APRENDIZADO DE MÁQUINAS PARA FÍSICA

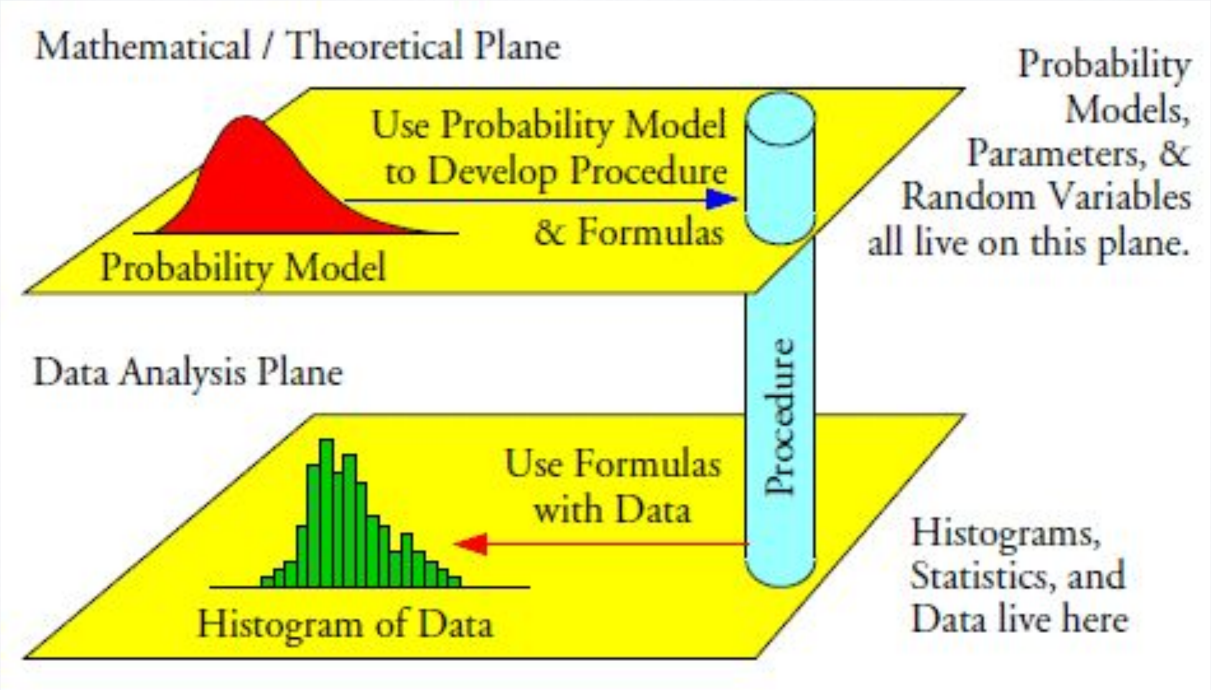
Aula III

Estimativa de ponto e intervalos de confiança

DESCRIÇÃO ESTATÍSTICA DOS DADOS

A medida de uma grandeza parte de uma *amostragem* da distribuição de probabilidade (PDF) base:

- Podemos estimar um parâmetro a partir de uma amostra de dados (POINT ESTIMATION)
- Podemos estimar um intervalo de confiança a partir de uma amostra de dados (INTERVAL ESTIMATION)



ESTIMADOR DE PONTO

- Estimamos o valor de um parâmetro desconhecido θ usando uma amostra dos dados $\{x_i\}$.
- Estimadores $\hat{\theta}$ são funções dos dados, com isto também são variáveis aleatórias que seguem sua própria PDF. Ao repetir as medições várias vezes as estimativas deveriam seguir a função de distribuição $g(\hat{\theta}|\theta)$

Queremos ter uma variância (erro estatístico) pequena

$$V = E\left[(\hat{\theta} - \theta)^2\right]$$

Queremos ter um viés (erro sistemático) pequeno

$$b = E[\hat{\theta}] - \theta$$

MÉDIA E VARIÂNCIA

A média μ = E[x] para a PDF base (desconhecida).

- Um estimador para o valor de μ é a média das medidas de uma amostra

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \bar{x}$$

- Se usarmos várias amostras, e calcularmos a média de cada uma, elas seguem uma distribuição. A variância do estimador médio entre as amostras é

$$V_{\hat{\mu}} = \frac{V_x}{n} \Rightarrow \sigma_{\hat{\mu}} = \frac{\sigma_x}{\sqrt{n}}$$

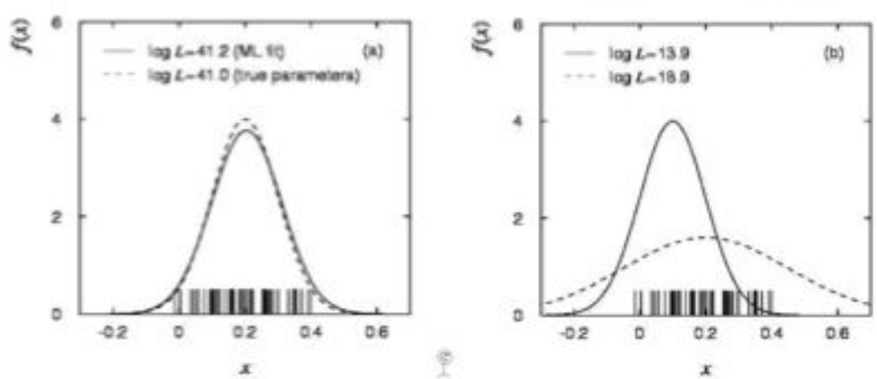
Desvio padrão da PDF base

MÁXIMA VEROSSIMILHANÇA

- Um experimento está modelado por uma PDF $f(x|\theta)$, e as medidas são $\{x_i\}$. O estimador de máxima verossimilhança do parâmetro θ está definido como o valor que faz a verossimilhança $L(\theta)$ avaliada na amostra ser máxima

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

If the hypothesized θ is close to the true value, there's a high probability to get data like we actually observe !



The value of $\hat{\theta}$ that maximizes the likelihood is not the most likely value of θ. It's the value of θ that makes your data most likely

Trabalhando com MV

- Em vez de tentar maximizar a função $L(\theta)$, vamos minimizar $-\ln(L(\theta))$ (equivalente e mais fácil).

$$-\ln(L(\theta)) = -\sum_{i=1}^{n} \ln(f(x_i|\theta))$$

- Pode ser feito numericamente.
- Para amostras grandes (prop. asintóticas) o estimador $\hat{\theta}$:
  - Não apresenta viés
  - Segue uma distribuição gaussiana
  - A variância pode ser estimada como $\hat{V}(\hat{\theta}) = -\left(\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right)^{-1}\Bigg|_{\theta=\hat{\theta}}$

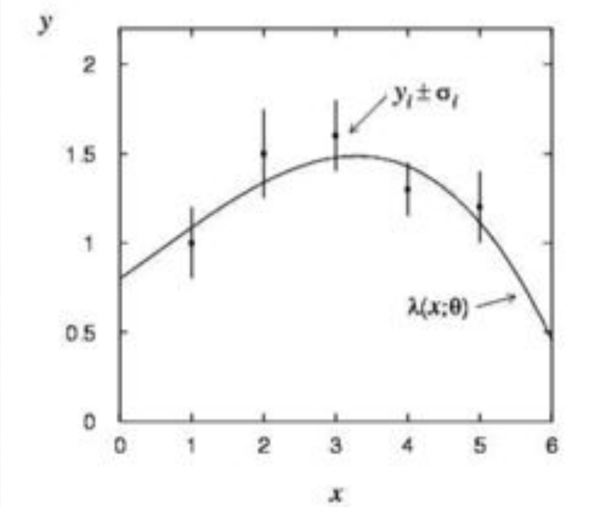
Mínimos Quadrados

Se há n medidas $\{y_i\}$ de um valor $\lambda(x_i,\theta)$, que seguem distribuições gaussianas com variâncias σ_i^2

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - \lambda(x_i,\theta))^2}{2\sigma_i^2}\right] \Rightarrow -\ln L(\theta) = \sum_{i=1}^{n} \frac{(y_i - \lambda(x_i,\theta))^2}{2\sigma_i^2} + C$$

Nesse caso, o estimador dos mínimos quadrados, χ^2 , é obtido a partir da maximização de $L(\theta)$

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \lambda(x_i,\theta))^2}{\sigma_i^2}$$



Estimando incertezas

A incerteza pelo erro estatístico na estimativa do nosso parâmetro pode ser avaliada

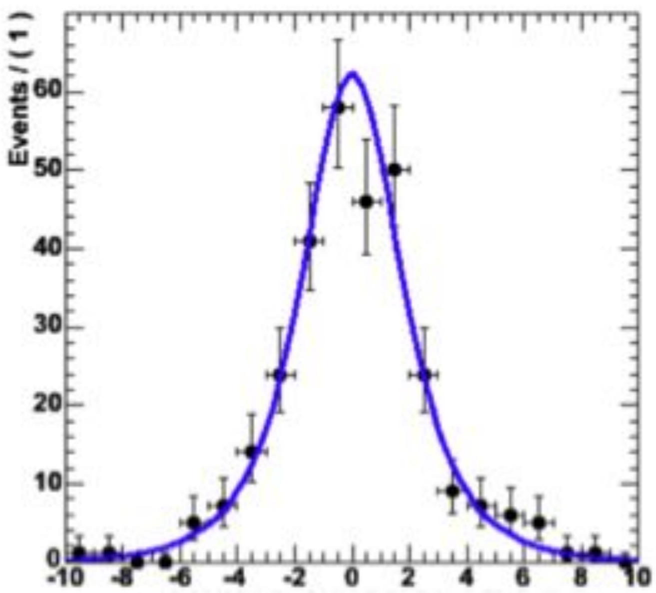
- Através de *simulações repetidas* (pseudo-experimentos) onde obtemos uma distribuição das estimativas
- De forma *analítica* se conhecemos a PDF $V = \mathrm{E}[\theta^2] - (\mathrm{E}[\theta])^2$
- De forma *gráfica* aproximando o contorno mínimo da variância

Monte Carlo

We can use simplified simulated experiments ("toy Monte Carlo") to understand the distribution of the ML estimators.

Variance of Estimators - MC Simulation

1. Choose a plausible true value of the parameter θ
2. Generate several sets of simulated data $\{x_i\}$ (experiments) by random sampling the model PDF $f(x, \theta)$
3. Maximize the likelihood in each set to estimate the parameter θ
4. Look at the distribution of the estimator $\hat{\theta}$
5. Repeat for all relevant choices of the parameter θ values



Estimativa gráfica

Expanding the likelihood in a Taylor series around its maximum $\hat{\theta}$ we have

$$log\ L(\theta) = log\ L(\hat{\theta}) + \left[\frac{\partial log\ L}{\partial \theta}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2!}\left[\frac{\partial^2 log\ L}{\partial \theta^2}\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + \ldots$$

First term is L_{max} , the second is zero and for the third we use information inequality

$$log\ L(\theta) = log\ L_{max} - \frac{(\theta - \hat{\theta})^2}{2\sigma^2_{\hat{\theta}}}$$

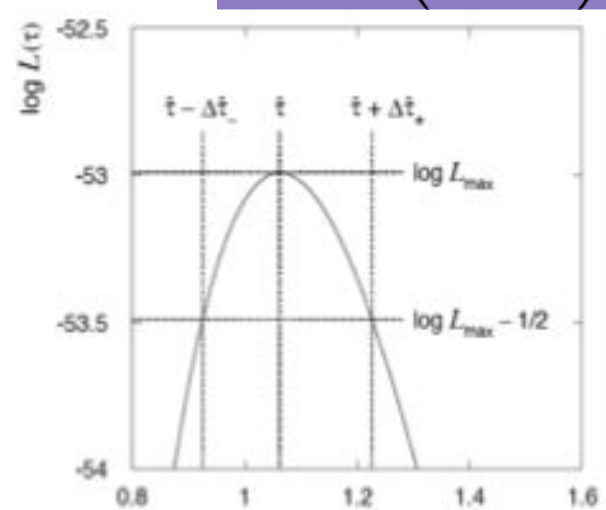
Limite de Cramér-Rao

$$V \gtrsim -\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1}$$

Variance of Estimators - Graphical

So , displacing θ by one standard deviation , $\theta \rightarrow \hat{\theta} \pm \sigma_{\hat{\theta}}$ we have that *log L* decreases by 1/2 from its maximum *log* L_{max}

$$log\ L(\hat{\theta} \pm \sigma_{\hat{\theta}}) = log\ L_{max} - \frac{1}{2}$$



Intervalo de confiança

- A definição da incerteza em base à variância ou desvio padrão é válida para gaussianas, ou distribuições simétricas.
- Para outras distribuições usamos intervalos de confiança, o que pode levar a erros *asimétricos*
- Um intervalo para μ com x% de confiança pode ser definido tal que *o verdadeiro valor de* μ está contido no intervalo uma fração x% das vezes (ou seja, temos x% de chance do μ estar no intervalo)
- Pode se escolher intervalo com 2 contornos (medição) ou só um (limite máximo ou mínimo para a grandeza)

Exercícios e notebooks

Amostragem por inversão, Aceitação-Rejeição e Aceitação eficiente

https://github.com/clemencia/ML4PPGF_UERJ/blob/master/Amostragem_e_integracao_MC.ipynb

Poisson e Binomial: ex. 3.8.4 livro Oguri:
Um teste tem 15 perguntas com 4 alternativas cada uma e apenas uma é correta, assim a probabilidade a priori de acerto ao acaso de uma questão é

$$p = 1/4$$

a. Determine a distribuição de probabilidades de acerto ao acaso de m das 15 questões (m=0,1,2,…m)
b. Represente em um histograma
c. Se 1000 alunos fizerem o teste respondendo as questões ao acaso, quantos, em média, acertarão pelo menos 3 questões?