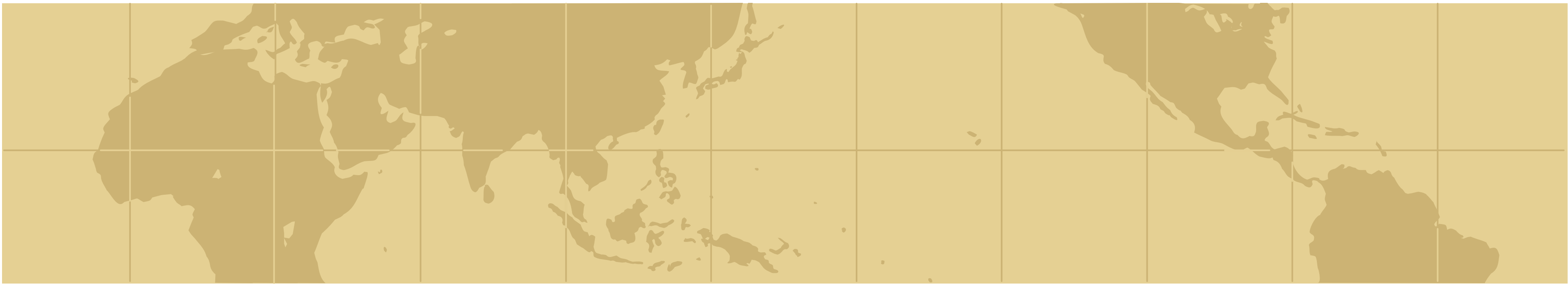
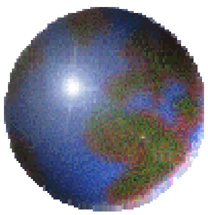




實驗設計發展史

林共進

美國賓州州立大學管理科學系

Design of Experiment

How to collect useful information?

	Agricultural Experiments	**Industrial Experiments**
Number of factors	Small	Large
Number of Runs	Large	Small
Reproducibility	Large	Small
Time taken	Long	Short
Blocking	Nature	Not obvious
Missing values	Often	Seldom
Randomization	Important	OK
Other		

Designing industrial experiments is very different
From designing agricultural experiments

Design Objectives

- Treatment Comparison
- Screening
- Model Building
- Parameter Estimation
- Optimization
- Prediction
- Confirmation
- Discovery (Random Shot)
- etc.

Design Methodology

- Treatment Comparison
- Fractional & Full Factorial Design
- Combinatorics Design
- Coding Theory
- Response Surface Methodology
- ANOVA type Design
- Optimal Design
- Bayesian (Optimal) Design

Design Methodology *(Continued)*

- Saturated (Minimal Point) Design
- Taguchi Product (Robust) Design
- Mixture Experiment
- Computer Experiment
- Supersaturated Design
- Uniform Design
- MicroArray Design

Summary: Design of Experiment

- Model is known
  - Optimal design
  - Optimality Criteria
    (e.g., alphabetical optimalities)
- Model is unknown
  (or is not completely known)
  - Bayesian Design
  - Robustness
  - Robustness Criterion
  - Representative Points

Design of Experiment: Looking Ahead

- Theoretical
  - Multiple response
  - Higher (mixed) level combinatorics
  - Analysis Methods (ANOVA→Regression→???)
- Practical (Restrictions)
  - Mixture
  - Error in variables
  - Run size, number of level consideration
  - Order in level values

Design of Experiment: Looking Ahead

- Applications
  - Information Technology

    (Computer Experiment)
  - Micro-Array

    (Gene Expression)
  - Data Mining—Data Squashing, Dimension Reduction
- Related Areas
  - Number Theory
  - Combinatorics
  - Coding Theory

Design of Experiment (Lin)

- Multiple Response Problems
  - Optimization: Kim and Lin (*JRSS-C*, 2000)
  - Design: Chang, Lo, Lin & Young (*JSPI*, 2001)
- Computer Experiment
  - Beattie and Lin (1998)
- Dispersion Effect
  - McGrath and Lin (*Technometrics*, 2002)
- Foldover Plan
  - Li and Lin (*Technometrics*, 2003)
- Supersaturated Designs
  - Lin (*Technometrics*, 1993, 1995, 2001) and others
- Uniform Designs
  - Fang, Lin, Winker & Yang (*Technometrics*, 1999)

致命的錯誤假設

- 陽光 ， 空氣 ， 水 ，
  取之不盡,用之不竭.

- 正確的數字.
  取之不盡,用之不竭.

正確的數字，需要投資

不可能從天上掉下來

Data

- Data Collection
- Data Preparation
- Data Quality
- Data Understanding
- Data Description
- Data Visualization
- Data Analysis

資料收集

- 抽樣理論 (Sampling)

  萬物有常,世事多變

- 實驗設計 (Design of Experiment)

Industrial Statistics

- Statistical Process Control
- Reliability
- Design of Experiments
- Others:
  - Data Mining, Neural Networks
  - Information Technology, Fuzzy
  - Marketing, Six-Sigma
  - etc. *--- Lin(1995)*

- Why Experiment?
  - Confirm OLD theory

    Newton's three laws
  - Discover NEW theory
- Why Design Experiments?
  - A lesson from Edison
  - 0 - 1 rule
- Statistics
  - Not to teach them *how to improve* but to teach them *how to speed up the improvement*

How to collect "useful" information?

"Design" occurs *before* "data analysis".

- **What do you have?**

- **What do you want?**

實驗設計的目的:

- ❖Treatment Comparison
- ❖Model Building
- ❖Parameter Estimation
- ❖Confirmation
- ❖Optimization
- ❖Screening
- ❖Discovery
- ❖etc.

Before Experiment

$$y = f(x_1,\ldots,x_p,\underbrace{x_{p+1},\ldots,x_k}) + \varepsilon$$

After Experiment

$$y = f(x_1,\ldots,x_p) + \varepsilon(x_{p+1},\ldots,x_k)$$

$$p << k$$

Saturated Design of First-Order Model

- Plackett & Burman Designs (1946)
- Regular Simplex Design (Box, 1952)
- Optimal Design
- p-efficient Design (Lin, 1993)
- Cyclic Orthogonal Design (Lin & Chang, 2000)
- Summary and Comparisons
- Minimal Point Designs

First-Order Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Higher-order terms, nonlinearity, noise, etc.

Testing

$$H_0^{(i)} : \beta_i = 0 \quad \text{vs.} \quad H_1^{(i)} : \beta_i \neq 0$$

Design Principle

- Simplicity
- Efficiency

實驗設計方法:

- 分析方法
- 實驗目的
- 優點
- 缺點
- 此設計是否能解決您的問題?

(1) Treatment Comparison

醫藥: pick-the-winner (直接擇優)

treatment assignment

有效樣本，實驗次數, Block Design

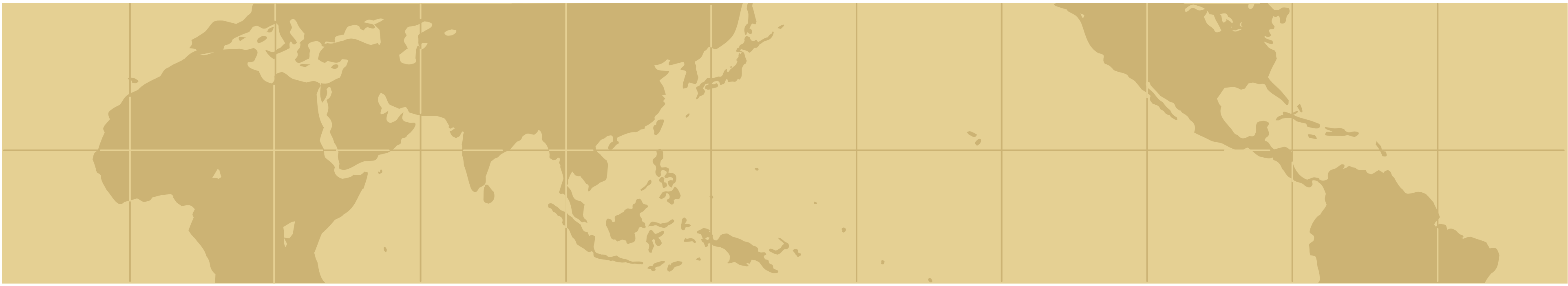
單因子

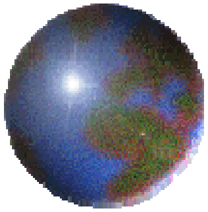
T_1: $y_{11}\ y_{12}\ ..\ y_{1n_1}$

T_2: $y_{21}\ y_{22}...y_{2n_2}$

T_3: $y_{31}\ y_{32}...y_{3n_3}$

Block: $B_1\ B_2...$





Type I Error: 天下本無事,庸人自擾之
(緊張大師)

Type II Error:不見棺材不掉淚
(麻木不仁)

(2) 所有可能性組合

- 全因子設計 (Factorial Design)，　多因子

例: X_1 有二種可能性 (1或2)，X_2 有二種可能性 (1或2)
X_3 有二種可能性 (1或2)

X_1	X_2	X_3
1	1	1
2	1	1
1	2	1
2	2	1
1	1	2
2	1	2
1	2	2
2	2	2

共計2^3次實驗

- 應用範圍極廣(尤其是人文科學上之研究)

(3) 部份可能性組合

部份因子設計 (Fractional Factorial Design)

全因子

X_1	X_2	X_3
1	1	1
2	1	1
1	2	1
2	2	1
1	1	2
2	1	2
1	2	2
2	2	2

部份因子

X_1	X_2	X_3
1	1	1
2	1	2
1	2	2
2	2	1

正交之觀念

- ❖R.A. Fisher (1920)
- ❖F. Yates

(4) 組合設計

Combinatorics & Coding Theory

- 正交條件
- 組合性質
- Blocking

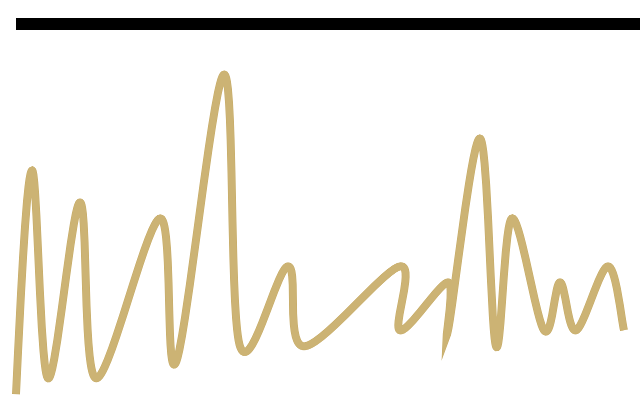
- Column-Row Design
- Weighing Design
- BIBD
- PBIBD
- Coding

- *R. C. Bose*
- *J. N. Srivastava*

(5) 反應曲面
Response Surface Methodology

將感興趣的變數視爲一未知的曲面分布

- ❖如何佈點
- ❖如何估計此曲面
- ❖如何找尋曲面的最高點



- ❖ G. E. P. Box
- ❖ N. R. Draper
- ❖ R. H. Myers

(6) 變異數表型設計

ANOVA type Design

假想實驗結果爲 $y_1\ y_2 \ldots y_n$

總變異數爲(Total Sum Squares)

$$TSS=(y_1-\overline{y})+(y_2-\overline{y})+\ldots+(y_n-\overline{y})$$

$\overline{y}$ 爲平均值

此變異數可能的來源爲何?

Source	Sum squares	Degree Freedom
變數 1		
變數 2		
Block 1		
交互作用		
Total	TSS	n-1

估計其理論分布爲何?

Split-plot Design
Nested Design
Factorial Design
etc.

❖V. L. Anderson (Purdue)
❖C. R. Hicks

Randomization
Missing values
etc.

(7) 最優設計
(Optimal Design)

Design

$X(m)=(x_1(m),x_2(m),\ldots,x_k(m))$ *in* I^k

Model

$Y(X(m))=f(X(m))'\beta + \varepsilon_m$ m=1,2,...n

Let ξ be the probability measure on I^k

$$M(\xi)= \int_{I^k} f(x)' f(x) d\xi(x)$$

Information matrix

- J. Keifer
- C. S. Cheng (鄭清水)

(8) 貝斯設計 (Bayesian Design)

- 針對 optimal Design
  加入先驗函數(prior)
- 另類 Optimal Design
- Optimal in term of a group of models

Optimal Design

- Specify the model
- Specify the optimality criterion
- Construct the design

- For Bayesian Design, add the prior distribution to the model information.

(9) Saturated Design (Minimal-Point Design)

- Given the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- The design matrix is *n* x *(k+1)*
  Find a (k+1) x (k+1) matrix which is "optimal".
- For D-opt, this is a Det-max matrix.

(9) Saturated Design (Minimal-Point Design)

- Given a n x n matrix, with 2 symbols, what is its maximal determinant possible?
  - (Hadamard, 1893).
  - Hadamard Matrix, for n=1, 2, and 4t, is also known as Plackett and Burman Design.
- P-efficient design

  Lin (1993).
- Extension: Saturated Second-order designs...
- Non-orthogonality.

(10) 田口設計 (Taguchi Design)

三次設計(系統設計.參數設計.容差設計)

基本上採用

組合設計

部份(全)因子設計

內.外正交表

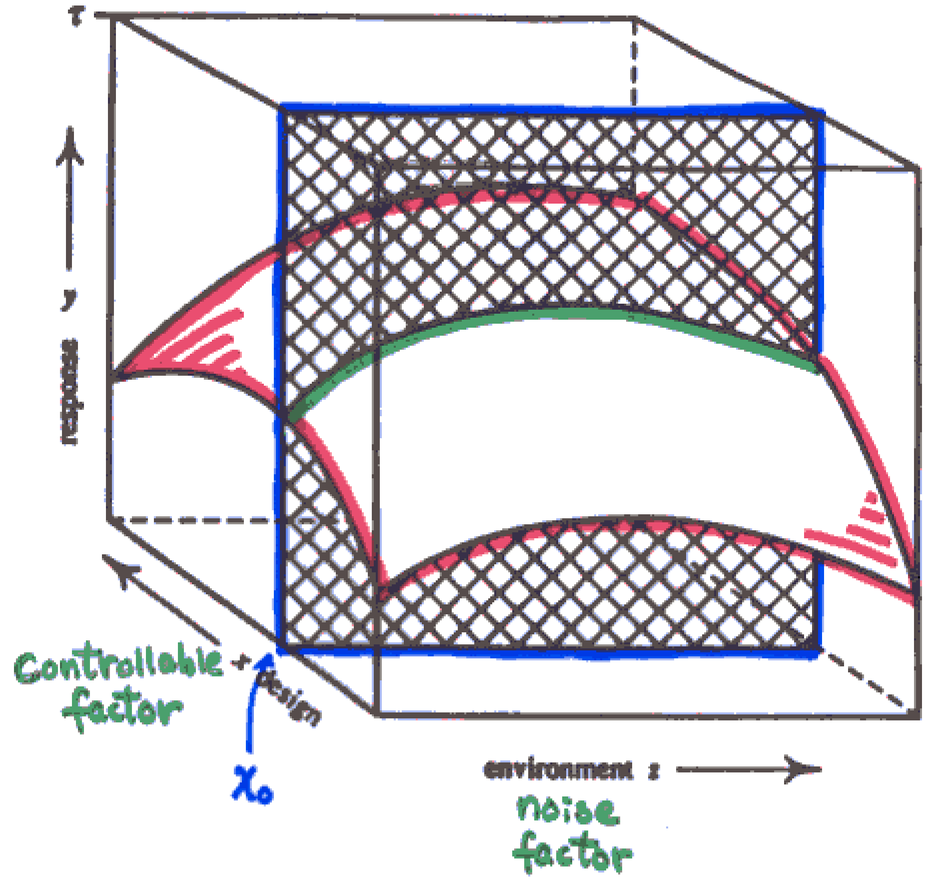
實驗次數偏多

可控因子: X1 X2 X3

不可控因子: Z1 Z2

X_1	X_2	X_3	Z_1	Z_2
1	1	1	1	1
			2	1
			1	2
			2	2
2	1	1	1	1
			2	1
			1	2
			2	2
1	2	1	1	1
			2	1
			1	2
			2	2
			…	…
2	2	2	1	1
			2	1
			1	2
			2	2

- G. Taguch (田口玄一)
- 張里千


response y
Controllable factor
design
x_0
environment z
noise factor

X_1	X_2	X_3
1	1	1
2	1	1
1	2	1
2	2	1
1	1	2
2	1	2
1	2	2
2	2	2

- One Observation
  Y_1

- Several Replicates
  $Y_{11}, Y_{12}, \ldots, Y_{1n}$

- Designing these Replicates

Z_1	Z_2	**Obs**
1	1	Y_{11}
2	1	Y_{12}
1	2	Y_{13}
2	2	Y_{14}

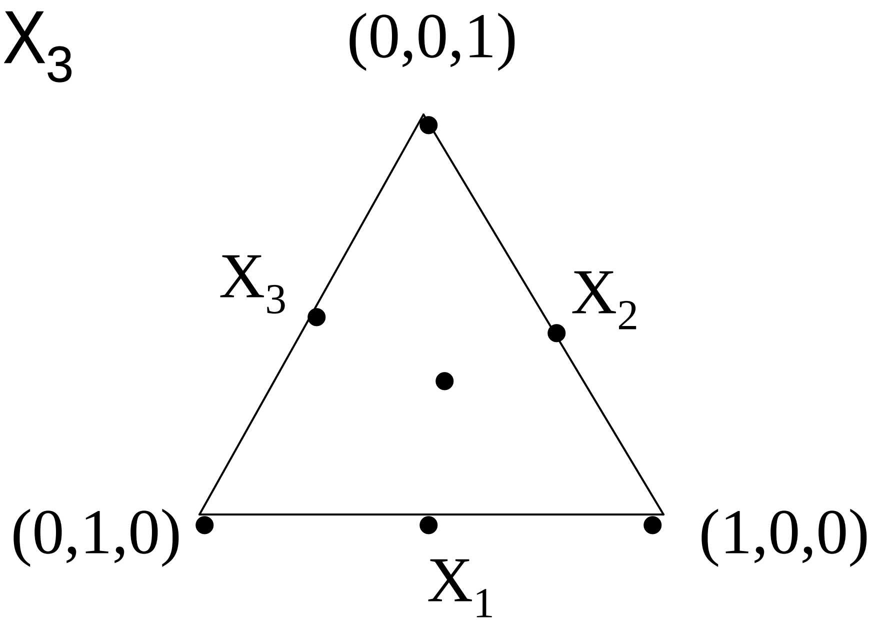
- Compound Orthogonal Arrays

- Uniform Design

(11) 混合設計
(Mixture Experiment)

多因子$X_{1,}$ $X_{2,}$ ... X_k，　但$X_1+X_2+\ldots+X_k=T$ (固定值)

例:X_1 X_2 X_3



化工上應用極為普遍

Scheffe (1958)，J. A. Cornell (1990)

(12) 均勻設計 (Uniform Design)

$\hat{F}_n(x)$ = **Empirical** Cumulative Distribution Function

$\hat{F}(x)$ = **Uniform** Cumulative Distribution Function

Find $x = (x_1, x_2, ..., x_n)$

such that $\hat{F}_n(x)$ is closest to $\hat{F}(x)$

Discrepancy

$$D = \left[\int_{\Omega} \left\| \hat{F}_n(x) - F(x) \right\|^p dx \right]^{1/p}$$

- 華羅庚，王元（數論）
- 方開泰

One-Dimension Optimal Representing Points in [0,1]

n	Uniform	Normal	Exponential
2	0.25	0.3876	0.0575
	0.75	0.6124	0.2773
3	0.1667	0.3388	0.0365
	0.5000	0.5000	0.1386
	0.8333	0.6612	0.3584
4	0.125	0.3083	0.0267
	0.375	0.4470	0.0940
	0.625	0.5530	0.1962
	0.875	0.6917	0.4159
5	0.1	0.2864	0.0211
	0.3	0.4127	0.0713
	0.5	0.5000	0.1386
	0.7	0.5873	0.2408
	0.9	0.7136	0.4605
6	0.0833	0.2695	0.0174
	0.2500	0.3876	0.0575
	0.4167	0.4650	0.1078
	0.5833	0.5350	0.1751
	0.7500	0.6124	0.2773
	0.9167	0.7305	0.4970

Wang, Lin, and Fang (1993)

The centered L_p-discrepancy is invariant under exchanging coordinates from x to 1-x. Especially, the centered L_2-discrepancy, denoted by CL_2, has the following computation formula:

$$(CL_2(\mathrm{P}))^2$$

$$= \left(\frac{13}{12}\right)^s - \frac{2}{n}\sum_{k=1}^{n}\prod_{i=1}^{s}\left(1 + \frac{1}{2}|x_{ki} - \frac{1}{2}| - \frac{1}{2}|x_{ki} - \frac{1}{2}|^2\right)$$

$$+ \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{n}\prod_{i=1}^{s}\left[1 + \frac{1}{2}|x_{ki} - \frac{1}{2}| + \frac{1}{2}|x_{ji} - \frac{1}{2}| - \frac{1}{2}\left|x_{ki} - x_{ji}\right|\right].$$

均勻設計未來定位:

特性:

- 幾何代表性(均勻性)
- 穩健性(Robust)
- 實驗次數可大可小
- 水平數(Level) 可多可少

均勻設計未來定位:

範圍:

- 可計算實驗(Computer Experiment)
- 函數性質已知,但過於複雜
- 函數性質不明確
- 全面了解,而非水平組合比較
- 模型更有彈性

  All models are wrong, some are useful.
- 均勻測度並非統計量?

  無法與傳統統統計分析方法結合

(13) 超飽和設計 (Supersaturated Design)

實驗次數少於因子個數

#	X_1	X_2	….	X_{24}	結果
1					y_1
2					y_2
.					.
.					.
.					.
14					y_{14}

Dennis Lin

Supersaturated Design From Hadamard Matrix of Order 12
(Using 11 as the branching column)

Run	Factors											
No.	I	1	2	3	4	5	6	7	8	9	10	(11)
1	+	+	+	-	+	+	+	-	-	-	+	-
2	+	+	-	+	+	+	-	-	-	+	-	+
3	+	-	+	+	+	-	-	-	+	-	+	+
4	+	+	+	+	-	-	-	+	-	+	+	-
5	+	+	+	-	-	-	+	-	+	+	-	+
6	+	+	-	-	-	+	-	+	+	-	+	+
7	+	-	-	-	+	-	+	+	-	+	+	+
8	+	-	-	+	-	+	+	-	+	+	+	-
9	+	-	+	-	+	+	-	+	+	+	-	-
10	+	+	-	+	+	-	+	+	+	-	-	-
11	+	-	+	+	-	+	+	+	-	-	-	+
12	+	-	-	-	-	-	-	-	-	-	-	-

Half Fraction Hadamard Matrix

$(n, k) = (2t, 4t\text{-}2)$

Balanced Incomplete Block Design

$v = 2t\text{-}1$

$b = 4t\text{-}2$

$r = 2t\text{-}2$

$k = t\text{-}1$

$\mathrm{ave}(s^2) = n^2/(2n\text{-}3)$ proved to be $E(s^2)$-optimal!

Non-isomorphic class exists!

(14) 可計算設計 (Computer Experiment)

- Expensive simulation
- 當Monte Carlo不可行時
  如何設計Simulation?
- Latin Hypercube

Computer Experiments

- The problem
- Latin Hypercube (LHC)
- LHC with constraints
- Rotated Factorial Designs
- Uniform Design
- Summary and Comparisons

Goal

- Confirmation
- Sensitivity Analysis
- Empirical Model Building
- Optimization
- Model Validation
- High Dimension Integration

Irrelevant Issues

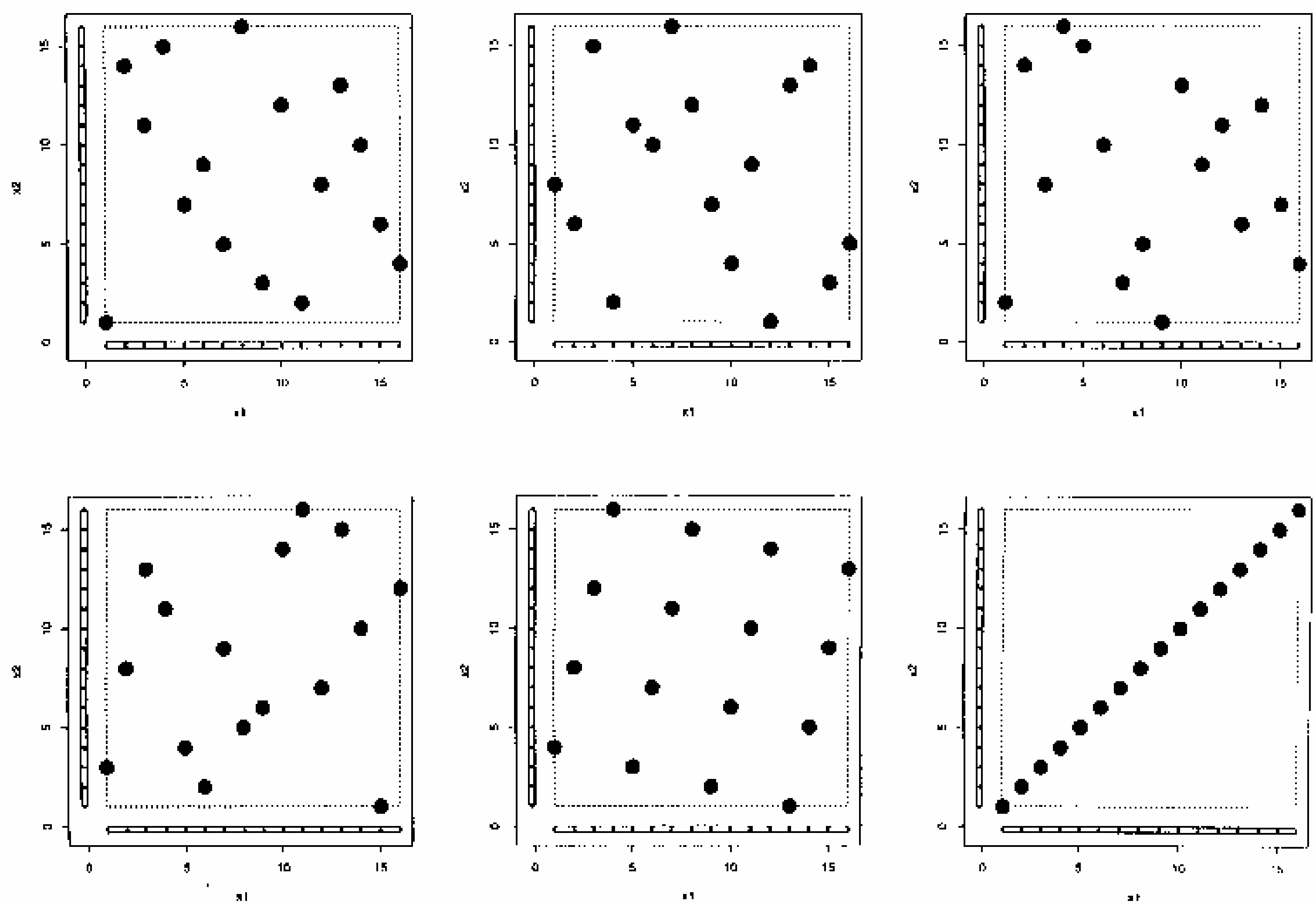
- Replicates
- Blocking
- Randomization

Question: How can a computer experiment be run in an efficient manner?

Current Approaches to Experimental Design

- Geometric (Frequentist) Designs
  - Full and Fractional Factorial Designs
  - Other Traditional Designs
  - Latin Hypercube Designs (McKay, Beckman, and Conover (1979))
- Computer-Generated (Bayesian) Designs
  - Maximin Distance Designs (Johnson, Moore, and Ylvisaker (1990))
- Combination Designs (Computer-Generated Geometric)
  - Maximin Latin Hypercube Designs (Morris and Mitchell (1992))
  - Orthogonal Array-based LHs (Tang (1993), Owen (1992))
  - Rotated Factorial Designs (Beattie and Lin, 1997)

Some Latin Hypercube Designs



A special class of LHC

$$
\begin{array}{cc}
x_1 & x_2 \\
\left[\begin{array}{cc}
1 & \tau_1 \\
2 & \tau_2 \\
3 & \tau_3 \\
4 & \tau_4 \\
. & . \\
. & . \\
. & . \\
16 & \tau_{16}
\end{array}\right]
\end{array}
$$

τ_i: permutation of {1, …, 16}

16!

n! for size *n* &

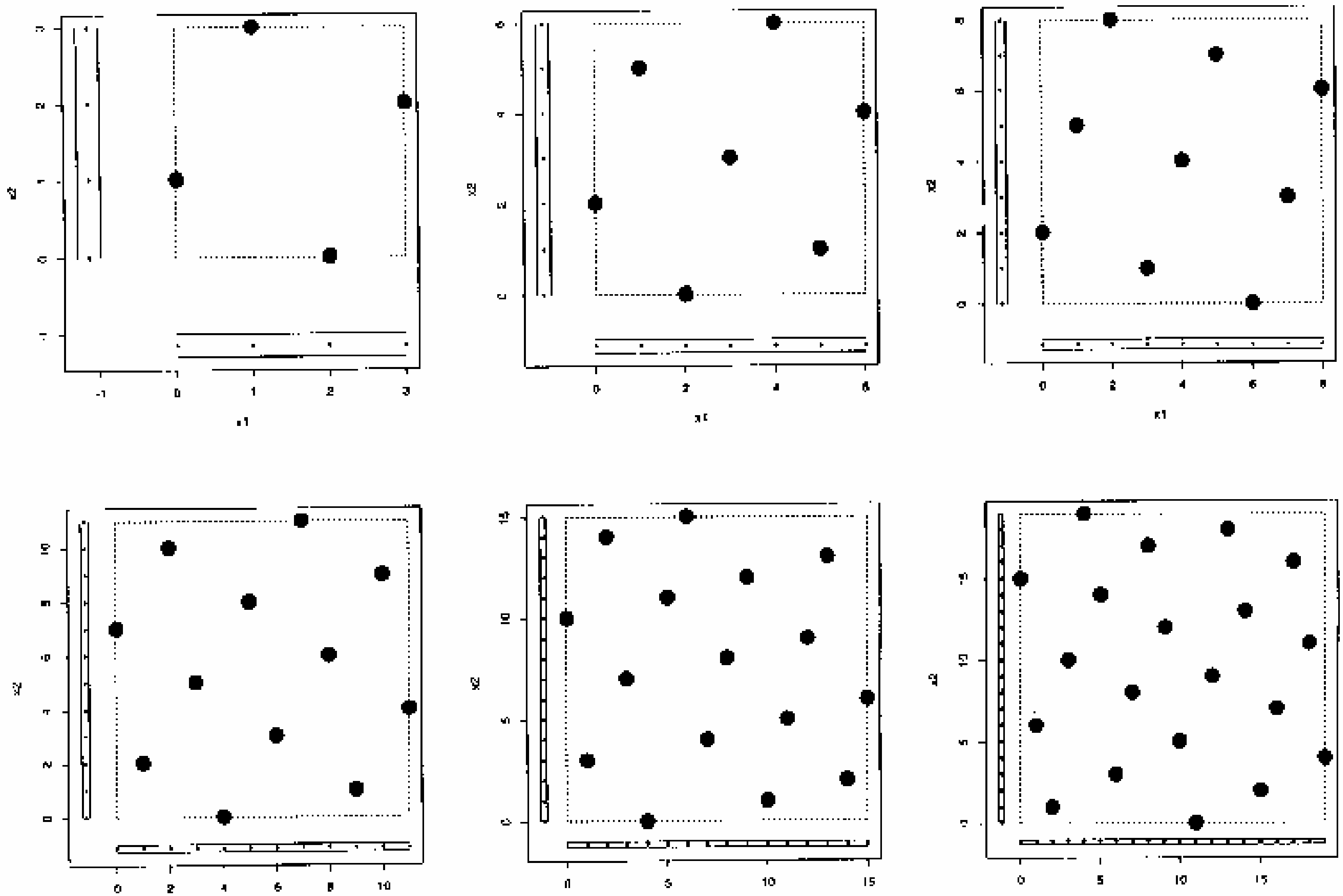
$(n!)^{d-1}$ for *d*-dim

Bayesian Designs

- Maximin Distance Designs, Johnson, Moore,and Ylvisaker (1990)
- Maximizes the Minimum Interpoint Distance (MID)
- Moves design points as far apart as possible in design space $MID = \min_{x_1, x_2 \in D} d(x_1, x_2)$
- D* is a Maximin Distance Design if

$$MID = \min_{x_1, x_2 \in D^*} d(x_1, x_2) = \max_{D} \min_{x_1, x_2 \in D} d(x_1, x_2)$$

Maximin Latin Hypercube Designs



Rotated Factorial Designs

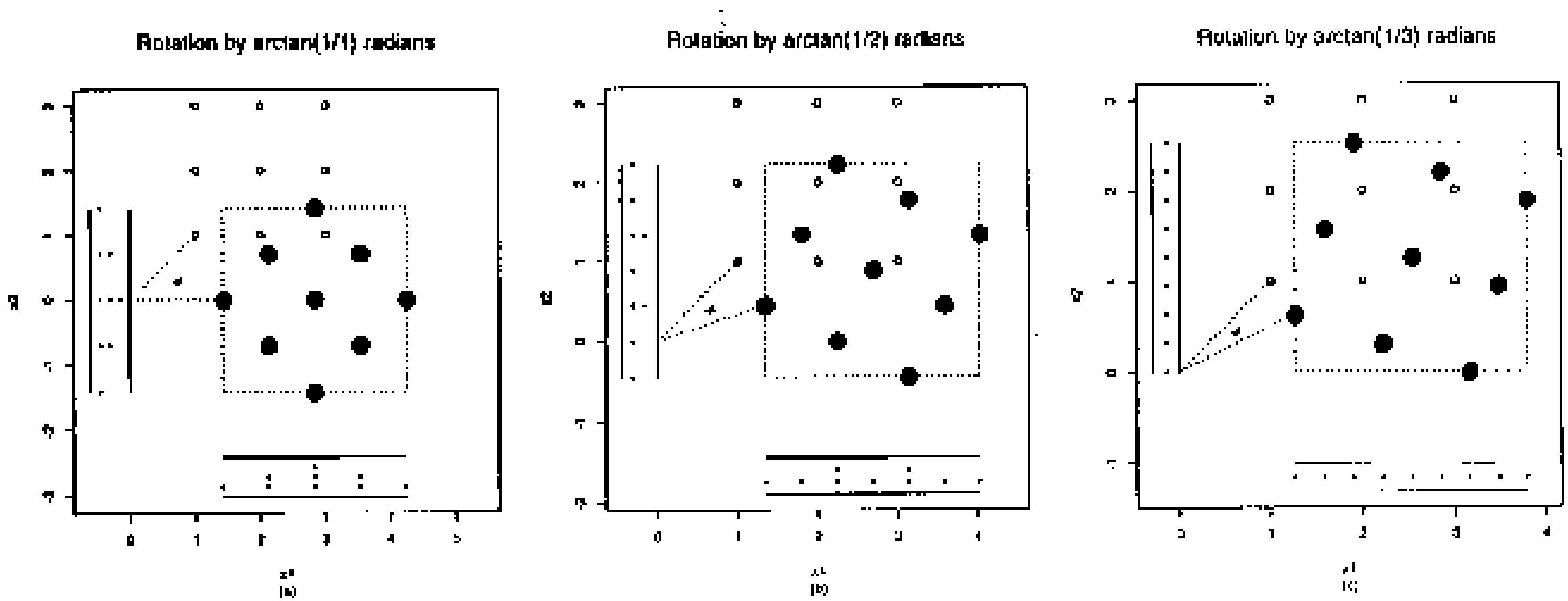
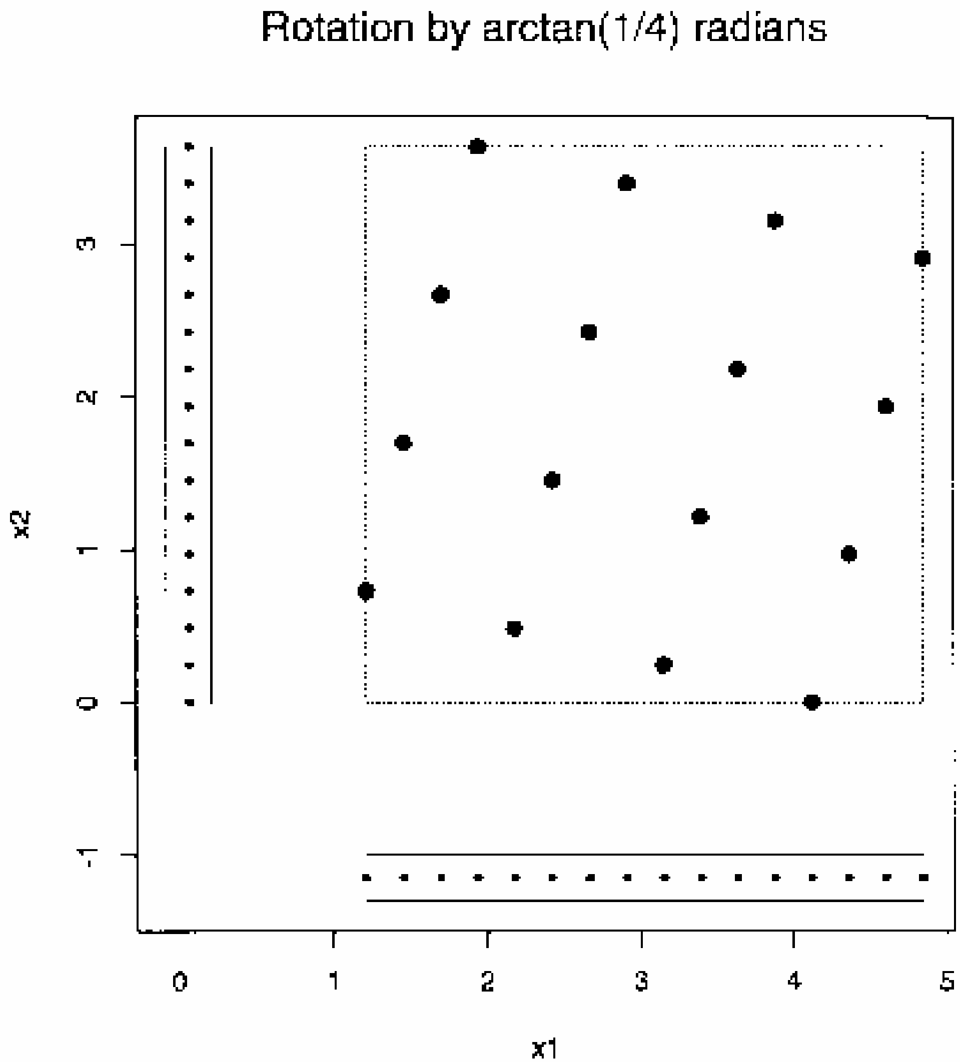


Figure 2: Three rotations of a standard 3^2 factorial design:

(a) $w = \tan^{-1}(1)$, (b) $w = \tan^{-1}(1/2)$, (c) $w = \tan^{-1}(1/3)$



- Rotation Theorem
- Orthogonality Theorem

Rotated Factorial Designs

- Computer experiments are gaining in popularity
  - main research area of the next 10 years
- Rotated factorial designs
  - good factorial design properties (orthogonality and structure)
  - good Latin hypercube properties (unique and equally-spaced projections)
  - easy to construct
  - comparable by Bayesian criteria
  - very suitable for computer experiments

(15) Micro Array Design

- Coming Soon...

Some Personal Views

- i and e
- Multiple response problem
- Classical design is as important as it was, but there are new problems requiring new designs
- Business world: experimental economics, supply chain design, electronic commerce, etc.
- Large data set problems (data mining, data warehouse, etc.): Design and Analysis
- Your conclusion is only as good as your assumption

Summary: DOE

- Model is known
  - Optimal design
  - Optimality Criteria (alphabetical optimality)
- Model is unknown

  (or is not completely known)
  - Robustness
  - Robustness Criterion
  - Representative Points

- 知識搬運業
- 知識宅急便
- 知識加工業
- 知識創造業