Alunos PPGMNE-2009

Estudando Verossimilhança com apoio computacional

Curitiba

1 de Maio de 2009

Alunos PPGMNE-2009

Estudando Verossimilhança com apoio computacional

Orientador:
Paulo Justiniano Ribeiro Jr.

UNIVERSIDADE FEDERAL DO PARANÁ

Curitiba

1 de Maio de 2009

CONTEÚDO

CAPÍTULO 1

INTRODUÇÃO E REVISÃO

1.1 Inferência Estatística

Este texto trata sobre a teoria da inferência estatística, ou pelo menos sobre uma consideravel parte. Para explicar o significado do termo ”inferência estatística”, vamos começar com o que a enciclopedia fala sobre isso.

Estatística *Pode ser definido como o corpo de uma concentração de métodos com uma abstração da informação sintetizada dos dados observados, com o propósito de caracterizar os aspectos do fenômeno de interesse que são relevantes para um particular objetivo. Assim, as estatísticas possuem um amplo campo de aplicações no estudo de todos os fenômenos onde se supõem que algum fator de erro esteja presente em adição em alguns fatores sistemáticos cuja os efeitos estão em destaque; Acontece que o resultado destes fatores, que podem ser descritos por alguma “lei matemática”, é sobreposto por um componente que transforma esta lei em “regularidade estatística”. Para examinar as características essênciais do fenômeno, a teoria estatística faz um amplo uso de técnicas de teoria das probabilidades, especialmente quando as obesrvações disponíveis não cobrem todas as possibilidades do fenômeno sob observação. Neste caso, o problema de “inferência estatística” chega, onde o objetivo é inferir sobre características de tod ocom a observação de uma parte disto.*

Portanto, a fim de realizar uma investigação estatística, é necessário estabelecer qual é o fenômeno exato sob consideração, e estabelecer explicitamente quais são as características observaveis, chamadas variáveis relevantes para nosso propósito . Deve-se também especificar um subconjunto de casos nos quais queremos focar. Por exemplo, suponha que a variável de interesse é a altura das pessoas. Talvez considerar o total geral(para toda população humana) ou focar em certo contexto geográfico (como uma nação) e/ou em certo tempo. Este conjunto de casos é chamado de população, e cada caso é chamado

de indivíduo; algumas vezes termos casos e unidades são usados. Esta derivação desta terminologia é histórica, a estatística originalmente se desenvolveujunto e em interação com a demográfia; hoje em dia, os termos “indivíduo e população” não se referem necessariamentea seres humanos, quando eles são usados no sentido técnico. Em alguns estudos, a população toda é examinada, isto é chamado de censo. Em outros casos, somente uma *amostra* i.e um subconjunto da população é examinado, mas o objetivo continua sendo investigar toda a população. *Inferência Estatística* constitue a operação através da qual a informação fornecida pela amostra é usada para tirar conclusões sobre características da população. Este tipo de procedimento é um passo a diante no que diz respeito a aproximação estatística puramente descritiva, que simplesmente tenta resumir as características mais relevantes dos daos observados. O texto presente é sobre a teoria e os métodos que direciona as operações de inferência estatística.

Exemplo 1.1. Para ilustrar os conceitos introduzidos, vamos introduzir um exemplo bastante simples. Uma indústria que produz bombas hidraulicas compra de diferentes fornecedores muitos componetes necessários para sua produção. em particular, juntas de plástico usados para unir elementos mecânicos são fornecidos por uma companhia em lotes de 5000. Obviamente, o comprador precisa avaliar a qualidade das juntas fornecidas a fim de eliminar, ou ao menos reduzir significativamente, a possibilidade de uma junta com defeito ser usada. Sendo que o custo de reparo de uma banheira encontrada com defeito seja muito maior do que o custo da junta em si, poderia ser desejável testar todas as juntas colocando-as para trabalhar sob uma pressão por um curto tempo, antes de adotar para produção. Por outro lado, o tempo necessário para criar e executar o teste das juntas representa um custo.
Uma saida para não examinar todas as juntas fornecidas mas apenas algumas, digamos 50, e usar a informação fornecida por estes testes para avaliar o número de juntas com defeito no lote todo, e para decidir sobre possibilidade de devolver o lote ao fornecedor se o resultado não for satisfatório. Ao fazer isso, nós temos que considerar que o subconjunto testado no geral não irá conter a quantidade de juntas com defeito exatamente na mesma proporção que o lote. Neste exemplo, podemos considerar o lote de 5000 juntascomoa população de interesse ,e cada junta como um individúo. Neste momento, nosso interesse no indivíduo esta restrito a um aspecto em específico, sobre se está “conforme” ou “não conforme” as especificações . As características observadas nos elementos daa amostra não são de interesse direto, exeto como um meio de fazer inferência a respeito de características da população como um todo.
Vamos agora discutir o motivo de frequentemente examinarmos apenas uma amostra ao invéz de toda a população, o que é aperentemente é preferível, uma vez que evitaria qualquer indeterminação em nossas conclusões (quando realizado com plena exatidão).

- O custo da inspeção na população toda pode ser muito alto, seja pelo grande número de indivíduos ou pelo alto custo de cada inspeção. Mesmo quando recursos economicos possam estar disponíveis para o censo da população, o dano causado por um estudo amostral pode não exeder o custo do censo.

- Uma vez que a investigação completa da população frequentemente leva muito tempo, isso pode facilmente entrar em conflito com pedidos urgentes.
  Por exemplo, o censo geral de uma população humana é feito em intervalos muito longos (normalmente a cada dez anos), e os resultados destas pesquisas são publicados muito depois. Por outro lado, existem muitos problemas econômicos e socias, tais como os relacionados ao custo de vida ou desemprego, para os quais informações defasadas não são aceitáveis , uma vez que o

governo e as agências devem tomar suas decisões muito mais rapidamente.

- Em muitos casos, existe uma população virtual que é efetivamente infinita esta situação ocorre quando casos do fenomeno sob estudo pode ser replicado tantas vezes quanto se queira . Suponha, por exemplo, que queiramos estudara capacidade de uma droga baixar a prssão sanguìnea em seres humanos; a população relevante é então formadas por todas as pessoas a quem a droga poderia possivelmente ser dada, i.e. toda raça humana, presente e futura. Claramente apenas em estudo amostral e exquecível aqui:
  Observe que a maoir parte dos experimentos científicos e tcnológicos caem nesta situação.

- Em alguns casos, a inspeção de unidades amostrais destrói a propria unidade. Por exemplo, a análise da duração de um lote de lâmpadas envolve deixa-las ligadas por muito tempo, possivelmente até queimarem.
  Contudo, no fim da inspeção,as lâmpadas estão comprometidas. Sea população for finita, claramente o censo da população é invariável, a não ser que a população depois do teste seja irrelevante.

1.2 Amostragem

É evidente que a amostra deve representar, tanto quanto possível, as características da população, afim de permitir a expansão das características da amostra para a população. Este requisito algumas vezes é referido dizendo que amostra deve ser representativa.
Para ilustrar este ponto, suponha por exemplo que membrosde uma sociedade cientifica amadora desejam levantar uma pesquisa sobre atitudes e comportamentos dos habitantes da cidade deles a cerca de problemas raciais.
Inicialmente, os membros da sociedade podem pensar em recorrer aos seus conhecidos , e administrar a estas pessoas m questionário adequado. Isto é , entretanto fácil de ver porque esta conduta é grsseiramente inadequada. Os membros da sociedade formam um grupo de pessoas compartilhando, para alguma projeção, características relevantes tais como, culturais, status sociais, e possivelmente inclinação política. Estes elementos em comum podem ser esperados entre seus parentes, amigos, vizinhos e colegas, mesmo a mais fraca conexão. Contudo, este procedimento de seleção de elementos da amostra tendo a selecionar integrantes com características relacionadas, que em adição não são independentes quanto ao tópico do questionamento. Em poucas palavras, este método poderia gerar uma amostra não representativa.
Contudo, mpara obter uma amostra adequada, o grupo poderia ser selecionado para inclusão na amostra nas bases da característica que são independentes das características em estudos.
Por outro lado, isso é um tanto difícil, e em alguns casos virtualmente impossível, escolher características que envolvam esta propriedade.
Uma solução para o problema é solecionar os elementos da amostra aleatóriamente, sendo por definição independente de qualquer característica. Para visualizarum processo de seleção aleatório, é conveniente

imaginar uma urna com bolas numeradas. Cada uma dessas é associada a um elemento da população. Bolas são tiradas da urna, e elementos asociados a cada bola irão continuar a amostra *****.
Mesmo este regime super simples nos leva duas variantes, com ou sem reposição das bolas. Além disso, aleatóriamente não siginfica selecionar com a probabilidade constante; Contudo, diferentes variantes chegam a associar diferentes probabilidades de seleção para as bolas. Todas essas variações ocasionam diferentes tipos de amostragem, das quais apenas a mais simples (mas não a mais importante) será discutida aqui com atenção especial ao caso de uma população "infinita". Nós antecipamos que o procedimento amostral adotado produz efeito no procedimento de inferência.

1.3 Estatística e Probabilidade

Nós já apontamos que a inferência é um assunto com um certo inevitável grau de incerteza,pois ela é completamente irrealista, rxeto em circunstâncias exepcionais, imaginar que a amostra representa exatamente todas as características da população.
O grau de proximidade entre as características da amostra e da população é uma entidade variável que pode ser estudado com ferramentas de teoria de probabilidade. Esta área da matemática joga, então, uma tacada essêncial no desenvolvimento da teoria estatística, assim algumas vezes se torna difícil se estabelecer fronteiras entre as duas disciplinas.
Observe que é na verdadeo fatoda seleção de amostrade acordo com um esquema aleatório que permite análise matemática do grau de correspondência entre amostra e a população. Assim uma análise não poderia usar diferente critério de seleção, assim como a seleção subjetiva das unidades amostrais decididas pelo experimentador, ou uma auto-seleção da amostra, i.e., situação onde as unidades se colocam a frente para entrevista.
Para começar a explicar como a teoria da probabilidade entra na teoria de inferência estatística vamos voltar para nossa amostra de 50 juntas de um lote de 5000.
É conveniente pensar no lote como uma urna com 5000 bolas numeradas, onde uma proporção θ desconhecida são pretas, e a proporção restante são brancas. Das 50 bolas retiradas, um certo número Y são pretas; suponha $Y = 4$.
Para estimar o total X de bolas pretas na urna, ou equivalente para estimar a proporção θ, é natural considerar a igualdade.

$$4 \div 50 = x \div 5000 \tag{1.1}$$

que nos leva a

$$\theta = 4/50$$

- Existem alternativas razoáveis por este raciocínio, por esta escolha de θ?

- Quão preciso é a presente estimativa de θ, comparado ao verdadeiro valor? Podemos obter um intervalo dos valores plausíveis de θ?

- Se o fornecedor garantiu que o lote contém uma proporção de defeitos não maior que 5%, existe evidência suficiente para afirmar que o lote seja inadequado e se retornar para o fornecedor ? Leve em consideração que o fornecedor pode facilmente alegar que nós simplismente retiramos uma amostra ”ruim”que 5% não é diferente de 8%, e que o teste no lote todo poderia certamente resultar que existem menos de 5% de defeitos.

- A equação 1.1 é certamente um critério razoavel, mas pode ser usada somente para avaliar o tamanho de uma sub-população, X neste caso, ou equivalentemente uma proporção, θ neste caso. Podemos construir um critério geral usavel em situações praticas,mesmo para uma situação mais complicada que nosso teste de juntas? Um exemplo de um problema estatístico não-trivial é o seguinte: em um estudo médico, um gupo de pacientes afetados por um certo tipo de câncer passando por quimioterapia usando uma combinação apropriada de diferentes drogas. Idealmente, as doses de drogas deveriam ser suficientemente altas para combater o tumor, mas não tão alta a ponto de gerar toxidade (definido como colocar a vida do paciente em risco); Se necessário, os ciclos do tratamento são repetidos até que a doença desapareça. Depois que todos os pacientes são tratados e exonerados os dados são analisados para estudar a relação entre a intensidade de toxidade e a dosagem da droga, onde o alvo é identificar a dosagem miníma para uma mistura otimizada de drogas necessárias para remover o tumor, evitando toxidade desnecessária de conhecidos fatores concominantes como sexo do paciente, idade, estágio do tumor, tipo histologico do tumor também poderiam ser incluidos no estudo. Claramente, um problema desta espécie não pode ser trabalhado com o resultado da equação 1.1.

Retornando para nosso exemplo simples de controle de qualidade, do ponto de vista da teoria de probabilidade, da relação entre a proporção desconhecida θ e os números observados de defeitos, Y. Neste contexto, o número observado Y pode ser considerado como um valor amostrado variavel aleatória Y a qual a distribuição de probabilidade depende de θ e do esquema amostral adotado. Especificamente, a distribuição de Y é binomial ou hipergeométrica, dependendo se a amostra é com ou sem suposição,desde que a população seja finita. Contudo, estas distibuiçoes são bem próximas neste caso, por causa da pequena fração amostrada, 50 tiradas em 5000. Assim, por simplicidade, podemos restringir a discussão ao caso binomial.
Uma priori com respeito a amostra, a probabilidade que Y resulte o valor observado Y é

$$P\{Y = y\} = \left(\begin{array}{c} 50 \\ y \end{array} \right) \theta^y (1-\theta)^{50-y} \tag{1.2}$$

Onde Y é um inteiro entre 0 e 50. O conjunto de possíveis valores de θ é o conjunto de razões do tipo K/500, onde $KE0,1,...,5000$; Contudo este conjunto pode ser razoavelmente bem aproximado pelo conjunto de todos os números em $[0,1]$ Como θ varia em $[0,1]$, a aproximação 1.2 abrange uma família toda de distribuições, apesar de termos apenas um valor de θ que atualmente controla a geração de dados.
Desde o ponto em diante, inferência pode ser considerada como uma operação distrindo o parâmetro θ que identifica a verdadeira distribuição de probabilidade de Y dentro da família de distribuições 1.2

Consequêntemente, de agora em diante, nós não iremos mais falar sobre populações, mas preferivelmente sobre variaveis aleatórias e a inferência sobre os parâmetros das distribuições é entendido como uma conexão entre populações e variáveis aleatórias.

1.4 Alguns Problemas Típicos

O foco de todo este texto é , na verdade, para discutir quesatões que aprareceram na seção 1.3. Apesar de uma discussão mais sistemática ser representada nos próximos capítulos, é util introduzir aqui algumas idéias chaves inicialmente em um padrão informal.

1.4.1 Verossimilhança e estimativas

Uma das questões na seção 1.3 perguntava por um critério geral para construção das estimativas pode ser informalmente tratavel como um valor plausivel , digamos $\hat{\theta}$, no lugar do valor desconhecido θ.
Uma vez que o dado amostral é dado por, $Y = 4$ em nosso exemplo, 1.2, é uma função de θ, a saber

$$L(\theta) = \left(\begin{array}{c} 50 \\ 4 \end{array} \right) \theta^4 (1-\theta)^{46} \quad (0 \leq \theta \leq 1)$$

Esta função é mostrada na figura 1.4.1

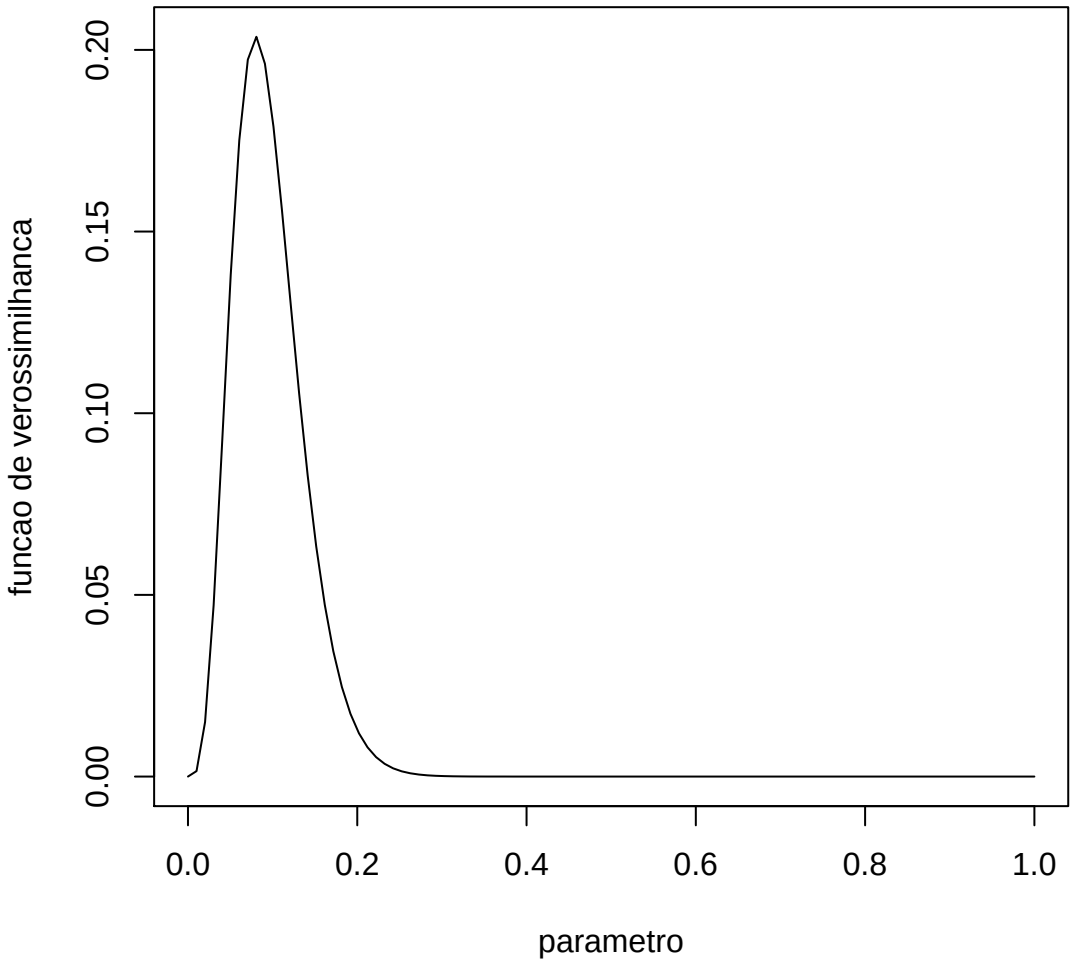
Esta expressão fornece, como função de θ, uma probabilidade a priori da observação que tem sido realmente observador. Contrariamente, isso pode ser considerado como uma medida de concordância entre qualquer valor nominal de θ e as observações, consequentemente explicando o nome verossimilhança dado para $L(\theta)$. Note que, apesar de todos os valores de $L(\theta)$ serem probabilidades a função $L(\theta)$ em si não é uma distribuição de probabilidade.
É então um tanto natural selecionar o ponto de maior verosimilhança, caso uma estimativa de θ seja requisirada. Um simples exercício matemático mostra que o máximo de $L(\theta)$ ocorre em

$$\hat{\theta} = 4/50,$$

Isso é chamado, por razão obvia de estimativa de máxima verossimilhança
Podemos imediantamente expandir o método para uma situação um pouco mais geral considerando um número genérico n de elementos amostrados, com um número de defeitos observados denotados por Y, um inteiro entre 0 e n. Então a variável aleatória Y ossui distribuição binomial com indice n e parâmetro θ, e a verossimilhança é agora

$$l = L(\theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y} \tag{1.3}$$

que corresponde a estimar


funcao de verossimilhanca
0.00
0.05
0.10
0.15
0.20
0.0
0.2
0.4
0.6
0.8
1.0
parametro

$$\hat{\theta} = y/n \tag{1.4}$$

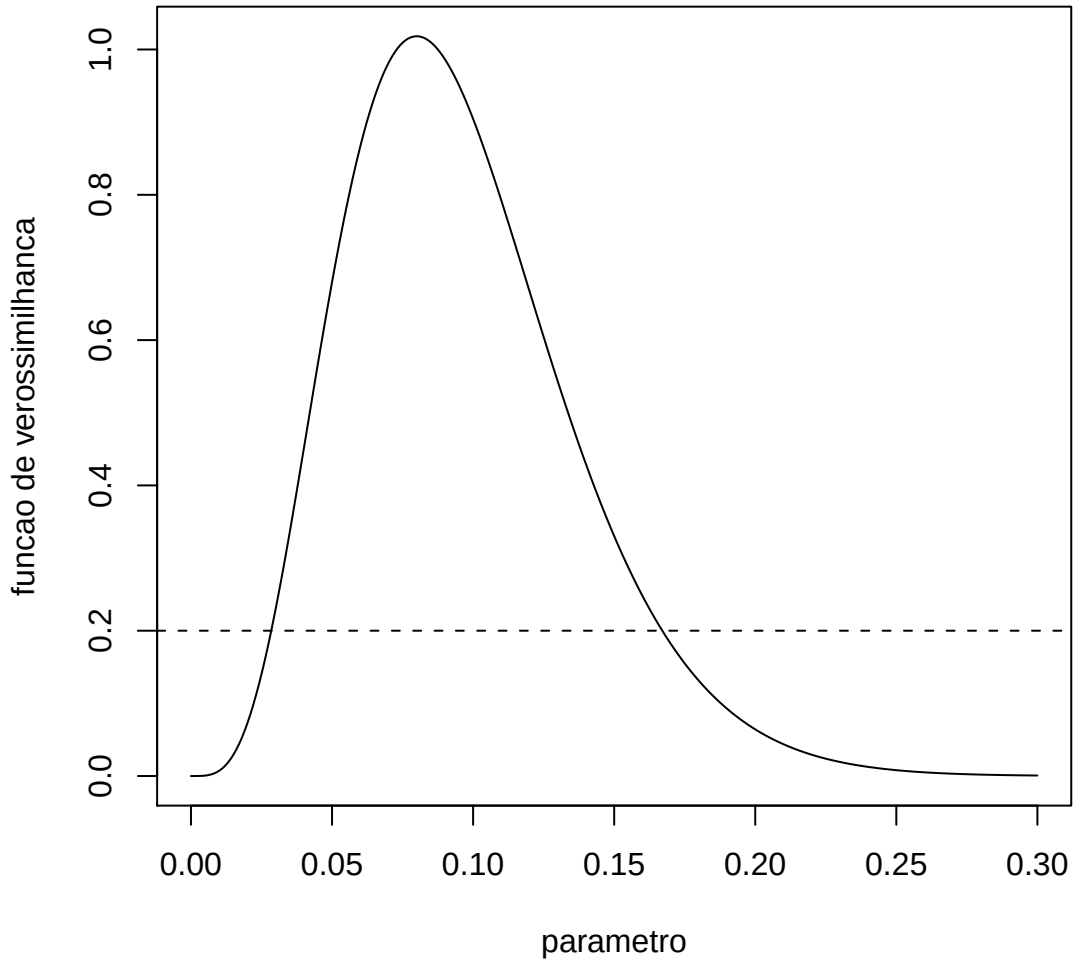
Apesar de o critério de maximização da verossimilhança não ter produzido nada diferente do que obtivemos antes por um mais simples, mais direto argumento, a presente critério possui vantagem que pode ser aplicado a problemas onde 1.1 não faz sentido, na verdade, isso pode ser aplicado onde passamos escrever a verossimilahnça, assim , cobrindo um número enorme de situações práticas, como será demostrado nos capitulos seguintes.

1.5 Estimação de intervalo e teste de Hipotese

Num primeiro momento, pode parecer que a etimação é o únio atributo da análise de dados.
A discussão na seção 1.3 mostra o erro desta crença, levantando, un número adicional de questões, cuja a discussão nós gostariamos de iniciar.
Sabendo que não é esperado que a estimativa tinha coincidência com o verdadeiro valor do parâmetro, é natural olhar para um intervalo que deve ser presumivelmente conter o parâmetro. No exemplo anterior com $\hat{\theta}4/50$ nós olhamos para um intervalo em termo de $4/50$ no qual é plausível que θ esteja. Como um intervalo no chamado intervalo estimado de θ.
Um diferente, mas relacionado, problema aparece quando é necessário comparar a estimativa, ou mais genericamente as observações, com um valor de referência do parâmetro. Por exemplo, na discussão da seção 1.3, a questão considerada de querer decidir se o dado fornecia evidencia a favor ou contra a alegação de que a proporção de defeitos no lote não exeda 5%. Problemas deste tipo são chamados de teste de hipóteses.
Vamos explorar, previamente seguindo informalmente, por que intervalo de estimação e teste de hipóteses são problemas similares, na compreensão. Se nós tivermos selecionado algum intervalo estimado (θ_1, θ_2) para θ então qualquer valor de θ contido neste intervalo e de alguma forma compatível com a evidência empirica contida na amostra. Por outro lado, se o hipotético valor $\theta = 5\%$ considerado no problema de testes de hipóteses está no intervalo $(\theta_1\theta_2)$, então, a hipótese sobre θ não é conflitante com o dado.
Pra escolher o intervalo $(\theta_1\theta_2)$, nós novamente fazemos uso da função de verossimilhança. Se $4/50$ é o ponto preferido por ter a maior verossimilança, então também é verdade que outros pontos no intervalo $[0,1]$ são mais ou menos plausíveis dependendo de seu valor de verossimilhança. Esta observação implica que o critério de inclusão de um ponto no intervalo de confiança deve ser baseado no valor da verossimilhança naquele ponto, comparado ao valor máximo. Somos então levados a considerar a verossimilhança relativa.

$$\tilde{L}(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$$

Que varia em $[0,1]$. Figura 1.5 representa a verossimilhança relativa correspondente a verossimilhança da Figura 1.4.1, dicartando a proporção do intervalo $[0,1]$ onde a verossimilhança em si é descartada. Sendo $L(\hat{\theta})$ constante com respeito a θ, o gráfico é essencialmente os mesmos que o anterior, exeto pelo fator escala.



Se tivermos que escolher um intervalo para θ, poderiamos selecionar um conjunto de pontos satisfazendo a condição.

$$\tilde{L}(\theta) > 1/2$$

Ou outra constante semelhante, por exemplo

$$\tilde{L}(\theta) > 1/5$$

Para esta escolha a Figura 1.5 mostra o intervalo de confiança correspondente.
Considere agora a questão de decidir sobre o valor referência $\theta = 5\%$. A verossimilhança relativa

para este ponto é $0,668$, um valor razoavelmente alto, então 5% não parece em conflito com o dado observado. Uma inspeção na Figura 1.5 mostra a conexão entre intervalo de confiança e teste de hipóteses: a plausividade do valor 5% deriva da verossimilhança relativa, e está em correspondencia com o fato de que 5% esteja no intervalo de confiança.

Isto é interessante para examinar o efeito do aumento do tamanho da amostra, consequentemente "o montante de informação", sem mudar a estimativa do parâmetro. Em nosso exemplo, suponha que ambos, o tamanho da amostra e quantidade de defeitos, são dobrados. Apesar de $\hat{\theta}$ não mudar, a função de verossimilhança e a verossimilhança relativa mudam; A Figura 1.5 mostra o delineamento da verossimilhança relativa para ambos os casos $n = 50$ e $n = 100$. Para a nova função da verossimilhança relativa, valores de θ diferentes de $\hat{\theta}$ são agora menos plausiveis; a nova função de verossimilhança é dita ser "mais concentrada" em torno de $\hat{\theta}$.

Em particular a verossimilahnça relativa em $\theta = 5\%$ vai de $0,668$ para $0,446$. Apesar de não ser extremamente baixo, o novo valor é um tanto menos razoavel que o anterior. A questão é séria ou não este decréscimo de confiança no valor de 5% suficiente para rejeitar com o valor plausivel; Este problema é essencialmente relatado como uma escolha do valor limitante para a verossimilhança relativa, $1/5$ na discussão acima e na Figura 1.5. Esta discussão será ampliada nos próximos capítulos, por enquanto, estamos restritos a colocar de lado alguns elementos básicos.

1.6 Princípio da Amostragem por repetição

Até agora a motivação dos principios e dos métodos dicutidos foram inteiramente intuitivos. Isto é, entretanto, possível examinar estes métodos por uma perspectiva diferente, estudando suas propriedades matemáticas.

Fazendo isso, adotamos o critério de considerar estimativas e outras quantidades relacionadas as técnicas estatísticas como valores amostrados de variaveis aleatórias. Por exemplo a estimativa $\hat{\theta} = y/n$ depende de Y, que é um valor amostrado de variavel aleatória; $\hat{\theta}$ é da mesma natureza. Este ponto de vista faz chegar no principio da amostragem por repetição que consiste em considerar $\hat{\theta}$ como uma quantidade amostrada variavel aleatória e no estudo das propriedades desta variavel aleatória.

A palavra repetição surge da seguinte consideração.

Se a amostra for retirada, um grande número de vezes e, para cada amostra, a etimativa de máxima verossimilhança for computada, então cada valor indivídual da estimativa poderia realmente ser considerado como uma amostra de variável aleatória. No exemplo discutido até agora, o lote específico de juntas poderia ser um de um fornecimento regular; Assim essa situação poderia cair facilmente dentro do princípio da amostragem por repetição. Em muitos outros casos, esta replicação de amostras não acontecem realmente, mas ainda argumentamos como se ocorressem.

Adotando o principio da amostra por repetição é uma escolha completamente separada do critério que foi produzido pelo método estatístico sob consideração; o mesmo principio pode ser aplicado para técnicas não relacionados com a verosimilhança, como métodos podem ser adotados puramente de considerações quantitativas , não envolvendo o príncipio da amostragem por repetição.

Para demostrar como o principio da amostragem por repetição funciona na prática, considere 1.1 onde Y é amostrado de uma amostra aleatória Y com distribuição binomial com indice N e o parâmetro θ. Usando resultados elementares da teoria de probabilidades, imediatamente obtemos.

$$E[\hat{\theta}] = \left[\frac{Y}{n}\right] = \frac{E[Y]}{n} = \theta$$

Que diz que em média $\hat{\theta}$ é concentrado em θ, qual seja este valor claramente, nós não temos certeza de que o valor observado de $\hat{\theta}$ seja igual a θ, de fato na maioria dos casos isto não pe verdade, mas é razoável saber que o método não possui um vício sistemático.
Para saber o grau de precisão de estimativa, nós podemos considerar uma medida de variabilidade de $\hat{\theta}$, em particular.

$$var\left[\hat{\theta}\right] = var\left[\frac{Y}{n}\right] = \frac{var[Y]}{n^2} = \frac{\theta(1-\theta)}{n}$$

Uma importante implicação deste resultado é que a variância vai para 0 quando $n \Rightarrow \infty$, uma característica altamente desejada. Se uma quantificação numérica da variância acima for necessária, é obtida substituindo θ pelo valor observado $\hat{\theta}$ na expressão final.

1.7 Estatística e Problemas reais

A teoria da inferência estatística é sobre príncipios gerais e critérios que motivam certas construções matemáticas, em particular métodos que possam ser usados para abranger problemas encontrados no mundo real.
O caminho lógico desses princípios básicos para solução de problemas reais é extremamente longo, e cruza diversos territórios. A primeira parte é inteiramente dentro do domínio da matemática no estágio final, deve-se entrar no mundo da evidência empírica.
Este texto abrange apenas o primeiro estágio, exeto por alguns componentes resumindos em outros aspectos. Quando desejamos mover para aplicação de métodos estatísticos em problemas reais, como temos que a tarefa de unir os dois mundos da dedução formal e da evidência empírica, este é o campo da estística aplicada.
No decorrer dos anos, esttísticos aplicados desenvolveram um amplo apanhado de técnicas para ajudar esta operação. Algumas dessas ferramentas são extremamente poderosas e sofisticadas ao ponto de envolver métodos de inteligencia artificial. Contudo, estas ferramentas não são auto-suficientes para resolver estes problemas, uma vez que muitas aplicações do métodos de inferência estatística também envolvem algum desentendimento do fenômeno sob consideração.
Nos casos mais simples, o conhecimento comum pode bastar, nos casos menos triviais, interação com o especialista será necessária. Uma integração frutífera de diferentes contextos acadêmicos não é, no geral, algo fácil de se alcançar. Isso requer de ambos ods lados, estatístico e pesquisador, uma mente aberta para entender e aceitar os princípios e métodos de uma área diferente.
Trabalhando a estatística sozinho ou em parceria com algum pesquisador, é muito importante que os métodos estatísticos sejam aplicados levando em conta os princípios básicos do contexto da área.

Em algumas aplicações, a motivação da análise estatística é mais pragmática do que exploratória do fenômeno, e muitas vezes não existe um “contexto da área” próprio; Contudo, ao menos alguma forma de informação acerca do fenômeno sob estudo está sempre virtualmente presente. De qualquer forma alguma informação está disponível e o estatístico deve fazer uso dela, evitando a tentação de uma manipulação numérica cega dos dados.

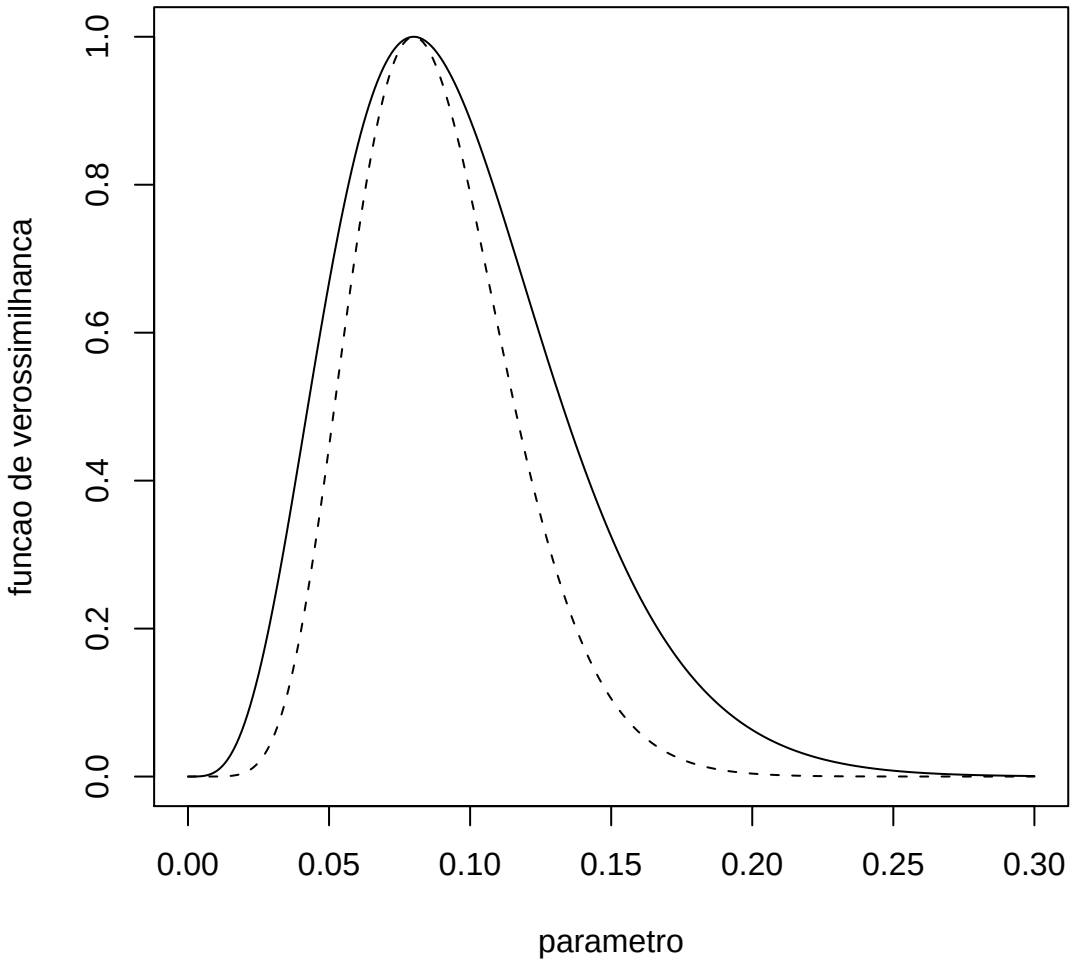


Figura 1.1: Duas funções de verossimilhança relativa associada com $Y = 4, n = 50 e Y = 8, n = 100$

CAPÍTULO 2

INTRODUÇÃO

Este curso é sobre inferência estatística, ou seja, aprender o máximo possível com um particular conjunto de dados, sobre o real mecanismo ou processo que o gerou. Nos vamos considerar problemas envolvendo modelos probabilísticos complexos, potencialmente com muitos parâmetros desconhecidos. Assim a classe de problemas que nos vamos estudar é diferente daqueles vindos de M232 porque aqui nos vamos ter muitos parâmetros desconhecidos, e ainda difere daquela vinda de M235, porque nos não nos restringiremos a uma simples classe de modelos. Nos vamos falar um uma simples e unificada abordagem para fazer inferência baseada na função de verossimilhança e suas propriedades, isto é, a abordagem que gerou o termo Inferência baseada em verossimilhança.

2.1 Exemplos motivacionais

Exemplo 2.1. $X_1, \ldots, X_n$ são i.i.d v.a com distribuição $N(\theta, 1)$, onde θ é um parâmetro desconhecido. Qual é o melhor estimador de θ, se a amostra observada é $x_1, \ldots, x_n$ qual a melhor estimativa ?

Este problema envolve um simples parâmetro e variáveis i.i.d, assim foi mostrado em M232, que o melhor estimador é $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$ com um intervalo de confiança de 95% dado por $\bar{x} \pm \frac{1.96}{\sqrt{n}}$, com suas respectivas estimativas. Relembre a diferença entre estimador e estimatia, um estimador é uma v.a, e uma estimativa é um número. Assim, o estimador de um intervalo de confiança de 95% é um intervalo aleatório que contém o valor verdadeiro de θ com probabilidade 0.95, ou seja, repetindo-se a avaliação da amostra para esta variável, os intervalos que contém o verdadeiro valor de θ com probabilidade 0.95, onde a estimativa é um intervalo que pode conter o verdadeiro valor de θ ou não.

Exemplo 2.2. $X_1, \ldots, X_n$ são i.i.d v.a com distribuição de X_i sendo $N(\alpha + \beta_i, \sigma^2)$ para $i = 1, \ldots, n$ onde $\theta = (\alpha, \beta, \sigma)$ é um vetor de parâmetros desconhecidos. Suponha que estamos interessado em encontrar uma aproximação para θ. Como o problema é multi-parâmetros, ou seja, θ é um vetor, isto esta dentro

do escopo de M232, mais desde que o modelo probabilístico seja:

$$X_i = \alpha + \beta_i + \varepsilon_i$$

onde $\varepsilon_i \sim N(0, \sigma^2)$, ou seja, uma regressão linear com erros normais. Este problema foi resolvido em M235 dando os melhores estimadores para o vetor de parâmetros θ.

Exemplo 2.3. $X_1, \ldots, X_n$ são i.i.d com distribuição $N(\mu, \sigma^2)$, onde $\theta = (\mu, \sigma^2)$ é um vetor de parâmetros desconhecidos. Suponha que os $X_i's$ correspondem a mensurações de algum componente manufaturado, que falha quando combinado com outro componentem, se este é maior do que um nível u. Em vez de estar interessado em encontrar θ aproximadamente, aqui a inferência estatística é necessária para uma combinação de parâmetros:

- Se u é fixo, então a probabilidade do componente ser inútil é, $P(X_i \geq u) = p$, é o parâmetro de interesse, mais
$$p = 1 - \phi(\frac{u-\mu}{\sigma}) = g_1(\theta)$$
onde $g_1(\theta)$ é uma função de θ.

- Se u pode ser designado somente dado o valor da probabilidade de falha, chamada p_1, então o parâmetro de interesse é u satisfazendo
$$u = \mu + \sigma^{-1}\Phi{-}1(1 - p_0) = g_2(\theta)$$
onde $g_2(\theta)$ é uma função de θ.

Exemplo 2.4. $X_1, \ldots, X_n$ são v.a independentes com distribuição sendo

$$N(\alpha + \sum_{i=1}^{n} \beta_j w_{i,j}, \sigma^2)$$

para $i = 1, \ldots, n$ onde $w_{i,j}$ são covariáveis (variáveis explanatórias), $\theta = (\alpha, \beta_1, \ldots, \beta_d, \sigma)$ é um vetor de parâmetros desconhecidos. Suponha X_i representa a idade de morte de uma pessoa i com atributos $w_{i,1}, \ldots wi, d$ correspondente (fuma ou não, sexo, peso, etc), ou seja,

$$w_{i,1} = \begin{cases} 0 & \text{se a i-ésima pessoa fuma} \\ 1 & \text{se a i-ésima pessoa não fuma} \end{cases}$$

Consequentemente, os parâmetros $\beta_j, j = 1, \ldots, d$ mostra o impacto no tempo de vida, destes fatores explanatórios. Um médico pode estar interessado no impacto do fumo. Assim, β_1 mostra o impacto que o fumo tem sobre o tempo de vida, dado todos as outras variáveis explanatórias, mais o interesse é só em β_j. Dado a estimativa de θ é sua incerteza o que pode ser dito apenas sobre β_1 ?

2.2 Instalação geral

Suponha que $X_1,\ldots,X_n$ são v.a observada tendo distribuição conjunta que depende de parâmetros desconhecidos $\theta = (\theta_1,\ldots,\theta_2)$, que tem espaço paramétrico Ω, assim $\theta \in \Omega$. Tomada $\mathbf{X} = (X_1,\ldots,X_n)$ a distribuição conjunta de $\mathbf{X}$ é dada para variáveis aleatórias contínuas por

$$f(x;\theta) \text{ a função densidade de probabilidade conjunta de } \mathbf{X}$$

e para v.a discreta por

$$p(x;\theta) \text{ a função de probabilidade de } \mathbf{X}$$

Esta função avaliada nos dados observados $\mathbf{x} = (x_1,\ldots,x_n)$ (uma realização de $\mathbf{X}$) é chamada de função de verossimilhança de θ. Como esta função é uma função estritamente de $\mathbf{x}$ e θ, isto pode ser denotado por $L(\theta,x)$, mais os dados $\mathbf{x}$ são conhecidos e fixos, assim, nos estamos interessados somente em como varia a verossimilhança com θ, isto gera uma notação simplificada $L(\theta)$. Consequentemente a função de verossimilhança $L(\theta)$ é

$$L(\theta) = \begin{cases} f(x;\theta) & \text{caso contínuo} \\ p(x;\theta) & \text{caso discreto} \end{cases}$$

Para evitar repetições, na discussão geral da função de verossimilhança nos usamos $f(x;\theta)$ irrespectivamente, para variáveis contínuas e discretas. Nos ainda vamos trabalhar com o log da função de verossimilhança, a log-verossimilhança, $l(\theta)$ é

$$l(\theta) = \log L(\theta)$$

Como log é um função monótona

$$L(\theta_1) > L(\theta_2) \text{ se e somente se } l(\theta_1) > l(\theta_2) \tag{2.1}$$

1. Nos podemos pensar $L(\theta)$ e $l(\theta)$, como uma mensuração do quanto os dados suportam os valores de θ, como sendo os valores verdadeiros do vetor paramétrico, de outra forma, podemos pensar em quanto "plausível"que θ tenha gerado os valores observados.

   - Se $\mathbf{X}$ é discreta, para cada θ, $L(\theta)$ da a probabilidade de observar $\mathbf{x}$ se θ for o verdadeiro valor do parâmetro.
   - No caso contínuo $L(\theta)$ é a probabilidade do elemento

$$\lim_{\delta x \to 0} \frac{P(x_i \leq X_i < x_i + \delta x_i, i = 1,\ldots,n;\theta)}{\delta x_1,\ldots,x_n}$$

2. $\frac{L(\theta_1)}{L(\theta_2)}$ é uma mensuração da verossimilhança relativa de θ_1 e θ_2. Usualmente, a verossimilhança relativa de parâmetros e estudada por examinar $\log \frac{L(\theta_1)}{L(\theta_2)} = l(\theta_1) - l(\theta_2)$, ou seja, a diferença em log-verossimilhança. O valor absoluto da verossimilhança tem um importante papel para comparar diferentes θ.

3. Um importante caso especial para o qual a função de verossimilhança é facilmente obtida é quando

$X_1, \ldots, X_n$ são i.i.d v.a sendo assim

$$L(\theta) = \prod_{i=1}^{n} f(x;\theta) \quad \theta \in \Omega$$

Outro caso simples é quando $X_1, \ldots, X_n$ são independentes v.a mais não são identicamente distribuídas, ou seja, X_i tem função densidade de probabilidade $f_{x_i}(x_i;\theta)$ então

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) \quad \theta \in \Omega$$

Dado esta definição a verossimilhança parece apreciável, para trazer bons estimadores para o verdadeiro valor de θ, sendo que este valor de $\theta \in \Omega$ que maximiza a verossimilhança $L(\theta)$. Este estimador, $\bar{\theta}$, é denominado estimador de Máxima Verossimilhança de θ e satisfaz;

$$L(\bar{\theta}) = \max_{\theta \in \Omega} L(\theta) \tag{2.2}$$

Similarmente, vinda da propriedade 2.1 $\bar{\theta}$ satisfaz

$$l(\bar{\theta}) = \max_{\theta \in \Omega} l(\theta)$$

É apreciável tomar o conjunto de valores de parâmetros que são mais consistente com os dados observados, ou seja, intervalo/região de confiança para θ, sendo aqueles valores de θ para os quais a verossimilhança é relativamente grande, sendo possíveis pontos de máximo. Tomamos como região de confiança aquele conjunto de valores de θ para os quais

$$1 \leq \frac{L(\bar{\theta})}{L(\theta)} \leq c' \quad \text{ou equivalentemente} \quad l(\bar{\theta}) - l(\theta) = \log c' = c$$

para alguns valores de c', ou equivalente c. Estes argumentos para a estimação e regiões de confiança para θ, são toda a inferência baseada em verossimilhança. Todo o resto é tecnicalidades que as vezes tiram o foco da simplicidade das idéias básicas.

CAPÍTULO 3

VEROSSIMILHANÇA PARA UM PARÂMETRO

Neste capítulo nos vamos considerar problemas envolvendo um único parâmetro, ou seja, $\theta = \theta$, e para simplificar as discussões fazemos $X_1, \ldots, X_n$ são i.i.d variáveis aleatórias vindas de um modelo probabilístico $f(x;\theta)$, com valores observados $x_1, \ldots, x_n$. Assim, o conjunto geral é como em M232 capítulo 4, assim a maioria desta seção é uma revisão natural.

3.1 Princípios

Para o problema de inferência na situação, a verossimilhança $L(\theta)$, e a log-verossimilhança, $l(\theta)$ são função uni-dimensionais, assim todas as propriedades que são de interesse podem ser discutidas geometricamente. Desde que as variáveis sejam independentes

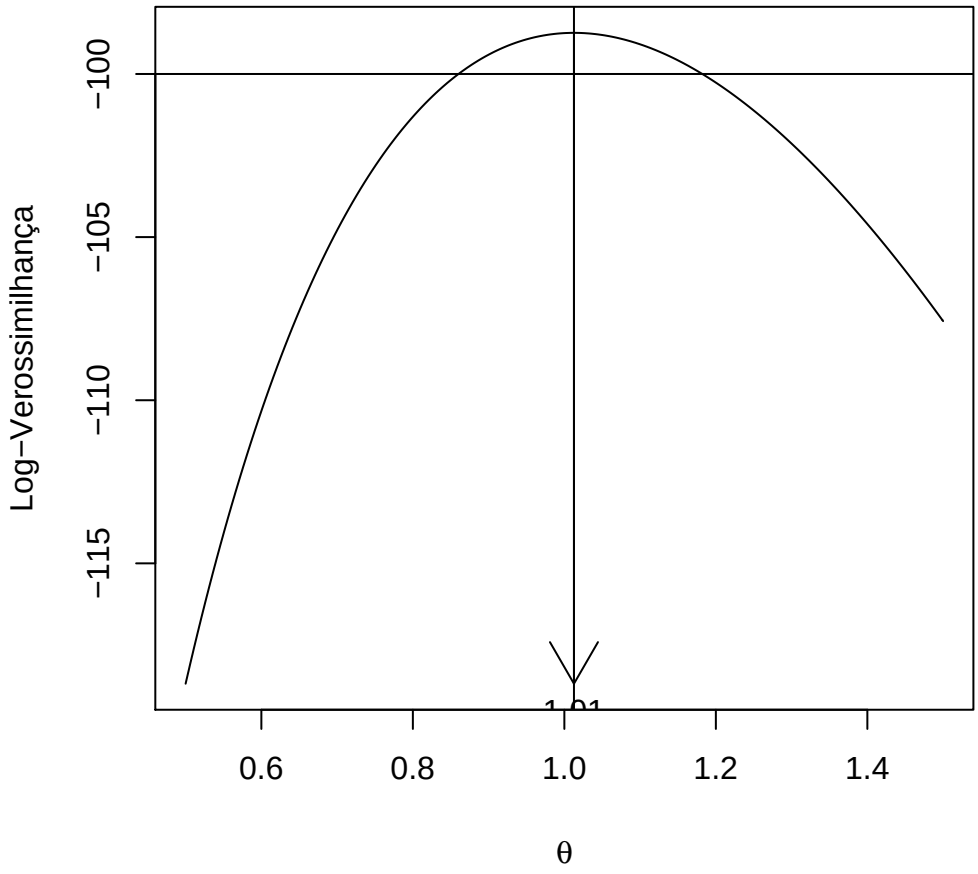
$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) \tag{3.1}$$

e

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i;\theta) \tag{3.2}$$

Um gráfico típico da função de log-verossimilhança é dado como segue:

Toda a informação vinda dos dados sobre o verdadeiro valor de θ está na função de verossimilhança ou log-verossimilhança. A estatística trata de assuntos onde a informação (dados) são resumidas de forma mais compreensível possível, inferência estatística não é diferente, vindo da função de verossimilhança nos queremos procurar o melhor estimador possível para θ, e o conjunto de valores



de θ mais consistente com os dados observados. Seguindo a discussão introduzimos:

- O melhor estimador de θ provavelmente é o estimador de máxima verossimilhança. $\hat{\theta}$, que maximiza a verossimilhança, veja o gráfico.
- O melhor estimador é o conjunto de valores mais prováveis, denominado de intervalo de confiança, é o conjunto $(\hat{\theta}_l, \hat{\theta}_u)$ onde

$$\frac{L(\hat{\theta})}{L(\hat{\theta}_l)} = \frac{L(\hat{\theta})}{L(\hat{\theta}_u)} = c \quad (\hat{\theta}_l \leq \hat{\theta} \leq \hat{\theta}_u)$$

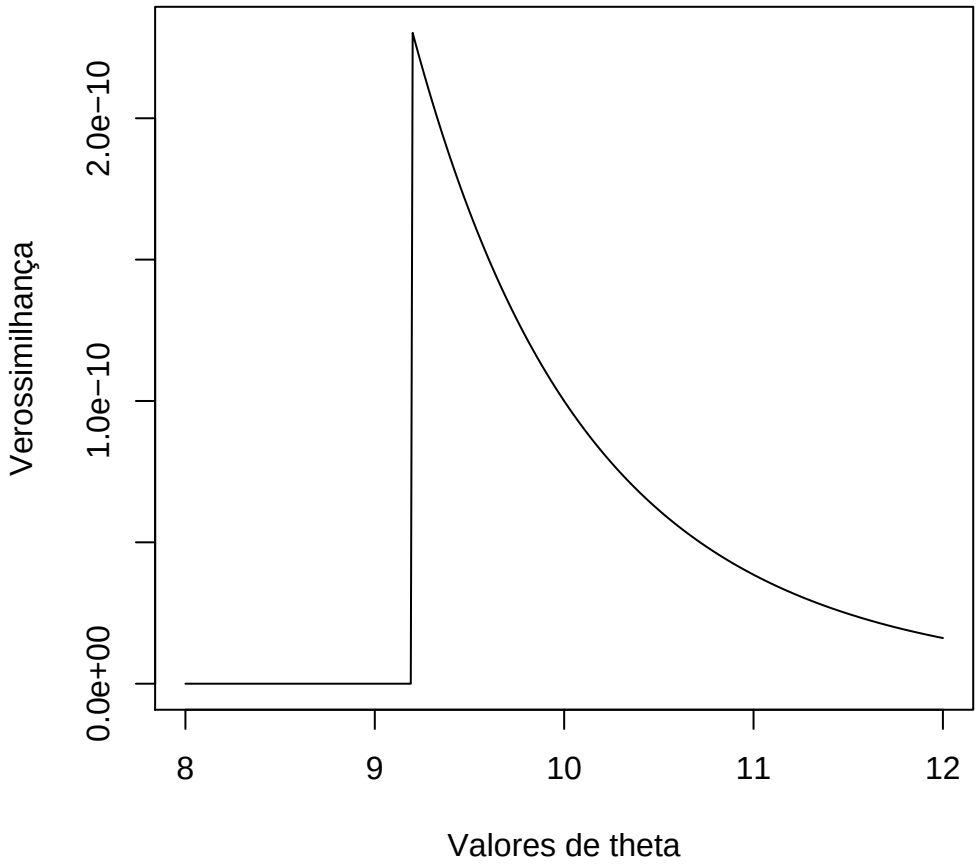
  Para alguma escolha conveniente de c, equivalentemente

$$l(\hat{\theta}) - l(\hat{\theta}_l) = l(\hat{\theta}) - l(\hat{\theta}_u) = \log c' = c$$

  veja o gráfico. Claramente, aumentando o c alarga o intervalo de confiança.

- O termo

$$D(\theta) = 2\{l(\hat{\theta}) - l(\theta)\} \geq 0$$



frquentemente no curso, vamos chamar este termo de função deviance. Aqui o intervalo de confiança é o conjunto

$$\{\theta \in \Omega : D(\theta) < c^*\}$$

onde $C^* = 2c$. A função de verossimilhança mostrada acima é típica mais em alguns exemplos não usuais podem ocorrer, a verossimilhança pode apresentar diferentes formas, como a seguir:

Aqui em (a) $\hat{\theta}_l = \hat{\theta}$, assim o melhor estimador coincide com um ponto final do intervalo, enquanto em (b) para algum valor de c o intervalo de confiança é a união de dois intervalos disjuntos, mais para outros valores de c é um simples intervalo.

Para um problema básico de inferência a abordagem básica é derivar a função de verossimilhança e obter a curvatura para os valores de $\hat{theta}$, $\hat{\theta}_L$ e $\hat{\theta}_U$. Isto nem sempre é satisfatório, e as vezes requer soluções númericas, isso é um tanto insatisfatório, porque não nos permite sentir qual a influência da amostra sobre as estiamtivas. Nos vamos contornar isso, no início considerando duas abordagens:

- Olhar exemplos onde se pode trabalhar analiticamente.

- Aproximar a verossimilhança, usando resultados para quando $n \to \infty$ e ainda trabalhar analiticamente.

Nos consideramos o primeiro caso aqui, e o segundo é estudado na seção 2.3.

Exemplo 3.1. $X_i \sim N(\theta, 1)$ assim

$$L(\theta) = \prod_{i=1}^{n} (2\pi)^{-\frac{1}{2}} \exp^{-\frac{1}{2}(x_i-\theta)^2} = 2\pi^{-\frac{n}{2}} exp^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2}$$

para $\theta \in (-\infty, \infty)$, portanto $\Omega = (-\infty, \infty)$, sim

$$l(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 = C - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2$$

onde C é uma constante que não depende de θ. Assim a função de log-verossimilhança é uma função quadrática exata, e é completamente definida por:

- $\hat{\theta}$ a posição do máximo.
- $l(\hat{\theta})$, a altura da log-verossimilhança no máximo.
- $\frac{\partial^2 l(\hat{\theta})}{\partial \theta^2} = l''(\hat{\theta})$, a curvatura da log-verossimilhança no máximo.

Voltando ao exemplo, temos

$$\frac{\partial l(\theta)}{\partial \theta} = l'(\theta) = \sum_{i=1}^{n}(x_i-\theta)^2$$

e

$$\frac{\partial^2 l(\theta)}{\partial \theta^2} = \frac{\partial l'(\theta)}{\partial \theta} = -n$$

assim a log-verossimilhança é exatamente uma função quadrática de θ. Resolvendo $l'(\theta) = 0$ temos

$$l'(\theta) = \sum_{i=1}^{n}(x_i-\hat{\theta}) = 0 \to \quad \hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Que é um ponto de máximo, pois $l''(\theta) < 0$. Resolvendo $D(\theta) = c^*$ dados $\hat{\theta}_L$ e $\hat{\theta}_U$. Temos que

$$D(\theta) = \sum_{i=1}^{n}(x_i-\theta)^2 - \sum_{i=1}^{n}(x_i-\bar{x}) = n(\bar{x}-\theta)$$

Assim, $\hat{\theta}_L = \bar{x} - \sqrt{\frac{c^*}{n}}$ e $\hat{\theta}_U = \bar{x} + \sqrt{\frac{c^*}{n}}$.

Nota. 1. O intervalo de confiança é simétrico sobre $\hat{\theta}$ desde que a log-verossimilhança seja simétrica sobre $\hat{\theta}$. Para o intervalo acima ser de 95% de confiança $\sqrt{c^*} = 1.96$, ou seja, $c^* = 3.84$.

2. A distribuição de $D(\theta)$ é uma χ^2_1, uma Qui-Quadrado com 1 grau de liberdade, assim $\bar{x} \sim N(\theta, \frac{1}{n})$, ou equivalentemente $\sqrt{n}(\bar{x} - \theta) \sim N(0,1)$.

3. Como o quadrado de uma Normal padrão é uma Qui-Quadrado segue o resultado. Note que $P(\chi^2_1 \leq 3.84) = 0.95$

Exemplo 3.2. $X_i \sim U(0,\theta)$, assim

$$f(x;\theta) = \begin{cases} \theta^{-1} & : \quad \theta > \max(x_i) \\ 0 & : \quad \theta < \max(x_i) \end{cases}$$

Assim

$$l(\theta) = \begin{cases} \theta^{-n} & : \quad \theta > \max(x_i) \\ 0 & : \quad \theta < \max(x_i) \end{cases}$$

ou seja, $\Omega = [\max(x_1,\ldots,x_n), \infty]$, assim a verossimilhança é como segue

Note que neste caso a posição do máximo é óbvia olhando para o gráfico, $\hat{\theta} = \max(x_1,\ldots,x_n)$, mais não é dado por resolver a equação $l'(\theta) = 0$. Por que ?
Aqui $\hat{\theta}_L = \max(x_1,\ldots,x_n)$ assim, a

$$D(\hat{\theta}_U) = 2n(\log(\hat{\theta}_U) - \log(\hat{\theta})) = c^*$$

Nota. 1. A log-verossimilhança não é quadrática, nem tornando $n \to \infty$. A razão para isto é que há uma informação desigual sobre θ em um lado da verossimilhança. Nos sabemos que o θ não pode ser menor que o maior valor observado, mais isto é pouca evidência, vinda dos dados sobre θ em $\hat{\theta}$.

2. A função de verossimilhança depende do $\max(x_1,\ldots,x_n)$, e n, somente vem de uma amostra.

3. A dimensão da variável é dado pelo parâmetro.

Exemplo 3.3. $X_i \sim E(\theta)$, assim

$$f(x;\theta) = \theta \exp^{-\theta x}, \quad x \geq 0$$

$$L(\theta) = \prod_{i=1}^{n} \theta \exp^{-\theta x_i}$$

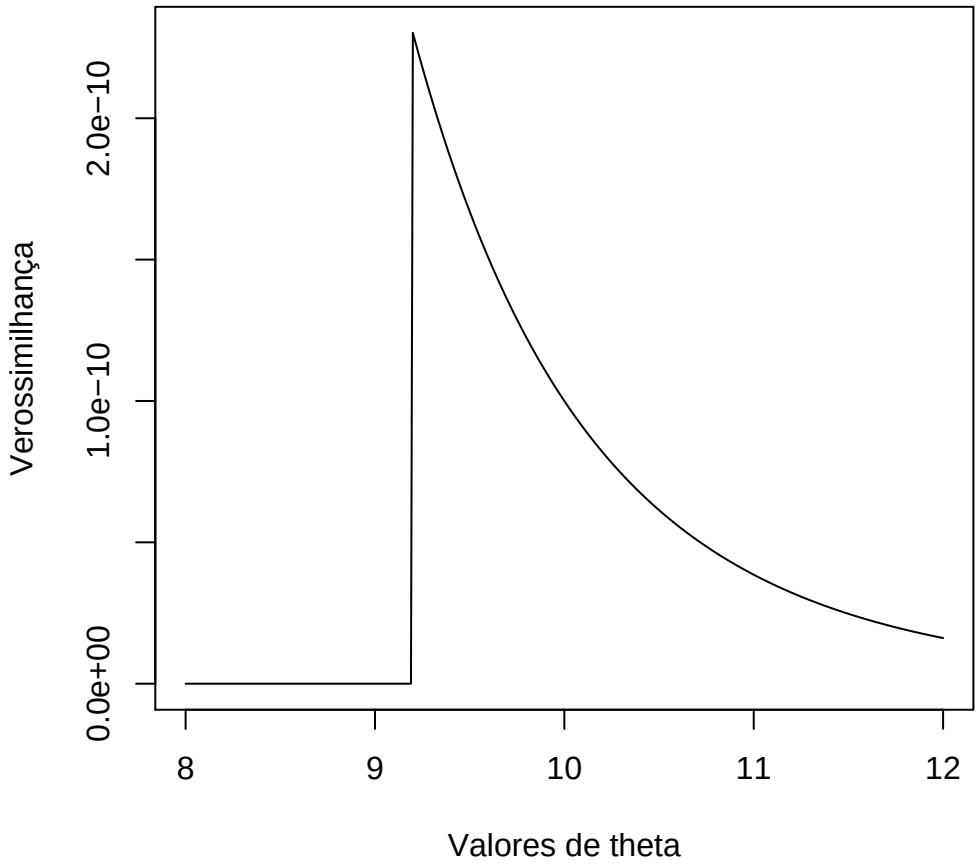
com $\Omega \in (0,\infty)$, assim

$$l(\theta) = n\log(\theta) - \theta \sum_{i=1}^{n} x_i \to n\log(\theta) - \theta n\bar{x}$$

Aqui temos que,

$$l'(\theta) = \frac{n}{\theta} - n\bar{x}$$

$$l''(\theta) = -\frac{n}{\theta^2}$$



assim a log-verossimilhança é côncava, de ordem infinita, é uma função polinomial. Resolvendo $l'(\theta) = 0$, temos

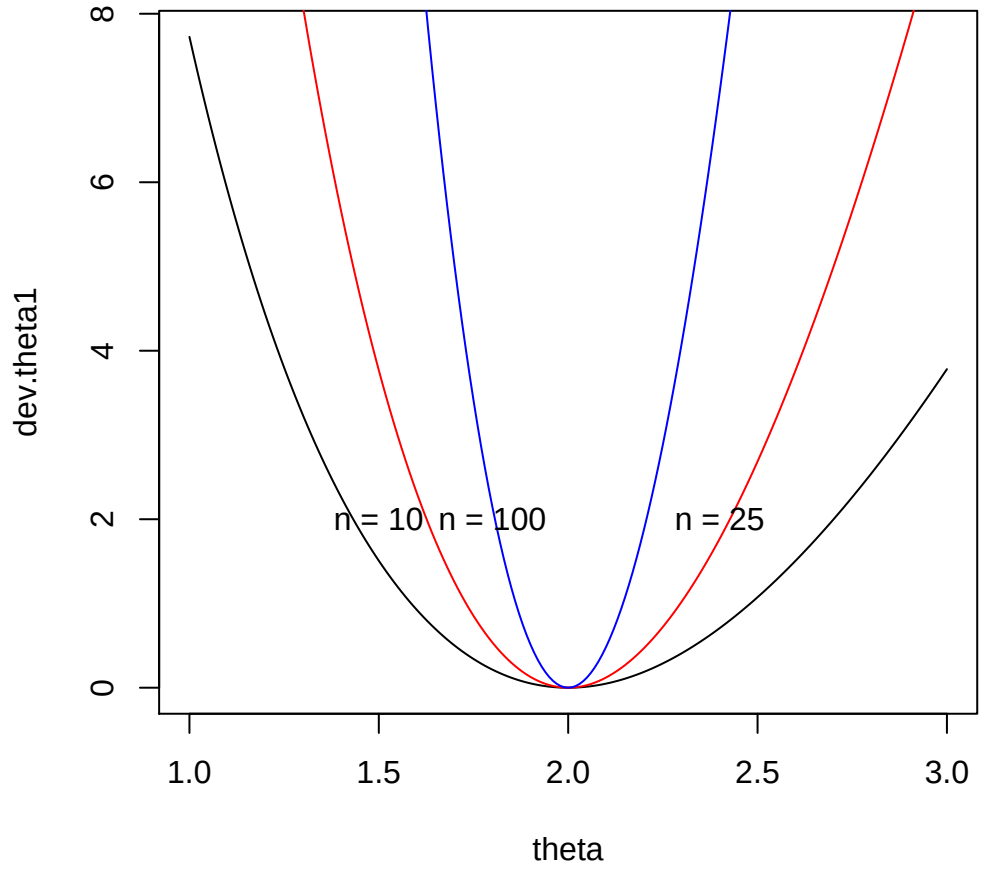
$$l'(\theta) = \frac{n}{\hat{\theta}} - n\overline{x} = 0 \rightarrow \hat{\theta} = \frac{1}{\overline{x}}$$

Que é o ponto de máximo, já que, $l''(\theta) < 0$. Resolvendo $D(\theta) = c^*$ dado $\hat{\theta}_L$ e $\hat{\theta}_U$, temos

$$D(\theta) = 2(l(\hat{\theta}) - l(\theta)) =$$

$$2n(\log(\frac{\hat{\theta}}{\theta}) - \overline{x}(\hat{\theta} - \theta))$$

mesmo este sendo um exemplo simples, uma solução numérica é necessária para encontrar $\hat{\theta}_L$ e $\hat{\theta}_U$. Para isso é útil examinar a forma de $D(\theta)$ para ver o que ela diz sobre os valores. A figura 3.1 mostra a $D(\theta)$ para $n = 10$, $n = 25$ e $n = 100$.



Nota. 1. O gráfico acima mostra que quando n é grande a vizinhança do máximo é aproximadamente quadrática, ou seja, nesta região

$$D(\theta) = 2(l(\hat{\theta})) - [l(\hat{\theta}) + (\theta - \hat{\theta})l'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 l''(\hat{\theta})]) =$$

$$n\left(\frac{\theta - \hat{\theta}}{\theta}\right)^2$$

assim os intervalos ficam dados por:

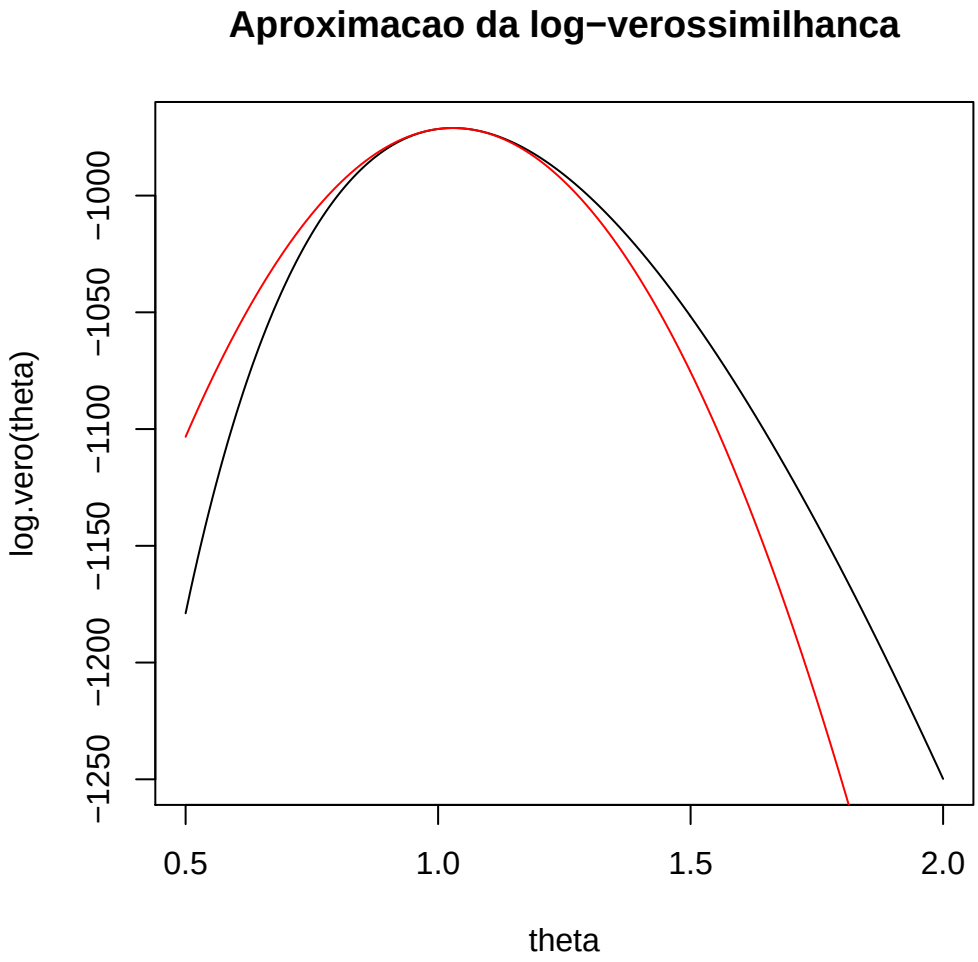
$$\hat{\theta}_U = \hat{\theta}(1 + \sqrt{\frac{c^*}{n}}$$

e

$$\hat{\theta}_L = \hat{\theta}(1 - \sqrt{\frac{c^*}{n}}$$

A figura **??** mostra a função de log-verossimilhança e a aproximação quadrática usada para construir o intervalo.

2. Para *n* pequeno a função de log-verossimilhança é assimétrica sobre o m.l.e, consequentemente é assim o intervalo de confiança exato.



3.2 Invariância

Em aplicações pode ser de interesse alguma função do parâmetro do modelo, por exemplo, na exponencial (θ) (Exemplo 3.3) o interesse pode ser em

$$\phi = P(X \leq u)1 - \exp^{-\theta u}$$

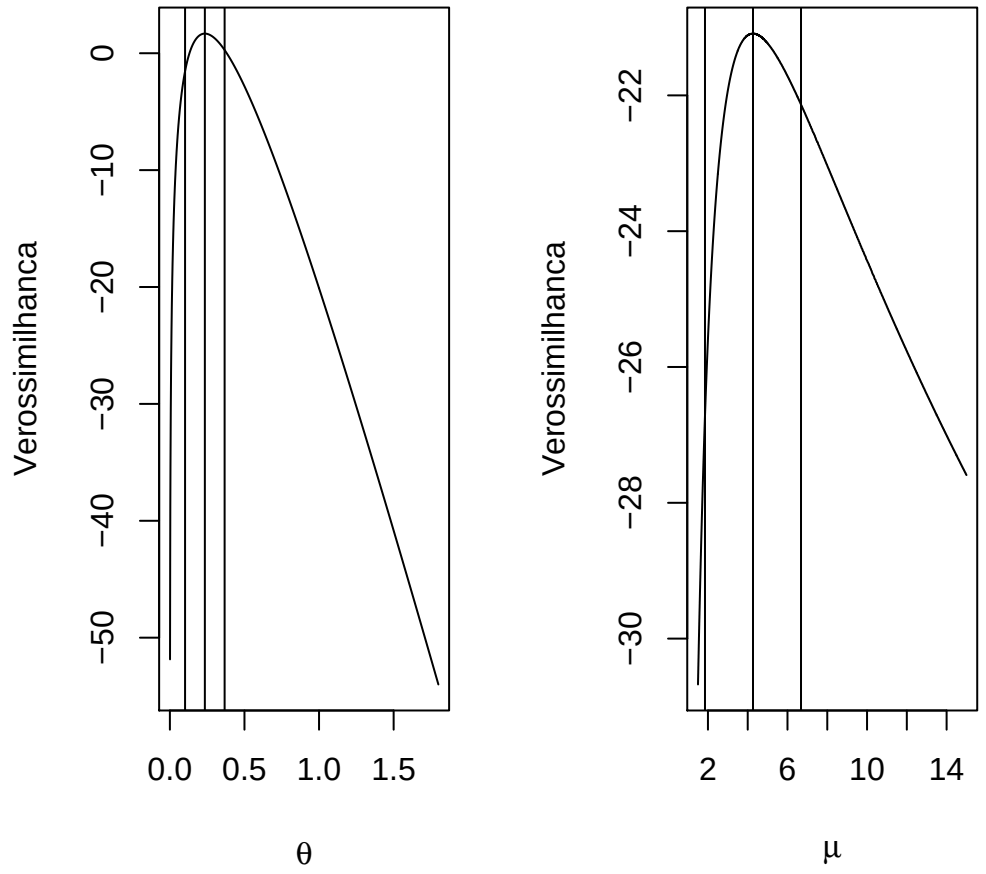
Para algum nível u fixado. Ou em alguma parametrização alternativa do modelo probabilístico.

$$f(x;\phi) = \frac{1}{\phi}\exp^{-\frac{x}{\phi}}$$

onde $\phi = \frac{1}{\theta} = E[X]$. Nesta seção nos mostramos que o estimador e o intervalo de confiança para este parâmetro pode ser derivado diretamente da função de verossimilhança para θ, e especificamente vindo de $\hat{\theta}$, $\hat{\theta}_L$ e $\hat{\theta}_U$.

Geralmente, supomos que estamos interessados em algum parâmetro $\phi = g(\theta)$, uma função um-a-um de θ, e que a verossimilhança, o m.l.e e o intervalo de confiança são requeridos para ϕ, ou seja, $L(\phi)$, $\hat{\phi}$, $\hat{\phi}_L$ e $\hat{\phi}_U$.

Desde que $L(\phi) = L(g(\theta))$, a função de verossimilhança para ϕ é obtida vinda da função de verossimilhança de θ por transformação na escala de x, no gráfico:



Consequentemente, vindo destes gráficos fica claro que $\hat{\phi} = g(\hat{\theta})$; $\hat{\theta}_L = g(\hat{\theta}_L)$; $\hat{\phi}_U = g(\hat{\theta}_U)$. Para ilustrar isto a figura 3.2 mostra isto para a exponencial(θ) (Exemplo 3.3) com $n = 25$ onde $\phi = \frac{1}{\theta}$.

Nota. 1. Invariância é uma propriedade útil quando somente uma maximização da verossimilhança e derivação de intervalos de confiança são necessário mesmo que o interesse seja em muitas funções do parâmetro desconhecido. Nos consequentemente podemos escolher a parametrização mais conveniente, para posteriormente derivar os estimadores e intervalos de confiança para os parâmetros de interesse.

2. Embora nos possamos ter uma razoável aproximação quadrática da log-verossimilhança para θ, isto pode ser difícil para $g(\theta)$ e vice-versa.

3.3 Aproximações assintóticas da verossimilhança

O exemplo da seção 3.1 mostra que mesmo para problemas simples, soluções analíticas são complicadas de serem obtidas, quando não são impossíveis. Quando é necessário usar um método numérico para encontrar a estimativa, não temos como entender como as características amostrais influenciam na fórmula do intervalo de confiança. Além disso, nos ainda não temos nenhuma idéia de quais valores de c^* são necessários para fazer intervalos de confiança baseados na verossimilhança, dados por:

$$\theta \in \Omega : D(\theta) < C^*$$

sendo de 95% ou 99% de confiança. Nesta seção nos usamos aproximações da verossimilhança que ocorre quando $n \to \infty$. Todos os resultados vem da seguinte expansão em séries de Taylor, em torno do m.l.e.

$$l(\theta) = l(\hat{\theta}) + (\theta - \hat{\theta})l'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 l''(\theta^+) \text{ para } |\theta^+ - \theta| \leq |\hat{\theta} - \theta| \tag{3.3}$$

$$l'(\theta) = l'(\hat{\theta}) + (\theta - \hat{\theta})l''(\theta^+) \text{ para } |\theta^+ - \theta| \leq |\hat{\theta} - \theta| \tag{3.4}$$

Para simplificar a notação nos fazemos $U(\theta) = l'(\theta)$, denominamos de função escore, e fazemos $I_o(\theta) = -l''(\theta)$, denominamos de informação observada, e

$$I_E(\theta) = E\{I_o(\theta)\} = E\{-\frac{\partial^2 \log L(\theta; x)}{\partial \theta^2}\},$$

é denominada de informação esperada, ou informação de Fisher. Note que aqui nos olhamos as propriedades de U e I_o, considerando propriedades amostrais, ou seja, trocando x por X. Estes dois termos são fundamentais para determinar o comportamento das expansões da verossimilhança (3.3) e (eq:l.deriv.taylor). Assim, a função de verossimilhança e log-verossimilhança são variáveis aleatórias, determinando a posição do máximo, θ, o gradiente da função $U(\theta)$ e a curvatura $I_o(\theta)$.

3.3.1 Resultados e definições

Definição 3.1. Se $T = T(x)$ tem a propriedade $E(T) = g(\theta)$, para alguma função g de θ, então T é chamado de estimador não viciado de $g(\theta)$.

Lemma 3.1. *$E(U(\theta)) = 0$ e $V(U(\theta)) = I_E(\theta)$.*

Lemma 3.2. *Se $T = T(x)$ é um estimador não viciado para θ, então $E[T] = \theta$, e $V[T] \geq [I_E(\theta)]^{-1}$, é o denominado limite de Cramér-Rao.*

Lemma 3.3. *Se $T = T(x)$ é um estimador não viciado para $g(\theta)$, então $V[T] \geq [g'(\theta)]^2[I_E(\theta)]^{-1}$.*

Prova. Sabemos que

$$I_E(\phi) = E\left\{-\frac{\partial^2 l(\phi)}{\partial \phi^2}\right\} \tag{3.5}$$

Entretanto,

$$\frac{\partial l}{\partial \phi} = \frac{\partial \theta}{\partial \phi}\frac{\partial l}{\partial \theta}$$

assim

$$\frac{\partial^2 l}{\partial \phi^2} = \frac{\partial^2 \theta}{\partial \phi^2}\frac{\partial l}{\partial \theta} + \left(\frac{\partial \theta}{\partial \phi}\right)^2$$

$$= \frac{\partial^2 \theta}{\partial \phi^2}U(\theta) + \left(\frac{1}{g'(\theta)}\right)^2 l''(\theta$$

segue então, que

$$E\left\{-\frac{\partial^2 l(\phi)}{\partial \phi^2}\right\} = 0 + \frac{I_E(\theta)}{[g'(\theta)]^2}$$

Aqui o último passo é baseado no lema 3.3.2 e na definição de informação esperada. O resultado é obtido por inserir a expressão da derivada em 3.5.

3.3.2 Principais resultados

Embora os resultados clássicos da teoria de verossimilhança, que são dados nesta seção, são usualmente aplicados, eles não são válidos universalmente, nos devemos assumir as seguintes condições de regularidade:

1. Ω é finito dimensional e o verdadeiro valor de θ é um ponto interior de Ω.
2. Para nenhum dois valores diferentes de θ, a verossimilhança é igual.
3. As primeiras 3 derivadas de $l(\theta)$ existem em uma vizinhança do verdadeiro valor de θ.

Estas condições em essência, dizem que a dimensão da distribuição não pode depender de θ, e assegura que com $n \to \infty$ a log-verossimilhança converge para uma forma quadrática em uma vizinhança do m.l.e, e assim depende somente da posição do m.l.e e da curvatura da verossimilhança no m.l.e.

Teorema 3.4. *Para um problema de estimação regular, no limite quando $n \to \infty$, se θ é o verdadeiro valor do parâmetro então*

$$\sqrt{I_E(\theta)}(\hat{\theta} - \theta) \sim N(0,1)$$

ou seja, $\hat{\theta} \sim N(\theta, I_E(\theta)^{-1})$, nos denominamos isso de distribuição assintótica de θ.

Prova.

$$U(\theta) = -(\theta - \hat{\theta})I_o(\theta^+)$$

Desde que $\hat{\theta} \to \theta$ com $n \to \infty$, então $\theta^+ \to \theta$, também. Nos podemos aproximar $I_o(\theta)$ por $I_E(\theta)$ então

$$\frac{I_E(\theta)}{I_o(\theta)} \to 1$$

Como $U(\theta)$ é a soma de variáveis aleatórias, aplicando o teorema do limite central, vindo do lema ,

$$\frac{U(\theta)}{[I_E(\theta)]^{-\frac{1}{2}}} \to Z \quad \text{onde} Z \sim N(0,1)$$

Desde que

$$\sqrt{I_E(\theta)}(\theta - \hat{\theta}) = \frac{U(\theta)}{[I_E(\theta)]^{-\frac{1}{2}}} \frac{I_E(\theta)}{I_0(\theta)}$$
$$= Z \quad x \quad 1 = Z$$

Corolário. *Se $\phi = g(\theta)$ então $\hat{\phi} \sim N(\phi, [g'(\theta)]^2 I_E(\phi)^{-1}$ com $n \to \infty$. A consequência importante do corolário 1, é que $V(\hat{\phi}) = [g'(\theta)]^2 V(\hat{\theta})$.*

Corolário. *Em termos assintóticos são equivalentes a $I_E(\theta)$ e podem ser usadas no teorema, as quantidades ;*

$$\sqrt{I_E(\hat{\theta})}(\hat{\theta} - \theta) \sim N(0,1)$$
$$\sqrt{I_o(\theta)}(\hat{\theta} - \theta) \sim N(0,1)$$
$$\sqrt{I_o(\hat{\theta})}(\hat{\theta} - \theta) \sim N(0,1)$$

mais as propriedades de convergência são diferentes em cada caso.

Teorema 3.5. *Para um problema de estimação regular, no limite com $n \to \infty$, se θ é o verdadeiro valor do parâmetro, temos*

$$D(\theta) = 2[l(\hat{\theta} - l(\theta)] \sim \chi^2$$

Prova. Este resultado é baseado na expansão de Taylor (3.3).

$$D(\theta) = 2[l(\hat{\theta}) - [l(\hat{\theta}) + (\hat{\theta} + \theta)U(\hat{\theta})$$
$$(\hat{\theta} - \theta)^2 I_o(\theta^+)$$

Como na prova do teorema 3.4, $\theta^+ \to \theta$ assim

$$D(\theta) = [\sqrt{I_E(\theta)}(\hat{\theta} - \theta)]^2 \frac{I_o(\theta^+)}{I_E(\theta)}$$

$$\approx Z^2 \quad x \quad 1 \quad = Z^2$$

Vindo do teorema 3.4, $Z \sim N(0,1)$ *e* $Z^2 \sim \chi^2_1$.

3.3.3 Discussão dos principais resultados

O teorema thm:assim.dist assoaciado com os corolários dão que:

1. O m.l.e, $\hat{\theta}$ de θ é assintóticamente não-viciado, $E(\hat{\theta}) \to \theta$.
2. $\hat{\theta}$ atinge assintóticamente o limite de Crámer-Rao, $V(\hat{\theta}) \to [I_E(\theta)]^{-1}$.
3. Nos denominamos $[I_E(\theta)]^{-\frac{1}{2}}$ de erro padrão de $\hat{\theta}$, as vezes escrevemos $se(\hat{\theta})$.
4. Nos podemos construir intervalos de confiança válidos assintóticamente com $100(1-\alpha)\%$ de confiança para θ na forma $\hat{\theta} \pm Z_{\frac{\alpha}{2}}$, onde $Z_{\frac{\alpha}{2}}$ é o quantil da distribuição Normal, por exemplo, se $\alpha = 0.05$ então $Z_{\frac{\alpha}{2}} = 1.96$, nos podemos denotar este intervalo por $(\hat{\theta}_L, \hat{\theta}_U)$.
5. O estimador de Máxima Verossimilhança $\hat{\phi} = g(\hat{\theta})$ de $\phi = g(\theta)$ é assintóticamente não viciado, ou seja, $E(g(\hat{\theta})) \to g(\theta)$.
6. $\hat{\phi} = g(\hat{\theta})$ atinge assintóticamente o limite de Crámer-Rao, ou seja,
$$V(g(\hat{\theta})) \to [g'(\theta)]^2 [I_E(\theta)]^{-1} = I_E(\phi)^{-1}$$
7. Denominamos o termo $|g'(\theta)|[I_E(\theta)]^{-\frac{1}{2}}$ de erro padrão de ϕ.
8. Nos podemos construir intervalos de confiança de $100(1-\alpha)\%$ de confiança para $\phi = g(\theta)$ da forma $g(\hat{\theta}) \pm Z_{\frac{\alpha}{2}} |g'(\theta)|[I_E(\theta)]^{-\frac{1}{2}}$, onde $Z_{\frac{\alpha}{2}}$ é o quantil da distribuição Normal.
9. Em todas as expressões $I_E(\theta)$ pode ser trocada pelos termos assintóticamente equivalentes $I_E(\hat{\theta})$, $I_o(\hat{\theta})$, $I_o(\theta)$, similarmente $g'(\theta)$ pode ser substituída por $g'(\hat{\theta})$. Muito estudo foi dedicado para dizer que estes são os melhores termos para serem usados na prática. E foi verificado que $I_o(\hat{\theta})$ tem as melhores propriedades. Isto entretanto é óbvio, como nos queremos obter a melhor aproximação da função de verossimilhança, apenas usando a curvatura observada. Além disso, a informação observada é a mais fácil das quatro opções de ser obtida, o que é sempre uma propriedade muito boa.

Correspondente ao teorema 3.5.

1. Pelos argumentos da seção 3.3 nos temos que o mais interessante estimador de um intervalo de confiança é da forma $\theta \in \Omega : D(\theta) < c^*$, para algum valor de c^*. Para dizer que o intervalo é exatamente de $100(1-\alpha)\%$ nos precisamos escolher c^* de tal forma que se retirarmos repetidas amostras e construirmos intervalos de confiança, estes contenham o verdadeiro valor de θ, com proporção $1-\alpha$. Isto é geralmente impossível de fazer, a menos se usar métodos computacionais. Entretanto, com n tomado grande, o Teorema 3.5 sugestiona que uma escolha conveniente é $c^* = c$,

com $Pr\chi_1^2 \geq c_\alpha = \alpha$ por exemplo, se $\alpha = 0.05$, então $c_\alpha = 3.84$. Uma simples e conveniente abordagem é usar c^* mesmo que n seja pequeno, e fazer disto uma aproximação do intervalo de $100(1-\alpha)\%$. Claramente ainda é um intervalo baseado na verossimilhança mais para amostras pequenas o valor verdadeiro de α pode ser ligeiramente diferente do que o valor de referência.

2. Note que vindo dos teoremas 3.4 e 3.5 nos temos duas abordagens para obter intervalos de confiança: a baseado no teorema thm:assim.dist que dá intervalos $(\tilde{\theta}_L, \tilde{\theta}_U)$, e a baseada no teorema 3.5 que dá $(\hat{\theta}_L, \hat{\theta}_U)$ para o parâmetro θ. O intervalo baseado no teorema 3.4, são simples de serem construídos, mais não possuem a importante propriedade de invariância, sendo assim, se $\phi = g(\theta)$ então

$$\{g(\tilde{\theta}_L), g(\tilde{\theta}_U\} = \left\{g(\hat{\theta} - Z_{\frac{\alpha}{2}}[I_E(\theta)]^{-\frac{1}{2}}); g(\hat{\theta} + Z_{\frac{\alpha}{2}}[I_E(\theta)]^{-\frac{1}{2}})\right\} \neq$$

$$\left\{g(\hat{\theta}) - Z_{\frac{\alpha}{2}}|g'(\theta)|[I_E(\theta)]^{-\frac{1}{2}}; g(\hat{\theta}) + Z_{\frac{\alpha}{2}}|g'(\theta)|[I_E(\theta)]^{-\frac{1}{2}}\right\}$$

   a menos que g seja linear.

3. Intervalos de confiança construídos usando o teorema 3.4, são baseados na verossimilhança, mais eles são baseados em uma log-verossimilhança correspondente a uma aproximação quadrática da função log-verossimilhança exata. Entretanto nos podemos ver o quanto bom o intervalo de confiança $(\tilde{\theta}_L, \tilde{\theta}_U)$ como uma aproximação para $(\hat{\theta}_L, \hat{\theta}_U)$ por olhar o fechamento da log-verossimilhança quadrática para θ. Mesmo se a aproximação for boa é importante lembrar que se o interesse é em algum outro parâmetro $\phi = g(\theta)$ e se considerar necessário intervalo de confiança para ϕ. A melhor abordagem é usar o intervalo baseado na verossimilhança $(\hat{\phi}_L, \hat{\phi}_U) = (g(\hat{\theta}_L), g(\hat{\theta}_U)$, mais desde que $\hat{\theta}_L$ e $\hat{\theta}_U$ não possam ser encontrados exatamente uma aproximação será necessária. Examinando a log-verossimilhança para ϕ acerca do m.l.e nos podemos ver que se $l(\phi)$ é mais enviesada que $l(\theta)$.

- Se $l(\phi)$ é menos enviesada a melhor abordagem é usar o intervalo aproximado $(\tilde{\phi}_l, \tilde{\phi}_U)$.
- Se $l(\theta)$ é menos enviesada, então a melhor abordagem é usar o intervalo $g(\tilde{\theta}_L); g(\tilde{\theta}_U)$.

3.3.4 Exemplos

Nesta seção nos consideramos com mais detalhes os exemplos 3.1 - 3.1 e uma família de distribuições denominada de 'família exponencial'.

Exemplo 3.1 (Cont): Normal com média desconhecida. Relembre

$$l(\theta) = C - \frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2$$

e $\hat{\theta} = \bar{x}$, $I_o(\theta) = -l''(\theta) = n$. Note que sempre que $I_o(\theta)$ é uma constante $I_E(\theta) = I_o(\theta)$. Aqui a log-verossimilhança é uma forma quadrática perfeita, assim

- a forma dos dois intervalos de confiança são iguais, $(\hat{\theta}_L, \hat{\theta}_U) = (\tilde{\theta}_L, \tilde{\theta}_U)$.

- Ambos intervalos são exatos,

consequentemente, o intervalo de $100(1-\alpha)$ de confiança tem extremos em:

$$\hat{\theta}_L = \bar{x} - \sqrt{\frac{c_\alpha}{n}} \quad \text{e} \quad \hat{\theta}_U = \bar{x} + \sqrt{\frac{c_\alpha}{n}}$$

onde a $Pr\chi^2 \geq c_\alpha = \alpha$. Para uma amostra de $n = 25$, a verossimilhança e o intervalo de confiança são mostrados na figura 5.

Suponha agora que o interesse está em $\phi = PrX \leq u = \Phi(u-\theta)$ para u desconhecido, então $\hat{\phi} = \Phi(u-\hat{\theta})$ e $(\hat{\phi}_L, \hat{\phi}_U) = \Phi(u-\hat{\theta}_L), \Phi(u-\hat{\theta}_U)$. A figura 6 mostra a log-verossimilhança ϕ e o intervalo de confiança $(\hat{\phi}_L, \hat{\phi}_L)$ e $(\tilde{\phi}_L, \tilde{\phi}_U)$. O trabalho para o intervalo $(\hat{\phi}_L, \hat{\phi}_L)$ é como segue

$$I_o(\hat{\theta}) = n \quad g(\theta) = \Phi(u-\theta)$$

assim

$$V[\hat{\phi}] = [g'(\theta)]^2 [I_E(\theta)]^{-1} \approx [g'(\hat{\theta})]^2 [I_o(\hat{\theta})]^{-1} =$$
$$\left[-\frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}(u-\hat{\theta})^2} \right]^2 \frac{1}{n}$$

Consequentemente, o erro padrão é dado por

$$se(\hat{\phi}) = \left[-\frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}(u-\hat{\theta})^2} \right] \frac{1}{\sqrt{n}}$$

Resultados para $\tilde{\phi}_L$ e $\tilde{\phi}_U$ segue de sua definição.
Exemplo ?? - Exponencial Relembre que

$$l(\theta) = n \log \theta - \theta \bar{x}$$

e $\widehat{\theta} = 1/\bar{x}$, $I_o(\theta) = n\theta^{-2}$. Note que desde que $I_o(\theta)$ seja uma constante $I_E(\theta) = I_o(\theta)$. Aqui a log-verossimilhança é ligeiramente enviesada assim o intervalo de confiança aproximado $(\tilde{\theta}_l, \tilde{\theta})$ não vai ser exatamente igual a $(\widehat{\theta}_l, \widehat{\theta}_u)$. O intervalo aproximado, obtido usando a informação observada, acerca do m.l.e, é

$$\tilde{\theta}_l, \tilde{\theta}_l = \widehat{\theta}(1 - z_{\frac{z}{\alpha}}/\sqrt{n}), \widehat{\theta}(1 + z_{\frac{z}{\alpha}}/\sqrt{n}),$$

ou seja, como encontrado antes usando a abordagem de aproximação. O intervalo de confiança baseado na verossimilhança precisa de solução numerica, veja a Figuras 3.1. Claramente quando n aumenta a aproximação melhora, tomando uma forma razoável quando $n > 100$. Agora suponha $\phi = P(X \leq u) = 1 - \exp -\theta u$ é o interesse. A figura 7 mostra os três intervalos de confiança $(\widehat{\phi}_l, \widehat{\phi}_l)$, $(\tilde{\phi}_l, \tilde{\phi}_u)$ e $(1 - \exp -\tilde{\theta}_u u, 1 - \exp -tilde\theta_l u)$ (o menor intervalo) para $n = 25$ dados examinados e $u = 0.75$.

Exemplo 3.4. A variável aleatória, X, é chamada pertencente a familia exponencial uniparamétrica se

$$f(x;\theta) = \exp h(\theta) + k(x) + c(\theta)t(x)$$

onde h,k,c e t são funções tais que $\int_R f(x;\theta)dx = 1$, onde R é a dimensão de X. E neste caso a

verossimilhança para uma amostra i.i.d é

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} \exp h(\theta) + k(x_i) + c(\theta)t(x_i) \\ &= \exp nh(\theta) + \sum_{i=1}^{n} k(x_i) + c(\theta)\sum_{i=1}^{n} t(x_i) \\ &= \exp nh(\theta) + k(\mathbf{x}) + c(\theta)t(\mathbf{x}) \end{aligned}$$

Assim,

$$\begin{aligned} l(\theta) &= nh(\theta) + k(\mathbf{x}) + c(\theta)t(\mathbf{x}), \quad \text{e} \\ U(\theta) &= l'(\theta) = nh'(\theta) + c'(\theta)t(\mathbf{x}) \\ &= c'(\theta)\left[t(\mathbf{x}) + \frac{nh'(\theta)}{c'(\theta)}\right] \\ &= c'(\theta)[t(\mathbf{x}) - \tau(\theta)] \end{aligned}$$

onde $\tau(\theta) = \frac{-nh'(\theta)}{c'(\theta)}$. Denomina-se o termo $c(\theta)$ de termo canônico e $t(\mathbf{x})$ de estatística suficiente. Vindo do lema 1, $0 = EU(\theta) = c'(\theta)[Et(\mathbf{X}) - \tau(\theta)]$ assim $Et(\mathbf{X}) = \tau(\theta)$. A principal propriedade da família exponencial é que

Teorema 3.6. *Se X é pertencente a famíla exponecial, então $t(\boldsymbol{X})$ atinge o limite de Cramér-Rao, ou seja, $Et(\boldsymbol{X}) = \tau(\theta)$ e*

$$Vart(\boldsymbol{X}) = [\tau'(\theta)]^2 [I_E(\theta)]^{-1}$$

Isto é geralmente, visto como uma boa propriedade mais é bastante restritiva e nem sempre assegura boas propriedades da verossimilhança. A caso particular onde isso é bom é quando $\phi = c(\theta)$ é o parâmetro de interesse, este parâmetro é denominado de parâmetro canônico. para ilustrar isto tome $c(\theta) = \theta$ então

$$l'(\theta) = nh'(\theta) + t(\boldsymbol{x}) \quad e \quad l''(\theta) = nh''(\theta)$$

assim a posição (altura) do máximo depende dos dados, mas a curvatura e as derivadas de maior ordem são independentes dos dados, dependendo apenas do tamanho da amostra, assim $I_E(\theta) = I_o(\theta)$.

Exemplo $X \sim N(\theta, 1)$ então,

$$\begin{aligned} f(x;\theta) &= \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}(x-\theta)^2} \\ &= \exp\{-\frac{1}{2}\log(2\pi) - \frac{1}{2}(x^2 - 2x\theta + \theta^2)\} \end{aligned}$$

aqui $h(\theta) = -\frac{\theta^2}{2}$, $k(x) = [-\log(2\pi) - x^2]/2$, $c(\theta) = \theta$ e $t(x) = x$. Aqui o parâmetro canônico é θ assim a log-verossimilhança para θ tem a curvatura independente dos valores amostrais.

Exemplo Uniforme: Relembre que a função de log-verossimilhança é não quadrática para todo n.

Consequentemente as aproximações 3.3 e **??** não podem ser usadas. Além dos mais este problema de inferência é não regular, com as condições de regularidade R! e R3 não são atendidas. Métodos alternativos são necessários aqui e dado $c^* = -2\log(1-\alpha)$.

Prova. Desde que $\widehat{\theta}_l = \widehat{\theta} < \widehat{\theta}_u$ nos precisamos $P(\widehat{\theta}_u \leq \theta) = 1-\alpha$. Entretando,

$$\begin{aligned}
P(\widehat{\theta}_u \leq \theta) &= P(\widehat{\theta}\exp c^*/(2n) \leq \theta) \\
&= P(\widehat{\theta} \leq \theta d) \quad \text{onde} \quad d = \exp -c^*/(2n) \\
&= P(\max(X_1,\ldots,X_n) \leq \theta d) \\
&= P(X_1 \leq \theta d,\ldots,X_n \leq \theta d) \\
&= [P(X_1 \leq \theta d)]^n \quad \text{por independência} \\
&= \left[\frac{\theta d}{\theta}\right]^n \\
&= d^n = 1-\alpha
\end{aligned}$$

assim $\exp -c^*/2 = 1-\alpha$ e segue o resultado.

3.4 Teste da razão de verossimilhança

Em alguns casos nos estamos interessados em testar se um certo valor de um parâmetro é consistente com os dados. Este valor, θ_0, vem tipicamente tem alguma sugestão fisica conhecida do fenômeno ou de análises anteriores. Nos podemos testar a hipótese que

$$\theta = \theta_0 \quad \text{com a restrição de} \quad \theta \in \Omega$$

por comparar a verossimilhança sob cada uma das hipóteses. A verossimilhança sob o reinvindicação é $L(\theta_0)$ considerando que o máximo da verossimilhança para $\theta \in \Omega$ é

$$\max_{\theta\in\Omega} L(\theta) = L(\widehat{\theta}).$$

Assim comparando a verossimilhança nos temos que a verossimilhança relativa para a hipótese é

$$\frac{L(\widehat{\theta}}{L(\theta_0)} \quad \text{ou} \log\left(\frac{L(\widehat{\theta}}{L(\theta_0)}\right) = l(\widehat{\theta} - l(\theta_0) = D(\theta_0)/2$$

assim a evidência para θ_0 é testada vindo da deviance calculada no valor reinvidicado. Como intervalos de confiança baseado na verossimilhança são derivados usando a deviance, segue que testar a hipótese que $\theta = \theta_0$ é equivalente a construir um intervalo de confiança e ver se o valor reinvidicado está dentro do intervalo. Tomando como exemplo o caso da $N(\theta,1)$ (Exemplo 3.1). Se $\overline{x} = 6.3$ e $n = 100$ testar a hipótese $\theta = 6$ ao nível de 5%. Aqui nos temos duas abordagens: uma construir o intervalo de

confiança de 95% ou calcular a deviance em (5). O intervalo é

$$\bar{x} \pm 1.96/\sqrt{n} = 6.3 \pm 0.196 = (6.104, 6.496)$$

assim exclui a hipótese de $\theta = 6$ assim a hipótese é rejeitada ao nível de 5%. Equivalentemente, a deviance é

$$D(6) = n(\bar{x} - 6)^2 = 100(6.3 - 6)^2 = 9 > 3.94$$

assim é uma deviance significativa ao nível de 5%. A conclusão é que os valores de θ contidos no intervalo de confiança de 95% são todos os valores de θ que são aceitos em um teste de hipótese com nível de 95%. Assim nos não precisamos examinar testes de hipóteses em mais detalhes, porque elas são idênticos a construir intervalos baseados na verossimilhança.

3.5 Conclusões

Nos temos visto que a construção de estimadores de Máxima verossimilhança e intervalos de confiança associados, apresentam boas propriedades vinda da função de verossimilhança (ou log-verossimilhança). Frequentemente calcular $\widehat{\theta}, \widehat{\theta}_l, \widehat{\theta}_u$ requer métodos numéricos, assim nos estudamos aproximações para a função de log-verossimilhança, usando aproximações quadráticas que são justificadas com $n \to \infty$ para problemas de estimação regulares. Além disso, nos temos mostrado que além de ser um estimador lógico, o estimador de Máxima verossimilhança é assintóticamente o *melhor* estimador, e com $n \to \infty$ é não - viesado além de atingir o limite minímo de Cramér-Rao.

CAPÍTULO 4

VEROSSIMILHANÇA COM MULTI-PARÂMETROS

Neste capítulo ilustramos o corpo da função de verossimilhança para problemas de dimensão Z. O conceito de estimados de máxima verossimilhança é extendido de forma simples de um para dois parâmetros: no caso de dois parâmetros a posição de ponto de máximo na superfie de verossimilhança sobre o espaço paramétrico. A extenção de um intervalo de confiança, para dois parâmetros requer o conceito de região de confiança, ou seja, um conjunto bidimensinal. A enfase aqui é examinar o corpo que a região de confiança tome em uma dimensão de problemas. Durante todo o capítulo toma-se $X_1, ..., X_n$ sendo independentes v.a.

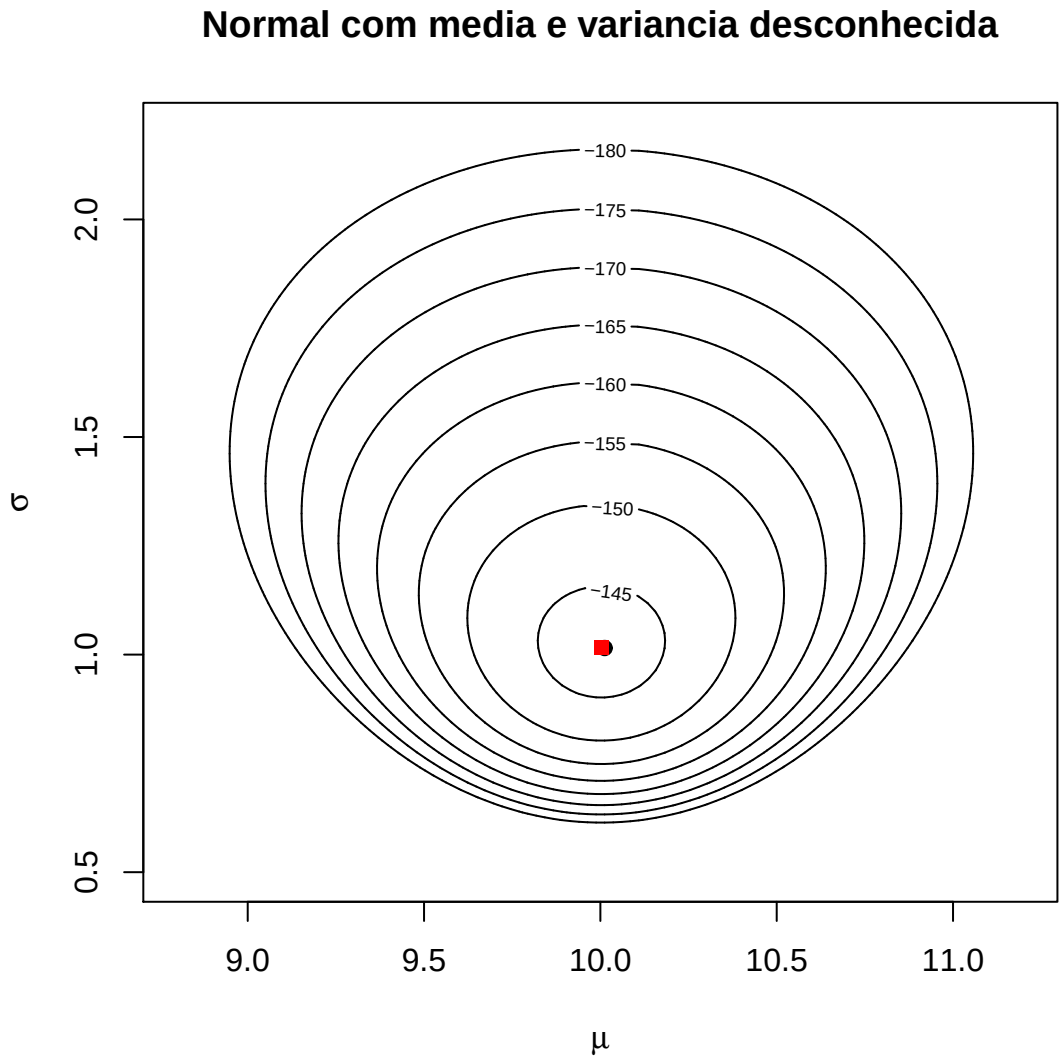
Exemplo 4.1. $X_I \sim N(\mu, \sigma^2)$ com $\theta = (\mu, \sigma)$ e aspaço paramétrico $\Omega = (-\infty, \infty) \times (0, \infty)$. Consequentemente

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}\left\{(x_i-\mu)^2\right\}}$$

$$l(\theta) = \sum_{i=1}^{n} -\frac{1}{2} log 2\pi - log\ \sigma - \frac{1}{2\sigma}^{2} (x_i - \mu)^2$$

$$l(\theta) = -\frac{n}{2} log(2\pi) - nlog\ \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{4.1}$$

Para uma amostra simulada com n=25 $\mu = 10, \sigma - 1$ A Figura 4.1 mostra a log-verossimilhança,sob a forma de contornos, traçado em $l(\theta) = l(\hat{\theta}) - j$ para j=1,.....,5



Note que os contornos são razoavelmente elípticos, com maior e menos ares pralelas os parâmetros, o verdadeiro valor aparece no contorno mais forte, a log-verossimilhança é quadrática com respeito a μ,assim é simétrica sobre $\hat{\mu}$, já a função é assimétrica com respeito a σ. Pode-se ver um único parâmetro, fixando um parâmetro e olhando a secção transversal para o outro. É útil considerar isto porque esta superfície tem contornos razoavelmente elípticos: neste caso mudar a média dos dados não muda a variação, assim não existe relacionamento entre os parâmetros

Exemplo 4.2. $X_i \sim Uniforme(a,b)$, então $\theta = (a,b)$ com $\Omega = \{-\infty < a < b < \infty\}$

$$f(x;\theta) = \begin{cases} (b-a)^{-1} & \text{para } a \leq x \leq b \\ 0 & \text{caso contrário} \end{cases}$$

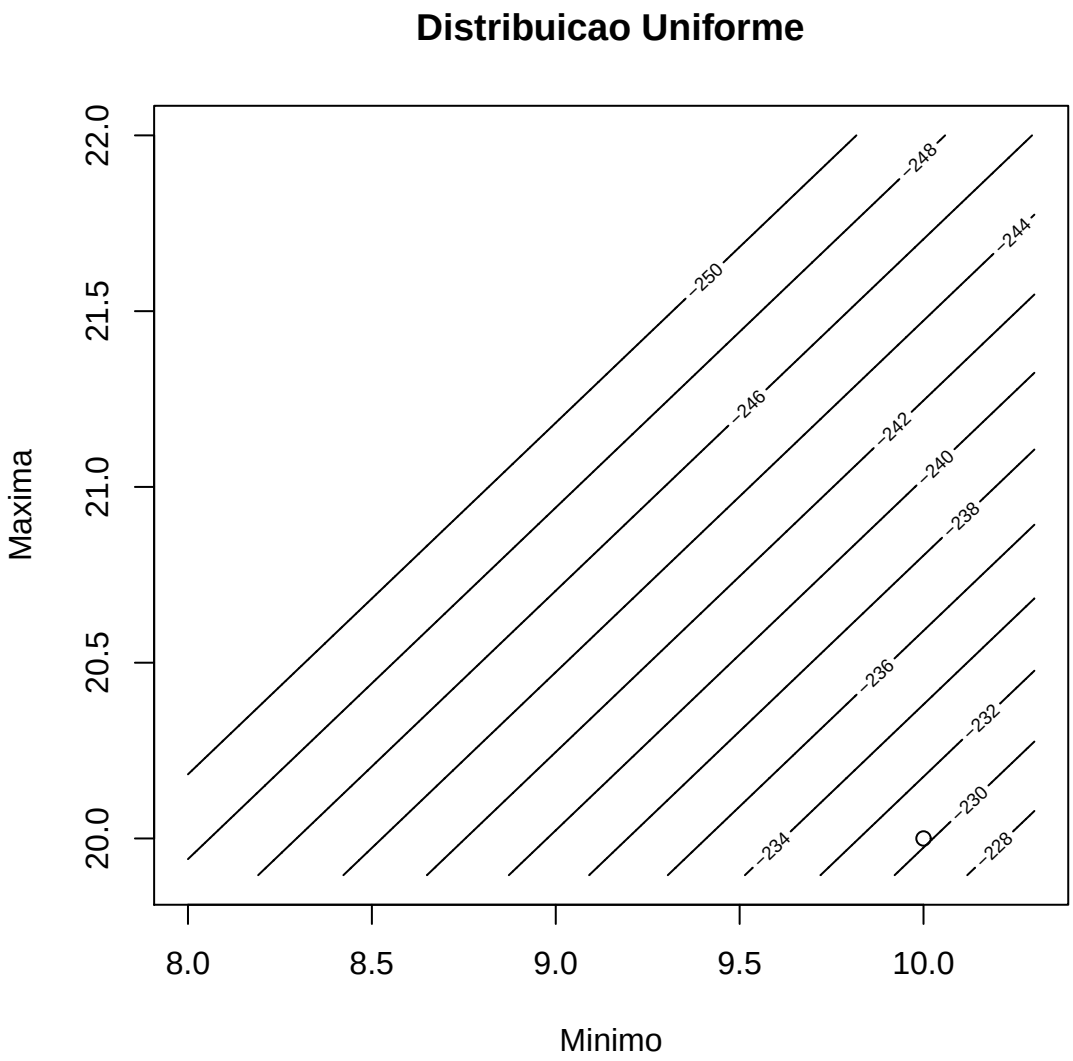
Assim

$$L(\theta) = \begin{cases} (b-a)^{-n} \text{ para } a \leq x_i \leq b \text{ para todo } i = 1, \ldots, n \\ 0 \quad \text{caso contrário} \end{cases}$$

i.e.

$$l(\theta) = \begin{cases} -n\log(b-a) \quad \text{se } a \leq min(x_1, \ldots, x_n) \text{ e } max(x_1, \ldots, x_n) \leq b \\ -\infty \quad \text{caso contrário} \end{cases}$$

Para uma amostra simulada com $n = 25$ e $(a = 0b = 1)$. A Figura 4.2 mostra a log-verossimilhança com seus contornos, delineadas com $l(\theta) = l(\hat{\theta}) - j$ para $j = 1, ...5$.



Aqui $\hat{\theta}\{\hat{a}, \hat{b}\}$ onde $\hat{a} = minx_1, ..., x_n$ e $\hat{b}maxx_1, ..., x_n$e os contornos estão longe de ser elípticos, na

verdade são linhas constantes nos valores de b-a. A linha pontilhada abaixo é a verossimilhança em $-infty$. Para a região de confiança precisamosda deviance e do valor d c^* . Aqui

$$D(\theta) = n\{\log(b-a) - \log[max(x_1,\ldots,x_n - min(x_1,\ldots,x_n))]\} \leq c^*$$

Assim a linha $b = a + \{max(x_1,...,x_n) - min(x_1,...,x_n)\}exp\{c^*/2n\}$ da forma ao lado da região, ou seja, a região da confiança é triangular com a melhor estimativa sendo um cantodestaregião. O valor de c^* não pode ser obtida por método simples para esse exemplo, para resolver precisamos de métodos similares aos usados para $U(0,\theta)$ como no Exemplo 3.2 com θ=6.

Exemplo 4.3.

$$X \sim \Gamma(\alpha,\beta)$$

$$\theta = (\alpha,\beta) \quad \text{e} \quad \Omega = (0,\infty)\times(0,\infty)$$

$$f(x,\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{\beta x}$$

onde

$$\Gamma(\alpha) = \int_0^\infty s^{\alpha-1}e^{-s}ds \quad \text{e} \quad E(X) = \frac{\alpha}{\beta}$$

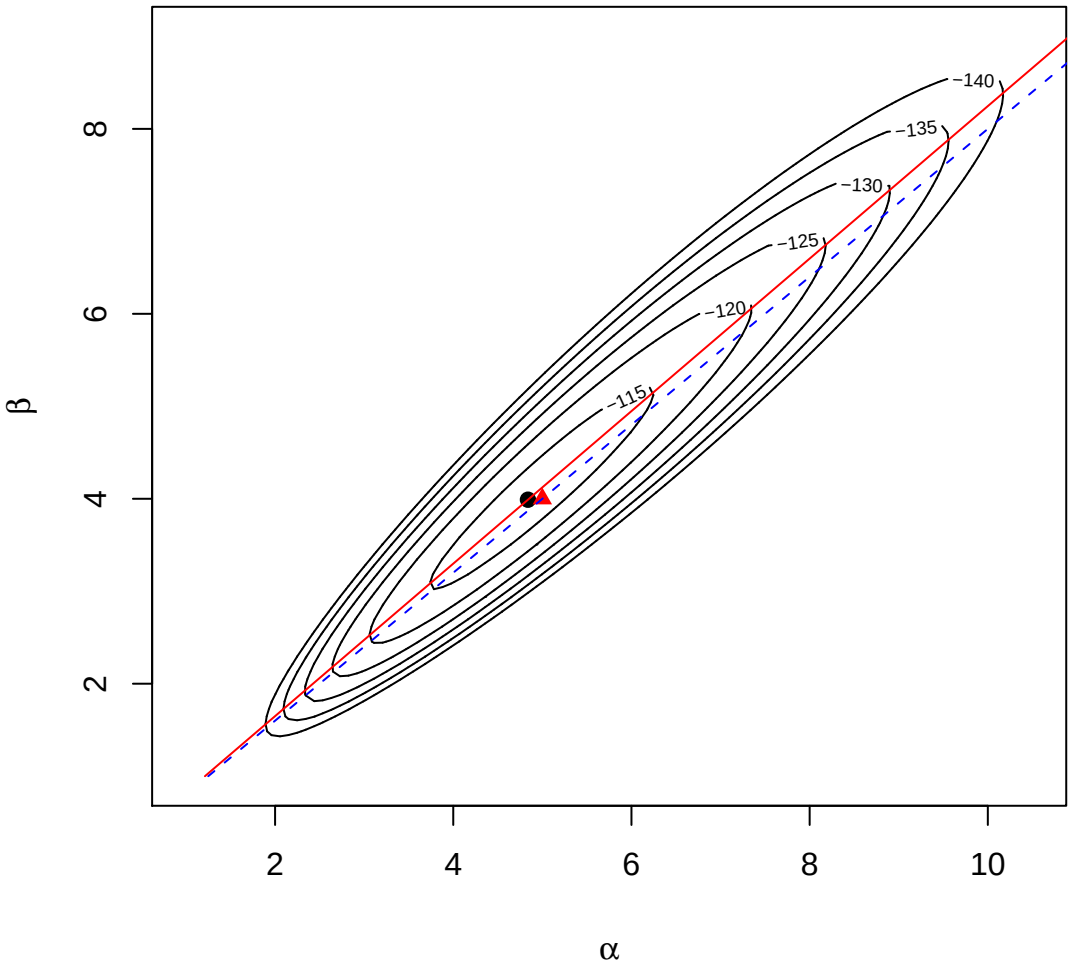
Consequêntemente

$$\begin{aligned} L(\theta) &= \prod_{i=1}^n \left\{ \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{\beta x} \right\} \\ &= \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n e^{-\beta\Sigma x}\prod_{i=1}^n x^{\alpha-1} \end{aligned}$$

$$l(\theta) = n\alpha log(\beta) - nlog[\Gamma(\alpha)] - \beta\sum_{i=1}^n x_i + (\alpha-1)\sum_{i=1}^n log(x_i) \tag{4.2}$$

Para uma amostra simulada com $n = 25$ e $\alpha = 5$ $\beta = 4$. A Figura 4.3 mostra a log-verossimilhança e suas curvas de níveis em forma de contorno em $l(\theta) = l(\hat{\theta}) - j$ para $j = 1,...,5$.

Aqui o valor verdadeiro está dentro do contorno superior, os contornos são elipticos com os maiores e menores eixos não alinhados com o eixo dos parâmetros.
Para $X \sim \Gamma(\alpha,\beta)$ temos.



$$E(X) = \frac{\alpha}{\beta} = \frac{4}{5}$$

Para manter constante o valor de $E[X] = \alpha/\beta$ ao aumentarmos α nós devemos aumentar β igualmente ,assim nós podemos esperar o relacionamento entre estes parâmetros , como mostra no gráfico.

Nota. Se fixarmos $\alpha = 5$ então a abrangencia de valores plausiveis para β será substancialmente reduzido.

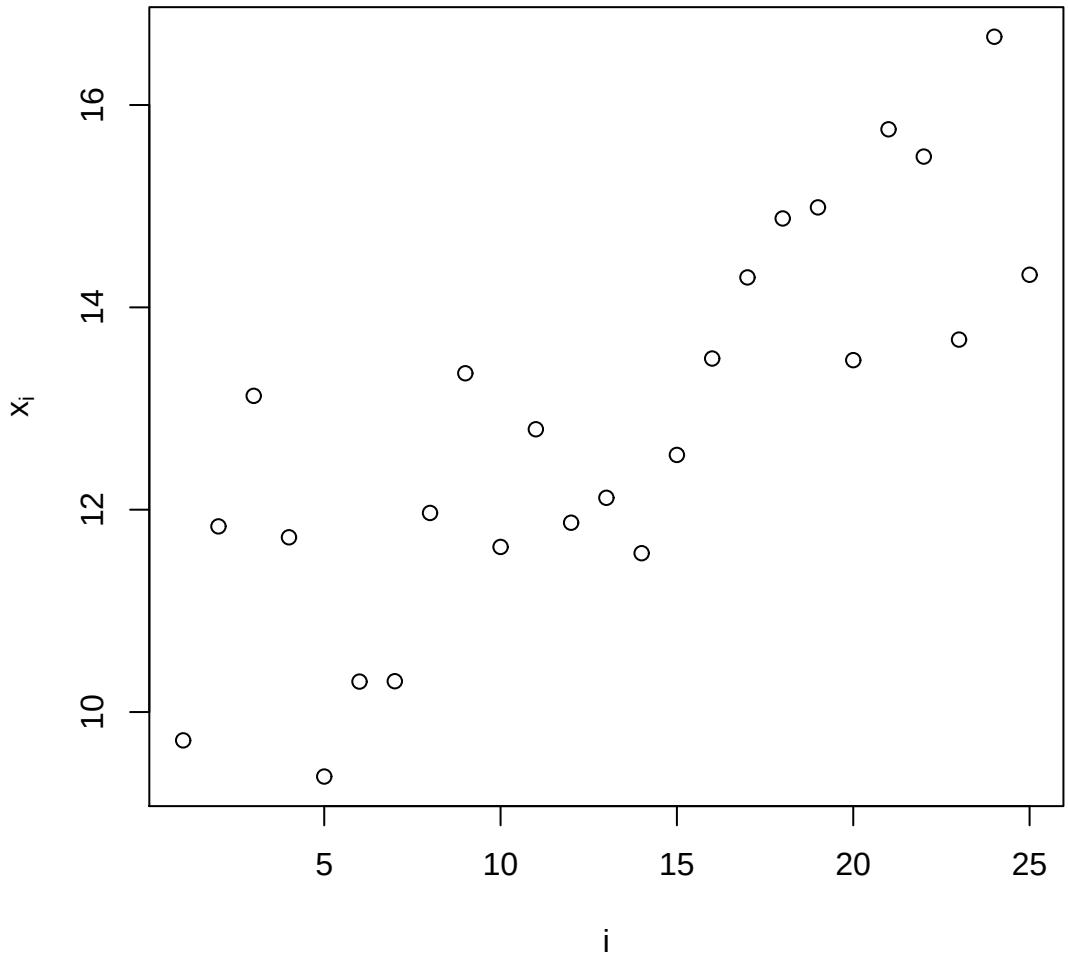
Exemplo 4.4. $X_i \sim N(\alpha + \beta z, \sigma^2)$ i.e. um modelo de regressão com variáveis (conhecidas) explanatórias $(z_1, \ldots, z_n)$, então $\theta = (\alpha, \beta, \sigma)$ e $\Omega = (-\infty, \infty) \times (0, \infty)$. Assim

$$L(\theta) = \prod_{i=1}^{n} \left\{ \frac{1}{(2\pi)^{1/2}\sigma} exp\{-\frac{1}{2\sigma^2}(x_i - \alpha - \beta z_i)^2\} \right\}$$

$$l(\theta) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}(x_i - \alpha - \beta z_i)^2 \right\}$$

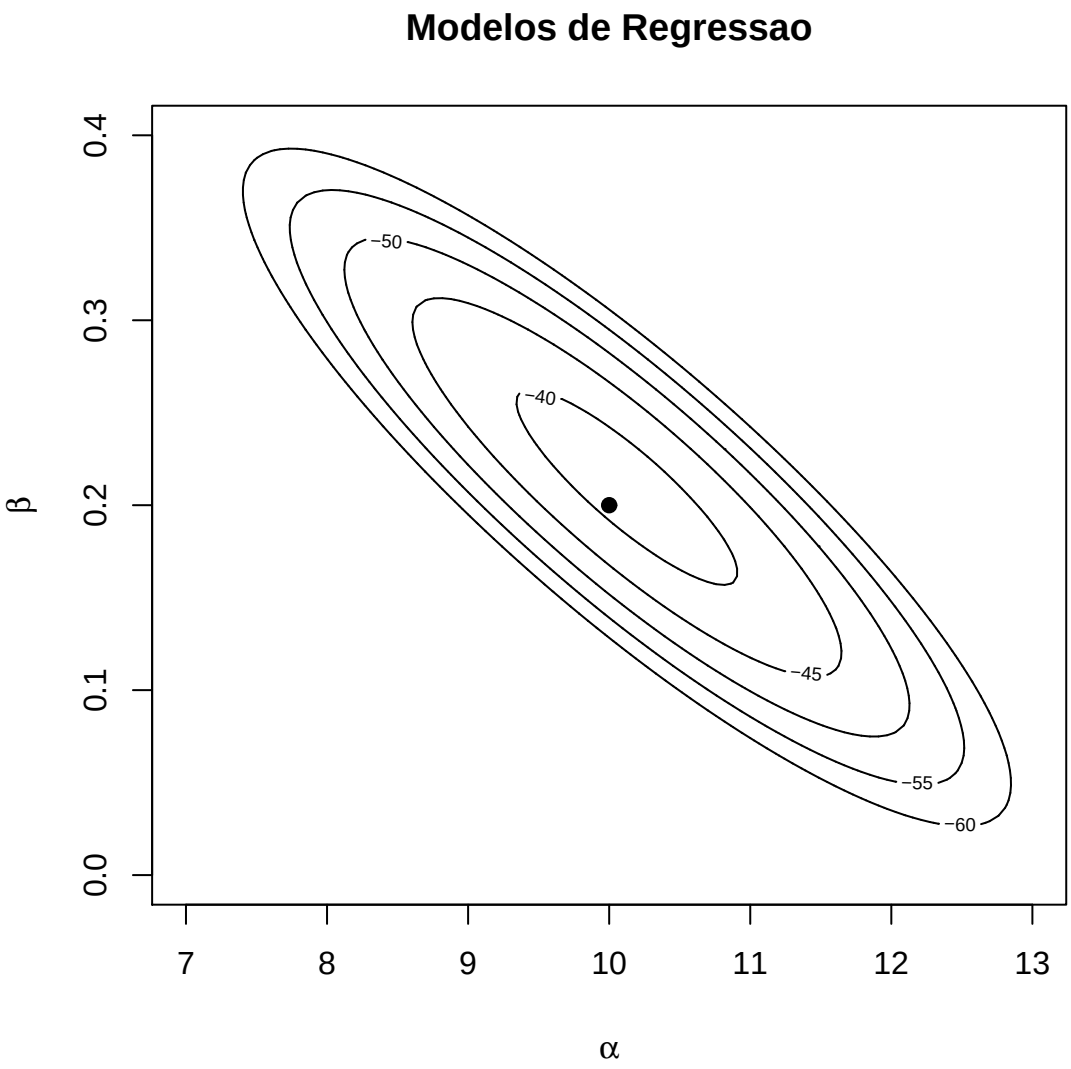
$$l(\theta) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \alpha - \beta z_i)^2 \tag{4.3}$$

Para uma amostra simulada com n=25, $Z_i = i$para todo i,e $(\alpha = 10, \beta = 0,2, \sigma = 1)$. A Figura 4.4 mostra os dados .



Embora este seja um problema de três parâmetros ilustra-se a superfície da log-verossimilhança separadamente para $(\alpha, \beta), (\alpha, \sigma)$ nas Figuras 4.4,4.4 e 4.4 respectivamente. Em cada figura os três parâmetros são tomados como sendo as verdadeiras valores para o determinado parâmetro. Em cada

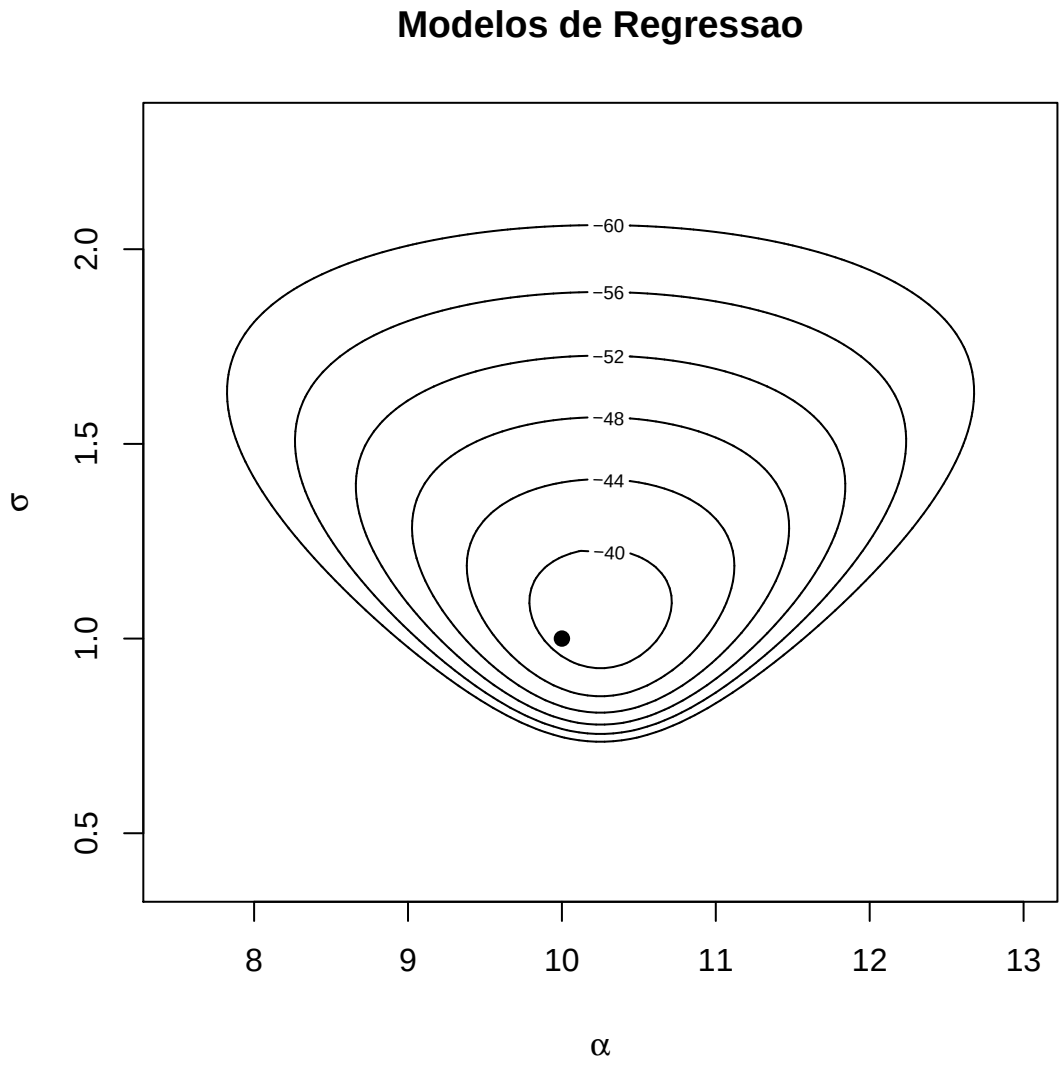
gráfico a função de log-verossimilhança é mostrada em termos de suas curvas de nível, delineadas em $l(\theta) = l(\hat{\theta}) - j$ para j=1,...,5.



Para $(\alpha_i \beta)$ os contornos são elípticos, mas os maiores e menores eixos não são paralelos ao eixo dos parâmetros, e os verdadeiros valores estão no espaço contido na segunda curva de nível.A razão para o relacionamento entre $\alpha e \beta$ é tal que

$$E(X_i) = \alpha + \beta z_i \qquad (= \alpha + \beta_i)$$

Para manter a média constante é necessário que se α é almentado então β tem que diminuir e vice-versa.
Para $(\alpha, \sigma) e (\beta, \sigma)$ as curvas de nível são razoavelmente elípticas como menos e mais valores do eixo, e são paralelos aos parâmetros.



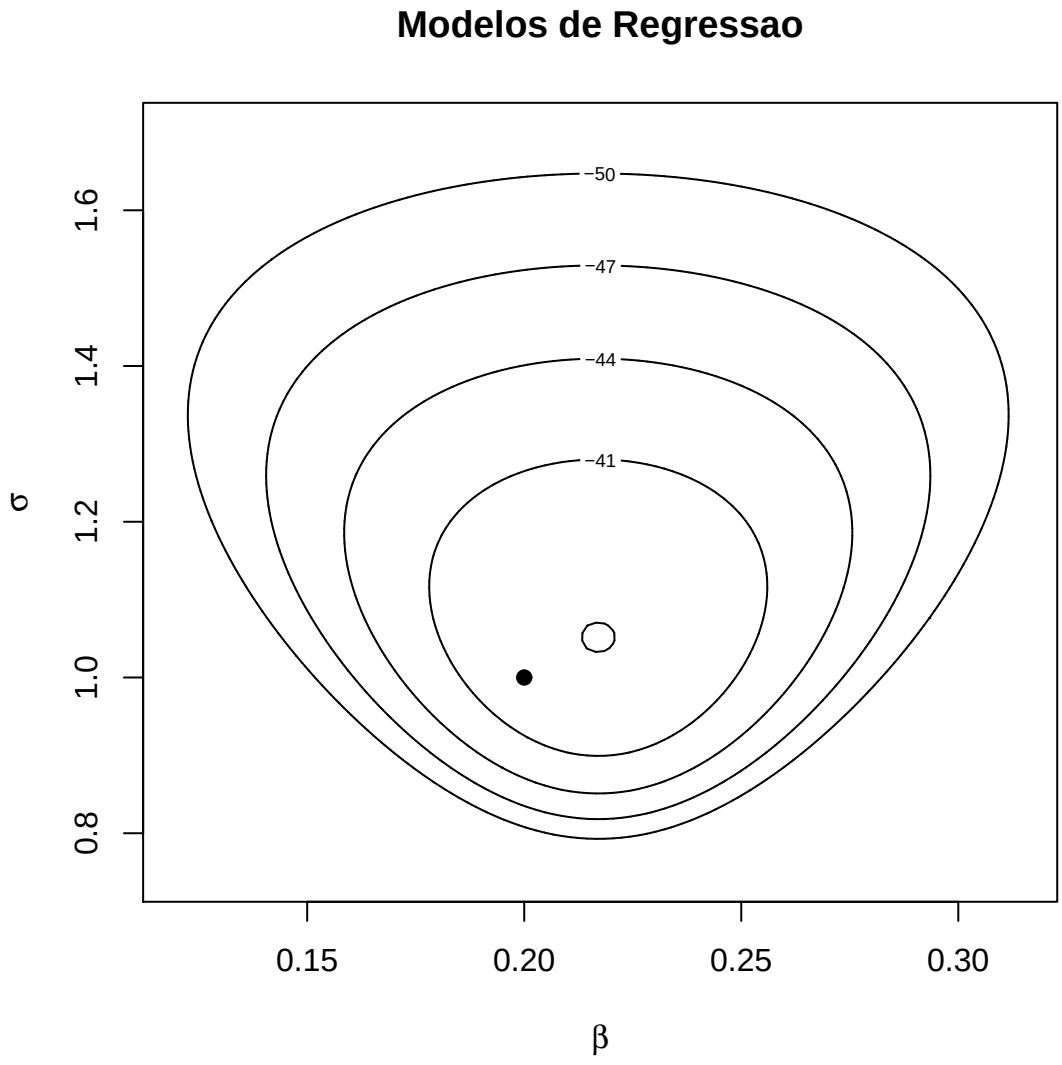
Exemplo 4.5. $X_i \sim N(\alpha^* + \beta^*(z_i - \overline{z}), \sigma^2)$, com $\overline{z} = \sum_{i=1}^{n} z_i$ i.e. um modelo de regressão com a mesma estrutura do modelo no Exemplo 4.4, mas aqui parametrizado por:

$$\beta^* = \beta \text{ e } \alpha^* = \alpha + \beta\overline{z},$$

Com $\theta = (\alpha^*, \beta^*, \sigma)$ e $\Omega = (-\infty, \infty) \times (0, \infty)$.
Aqui α^* é o valor esperado para a média da variável,

$$E(X_{\overline{z}}) = \alpha^*$$

o qual para dados simulados com $Z_i = i$ e n=25 fornece $E(X_{13}) = \alpha^*$. Este parâmetro está menos dependente de β que α, o intercepto. De 4.3 resulta

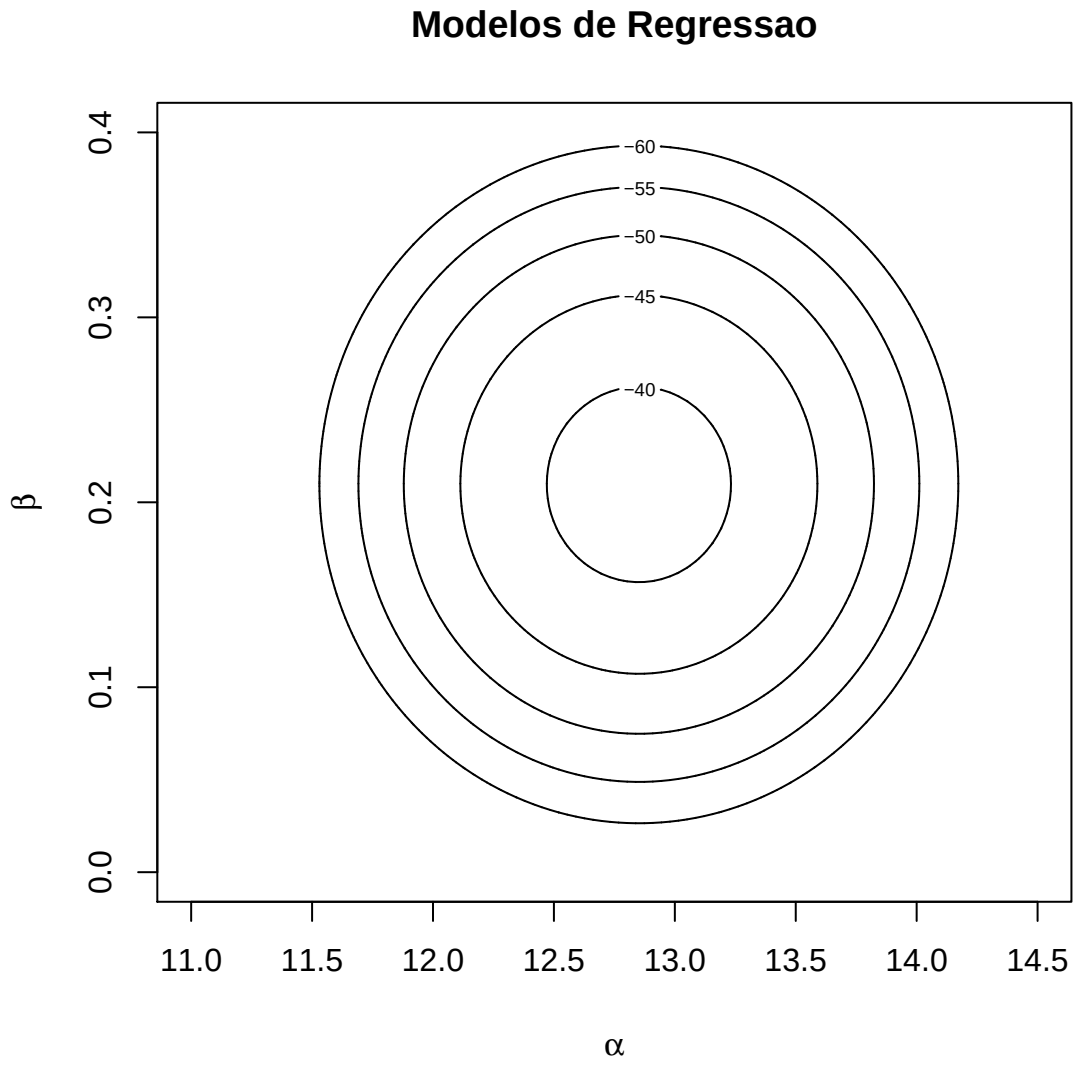


$$l(\theta) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}[x_i - \alpha^* - \beta^*(z_i - \bar{z})]^2$$

Sendo que (α^*, β^*) são uma simples reperametrização de (α, β). A Figura 4.5 mostra a log-verossimilhança baseado nos dados de Exemplo 4.4 com as curvas de nível delineadas em $l(\theta) = l(\hat{\theta}) - j$ para j=1,...,5.

Aqui a log-verossimilhança é muito simples então (α, β) como os contornos são elípticos, mas aqui os eixos maiores e menores são paralelos: isto é porque nossa nova parametrização tomou os eixos na mesma proporção.

Exemplo 4.6. (N_1, N_2, N_3) são variáveis aleatórias multinomiais $(\theta_1, \theta_2, \theta_3)$ onde $\theta_1 + \theta_2 + \theta_3 = 1$, e



$0 \leq \theta_i \leq 1$ para $i = 1,2,3$. Isso possue a função massa de probabilidade.

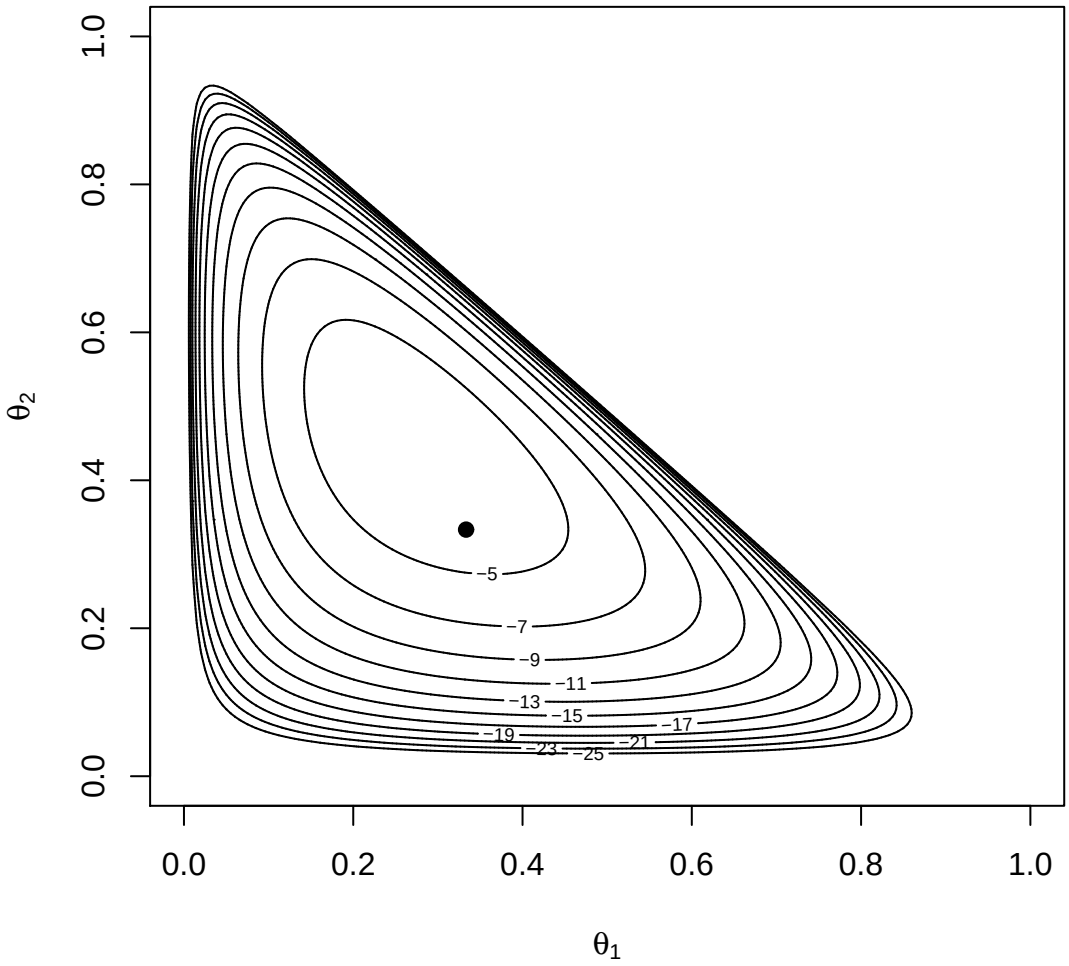
$$p(n_1,n_2,n_3) = \frac{n!}{n_1!n_2!n_3!}\theta_1^{n1}\theta_2^{n2}\theta_3^{n3}$$

onde $n = n_1 + n_2 + n_3$.

MOTIVAÇÃO: Se um experimento tem 32 possíveis resultados com probablidades $\theta_1, \theta_2, \theta_3$. Em n dependentes e identicos ensaios se Ni dentro o número de vezes que ocorre o resultado é então (N_1, N_2, N_3) seguem a distribuição trinomial, $(N_1 + N_2 + N_3) = N$, os valores observados N_1, N_2, N_3.
Note que se existem 2 possibilidades de resultado então a distribuição é Binomial.
A verossimilhança é $L(\theta) = p(N_1, N_2, N_3)$ assim

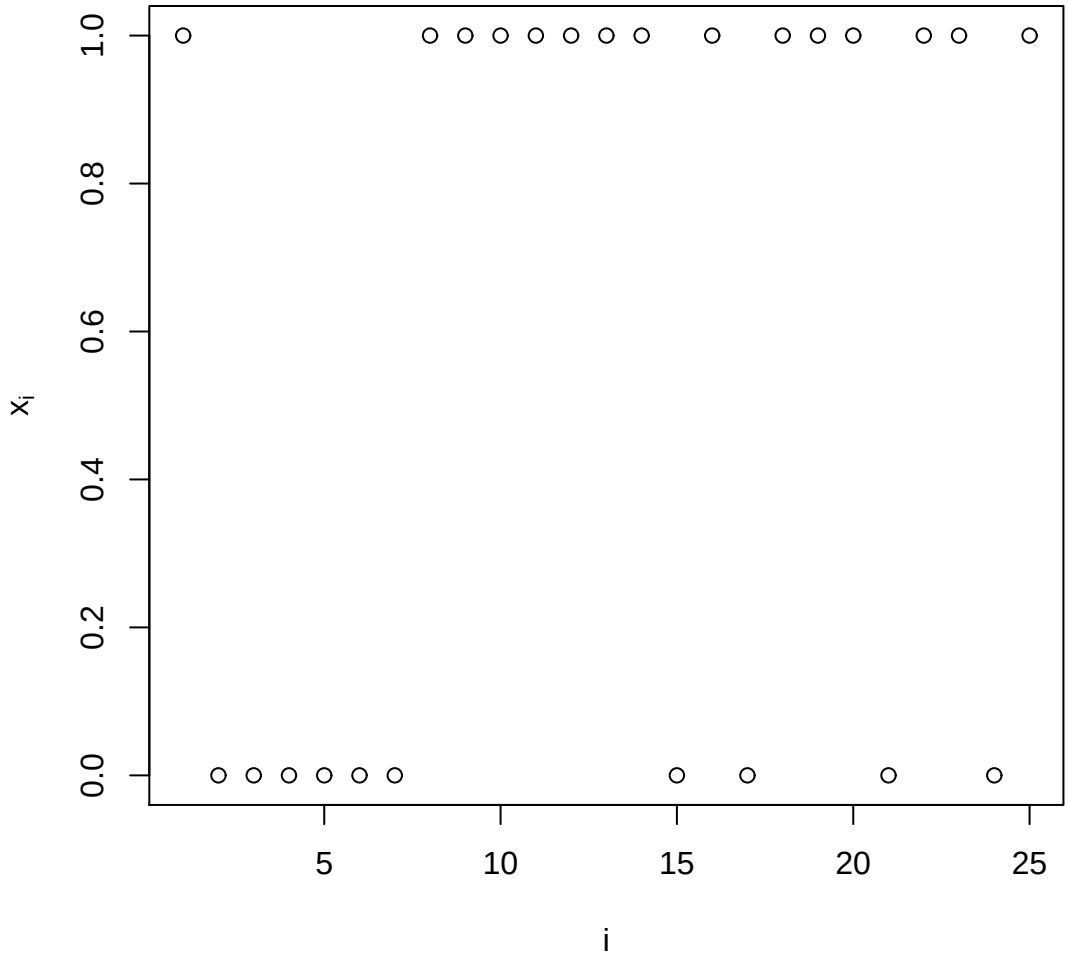
$$l(\theta) = log(n!) - log(n_1!) - log(n_2!) - log(n_3!) + n_1 \log(\theta_1) + n_2 \log(\theta_2) + n_3 \log(1 - \theta_1 - \theta_2).$$

Com $\theta = (\theta_1, \theta_2)$ e $\Omega = \{(\theta_1, \theta_2) : 0 \leq \theta_i \leq 1 i = 1, 2, 3\}$.
Para uma amostra simulada com n=25 e $\theta_1 = 1/3, \theta_2 = 1/3$ a Figura 4.6 mostra a log-verossimilhança, e suas curvas de nível delineadas com $l(\theta) = l(\hat{\theta}) - j$ para j=1,...,5.
Aqui por causa das restrições no espaço paramétrico as curvas de nível são parecidos com um triângulo, mas proximo do máximo os contornos ainda são razoavelmente elípticos.



Exemplo 4.7. Considere os dados mostrados na Figura 4.7. Aqui os valores de X representam uma morte $(n_i - 1)$ ou não $(x_i = 0)$ que pode ocorrer depois de um tratamento e o Z_i são variáveis explicativas, da pessoa i . Aqui a variável explanatória pode ser o peso do indivíduo. Assim na Figura 4.7 obesrva-se

que indivíduos mais pesados tem maior probabilidade de morrer.
Um modelo de regressão comum não apropriado aqui porque a variável observada certamente nãois egue o modelo normal. A variável observada X_i é uma v.a Bernoulli com probabilidade de 1 sendo p.



$$X_i = \begin{cases} 1 & \text{com probabilidade } p_i \\ 0 & \text{com probabilidade } 1-p_i \end{cases}$$

Assim $E(X_i) = 1 \times p_i + 0 \times (1-p_i) = p_i$. Consequentemente a tendência no valor esperado corresponde a tendência na probabilidade, p_i não pode ser uma função linear, porque é restrita a estar non intervalo $[0,1]$. Uma abordagem apropriada é o uso de uma função de ligação para transformar[0,1] em $(-\infty,\infty)$. aqui toma-se a função de ligação como sendo:

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) &= \alpha+\beta z_i \\ &= \alpha^*+\beta^*(z_i-\overline{z})\end{aligned}$$

Usa-se a parametrização alternativa para mostrar uma aplicação do Exemplo **??** assim

$$p_i = \frac{exp\{\alpha^*+\beta^*(z_i-\overline{z})\}}{1+\exp\{\alpha^*+\beta^*(z_i-\overline{z})\}}$$

A verossimilhançapara um indivíduo ser classificado é

$$p_i^{I_i}(1-p_i)^{1-I_i} = \begin{cases} p_i & \text{se } I_i = 1 \\ 1-p_i & \text{se } I_i = 1 \end{cases}$$

onde

$$I_i = \begin{cases} 1 \text{ se morte é a i-esima classificação} \\ 0 \text{ se a i-esima classificação é vida} \end{cases}$$

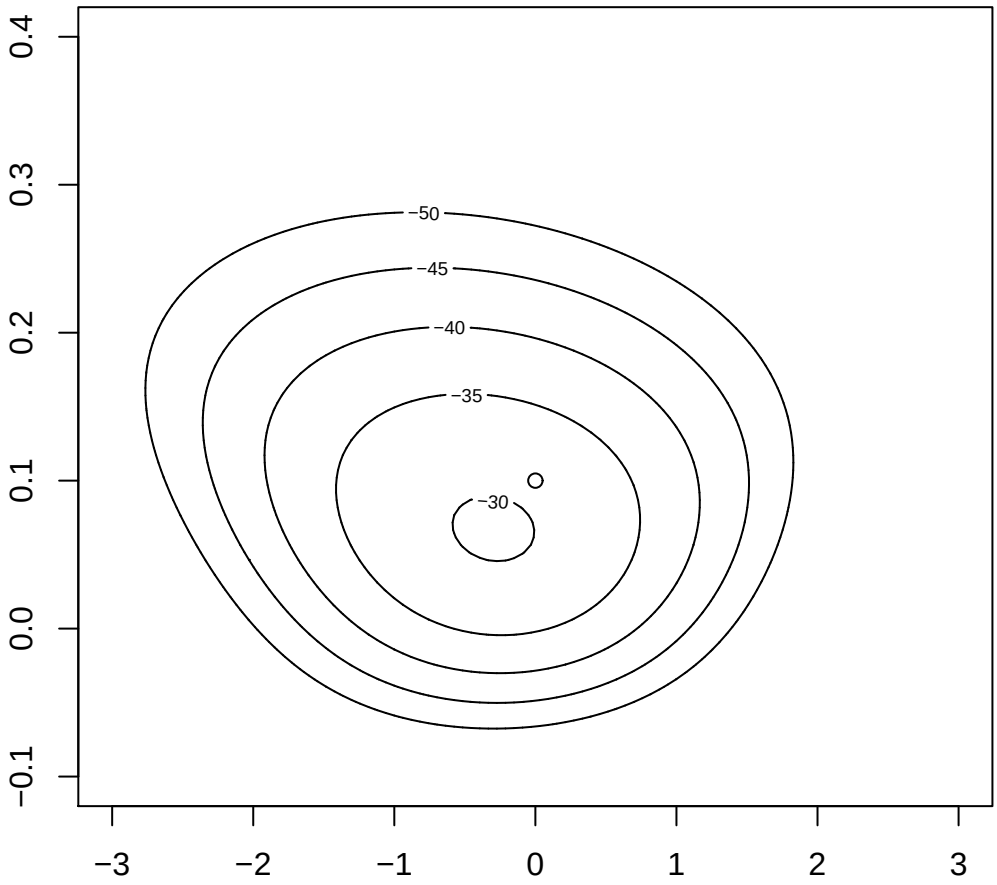
Consequentemente a verossimilhança para classificar todo os indivíduos é

$$L(\theta) = \prod_{i=1}^{n} p_i^{I_i}(1-p)^{1-I_i}$$

Assim

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n}\{I_i log p_i + (1-I_i) log(1-p_i)\} \\ \sum_1 log p_i + \sum_0 log(1-p_i) \end{aligned}$$

Onde $\sum_1$ é a soma dos indivíduos que morreram e $\sum_0$ é a soma dos indivíduos que não morreram. Para uma amostra simulado com $n = 50$ e $(\alpha^* = 0, \beta^* = 0.1)$. A Figura 4.7 mostra a log-verossimilhança e suas curvas de nível delineadas com $l(\theta) = l(\hat{\theta}) - j$ para j=1,....5.
Aqui o verdadeiro valor está dentro da curva superior, e as curvas são são elípticas e praticamente paralelas ao eixo.


−50
−45
−40
−35
−30
0.4
0.3
0.2
0.1
0.0
−0.1
−3
−2
−1
0
1
2
3

CAPÍTULO 5

RESULTADOS PARA VEROSSIMILHANÇA COM MÚLTIPLOS PARÂMETROS

No capítulo 4 examina-se a superfície da verossimilhança para uma variedade de problemas. Vinda do capítulo 2 a definição básica de intervalo de confiança baseado na verossimilhança é o conjunto

$$\{\theta \in \Omega : D(\theta) \leq c^*\}$$

Entretanto, ainda não se sabe como escolher os valores de c^*, no caso d-dimensional; além disso, não se sabe quais características da amostra influenciam a superfície da verossimilhança, a posição do estimador de máxima verossimilhança, e a região de confiança. Enfrentou-se este problema no Capítulo 3 e tem-se um progresso através da função de log-verossimilhança aproximada, usando resultados para quando $n \to \infty$. .Isto corresponde a aproximar a superfície de verossimilhança por uma versão d- dimensional da aproximação quadrática usada para verossimilhança simples. Ilustra-se isso para os problemas com dois parâmetros.

Exemplo 5.1. $X_i \sim N(\mu, \sigma^2)$, então $\theta = (\mu, \sigma)$. Da equação ref()

$$l(\theta) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

e da Figura 4.1observa-se que os contornos são razoavelmente elipticos, que corresponde a uma superfície sendo aproximada por uma superfície quadrática bidimensional.Assim, considera-se a aproximação da log-verossimilhança usando séries de Taylor em torno do estimador de máxima verossimilhança. Primeiro encontra-se o estimador de máxima verossimilhança e as suas derivadas dadas por:

$$\frac{\partial}{\partial\mu}l(\theta)=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)$$
$$\frac{\partial}{\partial\mu}l(\theta)=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2$$

Para obter o máximo precisa-se $\theta=\hat{\theta}$ onde

$$\frac{\partial}{\partial\mu}l(\hat{\theta})=0 \text{ e } \frac{\partial}{\partial\sigma}l(\hat{\theta})=0.$$

Consequentemente tem-se que

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu})=0 \text{ então } \hat{\mu}=\bar{x}$$

e para σ tem-se

$$0=-\frac{n}{\hat{\sigma}}+\frac{1}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i-\hat{\mu})^2=0 \text{ então } \hat{\sigma}=\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right)^{1/2}.$$

As segundas derivadas são:

$$\frac{\partial^2}{\partial\mu^2}l(\theta)=-\frac{n}{\sigma^2}$$
$$\frac{\partial^2}{\partial\sigma^2}l(\theta)=\frac{n}{\sigma^2}-\frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$
$$\frac{\partial^2}{\partial\mu\sigma}l(\theta)=-\frac{2}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu),$$

que avaliados nos estimadores de máximaverossimilhança são

$$\frac{\partial^2}{\partial\mu^2}l(\hat{\theta})=-\frac{n}{\hat{\sigma}^2}$$
$$\frac{\partial^2}{\partial\sigma^2}l(\hat{\theta})=\frac{n}{\hat{\sigma}^2}-\frac{3}{\hat{\sigma}^4}\sum_{i=1}^{n}(x_i-\mu)^2=\frac{n}{\hat{\sigma}^2}-\frac{3n}{\hat{\sigma}^4}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{n}{\hat{\sigma}^2}-\frac{3n\hat{\sigma}^2}{\hat{\sigma}^4}=-\frac{2n}{\hat{\sigma}^2}$$
$$\frac{\partial^2}{\partial\mu\sigma}l(\hat{\theta})=-\frac{2}{\hat{\sigma}^3}\sum_{i=1}^{n}(x_i-\hat{\mu}),$$

Assim a aproximação em séries de Taylor é

$$\begin{aligned}l(\theta) &\approx l(\hat{\theta})+(\mu-\hat{\mu})\frac{\partial}{\partial\mu}l(\hat{\theta})+(\sigma-\hat{\sigma})\frac{\partial}{\partial\sigma}l(\hat{\theta})+\\ &\quad+\frac{1}{2}(\mu-\hat{\mu})^2\frac{\partial^2}{\partial\mu^2}l(\hat{\theta})+\frac{1}{2}(\sigma-\hat{\sigma})^2\frac{\partial^2}{\partial\sigma^2}l(\hat{\theta})+(\mu-\hat{\mu})(\sigma-\hat{\sigma})\frac{\partial^2}{\partial\mu\partial\sigma}l(\hat{\theta})\\ &\approx l(\hat{\theta})+0+0-\frac{n}{2\hat{\sigma}^2}(\mu-\hat{\mu})^2-\frac{n}{\hat{\sigma}^2}(\sigma-\hat{\sigma})^2+0.\end{aligned}$$

Sendo assim a *deviance* é

$$D(\theta) \approx \frac{n}{\hat{\sigma}^2}(\mu - \hat{\mu})^2 + \frac{2n}{\hat{\sigma}^2}(\sigma - \hat{\sigma})^2,$$

que é a equação de uma elipse com os maiores e menores eixos paralelos ao eixo dos parâmetros. Note que comprimento dos eixos maior e menor são determinados pelas segundas derivadas em relação a cada um dos parâmetros.
Os contornos mostrados para $l(\theta) = l(\hat{\theta}) - j$ são equivalentes para $D(\theta)/2 = 1, \ldots, 5$. Do apêndice sobre elipses (veja Seção 6.2), parece que a razão para os contornos da log-verossimilhança serem paralelos aos eixos paramétricos é que o termo do produto cruzado não está presente em $D(\theta)$ o qual deve-se porque

$$\frac{\partial^2}{\partial\mu\partial\sigma} l(\hat{\theta}) = 0,$$

a esta propriedade denomina-se *ortogonalidade* entre os parâmetros.

Exemplo 5.2. $X_i \sim N(\alpha + \beta z_i, \sigma^2)$. Da equação ref(), fizando $\sigma = 1$ tem-se

$$l(\theta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \alpha - \beta z_i)^2.$$

No Capítulo 3 explicou-se que suas curvas de nível eram elípticas, mas não paralelas aos eixos. Aqui examinaremos como encontrar uma superfície para o estimador de máxima verossimilhança através da aproximação da função de log-verossimilhança. Inicialmente obtem-se os estimadores de máxima verossimilhança e suas derivadas por:

Para obter o máximo precisaremos $\theta = \hat{\theta}$ onde

$$\frac{\partial}{\partial\alpha} l(\hat{\theta}) = 0 \text{ e } \frac{\partial}{\partial\beta} l(\hat{\theta}) = 0.$$

Aqui tem-se que

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(z_i - \overline{z})}{\sum_{i=1}^{n}(z_i - \overline{z})^2}$$

e

$$\hat{\alpha} = \overline{x} - \hat{\beta}\overline{z}$$

(veja MATH 235). As segundas derivadas são

$$\frac{\partial^2}{\partial\alpha^2}l(\theta) = -n$$
$$\frac{\partial^2}{\beta^2}l(\theta) = -\sum_{i=1}^{n} z_i^2$$
$$\frac{\partial^2}{\partial\alpha\partial\beta}l(\theta) = -\sum_{i=1}^{n} z_i$$

Além disso a série de Taylor aproximada é

$$\begin{aligned} l(\theta) &\approx l(\hat{\theta}) + (\alpha-\hat{\alpha})\frac{\partial}{\partial\alpha}l(\hat{\theta}) + (\beta-\hat{\beta})\frac{\partial}{\partial\beta}l(\hat{\theta}) + \\ &\quad + \frac{1}{2}(\alpha-\hat{\alpha})^2\frac{\partial^2}{\partial\alpha^2}l(\hat{\theta}) + \frac{1}{2}(\beta-\hat{\beta})^2\frac{\partial^2}{\partial\beta^2}l(\hat{\theta}) + (\alpha-\hat{\alpha})(\beta-\hat{\beta})\frac{\partial^2}{\partial\alpha\partial\beta}l(\hat{\theta}) \\ &\approx l(\hat{\theta}) + 0 + 0 - \frac{n}{2}(\alpha-\hat{\alpha})^2 - \frac{1}{2}\sum_{i=1}^{n} z_i^2(\beta-\hat{\beta})^2 - \sum_{i=1}^{n} z_i(\alpha-\hat{\alpha})(\beta-\hat{\beta}). \end{aligned}$$

Assim a Deviance é

$$D(\theta) \approx n(\alpha-\hat{\alpha})^2 + \sum_{i=1}^{n} z_i^2(\beta-\hat{\beta})^2 + 2\sum_{i=1}^{n} z_i(\alpha-\hat{\alpha})(\beta-\hat{\beta}).$$

Veja que aqui o produto cruzado da equação da elipse não é zero, esta é a razão para as curvas de níveis não serem paralelas aos eixos dos parâmetros.

Exemplo 5.3. $X_i \sim N(\alpha+\beta(z_i-\bar{z}),\sigma^2)$. Com $\sigma = 1$ fixo tem-se

$$l(\theta) = -\frac{n}{2}\log(2\pi)\sum_{i=1}^{n}(x_i-\alpha^*-\beta^*(z_i-\bar{z}))^2.$$

Segue que

$$\frac{\partial^2}{\partial\alpha^*\partial\beta^*}l(\theta) = -\sum_{i=1}^{n}(z_i-\bar{z}) = 0,$$

que explica o porque que esta reparametrização faz a log-verosimilhança serem curvas de níveis paralelos aos eixos.

Exemplo 5.4. $X_i \sim \Gamma(\alpha,\beta)$. Recordando a expressão 4.2 tem-se

$$l(\theta) = n\alpha\log\beta + (\alpha-1)\sum_{i=1}^{n}\log x_i - \beta\sum_{i=1}^{n} x_i - n\log\Gamma(\alpha).$$

Aqui

$$\frac{\partial}{\partial\alpha}l(\theta) = n\log\beta + \sum_{i=1}^{n}\log x_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)},$$

e

$$\frac{\partial}{\partial\beta}l(\theta) = \frac{n\alpha}{\beta} - \sum_{i=1}^{n} x_i,$$

então para encontrar o estimador de máxima verossimilhança precisa-se resolver

$$\frac{\partial}{\partial\alpha}l(\hat{\theta}) = 0 \text{ e } \frac{\partial}{\partial\beta}l(\hat{\theta}) = 0.$$

Aqui

$$\hat{\beta} = \frac{\hat{\alpha}}{\bar{x}},$$

e $\hat{\alpha}$ satisfaz

$$n\log\hat{\alpha} - n\log\bar{x} + \sum_{i=1}^{n}\log x_i - n\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0,$$

que necessite claramente de alguma solução numérica. As segundas derivadas em $\hat{\theta}$ são

$$\begin{aligned}
\frac{\partial^2}{\partial\alpha^2}l(\hat{\theta}) &= -n\left[\frac{\Gamma''(\hat{\alpha})}{\Gamma(\hat{\alpha})} - \left(\frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}\right)^2\right] \\
\frac{\partial^2}{\beta^2}l(\hat{\theta}) &= -\frac{n\hat{\alpha}}{\hat{\beta}^2} \\
\frac{\partial^2}{\partial\alpha\partial\beta}l(\hat{\theta}) &= \frac{n}{\hat{\beta}}.
\end{aligned}$$

Note que aqui a derivada mista é positiva, enquanto que a derivada mista no Exemplo 4.4 foi negativa, explicando a diferença na direçao do gradiente em relação aos maiores eixos nos dois exemplos.

5.1 Resultados e Notação

Os resultados são muito similares aos da Seção 2.3, exceto a notação da função escore que é mais difícil, a função $U(\theta)$ é substituída por um vetor $\boldsymbol{U}(\theta)$

$$\begin{aligned}
\boldsymbol{U}(\theta) &= (U_1(\theta),\ldots,U_d(\theta))^T \\
&= \left(\frac{\partial}{\partial\theta_1}l(\theta),\ldots,\frac{\partial}{\partial\theta_d}l(\theta)\right)^T,
\end{aligned}$$

é o vetor gradiente para a log-verossimilhança. Também os termos informação $I_O(\theta)$ e $I_E(\theta)$ são substituídos pelas matrizes

$$\mathbf{I}_O(\theta) = \begin{pmatrix} -\frac{\partial^2}{\partial\theta_1^2}l(\theta) & \dots & \dots & -\frac{\partial^2}{\partial\theta_1\partial\theta_d}l(\theta) \\ \dots & \dots & -\frac{\partial^2}{\partial\theta_i\partial\theta_j}l(\theta) & \dots \\ \dots & -\frac{\partial^2}{\partial\theta_j\partial\theta_i}l(\theta) & \dots & \dots \\ -\frac{\partial^2}{\partial\theta_d\partial\theta_1}l(\theta) & \dots & \dots & -\frac{\partial^2}{\partial\theta_d^2}l(\theta) \end{pmatrix}$$

e

$$\mathbf{I}_E(\theta) = \begin{pmatrix} E\left\{-\frac{\partial^2}{\partial\theta_1^2}l(\theta)\right\} & \dots & \dots & E\left\{-\frac{\partial^2}{\partial\theta_1\partial\theta_d}l(\theta)\right\} \\ \dots & \dots & E\left\{-\frac{\partial^2}{\partial\theta_i\partial\theta_j}l(\theta)\right\} & \dots \\ \dots & E\left\{-\frac{\partial^2}{\partial\theta_j\partial\theta_i}l(\theta)\right\} & \dots & \dots \\ E\left\{-\frac{\partial^2}{\partial\theta_d\partial\theta_1}l(\theta)\right\} & \dots & \dots & E\left\{-\frac{\partial^2}{\partial\theta_d^2}l(\theta)\right\} \end{pmatrix}$$

a matriz Hessiana, a esperança da matriz Hessiana, para a log-verossimilhança.
Lema 4: $E\{\mathbf{U}(\theta)\} = 0$ e $Var\mathbf{U}(\theta) = E\{\mathbf{I}_O(\theta)\} = \mathbf{I}_E(\theta)$.
Prova: Similar ao Lema 1.

Note que a variância do vetor $(U_1, ...U_d)^T$ é a matriz com entradas

$$\begin{pmatrix} Cov(U_1,U_1) & \dots & \dots & Cov(U_1,U_d) \\ \dots & \dots & Cov(U_i,U_j) & \dots \\ \dots & Cov(U_j,U_i) & \dots & \dots \\ Cov(U_d,U_1) & \dots & \dots & Cov(U_d,U_d) \end{pmatrix}$$

onde $cov(U_i,U_i) = Var(U_i)$. Uma propriedade geral de $\mathbf{I}_O(\hat{\theta})$ e $\mathbf{I}_E(\theta)$ é que elas são matrizes definidaz positivas, as quais mensuram a curvatura observada de $\hat{\theta}$ e a curvatura esperada, em θ, nas respectivas superfícies de log-verossimilhança.

Exemplo 5.5. $X_i \sim N(\mu,\sigma^2)$ então $\theta = (\mu,\sigma)$. Aqui

$$\mathbf{U}(\theta) = \left\{ \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu), -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2 \right\}^T$$

e

$$\mathbf{I}_O(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{2}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu) \\ \frac{2}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu) & \frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{\sigma^2} \end{pmatrix}$$

então

$$\mathbf{I}_O(\hat{\theta}) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\hat{\sigma}^2} \end{pmatrix}.$$

Desde que

$$E\{\sum_{i=1}^{n}(X_i-\mu)\} = \sum_{i=1}^{n}(E(X_i)-\mu) = 0$$

e

$$E\{\sum_{i=1}^{n}(X_i-\mu)^2\} = \sum_{i=1}^{n}E\{(X_i-\mu)^2\} = n\sigma^2,$$

tem-se que

$$\mathbf{I}_E(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix}.$$

Exemplo 5.6. $X_i \sim \Gamma(\alpha,\beta)$. Aqui

$$\mathbf{U}(\theta) = \left\{ n\log\beta + \sum_{i=1}^{n}\log x_i - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}, \frac{n\alpha}{\beta} - \sum_{i=1}^{n}x_i \right\}^T$$

e

$$\mathbf{I}_O(\theta) = \mathbf{I}_E(\theta) = \begin{pmatrix} n\left[\Gamma''(\alpha)/\Gamma(\alpha) - (\Gamma'(\alpha)/\Gamma(\alpha))^2\right] & -n/\beta \\ -n/\beta & n\alpha/\beta \end{pmatrix}.$$

Finalmente, retornando ao caso geral multi-parâmetrico precisa-se das extensões (2.3) e (2.4) para $l(\theta)$, ou seja, a expansão de Taylor multi-parâmetros. As correspondentes expansões, dadas na Seção 6.1, em termos de notação de verossimilhança são:

$$l(\theta) \approx l(\hat{\theta}) + (\theta-\hat{\theta})^T\mathbf{U}(\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})^T\mathbf{I}_O(\theta^*)(\theta-\hat{\theta}) \text{ para } |\theta^*-\theta| \le |\hat{\theta}-\theta| \tag{5.1}$$

$$\mathbf{U}(\theta) = -\mathbf{U}(\hat{\theta}) - (\theta-\hat{\theta})^T\mathbf{I}_O(\theta^+) \text{ para } |\theta^+-\theta| \le |\hat{\theta}-\theta|, \tag{5.2}$$

onde (5.2) é uma equação vetorial. Da equação (5.1) tem-se que a deviance satisfaz

$$\begin{aligned} D(\theta) &\approx 2[l(\hat{\theta}) - \{l(\hat{\theta}) + (\theta-\hat{\theta})^T\mathbf{U}(\hat{\theta}) - \frac{1}{2}(\theta-\hat{\theta})^T\mathbf{I}_O(\theta^*)(\theta-\hat{\theta})\}] \\ &\approx (\theta-\hat{\theta})^T\mathbf{I}_O(\theta^*)(\theta-\hat{\theta}) \\ &\approx (\theta-\hat{\theta})^T\mathbf{I}_O(\hat{\theta})(\theta-\hat{\theta}), \end{aligned}$$

então $D(\theta)$ pode ser positiva desde que $\mathbf{I}_O(\hat{\theta})$ seja uma matriz definida positiva.

5.2 Principais Resultados

Esta seção foca-se em resultados de teoremas, mas não através de demonstrações as quais apenas serão dadas completamente e como referência na Seção 4.2.1.

Teorema 5.1. - *Para um problema de estimação regular, no limite com $n \Rightarrow \infty$, se θ é o verdadeiro valor do parâmetro, então*

$$\hat{\theta} \sim NM_d(\theta, \boldsymbol{I}_E(\theta)^{-1}),$$

ou seja, a distribuição assintótica de $\hat{\theta}$ é uma normal multivariada com matriz de variância-covariância dada pela inversa da matriz de informação esperada.

Corolário. - *Qualquer termo assintoticamente equivalente a $\boldsymbol{I}_E(\theta)$ pode ser usado no Teorema 4, então*

$$\begin{aligned}\hat{\theta} &\sim NM_d(\theta, \boldsymbol{I}_E(\hat{\theta})^{-1}) \\ \hat{\theta} &\sim NM_d(\theta, \boldsymbol{I}_O(\theta)^{-1}) \\ \hat{\theta} &\sim NM_d(\theta, \boldsymbol{I}_E(\hat{\theta})^{-1}).\end{aligned}$$

Teorema 5.2. - *Para um problema regular de estimação, no limite com $n \rightarrow \infty$, se θ é o verdadeiro valor do parâmetro então*

$$D(\theta) = 2[l(\hat{\theta}) - l(\theta)] \sim \chi^2_d.$$

5.2.1 Prova do Principais Resultados

Três resultados técnicos para vetores aleatórios são necessários para as provas dos Teoremas 5.1 e 5.2.

Resultado 1: Para um vetor aleatório d-dimensional Y com matriz de variância-covariância Σ, e A uma matriz dxd então

$$Var(AY) = A\Sigma A^T,$$

então, por exemplo, se $A = \Sigma^{-1}$ então

$$\begin{aligned}Var(\Sigma^{-1}Y) &= \Sigma^{-1}\Sigma(\Sigma^{-1})^T \\ &= \Sigma^{-1}\Sigma(\Sigma^{-1}) \text{ desde que } \Sigma^{-1} \text{ seja simétrica} \\ &= \Sigma^{-1}.\end{aligned}$$

Resultado 2: (Teorema Central de Limite Multivariado) Suponha que Y é um vetor d-dimensional com vetor de média μ e matriz de variância-covariância Σ. Se $Y_1, \ldots, Y_n$, é uma sequência i.i.d de valores aleatórios seguindo a mesma distribuição de Y, então se

$$S_n = \sum_{i=1}^{n} Y_i \text{ (componente soma do primeiro } n)$$

tem-se que

$$n^{-1/2}[S_n - E(S_n)] = n^{-1/2}[S_n - n\mu] \sim NM_d(0,\Sigma) \text{ com } n \to \infty.$$

Resultado 3: Se $Y \sim NM_d(0,\Sigma)$ então $Y^T\Sigma^{-1}Y \sim \chi^2_d$.
Prova do Teorema 5.1: Usando a expressão (5.2), desde que $\hat{\theta} \to \theta$ então $\theta^+ \to \theta$ e $\mathbf{U}(\hat{\theta}) = 0$, tem-se que

$$\mathbf{U}(\theta) \approx (\hat{\theta} - \theta)^T \mathbf{I}_O(\theta).$$

Usando a aproximação $\mathbf{I}_E(\theta) \approx \mathbf{I}_O(\theta)$ tem-se que

$$(\hat{\theta} - \theta) \approx \mathbf{I}_E(\theta)^{-1}\mathbf{U}(\theta).$$

Do Resultado 1 e do Lema 4 tem-se que a variância do vetor no lado direito é $\mathbf{I}_E(\theta)$ e a média é 0. Como **U** é a soma de variáveis aleatórias então o resultado do Teorema Central do Limite Multivariado (resultado 2).
Prova do Teorema 5.2: Usando a expressão (5.1) desde que $\hat{\theta} \to \theta$ então $\theta^* \to \theta$ e $mbox\mathbf{U}(\hat{\theta}) = 0$, tem-se

$$D(\theta) \approx (\hat{\theta} - \theta)^T \mathbf{I}_E(\theta)(\hat{\theta} - \theta).$$

Do Teorema 5.1 $\hat{\theta} - \theta \sim NM_d(0, \mathbf{I}_E(\theta)^-1)$, então usando o Resultado 3 segue o resultado.

5.3 Discussão dos Principais Resultados

Os Teoremas 5.1 e 5.2 resultam que:

- O estimador de máxima verossimilhança $\hat{\theta}$ de θ é assintóticamente não-viciado, isto é, $E(\hat{\theta}) \to \theta$.
- Assintoticamente $V(\hat{\theta}) \to \mathbf{I}_E(\theta)^{-1}$, o qual, por uma versão multivariada do limite de Cramer-Rao, é o melhor possível.
- Se $J = \mathbf{I}_E(\theta)^{-1}$, então $Var(\hat{\theta}) = J$, sendo que J é uma matriz simétrica e definida positiva, com elementos $J_{i,j} = Cov(\hat{\theta}_i, \hat{\theta}_j)$ então $J_{i,i}$ é a variância de $\hat{\theta}_i$. Denomina-se $J_{i,i}^{1/2}$, isto é, a raiz quadrada do i-ésimo termo da diagonal principal, o desvio-padrão de $\hat{\theta}_i$.
- Geralmente o desvio-padrão de $\hat{\theta}_i$ não é dado por

$$\left\{E\left[-\frac{\partial^2}{\partial\sigma_i^2}l(\theta)\right]\right\}^{-1/2} = [\mathbf{I}_E(\theta)_{i,i}]^{-1/2},$$

onde $\mathbf{I}_E(\theta)_{i,i}$ é o i-ésimo elemento da diagonal da matriz de informação esperada, ou seja, a raiz quadrada da matriz de informação esperada sob θ_i, pressumindo que os outros parâmetros são conhecidos. A diferença entre as duas expressões para o desvio-padrão, dado aqui e no Capítulo 3, aparece vindo da incerteza adicional em $\hat{\theta}_i$, vindo das estimativas de todos os outros elementos do vetor θ.

- Ortogonalidade: Quando

$$E\left[-\frac{\partial^2}{\partial\sigma_i\partial\sigma_j}l(\theta)\right] = 0, \text{ para todo } i \text{ e } j(i \neq j).$$

ou

$$-\frac{\partial^2}{\partial\sigma_i\partial\sigma_j}l(\hat{\theta}) = 0, \text{ para todo } i \text{ e } j(i \neq j).$$

é dito que os parâmetros são ortogonais . Nestes casos a matriz de informação esperada, ou observada, é uma matriz diagonal e assim o desvio-padrão de $\hat{\theta}_i$ é dado por $[\mathbf{I}_E(\theta)i,i]^{-1/2}$. O bom da ortogonalidadde é que a falta de conhecimento sobre θ_j não influência a precisão na estimação de $\hat{\theta}_i$.

- Pelos argumentos da Seção 1.3, tem-se que o estimador mais sensível de uma região de confiança é da forma

$$\{\theta \in \Omega : D(\theta) < c^*\}$$

para alguns valores de c^*. Pode-se escolher c^* baseado em justificativas assintóticas de que $D(\theta) \sim \chi_d^2$ é uma escolha razoavel para $c^8 = c_\alpha$, com $\Pr\{\chi_d^2 \geq c_\alpha\} = \alpha$, por exemplo se $\alpha = 0.05$ então

d	c_α	$l(\hat{\theta})$
1	3.84	1.92
2	5.99	3.00
3	7.81	3.90
4	9.49	4.75
5	11.07	5.53
6	12.59	6.25

Estas são as regiões de confiança de aproximadamente $100(1-\alpha)\%$. Assim regiões de confiança foram discutidas no Capítulo 3, correspondendo as curvas de nível apresentadas nas figuras.

- Regiões de confiança obtidos usando este resultado são as elípses obtidas no início deste capítulo.

5.4 Exemplos

Exemplo 5.7. $X_i \sim N(\mu, \sigma^2)$. Do Teorema 4 tem-se que $\hat{\theta} = (\hat{\mu}, \hat{\sigma})^T$ é assintoticamente

$$\hat{\theta} \sim NM_2\left(\theta, \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix}\right),$$

desde que

$$\mathbf{I}_E(\theta)^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix}.$$

Disto segue que o erro-padrão para $\hat{\mu}$ é $\sigma/n^{1/2}$ e que o desvio-padrão para $\hat{\sigma}$ é $\sigma/(2n)^{1/2}$. Note que eles são os mesmos para cada um dos parâmetros separadamente se o outro for considerado conhecido. Esta agradável característica de Não aumentar a incerteza se os parâmetros são conhecidos ou não é devido a ortogonalidade dos parâmetros.
Note também que $\hat{\mu}$ e $\hat{\sigma}$ são variáveis não-correlacionadas desde que os elementos fora da diagonal sejam zero, outra vez uma característica da ortongalidade.

Exemplo 5.8. $X_i \sim \Gamma(\alpha, \beta)$. Do Teorema 4 tem-se que $\hat{\sigma} = (\hat{\alpha}, \hat{\beta})^T$ é assintoticamente

$$\hat{\theta} \sim NM_2\left(\theta, det^{-1}\begin{pmatrix} \frac{n\sigma}{\beta^2} & \frac{n}{\beta} \\ \frac{n}{\beta} & n[\Gamma''(\alpha)/\Gamma(\alpha) - (\Gamma'(\alpha)/\Gamma(\alpha))^2] \end{pmatrix}\right),$$

onde

$$det = \left(\frac{n}{\beta}\right)^2 (\alpha[\Gamma''(\alpha)/\Gamma(\alpha) - (\Gamma'(\alpha)/\Gamma(\alpha))^2] - 1),$$

o determinante de $\mathbf{I}_E(\theta)$. Dele segue que o desvio-padrão de $\hat{\theta}$ é

$$\left(\frac{\alpha}{\alpha[\Gamma''(\alpha)/\Gamma(\alpha) - (\Gamma'(\alpha)/\Gamma(\alpha))^2] - 1}\right)^{1/2} n^{-1/2},$$

e o desvio-padrão de $\hat{\beta}$ é

$$\left(\frac{\beta[\Gamma''(\alpha)\Gamma(\alpha) - (\Gamma'(\alpha)/\Gamma(\alpha))^2]}{\alpha[\Gamma''(\alpha)/\Gamma(\alpha) - (\Gamma'(\alpha)/\Gamma(\alpha))^2] - 1}\right)^{1/2} n^{-1/2},$$

note que o desvio-padrão para cada parâmetro é diferente quando o outro parâmetro é conhecido, ou seja, se α é conhecido o desvio-padrão de $\hat{\beta}$ será

$$\left(\frac{\beta}{n\alpha}\right)^{1/2}.$$

Finalmente, note também que a correlação entre $\hat{\alpha}$ e $\hat{\beta}$ é positiva, a qual é consistente com as curvas de nível da Figura 4.3.

CAPÍTULO 6

PARÂMETRO DE INTERESSE

No Capítulo 2 cosiderou-se uma variedade de exemplos motivacionais, destes exemplos 1.2,1.3 e 1.4 são multiparâmetros. Por exemplo 1.2 os valores de todos os parâmetros desconhecidos foram de interesse, no exemplo 1.3 uma função $g(\theta)$ do parâmetro desconhecido foi de interesse, no exemplo 1.4 um componente do vetor paramâtrico foi o interesse.
No capítulo 4 tem muitos exemplos similares ao exemplo 1.2, mas não fora discutidos as versões particulares dos outros dois exemplos .
Ambos exemplos 1.3 e 1.4 são casos especiais de problemas de parâmetros de interesse , que são formulados como segue:
Suponha que observasões são realizadas de uma variável aleatória com distribuição de probabilidade com parâmetro desconhecido θ. Se somente um parâmetro $\phi = g(\theta)$, chamado de parâmetro de interesses , então restante dos parâmetros são denominados de parâmetros perfilhados, que podem ser ecrito como $\lambda = (\lambda_1, ..., \lambda_{d-1})$, ou seja, o parâmetro desconhecido $\theta = (\phi, \lambda)$. Não se obteve real interesse nas estimativas para λ ou da precisão destas estimativas entretanto desde que são desconhecidos, devem-se levar em considerações certas caracteristicas da estimação de λ ao estimar a variabilidade de $\hat{\phi}$.

Exemplo 6.1. $\phi = (\mu, \sigma)$. No primeiro caso

$$\phi = p = 1 - \phi(\frac{u-\mu}{\sigma}) = g(\theta)$$

logo $\lambda = \sigma$. No segundo caso

$$\phi = u = \mu + \sigma^{\phi^{-1}}(1 - p_0) = g(\theta), \text{ e } \lambda = \sigma$$

Exemplo 6.2. $\theta = (\alpha, \beta_1, ..., \beta_d, \sigma)$, agora

$$\phi = \beta_1 = g(\theta) \text{ e } \lambda = (\alpha, \beta_2, ..., \beta_d, \sigma).$$

6.1 Resultados

Nesta sessão foi examinada duas abordagens para obter um intervalo de confiança para o parâmetro de interesse $\phi = g(\theta)$, equanto este é um vetor de parâmetro desconhecido, θ, no modelo probabilistico. As duas abordagens são extensões multiparametrica dos metodos baseados nos teoremas 1 e 2 para funções de um parâmetro θ.
Primeiro considere estimando $\phi = g(\theta)$. A partir da estensões, para o caso multiparametrico, os argumentos de invariância dados na Seção 2.2, segue que o estimador máximo da verossimilhança $\hat{\phi}$ de ϕ e $g = (\hat{\theta})$
Agora considere obter a variância de $\hat{\phi}$. Vindo da extensão multiparamétrica do Cololário 1 do Teorema 1 tem-se que

RESULTADO 4-

$$Var(\hat{\phi}) = Var(g(\hat{\theta})) = \nabla g(\theta)^T I_E(\theta)^{-1} \nabla g(\theta)$$

onde

$$\nabla g(\theta) = \frac{\partial}{\partial \theta_1} g(\theta), ..., \frac{\partial}{\partial \theta_d} g(\theta)^T$$

Onde d=1 temos $\nabla g(\theta) = g^{'}(\theta)$ e $I_E(\theta)^{-1}$ é a escalar (matriz 1 x 1), assim

$$Var(\hat{\phi}) = [g^{'}(\theta)]^2 I_E(\theta)^{-1},$$

Dado o resultado do Teorema 1 e Corolário 1. A prova deste resultado é similar ao Lema 3.
O que a equação do Resultado 4 está a dar é a curvatura na direção $\bigtriangledown g(\theta)$ multiplicado por um fator associado com o tamanho do vetor.

RESULTADO 5- Para um problema de etimação regular se $\phi = g(\theta)$ são o verdadeiro valor, então quando $n \Rightarrow \infty$ tem-se que

$$\hat{\phi} \sim N(\phi, \nabla g(\theta)^T I_E(\theta)^{-1} \nabla g(\theta))$$

Vindo do Resultado 5, intervalo de confiança para $\hat{\phi}$ podem ser construidos, no caso uni-paramétricos isto é o $100(1-\alpha)\%$ o intervalo de confiança é:

$$(\hat{\phi} - Z_{\frac{\alpha}{2}} se(\hat{\phi}), \hat{\phi} + Z_{\frac{\alpha}{2}} se(\hat{\phi})),$$

onde se $(\hat{\phi}) = [Var(\hat{\phi})]^{1/2}$.
Finalmente, para comparar com a abordagem alternativa são baseada no seguinte resultado:

RESULTADO 6 -Para um problema de estimação regular, se $\phi = g(\theta)$ são o valor verdadeiro então com $n \Rightarrow \infty$ tem-se

$$D^*(\phi) = 2[l(\hat{\phi} - Pl(\phi)] \sim x_1^2.$$

Onde $Pl(\phi)$ determina a log-verossimilhança perfilhada avaliada em ϕ, é
$Pl(\phi) = max\, l(\phi, \lambda)$, isto é o máximo no conjunto de valores possíveis de λ sujeito a ϕ sendo fixado

Exemplo 6.3. i.e. $X_i \sim N(\mu, \sigma^2)$. Derive a log-verossimilhança perfilhada para μ
Aqui $\lambda = \sigma$ e $\phi = \mu$. Assim

$$l(\mu, \sigma) = -\frac{n}{2} log(2\pi) - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Maximizando isto com respeito a σ para μ fixo obtem-se que o máximo ocorre em $(\hat{\phi})$ quando

$$l_\sigma(\mu, \hat{\sigma}_\mu) = 0$$

isto é

$$0 = -\frac{n}{(\hat{\sigma})_\mu} + \frac{1}{(\hat{\sigma})_\mu^3} \sum_{i=1}^{n} (x_i - \mu)^2$$

então

$$\hat{\sigma}_\mu = (\frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2)^{1/2}$$

deste modo

$$D^*(\mu) = -2nlog(\hat{\sigma}/\hat{\sigma}_\mu) + n\frac{1}{(\hat{\sigma})_\mu^2} \sum_{i=1}^{n} (x_i - \mu)^2 - n\frac{1}{(\hat{\sigma})^2} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

então

$$D^*(\mu) = -2nlog(\hat{\sigma}/\hat{\sigma}_\mu)$$

Isto exige muito trabalho analítico, então geralmente não se considera exemplos desta forma. Desenhar isto é muito fácil, particularmente se ϕ é um elemento do vetor θ. Então $Pl(\phi)$ é o perfil do gráfico do eixo de ϕ.
Note que o valor máximo de $Pl(\phi)$ ocorre quando $\phi = \hat{\phi}$, assim $D^\alpha(\phi)$ é sempre positiva , e assim intervalos de confiança de 95% são dados pelo valor de 1.92 em $Pl(\phi)$ vindo do máximo.
Veja figura 8 e 10 onde tem-se evidência do perfil da verossimilhança

6.2 Exemplos

Agora ilustra-se a discussão teórica com aplicações na $N(\mu,\sigma^2)$ e problemas da Gamma(α,β)

Exemplo 6.4. i.e. $X_i \sim N(\mu,\sigma)$ e temos $I_E(\theta)^{-1} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix}$

1. Se $\phi = g(\theta) = \mu$ quando $\nabla g(\theta) = (1,0)$ então $Var(\hat{\phi}) = (1,0)\begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix}\begin{pmatrix} 1 \\ 0 \end{pmatrix} = \frac{\sigma^2}{n}$
2. Se $\phi = g(\theta) = \sigma^2$ quando $\nabla g(\theta) = (0,2\sigma)$então $Var(\hat{\phi}) = (0,2\sigma)\begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix}\begin{pmatrix} 0 \\ 2\sigma \end{pmatrix} = \frac{2\sigma^4}{n}$.
3. Se $\phi = g(\theta) = \mu + \sigma^{\phi^{-1}}(1-p_0)$ quando $\nabla g(\theta) = (1,\phi)^{-1}(1-p_0)$ então $Var(\hat{\phi}) = (1,\phi)^{-1}(1-p_0)\begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix} = \begin{pmatrix} 1 \\ \phi^{-1}(1-p_0) \end{pmatrix} = \frac{\sigma^2}{2n}(2+[\phi^{-1}(1-p_0)]^2)$.

Exemplo 6.5. i.e. $X_i \sim \Gamma(\alpha,\beta)$. Nota que $Var(\hat{\phi})$ se $\phi = E(X) = \alpha/\beta$.

CAPÍTULO 7

APÊNDICE

7.1 Três Testes Estatísticos Relacionados a Verossimilhança

A atribuição de hipoteses a seguir é essêncialmente de interesse teórico,

$$\begin{cases} H_0 : \theta_* = \theta_0 \\ H_1 : \theta_* = \theta_1 \end{cases}$$

Em situações práticas estamos mais sujeitos a deparar com casos mais complexos; por exemplo um caso que frequêntemente ocorre é

$$\begin{cases} H_0 : \theta_* = \theta_0 \\ H_1 : \theta_* \neq \theta_0 \end{cases}$$

Para este problema de teste, uma razoável tratativa, é o uso da razão de verossimilhança

$$\begin{aligned}\lambda(y) = \lambda &= \frac{L(\theta)}{sup_{\theta \neq \theta_0} L(\theta; y)} \\ &= \frac{L(\theta)}{sup_{\theta \in \Theta} L(\theta; y)} \\ &= \frac{L(\theta_0)}{L(\hat{\theta})}\end{aligned}$$

Onde assumimos a continuidade de $L(\theta)$ em respeito a θ para todo $y \in Y$. Observe que $\lambda(y) = (\tilde{\theta_0})$ é a verossimilhança relativa.
Uma transformação monótona do teste estatístico junto com as transformações correspondentes dos valores críticos não mudam a parte do espaço amostral na região de rejeição ou de aceitação, uma vez que não muda o procedimento de teste. Assim, teste estístico equivalente é

$$W(y) = 2ln\ \lambda(y) \tag{7.1}$$

Que é chamado, com um pequeno abuso da terminologia, de teste de razão de verossimilhança.
Obviamente a interpretação de valores extremos de $W(y)$ serão invertidos em relação a $\lambda(y)$, uma vez que a transformação é decrescente. Uma vez que $P\{0 < \lambda(Y) < 1\} = 1$ na maioria dos casos, então efetivamente $W(y) \in (0, \infty)$.
Nós agora gostariamos de derivar alguns outros testes estatísticos associados a função de verossimilhança e aproximadamente relacionados com $W(y)$. Fazendo a expanção por Série de Taylor de θ, obtemos

$$\begin{aligned}W(y) &= -2l(\theta) - l(\hat{\theta}) \\ &\cong -2\{(\theta_0\ \hat{\theta})l'(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^2 l''(\tilde{\theta})\}\end{aligned}$$

Onde $\tilde{\theta} \in (\tilde{\theta}, \theta_0)$ e $l'(\hat{\theta} = 0)$. Uma vez que $\hat{\theta}$ é um estimador consistente de θ_0 sob H_0, então $\tilde{\theta}$ está conduzindo a expanção.

$$\begin{aligned}W(y) &= -n(\hat{\theta} - \theta_0)^2)U'(\theta_0) + Op(1) \\ &= n(\hat{\theta} - \theta_0)^2 i(\theta_0) + Op(1)\ (4.12) \\ &= n(\hat{\theta} - \theta_0)^2 i(\theta_0) + Op(1)\end{aligned}$$

Se $i(\theta)$ for uma função contínua, então $i(\theta_0) = i(\hat{\theta}) + Op(1)$ sob H_O.
Então obtemos a aproximação

$$W(y) = W_E(y) + Op(1)$$

onde

$$W_E(y) = n(\hat{\theta} - \theta_0)^2 i\hat{\theta}$$

Que é chamado de estatística de teste de Wald. De onde obtemos sob H_0
$\sqrt{n}(\hat{\theta} - \theta_0) = U(\theta_0)\sqrt{n}i(\theta_0) + Op(1)$
Que substituindo em (4.12) nos dará outra aproximação para $W(y)$,

$$W(y) = W_u(y) + Op(1)$$

onde

$$W_u(y) = \frac{U(\theta_0)^2}{n_i(\theta_0)}$$

A qual chamamos de estatística de teste score.
Agora são conhecidos três funções de teste, diferente da outra pelos termos $Op(1)$. Estas funções de testes quantificavam três aspectos das funções de log-verossimilhança como ilustrado na figura 4.2

- $W(y)$ mensura a diferença no eixo ordenado entre a log-verossimilhança calculada em $\hat{\theta}$ *e em* θ_0.

- $W_E(y)$ mensura o desvio na abcissa de $\hat{\theta}$ *e* θ_0, adequadamente normalizados.

- $W_u(y)$ mensura a inclinação da log-verossimilhança em θ_0, novamente adequadamente e normalizado.

Assim podemos resumir estes três testes da seguinte forma:

Razão de verossimilhança $\Rightarrow W(y) = -2\{l(\theta_0) - l(\hat{\theta})\}$

Teste de Wald $\Rightarrow W_E(y) = \dfrac{(\hat{\theta} - \theta_0)^2}{i^{-1}(\theta_0)/n}$

Teste Score $\Rightarrow W_u(y) = \dfrac{U(\theta_0)^2}{n_i(\theta_0)}$

onde

$$(\hat{\theta})I(\theta_0)(\hat{theta} - theta_0) \sim \chi^2_p \text{ e } U'(\theta_0)I^{-1}(\theta)U(\theta_0) \sim \chi^2_p$$

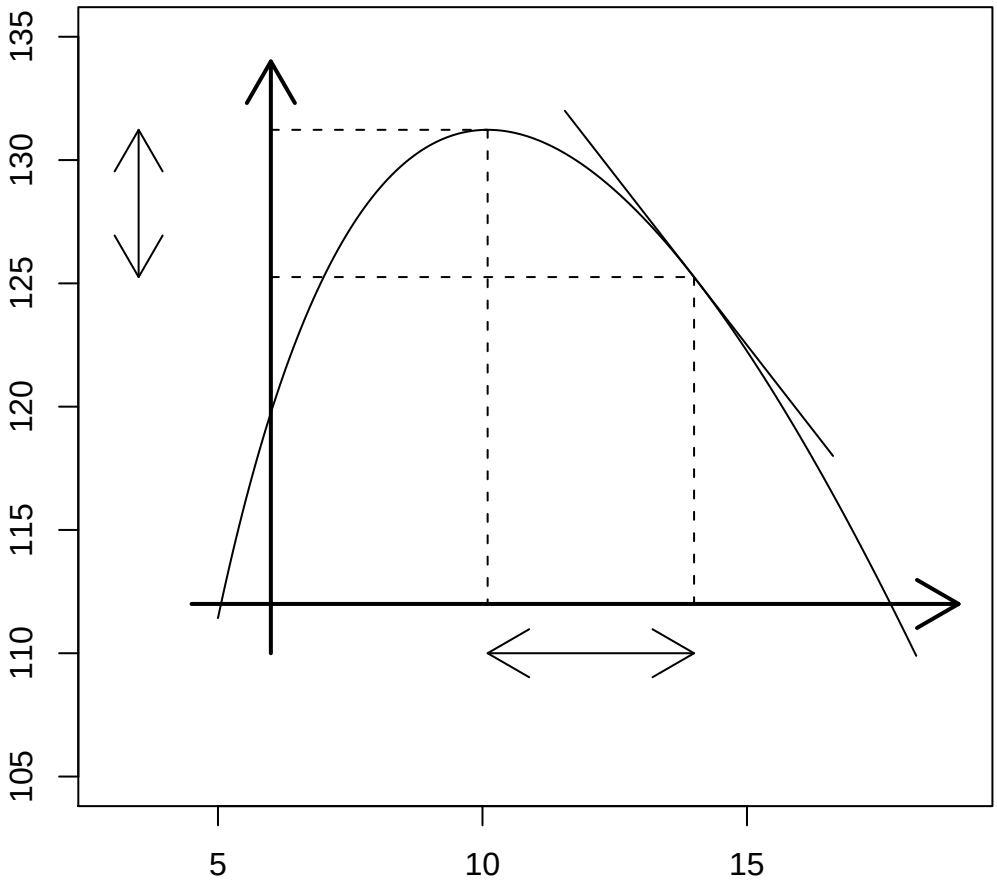


Figura 7.1: Três funções de teste conectados pela verossimilhança.