Question

What is Statistics?

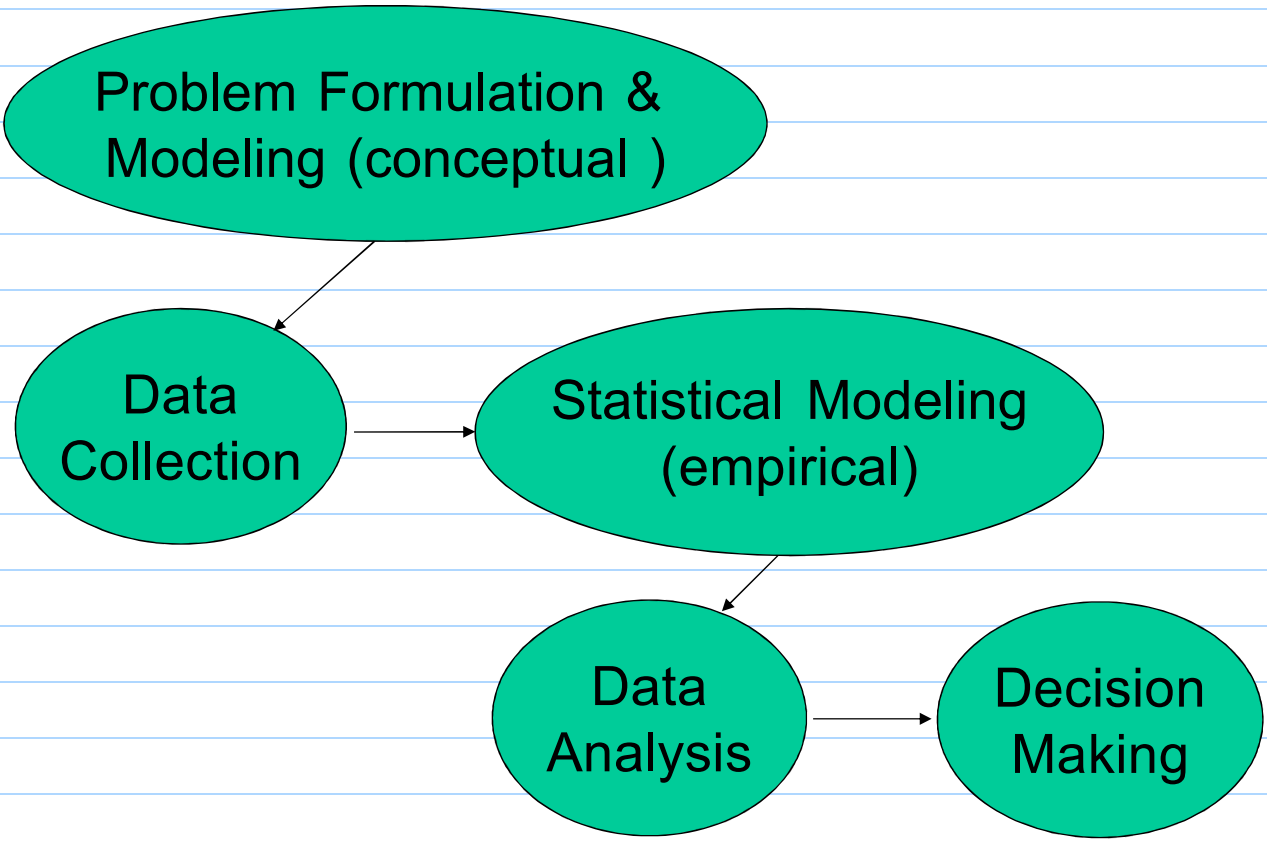
- A branch of math --- calculation, derivative, proof, …
- A collection of many statistics (formula)
- A useful tools for extracting information/knowledge from the data

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

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

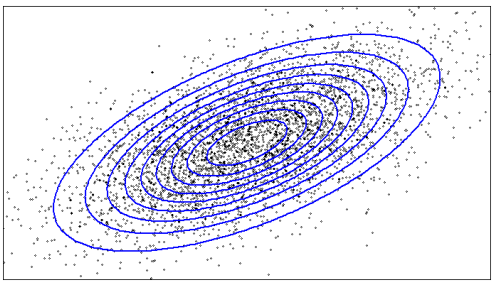
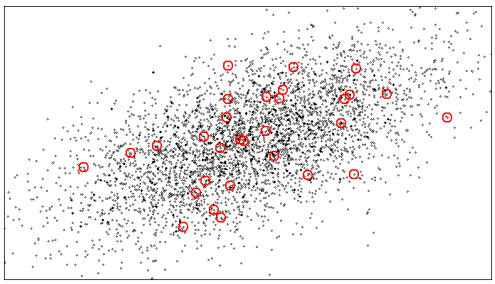
Basic Procedures of Statistics

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

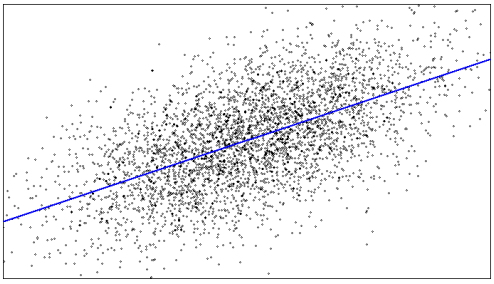
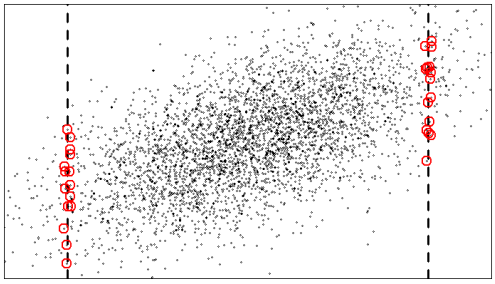


Distinction between Survey Sampling and DOE

- Survey sampling: interested in "distribution"

- DOE: interested in "relationship"

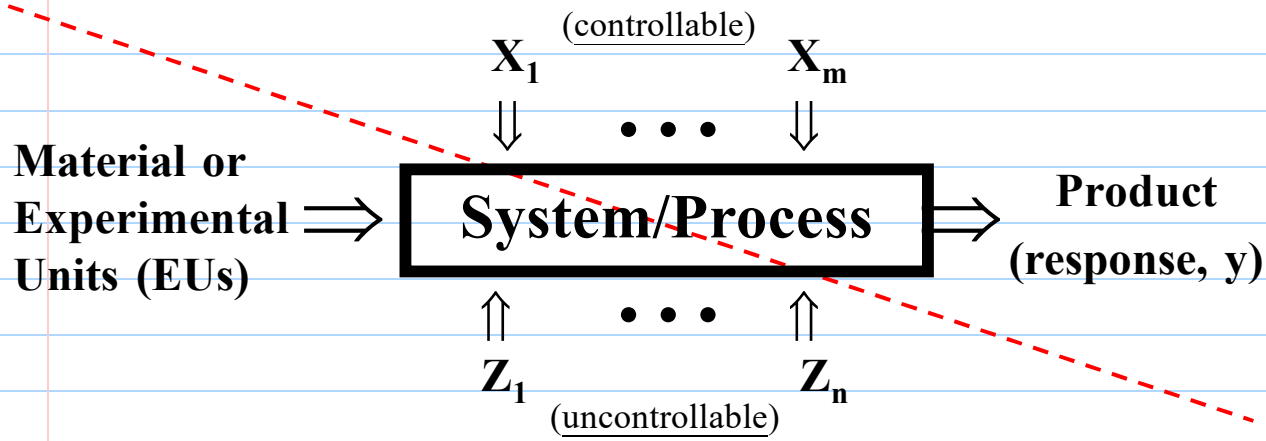



Experimentation

- Experimentation: one of the most common activities that people engage in
  - ➢<神農本草經>:

    "神農嘗百草，日遇七十二毒，得荼而解之"
  - ➢Experiment: a learning process for
    - knowledge gathering
    - problem solving
    - conjecture testing
- Modern experimental design dates back to 1930s
- The results of an experiment often contains two parts:
  - ➢ *deterministic* component
  - ➢ *random* component

Conceptual model



$$y = f(X_1, X_2, \ldots, X_m) + \epsilon$$

Deterministic/systematic component

error: **random component**

Historical perspectives

- **Agricultural Experiments :** Comparisons and selection of varieties (and/or treatments) in the presence of uncontrollable field conditions, Fisher's pioneering work on design of experiments and analysis of variance (ANOVA).

- **Industrial Era :** Process modeling and optimization, Large batch of materials, large equipments, Box's work motivated in chemical industries and applicable to other processing industries, regression modeling and response surface methodology.

Historical perspectives (Contd.)

- **Quality Revolution :** Quality and productivity improvement, variation reduction, total quality management, Taguchi's work on robust parameter design, Six-sigma movement.
- A lot of successful applications in manufacturing (cars, electronics, home appliances, etc.)
- **Current Trends and Potential New Areas :** Computer modelling and experiments, large and complex systems, applications to biotechnology, nanotechnology, material development, etc.

Some Definitions

- **Factor :** variable whose influence upon a response variable is being studied in the experiment.
- **Factor Level :** numerical values or settings for a factor.
- **Experimental unit :** object to which a treatment is applied.
- **Trial** (or **run**) : application of a treatment to an experimental unit.
- **Treatment or level combination :** set of values for all factors in a trial.
- **Randomization :** using a chance mechanism to assign treatments to experimental units or run order.

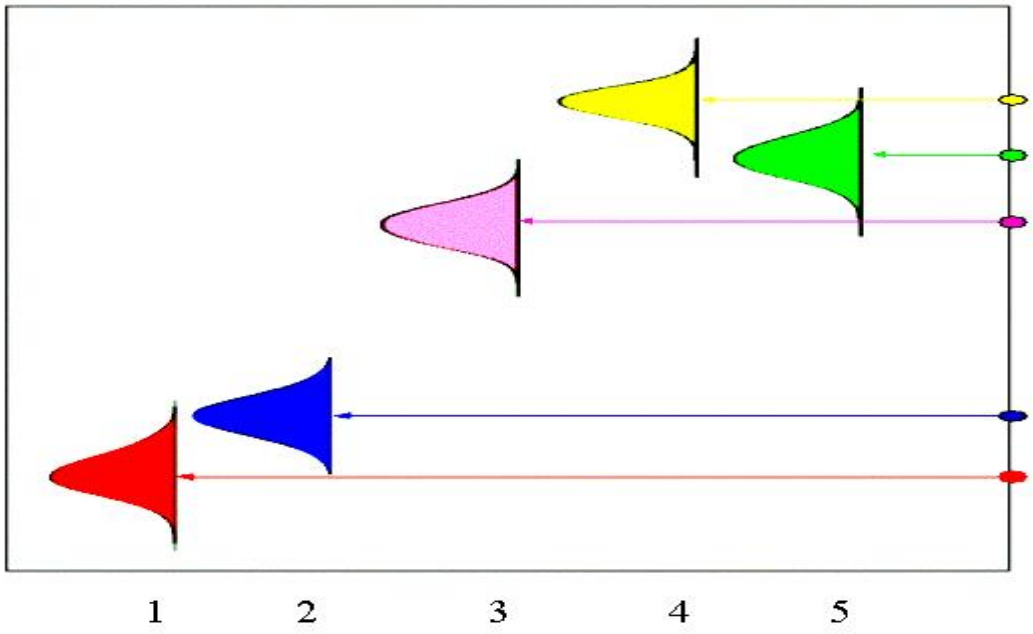
Types of Experiments

- **Treatment Comparisons :** Purpose is to compare several treatments of a factor (have 4 rice varieties and would like to see if they are different in terms of yield and drought resistence).
- **Variable Screening :** Have a large number of factors, but only a few are important. Experiment should identify the important few.
- **Response Surface Exploration :** After important factors have been identified, their impact on the system is explored; regression model building.

Types of Experiments (Contd.)

- **System Optimization :** Interested in determining the optimum conditions (e.g., maximize yield of semiconductor manufacturing or minimize defects).
- **System Robustness :** Wish to optimize a system and also reduce the impact of uncontrollable (noise) factors. (e.g., would like cars to run well in different road conditions and different driving habits; an IC fabrication process to work well in different conditions of humidity and dust levels).

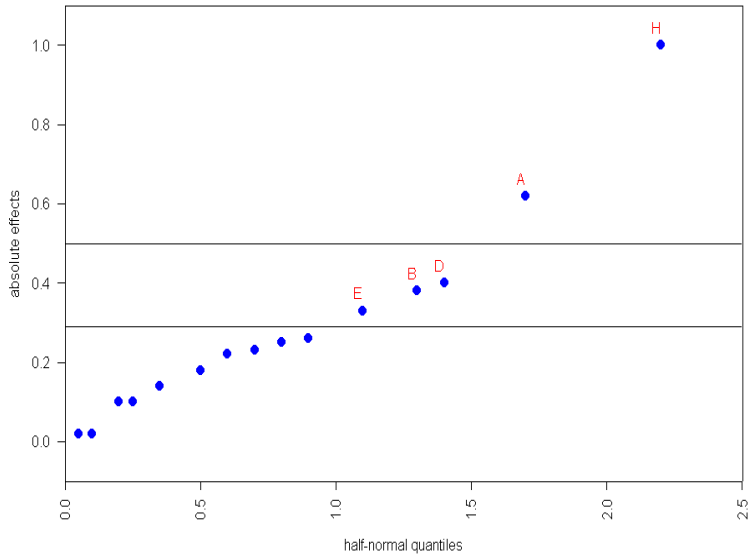
Treatment Comparison



treatment: a combination of factor levels

Variable (Factor) Screening

Objective: identify important factors or screen out unimportant factors



H, A: important

D, B, E: moderate

H+A: 50%

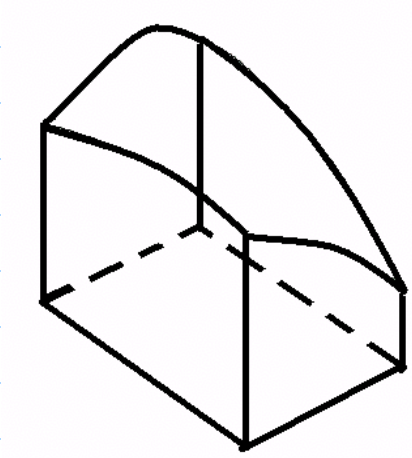
H+A+D+B+E: 65%

When to use?: usually in the *preliminary* stage of the study of a system/process

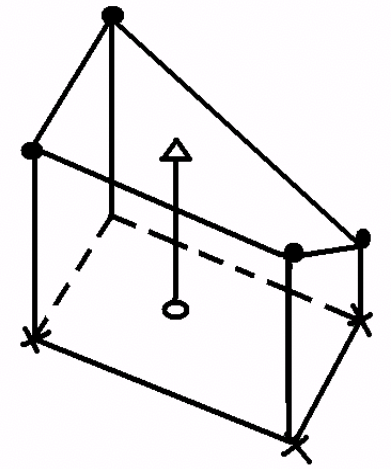
Response Surface Approximation

response surface: the relationship between a response and the factors

Objective:
develop a good approximation of the response surface



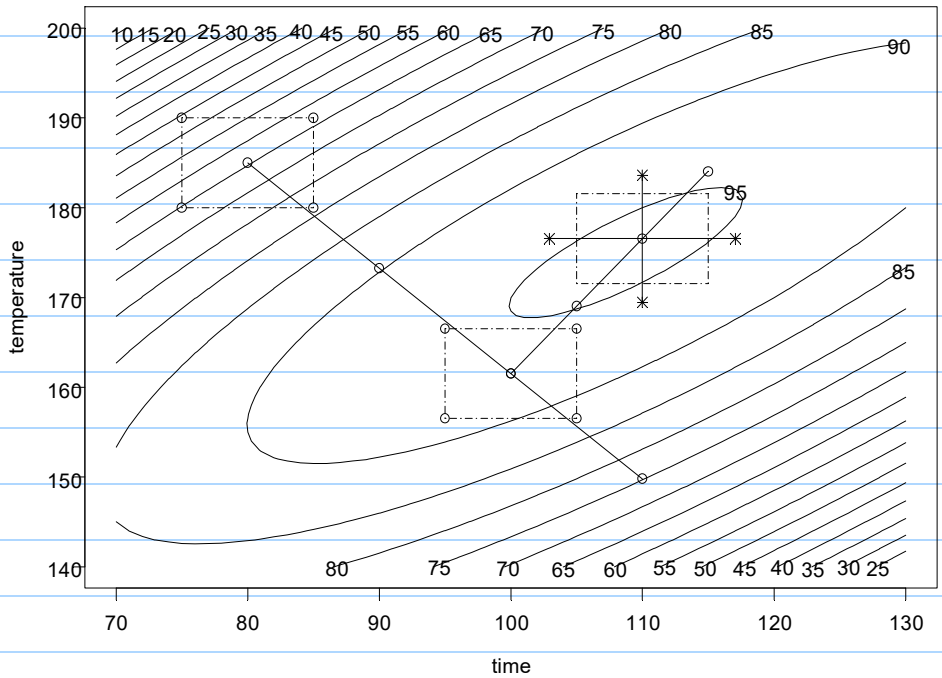
true response surface



approximate response surface

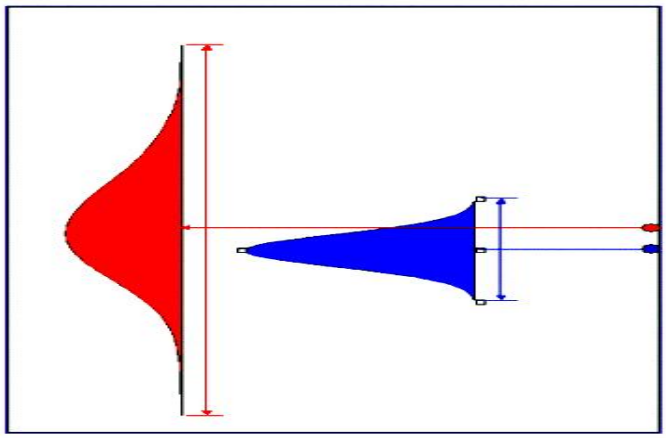
System/Process Optimization

Objective: obtain optimal setting (of minimum/maximum response)



Variation Reduction

Objective: adjust treatment factors to make the system/process robust against noise variation



Q: which one will you choose?

Concept: *Besides optimizing the response, variation reduction is important in quality improvement.*

Systematic Approach to Experimentation

1. State the objective of the study.
2. Choose the response variable … should correspond to the purpose of the study.
   - Nominal-the-best, larger-the-better or smaller-the-better.
3. Choose factors, levels, experimental region.
   - Use flow chart or cause-and-effect diagram.
4. Choose experimental design (i.e., plan).
5. Perform the experiment (use a planning matrix to determine the set of treatments and the order to be run).
6. Analyze data (design should be selected to meet objective so that the analysis is efficient and easy).
7. Draw conclusions.

Fundamental Principles : Replication, randomization, and blocking

Replication

- Each treatment is applied to units that are representative of the population (example : measurements of 3 units vs. 3 repeated measurements of 1 unit).
- Replication vs Repetition (i.e., repeated measurements).
- Enable the estimation of experimental error. Use sample standard deviation.
- Decrease variance of estimates and increase the power to detect significant differences : for independent y_i's,

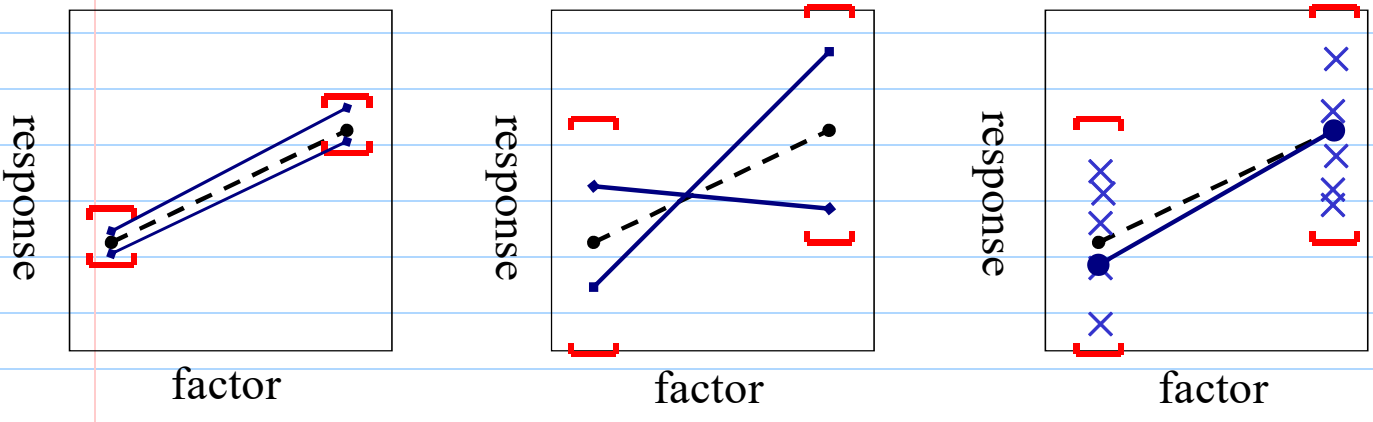
$$Var(\frac{1}{N}\sum_{i=1}^{N} y_i) = \frac{1}{N} Var(y_1).$$

Replicates and Experimental Errors

replicate: replication of same treatment

$$y = f(X_1, X_2, \ldots, X_m) + \epsilon$$
$$= \hat{f}(X_1, X_2, \ldots, X_m) + \hat{\epsilon}$$

- Q: Why do we need to understand the magnitude of exp'tal error?



Randomization

Use of a chance mechanism (e.g., random number generators) to assign treatments to units or to run order. It has the following advantages.

- Protect against latent variables or "lurking" variables (give an example).
- Reduce influence of subjective bias in treatment assignments (e.g., clinical trials).
- Ensure validity of statistical inference (This is more technical; will not be discussed in the book. See Chapter 4 of "*Statistics for Experimenters*" by Box, Hunter, Hunter for discussion on randomization distribution.)

Effect Aliasing/Confounding

	A	B	C	Operator
1	low	low	low	Peter
2	low	low	high	Peter
3	low	high	low	Peter
4	low	high	high	Peter
5	high	low	low	John
6	high	low	high	John
7	high	high	low	John
8	high	high	high	John

Q: what if operators have an effect on response?

	A	B	C	AB
2	low	low	high	high
3	low	high	low	low
5	high	low	low	low
8	high	high	high	high

- Q: Is aliasing/confounding always a bad thing?
  - pros & cons

Randomization

Q: what if operators have an effect on response?

	A	B	C	operator
1	low	low	low	Peter
2	low	low	high	Peter
3	low	high	low	Peter
4	low	high	high	Peter
5	high	low	low	John
6	high	low	high	John
7	high	high	low	John
8	high	high	high	John

	A	B	C	operator
5	high	low	low	Peter
2	low	low	high	Peter
8	high	high	high	Peter
4	low	high	high	Peter
3	low	high	low	John
1	low	low	low	John
6	high	low	high	John
7	high	high	low	John

Randomization provides protection against *extraneous factors* that are unknown to the experimenter, but may impact the response

- what should be randomized?
  - ➢ allocation of exp'tal materials to treatments; the order of applying treatments; the order of measuring responses; ...

Blocking

A **block** refers to a collection of homogeneous units. Effective blocking : larger between-block variations than within-block variations.
(Examples: hours, batches, lots, street blocks, pairs of twins.)

- Run and compare treatments within the same blocks. (Use randomization within blocks.) It can eliminate block-block variation and reduce variability of treatment effects estimates.
- **Block what you can and randomize what you cannot.**
- Discuss **typing experiment** to demonstrate possible elaboration of the blocking idea. See LNp.1-24.

Blocking

Q: If operator effect is identified as significant before exp't, what can we do?

	A	B	C	operator
1	low	low	low	Peter
2	low	low	high	John
3	low	high	low	John
4	low	high	high	Peter
5	high	low	low	John
6	high	low	high	Peter
7	high	high	low	Peter
8	high	high	high	John

Orthogonality

block factor : factors that are controllable and may influence the response but in which we are ***not directly interested***.
(cf. treatment factors)

Examples of blocking factors:

lot-to-lot, brand-to-brand, operator-to-operator, day-to-day, ...

block what you can & randomize what you cannot

Illustration: Typing Experiment

- To compare two keyboards A and B in terms of typing efficiency. Six manuscripts 1-6 are given to the same typist.
- Several designs (i.e., orders of test sequence) are considered:

1.

$$1.\,A,B,\ 2.\,A,B,\ 3.\,A,B,\ 4.\,A,B,\ 5.\,A,B,\ 6.\,A,B.$$

(A always followed by B, why bad ?)

2. Randomizing the order leads to a new sequence like this

$$1.\,A,B,\ 2.\,B,A,\ 3.\,A,B,\ 4.\,B,A,\ 5.\,A,B,\ 6.\,A,B.$$

(an improvement, but there are four with A,B and two with B,A. Why is this not desirable? Impact of *learning effect*.)

3. *Balanced randomization*: To mitigate the learning effect, randomly choose three with A,B and three with B,A. (Produce one such plan on your own).
4. Other improved plans?

KB	manu.	order
A	1	I
B	1	II
A	2	I
B	2	II
A	3	I
B	3	II
A	4	I
B	4	II
A	5	I
B	5	II
A	6	I
B	6	II

KB	manu.	order
A	1	I
B	1	II
B	2	I
A	2	II
A	3	I
B	3	II
B	4	I
A	4	II
A	5	I
B	5	II
A	6	I
B	6	II

KB	manu.	order
A	1	I
B	1	II
B	2	I
A	2	II
B	3	I
A	3	II
A	4	I
B	4	II
B	5	I
A	5	II
A	6	I
B	6	II

❖ **Reading**: textbook, 1.1, 1.2, 1.3