

## Question

What is Statistics?

統計 ← Data Science

方法	哈利波特	Real Life	
資訊	占卜學	Statistics	
載體	崔老妮	Statisticians	
產品	水晶球	Data	數據資料
	未來的資訊	Information	

aim of statistics: provide insight by means of data

into → see, watch

對現象、狀況、系統、...

## Basic Procedures of Statistics

- Statistics divides the study of data into four steps:

Data:  $X_1, \dots, X_n$  (random variables)

Data Collection

Statistical Modeling

➤ Q: What is a statistical model? 模型

➤  $X_1, \dots, X_n \sim$  joint cdf  $F_X$ /pdf  $f_X$ /pmf  $p_X$  with unknown parameters  $\Theta$  ← 規律

Transformations

 $g_1(X_1, \dots, X_n),$ 

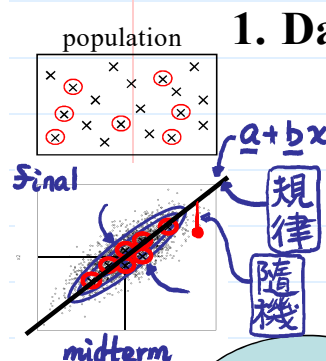
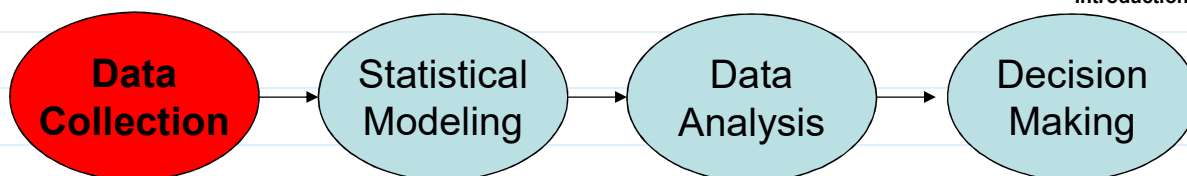
...

 $g_k(X_1, \dots, X_n)$  $\Theta$ 

Extract Information

Data Analysis

Decision Making

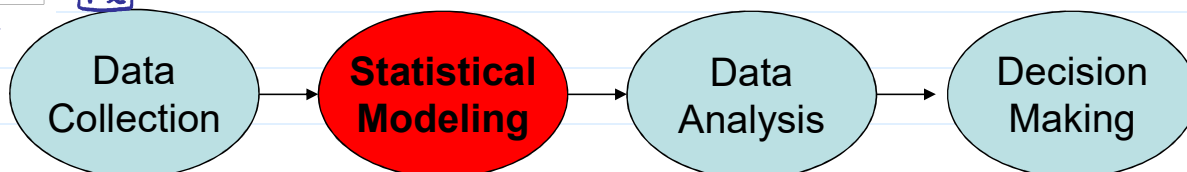


**1. Data collection:** producing representative data for drawing correct information

有代表性的

- design of experiment ← 實驗設計
- survey sampling ← 抽樣調查
- observational data ← uncontrollable

controllable



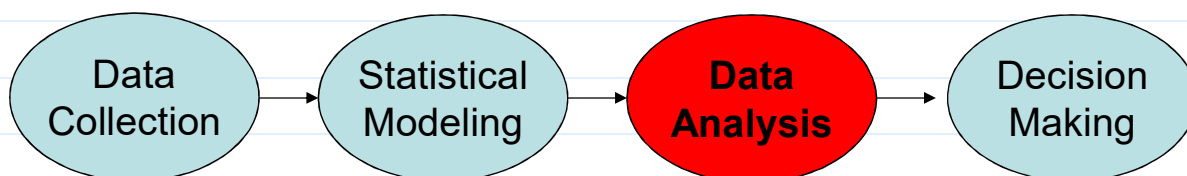
parameters (unknown)

**2. Statistical modeling:** using the information that we possess to develop a representation of the underlying system, which also accounts for uncertainty in data

probability

- Systematic part (規律)
- random part (隨機)

- a statistical model is a description of the joint distribution of data

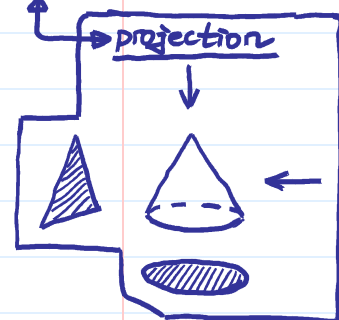


transformation (Ln, 1-2)

**3. Data analysis:** mining information from data

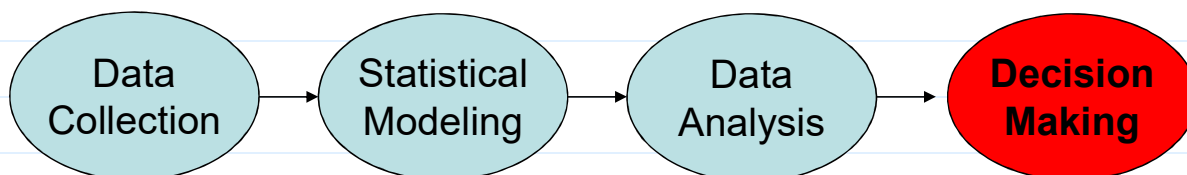
about parameters (規律)

- graphical methods
- numerical methods



- estimation ← 估計
- hypothesis testing ← 假設檢定

point estimation  
interval estimation



**4. Decision making:** drawing conclusions & answering questions based on results obtained in 3.

## Data collection

### Example (heat of fusion of ice, TBp. 423)

(Natrella, 1996) Two methods, A and B, were used in a determination of the latent heat of fusion of ice. The following table gives the change in total heat from ice at  $-0.72^{\circ}\text{C}$  to water  $0^{\circ}\text{C}$  in calories per gram of mass:

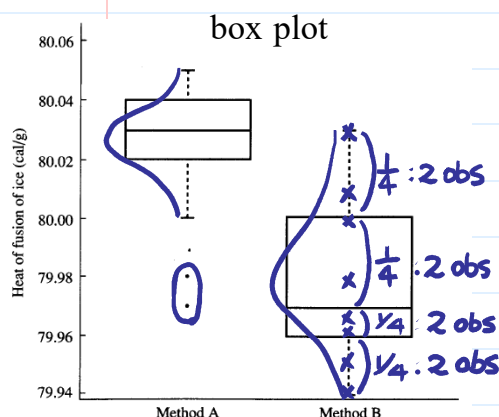
Method A	79.98	(80.04	80.02	80.04	80.03	80.03	80.04)	79.97
		(80.05	80.03	80.02	80.00	80.02)		
Method B	(80.02)	79.94	79.98	79.97	79.97	(80.03)	79.95	79.97

The investigators wished to find out:

how much the two methods “differ”?

- **Q:** Why not all the values from Method A/B are identical?  
*because of uncertainty.*
- **Q:** Beyond the uncertainty existing in the data, are there some “certain” information?  
*↑ 規律, systematic part in data.*

## Data analysis - graphical method



**Q:** From the plot, the two methods are different? or not different? and why?

### Question

How to model the data and the question, i.e., state them in a mathematical/statistical language?

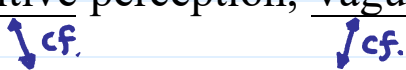
## Statistical modeling

- Let  $X_1, \dots, X_n$  be the  $n$  observations from method A
- Let  $Y_1, \dots, Y_m$  be the  $m$  observations from method B
- To account for the uncertainty in data,  
regard  $X_1, \dots, X_n$  and  $Y_1, \dots, Y_m$  as random variables.
- Assign distribution to random variables *parameters*  
*assumptions* { method A:  $X_1, \dots, X_n \sim \text{i.i.d. Normal}(\mu_X, \sigma^2)$   
method B:  $Y_1, \dots, Y_m \sim \text{i.i.d. Normal}(\mu_Y, \sigma^2)$  }

$$\mu_X = \mu_Y?$$

## Data analysis - numerical methods

- Estimation: what are the values of  $\mu_X, \mu_Y, \sigma^2$ ?
- Hypothesis testing:  $\mu_X = \mu_Y$ ? true or false? how confident?
  - $\hat{\mu}_X = 80.02, \hat{\mu}_Y = 79.98, \hat{\sigma}^2 = 0.0007178$
  - $p\text{-value} < 0.01, H_0: \mu_X = \mu_Y$  is rejected under significance level 0.99.

- 
- Compare the graphical and numerical methods
    - graphical methods: intuitive perception, vague conclusion  

    - numerical methods: lack of intuition, accurate conclusion

## Decision making

- There is a (statistically significant) difference between the means of the 2 methods:  $\mu_X > \mu_Y$
- level of evidence?

- Some other examples of statistical applications
  - Election: survey on voting
  - Lung cancer  $\longleftrightarrow$  Smoking
  - Moneyball (魔球)
  - Thinking, fast and slow (快思慢想)
  - The signal and the noise (精準預測)
  - Big data
  - Data-based AI
  - ...

- Materials to be covered in this course

- Probability – A Review: Chapters 1~6
- Estimation: Chapter 8
- Hypothesis Testing: Chapter 9
- Decision Theory: Chapter 15 (Rice, 1995, 2<sup>nd</sup> Edition)
- Applications:

method-  
ology

applied

- Survey Sampling: Chapter 7
- Two-Sample Comparison: Chapter 11
- Analysis of Variance: Chapter 12
- Some Graphical Methods from Chapter 10

Website of my mathematical  
statistics course

<http://www.stat.nthu.edu.tw/~swcheng/Teaching/stat3875/index.php>

**❖ Further reading:**

- Lewis (2004), Moneyball (中譯：魔球).
- Kahneman (2011), Thinking, Fast and Slow (中譯：快思慢想).
- Silver (2012), The Signal and the Noise (中譯：精準預測).
- Stigler (2016), The Seven Pillars of Statistical Wisdom.