



UNIVERSIDADE FEDERAL
DO RIO DE JANEIRO



Astroestatística

Aula 005 -- Prof. Walter Martins-Filho

Frequentista ou
Bayesiana

Inferências Estatísticas

- Existem dois paradigmas em estatística: frequentista e a bayesiana
- A Inferência Bayesiana, por mais antiga comparada a outra, passou a ser utilizada nas ciências físicas de forma tardia
- Questão filosófica como os dois paradigmas são desenvolvidos.

Inferência Clássica



Paradigma Frequêntista

- As probabilidades são relativas a frequência de eventos. Os quais são propriedades INTRÍNSECAS do mundo
- Parâmetros são constantes fixas e imutáveis
- Procedimento de análise deve averiguar longa cadeia de eventos afim de determinar os parâmetros e as propriedades das distribuições associadas.

Estimativas

- Avaliando nosso cenário:
 - “parâmetros não possuem flutuações, fixados em nosso universo”
 - Podemos associar probabilidades no intervalo de confiança
 - Probabilidade é sempre associada a frequência de uma variável aleatória

Inferência

“Point
Estimation”

“Confidence
Estimation”

“Hypothesis
Testing”

Inferência

“Point
Estimation”

Qual é o melhor valor para parâmetro θ do modelo que estamos avaliando para dados amostrados?

Inferência

“Confidence
Estimation”

Quão preciso é valor
encontrado que melhor
representa o parâmetro θ ?

Inferência

“Hypothesis
Testing”

A amostra que possuímos é
consistente com as hipóteses
e modelo assumidos?

Estimação da Máxima Verossimilhança

- Em inglês, “Maximum Likelihood Estimation”, **MLE**
- O paradigma frequentista NÃO é dependente do MLE, porém vários casos são avaliados por esta análise.
- Ideia: quantificar como os dados são medidos, em outras palavras, QUANTIFICAR O PRÓPRIO PROCESSO DE AQUISIÇÃO

FUNÇÃO DE VEROSSIMILHANÇA

- **Função de verossimilhança** representa o conjunto quantitativo que **descreve o processo de aquisição dos dados**
- Ideia introduzida por Gauss e Laplace, e popularizada por Fisher
- Conhecendo-se a distribuição o qual os dados foram obtidos, podemos calcular a probabilidade (ou verossimilhança) de qualquer variável observada.

Exemplo

- Se a amostra $\{x_i\}$ foi obtida por uma distribuição normal $N(\mu, \sigma)$, então a sua função de verossimilhança será dada como:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Observação

- Mas e o ajuste por resíduos?
 - O Ajuste de resíduos segue uma outra função de verossimilhança, que também representa o que esperamos que tenham saído os nossos resíduos do ajuste. Em geral, “joga-se” esta nomenclatura antes, caso não se tenha uma visão mais perfeita de inferência.

Função de Verossimilhança

- Assumindo que os dados de $\{x_i\}$ foram coletados de forma independente, o conjunto L é o produto da função de verossimilhança de cada valor em particular:

$$L \equiv p(\{x_i\}|M(\theta)) = \prod_{i=1}^n p(x_i|M(\theta))$$

M, θ : modelo, parâmetros

Função de Verossimilhança

- $L \equiv p(\{x_i\}|M(\boldsymbol{\theta}))$ pode ser lido como a probabilidade dos dados $\{x_i\}$ dado o modelo $M(\boldsymbol{\theta})$, baseado no vetor de parâmetros $\boldsymbol{\theta} = (\theta_1 \ \theta_2 \cdots \theta_p)$
- Note que L não é distribuição normalizada, enquanto que cada PDF de x_i é uma função normalizada.
- Devido a não ser normalizada, teremos então valores extremamente pequenos ou grandes para L , o que causa ser melhor trabalhar com $\log L$
- L pode ser considerada também como a função de Modelos dos parâmetros.

Máxima Verossimilhança

- Assim, o valor máximo (ou mínimo) da função de Verossimilhança, pelas ideias de Gauss, Laplace e Fisher, é encontrado quando obtemos os valores corretos para parâmetros da distribuição assumida:

$$\frac{d}{d\theta} \ln L \Big|_{\theta=\theta^0} \equiv 0$$

Técnica de Maximização

1. Aquisição dos dados para modelo M , suposição de uma $p(D|M(\boldsymbol{\theta}))$
2. Procura-se os melhores valores para $\boldsymbol{\theta}$ que maximizem $p(D|M)$, obtendo $\boldsymbol{\theta}^0$
3. Estima-se a região de confiança para os parâmetros $\theta_p^0, p = 0, \dots, k$
 1. Por derivação matemática
 2. Por análise numérica (bootstrap, jackknife e/ou cross-validation)
4. Faça um teste de hipóteses, afim de averiguar outros modelos e estimadores pontuais.

Caso de estudo: Distribuição Normal

- Vamos assumir que temos uma amostra de fluxo $\{x_i\}$ com N valores
- Todos eles possuem erro s_i que são o mesmo $s_i = \sigma | \forall s_i$, caso homocedástico
- Objetivo: encontrar a Máxima Verossimilhança e o intervalo de confiança

Caso de estudo

- Cada fluxo obtido $\{x_i\}$ segue distribuição Normal
- ASSUMINDO que é mesmo objeto, e que não temos variabilidade, a média $\mu_i = \mu$, deve ser a mesma.

$$L \equiv p(\{x_i\} | M(\theta)) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Caso de estudo

- Assim, a ideia é que maximizando essa função, iremos encontrar os valores corretos para parâmetro μ :

$$\frac{d}{d\mu} \ln L \Bigg|_{\mu=\mu^0} \equiv 0$$

Obtendo o $\ln L$

$$\ln L = \text{cte} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2}$$

O que implica em:

$$\mu_0 = \frac{1}{N} \sum_{i=1}^N x_i$$

Exercício

- Prove as contas do caso de estudo da MLE para distribuição normal, derivando $\ln L$
- Liste quais foram as suposições iniciais que possibilitaram esta derivação e determine as falhas dela.

Suposições MLE

1. Dados vieram da mesma distribuição original
2. Distribuições são suaves
3. A derivada para parâmetro a ser encontrado deve existir.

Propriedades MLE

- Consistência – pode ser provado que conforme $N \rightarrow \infty$, o parâmetro θ converge
- Assintoticamente Normal – conforme $N \rightarrow \infty$, a distribuição do parâmetro θ tende a uma distribuição Normal, centralizada no parâmetro determinado pela MLE
- Mínima Variância – conforme $N \rightarrow \infty$, a distribuição θ alcança o variância mínima possível teoricamente, chamado de Limite de Cramer-Rao

Intervalos de Confiância

- O parâmetro θ é fixo no Universo, mas como tratamos de amostra, não podemos associá-lo a um valor sem erro
- Por isto, todo parâmetro encontrado deve ter uma incerteza associada
- Dado o MLE μ_0 do exemplo anterior, como podemos estimar a incerteza associada?

Intervalo de confiança

- Expandindo a função $\ln L$ por série de Taylor, podemos determinar a variância associada a cada parâmetro θ_k :

$$\sigma_{jk} = \left(- \frac{d^2 \ln L}{d\theta_j d\theta_k} \Big|_{\theta=\theta_0} \right)^{-1/2}$$

Problemas

- Em geral, para distribuições normais, a função $\ln L$ gera um elipsoide suave para as incertezas σ_{jk}
- Contudo, para outras distribuições podemos ter:
 - Não ser suave no espaço σ_{jk}
 - Ser multimodal, o que indetermina qual é valor exato de σ_{jk}
- NO caso de não termos certeza, é sempre boa ideia fazer gráfico da superfície de $\ln L$

Barra de erros marginais

- Os elementos σ_{ii} fornecem barras de erro marginais para parâmetro θ_i
- Se $\sigma_{jk} = 0 | \forall j \neq k$, então os valores inferidos para os parâmetros θ não possuem correlação, e tem a mesma condição de independência
- SE $\sigma_{jk} \neq 0 | \forall j \neq k$, então os erros dos parâmetros são correlacionados. É importante notar que aqui estamos tratando de correlações NA INCERTEZA!

Caso de estudo

- Voltando ao caso de $L = N(\mu, \sigma^2 | \theta)$, temos que a incerteza associada é dada por:

$$\sigma_\mu = \left(-\frac{d^2 \ln L}{d\mu^2} \Bigg|_{\mu=\mu_0} \right)^{-1/2} = \frac{\sigma}{\sqrt{N}}$$

Caso Heterocedástico

- No caso de estudo anterior, assumimos que os valores de incerteza dos dados coletados, ou seja, σ_k , são todos os mesmos, i.e., iguais.
- No caso deles serem diferentes, estamos tratando de um problema heterocedástico

Caso de estudo

- Vamos assumir agora que erros associados aos dados são diferentes
- Com isto, não podemos eliminar σ_i da derivada da função $\ln L$

$$\ln L = \text{cte} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma_i^2}$$

$$\mu_0 = \frac{1}{N} \frac{\sum_{i=1}^N w_i x_i}{\sum_{i=1}^N w_i} \quad \text{onde} \\ w_i = \sigma_i^{-2}$$

Resultado é que μ_0 é soma aritmética com pesos associados. No caso considerado. Com erro associado de:

$$\sigma_\mu = \left(\sum_{i=1}^N \frac{1}{\sigma_i^2} \right)^{-1/2} = \left(\sum_{I=1}^N w_i \right)$$

Pergunta

Baseado em
estimar modelo
aos dados, qual
maior problema
do método MLE?

Cost Function

- A função de Verossimilhança é denominada “Função de Custo”
- A ideia é que “Funções de Custo” quantizam algum custo associado ao estimação do parâmetro θ
- O valor esperado da Função Custo é denominado risco (em inglês, risk) e pode ser minimizado afim de obter os melhores ajustes (*best-fit parameters*)

Mean Integrated Square Error

- Uma função de custo usada é a Média Integrada do Erro ao Quadrado (MISE)
- Esta função de custo quantiza quão “próxima” a nossa pdf empírica $f(x)$ está da verdadeira pdf populacional $h(x)$:

$$\text{MISE} = \int_{-\infty}^{\infty} [f(x) - h(x)]^2 dx$$

Ajuste de Modelo

- Uma vez que obtemos os valores que maximizam a função de verossimilhança, temos pergunta a responder:
 - **Se θ_0 implica em $\max(L) \equiv L_0$, quão crível é usar os valores de best-fit para obter L_0 ?**
 - Veja bem, não estamos considerando L para obter θ_0 , mas sim o inverso. Se conhecemos o best-fit θ_0 com sua incerteza associada, qual é a probabilidade de obter L_0 .
 - Esse questionamento possibilita analisar diferentes modelos e respostas de θ_0 para mesmo modelo e identificar aquele que seja melhor.

Ajuste de Modelo

Vamos revisar o
caso de $\ln L$ para
distribuição Normal:

$$\ln L = \text{cte} - \sum_{i=1}^N \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$z_i^2 = \frac{(x_i - \mu)^2}{\sigma^2}$$

$$\ln L = \text{cte} - \frac{1}{2} \sum_{i=1}^N z_i^2$$

$$\ln L = \text{cte} - \frac{1}{2} \chi^2$$

Ajuste de Modelo

- Assim, temos que dado conjunto de parâmetros θ , a função de verossimilhança $\ln L$, baseada numa distribuição Normal para dados adquiridos $\{x_i\}$, seguirá uma distribuição de χ^2
- A distribuição de χ^2 terá $N - k$ graus de liberdade, aonde k é a quantidade de parâmetros a serem ajustados (dimensão de θ)
- Para caso da distribuição Normal, queremos ajustar μ , logo $k = 1$, o que implica em $N - 1$ graus de liberdade. (e essa é origem do termo de correção anteriormente comentado)

Referências

- Livros-textos
- <http://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/06/lecture-06.pdf>