呂浩宇

# Introduction to Data Analysis

2023. Mar. 03

## I. Purpose of modeling data:

1. Simplifying the description (e.g. describing as Gaussian distributions with certain mean values and standard deviations)
2. Constraining physical parameters and understanding the underlying physical principles.
   (e.g. comparing the power spectrum of CMB to inflation models)

## II. Good (necessary) practice when dealing with data:

1. knowing what we are doing.
2. Being able to tell people why we are doing so. (Not good to say "老師叫我們這樣做")
   Ideal to loose everybody at hello .....
3. Reproducibility
4. Portability (other people should easily understand our data product, and can utilize them. When you define quantities in your own way, or when you invent a new data format, make them transparent and acceptable)

## III. Concept: from Bayesian to frequentist (類神經網路及AI為其它沒時之經驗推斷)

↳ 由基本的機率公理出發

一般想知道並描述完整的 $p(\text{model}|\text{data})$ 分布,
如藉由 Markov Chain Monte Carlo (MCMC) method.

Bayes inference 處理的問題
為處理(已測得)數據時之最完整命題

當數據為某測量值時, model 為某特定 model 之機率

特定 model 之機率分布 (prior model distribution)

$$P(\text{model}|\text{data})\,P(\text{data}) = P(\text{data}|\text{model})\,P(\text{model})$$

出現某(組)數據的機率

某 model 為真時, data 為某(組)特定測定之機率

※ 難以真正在實作中做設對問題一無所知, 假設要合理
藉由對問題本身的初步知識選取或建構之模型

Prior predictive probability
某些情形下可視為用以維持機率適合為 1.0 的 normalization constant.

frequentist inference 處理的問題, 常用.
為在無法得知 prior, $P(\text{model})$ 之情形下採用之方法

一般情形: 選擇 model 之參數, 使得此機率為極大

- least square,
- $\chi^2$
- Maximum likelihood analysis.

Example: 已知男女在大二實驗課中得分之機率分布分別為如何。
當10個學生中100分有3人, 50分有5人, 0分有2人, 問此10人最有可能之性別分佈為何?

(可利用物理系女生佔20%之已知情報做為 prior probability distribution)

↳ "flat" prior 假設男女比1:1 視得合理

呂浩宇

# Introduction to Data Analysis 2023. Mar. 03

## IV. Goal of frequentist inference

1. describe data with certain model parameters.
2. knowing the uncertainty of parameters.
3. assessing goodness-of-fitting. (若 goodness-of-fitting 不好，則前兩步得知的信息不具任何參考價值)

## A. Least square analysis (frequentist)

(i) model：測量 $y$ 為變數 $x$ 的函數，並且此函數有 $M$ 個可變的參數 $a_j$, $j=0,\cdots M-1$

$$y(x) = y(x\,|\,a_0,\cdots,a_{M-1})$$

(ii) 如有 $N$ 組 $(y_i, x_i)$ 的測量，$(i=0,\cdots N-1)$，且僅 $y$ 有測量誤差，並且測量誤差可以用標準差為 $\sigma$ 之 normal distribution 描述

則

$$\underbrace{P(\text{data}\,|\,\text{model})}_{\text{likelihood}} \propto \prod_{i=0}^{N-1}\left\{\exp\left[-\frac{1}{2}\left(\frac{y_i - y(x_i)}{\sigma}\right)^2\right]\Delta y\right\}$$

($y_i$: measurement; $y(x_i)$: model; $\sigma$: measurement error.)

frequentist 推論就工作為調整參數使得 $P(\text{data}\,|\,\text{model})$ 為極大，亦即極小化

$$\left[\sum_{i=0}^{N-1}\frac{[y_i - y(x_i)]^2}{2\sigma^2}\right]$$

Bayesian 推論則討論 $P(\text{model}\,|\,\text{data}) \propto P(\text{data}\,|\,\text{model})\,P(\text{model})$ 之完整分布

→ $\chi^2$ (chi-square) fitting 假設每個測量可能有不同之誤差 $\sigma_i$，而極小化

$$\chi^2 \equiv \sum_{i=0}^{N-1}\left(\frac{y_i - y(x_i\,|\,a_0\cdots a_{M-1})}{\sigma_i}\right)^2$$

也就是找 $\frac{\partial}{\partial a_i}\chi^2$ 為 0 之位置

可藉由解析推導（線性問題）或 Monte-Carlo simulation 推估 $\chi^2$ 大於某特定值之機率，依此可定義 goodness-of-fitting. 例如，若 $\chi^2$ 大於 1000 的機率為 $10^{-18}$，而我們的 best-fit 給出之極小的 $\chi^2$ 為 2000，則可斷定 model 不恰當，在此情形推導出之參數及其 uncertainties 皆不具意義

※ 機率太高也有問題，一般代表 $\sigma$ 被高估，為常見之數據擬合作弊行為

best-fit 合理的 $\chi^2$ 值應接近 $N-M$

(i.e., 每個獨立的測量貢獻 ~1.0 到 $\chi^2$)