Question

What is Statistics?

- A branch of math --- calculation, derivative, proof, …
- A collection of many statistics (formula)
- A useful tools for extracting information/knowledge from the data

Data → Information → Knowledge → Wisdom
What? ← Information
How? ← Knowledge
Why? ← Wisdom

哈利波特	Real Life
占卜學	Statistics
崔老妮	Statisticians
水晶球	Data
未來的資訊	Information

Statistics ← (卸貨)方法 → transformation
Data ← 載體
Information ← 貨品
$\mathbb{R}^n$ (data) → $\mathbb{R}^1$, $\mathbb{R}^2$, ⋮, graphs

into → insight ← see, watch

***aim of statistics**: provide **insight** by means of data*

insight: 对現象、狀況、系統、…

Basic Procedures of Statistics

- Statistics divides the study of data into *five* steps:

仿製品 ‖ 建立模型

Problem Formulation & Modeling (conceptual)

Data: $X_1, \ldots, X_n$

Statistician may get involved from here

Data Collection

c.f.

Statistical Modeling (empirical)

Data Analysis

Decision Making

> ➢ **Q**: What is a statistical model?
>
> ➢ $X_1, \ldots, X_n$ (random variables) $\sim$ joint cdf F_X/pdf f_X/pmf p_X with parameters Θ

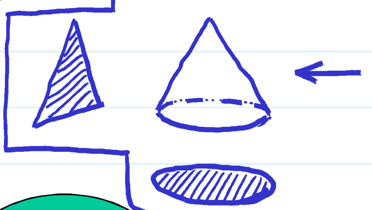
規律 (related to the answer to the question)

Q: How to find good transformation? Recall. mathematical statistics

statistics or graphs

Transformations		
$X_1, \ldots, X_n$	$g_1(X_1, \ldots, X_n)$, …, $g_k(X_1, \ldots, X_n)$	→ Θ
	Extract Information	

projection



1. Problem formulation & modeling (conceptual approach)

- Problem formulation: use statistical/probabilistic/ mathematical language to "clearly" define the problem and the objective of study

key

problem → formulation → statistical problem

- modeling (conceptual approach): use the information that we possessed *prior to obtaining data* to develop a *representation of the underlying system*, also account for uncertainty in data

domain knowledge

key: good surrogate

deterministic / systematic component + random component

規律 + 隨機

signal + noise

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

有代表性的

- survey sampling (抽樣調查)
- design of experiment (實驗設計)
- observational data

Data
① may contain wrong information,
② may not contain useful information about the problem.

3. Statistical modeling (empirical approach): use empirical information contained in the data to build a model or to justify/adjust the (conceptual) model developed in 1., also account for uncertainty in data

- a statistical model is a description of the joint distribution of data
- a statistical model may contain the following components:
  - nonparametric component
  - parametric component: (fixed, random) effects
  - distribution component

# of parameters $\overset{?}{=} \infty$

c.f. semi-parametric

deterministic component (規律)

random component (隨机)

4. data analysis: mining information from data

about parameters (規律)

- graphical methods
- numerical methods
  - (point, interval) estimation
  - hypothesis testing

objective probability

subjective probability

5. Inference/decision making: drawing conclusions & answering questions based on results obtained in 4.

- Example (from Gilchrist, *Statistical Modelling*, 1984):

  "A range of problems related to the positioning of stores and the planning of delivery routes requires information on the distances by road, $\boldsymbol{y}$, between different places. Where a large number of such places are involved, finding these distances by driving or by direct measurement along the roads on a map is time-consuming."

Q: problem "clearly" defined?

Problem: how to measure road distance y btw any 2 stores.

---

formulation

a statistical problem

"To avoid this problem, the usual approach is to relate the road distances $\boldsymbol{y}$ to the straight line distance, denoted by $\boldsymbol{x}$, as measured using a scale map. This relationship will be expressed mathematically and will enable us to predict a value of y given a corresponding value of x. This relationship will be our quantitative model of the situation. The fundamental question is: *how do we obtain this relationship (model)*."

$\hat{y} = \underbrace{f}_{\text{unknown}}(x)$

- how can we understand it using data?
- do we have some prior information about f?

Let's assume the following conditions (are they reasonable?):

a) $x=0 \Rightarrow y=0$ $x=12$, $y=12$ $x=12$, $y>12$

b) If there is a straight road between two points, then $x=y$; otherwise, $y \geq x$

c) Generally, y should increase with x. However, because of randomness in road patterns, places with same x's may have different y's.

d) Under similar situations, e.g. urban roads, the form of the relationship should not depend strongly on the distances involved, i.e., if x is, say, doubled, we would expect y is also approximately doubled.

Consider the following relationships (models):

1.	$y=x$	[satisfies a) and d), but not b) or c)]
2.	$y=x+\varepsilon$, ε: random component	[now allows c), but not b)]
3.	$y=\alpha+x+\varepsilon$, α: a constant	[helps with b), but a) fails]
4.	$y=\underbrace{\beta x}_{\text{systematic component}}+\underbrace{\varepsilon}_{\text{random component}}$, β: a constant ≥ 1	[satisfies all four conditions. true?]
5.	distribution assumption can be added on the ε in *4*, e.g., $\varepsilon \sim N(0, \sigma^2)$	

Note: The above (conceptual) model is derived without any data provided.

Problem formulation: Estimate and test parameters in $y=\beta x+\varepsilon$, where $\beta \geq 1$

a statistical problem "clearly" defined.

variables in data set

Some Notes in Problem formulation & modeling (conceptual approach)

statistics: 輔助科學

- understand the physical/social/political/biological/medical/... background to avoid the missing of important conditions that should be included in model

domain knowledge

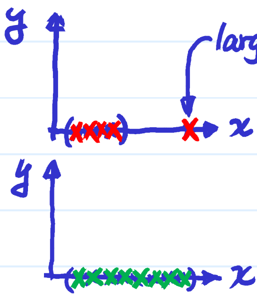
- understand the objective → e.g. Cox proportional hazard model

semi-parametric
- parametric part (規律 of interest)
- nonparametric part (規律 not of direct interest)

- make sure you know what the client wants
- state the problem in "statistical language"

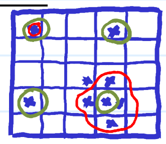
> **Albert Einstein**. *The formulation of a problem is often more essential than its solution which may be merely a matter of mathematical or experimental skill.*

Example (cont.):



y	x
10.7	9.5
6.5	5
29.4	23
17.2	15.2
18.4	11.4
19.7	11.8
16.6	12.1
29	22
40.5	28.2
14.2	12.1
11.7	9.8
25.6	19
16.3	14.6
9.5	8.3
28.8	21.6
31.2	26.5
6.5	4.8
25.7	21.7
26.5	18
33.1	28

- the collected data are given in the tabular. Is it a "representative" data set?



e.g. record data on some day

- observational or experimental data?

e.g., uniformly choose stores on the map.

- **Q**: If you can design the experiment, what are the data collection issues that should be concerned in the example?
- Consider the following situations:
  - ➢ if there are hundred/thousand of places, how to choose a small number of appropriate locations? geometrically uniform allocation? stratified sampling?

  - ➢ what if there are many routes that link any two places? replication required? → can be used to understand "information" about the variation caused by routes
  - ➢ who should be assigned to measure these y's by driving? randomization? blocking?

Some Notes in Data Collection

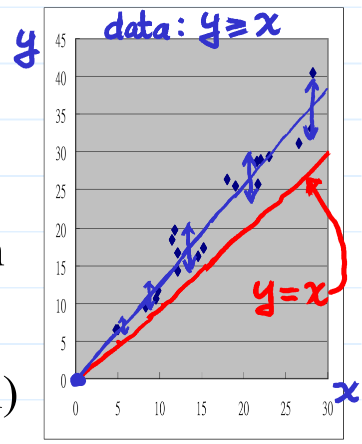
- are the data observational or experimental?
- how to collect a representative data?
- is there non-response? It's also informative.
- are there missing values? missing information. MCAR, MAR, MNAR
- qualitative or quantitative?
  類別型 連續型
- how are the data coded?
- what are the units of measurement?
- beware of data entry errors ← Data sanity check, Data cleaning

Example (cont.):

- What empirical model will you suggest after examining the plot?
- should empirical model be identical to conceptual model?
- if the plot (or numeric analysis) reveals different patterns ...
  - ➢ what if you find curvature or jump relationship existing between x and y?
  - ➢ what if you find non-constant variance?

  how should the conceptual model be adjusted?

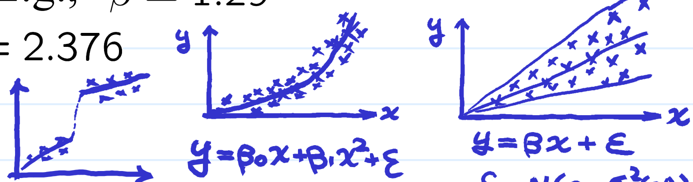
$y=\beta x+\varepsilon$, $\beta \geq 1$



- graphic analyses offer vivid and intuitive perception
- numeric analyses present numeric summaries (such as estimation and testing of parameters in the model) for making concrete conclusions. E.g., $\hat{\beta}=1.29$ and is significant in t-test, and $\hat{\sigma}=2.376$

final fitted model

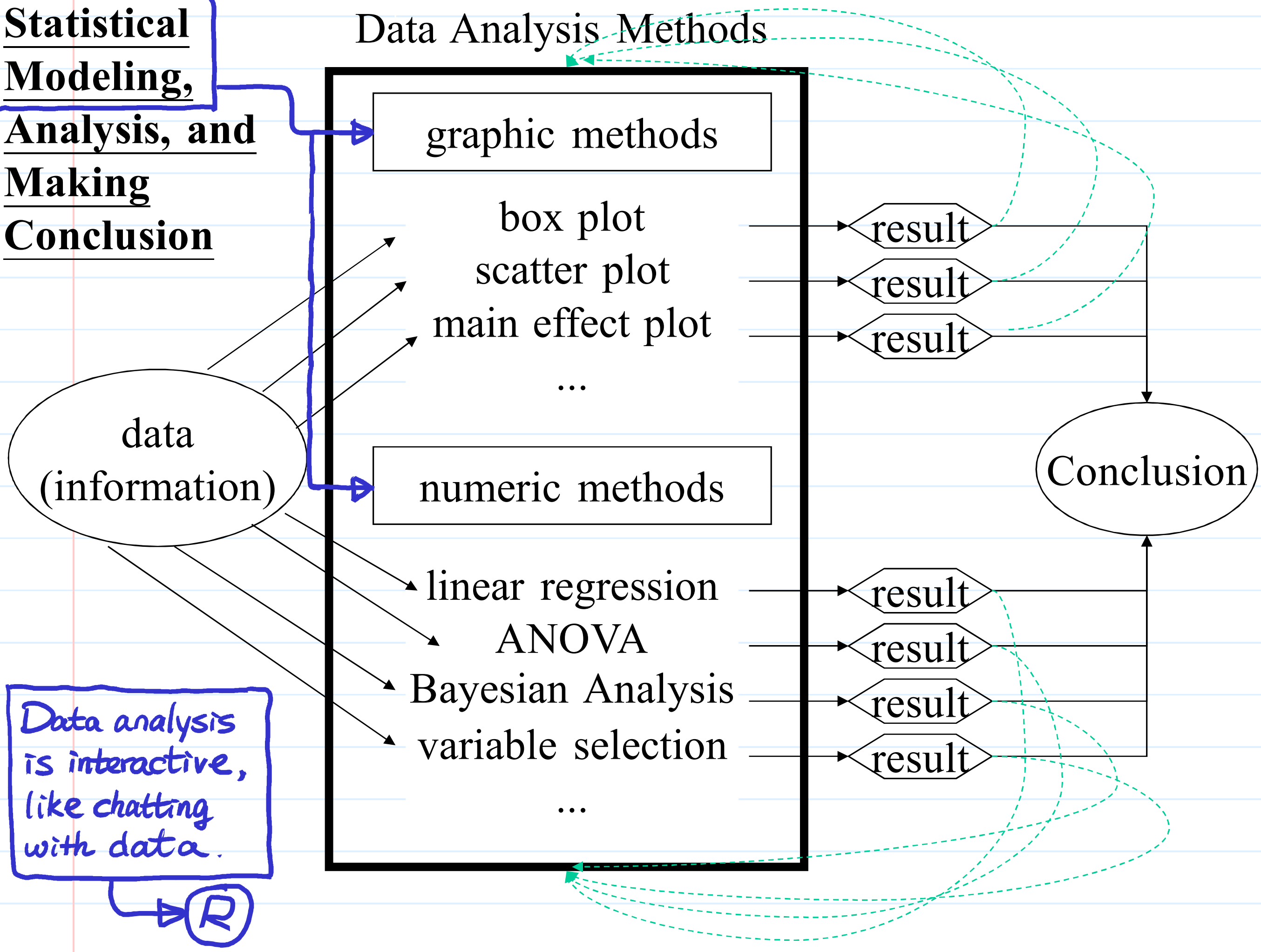
⊙ Conclusion: $\hat{y}=1.29x$ (or offer confident interval of $\hat{y}$)



$y=\beta_0 x+\beta_1 x^2+\varepsilon$

$y=\beta x+\varepsilon$, $\varepsilon \sim N(0, \sigma^2(x))$

Statistical Modeling, Analysis, and Making Conclusion



Some Notes in Statistical modeling, Data analysis, and Decision making

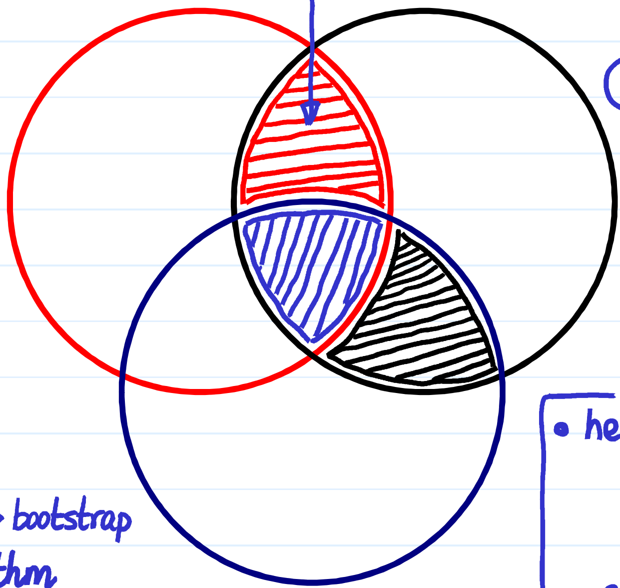
- If possible, most available analysis methods should be performed. Data analysis ⇔ projection from different angles
- Assumptions and analysis results between different methods could be (slightly) different
  → hidden in statistical models
- Data analysis is inherently interactive
- Conclusions should be summarized based on consistent results. → level of evidence
- Important information usually consistently appear in the results of every methods
- quantitative (定量) and qualitative (定性) conclusions

A successful data analysis usually requires a mix of the three components:

an AI is a fitted model

statistical learning (machine learning)

Computation
- programming
- numerical optimization
- processing & managing massive amounts of complex data



Statistics
- statistical modeling
- transformation of data
- random mechanism

★ (exact & asymptotic) sampling distribution → bootstrap

★ MLE → EM algorithm

★ posterior distribution → MCMC

⋮

Domain knowledge
- help interpretation of analysis results (connecting the results with phenonmena)
- offer information that might not be always available in data

required ❖ **Reading**: Faraway (2005, 1st edition), 1.1

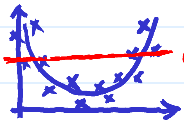
optional ❖ **Further reading**: D&S : Draper and Smith (1998)

- Statistical modelling (Gilchrist, 1984)
- Statistics: a guide to the unknown (edited by Tanur et al., 1972, 1978, 1989; Peck et al., 2005)
- Applied statistics: principles and examples (Cox & Snell, 1981)

❖ **Some other 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).

What aspects you should focus on in this course?

1. **Understand** analysis methods
   - objective is ...?
   
   model: $y=\beta_0+\beta_1 x+\varepsilon$
   $H_0: \beta_1=0$
   - for an estimator (parameter), what's its meaning?
   - for a test, what are its H_0 and H_1?
   collection of models
   $H_0 \cup H_1$: collection of all the models (parameters) considered in the test.
   - how to find statistically significant results in outputs?
   - assumptions and limitations in a statistical model?
   - ...

   not only p-value > or < 0.05
   level of evidence (LN p. 1-12)
2. **Interpretation**: for those significant results, how to interpret them in the language that your clients use
3. **How to implement** the analysis method in softwares, such as R, Splus, SAS, ...?