais, dado a hipótese A.

Este mesmo princípio se aplica à verossimilhança. A função de verossimilhança de uma amostra composta de observações independentes será o produto das funções de verossimilhança das observações individuais:

$$\begin{aligned}\mathcal{L}\{A|\mathbf{X}_n\} &= \mathcal{L}\{A|X=x_1\}\cdot\mathcal{L}\{A|X=x_2\}\cdot\ldots\cdot\mathcal{L}\{A|X=x_n\}\\ \mathcal{L}\{A|\mathbf{X}_n\} &= \prod_{i=1}^{n}\mathcal{L}\{A|X=x_i\}\end{aligned}$$

2.3.1 Exemplo da Regeneração Natural Revisitado

Considere um exemplo mais realista de um levantamento de regeneração natural onde a amostrada foi composta de 10 parcelas, onde se realizou a contagem do número de plântulas. Os resultados são apresentados na tabela 1.

Note que a amostra com 10 parcelas continua a indicar a hipótese A ($\mu = 16$) como mais verossímil, mas com uma margem bem menor. Outro fato que chama atenção é que a verossimilhança da amostra é um número muito pequeno da ordem

Tabela 1: Números de plântulas observados em 10 parcelas de regeneração natural, seguidos dos valores de verossimilhança calculados segundo a distribuição de Poisson.

PARCELA	NO. DE PLÂNTULAS	VEROSSIMILHANÇA	
(i)	$(X = x_i)$	$(\mu = 16)$	$(\mu = 35)$
1	24	0.0144	0.0116
2	27	0.0034	0.0283
3	23	0.0216	0.0080
4	28	0.0019	0.0354
5	26	0.0057	0.0219
6	24	0.0144	0.0116
7	17	0.0934	0.0003
8	23	0.0216	0.0080
9	24	0.0144	0.0116
10	27	0.0034	0.0283
	$\prod_{i=1}^{10}$	2.2574×10^{-22}	2.2500×10^{-22}

de 10^{-22}. Como a verossimilhança da amostra é o **produto** da verossimilhança das observações, ela rapidamente se aproxima do zero quando o tamanho da amostra cresce.

2.4 Log-Verossimilhança Negativa

Para tornar mais fácil a manipulação matemática da verossimilhança se utiliza a função de *log-verossimilhança negativa*, que consistem aplicar a função logaritmo, geralmente logaritmo natural ou neperiano, e transformar o sinal:

$$\mathbf{L}\{\mu|X\} = -\log\left[\,\mathcal{L}\{\mu|X\}\,\right].$$

Como o valor numérico da verossimilhança é geralmente (mas não necessariamente) menor que um, o logaritmo desse valor é negativo. Assim a transformação do sinal é realizada para que a log-verossimilhança negativa seja um valor positivo, o que geralmente (mas não necessariamente) ocorre. Assim, se o valor da verossimilhança de uma amostra com muitas observações é um *número positivo muito pequeno*, o valor da log-verossimilhança negativa será um *número positivo* numa escala mais fácil de trabalhar.

Por outro lado, o fato da transformação incluir uma *mudança de sinal* implica que o comportamento da função de log-verossimilhança negativa é oposto ao comportamento da função de verossimilhança. Isso significa que a hipótese com **maior** verossimilhança terá **menor** log-verossimilhança negativa.

A transformação logarítmica também auxilia muito a *tratabilidade matemática* da função de verossimilhança, pois ela faz com que a log-verossimilhança negativa de uma amostra seja o *somatório* das log-verossimilhanças negativas das observações independentes individuais:

$$\begin{aligned}
\mathbf{L}\{\mu|\mathbf{X}_n\} &= -\log\left[\,\mathcal{L}\{\mu|\mathbf{X}_n\}\,\right] \\
&= -\log\left[\,\prod_{i=1}^{n}\mathcal{L}\{\mu|X=x_i\}\,\right] \\
&= \sum_{i=1}^{n} -\log\left[\,\mathcal{L}\{\mu|X=x_i\}\,\right] \\
\mathbf{L}\{\mu|\mathbf{X}_n\} &= \sum_{i=1}^{n}\mathbf{L}\{\mu|X_i\}
\end{aligned}$$

2.4.1 Exemplo da Regeneração Natural Revisitado Novamente

A aplicação da log-verossimilhança negativa no exemplo regeneração natural é apresentada na tabela 2. Os valores de log-verossimilhança negativa para amostra estão agora numa escala bem mais fácil de trabalhar. Por outro lado, a hipótese mais provável ($\mu = 16$) apresenta agora a *menor* log-verossimilhança negativa.

Para cálculo da verossimilhança de uma amostra não é necessário calcular a verossimilhança para cada observação para depois obter o valor para amostra. Com um pouco de tratamento algébrico pode se obter a expressão da função de verossimilhança para a amostra, mas a expressão da função de log-verossimilhança negativa é sempre mais conveniente.

No caso da distribuição Poisson, a expressão da verossimilhança da amostra de tamanho n é:

$$\mathcal{L}\{\mu|X_n\} \;=\; \prod_{i=1}^{n} \frac{\mathrm{e}^{-\mu}\mu^{x_i}}{x_i!}. = \mathrm{e}^{-\mu} \prod_{i=1}^{n} \frac{\mu^{x_i}}{x_i!}.$$

Nessa expressão, o cálculo da verossimilhança para cada valor do parâmetro (μ) envolverá necessariamente cálculos com cada observação individual (x_i).

Tabela 2: Números de plântulas observados em 10 parcelas de regeneração natural, seguidos dos valores de *log-verossimilhança negativa* calculados segundo a distribuição de Poisson.

PARCELA (i)	NO. DE PLÂNTULAS ($X = x_i$)	LOG-VEROSSIMILHANÇA NEG. ($\mu = 16$)	($\mu = 35$)
1	24	4.2426	4.4564
2	27	5.6976	3.5631
3	23	3.8371	4.8337
4	28	6.2573	3.3400
5	26	5.1744	3.8227
6	24	4.2426	4.4564
7	17	2.3711	8.0642
8	23	3.8371	4.8337
9	24	4.2426	4.4564
10	27	5.6976	3.5631
	$\sum_{i=1}^{10}$	49.8427	49.8459

A transformação gera a seguinte função log-verossimilhança negativa para distribuição Poisson:

$$\mathbf{L}\{\mu|X_n\} \quad = \quad n\mu - \log(\mu)\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\log(x_i!)$$

Nessa expressão os dois somatórios envolvem exclusivamente os valores das observações, sendo, portanto, constantes para uma determinada amostra. Assim, o cálculo da log-verossimilhança se torna bem mais simples na seguinte forma:

$$\mathbf{L}\{\mu|X_n\} \quad = \quad n\mu - \log(\mu)\, k_1 + k_2$$

onde $k_1 = \sum_{i=1}^{n} x_i$ e $k_2 = \sum_{i=1}^{n}\log(x_i!)$.

A figura 3 ilustra o comportamento da função de verossimilhança e da função de log-verossimilhança negativa do modelo Poisson com os dados de regeneração natural.

3 *Método da Máxima Verossimilhança*

O método da **Máxima Verossimilhança** consiste em estimar os parâmetros de um modelo utilizando as estimativas que tornam *máximo* o valor da função de verossimilhança. Isso é equivalente a encontrar o valor para o parâmetro que torna *mínima* a função de log-verossimilhança negativa. Olhando o gráfico da função da verossimilhança ou da log-verossimilhança negativa (figura 4) fica claro onde esse ponto se encontra.

Utilizando o cálculo diferencial, podemos encontrar o ponto de mínimo de uma função igualando a zero a primeira derivada da função e solucionando a expressão. No caso da distribuição Poisson, a função log-verossimilhança negativa é:

$$\mathbf{L}\{\mu|X_n\} \quad = \quad n\mu - \log(\mu)\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\log(x_i!).$$

Encontrando a primeira derivada de $\mathbf{L}\{\mu|X_n\}$ e igualando a zero temos:

$$\frac{d\mathbf{L}\{\mu|X_n\}}{d\mu} \quad = \quad n - \frac{\sum_{i=1}^{n} x_i}{\mu} = 0 \quad \Rightarrow \quad \widehat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

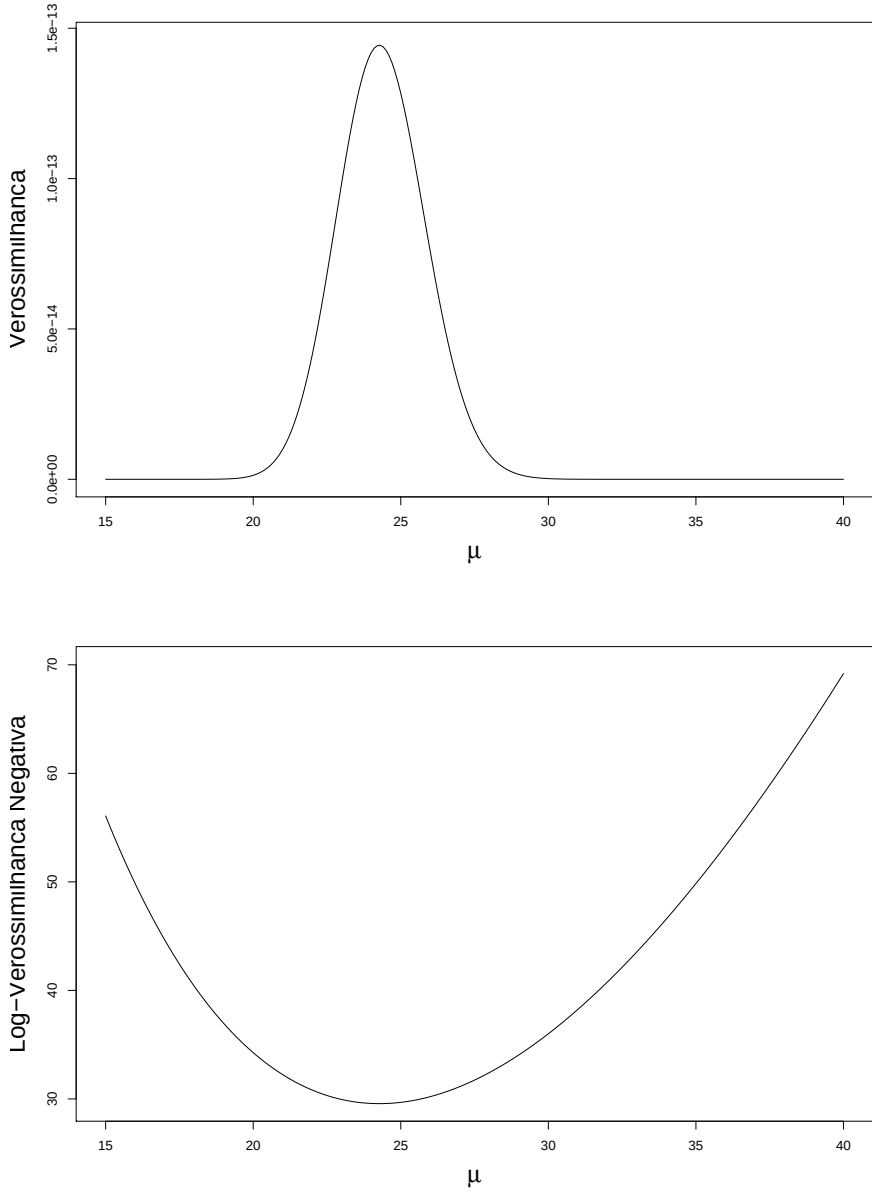

Figura 3: Função de verossimilhança e de log-verossimilhança negativa da distribuição Poisson para uma amostra de tamanho 10: $X = \{24, 27, 23, 28, 26, 24, 17, 23, 24, 27\}$.

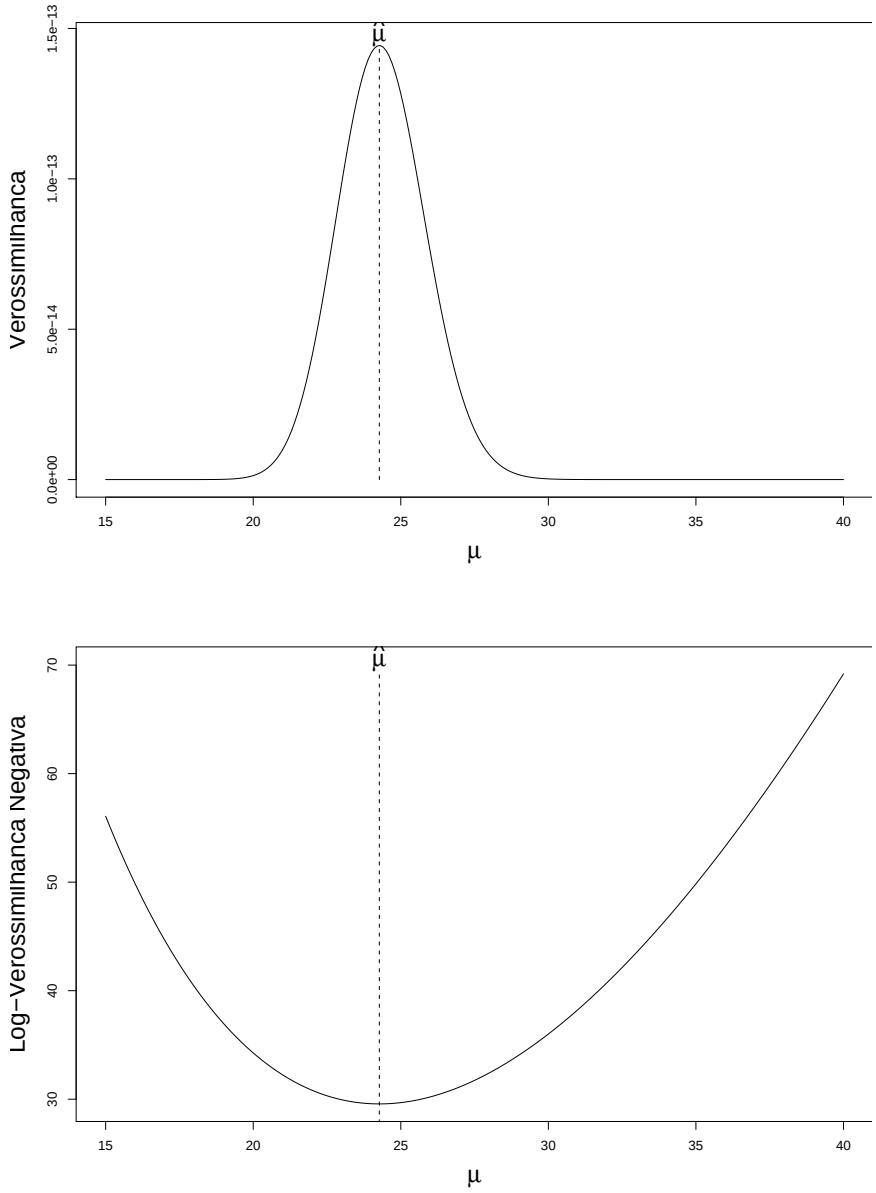

Figura 4: Função de verossimilhança e log-verossimilhança negativa da distribuição Poisson para uma amostra de tamanho 10: $X = \{24, 27, 23, 28, 26, 24, 17, 23, 24, 27\}$. A linha vertical indica a possição da estimativa de máxima verossimilhança.

Portanto, a estimativa de máxima verossimilhança do parâmetro μ da distribuição de Poisson nada mais é que a média amostral. Assim no exemplo de regeneração de plântulas temos como estimativa de máxima verossimilhança:

$$\widehat{\mu} \;=\; \frac{24+27+23+28+26+24+17+23+24+27}{10} = 24.3$$

3.1 Propriedades das Estimativas de Máxima Verossimilhança

O método da máxima verossimilhança é um método estabelecido de estimação de parâmetros de modelos estatístico, sendo utilizado por estatísticos teóricos e práticos de todas as tribos. O uso generalizado do método se deve às propriedades probabilísticas das estimativas produzidas por ele.

As estimativas de máxima verossimilhança são chamadas em inglês de *maximum likelihood estimates*, sendo portanto designadas pela sigla MLE. Assumindo-se que a função de verossimilhança satisfaz algumas propriedades matemáticas básicas, que são frequentemente alcançadas pelos modelos utilizados em Ecologia, as MLE tem as seguintes propriedades:

Consistência: as MLE são consistentes, i.e., elas *convergem em probabilidade* para o valor do parâmetro. Ou seja, para grandes amostras ($n \to \infty$) as MLE, para efeitos práticos, são não-viciadas (não enviesadas).

Eficiência Assintótica: O Teorema do Limite Inferior de Cramer-Rao afirma que, para um dado parâmetro qualquer, existe um limite inferior para a variância das estimativas não-viciadas. Para grandes amostras, as MLE atingem esse limite e, portanto, têm a menor variância possível dentre as estimativas não-viciadas.

Normalidade Assimptótica: As MLE convergem em distribuição para distribuição Gaussiana. Para grandes amostras, os MLE tem distribuiçõa aproximadamente gaussiana.

Invariância: as MLE são invariantes sob transformações monotônicas. Por exemplo, seja $\widehat{\mu}$ uma MLE que pode ser transformada para:

- $\widehat{\theta}_1 = \log(\widehat{\mu})$,
- $\widehat{\theta}_2 = \sqrt{\widehat{\mu}}$ e
- $\widehat{\theta}_3 = \mathrm{e}^{\widehat{\mu}}$,

então as estimativas $\widehat{\theta}_1$, $\widehat{\theta}_2$ e $\widehat{\theta}_3$ também são MLE.

Uma aspecto muito importante que deve ser frisado é que as três primeiras propriedades são válidas para ***grandes amostras***.

4 Intervalo de Verossimilhança

O intervalo de verossimilhança corresponde a um intervalo ao redor da MLE, onde a razão das verossimilhanças dos valores dentro do intervalo para a verossimilhança da MLE não ultrapassa um certo limite. A figura 5 mostra um intervalo de verossimilhança para o exemplo da regeneração natural onde o limite é 8. Ou seja, dentro deste intervalo, a razão de verossimilhança é definida como:

$$\frac{\mathcal{L}\{\widehat{\mu}|X_{10}\}}{\mathcal{L}\{\mu|X_{10}\}} \leq 8 \implies \mathcal{L}\{\mu|X_{10}\} \geq \frac{\mathcal{L}\{\widehat{\mu}|X_{10}\}}{8}$$

onde

X_{10} se refe a amostra com 10 parcelas (unidades amostrais),

$\widehat{\mu}$ é a MLE, e

μ é o parâmetro variando dentro do intervalo.

Como na prática trabalhamos com a função de log-verossimilhança negativa, o intervalo de verossimilhança também pode ser construido na forma de um *intervalo de log-verossimilhança negativa*, com interpretação análoga. Para isso é necessário aplicar a transformação à razão de verossimilhança:

$$\begin{aligned} -\log\left\{\frac{\mathcal{L}\{\widehat{\mu}|X_{10}\}}{\mathcal{L}\{\mu|X_{10}\}}\right\} \geq -\log(8) &\implies \mathbf{L}\{\widehat{\mu}|X_{10}\} - \mathbf{L}\{\mu|X_{10}\} \geq -\log(8) \\ &\implies \mathbf{L}\{\mu|X_{10}\} \leq \mathbf{L}\{\widehat{\mu}|X_{10}\} + \log(8) \end{aligned}$$

Em termos de log-verossimilhança negativa a razão de 8, se torna uma **diferença** de $\log(8)$.

Analisando a curva de verossimilhança ou log-verossimilhança negativa verifica-se que ela é ligeiramente assimétrica em relação à MLE, por isso, o intervalo de verossimilhança também é ligeiramente assimétrico. Portanto, o intervalo de verossimilhança é determinado pela forma da curva de verossimilhança, ao contrário do *intervalo de confiança* na estatística frequentista que é simétrico por definição.

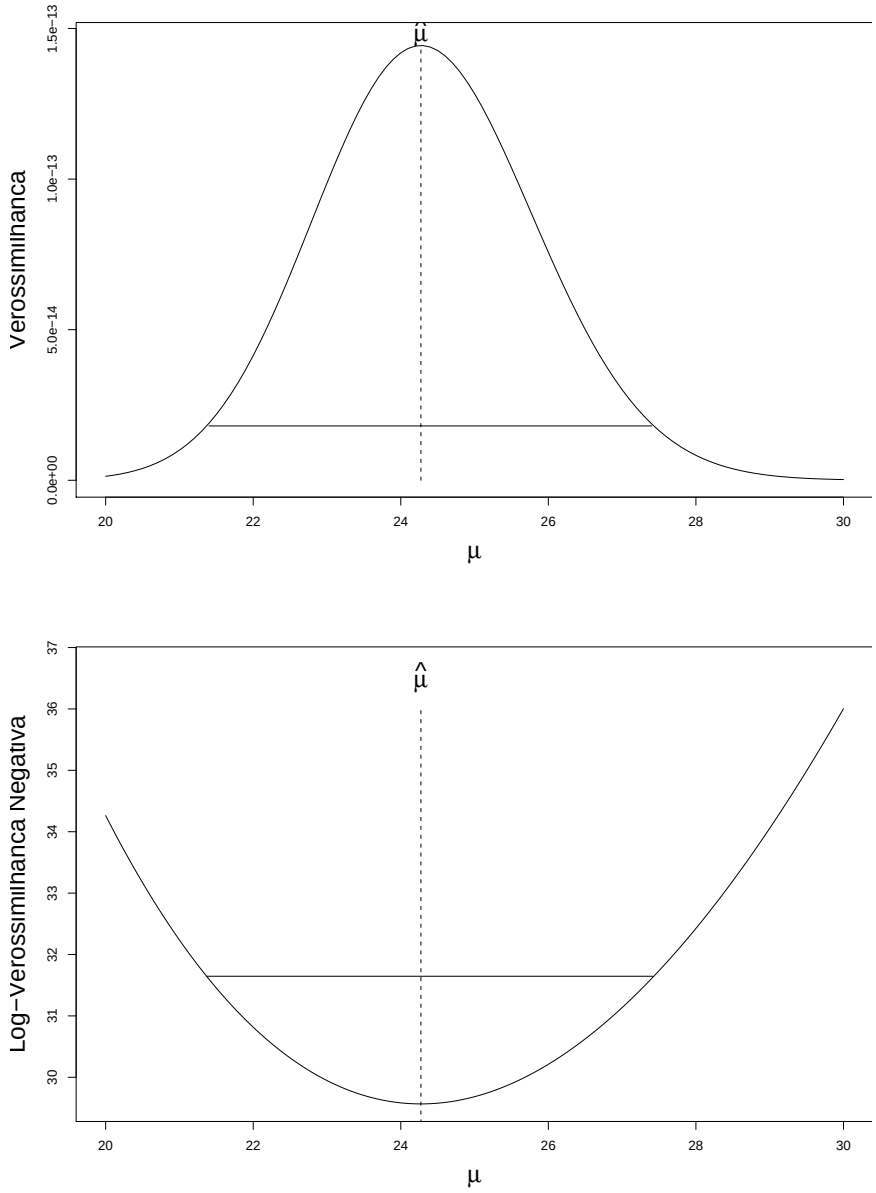

Figura 5: Função de verossimilhança e log-verossimilhança negativa da distribuição Poisson para uma amostra de tamanho 10: $X = \{24, 27, 23, 28, 26, 24, 17, 23, 24, 27\}$. A linha vertical indica a possição da MLE. A linha horizontal indica um intervalo de verossimilhança para uma razão de 8.

O intervalo de verossimilhança também redefine a forma de apressentação da curva de verossimilhança. Como o valor da verossimilhança, ou da log-verossimilhança negativa, não pode ser interpretado de forma direta e absoluta, faz sentido apresentar as curvas sempre *em termos relativos à máxima verossimilhança*.

A figura 6 apresenta os mesmos gráficos da figura 5, mas com a escala relativa. Na escala relativa, a forma da curva não se altera, mas a visualização do intervalo fica mais fácil. No gráfico da razão de verossimilhança, o intervalo de verossimilhança para a razão $R = 8$, será sempre definido por uma linha horizontal na altura $1/R = 1/8$. No gráfico da diferença de log-verossimilhança negativa, o intervalo de verossimilhança para a razão $R = 8$, será sempre definido por uma linha horizontal na altura $\log(R) = \log(8) = 2,0794$.

4.1 Exemplo da Distribuição Binomial em Eventos Raros

Num levantamento florestal foram selecionadas aleatoriamente 100 árvores para verificar a proporção de árvores doentes. Entretanto, nenhuma delas apresentou a doença em questão.

Utilizando-se a abordagem tradicional a estimativa da probabilidade de ocorrência e sua variância são:

$$\widehat{p} = \frac{0}{100} = 0 \qquad s^2_{\widehat{p}} = \frac{\widehat{p}(1-\widehat{p})}{n} = \frac{0(1-0)}{100} = 0$$

Logo não é possível acessar a qualidade da estimativa $\widehat{p} = 0$.

Utilizando a curva de verossimilhança não há problema, pois esta curva é dada por

$$\mathcal{L}\{p\} = \left(\begin{array}{c} n \\ 0 \end{array} \right) p^0(1-p)^n = (1-p)^n$$

A figura 7 apresenta esta curva para diferentes tamanhos de amostra. Nota-se que, independentemente do valor estimado ($\widehat{p} = 0$), o intervalo de verossimilhança pode ser facilmente definido. A figura também deixa claro o forte efeito do tamanho da amostra sobre a amplitude do intervalo.

Como o tamanho da amostra efetivamente tomada em campo foi $n = 100$, verificamos na figura 7 que o intervalo de verossimilhança para a razão de 8 é $[0.00, 0.02]$. Os dados indicam que, se a doença procurada ocorre de fato na floresta, a proporção de árvores doentes é no máximo 2%.

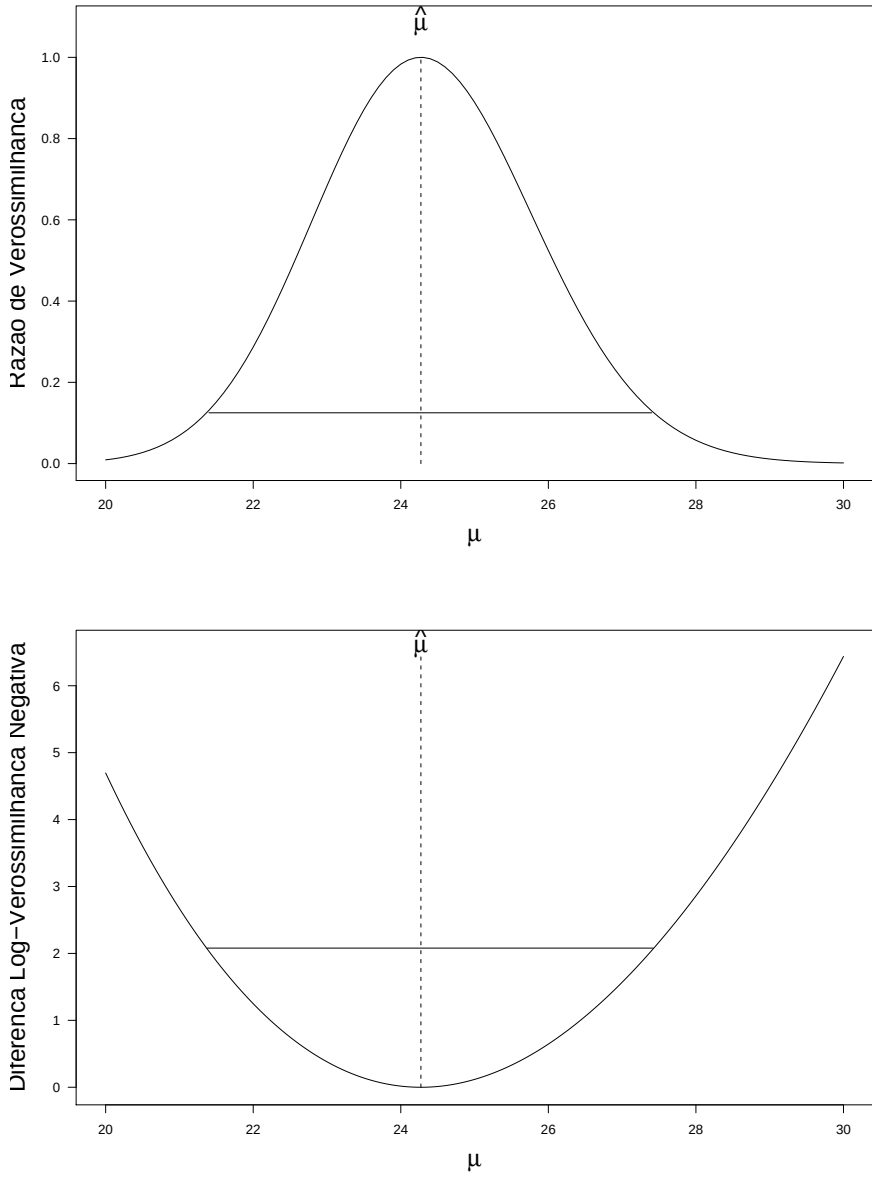

Figura 6: Curva **relativa** da verossimilhança e da log-verossimilhança negativa da distribuição Poisson para uma amostra de tamanho 10.

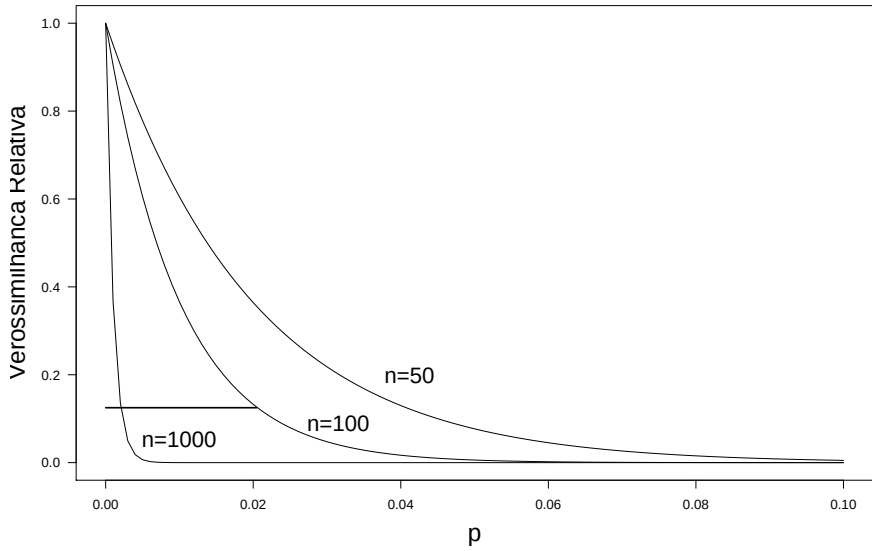

Figura 7: Curva de verossimilhança para o modelo binomial, tomando 0 sucessos observados numa amostra de tamanho n. A linha horizontal representa a fração $1/8$.

5 Superfícies e Curvas de Verossimilhança

O exemplo que vem sendo apresentado utiliza a distribuição de Poisson, que possui apenas um parâmetro (μ). No caso das distribuições estatísticas com dois ou mais parâmetros, a *curva de verossimilhança* se transforma numa *superfície de verossimilhança*.

A figura 8 ilustra como a superfície de log-verosimilhança negativa *relativa* para diversas amostras de uma distribuição Gaussiana (Normal). A distribuição Gaussiana possui dois parâmetros: média (μ) e desvio padrão (σ), de forma que apresentação foi feita na forma de um gráfico de contorno, onde as linhas representam isolinhas de log-verossimilhança negativa.

Tratando-se de dois parâmetros, não se tem mais um intervalo de verossimilhança, mas uma *região de verossimilhança*. Na figura 8, esta região é delimitada por uma linha mais espessa correspondente à isolinha para log-verorssimilhança negativa relativa igual a $\log(8)$. A figura também mostra o efeito do tamanho da amostra sobre o tamanho da região de verossimilhança: à medida que o tama-

nho da amostra aumenta o tamanho da região diminui, indicando o aumento da precisão das MLE.

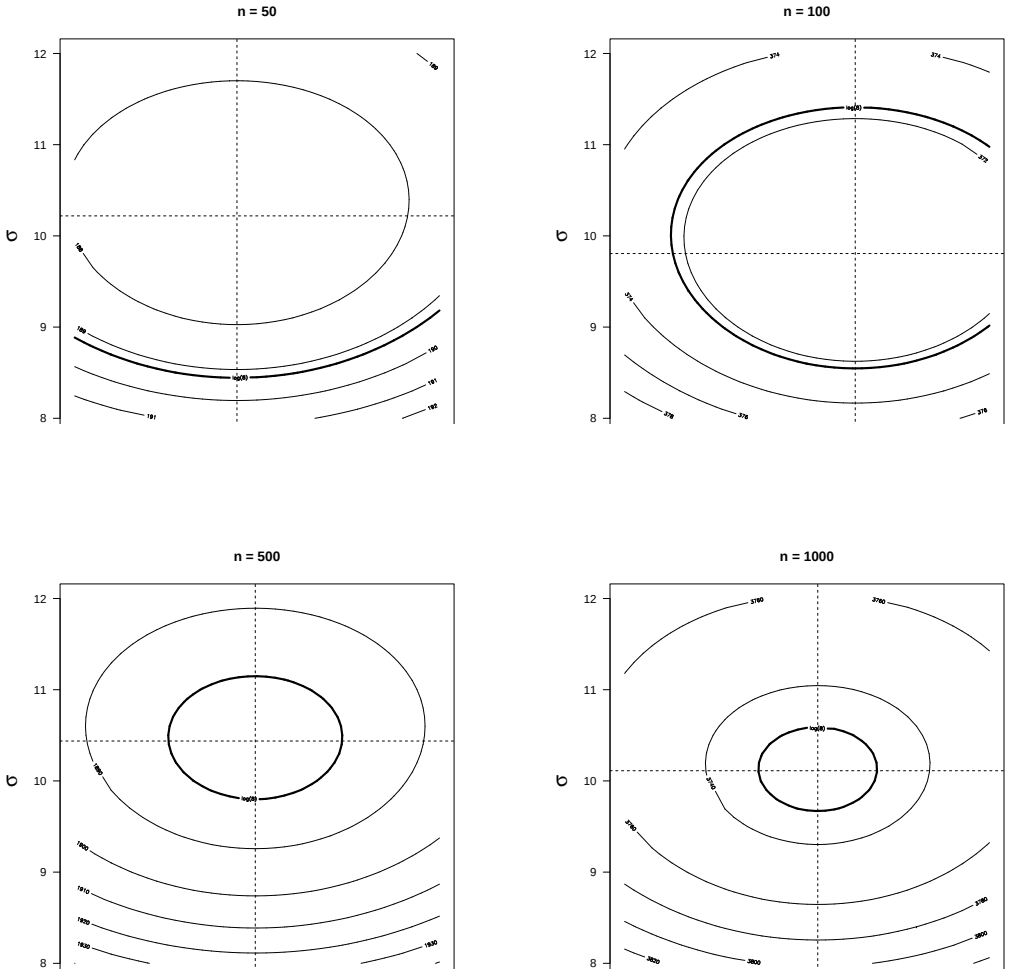

Figura 8: Gráfico de contorno com isolinhas da log-verossimilhança negativa relativa para amostras de diferentes tamanhos (n) da distribuição Gaussiana (Normal) com $\mu = 25$ e $\sigma = 10$. As linhas pontilhadas indicam os valores das MLE: $\widehat{\mu}$ e $\widehat{\sigma}$. A isolinha mais espessa indica log-verossimilhança negativa relativa igual a $\log(8)$.

5.1 De Superfície para Curvas

Frequentemente quando se trabalha com modelos de variáveis aleatórias que possuem vários parâmetros, apenas um ou alguns deles são de interesse, sendo que os demais não são considerados na análise, mas fazem parte do modelo. Por exemplo, a distribuição Gaussiana é frequentemente utilizada para se interpretar o comportamento da média (μ), independentemente do comportamento do desvio padrão (σ). Os parâmetros necessários ao modelo, mas sem interesse para interpretação, são parâmetros inconvenientes (*nuisance parameters*). Na sua presença, não é conveniente se analisar as regiões de verossimilhança, pois a falta de interpretação para os parâmetros inconvenientes só faz aumentar a complexidade da interpretação da região de verossimilhança. Assim, se faz necessário uma forma de analisar os parâmetros de interesse sem a interferência dos parâmetros inconvenientes.

Mesmo quando o número de parâmetros de interesse é maior do que dois, a interpretação gráfica da região de verossimilhança é extremamente complexa. Logo é necessário uma forma conveniente de se estudar o comportamento das MLE dos parâmetros em modelos com muitos parâmetros.

A forma de realizar esse estudo é substituir a análise da supefície de verossimilhança de um modelo pela análise de uma série de curvas de verossimilhança, cada uma relativa a um parâmetro de interesse no modelo.

5.2 Verossimilhança Estimada

A *Verossimilhança Estimada* é uma técnica se constroe uma curva de verossimilhança para cada parâmetro de interesse, mantendo os demais parâmetros (de interesse ou inconvenientes) num valor constante. O melhor valor para manter demais parâmetros é a a estimativa de máxima verossimilhança deles.

No exemplo da distribuição Gaussiana, a curva de verossimilhança estimada para média é obtida como

$$\mathbf{L}_E\{\mu\} = \mathbf{L}\{\mu, \widehat{\sigma}\}$$

onde $\widehat{\sigma}$ é a MLE do desvio padrão.

A figura 9 apresenta as curvas de verossimilhança estimada para o exemplo da distribuição Gaussiana, utilizando os mesmos dados apresentados na figura 8. Para facilitar a visualização, a curva é apresentada em função da diferença $\mu - \widehat{\mu}$, pois as amostras geraram estimativas $\widehat{\mu}$ ligeiramente distintas. O gráfico mostra o efeito do tamanho da amostra aumentando a curvatura da superfície de verossimilhança e, consequentemente, reduzindo o tamanho do intervalo de verossimi-

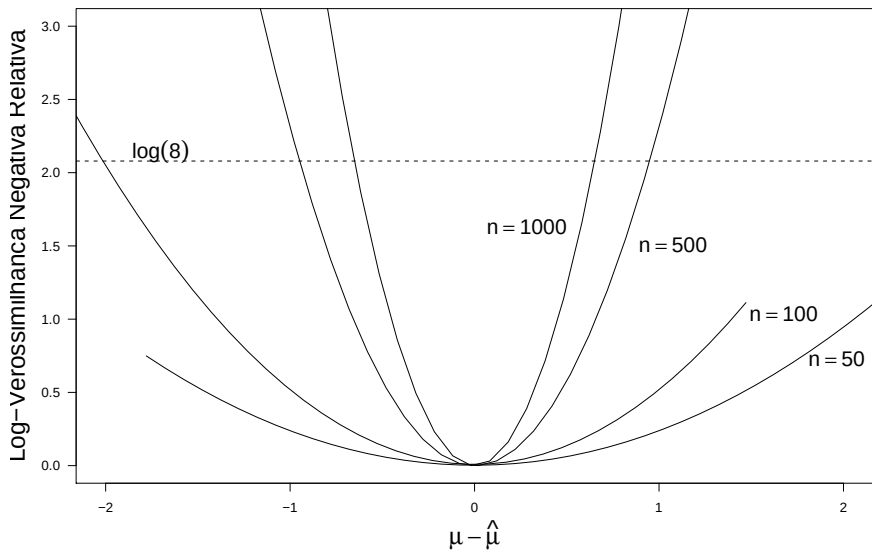

Figura 9: Log-verossimilhança negativa relativa **estimada** para amostras de diferentes tamanhos (n) da distribuição Gaussiana com $\mu = 25$ e $\sigma = 10$. As curvas foram traçadas para diferentes valores de μ, mantendo o valor do desvio padrão (σ) fixo no valor da sua MLE ($\widehat{\sigma}$).

lhança da MLE da média.

5.3 Verossimilhança Perfilhada

Outra forma de construir curvas de verossimilhança é a *Verossimilhança Perfilhada*. Essa técnica consiste em variar o parâmetro de interesse e, *para cada valor dele*, encontar as MLE dos demais parâmetros do modelo substituindo-os na função de verossimilhança para calcular a verossimilhança ponto-a-ponto do parâmetro de interesse.

Tomando o exemplo da distribuição Gaussiana, a curva de log-verossimilhança negativa perfilhada para média é construída tornando a MLE do desvio padrão (ou da variância), para cada valor da média:

$$\mathbf{L}_P\{\mu\} = \mathbf{L}\{\mu, \sigma^2 = \widehat{\sigma^2}\} = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln(\widehat{\sigma^2}) + \frac{1}{2\widehat{\sigma^2}}\sum_{i=1}^{n}(y_i - \mu)^2$$

Mas a MLE da variância é uma função da própria média:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$$

Inserindo essa expressão na função de log-verossimilhança negativa se obtem a verossimilhança perfilhada da média:

$$\mathbf{L}_P\{\mu\} = \frac{n}{2}\ln(2\pi) + \frac{n}{2}\ln\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2\right) + \frac{n}{2}.$$

A figura 10 apresenta uma comparação da verossimilhança perfilhada e estimada. Ambas têm o mesmo ponto de mínimo (ponto da MLE), e na vizinhança desse ponto as curvas são coincidentes. Entretanto, à medida que se afasta do ponto de mínimo a curva da verossimilhança perfilada é mais aberta, o que indica um maior grau de incerteza sobre a estimativa da média. A verossimilhança perfilada é mais coerente e realista, uma vez que para cada valor de μ no gráfico, ela busca o MLE da variância assumindo aquele valor de média.

6 O Critério de Akaike na Comparação de Modelos

A utilização da razão da verossimilhança não considera que dois modelos sendo comparados podem diferir bastante no número de parâmetros ajustados. Geralmente, espera-se que um modelo com mais parâmetros tenda a apresentar um ajuste melhor que um com menos parâmetros, pois os parâmetros são os elementos que nos permitem explicar o comportamento dos dados. Pelo *Princípio da Parsimônia*, ou *Princípio de Okham*, entre dois modelos com igual poder explicativo opta-se pelo modelo mais simples, o que geralmente é entendido como o modelo com o menor número de parâmetros.

Um critério baseado da log-verossimilhança negativa que considera o número de parâmetros é o **Critério de Informação de Akaike** (AIC), que penaliza a log-verossimilhança negativa com duas vêzes o número de parâmetros:

$$AIC = -2\ln\left[\mathcal{L}\{\theta\}\right] + 2p = 2\mathbf{L}\{\theta\} + 2p \tag{1}$$

No caso da distribuição Gaussiana o AIC pode ser detalhado como:

$$AIC \;=\; 2\mathbf{L}\{\widehat{\mu}, \widehat{\sigma}\} + 2p$$

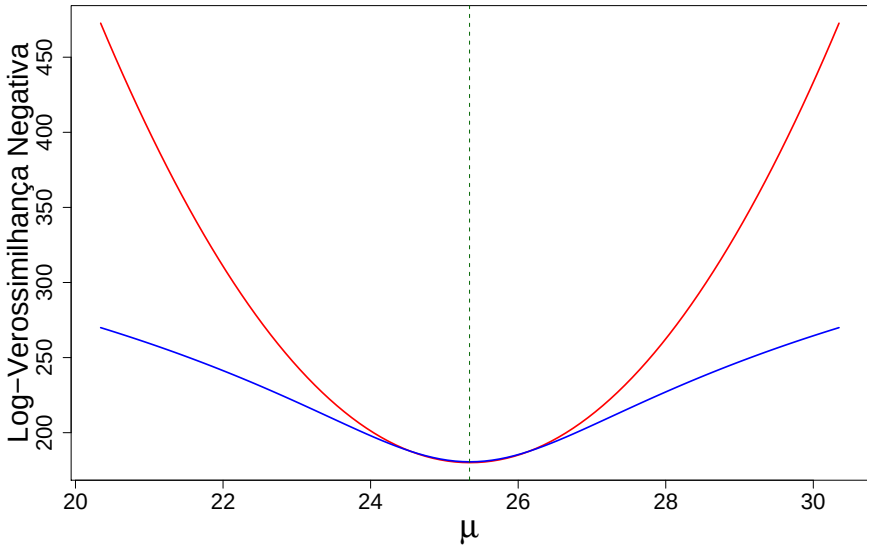

Figura 10: Função de log-verossimilhança negativa *perfilada* (em azul) e log-verossimilhança negativa *estimada* (em vermelho) para a média de uma distribuição Gaussiana, applicada a dados do DAP médio por parcela em floresta nativa do Maranhão.

$$
\begin{aligned}
&= 2\left[\frac{n}{2}\ln 2\pi + n\ln\widehat{\sigma} + \frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{n}(x_i-\widehat{\widehat{\mu}})^2\right] + 2p \\
&= 2\left[\frac{n}{2}\ln 2\pi + n\ln\left(\sqrt{\frac{\sum_{i=1}^{n}(x_i-\widehat{\mu})^2}{n}}\right) + \frac{1}{2\sum_{i=1}^{n}(x_i-\widehat{\mu})^2}\sum_{i=1}^{n}(x_i-\widehat{\mu})^2\right] + 2p
\end{aligned}
$$

Simplificando-se a expressão:

$$
AIC = n\ln(2\pi) + n\ln\left(\sum_{i=1}^{n}(x_i-\widehat{\mu})^2\right) + 2p
$$

No modelo Gaussiano, o AIC é obtido a partir da soma dos desvios quadráticos das observações em relação à média estimada (soma de quadrados do resíduo).

Anexo
MLE para alguns Modelos

A sigla MLE pode designar *estimativas* ou *estimadores*. Entende-se por *estimativa* o valor numérico atribuido a um parâmetro na situação particular de um certo conjunto de dados. Entende-se por *estimador* a expressão ou processo matemático que permite encontrar a estimativa nos casos particulares. A expressão ***as*** MLE se refe às estimativas, a expressão ***os*** MLE se refere aos estimadores.

Nessa secção são apresentados os MLE para alguns modelos de distribuição aleatória.

A Distribuição Gaussiana (Distribuição Normal)

Assumindo que um conjunto de variáveis aleatórias $X_1, X_2, \ldots, X_n$ são independentes e identicamente distribuidas seguindo a distribuição Normal com média μ e variância σ^2 ($X_i \sim N(\mu, \sigma^2), \quad i = 1, \ldots, n$), a densidade conjunta destas variáveis é igual ao produto das densidades marginais. A função de verossimilhança, para uma amostra observada $X_i = x_i$, se torna:

$$\begin{aligned}
\mathcal{L}\{\mu, \sigma\} &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right] \\
\mathcal{L}\{\mu, \sigma\} &= (2\pi)^{-n/2}\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right]
\end{aligned}$$

A função de log-verossimilhança negativa é de forma mais simples:

$$\mathbf{L}\{\mu, \sigma\} = \frac{n}{2}\ln 2\pi + n\ln\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Igualando as derivadas parciais a zero e solucionando para os parâmetros de interesse, obtem-se:

$$\begin{aligned}
\frac{\partial \mathbf{L}\{\mu, \sigma\}}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0 \qquad &\Longrightarrow \qquad \widehat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \\
\frac{\partial \mathbf{L}\{\mu, \sigma\}}{\partial \sigma} = \frac{n}{\sigma} - \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \qquad &\Longrightarrow \qquad \widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}
\end{aligned}$$

O MLE para o desvio padrão é ligeiramente diferente do estimador convencional. O estimador convencional é não-viciada, mas o MLE é consistente, isto é, não-viciado para grandes amostras.

B Distribuição Weibull - 2 Parâmetros

Dado um conjunto de variáveis aleatórias $X_1, X_2, \ldots, X_n$, independentes e identicamente distribuidas de acordo com a distribuição Weibull, com parâmetro de escala β e parâmetro de forma γ, ($X_i \sim \text{Weibull}(\beta, \gamma), \quad \text{i} = 1, \ldots, \text{n}$), a densidade conjunta destas variáveis é igual a produto das densidades marginais. A função de verossimilhança, para uma amostra observada $X_i = x_i$, se torna:

$$\begin{aligned}\mathcal{L}\{\beta,\gamma\} &= \prod_{i=1}^{n} \frac{\gamma}{\beta^\gamma} x_i^{\gamma-1} \exp\left[-\frac{1}{\beta^\gamma} x_i^\gamma\right] \\ &= \left(\frac{\gamma}{\beta^\gamma}\right)^n \left[\prod_{i=1}^{n} x_i^{\gamma-1}\right] \exp\left[-\frac{1}{\beta^\gamma}\sum_{i=1}^{n} x_i^\gamma\right]\end{aligned}$$

A função de log-verossimilhança negativa assume a forma:

$$\mathbf{L}\{\mu,\sigma\} = n\ln\left(\frac{\gamma}{\beta^\gamma}\right) - n\gamma\ln(\beta) + (\gamma-1)\sum_{i=1}^{n}\ln(x_i) - \frac{1}{\beta^\gamma}\sum_{i=1}^{n} x_i^\gamma$$

A derivada parcial em relação ao parâmetro de escala fica:

$$\begin{aligned}\frac{\partial \mathbf{L}\{\beta,\gamma\}}{\partial\beta} = \frac{\gamma}{\beta}\left(-n + \frac{1}{\beta^\gamma}\sum_{i=1}^{n} x_i^\gamma\right) &= 0 \\ \beta^\gamma &= \frac{\sum_{i=1}^{n} x_i^\gamma}{n}\end{aligned}$$

resultando num estimador de máxima verossimilhança que é dependente do parâmetro de forma:

$$\widehat{\beta} = \left[\frac{\sum_{i=1}^{n} x_i^\gamma}{n}\right]^{1/\gamma}$$

A derivada parcial em relação ao parâmetro de forma fica:

$$\frac{\partial \mathbf{L}\{\beta,\gamma\}}{\partial\gamma} = \frac{n}{\gamma} - n\ln(\beta) + \sum_{i=1}^{n}\ln(x_i) - \frac{1}{\beta^\gamma}\sum_{i=1}^{n} x_i^\gamma \ln(\beta) = 0$$

Tomando o valor de β^γ está expressão é simplificada para:

$$\frac{n}{\gamma} + \sum_{i=1}^{n} \ln(x_i) - \frac{\sum_{i=1}^{n} x_i^\gamma}{\sum_{i=1}^{n} x_i^\gamma} = 0$$

a qual não pode ser simplificada mais. Assim, a MLE para o parâmetro da forma ($\widehat{\gamma}$) é obtido solucionando a expressão acima por métodos iterativos. Uma vez que $\widehat{\gamma}$ é encontrado, a estimativa do parâmetro de escala é obtida diretamente.

C Distribuição Weibull - 3 Parâmetros

Tomando-se novamente um conjunto de variáveis aleatórias $X_1, X_2, \ldots, X_n$, independentes e identicamente distribuidas de acordo com a distribuição Weibull, mas agora com três parâmetros: parâmetro de locação α, parâmetro de escala β e parâmetro de forma γ. ($X_i \sim \text{Weibull}(\alpha, \beta, \gamma), \quad \text{i} = 1, \ldots, \text{n}$). Novamente, a densidade conjunta destas variáveis é igual a produto das densidades marginais, logo a função de verossimilhança, para uma amostra observada $X_i = x_i$, se torna:

$$\begin{aligned}
\mathcal{L}\{\alpha, \beta, \gamma\} &= \prod_{i=1}^{n} \frac{\gamma}{\beta^\gamma}(x_i - \alpha)^{\gamma-1} \exp\left[-\frac{1}{\beta^\gamma}(x_i - \alpha)^\gamma\right] \\
&= \left(\frac{\gamma}{\beta^\gamma}\right)^n \left[\prod_{i=1}^{n}(x_i - \alpha)^{\gamma-1}\right] \exp\left[-\frac{1}{\beta^\gamma}\sum_{i=1}^{n}(x_i - \alpha)^\gamma\right]
\end{aligned}$$

A função de log-verossimilhança negativa assume a forma:

$$\mathbf{L}\{\mu, \sigma\} = n \ln\left(\frac{\gamma}{\beta^\gamma}\right) - n\gamma\ln(\beta) + (\gamma - 1)\sum_{i=1}^{n}\ln(x_i - \alpha) - \frac{1}{\beta^\gamma}\sum_{i=1}^{n}(x_i - \alpha)^\gamma$$

A derivada parcial em relação ao parâmetro de escala fica:

$$\begin{aligned}
\frac{\partial \mathbf{L}\{\alpha, \beta, \gamma\}}{\partial \beta} = \frac{\gamma}{\beta}\left(-n + \frac{1}{\beta^\gamma}\sum_{i=1}^{n}(x_i - \alpha)^\gamma\right) &= 0 \\
\beta^\gamma &= \frac{\sum_{i=1}^{n}(x_i - \alpha)^\gamma}{n}
\end{aligned}$$

resultando num estimador de máxima verossimilhança que é dependente do parâmetro de forma:

$$\widehat{\beta} = \left[\frac{\sum_{i=1}^{n}(x_i - \alpha)^\gamma}{n}\right]^{1/\gamma} \tag{2}$$

A derivada parcial em relação ao parâmetro de forma fica:

$$\frac{\partial \mathbf{L}\{\alpha,\beta,\gamma\}}{\partial \gamma} = \frac{n}{\gamma} - n\ln(\beta) + \sum_{i=1}^{n}\ln(x_i-\alpha) - \frac{1}{\beta^{\gamma}}\sum_{i=1}^{n}(x_i-\alpha)^{\gamma}\ln(\beta) = 0$$

Tomando o valor de β^{γ} está expressão é simplificada para:

$$\frac{n}{\gamma} + \sum_{i=1}^{n}\ln(x_i-\alpha) - \frac{\sum_{i=1}^{n}(x_i-\alpha)^{\gamma}}{\sum_{i=1}^{n}(x_i-\alpha)^{\gamma}} = 0 \tag{3}$$

a qual não pode ser simplificada mais.

Em relação ao parâmetro de locação, a derivada parcial fica:

$$\frac{\partial \mathbf{L}\{\alpha,\beta,\gamma\}}{\partial \alpha} = -(\gamma-1)\sum_{i=1}^{n}\frac{1}{x_i-\alpha} + \frac{\gamma}{\beta^{\gamma}}\sum_{i=1}^{n}(x_i-\alpha)^{\gamma-1} = 0$$

Substituindo-se o valor de β^{γ} em (2), obtemos:

$$\begin{aligned}
\frac{n\gamma}{\sum_{i=1}^{n}(x_i-\alpha)} - (\gamma-1)\sum_{i=1}^{n}\frac{1}{x_i-\alpha} &= 0 \\
-(\gamma-1)\left[\sum_{i=1}^{n}(x_i-\alpha)\right]\left[\sum_{i=1}^{n}\frac{1}{x_i-\alpha}\right] + n\gamma &= 0 \\
\left[\sum_{i=1}^{n}x_i - n\alpha\right]\left[\sum_{i=1}^{n}\frac{1}{x_i-\alpha}\right] - \frac{n\gamma}{\gamma-1} &= 0 \\
n(\overline{x}-\alpha)\sum_{i=1}^{n}\frac{1}{x_i-\alpha} - \frac{n\gamma}{\gamma-1} &= 0 \\
(\overline{x}-\alpha)\sum_{i=1}^{n}\frac{1}{x_i-\alpha} - \frac{\gamma}{\gamma-1} &= 0 \qquad (4)
\end{aligned}$$

onde $\overline{x}$ é a média da amostra.

Todas as expressões são dependentes do parâmetro de locação (α) e do parâmetro de forma (γ). Para se obter as MLE torna-se necessário utilizar um processo iterativo envolvendo essas expressões:

1. Inicia-se o processo com um valor arbitrário de α, (α_0) e encontra-se o valor inicial para γ (γ_1) fazendo a expressão (3) convergir por um processo iterativo.

2. Utilizando-se γ_1, obtem-se o valor incial de α (α_1), fazendo a expressão (4) convergir por um processo iterativo.

3. Repete-se os passos 1 e 2 acima, até que os valores obtidos para α e γ deixem de sofrer alterações marcantes. Os valores finais α_n e γ_n são as MLE $(\widehat{\alpha}, \widehat{\gamma})$.

4. Utilizando $\widehat{\alpha}$ e $\widehat{\gamma}$, obtem-se $\widehat{\beta}$ solucionando-se a expressão (2).