Primer on Analysis of Experimental Data and Design of Experiments

Lecture 8. Statistical Design of Experiments

Muhammad A. Alam
alam@purdue.edu

copyright 2018

Course Outline

$$\overline{y} = f(\overline{x}) \qquad \overline{x} = x_1, x_2, \dots x_n \qquad \overline{y} = y_1, y_2, \dots y_m$$

Lecture 1: Introduction
Lecture 2: Collecting and plotting $x_1, x_2, \dots x_n$
Lecture 3: Physical and empirical $f, F, df/dx, \dots$
Lecture 4: Model selection between $f_1, f_2, \dots$
Lecture 5: Model Selection: Cross-validation and Bootstrapping method
Lecture 6: Scaling theory with known f, $f(\overline{x}) = f(\overline{X})$
Lecture 7: Scaling theory with unknown f, $\overline{x} \rightarrow X$
Lecture 8: Design of experiments to determine $\overline{y}_{\text{max}} = f(\overline{x})$
Lecture 9: DOE and Anova
Lecture 10: Principle component analysis for classifying $\{y\}$.
Lecture 11: Machine learning … Statistical approach to learn f
Lecture 12: Interpretable ML: Physics-based machine learning $f = f_{\text{physics}} + \Delta f$
Lecture 13: Interpretable ML: System Equation Modeling
Lecture 14: Conclusions

Outline

1. Context and background
2. Single factor and full factorial method
3. Orthogonal vector analysis: Taguchi/Fisher model
4. Correlation in dependent parameters
5. Conclusions

Design of Experiments

- Set of guidelines for designing, conducting and analyzing experiments for system optimization
- Foundations of DOE were laid by Sir. R.A. Fisher in early 1920s (Analysis of Farm data, output as good as input).
- Concepts of Orthogonal arrays were introduced by Taguchi in 1950s. (Formalized the whole analysis)
- DOE has revolutionized quality control/reliability in all fields of science and technology (Toyota was one of the early adopter, most semiconductor companies use the method).

Problem definition

(A) Oxide Thickness

(B) Doping

(C) Anneal temp.

(D) Junction depth

(E) Gate Overlap

(F) Halo implant

(G) Supply voltage

MOSFET

Drain current

7 parameter optimization for a single objective function

Could be car, farm, airport

Philosophical shift with DOE

Before Fisher …

Experimentalist determine what experiments to do $\xrightarrow{\text{Results}}$ Statisticians/ Theorists/Expt collaborate to interpret results

After Fisher …

Statisticians/ Theorists/Expt plan what experiments to do $\xrightarrow{\text{Results}}$ Statisticians/ Theorists/Expt collaborate to interpret results

Output cannot be greater than input …..

Definition of terms

Factor	Level	Run/trial/replicate
Tox	1, 2, 3 nm	$(2\ \text{nm}, 10^{17}\ \text{cm}^{-3}, 4\ \mu\text{m})_{rep}$
Doping	10^{16}, 10^{17} cm^{-3}	
Lch	2, 3, 4 μm	

- 1 factor, 3 level, 4 replicate experiment
- 2 factor, 2 level, 3 replicate experiment
- 8 factor, 2 level, 1 replicate experiment

Puzzle Analogy: Many factors, 2 levels

Graeco-Latin Squares
Euler Squares

Land type ….A,B,C,D (Latin)
Fertilizer … a,b,c,d (Greek)

Aa	Bc	Cd	Db
Bb	Ad	Dc	Ca
Cc	Da	Ab	Bd
Dd	Cb	Aa	Ac

Balance and statistical content

Soduku

	2		7		4		9	
		5	6		9	2		
1								7
5			4		8			2
		2				6		
8			3		7			4
9								1
		8	1		2	3		
	4		9		5		8	

30 filled cells vs. 81 cells

3	2	6	7	8	4	1	9	5
7	8	5	6	1	9	2	4	3
1	9	4	2	5	3	8	6	7
5	1	7	4	6	8	9	3	2
4	3	2	5	9	1	6	7	8
8	6	9	3	2	7	5	1	4
9	5	3	8	7	6	4	2	1
6	7	8	1	4	2	3	5	9
2	4	1	9	3	5	7	8	6

Outline

1. Context and background
2. **Single factor and full factorial method**
3. Orthogonal vector analysis: Taguchi/Fisher model
4. Correlation in dependent parameters
5. Conclusions

7 Factor, 2 level: One factor at a time

	A	B	C	D	E	F	G	Output
Run 1	1	1	1	1	1	1	1	10
Run 2	2	1	1	1	1	1	1	15
Run 3	2	2	1	1	1	1	1	12
Run 4	2	1	2	1	1	1	1	9
Run 5	2	1	1	2	1	1	1	18
Run 6	**2**	**1**	**1**	**2**	**2**	**1**	**1**	**19**
Run 7	2	1	1	2	2	2	1	17
Run 8	2	1	1	2	2	1	2	13
Final	**2**	**1**	**1**	**2**	**2**	**1**	**1**	**19**

Simple, widely used, but non-optimum solutions

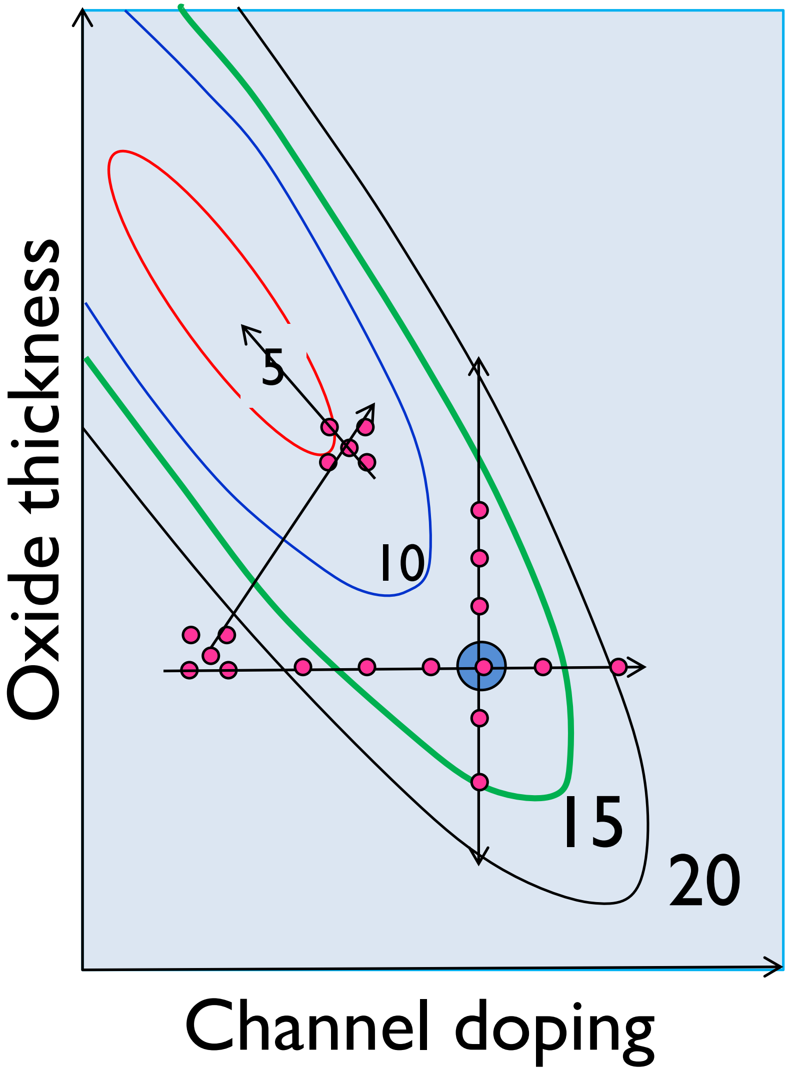
7 Factor, 2 Level: Full factorial analysis

				A_1				A_2			
				B_1		B_2		B_1		B_2	
				C_1	C_2	C_1	C_2	C_1	C_2	C1	C_2
D_1	E_1	F_1	G_1	R-1 (10)				R-2 (15)		R-3 (12)	
			G_2								
		F_2	G_1						R-4 (9)		
			G_2								
	E_2	F_1	G_1								
			G_2								
		F_2	G_1								
			G_2								
D_2	E_1	F_1	G_1					R-5 (18)			
			G_2								
		F_2	G_1								
			G_2								
	E_2	F_1	G_1					R-6 (19)			
			G_2					R-8 (13)			
		F_2	G_1					R-7 (17)			
			G_2								

Single parameter method is a fractional non-optimal factorial method: After A2 win, will never visit A1. After B2 loss, will never visit B2. Same for C2 Column, etc.

$$Level^{factor} = 2^7 = 128$$

The problem with one-at-a-time approach



Response surface
Orthogonal sampling

Outline

1. Context and background
2. Single factor and full factorial method
3. **Orthogonal vector analysis: Taguchi/Fisher model**
4. Correlation in dependent parameters
5. Conclusions

Uncorrelated main effect (forward/backward)

4 factors for simplicity

	Level 1	**Level 2**
A	0	**2**
B	0	**-6**
C	0	**4**
D	0	**-2**

	Level 1	**Level 2**	L2-L1
A	-2	**0**	**2**
B	2	**-4**	**-6**
C	-3	**1**	**4**
D	0	**-2**	**2**

Full factorial (Only A)

		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	0	**2**	**2**
	D_2	0	0	**2**	**2**
C_2	D_1	0	0	**2**	**2**
	D_2	0	0	**2**	**2**

Full factorial (A, B, C and D)

		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	-6	2	-4
	D_2	-2	-8	0	-6
C_2	D_1	4	-2	6	0
	D_2	2	-4	4	-2

Full factorial (A and B)

		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	-6	2	-4
	D_2	0	-6	2	-4
C_2	D_1	0	-6	2	-4
	D_2	0	-6	2	-4

Full factorial (A, B and C)

		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	-6	2	-4
	D_2	0	-6	2	-4
C_2	D_1	4	-2	6	0
	D_2	4	-2	6	0

Taguchi orthogonal array (L8 array)

	A	B	C	D	E	F	G
R-1	1	1	1	1	1	1	1
R-2	1	1	1	2	2	2	2
R-3	1	2	2	1	1	2	2
R-4	1	2	2	2	2	1	1
R-5	2	1	2	1	2	1	2
R-6	2	1	2	2	1	2	1
R-7	2	2	1	1	2	2	1
R-8	2	2	1	2	1	1	2

1) Check to see that for every factor, e.g. A, the rest of factors are fully randomized, e.g. every column sums to same number.
2) Does it remind you of Soduku?
3) For smaller system (4 factors, 2 levels), choose the first four columns, ignore the remaining 3 – still need 8 experiments. For other systems, see …

http://www.freequality.org/sites/www_freequality_org/documents/tools/Tagarray_files/tamatrix.htm

Orthogonal measurements (uncorrelated)

				A_1				A_2			
				B_1		B_2		B_1		B_2	
				C_1	C_2	C_1	C_2	C_1	C_2	C_1	C_2
D_1	E_1	F_1	G_1	R-1							
			G_2								
		F_2	G_1								
			G_2				R-3				
	E_2	F_1	G_1								
			G_2						R-5		
		F_2	G_1							R-7	
			G_2								
D_2	E_1	F_1	G_1								
			G_2							R-8	
		F_2	G_1						R-6		
			G_2								
	E_2	F_1	G_1				R-4				
			G_2								
		F_2	G_1								
			G_2	R-2							

$Y_{A1}=(R1+R3+R2+R4)/4$ $Y_{C2}=(R3+R4+R5+R6)/4$

If the system optimizes for (A1 B2 C2 D2 E2 F1 G2)

$Y=Y_M+(Y_{A1}-Y_M)+(Y_{B2}-Y_M)+(Y_{C2}-Y_M)+(Y_{D2}-Y_M)+ \ldots (Y_{G2}-Y_M)$

$Y_M=(Y_{A1}+Y_{A2}+Y_{B1}+Y_{B2}+\ldots.Y_{G1}+Y_{G2})/14.$

Outline

1. Context and background
2. Single factor and full factorial method
3. Orthogonal vector analysis: Taguchi/Fisher model
4. **Correlation in dependent parameters**
5. Conclusions

Correlated effect & level factor

4 parameters for simplicity …

	Level 1	Level 2
A	0	2
B	0	+6
C	0	4
D	0	-2
B if A	**0**	**-6**

Uncorrelated

		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	-6	2	-4
	D_2	-2	-8	0	-6
C_2	D_1	4	-2	6	0
	D_2	2	-4	4	-2

Correlated Table

		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	6	2	-4
	D_2	-2	4	0	-6
C_2	D_1	4	10	6	0
	D_2	2	8	4	-2

A2+ **B2(GivenA2)** +C2+D2

e.g. $A_1B_1=(0-2+4+2)/4=1$

Are the variables correlated?

Define Level factors $A_mB_m=(A\ B)_1$

Define $A_mB_n=(A\ B)_2$

Correlated effect & level factor

Pair Corr ...... $Corr_{AB} = \sum_{i,j=1,2} A_i B_j (-1)^{(i+j)}$

Third order ... $Corr_{ABC} = \sum_{i,j,k=1,2} A_i B_j C_k (-1)^{(i+j+k)}$

Fourth order ... $Corr_{ABCD} = \sum_{i,j,k,p=1,2} A_i B_j C_k D_p (-1)^{(i+j+k+p)}$

		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	6	2	-4
	D_2	-2	4	0	-6
C_2	D_1	4	10	6	0
	D_2	2	8	4	-2

$$Corr_{AB} = \sum_{i,j} A_i B_j (-1)^{(i+j)} = A_1B_1 - A_1B_2 - A_2B_1 + A_2B_2 = -12$$

$$Corr_{AC} = \sum_{i,j} A_i C_j (-1)^{(i+j)} = 0$$

$$Corr_{AD} = \sum_{i,j} A_i D_j (-1)^{(i+j)} = 0$$

$$Corr_{BC} = \sum_{i,j} B_i C_j (-1)^{(i+j)} = 0$$

$$Corr_{BD} = \sum_{i,j} B_i D_j (-1)^{(i+j)} = 0$$

$$Corr_{CD} = \sum_{i,j} C_i D_j (-1)^{(i+j)} = 0$$

Example of pair correlation:

$A_1B_1 = 1$, $A_1B_2 = 7$
$A_2B_1 = 3$, $A_2B_2 = -3$

A is correlated to B … There are no other pair correlation

Correlated effect & level factor

		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	6	2	-4
	D_2	-2	4	0	-6
C_2	D_1	4	10	6	0
	D_2	2	8	4	-2

Pair Corr …… $Corr_{AB} = \sum_{i,j=1,2} A_i B_j (-1)^{(i+j)}$

Third order … $Corr_{ABC} = \sum_{i,j,k=1,2} A_i B_j C_k (-1)^{(i+j+k)}$

Fourth order … $Corr_{ABCD} = \sum_{i,j,k,p=1,2} A_i B_j C_k D_p (-1)^{(i+j+k+p)}$

Third order correlation:

$A_2B_2\,C_2$ =-1, $A_2B_2\,C_1$ =-5, etc.

$$Corr_{ABC} = \sum_{i,j,k=1,2} A_i B_j C_k (-1)^{(i+j+k)} = A_2B_2C_2 - A_1B_2C_2....$$

$$= -1 - (+5) + (1) - (-5) + (+3) - (+9) - (-1) + (+5) = 0$$

$$Corr_{ABCD} = \sum_{i,j,k,p=1,2} A_i B_j C_k D_p (-1)^{(i+j+k)} = A_2B_2C_2D_2 - A_1B_2C_2D_2.... = 0$$

No third or fourth order correlation ….

How to fix for correlation

4 parameters for simplicity …

	Level 1	Level 2
A	0	2
B	0	+6
C	0	4
D	0	-2
B given A	**0**	**-6**

		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	6	2	-4
	D_2	-2	4	0	-6
C_2	D_1	4	10	6	0
	D_2	2	8	4	-2

$$Corrected = B_2 - B_1 - \frac{\sum_{A,C,D} \text{all B interaction}}{2}$$

$$= 2 - 2 - (-12/2) = +6$$

$$Corrected = A_2 - A_1 - \frac{\sum_{B,C,D} \text{all A interaction}}{2}$$

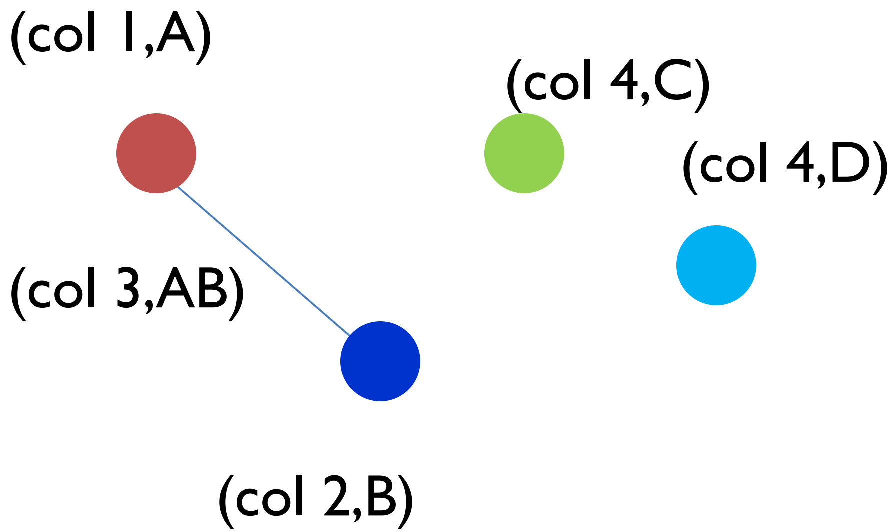
$$= 0 - 4 - (-12/2) = +2$$

$$Corr_{AB} = \sum_{i,j} A_i B_j (-1)^{(i+j)} = -12$$

Only (AB) pair correlation found, no other correlation ….

Aside: correlation linear graph

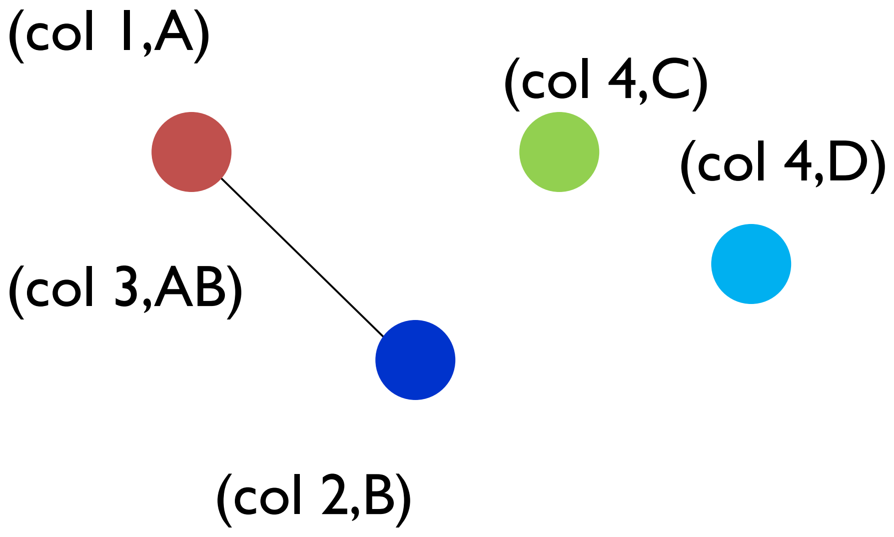
		A_1		A_2	
		B_1	B_2	B_1	B_2
C_1	D_1	0	6	2	-4
	D_2	-2	4	0	-6
C_2	D_1	4	10	6	0
	D_2	2	8	4	-2



Only (AB) pair correlation found, no other correlation ….

Main effect and interactions

	A	B	AxB	C	D
R-1	1	1	1	1	1
R-2	1	1	1	2	2
R-3	1	2	2	1	1
R-4	1	2	2	2	2
R-5	2	1	2	1	2
R-6	2	1	2	2	1
R-7	2	2	1	1	2
R-8	2	2	1	2	1



(AB)=1 means $A_1B_1=1$, $(AB)_2 = A_1B_2$
Expanded basis set and orthogonal vector set ….

(AB) is a dummy column, without it the C and D would have different arrangements …

Still need L8 array (4-7), other two-level arrays L4 (1-3) and L12(8-11)

Taguchi table: Continued

$L_4(2^3)$

Run	Columns		
	1	2	3
1	1	1	1
2	1	2	2
3	2	1	2
4	2	2	1

$L_8(2^7)$

Run	Columns						
	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

$L_{12}(2^{11})$

Run	Columns										
	1	2	3	4	5	6	7	8	9	10	11
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	2	2	2	2	2	2
3	1	1	2	2	2	1	1	1	2	2	2
4	1	2	1	2	2	1	2	2	1	1	2
5	1	2	2	1	2	2	1	2	1	2	1
6	1	2	2	2	1	2	2	1	2	1	1
7	2	1	2	2	1	1	2	2	1	2	1
8	2	1	2	1	2	2	2	1	1	1	2
9	2	1	1	2	2	2	1	2	2	1	1
10	2	2	2	1	1	1	1	2	2	1	2
11	2	2	1	2	1	2	1	1	1	2	2
12	2	2	1	1	2	1	2	1	2	2	1

$L_N(S^M)$

M … Factors
S … Levels
N … experiment > DOF = 1 + M(S-1)

Graeco-Latin Squares and Taguchi Tables

Land type ….A,B,C,D (Latin)
Fertilizer … a,b,c,d (Greek)

2 factor (M=2) & 4 levels (S=4)

Full factorial experiment
Involves $4^2 = 16$ experiment

$L_N(S^M)$

Choose a table with N=16
that has S=4 levels

$N = DOF = 1 + M(S-1) = 1 + 3 \times M$

Factors

Levels

1	1	1	1	1
1	2	2	2	2
1	3	3	3	3
1	4	4	4	4
2	1	2	3	4
2	2	1	4	3
2	3	4	1	2
2	4	3	2	1
3	1	3	4	2
3	2	4	3	1
3	3	1	2	4
3	4	2	1	3
4	1	4	2	3
4	2	3	1	4
4	3	2	4	1
4	4	1	3	2

A=1, B=2, C=3, D=4
a=1, b=2, c=3, d=4

Randomization

Aa	Bc	Cd	Db
Bb	Ad	Dc	Ca
Cc	Da	Ab	Bd
Dd	Cb	Aa	Ac

Web Design: 4 Factor, 5 level

4 factors (font, color, background, foreground) and 5 levels of each

$L_N(S^M)$

M … Factors
S … Levels
N … experiment
> DOF = 1 + M(S-1)

5^4 experiments
$DOF = 1 + 4 \times 4 = 17$
N=25
Partial factorial Design

Factors

Levels

X1	X2	X3	X4	X5	X6
-	-	-	-	-	-
1	1	1	1	1	1
1	2	2	2	2	2
1	3	3	3	3	3
1	4	4	4	4	4
1	5	5	5	5	5
2	1	2	3	4	5
2	2	3	4	5	1
2	3	4	5	1	2
2	4	5	1	2	3
2	5	1	2	3	4
3	1	3	5	2	4
3	2	4	1	3	5
3	3	5	2	4	1
3	4	1	3	5	2
3	5	2	4	1	3
4	1	4	2	5	3
4	2	5	3	1	4
4	3	1	4	2	5
4	4	2	5	3	1
4	5	3	1	4	2
5	1	5	4	3	2
5	2	1	5	4	3
5	3	2	1	5	4
5	4	3	2	1	5
5	5	4	3	2	1

Randomization

fjords	jawbox	phlegm	qiviut	zincky
zincky	fjords	jawbox	phlegm	qiviut
qiviut	zincky	fjords	jawbox	phlegm
phlegm	qiviut	zincky	fjords	jawbox
jawbox	phlegm	qiviut	zincky	fjords

https://www.mne.psu.edu/cimbala/me345/Lectures/Taguchi_orthogonal_arrays.pdf

https://www.york.ac.uk/depts/maths/tables/orthogonal.htm

Conclusions

1. Design of experiment is a powerful technique universally used in industry and in large scale field trials.

2. Taguchi/Fisher methods replace the older one-factor-at-a-time experiments with experiments based on orthogonal arrays; In this approach, only the effect of main factors remain; others are cancelled.

3. Understanding and analyzing correlation is important in design of experiments. Unless the correlation is well understood and incorporated through dummy variables, the analysis may lead to faulty conclusions.

References

I understood the essence of the problem of randomization from the book "Nets, Puzzle, and Postman", by Peter Higgins, Oxford University Press.

A wonderful set of lectures by Stuart Hunter is available in youtube, see …

http://www.youtube.com/watch?v=AVUAt0Qly60&list=PLWQ-BDMTHPQVH3IUGF7EM_3XHJWFD2EIP

For DOE based on Taguchi method, I liked Lloyd W Condra, "Reliability Improvement with design of experiments", Marcel Dekker Inc., 1993. Another good book is by Ranjith Roy, "A primer on the Taguchi Method", Van Nostrand Reinhold International Co. Ltd., 1990. Some of the examples are taken fromAT&T, "Statistical Quality Control Handbook".

Also, see the lectures on DOE by Hunter
http://www.youtube.com/watch?v=NoVIRAq0Uxs
http://www.youtube.com/watch?v=hTviHGsI5ag

The classical AVONA method is discussed in great detail in Chapter 13 and 14 of "Applied Statstics and Proability for Engineers, 3rd Edision, D.C. Montgomery and G. C. Runger, Wiley, 2003.

Hunter's lectures on AVONA is also very enjoybale
http://www.youtube.com/watch?v=k3n9iSB6Cns
http://www.youtube.com/watch?v=F05zZL3uyRo

A slightly different approach that also reduces the number of experiments greatly is based on the response surface approach. It uses Newton-like algorithm to find the peaks/valleys of the response surface, see R. H. Myers and D.C. Montgomery, "Response Surface Methodology", Wiley Interscience, 2002. This book discusses design of experiment in great detail.

For general reference see

Joan Fisher Box, "R. A. Fisher and the Design of Experiments, 1922-1926", *The American Statistician*, vol. 34, no. 1, pp. 1-7, Feb. 1980.

F. Yates, "Sir Ronald Fisher and the Design of Experiments", *Biometrics*, vol. 20, no. 2, In Memoriam: Ronald Aylmer Fisher, 1890-1962., pp. 307-321, (Jun. 1964.

Review Questions

1. What role did Fisher play in developing the design of experiment?
2. If you have 3 variables at two levels, what Taguchi array would you choose?
3. How does one find correlation among variables in Full factorial method?
4. What is the role of linear graphs in Taguchi method?
5. In what ways Fisher philosophy of change the ways experiments are done? Is there a down side of such analysis?
6. What is dummy variable? What does dummy variable to do in DOE?
7. Can you have 3rd or higher order correlation, if you do not have second order correlation?