Part II

DESIGN OF EXPERIMENTS

Box, Hunter & Hunter ~~Chapters~~ used in these notes are!

→ Chapter 15 (2^{nd} order CCD)

→ Chapter 9/10/11/12 (2^{x} factorials)

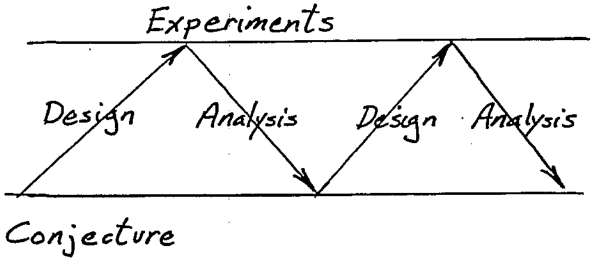
→ Chapter 13 Examples/Confounding

→ Chap 15 Response surfaces

Copied: Chap 12;

DESIGN OF EXPERIMENTS

Iterative Nature of Experimentation.



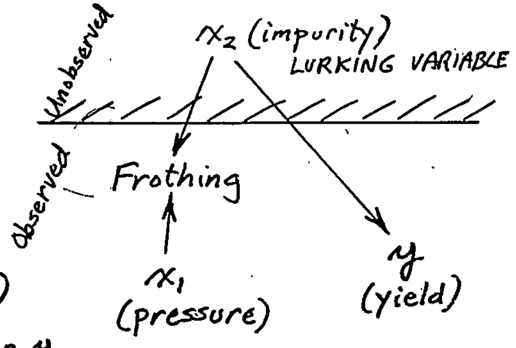
Why Design ?

- Reduces amount of experimentation needed.
- Ensures adequate range of variation in all x's
- Minimizes confounding of effects
- Enables one to infer cause + effect

Example : Chemical Process (BH^2, pg. ~~487~~ 493)

Observed that undesirable frothing in reactor could be reduced by increasing Pressure (x_1)

Operating Procedure:
Increase x_1 when frothing.



Truth (unknown):

(i) High impurity (x_2) causes frothing

(ii) High x_2 lowers yield (y)

(iii) Press(x_1) has no effect on y.

WHY DESIGN ?

1. Ensure adequate variability in all key variables.

2. Ensure identifiability of all important effects & interactions (unconfounded) separate temperature & pressure effects because they tend to move together

③. Maximize the information obtained in fewest number of experiments . Costly ito time/materials

4. Distinguish between causality and correlation

1. Adequate Variability

- Variable x may have very important effect on process performance.

- But, if variation in it is small relative to noise level, then may
  - Accept H_0: effect of $x = 0$
  - -obtain confidence interval on effect of x to include zero.

Data set specific – need to excite variable enough

- This does not mean that effect of x is not important - only that it isn't large enough in this particular data set to detect significance.

- Design of Experiments provides a form of guarantee that accepting H_0 implies that the effect is not important.

If the confidence interval includes zero, after a proper design, over the range of interest ⟹ that variable is NOT important – in that range

2. Identifiability of Effects

- DOE helps ensure that all important main effects and interactions can be identified - minimizes confounding.

DOE forces independant variation

- Our bad experimental habits arise from the nature of university laboratories:

-These labs aimed at demonstrating theoretical principles, not at building models, exploring for unknown effects, or optimizing processes.

Industry wants
- build models
- optimize
- explore unknowns

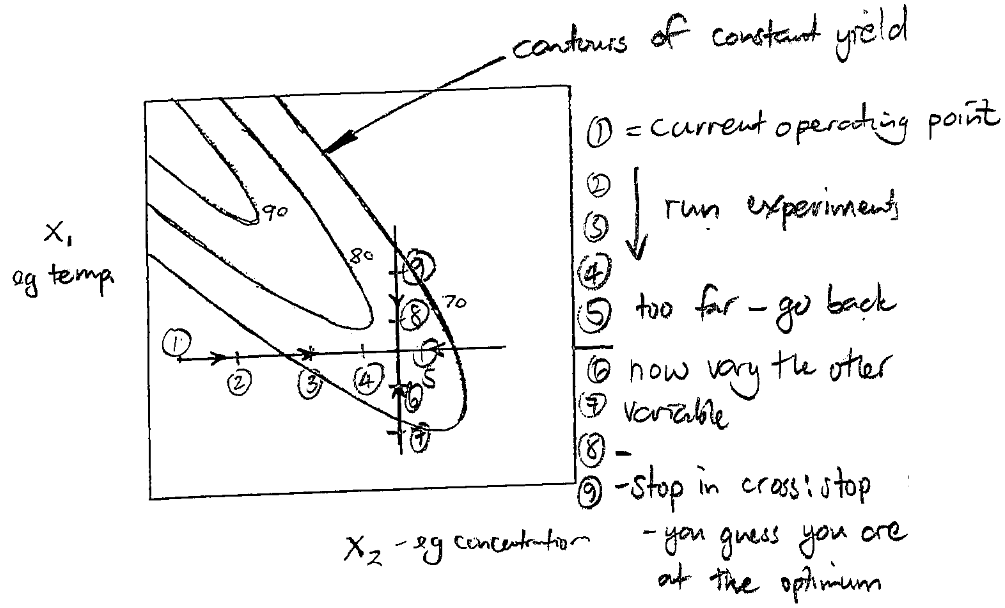
Eg. Demonstrate the effect of Temp. on reaction equilibrium
-Change Temp. holding all other variables constant !
Change a single variable at a time
-Experiments focus on changing one variable at a time to verify the particular principle being studied.

- COST approach is not good when searching for effects, building models, or optimization processes

11

Optimization

COST approach -- Changing One Single variable at a Time



Design of Experiments - Efficient ways of changing many variables at once

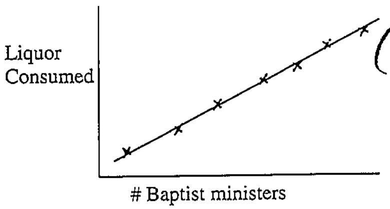
3. Maximizing the information obtained in the minimum number of experiments.

Example of industrial screening experiment (ICI, UK)

- Problem: In a new plant the cycle time in the filtration section was unacceptably long.
- Need to de-bottleneck
- Many factors suggested that might be responsible
- How to screen out important ones in fewest runs possible?

4. Distinguishing Between Correlation and Causation

- Data from Australia over many years on
  - Number of Baptist Ministers vs Amount of Liquor consumed

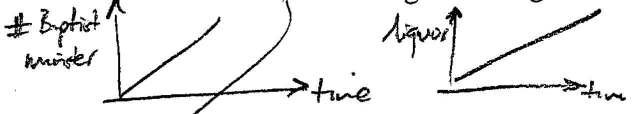


- Strong Correlation !
- Causal effect ?

→ so time is an indirect lurking variable

Cannot get a cause and effect relationship from it

- Smoking and Lung Cancer

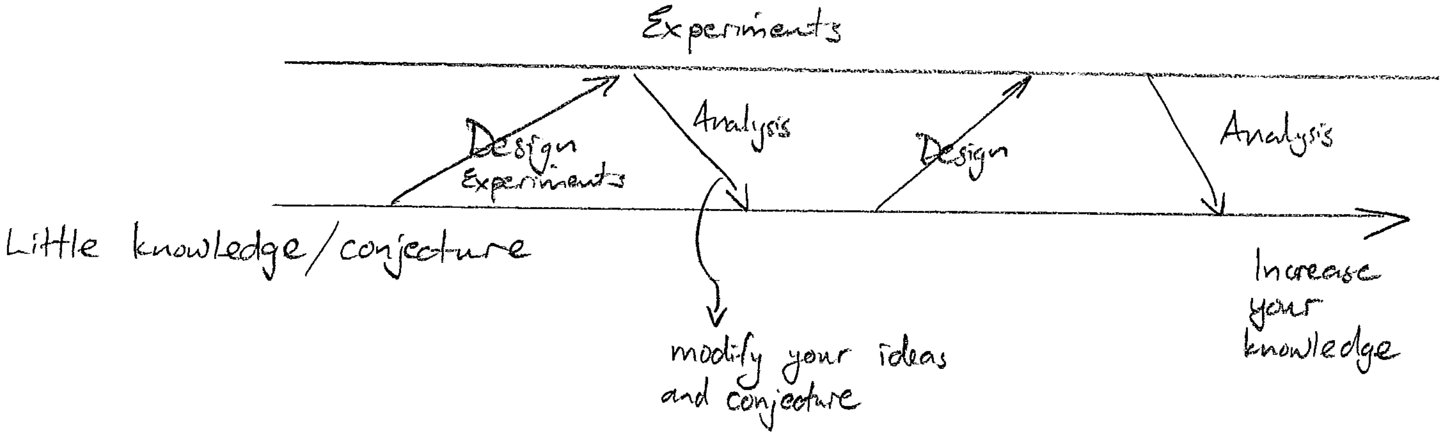


=> you can get the above correlation

  - Global warming and use of carbon fuels

→ have to use random samp' groups.

<u>Design of Experiments</u>



You have to do iterative/sequential design & analysis – no point to do 100 experiments based on little knowledge and then find little new knowledge.

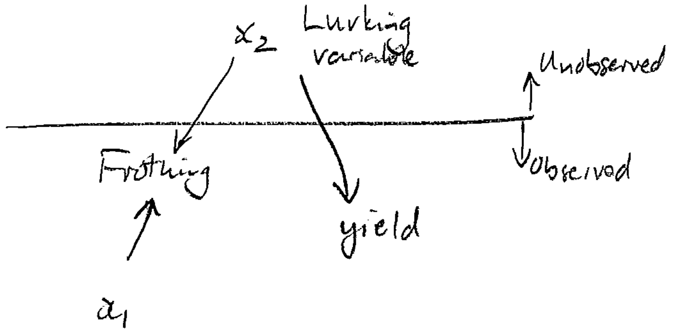
DOE puts information into data to be analysed, badly designed experiments cannot use very powerful methods to extract information from the data. A good design leads to a simple analysis

See notes on <u>Why Design</u> – go inbetween here

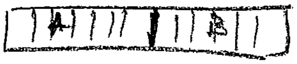
Correlation vs Causation: Box, Hunter & Hunter p487

Increase pressure reduces frother. So increase x_1 when frothing

Truth: High impurity x_2 cause frothing
High x_2 also lowers yield
x_1 has no effect on y
But x_1 and y are highly correlated
through ~~x_2~~ operating procedure.



Blocking & Randomization



Example: Two treatments on rubber: A and B
cut a section and run A
cut a section and run B

A	B
y_{A1}	y_{B1}
y_{A2}	y_{B2}
↓	↓
y_{AN}	y_{BN}
$\bar{y}_A$	$\bar{y}_B$ ← means

Test $H_0: \mu_A - \mu_B = 0$ hypothesis test

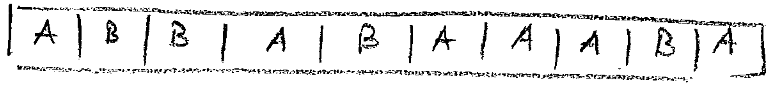
$H_1: \mu_A - \mu_B \neq 0$

Use a t-test (confidence interval approach)

$t_{\alpha, N_A + N_B - 2}$ and it uses a pooled variance

Potential problem: properties may vary along the rubber length such as thickness

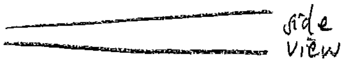
→ We should randomize the treatment

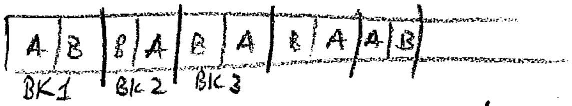
→ 

→ removes ~~effect~~ of unknown variation and other lurking variables.

→ called a randomized design

If you had ABCD then to compare $y_A - y_B$, $y_B - y_C$ etc use ANOVA Table
ANOVA with 2 variables is exactly same as t-test with 2 variables

If we know variation along the strip is gradual 

then we block two pieces 

but randomize in the blocks; reason: pieces next to each other are very similar

The relative difference between A and B in each block should be approximately the same

Block	A	B	Difference =(A-B)
BK1	y_{A1}	y_{B1}	$d_1 = y_{A1} - y_{B1}$
BK 2	y_{A2}	y_{B2}	$d_2 =$
BK 3	y_{A3}	y_{B3}	d_3
↓			$\bar{d} = \Sigma d_i / n$

PAIRED T-TEST

Now $\bar{d}$ is better measure of $\mu_A - \mu_B$

$H_0: \mathcal{E}(d) = 0 \quad H_1\ \mathcal{E}(d) \neq 0$

$S_{\bar{d}}^2$ is used in t-test $t_{\alpha, n-1}$ used

$S_{\bar{d}}^2 = S_d^2 / n$

Beer Example

- Improved by letting each one taste both beers
- Because tastes are individual, one has to use blocking
- Then look at differences — removes the bias.

example

A	10	6	3	5	$\bar{y}_A = 6$	$\bar{y}_A - \bar{y}_B = 1.0$
B	8	5	2	5	$\bar{y}_B = 5$	
diff	2	1	1	0	$\bar{y}_d = 1.0$	

But the variance in the data is large across A and B not so across the paired test. The standard error in comparing $\bar{y}_A - \bar{y}_B$ is larger than that of $\bar{y}_d$.

Difference in sensitivity ⟶

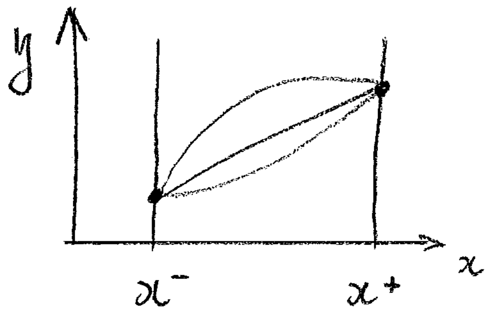
Designs for Empirical Studies

$\underbrace{\text{M\&R: Chap 12}}_{\text{ANOVA based}}$ $\qquad$ $\underbrace{\text{BHH Chap 9-12}}_{\text{Regression based.}}$

- Screening studies — which variables affect the response
- Empirical Model — approximate true $f(x_1\ x_2\ x_3) = y$ model unknown.
  - designs for linear models
  - designs for higher order models

2^k factorial design

— assume linear effects for now



2 experiments run at x^- and x^+ if linear model

2 level design in 1 variable

2^k levels design in k variables

$y = \hat{\beta}_0 + \hat{\beta}_1 x$ $\quad \hat{\beta}_1$ = effect of changing x by one unit

= main effect of x, a linear effect

include centre points to test for lack of fit.

Concepts in Design of Experiments

Randomization + Blocking

Simple Comparative experiment :

- Effect of two treatments on strength of rubbe

(i)

|A|A|A|A| - - - - |A|B|B|B| - - - - - |B|

A	B
y_{A1}	y_{B1}
⋮	
y_{An_A}	y_{Bn_B}
$\bar{y}_A$	$\bar{y}_B$

Hypothesis $n_A + n_B$ samples

$H_0 : \mu_A - \mu_B = 0$ How many DOF variance pooled?

$H_1 : \mu_A - \mu_B \neq 0$

Test : $\dfrac{\bar{y}_A - \bar{y}_B}{s_p^2\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} \sim t_{n_A+n_B-2}$ PTO

Problems with ~~this~~ ? → page over for calculation

(ii) What if strip of rubber had variations (eg. thickne along its length ?

Then $\bar{y}_A - \bar{y}_B$ might just be reflecting this diff.

ie. thickness = lurking variable

One solution → Randomize allocation of rubber pieces to treatments (A+B)

eg. flipping a coin

|A|B|B|A|B|A|B|B|B|A| - - - - - |A|B|B|

∴ Randomly allocates any lurking variable effect to A and B

- Ensures validity of hypothesis test.
- Randomized Design Problems ?

A more sensitive way to analyse the data is to pair the ~~test~~

⟹ compare within blocks

(iii) Suppose we expect variation in rubber to be progressive along length of the strip !

Then two adjacent pieces will be much more similar than two distant ones.

∴ BLOCK into pairs of adjacent pieces

Assign treatments (A, B) RANDOMLY within block

- Randomized Block Design

|B|A|A|B|A|B| ———— |B|A|

Bl_1 Bl_2 Bl_3 … Bl_n

Only compare within blocks

How many DOF associated with $\bar{d}$: n observations

−1 for its mean

ie: n−1 DOF

Block	A	B	difference $d = y_A - y_B$
Bl_1	y_{A1}	y_{B1}	$d_1 = y_{A1} - y_{B1}$
Bl_2	y_{A2}	y_{B2}	d_2
⋮			
Bl_n	y_{An}	y_{Bn}	d_n
			$\bar{d}$

} n measure of this difference

Blocking removes effect of possible uncontrolled variations along length of strip.

∴ $\bar{d}$ better measure of $\mu_A - \mu_B$ than $\bar{y}_A - \bar{y}_B$

$H_0 : E(d) = 0$ $\quad S_d^2 = \sum_{i=1}^{n}(d_i - \bar{d})^2 / n-1$

Paired t-test : $\dfrac{\bar{d} - 0}{s_{\bar{d}}} \sim t_{n-1}$

$s_{\bar{d}}^2 = \dfrac{s_d^2}{n}$

Sum of squares:

$$\underbrace{\sum_{i=1}^{N_A} \left(y_{Ai} - \bar{y}_A\right)^2}_{} + \underbrace{\sum_{i=1}^{n_B} \left(y_{Bi} - \bar{y}_B\right)^2}_{} = J$$

$$\text{DOF:} \quad N_A - 1 \quad + \quad N_B - 1 = N_A + N_B - 2$$

$$s_p^2 = \frac{J}{N_A + N_B - 2}$$

Designs for Empirical Studies

Montgomery + Runger : Chapt. 12
BH^2 : Chapt. 9, 10, 11, 12

1.) Screening Studies : Discovering which of a large number of variables affect response.

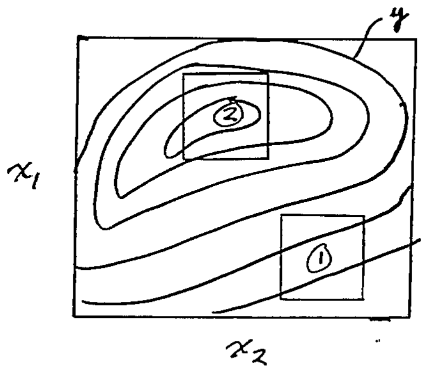
2.) Empirical model building studies

$$y = f(x_1, x_2 \cdots, x_k)$$

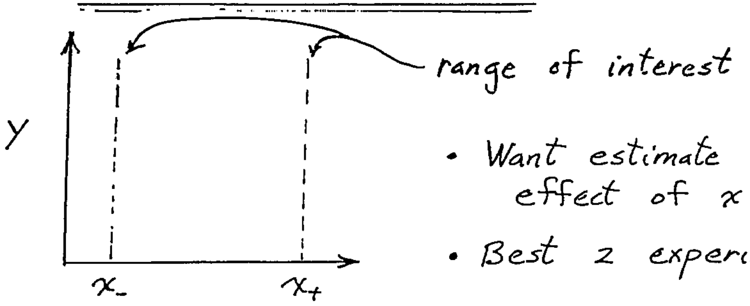
True model unknown. Use approx. models.

Region ① : Linear model OK

Region ② : Need model quadratic in x's



2^k FACTORIAL DESIGNS



- Want estimate of linear effect of x on y.
- Best 2 experiments?

Effect on y of changing x from x_- to x_+ is $(y_2 - y_1)$ ← Main effect of x

If fit LS model : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$\hat{\beta}_1$ = effect of changing x by one unit

- Linear effect only (two level experiment)

2^2 Factorial Design

2 independent variables :

	Range
temperature (T) :	160° C – 180° C
concentration (C) :	20% – 40%

Study effect of T + C on yield y.

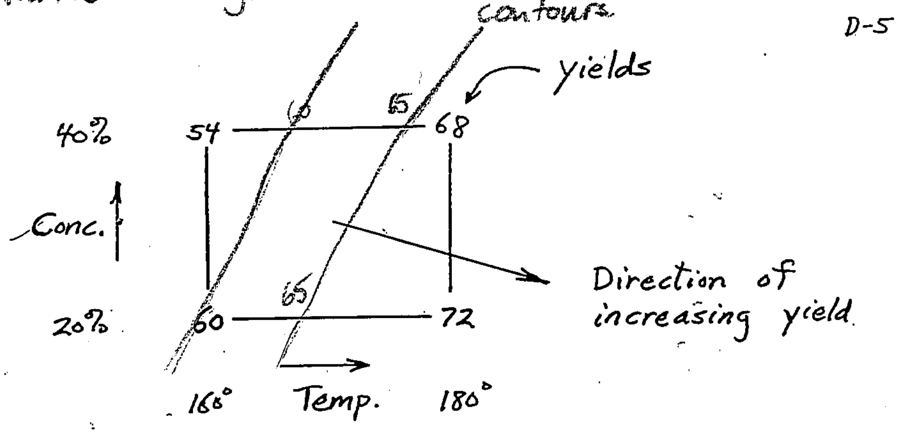
Design : 2^2 factorial in $2^2 = 4$ runs

two variables

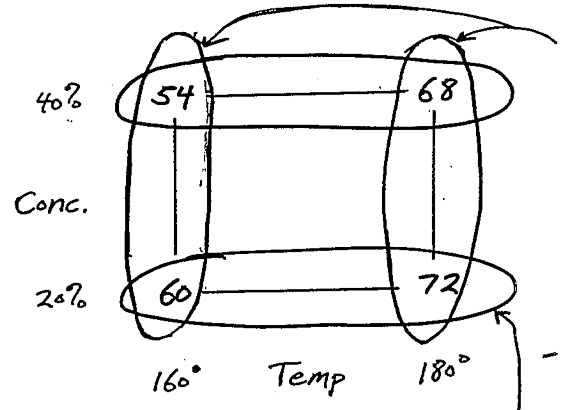
two levels

All possible combinations of 2 levels of 2 variables

If you have a good design you need to do very little further analysis



Main Effects of T + C



where T is constant

Two measures of effect of C

54 − 60 = −6

68 − 72 = −4

Avg. $\boxed{= -5}$

Main effect of C

−5% yield change by increasing concentration from 20% to 40%

Two measures of main effect of T

68 − 54 = 14

72 − 60 = 12

Avg $\boxed{= 13}$ %/20°C change in T

Main effect of T

If you calculate β_i: $\beta_0 = 63.5$ $\overset{\text{Temp}}{\beta_1} = 6.5$ $\beta_2 = -2.5$

$\beta_{12} = 0.5$ ~small (linear)

Interaction between T + C

Do variables T + C act independently on y?

Is effect of T same at both levels of C?

" " " C " " " " T?

If T depends on C or C depends on T

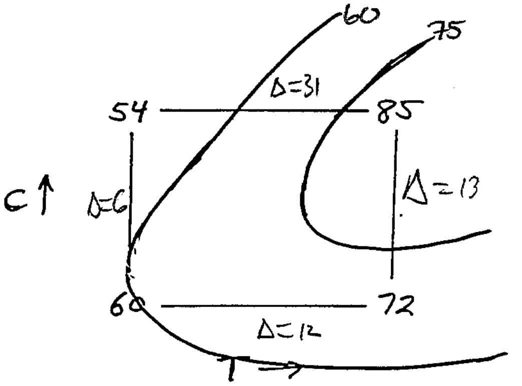
If effect is different ⇒ T×C interaction.

Temp by Conc interaction

Above example → very little interaction.

why, 14 vs 12 and −6 vs −4 are close ⇒ linear in this region

But change 68 → 85



T effect at high C

85 − 54 = 31

substantial difference

T effect at low C

72 − 60 = 12

$\beta_0 = 67.75$

$\beta_1 = 10.75$

$\beta_2 = 1.75$

$\beta_{12} = -4.75$

C effect at low T:

54 − 60 = −6

C effect at high T:

85 − 72 = +13

big difference

⇒ Large T × C interaction because of these large differences

⇒ non linear

PTO

We could fit a model of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \underset{\substack{\uparrow \\ \text{interaction term}}}{\beta_{12} x_1 x_2} + e$$

Consider the effect of x_1 at some fixed $x_2 = x_2^*$

$$\Rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^* + \beta_{12} x_2^* x_1$$

$$y = (\beta_0 + \beta_1 x_2^*) + (\beta_1 + \beta_{12} x_2^*) x_1$$

~~$y = \beta_3 + \beta_4 x_1$~~

- if $\beta_{12} = 0$

$\Rightarrow$ no interaction

$\Rightarrow x_1 \neq fn(x_2)$

- if $\beta_{12} \neq 0$

$\Rightarrow$ interaction and the effect of x_1 on y depends on the level of x_2

Design so far:

T	C	
160°	20	LL
180	20	HL
160	40	LH
180	40	HH

Centre condition $T = 170°$ $C = 30\%$

$\Longrightarrow$ usually your current operating point

Transform T + C to scaled variables

$$x_i = \frac{\text{Variable} - \text{Center point}}{\text{Range}/2}$$

$$T_{scaled} = x_1 = \frac{T-170}{10}$$

$$C_{scaled} = x_2 = \frac{C-30\%}{10}$$

Range of x_i's

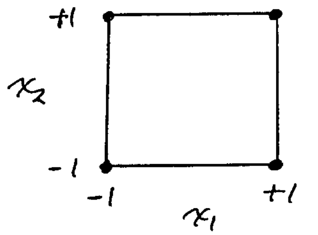
-1 to $+1$

-1 to $+1$

Design matrix becomes:

with low $= -1$ high $= +1$

x_1	x_2
-1	-1
$+1$	-1
-1	$+1$
$+1$	$+1$



Fit model : $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$

$\underline{\eta} = \underline{X}\,\underline{\beta}$

this is an orthogonal design $\rightarrow$ PTO

$\underline{\eta} = \underline{X}\,\underline{\beta}$ $\underline{X}$ = independent variable matrix

$$\underline{X} = \begin{array}{c} \begin{array}{cccc} x_0 & x_1 & x_2 & x_1x_2 \end{array} \\ \begin{bmatrix} +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 \\ +1 & -1 & +1 & -1 \\ +1 & +1 & +1 & +1 \end{bmatrix} \end{array} \qquad \underline{Y} = \overset{\text{responses}}{\begin{bmatrix} 60 \\ 72 \\ 54 \\ 68 \end{bmatrix}} \qquad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \end{bmatrix}$$

Fit model by LS : $\hat{\underline{\beta}} = (X^TX)^{-1}X^TY$

$$X^TX = \begin{bmatrix} \Sigma x_0^2 & 0 & 0 & 0 \\ 0 & \Sigma x_1^2 & 0 & 0 \\ 0 & 0 & \Sigma x_2^2 & 0 \\ 0 & 0 & 0 & \Sigma (x_1x_2)^2 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

$\rightarrow$ prove this for yourself $\Longrightarrow (X^TX) = (X^TX)^{-1}$

Columns of $\underline{X}$ are orthogonal ($\underline{x}_i^T\underline{x}_j = 0$)

ie. $\Sigma x_0x_1 = \Sigma x_0x_2 = \Sigma x_0(x_1x_2) = \Sigma x_1x_2$

$= \Sigma x_1(x_1x_2) = \Sigma x_2(x_1x_2) = 0$

$$\therefore \hat{\underline{\beta}} = \overset{(X^TX)^{-1}}{\begin{bmatrix} 1/\Sigma x_0^2 & 0 & 0 & 0 \\ & 1/\Sigma x_1^2 & 0 & 0 \\ & & 1/\Sigma x_2^2 & 0 \\ & & & 1/\Sigma x_1x_2 \end{bmatrix}} \overset{\text{times}}{} \overset{(X^TY)}{\begin{bmatrix} \Sigma x_0 y \\ \Sigma x_1 y \\ \Sigma x_2 y \\ \Sigma (x_1x_2) y \end{bmatrix}}$$

ie $\hat{\beta}_i = \dfrac{\Sigma x_i y}{\Sigma x_i^2}$

Each $\hat{\beta}_i$ can be calculated independently because X^TX is diagonal

eg $\hat{\beta}_1 = \dfrac{-y_1 + y_2 - y_3 + y_4}{4}$

$\hat{\beta}_0 = \dfrac{+y_1 + y_2 + y_3 + y_4}{4} = \bar{y}$

$\hat{\beta}_2 = \dfrac{-y_1 - y_2 + y_3 + y_4}{4}$

each column of X is orthogonal $\Rightarrow \underline{x}_1 \cdot \underline{x}_2 = 0$

$\Rightarrow \underline{x}_i \cdot \underline{x}_j = 0 \quad i \neq j$

$\Rightarrow \underline{x}_i^T \underline{x}_j = 0$

Factorial designs lead to orthogonal matrices of $\underline{\underline{X}}$

The literature and BH^2 values for $\hat{\beta}_i$ are twice the size of the ones we would obtain $\Rightarrow$ we should halve their values
$\Rightarrow$ we should double the D-9 values we calculate to compare to theirs

Factorial table in Matlab 2^3
(fullfact([2 2 2])-1).*2-1

Note: I will denote $\hat{\beta}_i$ = effect of variable x_i

($\hat{\beta}_i$ = effect on y of changing x_i from $0 \rightarrow +1$)
(or from -1 to 0)

Most texts (BH^2 + MR) denote "effect of x_i" as
= change in y due to changing x_i from -1 to +1

ie $= 2\hat{\beta}_i \hookrightarrow \hat{\beta}_1 = \frac{-y_1+y_2-y_3+y_4}{2}$ in literature

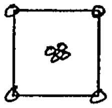
$Var(\hat{\underline{\beta}}) = (X^TX)^{-1}\sigma^2$ → because we will use linear least squares

$\therefore Var(\hat{\beta}_i) = \frac{\sigma^2}{\Sigma x_i^2} = \frac{\sigma^2}{4}$ → variance of the error
$y = \underline{\beta}\underline{x} + e$

ⓧ $\hat{\beta}_i$'s are all uncorrelated due to orthogonality of the design.

If σ^2 is unknown → estimate (s^2) from prior data or replicates

Possibilities : (i) Replicate whole factorial
(ii) Add replicate centre points



(iii) Normal Probability Plots
(see later)

95% CI on β_i's

$$\hat{\beta}_i \pm t_{\nu,0.025} \boxed{\sqrt{s^2/\Sigma x_i^2}}$$

standard error of the estimate → $\sqrt{\text{(estimated) variance}}$

If restricted - do runs that provide new information rather than replicates

ν = num of DOF from estimating s^2

If you are doing 2^2 factorial you estimate β_0 β_1 β_2 β_{12} (4 variables) and 4 runs $\Rightarrow$ 0 DOF left [PTO]

2^3 Factorial Design

2^3 ← 3 variables, ← 2 levels ($2^3 = 8$ runs)

Variables : T, C, Catalyst type (eg A, B) → on/off variable
↑ qualitative variable

Denote : $x_3 = -1$ for catalyst A
$= +1$ " " B

2^3 factorial = All combinations of the 2 levels of the 3 variables

Run Order	x_0	alternate x1 x_1	alternate x2 x_2	alternate x4 x_3	x_1x_2	x_1x_3	x_2x_3	$x_1x_2x_3$
6	+1	-1	-1	-1	+1	+1	+1	-1
3	+1	+1	-1	-1	-1	-1	+1	+1
1	+1	-1	+1	-1	-1	+1	-1	+1
7	+1	+1	+1	-1	+1	-1	-1	-1
2	+1	-1	-1	+1	+1	-1	-1	+1
8	+1	+1	-1	+1	-1	+1	-1	-1
5	+1	-1	+1	+1	-1	-1	+1	-1
4	+1	+1	+1	+1	+1	+1	+1	+1

Randomized → to reduce effect of lurking variables

Design Matrix

X (indep. var. matrix)

Could look at say 3 catalysts by introducing dummy variables we always need one less dummy variable eg 2 dummy variables

[PTO]

Practical considerations

If it is not feasible to randomize all variables, at least randomize the other variables – more of a risk.

DOF

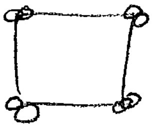
If you run the 4 runs and 4 runs at the centre point

~~⇒ 4 repeats~~ s^2 at the centre point has 3 DOF

use the variance at centre point as s^2

$t_{3,0.025}$ is used for $\hat{\beta_i} = \beta_i \pm t_{v,0.025} (SE)$

If you run 4 runs, twice ie:



$\boxed{\text{and if model is correct}} \Rightarrow \text{DOF} = 4 = n - p$

$= 8 - 4$

$p = 4$ parameters estimated

s^2 is obtained from the four repeats/replicates

σ^2: – prior data – if you used $n=100$ from prior data ⇒ $t_{100,95\%}$ is used for $\pm t_{v,0.025}\sqrt{\frac{\sigma^2}{n}}$

– replicates

PTO

$$\eta = \beta_0 + \overbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}^{\text{three main effects}} \quad \text{Linear Regression only}$$

$$+ \underbrace{\beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3}_{\text{two factor interactions}} + \underset{\uparrow \text{ 3 factor interaction}}{\beta_{123} x_1 x_2 x_3}$$

$\underline{\eta} = \underline{X}\underline{\beta}$ estimate 8 ~~effects~~

estimate 8 parameters/effects from 8 runs $\Rightarrow$ no degrees of freedom to calculate s^2

Again by LS: $\hat{\beta}_i = \dfrac{\sum x_i y}{\sum x_i^2}$

(This design analysed in assignment #2)

(Again note that all columns of $\underline{X}$ are orthogonal.)

- 2^k Factorial in k variables can easily be written down in standard form.

Desirable Features of Factorial Designs

(i) Orthogonal $\rightarrow$ easy calculations
$\rightarrow$ uncorrelated estimates $\hat{\beta}_i$

(ii) Good variation in all variables

(iii) Efficient use of all data points *

(iv) Well patterned design $\rightarrow$ Good visual appreciation

(v) Allows experiments to be performed in blocks (Fractional Factorials)

(vi) Allows designs of increasing order to be built up sequentially

Assessing Significance of Effects when we have no estimate of σ^2

Example: 2^4 factorial with no replicates (16 runs)

BH^2 page 327

If had estimate s^2 with ν df, then 95% CI's

$$\hat{\beta}_i \pm t_{\nu,0.025} \sqrt{\boxed{s^2/16}} \leftarrow Var(\hat{\beta}_i)$$

What to do if have no variance estimate s^2?

(1.) Use estimates of effects $(\hat{\beta}_j)$ that are expected to be small.

ie Estimate $Var(\hat{\beta}_i)$ from 3 + 4 factor interactions.

$\longrightarrow$

- But, how do know which interactions to use?
- If use only smallest ones $\longrightarrow$ under estimation of $Var(\hat{\beta}_i)$

Variance of $\hat{\beta}_i$ for 2^3 design

$$Var(\hat{\beta}_i) = (X^T X)^{-1} \sigma^2$$

$$= \frac{\sigma^2}{\sum x_i^2} = \frac{\sigma^2}{8}$$

Then if you have estimate of σ^2/s^2 you can get Conf. Intervals.

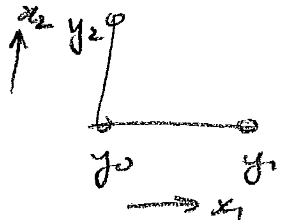
Use of data points

from p D-8 $\quad 2^2$ has 4 runs

every value of y is used to estimate every effect

2^5 - all 32 runs are used to estimate $\hat{\beta}_i$

If we had a bad design



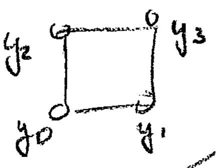
$\hat{\beta}_1 = \frac{y_1 - y_0}{2}$ $\quad \hat{\beta}_2 = \frac{y_2 - y_0}{2}$ ⟸ data points for all runs are not used this time

$$Var(\hat{\beta}_1) = \frac{1}{4} Var(y_1 - y_0)$$

$$= \frac{1}{4}(\sigma^2 + \sigma^2) \text{ because } y_1 \text{ and } y_0 \text{ are independent}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{2}$$ ⟸ have halved the variance

A factorial design



→ Also with the bad design one cannot estimate the interaction effects

$$Var(\hat{\beta}_i) = \frac{\sigma^2}{n} = \frac{\sigma^2}{4}$$

TABLE 10.7. Signs for calculating effects for a 2^4 factorial, process development example

24 = 2 & 4 interaction

I	1	2	3	4	12	13	14	23	24	34	123	124	134	234	1234	conversion (%)
+	−	−	−	−	+	+	+	+	+	+	−	−	−	−	+	71
+	+	−	−	−	−	−	−	+	+	+	+	+	+	−	−	61
+	−	+	−	−	−	+	+	−	−	+	+	+	−	+	−	90
+	+	+	−	−	+	−	−	−	−	+	−	−	+	+	+	82
+	−	−	+	−	+	−	+	−	+	−	+	−	+	+	−	68
+	+	−	+	−	−	+	−	−	+	−	−	+	−	+	+	61
+	−	+	+	−	−	−	+	+	−	−	−	+	+	−	+	87
+	+	+	+	−	+	+	−	+	−	−	+	−	−	−	−	80
+	−	−	−	+	+	+	−	+	−	−	−	+	+	+	−	61
+	+	−	−	+	−	−	+	+	−	−	+	−	−	+	+	50
+	−	+	−	+	−	+	−	−	+	−	+	−	+	−	+	89
+	+	+	−	+	+	−	+	−	+	−	−	+	−	−	−	83
+	−	−	+	+	+	−	−	−	−	+	+	+	−	−	+	59
+	+	−	+	+	−	+	+	−	−	+	−	−	+	−	−	51
+	−	+	+	+	−	−	−	+	+	+	−	−	−	+	−	85
+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	78
divisor 16	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	

Just looking at the first column α where α ± 8 there are some small values

Good assumption that the three order & fourth order interactions are truly zero. They are not zero at the moment due to measurement error

we calculate the variance of these 5 effects $\hat{\beta}_i$

Est $Var(\hat{\beta}_i) = \frac{1.50}{5} = 0.3$

ie we make the hypothesis that H_0 : Higher order interactions are zero

TABLE 10.8. Estimated effects from a 2^4 factorial design, process development example

effects	estimated effects ± standard error
average	72.25 ± 0.27
1	$\boxed{-8.00}$ ± 0.55
2	$\boxed{24.00}$ ± 0.55
3	−2.25 ± 0.55
4	$\boxed{-5.50}$ ± 0.55
12	1.00 ± 0.55
13	0.75 ± 0.55
14	0.00 ± 0.55
23	−1.25 ± 0.55
24	$\boxed{4.50}$ ± 0.55
34	−0.25 ± 0.55
123	−0.75 ± 0.55
124	0.50 ± 0.55
134	−0.25 ± 0.55
234	−0.75 ± 0.55
1234	−0.25 ± 0.55

Interactions: The 'levels of x_1 depends on the level of x_2 $\Rightarrow \beta_{12} \neq 0$

$\sigma_{\hat{\beta}} = 0.55 = SE(\hat{\beta}) \longrightarrow$ from $S^2 = \frac{1}{n-p}\sum(y_i - \hat{y}_i)^2$ then $\frac{S^2}{16} = 0.3$

Calculation of Standard Errors for Effects Using Higher-Order Interactions

No direct estimate of σ^2 is available from these 16 runs since there were no replicates. However we can obtain such an estimate if certain assumptions are made. In particular, if all three- and four-factor interactions are supposed negligible (an assumption made plausible by the earlier discussion of smoothness and similarity of response functions) these higher-order interactions would measure differences arising principally from experimental error.

They could thus provide an appropriate reference set for the remaining effects. We find:

	effect	effect^2
123	−0.75	0.5625
124	0.50	0.2500
134	−0.25	0.0625
234	−0.75	0.5625
1234	−0.25	0.0625
	sum	1.5000

$0.3 = \frac{1.5}{5}$ = estimate of variance with 5DOF $\Rightarrow$ ~~SE~~ standard error $= \sqrt{0.3} = 0.55$

(2.) Normal Probability Plots

Montgomery + Runger : pg. 449, 737-744
BH^2 : pg. 329-334
Draper + Smith : pg. 177-183

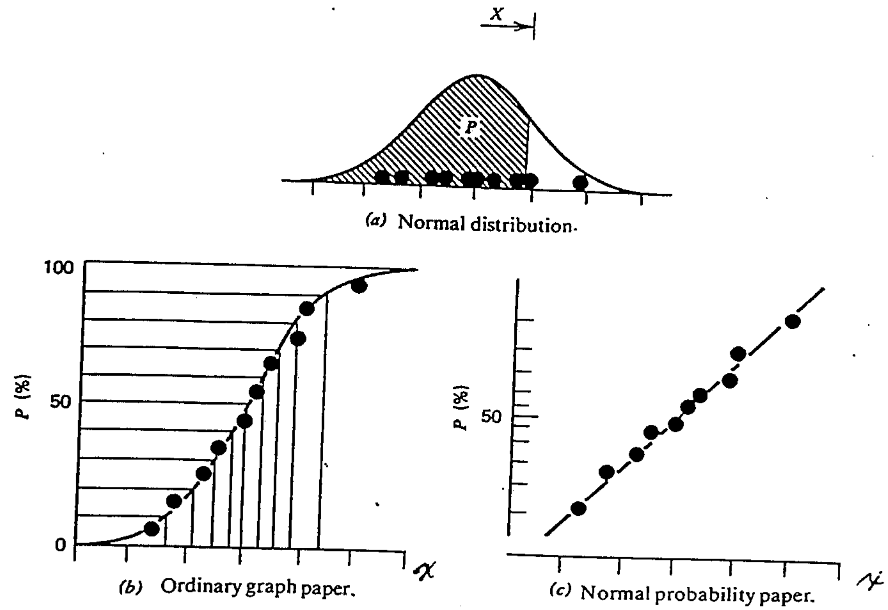
Example : 2^4 factorial in $n=16$ runs (BH^2)
Estimate 15 effects $\hat{\beta}_i$ $(i=1,2\cdots,15)$ + $\hat{\beta}_0$

Under H_0 : $\beta_i = 0$ for all i, we would expect that all $\hat{\beta}_i$'s would come from a Normal distribution $N(0, \frac{\sigma^2}{\Sigma x_i^2})$ ← $Var(\hat{\beta}_i)$

∴ Use Normal Probability Plot to see if this is true.

Idea of Normal Prob. Plots → Fig 10.8 BH^2

Why normal distributions?
- central limit theorem
- if y is normally distributed then so should β_i because they are just linear combinations of y



(a) Normal distribution.

(b) Ordinary graph paper.

(c) Normal probability paper.

FIGURE 10.8. Normal probability plots.

If they fall on a straight line ⇒ they come from a normal distribution

D-18

TABLE 10.9. The 15 ordered effects and the probability points P, process development example

The previous 2^4 design effects arranged in ascending order

order number i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
effect ?	−8.0	−5.5	−2.25	−1.25	−0.75	−0.75	−0.25	−.025	−0.25	0.00	0.50	0.75	1.00	4.50	24.00
identity of effects	*1*	*4*	*3*	*23*	*123*	*234*	*34*	*134*	*1234*	*14*	*124*	*13*	*12*	*24*	*2*
$P = 100(i - \frac{1}{2})/15$	3.3	10.0	16.7	23.3	30.0	36.7	43.3	50.0	56.7	63.3	70.0	76.7	83.3	90.0	96.7

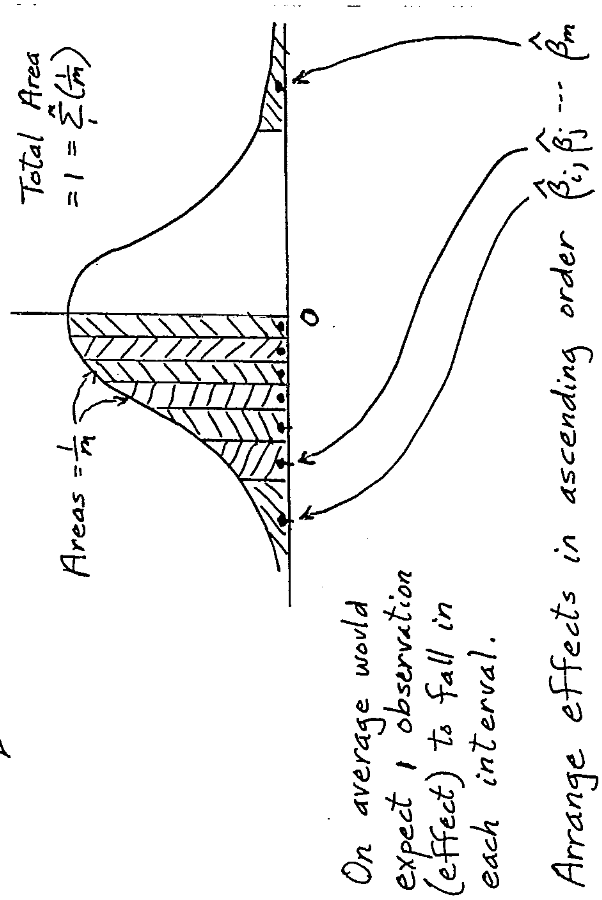
plot P_i ⇒ normal curve or straight line on probability paper

---

Each interval has equal area ≡ equal probability

D-17

If m effects are truly distributed Normally with mean = 0, then could divide area under Normal distribution into m intervals of equal area ($= 1/m$).



On average would expect 1 observation (effect) to fall in each interval.

Arrange effects in ascending order $\hat{\beta}_i, \hat{\beta}_j \text{ --- } \hat{\beta}_m$

Cumulative area up to centre of i-th interval $= (i-\frac{1}{2})/m$.

∴ Cumulative % probability for i-th largest effect $= 100(i-\frac{1}{2})/m$

∴ Plot $\hat{\beta}_i$'s in ascending order against $P_i = 100(i-\frac{1}{2})/m$ on Normal Probability Plot.

Box, Hunter & Hunter Fig. 10.8 Table 10.9
Fig 10.9

The X^TX matrix

- ordinary data: X^TX is not diagonal
  usually badly conditioned

- factorial design: X^TX is diagonal – best conditioning possible

for example:

x_1	x_2
−1	−1
+1	−1
−1	+1
+1	+1

$$X^TX = \begin{bmatrix} x_1^Tx_1 & x_1^Tx_2 \\ x_2^Tx_1 & x_2^Tx_2 \end{bmatrix}$$

$$X^TX = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$

→ if experiment is properly performed ie: x_1 was at +1 or −1

→ each column of the X matrix is orthogonal to the others

→ X^TX is deliberatly chosen to be diagonal to totally avoid the con-founding effects between x_i columns

- other designs: X^TX is "almost" diagonal

Scaling

$$x_1 = \frac{T-170}{10} \qquad x_2 = \frac{C-30}{10}$$

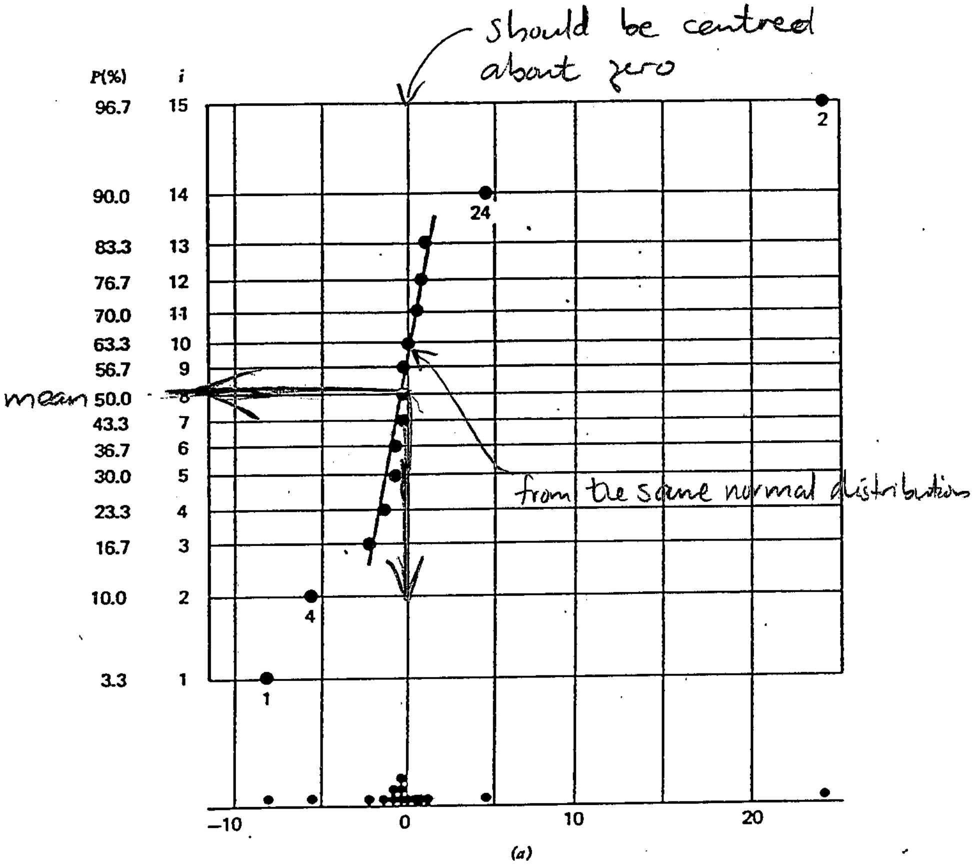
so we need to unscale the variables

Scaled regression: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2$

convert $= \hat{\beta}_0 + \hat{\beta}_1\left(\frac{T-170}{10}\right) + \hat{\beta}_2\left(\frac{C-30}{10}\right) + ---$

simplify $= \hat{\alpha}_0 + \hat{\alpha}_1 T + \hat{\alpha}_2 C + \hat{\alpha}_{12} TC$

Now we can go back and use the points along the line to estimate the variance — they are insignificant — all of them can be pooled that are not circled on p D-14



Should be centred about zero

mean

from the same normal distribution

FIGURE 10.9. (*a*) Normal plot of effects, process development example.

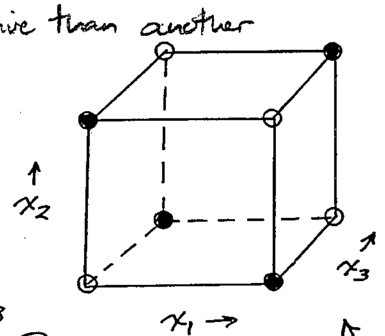
Effects $\hat{\beta}_2, \hat{\beta}_{24}, \hat{\beta}_1, \hat{\beta}_4$ appear to be too large to have come from same Normal distribution as rest of effects.

Blocking of a 2^3 Factorial

Want to examine 3 factors in a $2^3 = 8$ run design. But material to be used in experiment comes in batches sufficient for only 4 runs, and differences may exist between batches of material.
One batch may be more active than another

Can we split the design so that differences in the material will not affect the results? Split the design

PTO



Split on 123 column

1	2	3	12	13	23	123	
−	−	−	+			−	○
+	−	−	−			+	●
−	+	−	−			+	●
+	+	−	+			−	○
−	−	+	+			+	●
+	−	+	−			−	○
−	+	+	−			−	○
+	+	+	+			+	●

Run all exp. with + sign in 123 column in one block + all − signs in other!

(Randomize within blocks,

∴ Any block effect (ie. differences in material) will be CONFOUNDED with 3-factor interaction term x_{123}

The effect we expect to be least important is $\hat{\beta}_{123}$ − split on these runs. We may be tempted to run expts 1−4 on a batch and Expts 5 − 8 on another, but this will confound the effect of x_3, because $x_3 = -1$ for expts 1−4
$x_3 = +1$ for expts 5−8

→ don't want this situation

Split on β_{123} because we suspect least importance, so we intentionally confound variable $x_1 x_2 x_3$, or $\hat{\beta}_{123}$ may be confounded.

Because columns are orthogonal ⟹ for every other variable x_1, x_2, x_3, x_{12}, x_{13}, x_{23} two low levels of each of these variables is associated with Batch 1 and two high levels with Batch 2 ⟹ they cancel each other out.

$$+ \quad - \quad - \quad + \quad : \Sigma = 0$$

$$\hat{\beta}_1 = \frac{1}{8}\left(-\overset{\circ}{y}_1 + \overset{\bullet}{y}_2 + \overset{\bullet}{y}_3 + \overset{\circ}{y}_4 - \overset{\bullet}{y}_5 + \overset{\circ}{y}_6 - \overset{\circ}{y}_7 + \overset{\bullet}{y}_8\right)$$

even if there is a difference between Batch 1 and Batch 2, the effect will cancel out. This is true for every β_i

→ every y from the open ○ is slightly less than the y from the closed ●, but a constant k, say,

$$\hat{\beta}_1 = \frac{1}{8}(\text{—}) + k - k - k + k = \frac{1}{8}(\sim)$$

⟹ no difference } due to
⟹ perfect cancellation } orthogonality

$\therefore$ Confounded | Can't tell whether $\hat{\beta}_{123}$ is due to a real 123 interaction or a block effect (material)

ie. $\hat{\beta}_{123} = \underbrace{123 \text{ effect}}_{\text{expected to be small anyway.}} + \text{block effect}$ (Batch 1 vs Batch 2 effect)

Since 123 column is orthogonal to all other columns, any block effect will have no influence on them!
Any shift in level due to blocking is cancelled out

2^{k-1} FRACTIONAL FACTORIAL DESIGNS

($\frac{1}{2}$ fraction of a 2^k)

Suppose want to examine 3 variables in 4 runs.

Use 2^{3-1} fractional factorial ($2^{3-1} = 4$ runs)

To set up the design, write down a 2^2 in 4 runs

Design Matrix

3 = ← Let 12 column accomodate variable 3

1	2	12
−	−	+
+	−	−
−	+	−
+	+	+

Variable 3 = 12 interaction variable level

Run this 4 run design in 3 variables (1,2,3)

Note: This 2^{3-1} design corresponds to the $\frac{1}{2}$ fraction of 2^3 design given by ●'s. on p D-20

If had associated variable 3 with −12 interaction column we would have got other ½ fraction (ie. o's) open circles

But! Have only 4 runs and so can't estimate all
- 3 main effects
- 3 two-factor interactions
- 1 three factor interaction
- mean

} 8 effects

Confounding: ~~pattern~~

These four are actually performed

I	1	2	3	12	13	23	123
+	−	−	+	+	−	−	+
+	+	−	−	−	−	+	+
+	−	+	−	−	+	−	+
+	+	+	+	+	+	+	+

not orthogonal anymore

Note that:

Column 3 = 12 ← was deliberatly chosen
" 1 = 23
" 2 = 13
" I = 123
} these come from making that decision

$\therefore$ If fit model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

$\hat{\beta}_1 \longrightarrow 1 + 23$ effects
$\hat{\beta}_2 \longrightarrow 2 + 13$ "
$\hat{\beta}_3 \longrightarrow 3 + 12$ "
$\hat{\beta}_0 \longrightarrow$ ~~avg.~~ + 123 "
mean = I column ~~effect~~

} Effects are confounded.
If interactions are small we are still fine.

If k is large, how do we find the columns that are confounded?

<u>GENERATING RELATIONSHIP</u> for design : $I = 123$
we associated 3 with 12 interaction

Key to unraveling the confounding of effects is the <u>DEFINING RELATIONSHIP</u> which in case of single generator is the same as the generating relationship : $\boxed{I = 123}$ Defining Relationship also

x_0	$x_1x_2x_3$	x_1^2	x_2^2	x_3^2
I	123	1^2	2^2	3^2
+	+	+	+	+
+	+	+	+	+
+	+	+	+	+
+	+	+	+	+

Note also $I = 1^2 = 2^2 = 3^2$

Multiply <u>both sides</u> of defining relation by x_K $K=1,2,3$

$$\textcircled{1} \times I \rightarrow 1 = (1^2)23 = 23 \qquad \hat{\beta}_1 = 1+23$$

Identity

$$2 \times I \rightarrow 2 = 12^23 = 13 \qquad \hat{\beta}_2 = 2+13$$

$$3 \times I \rightarrow 3 = 123^2 = 12 \qquad \hat{\beta}_3 = 3+23$$

$$I = 123 \qquad \hat{\beta}_0 = avg + 123(+1^2+2^2+3^2)$$

$$x_1 I = x_1 = \underline{23}$$

$$\begin{array}{cc|c|c|c} - & + & & - & - \\ + & + & & + & + \\ - & + & = & - & = \; - \\ + & + & & + & + \end{array}$$

Quick way for a large number of runs

We might need to run the full factorial

→ the first half are NOT wasted

→ we complete them

After running the ½ fraction, if need more information (ie to break confounding of effects), one can add design given by other ½ fraction.

1	2	3 = $\boxed{-12}$
−	−	−
+	−	+
−	+	+
+	+	−

just the negative

This is other $\frac{1}{2}$ fraction given by O's.

Combining these two $\frac{1}{2}$ fractions (2^{3-1}) we get full 2^3 factorial in <u>two blocks</u> as before when we had two separate blocked batches example

The can estimate
- all 3 main effects
- all 3 2-fi's (factor interactions)
- $\underbrace{\text{block effect + 123 interaction}}_{\text{confounded}}$

2^{4-1} Fractional Factorial

						4 =	
1	2	3	12	13	23	123	34
−	−	−	+			−	+
+	−	−	−			+	−
−	+	−	−			+	−
+	+	−	+			−	+
−	−	+	+			+	+
+	−	+	−			−	−
−	+	+	−			−	−
+	+	+	+			+	+

- Write down full 2^3 factorial.
- Associate variable 4 with 123 column

Confounding of Effects ?

Generator : $I = 1234$ (Also defining relationship)

we associated 123 with 4 $\Rightarrow I = 1234$

$$\text{Confounding pattern}\left\{\begin{array}{lll} 1\times I = 1^2 234 : & \hat{\beta}_1 \rightarrow & 1 + 234 \\ \quad 12^2 34 & \hat{\beta}_2 \rightarrow & 2 + 134 \\ & \hat{\beta}_3 \rightarrow & 3 + 124 \\ & \hat{\beta}_4 \rightarrow & 4 + 123 \leftarrow \text{how we designed it} \\ 12\times I = 1^2 2^2 34 : & \hat{\beta}_{12} \rightarrow & 12 + 34 \\ & \hat{\beta}_{13} \rightarrow & 13 + 24 \\ & \hat{\beta}_{23} \rightarrow & 23 + 14 \\ & \hat{\beta}_0 \rightarrow & \text{avg.} + 1234 \end{array}\right.$$

$\bar{\beta}_i$ = actual, true value from full expt.

$\hat{\beta}_{12}$ could be small if $\bar{\beta}_{12}$ is big $\bar{\beta}_{34}$ is big but opposite sign

If, as is often true, 3 factor interactions are small, then will have estimates of
- All main effects
- Three combinations of 2 fi's (factor interactions)

Why not chose generator for 2^{4-1} design as
$I = 124$ (ie assign 4 to 12 column) ?

Poorer choice because main effects would then be confounded with 2fi's
(eg. 1+24, 2+14, - - -)
8+1234
12+34
24+1

D-optimal: We have run 8 runs, what is the next best run to remove 12+34 confounding

If ambiguities exist after this $\frac{1}{2}$ fraction then can add other $\frac{1}{2}$ fraction given by

$$I = -1234$$

ie. by associating 4 with -123 column.

Combined design is full 2^4 run in 2 blocks.

Iterative approach to design :
- Run one fraction
- Examine results
- Add another fraction if necessary.

Half Fractions of 2^k Designs

(i) Write down a full factorial in (k-1) var.

(ii) Assign k-th variable to an interaction column. Any interaction could be used, but highest order interaction will give design with best confounding pattern.

k	Design	Generator	# Exp.
3	2^{3-1}	$I = 123$	4
4	2^{4-1}	$I = 1234$	8
5	2^{5-1}	$I = 12345$	16

Variable 5 associated with 1234 interaction

For $k > 5$ # Experiments becomes large

$\therefore$ Fractionate further $\Rightarrow 2^{k-p}$ factorial
(2^{-p} fraction of 2^k design)

We want as high a resolution as possible

Resolution in 2^{k-p} Designs

Resolution III designs are those with :

(i) No main effects confounded with other main effects

(ii) But some main effects confounded with 2 fi's

$\text{eg. } 2^{4-1}_{\mathrm{III}} \quad I = 124$

difficult to interpret

Resolution $\overline{\underline{\mathrm{IV}}}$:

(i) No main effect confounded with other main effects or with 2 fi's.

Main effect confounded with 3 or higher interactions

(ii) But 2 fi's confounded with one another

same as example on previous page

$\text{eg. } 2^{4-1}_{\mathrm{IV}} \quad I = 1234$

4 associated with fi = 123

Resolution $\overline{\underline{\mathrm{V}}}$:

(i) No main effect <u>or</u> 2 fi confounded with any other main effect or 2 fi.

$\text{eg. } 2^{5-1}_{\mathrm{V}} \quad I = \pm 12345$

Can get main & 2fi. parameters

A design is of Resolution R if the smallest <u>word</u> in the <u>defining relation</u> is of length R.

Once we have run a saturated design we cleverly chose the next set of runs so that in the end we have not done the full factorial design but we have obtained the maximum information

A Special Class of Resolution III Designs:

Saturated Designs

Good for screening designs

Saturated Design : (N-1) variables in N runs

eg. 7 variables in 8 runs : 2^{7-4}

15 " " 16 " : 2^{15-11}

Example : 2^{7-4}_{III} (8 runs) Resolution = III

2^{-4} or $\frac{1}{16}$ fraction of a 2^7 factorial (128 runs)

$8 = \frac{1}{16}(128)$

To construct : Write down a 2^3 factorial (8 runs) + associate variables 4, 5, 6, 7 with interactions

1	2	3	4= 12	5= 13	6= 23	7= 123
−	−	−	+	+	+	−
+	−	−	−	−	+	+
−	+	−	−	+	−	+
+	+	−	+	−	−	−
−	−	+	+	−	−	+
+	−	+	−	+	−	−
−	+	+	−	−	+	−
+	+	+	+	+	+	+

Should still be run in random order

Generators : I = 124 = 135 = 236 = 1237

Num Generators = Num of assigned variables

Product of any generator = I with another generator

If I = 124 and I = 135 ,

then I = (124)(135) = 2345

$= (1)^2 2345$

∴ When a design has more than one generator the <u>defining relation</u> includes each generator and all possible products of them .

∴ Defining relation for this design is

I = 124 = 135 = 236 = 1237 (one generator at a time)

I = 2345 = 1346 = 347 = 1256 = 257 = 167 } multiplying 2 at a time

I = 456 = 1457 = 2467 = 3567 (3 at a time)

= 1234567 (4 at a time)

All these are identity columns

Resolution = III since smallest word length has only 3 characters

When had one generator ⟹ one defining relation

Many with ~~three~~ word relations.

Very fractionated ⇒ very confounding

Confounding Pattern of Variable 1

all of these are confounded with β_0

Multiply defining relation through by 1 giving

$1 = 24 = 35 = 1236 = 237 = 12345 = 346$

$= 1347 = 256 = 1257 = 67 = \cdots$

As number of runs ↓ confounding ↑

<u>Following estimated effects can be obtained</u>

(ignoring 3 f.i.'s and higher) hypothesis: these are not important

eg 1234, 12345 etc

$\hat{\beta}_0 =$ avg. of 8 runs

$\hat{\beta}_1 = 1 + 24 + 35 + 67$

$\hat{\beta}_2 = 2 + 14 + 36 + 57$

$\hat{\beta}_3 = 3 + 15 + 26 + 47$

$\hat{\beta}_4 = 4 + 12 + 56 + 37$

$\hat{\beta}_5 = 5 + 13 + 46 + 27$

$\hat{\beta}_6 = 6 + 23 + 45 + 17$

$\hat{\beta}_7 = 7 + 34 + 25 + 16$

obtained from the defining relation

First 8 runs (principal fraction)

(Ref. BH^2 page 392)

compare

$N =$ number of runs in satd design or $(N-1) =$ number of variables being screened

$2^{N-1}/N$ others

What other designs are possible?

Alternative Fractions from same family of this 2^{7-4}_{III} Design.

16 designs : $I = \pm 124 = \pm 135 = \pm 236 = \pm 1237$

4 associated with +12 or −12 etc

Example : Associate variable 5 with −13 column and 6 with −23 col.

Now the generator becomes: by our association

Generator : $I = 124 = -135 = -236 = 1237$

Defining Relation : take all possible products

$I = 124 = -135 = -236 = 1237$

$= -2345 = -1346 = 347 = 1256 = -257 = -167$

$= 456 = -1457 = -2467 = 3567$

$= 1234567$

Estimated Effects : $\boxed{\hat{\beta}_0' = \text{avg.}}$

Second 8 runs

$\hat{\beta}_1' = 1 + 24 - 35 - 67$

$\hat{\beta}_2' = 2 + 14 - 36 - 57$

$\hat{\beta}_3' = 3 - 15 - 26 + 47$

$\hat{\beta}_4' = 4 + 12 + 56 + 37$

$\hat{\beta}_5' = 5 - 13 + 46 - 27$

$\hat{\beta}_6' = 6 - 23 + 45 - 17$

$\hat{\beta}_7' = 7 + 34 - 25 - 16$

this is what they are estimating

New confounding pattern that we obtain

Run 1: $\beta_1 = 1 + 24 + 35 + 67$

Run 2: $\beta_1' = 1 + 24 - 35 - 67$

$\frac{\text{add}}{2}$ $\beta_1 = 1 + 24$ $\qquad \frac{\text{Subtract}}{2} = 35 + 67$

Combining the 2 Fractions

→ Look at defining relations/add & subtract

→ Combine and redo the least squares over

Take $\frac{1}{2}$ sum and $\frac{1}{2}$ difference of effects estimated from the 2 fractions to get:

$\frac{1}{16} + \frac{1}{16} = \frac{1}{8}$ fraction effectively

$\frac{1}{2}$ sum	$\frac{1}{2}$ difference
$\frac{1}{2}(\hat\beta_0 + \hat\beta_0') =$ avg.	$\frac{1}{2}(\hat\beta_0 - \hat\beta_0') =$ block effect
$\frac{1}{2}(\hat\beta_1 + \hat\beta_1') = 1 + 24$	$\frac{1}{2}(\hat\beta_1 - \hat\beta_1') = 35 + 67$
$\frac{1}{2}(\hat\beta_2 + \hat\beta_2') = 2 + 14$	$\frac{1}{2}(\hat\beta_3 - \hat\beta_3') = 36 + 57$
$3 + 47$	$15 + 26$
$4 + 12 + 56 + 37$	higher order interactions
$5 + 46$	$13 + 27$
$6 + 45$	$23 + 17$
$7 + 34$	$25 + 16$

Compare/Combine their defining relations

Defining Relationship of combined design

= Words common to defining relationship of both $\frac{1}{16}$ fractions

$= 124 = 1237 = 347 = 1256 = 456 = 3567 = 1234567$

$+135$ and -135 are not common

Adding fractions in sequence with suitably swithed signs

→ useful in resolving ambiguities that exist after a set of experiments has been run.

Two types of sign switching particularly useful

(i) Changing signs of one factor only

(ii) Changing signs of all factors

(i) Changing sign of one factor only

Suppose have already run 2^{7-4}_{III}

$I = 124 = 135 = 236 = 1237$ associated with all positive elements

Add new fraction with signs of variable 1 switched.

ie. all columns same except #1 → $\begin{matrix}+\\+\\\vdots\end{matrix}$

New design generators + defining relation:

Replace 1 by −1

ie 4 associated with −12; 5 with −13;

ie. $I = -124 = -135 = 236 = -1237$

Effects estimated:

$\hat\beta_1' = 1 - 24 \mp 35 - 67$

$\hat\beta_2' = 2 - 14 + 36 + 57$

$\hat\beta_3' = 3 - 15 + 26 + 47$

$\vdots$

$\hat\beta_7' = 7 + 34 + 25 - 16$

After the second set of runs we are in a position to estimate 16 parameters or effects.

We need to find what the 16 will be

- add & subtract the defining relationships
- compare / combine the defining relationships' words

The confounding that we break out is dependant on the second fraction we choose to run - which one do we choose to break certain confounding patterns.

Combining the 2 Fractions

Alternative :

- Put all 16 runs together + fit by least squares
- Use defining relationship of the combined design (words in common from defining relations of the 2 separate fractions)

$$I = 124 = 347 = 456 = 1237 = 1256 = 3567 = 1234567$$

∴ Ignoring 3 f.i.'s and higher

$\hat{\beta}_1 = 1 + 24$

$\hat{\beta}_2 = 2 + 14$

$\hat{\beta}_3 = 3 + 47$

$\hat{\beta}_4 = 4 + 12 + 37 + 56$

⋮

$\hat{\beta}_{35} = 35 + 67$

$\hat{\beta}_{36} = 36 + 57$

⋮

- Same result as got from adding + subtracting the results of the 2 separate fractions.
- But easier approach when have to combine 3 or more fractions.

If from the principal screening run it appears as if Variable 1 & its interactions appear to be important

Switching the sign of column 1

Combining 2 fractions ⇒ brings out the main effects and its two factor interaction

$\frac{1}{2}(\hat{\beta}_i - \hat{\beta}_i')$		$\frac{1}{2}(\hat{\beta}_i + \hat{\beta}_i')$	
24+35+67		$\boxed{1}$	main effect
14	two factor interactions	2 + 36 + 57	
15		3 + 26 + 47	
12		4 + 56 + 37	
13		5 + 46 + 27	
17		6 + 23 + 45	
16		7 + 34 + 25	

two higher order effects as well – not shown

ie. Adding to one fraction a second fraction with signs of a single variable switched, we isolate effect of that variable + all its 2 f.i.'s

(ii) Switching signs of all variables

1 → -1   2 → -2   - - -   7 → -7

$I = -124 = -135 = -236 = 1237$

work out new defining relationships

Effects estimated : $\hat{\beta}_0' = \text{avg.}$

$\hat{\beta}_1' = 1 - 24 - 35 - 67$

$\hat{\beta}_2' = 2 - 14 - 36 - 57$

⋮

$\hat{\beta}_7' = 7 - 34 - 25 - 16$

Combining 2 fractions

$\frac{1}{2}(\hat{\beta}_i + \hat{\beta}_i')$	$\frac{1}{2}(\hat{\beta}_i - \hat{\beta}_i')$
Block effect	Avg.
1	24 + 35 + 67
2	14 + 36 + 57
3	
4	⋮
5	
6	
7	34 + 25 + 16

ie. Adding to a fraction a further fraction with signs of all variables reversed, we separate main effects of all variables from the 2 f.i.'s.

So now have all main effects and clusters of two factor interactions

May stop experiments here if a normal probability plot shows the other terms to negligible

Could add a third fraction to remove the confounding from some of the two factor interactions

ICI plant, crystallization & filtration

D-38

BH² Pages 424 – 429 Screening design D-37

TABLE 13.4. Results of Example 3

variable	– Current operation	+ Alternative Operation
1 water supply	town reservoir	well
2 raw material	on site	other
3 temperature	low	high
4 recycle	yes	no
5 caustic soda	fast	slow
6 filter cloth	new Supplier	old supplier
7 holdup time	low	high

2^{7-4}_{III}

test	1	2	3	12 4	13 5	23 6	123 7	filtration time (min) y
1	–	–	–	+	+	+	–	68.4
2	+	–	–	–	–	+	+	77.7
3	–	+	–	–	+	–	+	66.4
4	+	+	–	+	–	–	–	81.0
5	–	–	+	+	–	–	+	78.6
6	+	–	+	–	+	–	–	41.2
7	–	+	+	–	–	+	–	68.7
8	+	+	+	+	+	+	+	38.7

Four Tentative Interpretations of Results

In Table 13.5 three of the calculated effects (l_1, l_3, and l_5) are large in absolute value and have been circled. There are several possible interpretations. Four of the most likely are:

1. Main effects *1*, *3*, and *5* are producing the effects.
2. Main effects *1* and *3* and interaction *13* are producing the effects.
3. Main effects *1* and *5* and interaction *15* are producing the effects.
4. Main effects *3* and *5* and interaction *35* are producing the effects.

cannot resolve this which is the best second fraction to run

TABLE 13.5. Calculated values and abbreviated confounding pattern for eight-run filtration experiment, Example 3 ignore 3 f.i.

MacGregor's Notation

$2\hat{\beta}_1 = l_1 = \boxed{-10.9} \rightarrow 1 + 24 + \boxed{35} + 67$

$2\hat{\beta}_2 = l_2 = -2.8 \rightarrow 2 + 14 + 36 + 57$

$l_3 = \boxed{-16.6} \rightarrow 3 + \boxed{15} + 26 + 47$

$l_4 = 3.2 \rightarrow 4 + 12 + 37 + 56$

$l_5 = \boxed{-22.8} \rightarrow 5 + \boxed{13} + 27 + 46$

$l_6 = -3.4 \rightarrow 6 + 17 + 23 + 45$

$l_7 = 0.5 \rightarrow 7 + 16 + 25 + 34$

TABLE 13.6. Results of second filtration experiment, Example 3

test	1	2	3	–12 4	–13 5	–23 6	123 7	filtration time (min) y
9	+	+	+	–	–	–	+	66.7
10	–	+	+	+	+	–	–	65.0
11	+	–	+	+	–	+	–	86.4
12	–	–	+	–	+	+	+	61.9
13	+	+	–	–	+	+	–	47.8
14	–	+	–	+	–	+	+	59.0
15	+	–	–	+	+	–	+	42.6
16	–	–	–	–	–	–	–	67.6

Break it out into the main effect, switch sign of all the variable – because these possibilities include two main effects

TABLE 13.7. **Calculated values and abbreviated confounding pattern for 16-run filtration experiment, Example 3**

$2\hat{\beta}_1 =$	$l_1 =$	$\boxed{-6.7}$	→ *1*
$2\hat{\beta}_2 =$	$l_2 =$	−3.9	→ *2*
	$l_3 =$	−0.4	→ *3*
	$l_4 =$	2.8	→ *4*
	$l_5 =$	$\boxed{-19.2}$	→ *5*
	$l_6 =$	0.1	→ *6*
	$l_7 =$	−4.4	→ *7*
	$l_{12} =$	0.5	→ *12 + 37 + 56*
	$l_{13} =$	−3.6	→ *13 + 27 + 46*
	$l_{14} =$	1.1	→ *14 + 36 + 57*
	$l_{15} =$	$\boxed{-16.2}$	→ *15 + 26 + 47*
	$l_{16} =$	4.9	→ *16 + 25 + 34*
	$l_{17} =$	−3.4	→ *17 + 23 + 45*
	$l_{24} =$	−4.2	→ *24 + 35 + 67*

Normal probability plot shows if these are statistically significant

1 = water
2 = caustic addition

$$y = \beta_0 - \frac{6.7}{2}x_1 - \frac{19.2}{2}x_5 - \frac{16.2}{2}(x_1 x_5)$$

To minimise y one sets $\left.\begin{array}{l} x_1 = +1 \\ x_5 = +1 \end{array}\right\} x_1 x_5 = 1$

$\Rightarrow$ minimises the filtration time

Going from − to + decreases filtration time by $\frac{1}{2}(6.7 + 19.2 + 16.2)$

Can see the interactions in the averages

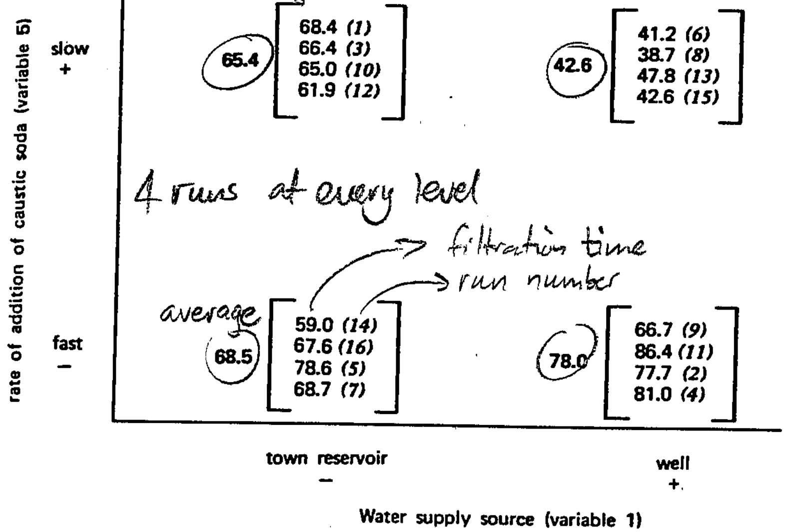


FIGURE 13.2. Results of the 16 trials in relation to two variables, water supply source (1) and rate of addition of caustic soda (5), Example 3. The average result under each one of the four conditions is in bold type. The test condition numbers are given in parentheses.

Equivalent to a 2^2 factorial if the other variables are not important want the highest resolution possible

III = main effects not confounded with each other

IV = no main effects confounded with 2 factor interactions

V = IV and no 2fi. are confounded either

D-50

Resolution III + IV Designs with Arbitrary Number of Variables

Commonly Used

# of variables	# of runs	Designs	Typical Generators
4	8	2_{IV}^{4-1}	shortest word length 1234
5	8	2_{III}^{5-2}	1234, 235
6	8	2_{III}^{6-3}	1234, 235, 136
7	8	2_{III}^{7-4}	1234, 125, 136, 237
15	16	2_{III}^{15-11}	
31	32	2_{III}^{31-26}	
8	16	2_{IV}^{8-4}	

Note : Best to set up design using symbols 1, 2, --- and then assign which variable will be 1, 2, 3 -- etc. only after looking at confounding patterns

ie: if you look at the confounding pattern and see that 1&2 are "lumped" together, then assign two variable that you know will not interact

Other Equivalent Saturated Designs

Taguchi Orthogonal Arrays

really equivalent to standard fractional factorial systems

2-Level Designs :

$L_4 \longrightarrow 2_{III}^{3-1}$

$L_8 \longrightarrow 2_{III}^{7-4}$

$L_{12} \longrightarrow$ 12 run Plackett + Burman

$L_{16} \longrightarrow 2_{III}^{15-11}$

normally go 4, 8, 16, 32 (can get a 12 run design)

3-Level Designs (Taguchi 3 level designs)

$L_9 \longrightarrow 3^2$ with 2 variables (main + 2fi + quad) +quadratic effect

3^{3-1} " 3 " (main + quad + group of 2fi)

9 runs

3^{4-2} " 4 " (main + quad) more confounded

Start with a standard (linear) design and if we feel there are non-linearities we expand the work done so far to include nonlinearities

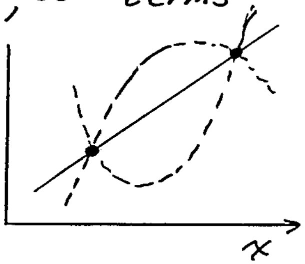
DESIGNS FOR 2^{ND} ORDER MODELS

BH^2 – Chapter 15
Montgomery + Runger: Chapter 12.9

If 1st order + interaction model exhibits Lack of Fit

→ include $x_1^2, x_2^2, \ldots$ terms

Need more than 2 level designs to capture the quadratic effects

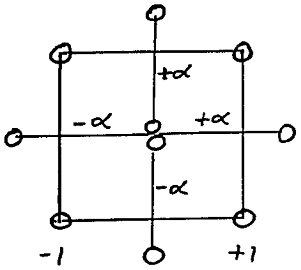


Central Composite Designs – a class from response surface designs

1. Start with 2^k or 2^{k-p} design with center points.

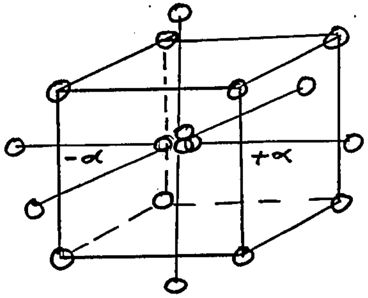
x_1	x_2	
-1	-1	2^2
+1	-1	
-1	+1	
+1	+1	
0	0	cp
$-\alpha$	0	star
$+\alpha$	0	
0	$-\alpha$	
0	$+\alpha$	
0	0	cp

2. Add vertices of star



For 2 variable design
$\alpha = \sqrt{2} = 1.414$ is good choice

3 Variables : $2^3 + cp + star$



$\alpha \doteq 1.68$

	x_1	x_2	x_3
2^3	-1	-1	-1
	+1	-1	-1
	-1	+1	-1
	+1	+1	-1
	-1	-1	+1
	+1	-1	+1
	-1	+1	+1
	+1	+1	+1
cp	0	0	0
	0	0	0
star	$-\alpha$	0	0
	$+\alpha$	0	0
	0	$-\alpha$	0
	0	$+\alpha$	0
	0	0	$-\alpha$
	0	0	$+\alpha$
cp	0	0	0
	0	0	0

For $k = 4$:
$2^4 + cp + star$

For $k > 4$
$2^{k-p} + cp + star$

k	Design	α (for rotatability)
2	2^2	$1.414 = 4^{1/4}$
3	2^3	$1.68 = 8^{1/4}$
4	2^4	$2.0 = 16^{1/4}$
5	2^{5-1}	$2.0 = 16^{1/4}$
6	2^{6-1}	$2.38 = 32^{1/4}$

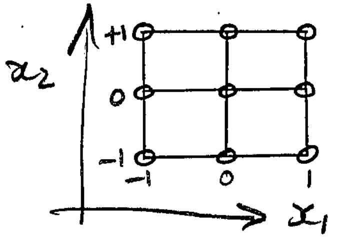
One should run all 3-level designs in a fractionated/incomplete form

for quadratic models we require at least 3 levels

3-LEVEL FACTORIALS

matches central composite design with $\alpha = 1$

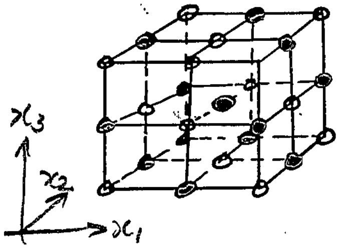
3^2 design : 2 variables at all combinations of 3 levels



9 runs

3^3 design :

27 runs



very dense

⇒ so we fractionate it

Box + Behnken and 3 Level Orthogonal Arrays (Taguchi)

→ Incomplete (fractional) 3 Level designs

only run the ● experiments

Run	x_1	x_2	x_3	
1	-1	-1	0	2^2
2	+1	-1	0	
3	-1	+1	0	
4	+1	+1	0	
5	-1	0	-1	2^2
6	+1	0	-1	
7	-1	0	+1	
8	+1	0	+1	

Run	x_1	x_2	x_3	
9	0	-1	-1	2^2
10	0	+1	-1	
11	0	-1	+1	
12	0	+1	+1	
13	0	0	0	repeats 1 cp for each factorial
14	0	0	0	
15	0	0	0	

minimum of 13 runs

Full quadratic model

$$\eta = b_0 + \overset{\substack{9-12\\1-4}}{b_1 x_1} + \overset{\substack{5-8\\1-4}}{b_2 x_2} + \overset{\substack{5-8\\9-12}}{b_3 x_3}$$

$$+ \overset{1-4}{b_{12} x_1 x_2} + \overset{5-8}{b_{13} x_1 x_3} + \overset{9-12}{b_{23} x_2 x_3}$$

$$+ \overset{1-4}{b_{11} x_1^2} + \overset{5-8}{b_{22} x_2^2} + \overset{9-12}{b_{33} x_3^2}$$

(10 parameters)

Allows for approximation of many responses.

Running all 27 runs one would get the same model, just with improved confidence limits

Most statistical software provides 2-D and 3-D plotting to examine response surface.

from runs 1 to 4 we can get effect of x_1 and x_2 and its interaction and because $x_3 = 0$ we can also estimate $b_{11} x_1^2$ ie total of 4 params b_1, b_2, b_{12} and b_{11}

RSM-1

RESPONSE SURFACE METHODS (RSM)

BH^2 Chapter 15
M+R Chapter 12.9

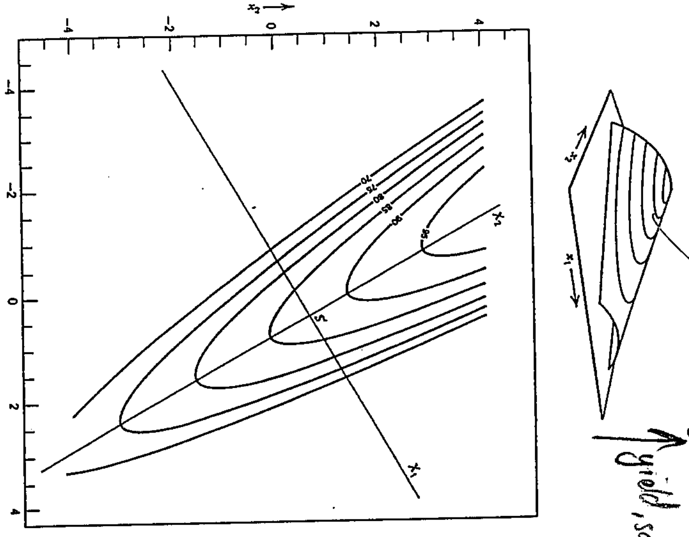
Empirical (data-driven) approach to process optimization.
processes that are too complex to model from first principles

1. Design experiment in region of interest
2. Build model from DOE : $\hat{y} = f(x_1, x_2 \cdot\cdot x_k)$
3. Use model to find new conditions $x_1, x_2 \cdot\cdot x_k$ that will improve a single response $\hat{y}_1$ or give good region for several responses.
4. Repeat steps 1., 2. and 3. until attain optimal conditions
   gradual move to a better operating point
   iterative: repeat and learn and improve

Problem with COST Approach

Ex. Maximize yield of a reaction by choice of - reaction time (t)
- reaction temp (T)

change a single variable at a time



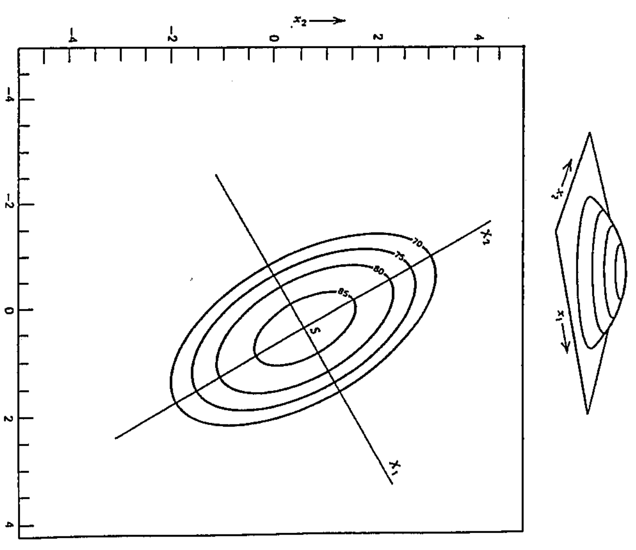
contours of constant yield

yield, say

(d)

FIGURE 15.7. (d) Example of a second-degree equation representing a rising ridge.

$\hat{y} = 82.71 + 8.80x_1 + 8.19x_2 - 6.95x_1^2 - 2.07x_2^2 - 7.59x_1x_2$
$\hat{y} - 87.69 = -9.02X_1^2 + 2.97X_2$



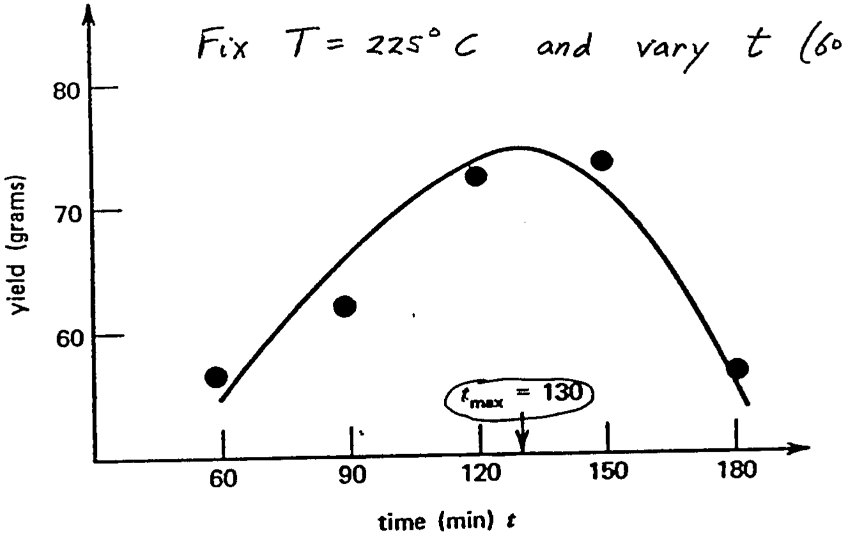
(a)

FIGURE 15.7. (a) Example of second-degree equations representing a simple maximum.

$\hat{y} = 83.57 + 9.39x_1 + 7.12x_2 - 7.44x_1^2 - 3.71x_2^2 - 5.80x_1x_2$
$\hat{y} - 87.69 = -9.02X_1^2 - 2.13X_2^2$

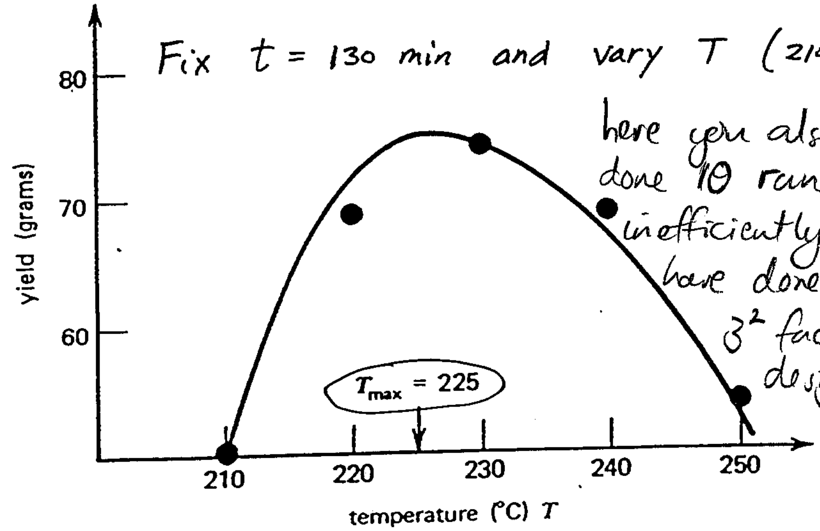
Other shapes: eg saddles - indicates interaction between variables
- local minima possible

Fix $T = 225°C$ and vary t (60 min → 180 min)



set of experiments: yield versus reaction time, temperature held fixed at 225°C.

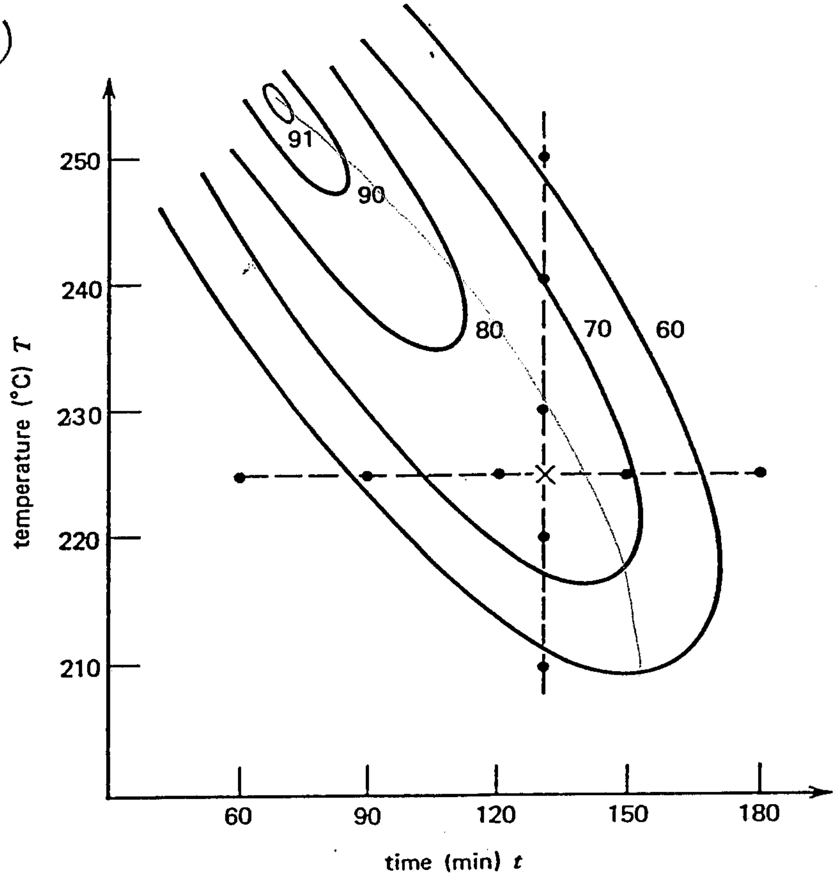
Fix $t = 130$ min and vary T (210° → 250°)



here you also have done 10 runs and inefficiently. Could have done a 3^2 factorial design instead

set of experiments: yield versus temperature, reaction time held fixed at 130 minutes.

IGURE 15.1. Hypothetical results from one-variable-at-a-time approach.

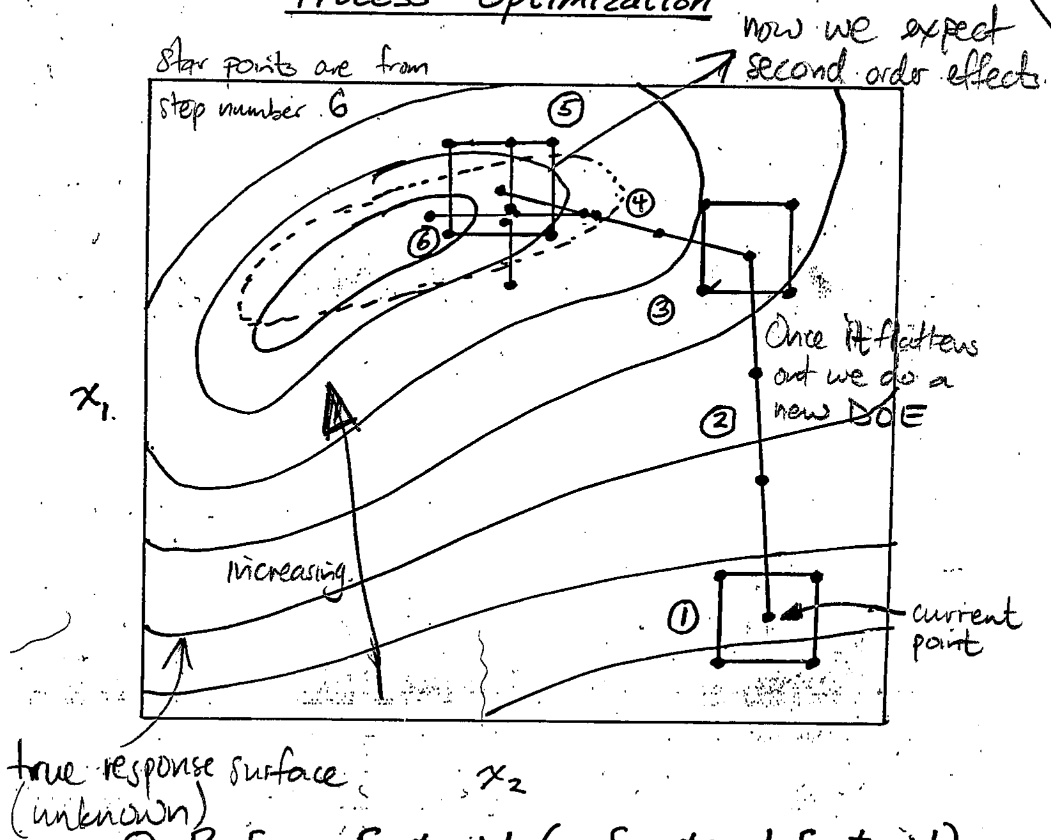


15.2. Possible true response surface representing yield versus reaction time and re, with points shown for one-variable-at-a-time approach.

eventually just land up on the ridge

Problem is caused by: not a simple process there is definitely $x_1 - x_2$ interaction here

RSM - Experimental Approach to Process Optimization



① Perform factorial (or fractional factorial) design about current operating conditions

Fit linear model

$$\hat{y} = \hat{\beta}_0 + \underset{\text{significant}}{\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2} + \underset{\text{small}}{\hat{\beta}_{12} x_1 x_2}$$

because we are in a linear region

units that you move in are SCALED units that you original chose to do designed experiments, ie units that lead to a square design

② Calculate direction of Steepest Ascent + perform experiments along this direction until <u>response</u> doesn't improve.

$\boxed{\text{Path of SA}}$: (If $\hat{\beta}_{12}$ term is small ie model is linear)

$$\frac{\partial \hat{y}}{\partial x_1} = \hat{\beta}_1$$

$$\frac{\partial \hat{y}}{\partial x_2} = \hat{\beta}_2$$

Move them both in a same proportion

∴ Move x_1 $\hat{\beta}_1$ units in direction of x_1 for every $\hat{\beta}_2$ units in direction of x_2

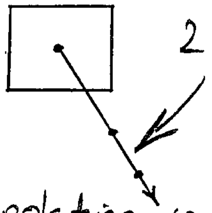
Unrelated to this specific example of temp & time vs. yield

$\boxed{\text{eg.}}$ $\hat{y} = 3.5 + 1.5 x_1 - 3.0 x_2 + 0 x_1 x_2$

Points along path of SA :

	x_1	x_2
Current pt. →	0	0
	1.0	-2.0
	1.5	-3.0
	3.0	-6.0

$2x_1 = x_2$



③ Lay down a new factorial design

extrapolating so you should do a factorial after a few steps

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2$$

④ If linear terms still large + $\hat{\beta}_{12}$ small again calculate direction of SA + perform experiments along it.

$\frac{\partial y}{\partial x_1} = \hat{\beta}_1 + \hat{\beta}_{12} x_2$ $\frac{\partial y}{\partial x_2} = \hat{\beta}_2 + \hat{\beta}_{12} x_1$

⑤ 3rd Factorial design.

→ Clear Lack of Fit of Linear Model

- Linear effects $\hat{\beta}_1, \hat{\beta}_2$ small
- Interaction term $\hat{\beta}_{12}$ large
- Check on curvature or quadratic effect $(\beta_{11} x_1^2 + \beta_{22} x_2^2$ terms$)$?

6 parameters from 6 runs

x_0	x_1	x_2	x_1^2	x_2^2	$x_1 x_2$
+	−	−	+	+	+
+	+	−	+	+	−
+	−	+	+	+	−
+	+	+	+	+	+
+	0	0	0	0	
+	0	0	0	0	

two ways to estimate

$$\hat{\beta}_0 = \frac{y_1+y_2+y_3+y_4}{4} = \bar{y}_f$$

$$\rightarrow \cancel{\text{avg.} + \hat{\beta}_{11} + \hat{\beta}_{22}}$$

$$\hat{y}(x=0) + \hat{\beta}_{11} + \hat{\beta}_{22}$$

$\bar{y}_{cp}$

If model were linear (ie $\hat{\beta}_{11} = \hat{\beta}_{12} = 0$) then $\hat{\beta}_0 = \bar{y}_f$ would be estimate of response at center of design.

$\therefore\ \bar{y}_f - \bar{y}_{cp} \rightarrow$ estimate of $(\hat{\beta}_{11} + \hat{\beta}_{22})$ (curvature)

⑥ If curvature and/or interaction large relative to main effects the add star points → 2nd order central composite design

linear in parameters, nonlinear in $\underline{x}$

⑥ Fit full second order model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2 + \hat{\beta}_{11} x_1^2 + \hat{\beta}_{22} x_2^2$$

now main effects are not really meaningful - do an overall quadratic optimization

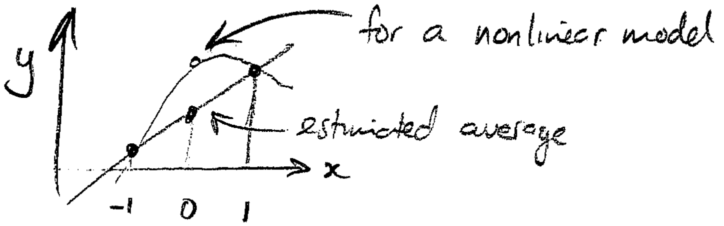
⑦ • Plot response contours.

2-D or 3-D plot routines in all statistical design software (MODDE)

- Examine response surface and move towards best conditions *

The dashed contour is an approximation made by the DOE software of the real contours

For $k > 3$ variables, use fractional factorials



If model were linear $\bar{y}$ = estimate of response at the centre point

→ compares the difference between actual centre points and our guess of the centre point, if it is linear model

→ if large difference ⇒ quadratic effects are important

Saddle points are tricky: improve by going in two directions ⇒ choose least cost option

DOE

Linear Empirical

- Factorial
- Fractional factorial
- Random block design

Nonlinear, mechanistic model

form known

now need to estimate the parameters in this model (still empirical really)

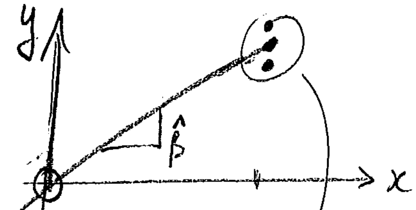
ie: which DOE will give precise parameter estimates

Optimal designs: ie $\max |X^T X|$

objective: good estimate of how precise we can get our parameter – minimize variance of $\hat{\beta}$

example: $y = \beta x$

$$\mathrm{var}(\hat{\beta}) = \frac{\sigma^2}{\Sigma x^2}$$



to minimize $\mathrm{var}(\hat{\beta}) \Rightarrow \max \sum x_u^2$

ie run all experiments at the maximum value of the x

form unknown

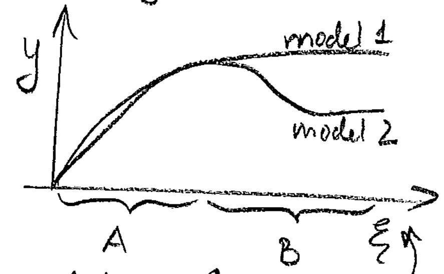
eg: Langmuir form

- diffusion step controlling
- reaction step controlling
- desorption step controlling

dependant on which on

Discriminate between them which one is most plausible

vast body of work on this

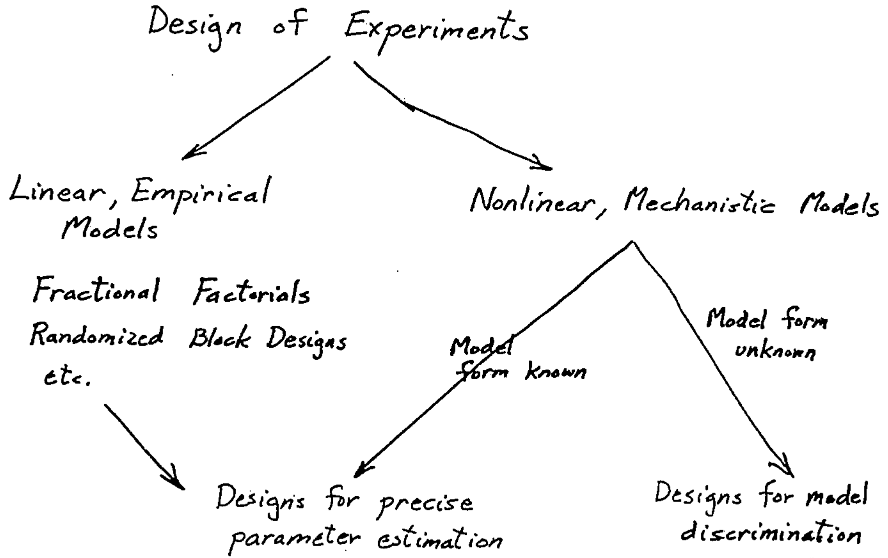


$\max_{\xi} (\hat{y}_1 - \hat{y}_2)^2$

one of the variables

should run experiments in part B, ie where the predictions deviate the most – put all models in maximum jeopardy

After this return to form known


Design of Experiments

Linear, Empirical Models

Fractional Factorials
Randomized Block Designs
etc.

Nonlinear, Mechanistic Models

Model form known

Model form unknown

Designs for precise parameter estimation

Optimal designs

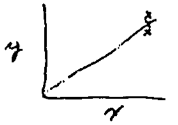
eg. D-optimal : Max $|X'X|$

$y = \beta x$

$Var(\hat{\beta}) = \frac{\sigma^2}{\Sigma x_u^2}$

$Max|X'X| \equiv Max\ \Sigma x_u^2$

$\equiv Min\ V\hat{a}r(\hat{\beta})$



Designs for model discrimination

$\underset{\xi}{Max} (\hat{y}_1 - \hat{y}_2)^2$

Assumptions: model is ~~perfect~~
- optimize to joint confidence region
- optimize for minimum variance in some appropriate sense

OPT-1

OPTIMAL DESIGNS

"Optimal Designs" are designs which optimize some objective function via the choice of design conditions $\underline{x}$ eg: settings of $\underline{x}$ our variables

Designs for Precise Parameter Estimation

Assume we know or have selected the structure of the model.

$$\left.\begin{array}{lll} \text{eg.} & \eta = \underline{X}\beta & (\text{linear}) \\ \text{or} & \eta = f(\underline{x}, \beta) & (\text{nonlinear}) \end{array}\right\} \text{model is already determined}$$

Consider linear model (in the parameters)

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

$$X = \begin{array}{c} \begin{array}{cccccc} 1 & x_1 & x_2 & x_1 x_2 & x_1^2 & x_2^2 \end{array} \\ \begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \end{array}$$

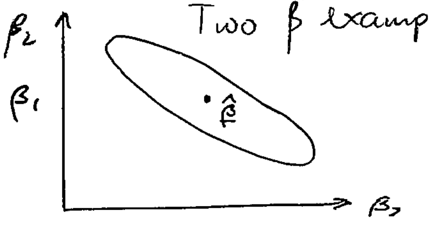
$Var(\hat{\beta}) = (X^T X)^{-1} \sigma^2$ variance/covariance matrix
P. Definite matrix

Joint Confidence Region given by ellipse

$$(\beta - \hat{\beta})^T (X^T X)(\beta - \hat{\beta}) = \overset{\longrightarrow \text{number of parameters estimated}}{p} s^2 F_\alpha(p, n-p)$$

quadratic equation in β_1 and β_2
want to minimise the area of this ellipse

Two $\hat{\beta}$ example



Example : Linear model $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$

Design "best" 4 experiments to give most precise estimates of β in D-optimal sense.

$$\underset{\substack{\text{exp1} \longrightarrow (x_{11}, x_{21})\\ \text{exp2} \longrightarrow (x_{12}, x_{22})\\ \text{exp3} \longrightarrow (x_{13}, x_{23})\\ \text{exp4} \longrightarrow (x_{14}, x_{24})}}{\text{Max}} |X^TX| = \text{Max} \begin{vmatrix} \overset{(1,1)}{4} & \overset{(1,2)}{\sum_{u=1}^{4} x_{1u}} & \overset{(1,3)}{\sum_{1}^{4} x_{2u}} & \overset{(1,4)}{\sum_{1}^{4} x_{1u}x_{2u}} \\ & \sum x_{1u}^2 & \sum x_{1u}x_{2u} & \sum x_{1u}^2 x_{2u} \\ \overset{(2,2)}{} & & \sum x_{2u}^2 & \sum x_{1u}x_{2u}^2 \\ \text{sym.} & & & \sum x_{1u}^2 x_{2u}^2 \end{vmatrix}$$

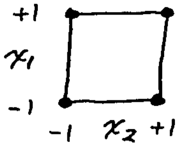
x_1 x_2

subject to constraints : $-1 \le x_1 \le 1$

$-1 \le x_2 \le 1$

Optimization routine → Search for 4 experiments

Solution → 2^2 Factorial



$$\text{Max} |X^TX| = \begin{vmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{vmatrix} = 64$$

Optimal 8 run design = Repeated 2^2

In general if have p parameters, D-opt design will place experiments at p locations, and then just start repeating these if want more than p experiments.

It will repeat the same design

This comes from the fact that one can tell upfront what your parameter variance will be before you do the experiments.

Objective : Design experiments to minimize some measure of uncertainty in $\hat{\beta}$.

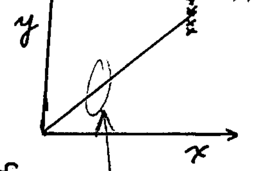
- eg. - Case of one parameter $\boxed{\eta = \beta x}$ assumes it is exact or correct

  $Var(\hat{\beta}) = \dfrac{\sigma^2}{\sum_u x_u^2}$ $u = 1 \ldots n$

  may want to include other points to check form of model

  Min. $Var(\hat{\beta})$ by maximizing $\sum_{u=1}^{n} x_u^2$

  $\Rightarrow$ Place all x_u's at upper limit of x.



running here we have more uncertainty in the slope

- More than one β.

  $Var(\hat{\beta}) = (X^TX)^{-1}\sigma^2$ $\neq f(y)$ it is totally independent of the response, only depends on where we place $\underline{x}$

  ↑ $p(p+1)/2$ terms !

Summarize all the uncertainty into one number

Need single measure of precision.

ie; we can predict our confidence intervals before we even do the experiments

A - Optimality : Min Trace $(X^TX)^{-1}\sigma^2$ dependant on the units used
average optimality

$$= \underset{\underline{x}_u}{\text{Min}} \sum_{i=1}^{p} Var(\hat{\beta}_i)$$

eg: kg vs g should scale them

= sum of variances of parameters

D - Optimality :
determinant-optimality

Volume (area) of elliptical joint confidence region for β $\propto |X^TX|^{-1/2}$ directly proportional to this value

↖ Determinant

∴ Design criterion : independant of scaling & units

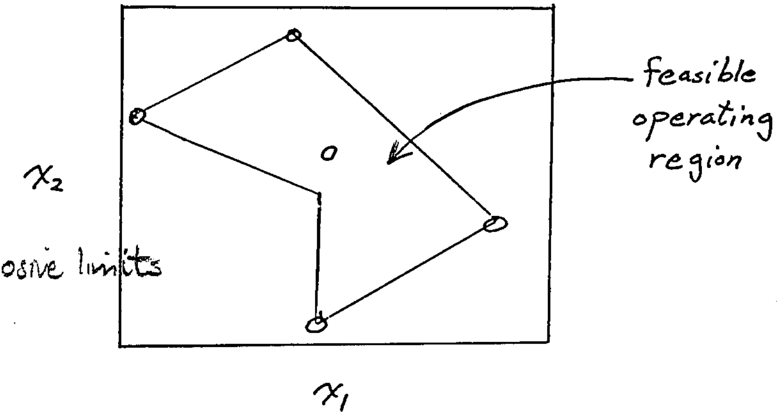
$$\boxed{\underset{\underline{x}_u,\, u=1,2\ldots}{\text{Max}} |X^TX|}$$

→ choice of experimental conditions

Important Areas for Optimal Designs

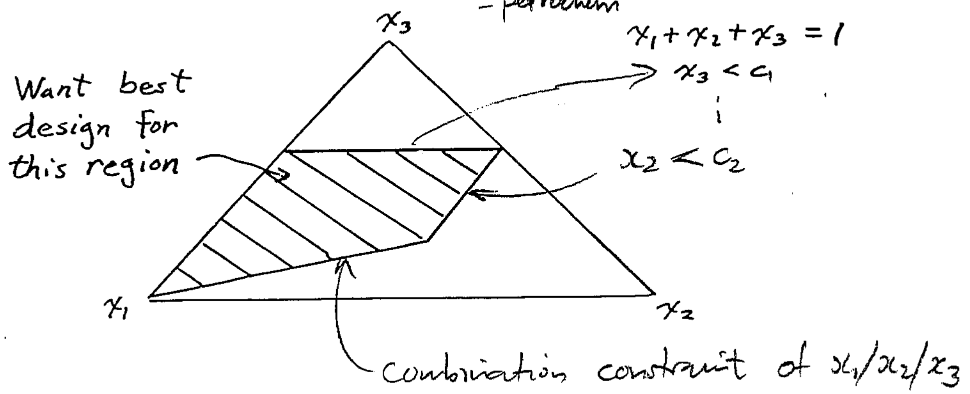
(1) Constrained Design Regions

Want to fit model $\eta = X\beta$ in a constrained experimental region



eg: explosive limits

Ex. Mixture Designs
- paints
- adhesives
- petrochem



(2) Sequential Designs

- Suppose have already performed n experiments.
- Now want to design $(n+1)$-st experiment ($\underline{x}_{n+1}$) (or set of m new experiments) which when added to existing n experiments will give us "best" estimates of β.

$$\text{Write} \quad \underset{(n+1)\times p}{\underline{X}_{n+1}} = \begin{bmatrix} \underline{X}_n \\ \underline{x}_{n+1} \end{bmatrix}$$

$\underline{X}_n$ ← already performed these known $n \times p$ conditions

$\underline{x}_{n+1}$ ← new row of conditions

ie: $x_{n+1} = 1 \quad x_1 \quad x_2 \quad x_1x_2 \quad x_1^2 \quad x_2^2$

$$\underset{\underline{x}_{n+1}}{\text{Max}} \left| X_{n+1}^T X_{n+1} \right| =$$

$$= \underset{\underline{x}_{n+1}}{\text{Max}} \left| \underbrace{X_n^T X_n}_{\text{fixed, known matrix}} + \underbrace{x_{n+1}^T x_{n+1}}_{\text{function of unknown conditions } x_{1,n+1},\, x_{2,n+1} \cdots} \right|$$

Commercial software makes (1) + (2) easy.

$$\left| \begin{bmatrix} X_n & x_{n+1} \end{bmatrix} \begin{bmatrix} X_n \\ x_{n+1} \end{bmatrix} \right. = X_n^T X_n + x_{n+1}^T x_{n+1}$$

ie; add new runs to a prexisting "bad" design to fix it up.
or add some runs to a fractional factorial

PTO

you have $\hat{\beta}_{34} \rightarrow x_3 x_4 + (x_1 x_5)$ interaction confounding

we want to separate this out to get

$$\rightarrow \hat{\beta}_{34} x_3 x_4 + \beta_{15} x_1 x_5$$

Add this to determine the next best runs

→ Software

(3) Optimal Designs for Nonlinear Models

Theoretical model (kinetics, mass transfer, ...)

$$\eta_u = \eta(\underline{\xi}_u, \underline{\beta})$$

$\underline{\beta}$ ← (p×1) vector of parameters

$\underline{\xi}_u$ ← vector of independent variables for u-th experiment (eg. Temp, Press, ...)

$$x_{iu} = \frac{\partial \eta(\underline{\xi}_u, \underline{\beta})}{\partial \beta_i} \qquad \begin{matrix} u = 1, 2, \cdots, n \\ i = 1, 2, \cdots, p \end{matrix}$$

$$\underset{n \times p}{\underline{X}} = \{x_{iu}\}$$

$$\underset{\underline{\xi}_1, \underline{\xi}_2 \cdots \underline{\xi}_n}{\text{Max}} |X^T X|$$

Dilemna : $x_{iu} = f(\text{unknown } \beta\text{'s})$

∴ Need to know $\underline{\beta}$ in order to design experiments!

Solution : Evaluate derivatives at current best estimates $\hat{\beta}$ ↑

Part III

STATISTICAL PROCESS CONTROL

CHAPTER 12

Fractional Factorial Designs at Two Levels

The number of runs required by a full 2^k factorial design increases geometrically as k is increased. It turns out, however, that when k is not small the desired information can often be obtained by performing only a fraction of the full factorial design. This chapter describes how suitable fractions can be generated and discusses their advantages and limitations.

12.1. REDUNDANCY

Consider a two-level design in seven variables. A complete factorial arrangement requires $2^7 = 128$ runs. From these runs 128 statistics can be calculated, which estimate the following effects:

		interactions					
average	main effects	2-factor	3-factor	4-factor	5-factor	6-factor	7-factor
1	7	21	35	35	21	7	1

Now the fact that all these effects can be estimated does not imply that they all are of appreciable size. There tends to be a certain hierarchy. In terms of absolute magnitude, main effects tend to be larger than two-factor interactions, which in turn tend to be larger than three-factor interactions, and so on. This fact relates directly to the properties of smoothness and similarity discussed earlier. (In particular, for quantitative variables the main effects and interactions can be associated with the terms of a Taylor series expansion of a response function. Ignoring, say, three-factor interactions corresponds to ignoring terms of third order in the Taylor expansion.)

It is often true, then, that at some point higher order interactions tend to become negligible and can properly be disregarded. Also, when a moderately large number of variables is introduced into a design, it often happens that some have no distinguishable effects at all. We can encompass both these ideas by saying that there tends to be redundancy in a 2^k design if k is not small—redundancy in terms of an excess number of interactions that can be estimated and sometimes in an excess number of variables that are studied. Fractional factorial designs exploit this redundancy. We begin by considering what effects can be estimated using only a half-fraction of a 2^5 factorial design.

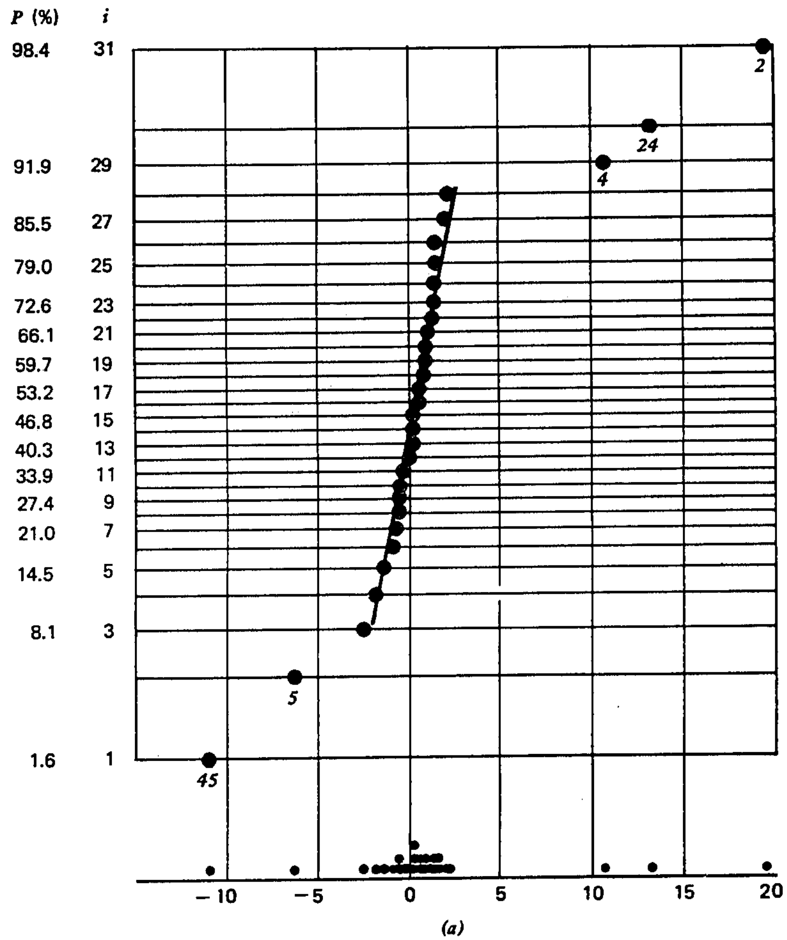


FIGURE 12.1. (*a*) Normal plot of effects from 2^5 factorial design, reactor example.

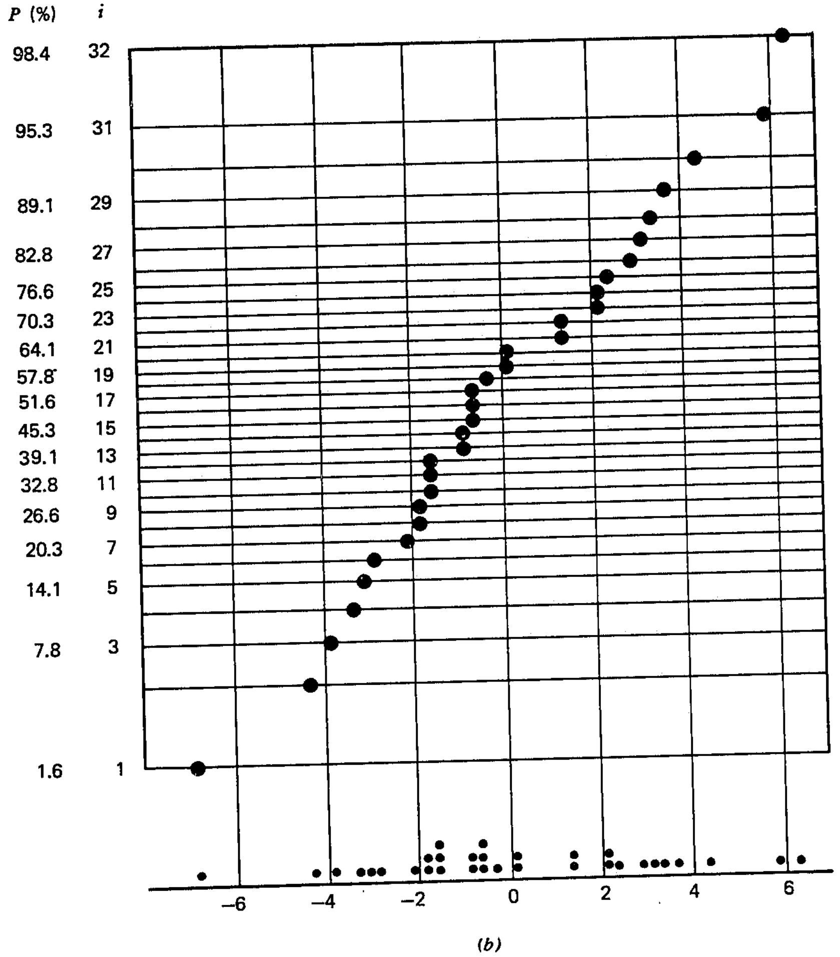


FIGURE 12.1. (*b*) Normal plot of residuals after eliminating *2*, *4*, *5*, *24*, and *45* from 2^5 factorial design, reactor example.

12.2. A HALF-FRACTION OF A 2^5 DESIGN: REACTOR EXAMPLE

Table 12.1*a* shows data from a complete 2^5 factorial design analyzed in Table 12.1*b*. Normal plots (Figure 12.1) indicate that over the ranges of the variables studied the main effects *2*, *4*, and *5* and interactions *24* and *45* are the only effects distinguishable from noise.

TABLE 12.1*a*. Results from 2^5 factorial design, reactor example

variable	−	+
1 feed rate (liters/min)	10	15
2 catalyst (%)	1	2
3 agitation rate (rpm)	100	120
4 temperature (°C)	140	180
5 concentration (%)	3	6

	variable					response (% reacted)
run	1	2	3	4	5	y
1	−	−	−	−	−	61
*2	+	−	−	−	−	53
*3	−	+	−	−	−	63
4	+	+	−	−	−	61
*5	−	−	+	−	−	53
6	+	−	+	−	−	56
7	−	+	+	−	−	54
*8	+	+	+	−	−	61
*9	−	−	−	+	−	69
10	+	−	−	+	−	61
11	−	+	−	+	−	94
*12	+	+	−	+	−	93
13	−	−	+	+	−	66
*14	+	−	+	+	−	60
*15	−	+	+	+	−	95
16	+	+	+	+	−	98
*17	−	−	−	−	+	56
18	+	−	−	−	+	63
19	−	+	−	−	+	70
*20	+	+	−	−	+	65
21	−	−	+	−	+	59
*22	+	−	+	−	+	55
*23	−	+	+	−	+	67
24	+	+	+	−	+	65
25	−	−	−	+	+	44
*26	+	−	−	+	+	45
*27	−	+	−	+	+	78
28	+	+	−	+	+	77
*29	−	−	+	+	+	49
30	+	−	+	+	+	42
31	−	+	+	+	+	81
*32	+	+	+	+	+	82

TABLE 12.1*b*. Analysis of 2^5 factorial design, reactor example

estimates of effects

average =	65.5		
1 =	− 1.375	*123* =	1.50
2 =	19.5	*124* =	1.375
3 =	− 0.625	*125* =	−1.875
4 =	10.75	*134* =	−0.75
5 =	− 6.25	*135* =	−2.50
		145 =	0.625
12 =	1.375	*235* =	0.125
13 =	0.75	*234* =	1.125
14 =	− 0.875	*245* =	−0.250
15 =	0.125	*345* =	0.125
23 =	0.875		
24 =	13.25	*1234* =	0.0
25 =	2.0	*1245* =	0.625
34 =	2.125	*2345* =	−0.625
35 =	0.875	*1235* =	1.5
45 =	−11.0	*1345* =	1.0
		12345 =	−0.25

The full 2^5 factorial requires 32 runs. Suppose that the experimenter had chosen to make only the 16 runs marked with asterisks in Table 12.1, so that only the data of Table 12.2 were available. When the 15 main effects and two-factor interactions are calculated from the reduced set of data in Table 12.2, they produce the estimates listed there, which are not very different from those obtained from the complete factorial design. Furthermore the normal plots of Figure 12.2 call attention to precisely the same effects: *2, 4, 24, 45* and *5*. Thus the essential information could have been obtained with only half the effort.

The 16-run design in Table 12.2 is called a *half-fraction*. It is often designated as a 2^{5-1} fractional factorial design since

$$\tfrac{1}{2}2^5 = 2^{-1}2^5 = 2^5 2^{-1} = 2^{5-1}$$

The notation tells us that the design accommodates five variables, each at two levels, but that only $2^{5-1} = 2^4 = 16$ runs are employed.

TABLE 12.2 Analysis of a half-fraction of the full 2^5 design: a 2^{5-1} fractional factorial design, reactor example

variable	−	+
1 feed rate (liters/min)	10	15
2 catalyst (%)	1	2
3 agitation rate (rpm)	100	120
4 temperature (°C)	140	180
5 concentration (%)	3	6

	design															response (% reacted)
run	**1**	**2**	**3**	**4**	**5**	**12**	**13**	**14**	**15**	**23**	**24**	**25**	**34**	**35**	**45**	y
17	−	−	−	−	+	+	+	+	−	+	+	−	+	−	−	56
2	+	−	−	−	−	−	−	−	−	+	+	+	+	+	+	53
3	−	+	−	−	−	−	+	+	+	−	−	−	+	+	+	63
20	+	+	−	−	+	+	−	−	+	−	−	+	+	−	−	65
5	−	−	+	−	−	+	−	+	+	−	+	+	−	−	+	53
22	+	−	+	−	+	−	+	−	+	−	+	−	−	+	−	55
23	−	+	+	−	+	−	−	+	−	+	−	+	−	+	−	67
8	+	+	+	−	−	+	+	−	−	+	−	−	−	−	+	61
9	−	−	−	+	−	+	+	−	+	+	−	+	−	+	−	69
26	+	−	−	+	+	−	−	+	+	+	−	−	−	−	+	45
27	−	+	−	+	+	−	+	−	−	−	+	+	−	−	+	78
12	+	+	−	+	−	+	−	+	−	−	+	−	−	+	−	93
29	−	−	+	+	+	+	−	−	−	−	−	−	+	+	+	49
14	+	−	+	+	−	−	+	+	−	−	−	+	+	−	−	60
15	−	+	+	+	−	−	−	−	+	+	+	−	+	−	−	95
32	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	82

estimates of effects
(assuming that three-factor and higher order interactions are negligible)

average =	65.25	*12* =	1.5
1 =	− 2.0	*13* =	0.5
2 =	20.5	*14* =	−0.75
3 =	0.0	*15* =	1.25
4 =	12.25	*23* =	1.50
5 =	− 6.25	*24* =	10.75
		25 =	1.25
		34 =	0.25
		35 =	2.25
		45 =	−9.50

16 effects for 2^{5-1} design

12.3. CONSTRUCTION AND ANALYSIS OF HALF-FRACTIONS: REACTOR EXAMPLE

How Were the16 Runs Chosen?

The 2^{5-1} design in Table 12.2 was constructed as follows:

1. A full 2^4 design was written for the four variables **1**, **2**, **3**, and **4**.
2. The column of signs for the **1234** interaction was written, and these were used to define the levels of variable **5**. Thus we made **5** = **1234**.

Exercise 12.1. By using this procedure, verify that the design obtained is the one given in Table 12.2.

The Anatomy of the Half-Fraction

At this point we seem to have gained something for nothing. It is natural to ask, Have we lost anything? Look again at the fractional factorial design of Table 12.2. We have made 16 runs and estimated 16 quantities: the mean, the 5 main effects, and the 10 two-factor interactions. But what happened to the remaining 16 effects we were able to estimate with the full factorial design—the 10 three-factor interactions, the 5 four-factor interactions, and the 1 five-factor interaction?

Let us try to estimate the value of the three-factor interaction *123*. Multiplying the signs in columns **1**, **2**, and **3**, we obtain the sequence (which, to save space, we write as a row rather than a column)

$$\mathbf{123} = - + + - + - - + - + + - + - - +$$

We notice that this is identical to

$$\mathbf{45} = - + + - + - - + - + + - + - - +$$

Thus **123** = **45**, and as a consequence the *123* and *45* interactions are confounded. Equivalently, in the fractional design the individual interactions *123* and *45* are said to be *aliases* of each other. Now suppose that we use the symbol l_{45} to denote the linear function of the observations which we used to estimate the *45* interaction:

$$l_{45} = \tfrac{1}{8}(-56 + 53 + 63 - 65 + 53 - 55 - 67 + 61 - 69 + 45 + 78$$
$$-93 + 49 - 60 - 95 + 82) = -9.50 \qquad (12.1)$$

We can call this the l_{45} *contrast* since it is the *difference* between two averages of eight results. Properly speaking, contrast l_{45} estimates the *sum* of the mean

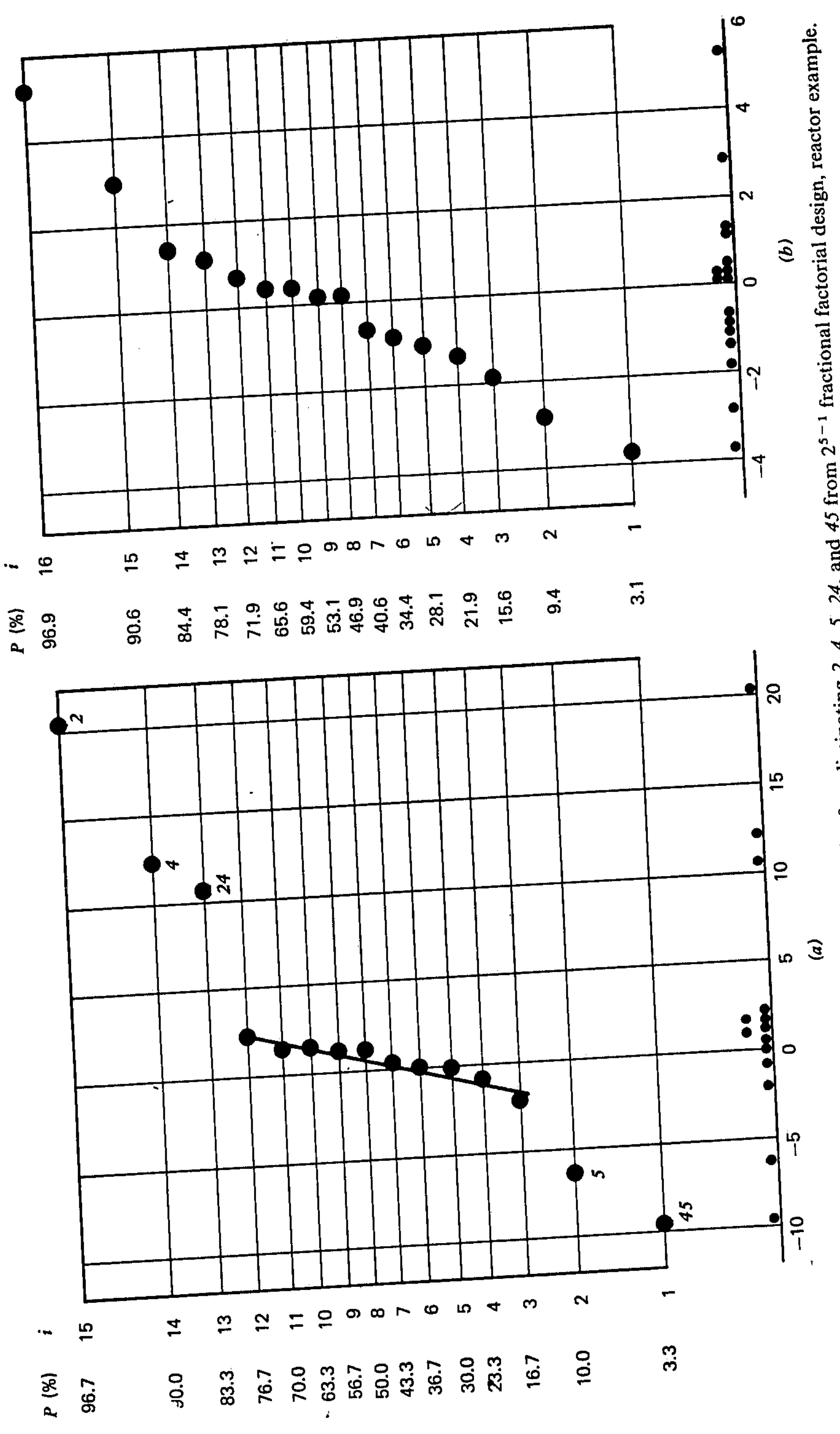


FIGURE 12.2. Normal plot of (*a*) effects and (*b*) residuals after eliminating 2, 4, 5, 24, and 45 from 2^{5-1} fractional factorial design, reactor example.

I, is actually obtained. The half-fraction is defined* by a single generator, so that the relation **I = 12345** also provides the *defining relation* of the design.

This defining relation is the key to the confounding pattern. For example, multiplying the defining relation on both sides by **1** yields

$$\mathbf{1} = \mathbf{2345}$$

In a similar way multiplying by **2** gives **2 = 1345** and so on to produce all the identities in the first column of Table 12.3.

The Complementary Half-Fraction

In the above example the generator **5 = 1234**, or, equivalently, **I = 12345**, produced the defining relation for the design. In other words, by generating a new column **5 = 1234** we obtained the half-fraction corresponding to the runs marked with asterisks in Table 12.1. The defining relation **I = 12345** provided by this generator immediately yields the confounding pattern of Table 12.3. The complementary half-fraction is generated by putting **5 = −1234**. We then obtain the half-fraction corresponding to the runs of the original 2^5 that are *not* marked with asterisks in Table 12.1. The defining relation for this design may be written as

$$\mathbf{I} = -\mathbf{12345}$$

In practice either half-fraction can equally well be used. For the data of Table 12.1 the complementary half-fraction would have given, for example,

$$l'_1 = -0.75 \to 1 - 2345$$
$$l'_2 = 18.50 \to 2 - 1345$$

Exercise 12.3. For the 16 runs in Table 12.1 that do *not* have asterisks, calculate the average and the 15 contrasts $l'_1, l'_2, \ldots, l'_{45}$. Show by making a normal plot that the conclusions that would result from this fraction would be similar to those obtained from the other one.

Answer: (average, *1, 2. 3, 4, 5, 12, 13, 14, 15, 23, 24, 25, 34, 35, 45*) = (65.75, −0.75, 18.5, −1.25, 9.25, −6.25, 1.25, 1.0, −1.0, −1.0, 0.25, 15.75, 2.75, 4.0, −0.5, −12.5).

Combining the Two Half-Fractions

Suppose that after completing one of the half-fractions the other was subsequently added, so that the whole factorial was available. Unconfounded estimates of all effects could then be obtained by analyzing the 32 runs as a full 2^5 factorial design run in two blocks of 16. The same result would be obtained by suitably adding and subtracting estimates from the two individual fractions. For example, we have

---

* When higher fractions are employed, there is more than one generator. For example, a quarter-fraction is defined by two generators. For more complicated fractions see Appendix 12A.

first fraction	second fraction
$l_2 = 20.5 \to 2 + 1345$	$l'_2 = 18.5 \to 2 - 1345$

whence

$$\begin{aligned} \tfrac{1}{2}(l_2 + l'_2) &= \tfrac{1}{2}(20.5 + 18.5) = 19.5 \to 2 \\ \tfrac{1}{2}(l_2 - l'_2) &= \tfrac{1}{2}(20.5 - 18.5) = 1.0 \to 1345 \end{aligned} \tag{12.6}$$

These values for *2* and *1345* agree with those given in Table 12.1 for the complete 2^5 design.

12.4. THE CONCEPT OF DESIGN RESOLUTION: REACTOR EXAMPLE

The 2^{5-1} fraction is called a *resolution V* design. Looking at the confounding pattern in Table 12.3, we see, for example, that $l_1 \to 1 + 2345$ and $l_{12} \to 12 + 345$. Thus main effects are confounded with four-factor interactions, and two-factor interactions with three-factor interactions.

In general, a design of resolution R is one in which no p-factor effect is confounded with any other effect containing less than $R - p$ factors. The resolution of a design is denoted by the appropriate Roman letter appended as a subscript. Thus we could refer to the design of Table 12.2 as a 2_V^{5-1} design. To illustrate:

1. A design of resolution R = III does not confound main effects with one another but does confound main effects with two-factor interactions.
2. A design of resolution R = IV does not confound main effects and two-factor interactions but does confound two-factor interactions with other two-factor interactions.
3. A design of resolution R = V does not confound main effects and two-factor interactions with each other, but does confound two-factor interactions with three-factor interactions, and so on.

In general, the *resolution* of a two-level fractional design is *the length of the shortest word in the defining relation.*

values of effects *45* and *123*. We indicate this by the notation $l_{45} \rightarrow 45 + 123$. If the columns of signs corresponding to all the other three-factor, four-factor, and five-factor interactions are obtained by multiplying signs, we get the results shown in Table 12.3.

TABLE 12.3. Confounding pattern and estimates from 2^{5-1} design of Table 12.2

$30+2=32=2^5$

relationship between column pairs	confounding pattern	estimate
1 = 2345	$l_1 \rightarrow 1 + 2345$	$l_1 = -2.0$
2 = 1345	$l_2 \rightarrow 2 + 1345$	$l_2 = 20.5$
3 = 1245	$l_3 \rightarrow 3 + 1245$	$l_3 = 0.0$
4 = 1235	$l_4 \rightarrow 4 + 1235$	$l_4 = 12.25$
5 = 1234	$l_5 \rightarrow 5 + 1234$	$l_5 = -6.25$
12 = 345	$l_{12} \rightarrow 12 + 345$	$l_{12} = 1.5$
13 = 245	$l_{13} \rightarrow 13 + 245$	$l_{13} = 0.5$
14 = 235	$l_{14} \rightarrow 14 + 235$	$l_{14} = -0.75$
15 = 234	$l_{15} \rightarrow 15 + 234$	$l_{15} = 1.25$
23 = 145	$l_{23} \rightarrow 23 + 145$	$l_{23} = 1.5$
24 = 135	$l_{24} \rightarrow 24 + 135$	$l_{24} = 10.75$
25 = 134	$l_{25} \rightarrow 25 + 134$	$l_{25} = 1.25$
34 = 125	$l_{34} \rightarrow 34 + 125$	$l_{34} = 0.25$
35 = 124	$l_{35} \rightarrow 35 + 124$	$l_{35} = 2.25$
45 = 123	$l_{45} \rightarrow 45 + 123$	$l_{45} = -9.50$
(I = 12345)	$[l_I \rightarrow \text{average} + \frac{1}{2}(12345)]$	$(l_I = 65.25)$

Exercise 12.2. As was done for columns **45** and **123**, verify that columns **24** and **135** are identical. Verify the identity of the other column pairs in Table 12.3.

A Justification for the Analysis

Evidently our earlier analysis would be justified if it could be assumed that effects of third and fourth order (represented by three-factor and four-factor interactions) could be ignored. In the reactor example the assumption was apparently justified. We shall see later that the analysis could also be justified on different and somewhat more subtle grounds (see the subsection entitled "An Alternative Rationale for the Half-Fraction Design in the Reactor Experiment").

How to Find the Confounding Patterns

In manipulating fractional factorials it is important to be able to obtain the confounding pattern for any given design. The method of associating like sign sequences is extremely tedious. Fortunately a much more expeditious route is available. To understand it remember the following four points:

1. Boldface numerals (e.g., **3** and **12**) refer to *columns* of plus and minus signs.
2. A product column is obtained by multiplication of the individual elements in the columns that make up that product. (The product column **124**, for instance, is obtained by multiplication of the individual elements in the corresponding columns, **1**, **2**, and **4**.)
3. Multiplying the elements in any column by a column of identical elements gives a column of plus signs, which is designated by the letter **I**, that is, $\mathbf{1} \times \mathbf{1} = \mathbf{1}^2 = \mathbf{I}$, $\mathbf{2}^2 = \mathbf{I}$, $\mathbf{3}^2 = \mathbf{I}$, $\mathbf{4}^2 = \mathbf{I}$, and so forth.
4. A contrast like l_{45} in Equation 12.1 is obtained by multiplying the observations by the appropriate plus and minus signs in column **45** and dividing by $N/2 = 8$ where N is the number of observations (16 in this case). Each quantity l is thus a contrast between two averages, each of $N/2$ observations. The single exception is $l_I = \bar{y}$, which is obtained by multiplying the observations by the column **I** of plus signs (i.e., summing the observations) and dividing the result by N (in this example $N = 16$).

Generator and Defining Relation

The 2^{5-1} design in Table 12.2 was constructed by setting

$$\mathbf{5} = \mathbf{1234} \tag{12.2}$$

This relation is called the *generator* of the design. Multiplying both sides by **5**, we obtain

$$\mathbf{5} \times \mathbf{5} = \mathbf{1234} \times \mathbf{5} \tag{12.3}$$

or

$$\mathbf{5}^2 = \mathbf{12345} \tag{12.4}$$

Thus the generator for the design can equivalently (and more conveniently) be written as

$$\mathbf{I} = \mathbf{12345} \tag{12.5}$$

This version of the identity is readily confirmed by multiplying together the elements in columns **1**, **2**, **3**, **4**, and **5**, and noting that a column of plus signs,

Resolutions of Some Half-Fractions

For any half-fraction the number of symbols on the right of the defining relation denotes the resolution of the design. Thus a 2^{5-1} half-fraction with defining relation **I** = $\pm$**12345** has resolution V. In Table 12.4 the 2^{3-1} half-fractions with defining relations **I** = $\pm$**123** have resolution III, and the 2^{4-1} fractions with defining relations **I** = $\pm$**1234** have resolution IV.

Half-Fractions of Highest Resolution

At the beginning of Section 12.3 we gave a procedure for constructing a 2^{5-1} design. In fact, it would have been possible to use *any* interaction or main effect column to accommodate the fifth variable. The choice we made yields a half-fraction with *highest possible resolution.* In general, to construct a 2^{k-1} fractional factorial design of highest possible resolution:

1. Write a full factorial design for the first $k - 1$ variables.
2. Associate the kth variable with plus or minus the interaction column **123...(k − 1)**.

Table 12.4 gives examples of 2^{3-1}_{III}, 2^{4-1}_{IV}, and 2^{5-1}_{V} half-fractions of this kind. The two 2^{3-1} half-fractions obtained by the above rule are shown geometrically in Table 12.4.

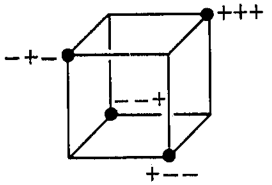
Exercise 12.4. Obtain the confounding pattern for a 2^{5-1} design generated by setting **5** = **123**. Discuss its properties. What is its resolution? Can you imagine circumstances in which it might be preferred to the resolution V design?

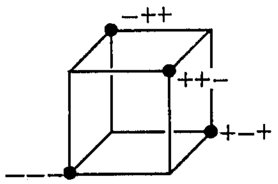
Partial answer: $l_1 \to 1 + 235$, $l_2 \to 2 + 135$, R = IV.

An Alternative Rationale for the Half-Fraction Design in the Reactor Experiment

Consider the 2^{5-1}_{V} half-fraction with **I** = **12345** given in Tables 12.2 and 12.4. Obviously (from its mode of construction), if we omit the fifth column of plus and minus signs from this design, we have a complete factorial in variables **1**, **2**, **3**, and **4**. But try omitting column **1** instead. There is now a complete factorial in variables **2**, **3**, **4**, and **5**! Indeed, a *complete* factorial in the remaining variables is obtained *whichever column is omitted.* We have already seen that the experimenter could justify the 2^{5-1} half-fraction on the assumption that three-factor, four-factor, and five-factor interactions could be ignored. An alternative justifying assumption is that at most only four of the five variables will produce detectable effects and the other will be essentially

TABLE 12.4. **Best half-fractions for $k = 3$, $k = 4$, and $k = 5$**





2^{3-1}_{III}
3 = 12
(I = 123)

1	2	3
−	−	+
+	−	−
−	+	−
+	+	+

2^{3-1}_{III}
3 = −12
(I = −123)

1	2	3
−	−	−
+	−	+
−	+	+
+	+	−

2^{4-1}_{IV}
4 = 123
(I = 1234)

1	2	3	4
−	−	−	−
+	−	−	+
−	+	−	+
+	+	−	−
−	−	+	+
+	−	+	−
−	+	+	−
+	+	+	+

2^{4-1}_{IV}
4 = −123
(I = −1234)

1	2	3	4
−	−	−	+
+	−	−	−
−	+	−	−
+	+	−	+
−	−	+	−
+	−	+	+
−	+	+	+
+	+	+	−

2^{5-1}_{V}
5 = 1234
(I = 12345)

1	2	3	4	5
−	−	−	−	+
+	−	−	−	−
−	+	−	−	−
+	+	−	−	+
−	−	+	−	−
+	−	+	−	+
−	+	+	−	+
+	+	+	−	−
−	−	−	+	−
+	−	−	+	+
−	+	−	+	+
+	+	−	+	−
−	−	+	+	+
+	−	+	+	−
−	+	+	+	−
+	+	+	+	+

2^{5-1}_{V}
5 = −1234
(I = −12345)

1	2	3	4	5
−	−	−	−	−
+	−	−	−	+
−	+	−	−	+
+	+	−	−	−
−	−	+	−	+
+	−	+	−	−
−	+	+	−	−
+	+	+	−	+
−	−	−	+	+
+	−	−	+	−
−	+	−	+	−
+	+	−	+	+
−	−	+	+	−
+	−	+	+	+
−	+	+	+	+
+	+	+	+	−

inert—it will have no detectable main effect or interaction with any other variable. On the assumption of one or more inert variables, the 2_V^{5-1} design will generate complete factorials in the remaining variables, *no matter which variables these are.*

In fact, our analysis for the reactor example suggests that only three of the variables had detectable effects: **2**, **4**, and **5** (catalyst, temperature, and concentration). Since variables **1** and **3** were effectively inert, we had a *replicated* 2^3 factorial in variables **2**, **4**, and **5**, and the results can be assembled as in Figure 12.3.

Factorials Embedded in Fractions:
The General Importance of the Concept of Resolution

In general, it can be shown that a fractional factorial design of resolution R contains complete factorials (possibly replicated) in every set of $R - 1$ variables. Suppose, then, that the experimenter has a number of candidate variables but believes that all but $R - 1$ of them (specific identity unknown) may have no detectable effects. Then, if he employs a design of resolution R and his conjecture is justified, he will have a complete factorial design in the effective variables. This idea is illustrated with the 2_{III}^{3-1} design in Figure 12.4, which projects a 2^2 pattern in every subspace of two dimensions.

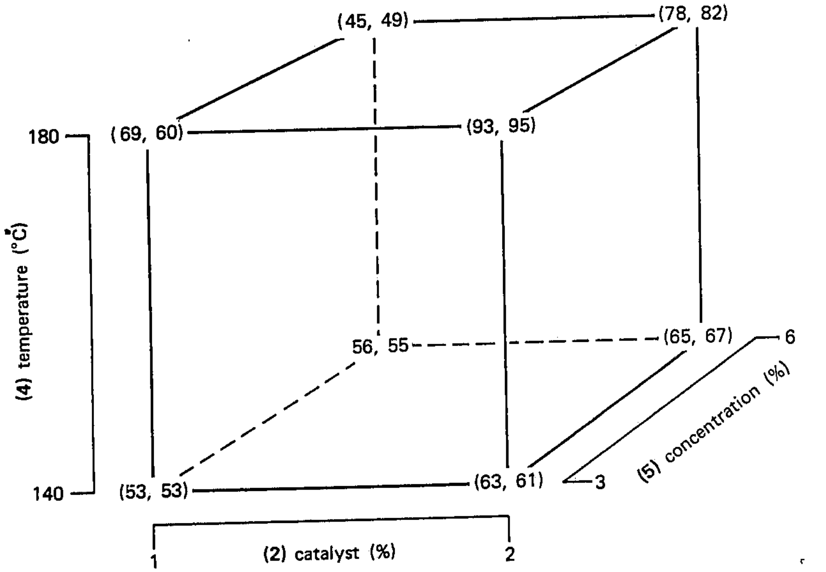


FIGURE 12.3. Data (% reacted) from a 2^{5-1} fraction, shown as replicated 2^3 factorial in variables *2*, *4*, and *5*, reactor example.

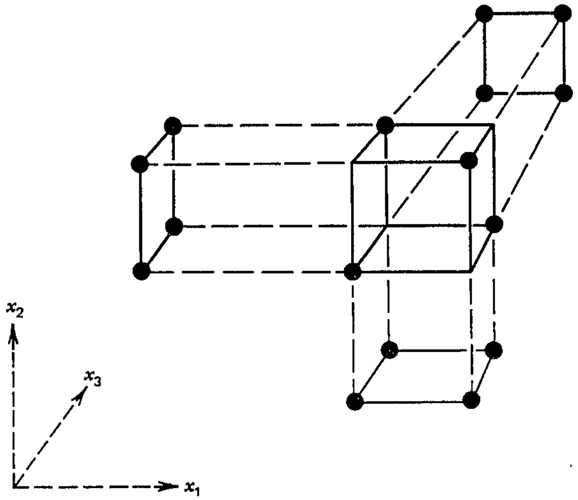


FIGURE 12.4. A 2_{III}^{3-1} design, showing projections into three 2^2 factorials.

Exercise 12.5. If a resolution R design gives a full factorial in every set of $R - 1$ variables, is it necessarily true that a full factorial is obtained in every subset containing fewer than $R - 1$ variables? *Answer*: Yes.

Exercise 12.6. A 2_V^{5-1} design gives full factorials in every subset of q variables. What is the value of q? *Answer*: 4, 3, 2, or 1 (for an example of $q = 3$ see Figure 12.3).

Economy in Experimentation Arising from the
Sequential Use of Fractional Designs

Suppose that an experimenter who can make his runs sequentially wishes to investigate five factors, each at two levels, and is contemplating a 2^5 design involving 32 runs. It is almost always better for him to run a half-fraction containing 16 runs first, analyze the results, and think about them. If necessary, he can always run the second fraction later to complete the full design. Frequently, however, the first half-fraction itself will allow him to proceed to the next stage of experimental iteration, which may involve, for example, the introduction of new variables or different levels of the old ones. Use of this sequential approach can thus greatly accelerate progress. It is worth noting that:

1. The experimenter should randomize within each fraction.
2. If eventually it is decided to run both fractions, these fractions will be randomized orthogonal blocks of the complete design.

3. No information will be "lost" except that concerning the interaction which is actually confounded with the block contrast.
4. The design run as two randomized fractions can give greater precision than the whole design run in random order because the block difference is eliminated.

Recapitulation

We began the chapter by discussing redundancy. It was pointed out that, for moderate k, a full factorial design frequently makes possible the estimation of many more effects than are detectably different from the noise. Sometimes these nondetectable effects are high-order interactions and sometimes they are all the effects associated with some inert variable or variables.

The fractional factorials discussed in this chapter are ideally suited to exploiting the probable existence of redundancy of one or both of these kinds for the following reason:

1. It can be arranged so that the confounding that occurs is between effects of high and low order,
2. A complete factorial design is available for whichever subset of $R - 1$ variables turns out to have appreciable effects.

In sequential experimentation, unless the total number of runs for a full or replicated factorial is needed to achieve sufficient precision, it is usually better to run fractional factorial designs. The fractions, used as building blocks, can build up to the full factorial design if this is necessary.

We now illustrate these ideas for designs of resolution III.

12.5. RESOLUTION III DESIGNS: BICYCLE EXAMPLE*

Suppose that the hypothetical data of Table 12.5 are times in seconds for a particular person to complete eight trial bicycle runs up a hill between fixed marks. These runs were performed in random order on eight successive days. The design is of resolution III and is a $\frac{8}{128} = \frac{1}{16}$ fraction of the full 2^7 factorial. Thus it is a 2_{III}^{7-4} design. (Note that $2^{7-4} = 2^7 2^{-4} = 2^{-4} 2^7 = \frac{1}{16} 2^7$.)

Table 12.6 gives the calculated contrasts. For example,

$$l_1 = \tfrac{1}{4}(-69 + 52 - 60 + 83 - 71 + 50 - 59 + 88) \qquad (12.7)$$

---

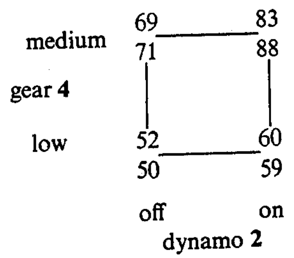
* This hypothetical example is an extension of the real one in Appendix 11A, but it is assumed now that both the rider and the bicycle are different.

TABLE 12.5. An eight-run experimental design for studying how time to cycle up a hill is affected by seven variables (I = 124, I = 135, I = 236, I = 1237).

run	seat up/down 1	dynamo off/on 2	handlebars up/down 3	gear low/medium 4 12	raincoat on/off 5 13	breakfast yes/no 6 23	tires hard/soft 7 123	time to climb hill (sec) y
1	−	−	−	+	+	+	−	69
2	+	−	−	−	−	+	+	52
3	−	+	−	−	+	−	+	60
4	+	+	−	+	−	−	−	83
5	−	−	+	+	−	−	+	71
6	+	−	+	−	+	−	−	50
7	−	+	+	−	−	+	−	59
8	+	+	+	+	+	+	+	88

TABLE 12.6. Calculated contrasts and abbreviated confounding pattern for data and design in Table 12.5

seat	$l_1 =$ 3.5	→ $1 + 24 + 35 + 67$
dynamo	$l_2 = \boxed{12.0}$	→ $2 + 14 + 36 + 57$
handlebars	$l_3 =$ 1.0	→ $3 + 15 + 26 + 47$
gear	$l_4 = \boxed{22.5}$	→ $4 + 12 + 56 + 37$
raincoat	$l_5 =$ 0.5	→ $5 + 13 + 46 + 27$
breakfast	$l_6 =$ 1.0	→ $6 + 23 + 45 + 17$
tires	$l_7 =$ 2.5	→ $7 + 34 + 25 + 16$
	($l_I =$ 66.5	→ average)



The table also gives an abbreviated* confounding pattern in which interactions between three or more factors have been ignored. Suppose that previous experience suggested that the standard deviation for repeated runs up the hill under the same conditions is about 3 seconds. Thus the calculated effects $l_1, l_2, \ldots, l_7$ have a standard error of about

$$\sqrt{\frac{3^2}{4}+\frac{3^2}{4}} = 2.1$$

Evidently only two contrasts, l_2 and l_4, are distinguishable from noise. Their values are circled in Table 12.6. The simplest interpretation of the results is that only two of the seven factors, the dynamo (**2**) and gear (**4**), exert a detectable influence, and they do so by way of their main effects. Having the dynamo on adds about 12 seconds to the time, and using medium gear instead of low gear adds about 22 seconds. On this interpretation we have in effect a replicated 2^2 design in the variables **2** and **4**, as indicated at the bottom of Table 12.6. There is, of course, some ambiguity in these conclusions. It is possible, for example, that l_4 is large, not because of a large main effect *4*, but because one or more of the interactions *12*, *56*, *37* are large. We see in Appendix 12B how sequential addition of further runs can resolve such ambiguities. However, for this example we suppose that the experimenter's knowledge of the nature of his bicycle suggests that the simpler explanation is likely to be right. The experimenter might well decide to proceed to the next stage of the investigation at this point.

---

* The method by which the confounding pattern has been obtained is given in Appendix 12A.

Because one use of resolution III designs is to determine the main effects of each of the factors, assuming that they do not interact, these arrangements have sometimes been called "main effect plans."

Embedded 2^2 Factorials in Resolution III designs

A resolution R design has a complete factorial (possibly replicated) in every subset of $R - 1$ variables. For the resolution III design of Table 12.5, for example, whichever two columns of the design are chosen, they form a complete 2^2 factorial replicated twice. Also notice what happens to the confounding pattern in Table 12.6 supposing that two variables, say **2** and **4**, are effective, and the rest, that is, **1**, **3**, **5**, **6**, and **7**, are essentially inert. Then all interactions and main effects containing these numbers vanish, $l_2 \to 2$, $l_4 \to 4$, and $l_1 \to 24$, and the remaining l's measure experimental error only.

Exercise 12.7. For the examples in Table 12.4, verify that any subset of $R - 1$ variables from a design of resolution R produces a full factorial design.

Construction of 2_{III}^{7-4} Design

The 2^{7-4} design in Table 12.5 can be constructed as follows:

1. Write a full factorial design for the three variables, **1**, **2**, and **3**.
2. Associate additional variables **4**, **5**, **6**, and **7** with all the interaction columns **12**, **13**, **23**, and **123**, respectively.

The design is obtained by associating every available contrast with a variable and is therefore sometimes called a *saturated design.**

---

* It is actually possible to construct supersaturated designs, but we do not recommend them in ordinary circumstances.

Other 2^{7-4} *Fractions*

In Table 12.5 a one-sixteenth fraction of a full 2^7 factorial design is shown. How can the other one-sixteenth fractions that make up the full factorial design be generated? The first design was generated by setting

$$\mathbf{4} = +\mathbf{12} \quad \mathbf{5} = +\mathbf{13} \quad \mathbf{6} = +\mathbf{23} \quad \mathbf{7} = +\mathbf{123} \qquad (12.8)$$

but, for example, we could equally well have used

$$\mathbf{4} = -\mathbf{12} \quad \mathbf{5} = +\mathbf{13} \quad \mathbf{6} = +\mathbf{23} \quad \mathbf{7} = +\mathbf{123} \qquad (12.9)$$

This gives a different one-sixteenth fraction, which is shown in Table 12.7 with further hypothetical data on times to cycle up the hill. Note that none of the runs in this new design is the same as any of those in the preceding design. Calculated contrasts for this design are shown in Table 12.8.

TABLE 12.7. A second 2_{III}^{7-4} fractional factorial design with times to cycle up a hill (I = − 124, I = 135, I = 236, I = 1,237).

run	seat 1	dynamo 2	handlebars 3	gear 4 −12	raincoat 5 13	breakfast 6 23	tires 7 123	time to climb hill (sec)
9	−	−	−	−	+	+	−	47
10	+	−	−	+	−	+	+	74
11	−	+.	−	+	+	−	+	84
12	+	+	−	−	−	−	−	62
13	−	−	+	−	−	−	+	53
14	+	−	+	+	+	−	−	78
15	−	+	+	+	−	+	−	87
16	+	+	+	−	+	+	+	60

What is the confounding pattern for the new fraction? Notice that the new fraction was obtained by switching signs for variable **4** in the first design (variable **4** was associated with −**12** instead of +**12**). The abbreviated confounding pattern for this new fraction may be obtained, therefore, by switching signs in the confounding pattern of Table 12.6. This gives the confounding pattern in Table 12.8.

For this set of data the contrasts calculated from the second fraction confirm the conclusions from the first fraction.

TABLE 12.8. Calculated contrasts and abbreviated confounding pattern for second design in bicycle experiment

$l'_1 = \ \ 0.8 \rightarrow 1 - 24 + 35 + 67$
$l'_2 = 10.2 \rightarrow 2 - 14 + 36 + 57$
$l'_3 = \ \ 2.7 \rightarrow 3 + 15 + 26 - 47$
$l'_4 = 25.2 \rightarrow 4 - 12 - 56 - 37$ (i.e., $l'_{-4} = -25.2 \rightarrow -4 + 12 + 56 + 37$)
$l'_5 = -1.7 \rightarrow 5 + 13 - 46 + 27$
$l'_6 = \ \ 2.2 \rightarrow 6 + 23 - 45 + 17$
$l'_7 = -0.7 \rightarrow 7 - 34 + 25 + 16$

The Sixteen Different Fractions

In all there are 16 different ways of allocating signs to the four generators:

$$\mathbf{4} = \pm\mathbf{12}, \qquad \mathbf{5} = \pm\mathbf{13}, \qquad \mathbf{6} = \pm\mathbf{23}, \qquad \mathbf{7} = \pm\mathbf{123} \qquad (12.10)$$

Thus appropriate sign switching in columns* **4**, **5**, **6**, and **7** of Table 12.5 produces 16 fractional factorial designs which together make up the complete 2^7 factorial design. Corresponding sign switching in Table 12.6 produces the 16 different confounding patterns.

Designing Two Fractions

Consider again the bicycle example. Suppose that the 16 results from the two 2_{III}^{7-4} fractionals were considered together. What conclusions could be drawn? Combining the results from Tables 12.6 and 12.8, we obtain Table 12.9.

Conclusions would now be somewhat more certain. In particular, the large main effect of factor **4** (gear) is now estimated free of bias from two-factor interactions, and has a value close to that conjectured earlier. The joint effect of the string of interactions *12 + 56 + 37* can now be estimated separately from the main effect *4*, and it is shown to be small. Most interestingly, all the two-factor interactions involving the important variable 4 are now *free of aliases*. (Of course we continue to assume all three-factor and higher order interactions to be zero.) For this particular set of data, however, none of these two-factor interactions is distinguishable from noise. Factor **2** (dynamo), somewhat less aliased than before, is showing an effect similar to that previously conjectured.

---

* The reader can confirm by experimentation that switching signs in other columns of the design only produces one or another of these basic 16 fractions. However, the *order* in which the runs appear can be different.

TABLE 12.9. Analysis of complete set of 16 runs, combining the results of the two fractions, bicycle example

seat	$\frac{1}{2}(l_1 + l'_1) = \frac{1}{2}(3.5 + 0.8)$	$= \quad 2.2 \rightarrow 1 + 35 + 67$
dynamo	$\frac{1}{2}(l_2 + l'_2) = \frac{1}{2}(12.0 + 10.2)$	$= \quad 11.1 \rightarrow 2 + 36 + 57$
handlebars	$\frac{1}{2}(l_3 + l'_3) = \frac{1}{2}(1.0 + 2.7)$	$= \quad 1.9 \rightarrow 3 + 15 + 26$
gear	$\frac{1}{2}(l_4 + l'_4) = \frac{1}{2}(22.5 + 25.2)$	$= \quad 23.9 \rightarrow 4$
raincoat	$\frac{1}{2}(l_5 + l'_5) = \frac{1}{2}(0.5 - 1.7)$	$= \quad -0.6 \rightarrow 5 + 13 + 27$
breakfast	$\frac{1}{2}(l_6 + l'_6) = \frac{1}{2}(1.0 + 2.2)$	$= \quad 1.8 \rightarrow 6 + 23 + 17$
tires	$\frac{1}{2}(l_7 + l'_7) = \frac{1}{2}(2.5 - 0.7)$	$= \quad 0.9 \rightarrow 7 + 25 + 16$
	$\frac{1}{2}(l_1 - l'_1) = \frac{1}{2}(3.5 - 0.8)$	$= \quad 1.3 \rightarrow 24$
	$\frac{1}{2}(l_2 - l'_2) = \frac{1}{2}(12.0 - 10.2)$	$= \quad 0.9 \rightarrow 14$
	$\frac{1}{2}(l_3 - l'_3) = \frac{1}{2}(1.0 - 2.7)$	$= \quad -0.9 \rightarrow 47$
	$\frac{1}{2}(l_4 - l'_4) = \frac{1}{2}(22.5 - 25.2)$	$= \quad -1.4 \rightarrow 12 + 56 + 37$
	$\frac{1}{2}(l_5 - l'_5) = \frac{1}{2}(0.5 + 1.7)$	$= \quad 1.1 \rightarrow 46$
	$\frac{1}{2}(l_6 - l'_6) = \frac{1}{2}(1.0 - 2.2)$	$= \quad -0.6 \rightarrow 45$
	$\frac{1}{2}(l_7 - l'_7) = \frac{1}{2}(2.5 + 0.7)$	$= \quad 1.6 \rightarrow 34$

Sequential Use of Highly Fractionated Designs

The preceding example illustrates a useful application of highly fractionated designs as sequential building blocks. Additional fractions may be selected to resolve ambiguities, which knowledge of the variables and data available so far suggest may be of importance. We explore two important applications of this idea. The reader can devise others to suit particular circumstances.

Addition of a Second Fraction to De-alias Any One Main Effect and All Its Associated Two-Factor Interactions

Consider the two fractions used in the bicycle experiment. The largest effect obtained from the first set of eight runs was associated with the choice of gear (variable **4**). It might have been argued, therefore, that if further runs were to be made, they could best be employed to de-alias **4** and all the interactions of other variables with **4**.

Table 12.9 shows that by adding a second fraction in which the sign of variable **4** has been switched, a design of 16 runs possessing the desired property is obtained. This ability to de-alias one effect and all its two-factor property is obtained. This ability to de-alias one effect and all its two-factor interactions by adding a second fraction with the appropriate column of signs switched is a handy device for the sequential use of these designs.

Adding a Second Fraction to De-alias All Main Effects

Consider Table 12.5 again, and suppose that a different second fraction is added in which signs are switched in *all* the columns. Then for the new fraction the first two rows in the confounding pattern (obtained by switching signs in Table 12.6) are

$$\begin{aligned} l'_1 &\rightarrow 1 - 24 - 35 - 67 \qquad (l'_{-1} \rightarrow -1 + 24 + 35 + 67) \\ l'_2 &\rightarrow 2 - 14 - 36 - 57 \qquad (l'_{-2} \rightarrow -2 + 14 + 36 + 57) \end{aligned} \tag{12.11}$$

By combining this second fraction with the original fraction, we obtain

$$\begin{aligned} \tfrac{1}{2}(l_1 + l'_1) &\rightarrow 1, \qquad \tfrac{1}{2}(l_1 - l'_1) \rightarrow 24 + 35 + 67 \\ \tfrac{1}{2}(l_2 + l'_2) &\rightarrow 2, \qquad \tfrac{1}{2}(l_2 - l'_2) \rightarrow 14 + 36 + 57 \end{aligned} \tag{12.12}$$

and so on.

This way of augmenting the design yields *all* main effects clear of all two-factor interactions, but the two-factor interactions themselves are still confounded in groups of three. An example of the use of this sequence is given in Section 13.3.

Exercise 12.8. Show that the second fraction obtained above by switching all signs may also be obtained (with runs in a different order) by switching signs in columns **4**, **5**, **6**, and **7** only. Can you find other ways to reproduce the second fraction? Explain the equivalences you find.

General Construction of Resolution III Designs

Resolution III designs for $2^k - 1$ variables may be obtained by saturating a 2^k factorial with additional variables. For example, to construct a saturated 16-run design in 15 variables first write a full factorial design for four variables and then associate the extra variables **5**, **6**, ..., **15** with the 11 interaction columns **12**, **13**, **14**, **23**, **24**, **34**, **123**, **124**, **134**, **234**, and **1234**, respectively. The resulting design is a 2_{III}^{15-11} fractional factorial design for 15 variables in 16 runs.

Exercise 12.9. Construct a two-level fractional factorial design for 31 variables in 32 runs. This is a 2^{k-p} design; what values do k and p have? *Answer*: $k = 31, p = 26$.

Exercise 12.10. Indicate how you could construct a 2^{63-57} fractional factorial design. Is this a saturated design? *Answer*: Yes.

Useful designs may be obtained by appropriately deleting columns from the saturated designs. For example, dropping columns **4** and **7** from the design matrix for a 2^{7-4} design yields a 2^{5-2} design, the defining relation for which can be obtained from that for the 2^{7-4} design by deleting all words containing **4** and **7**. The variables to be dropped are selected so as to obtain the most satisfactory alias arrangement.

Plackett and Burman Saturated Designs

The saturated fractional factorial designs have the following orthogonal* property: if we take any two columns, then, corresponding to the $N/2$ plus signs in the first column, there will be $N/4$ plus and $N/4$ minus signs in the second column, and similarly for the minus signs in the first column. Provided that all interactions are negligible, designs with this property allow unbiased estimation of all main effects of $N - 1$ variables with smallest possible variance. The fractional factorials so far discussed are available only if N is a power of 2. Plackett and Burman (1946) have obtained arrangements with this same orthogonal property when N is a multiple of 4. For example, their design for $k = 11$ factors in $N = 12$ runs is shown in Table 12.10. The fashion in which two-factor interactions confound main effects for most Plackett and Burman designs is complicated. However, fold-over pairs of any such orthogonal design are of resolution IV (see Box and Wilson, 1951).

TABLE 12.10. **Plackett and Burman design for study of 11 factors in 12 runs**

	variable										
run	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	$\overline{\mathbf{10}}$	$\overline{\mathbf{11}}$
1	+	−	+	−	−	−	+	+	+	−	+
2	+	+	−	+	−	−	−	+	+	+	−
3	−	+	+	−	+	−	−	−	+	+	+
4	+	−	+	+	−	+	−	−	−	+	+
5	+	+	−	+	+	−	+	−	−	−	+
6	+	+	+	−	+	+	−	+	−	−	−
7	−	+	+	+	−	+	+	−	+	−	−
8	−	−	+	+	+	−	+	+	−	+	−
9	−	−	−	+	+	+	−	+	+	−	+
10	+	−	−	−	+	+	+	−	+	+	−
11	−	+	−	−	−	+	+	+	−	+	+
12	−	−	−	−	−	−	−	−	−	−	−

12.6. RESOLUTION IV DESIGNS: INJECTION MOLDING EXAMPLE

We have seen that for designs of resolution V main effects are confounded only with four-factor interactions, and two-factor interactions only with three-factor interactions. Full factorial designs are generated by every subset

---

* If the level of the ith variable is represented by $x_i = \pm 1$ and that of the jth variable by $x_j = \pm 1$, then $\Sigma x_i = 0$, $\Sigma x_j = 0$, and $\Sigma x_i x_j = 0$ for every i and j.

of four variables. Designs of resolution III introduce much more serious confounding, with main effects having two-factor interactions as aliases. For these designs full factorial designs exist for every subset of two variables. Designs of resolution IV occupy an intermediate position. No main effect is confounded with any two-factor interaction, but two-factor interactions *are* confounded with each other. For these designs full factorial designs exist for every subset of three variables.

An Experiment on Injection Molding

In an injection molding experiment (Table 12.11) eight variables were studied in a 2_{IV}^{8-4} (a $\frac{1}{16}$ replicate of a 2^8 factorial of resolution IV). The normal plots, shown in Figure 12.5, suggest that the linear contrasts l_3, l_{15}, and l_5 are distinguishable from the noise. The largest remaining effect is l_8. The confounding pattern, assuming negligible interaction between three or more factors, is shown in Table 12.12. It seems likely that main effects associated with holding pressure (**3**) and booster pressure (**5**) exist. Also, the interactions most likely to explain the large size of l_{15} are perhaps *15* and *38*, since these involve factors **3** and **5**, which have large main effects. It is, however, possible that interactions exist between factors that have no main effects. Without further information the situation is uncertain. One way to proceed is to choose a further fraction of eight or 16 runs designed to resolve the ambiguity. However, in this particular example the large size of l_{15} suggested that the problem might be resolved with even fewer than eight runs. We show in Appendix 12B how four additional runs were chosen and used to discover and estimate the responsible interaction.

Construction of the Resolution IV Design by "Folding Over" a Resolution III Design

The sixteen-run 2_{IV}^{8-4} design in Table 12.11 was constructed as follows. The eight-run 2_{III}^{7-4} design was first written as in Table 12.5 for the seven variables **1**, **2**, **3**, . . . , **7**. A further column labeled **8** and consisting entirely of plus signs was then added. The remaining eight runs were obtained by switching all signs in the first set of eight runs. Thus run 9 was obtained by switching all signs in run 1 and so on.

The Alias Pattern

The alias pattern for the folded-over design given in Table 12.11 can be obtained from that of the resolution III design (Table 12.6) by the following argument. Suppose that we compute for the first set of eight runs

$$l_1 = \tfrac{1}{4}(-y_1 + y_2 \dots + y_8)$$

and for the second set of eight runs

$$-\bar{l}'_1 = \tfrac{1}{4}(-y_9 + y_{10} \cdots + y_{16})$$

Then using Table 12.6

$\bar{l}_1 \to 1 + 18 + 24 + 35 + 67$ and $-\bar{l}'_1 = -1 + 18 + 24 + 35 + 67$

Now the contrast l_1 for the complete set of 16 runs is

$$l_1 = \tfrac{1}{8}(-y_1 + y_2 + \cdots + y_8 + y_9 - y_{10} \cdots - y_{16}) = \tfrac{1}{2}(\bar{l}_1 + \bar{l}'_1)$$

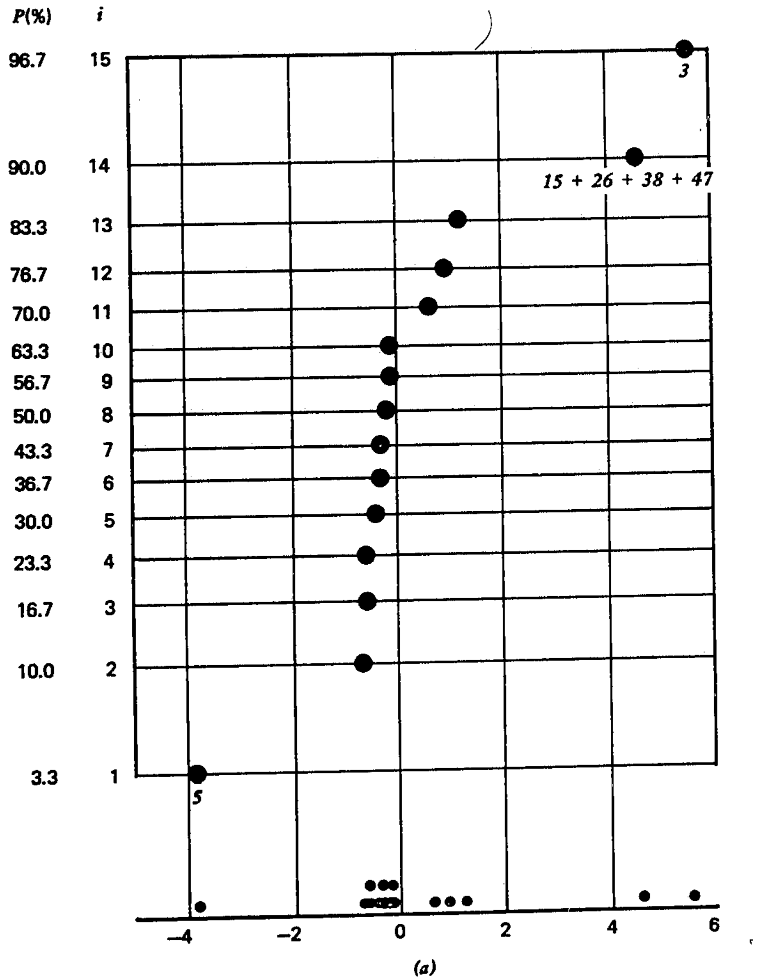


FIGURE 12.5. (*a*) Normal plot of contrasts, injection molding example.

Similarly for the contrast associated with the interaction *18* it is

$$l_{18} = \tfrac{1}{8}(-y_1 + y_2 + \cdots - y_8 - y_9 + y_{10} \cdots + y_{16}) = \tfrac{1}{2}(\bar{l}_1 - \bar{l}'_1).$$

Thus $l_1 \to 1$ and $l_{18} \to 18 + 24 + 35 + 67$. The same argument applied to the remaining contrasts yields the confounding pattern of Table 12.12. A more complete discussion is given in Appendix 12A.

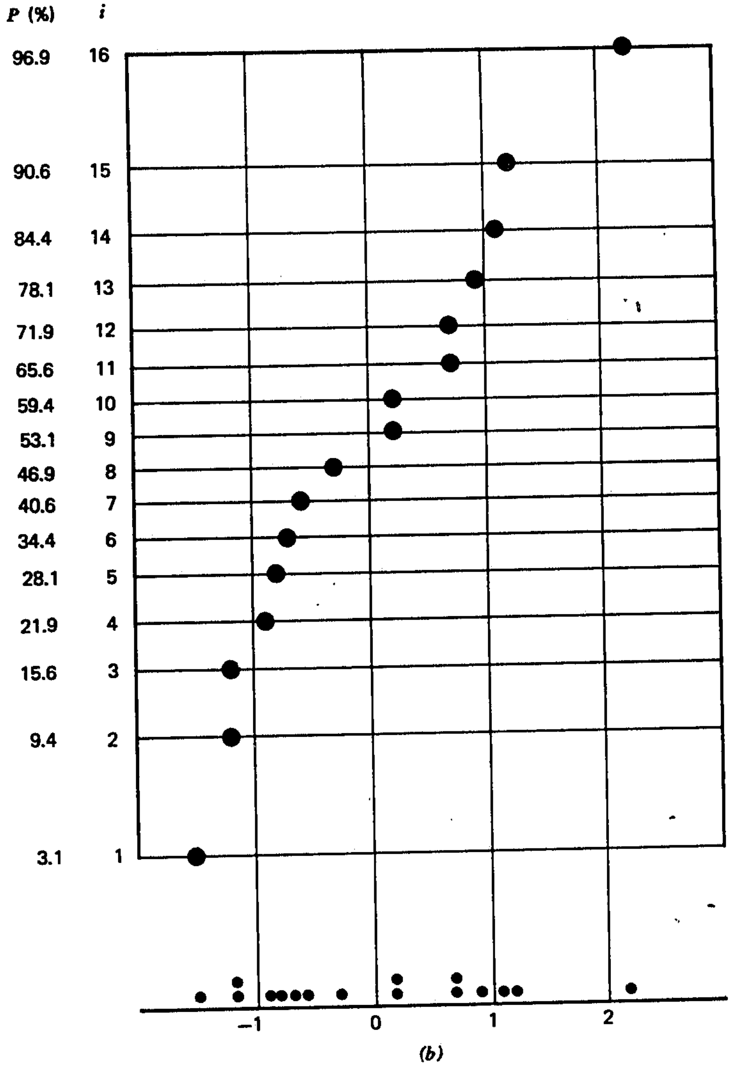


FIGURE 12.5. (*b*) Normal plot of residuals 2_{IV}^{8-4} design, injection molding example.

TABLE 12.11. A 2^{8-4} resolution IV design, molding example (I = 1248, I = 1358, I = 2368, I = 1237).

	mold temperature	moisture content	holding pressure	cavity thickness	booster pressure	cycle time	gate size	screw speed	shrinkage
run	1	2	3	4	5	6	7	8	y
1	−	−	−	+	+	+	−	+	14.0
2	+	−	−	−	−	+	+	+	16.8
3	−	+	−	−	+	−	+	+	15.0
4	+	+	−	+	−	−	−	+	15.4
5	−	−	+	+	−	−	+	+	27.6
6	+	−	+	−	+	−	−	+	24.0
7	−	+	+	−	−	+	−	+	27.4
8	+	+	+	+	+	+	+	+	22.6
9	+	+	+	−	−	−	+	−	22.3
10	−	+	+	+	+	−	−	−	17.1
11	+	−	+	+	−	+	−	−	21.5
12	−	−	+	−	+	+	+	−	17.5
13	+	+	−	−	+	+	−	−	15.9
14	−	+	−	+	−	+	+	−	21.9
15	+	−	−	+	+	−	+	−	16.7
16	−	−	−	−	−	−	−	−	20.3

TABLE 12.12. Calculated contrasts with their expected values: interactions between three or more factors ignored, molding example

$l_1 = -0.7 \rightarrow 1$

$l_2 = -0.1 \rightarrow 2$

$l_3 = 5.5 \rightarrow 3$

$l_4 = -0.3 \rightarrow 4$

$l_5 = -3.8 \rightarrow 5$

$l_6 = -0.1 \rightarrow 6$

$l_7 = 0.6 \rightarrow 7$

$l_8 = 1.2 \rightarrow 8$

$l_{12} = -0.6 \rightarrow 12 + 37 + 48 + 56$

$l_{13} = 0.9 \rightarrow 13 + 27 + 46 + 58$

$l_{14} = -0.4 \rightarrow 14 + 28 + 36 + 57$

$l_{15} = 4.6 \rightarrow 15 + 26 + 38 + 47$

$l_{16} = -0.3 \rightarrow 16 + 25 + 34 + 78$

$l_{17} = -0.2 \rightarrow 17 + 23 + 68 + 45$

$l_{18} = -0.6 \rightarrow 18 + 24 + 35 + 67$

average = 19.75

Alternative 2^{8-4} Fractions

Sixteen different 2_{IV}^{8-4} fractions are members of the family making up the complete 2^8 design. Individual members of the family may be generated by sign switching. Exactly as with the resolution III designs, the switching of signs in one or more columns will always yield a member of the family, and the associated confounding pattern is obtained by making the corresponding sign changes in the alias patterns of Table 12.12.

Building Blocks

Resolution IV designs may be used sequentially as were the resolution III designs. As before, sign switching may be used to eliminate particular confounding links.

General Construction of Resolution IV Designs

The construction of a resolution IV design containing 2^k variables follows exactly the pattern given for the 2_{IV}^{8-4} design:

1. Write a complete 2^k factorial with added columns for all interaction terms.
2. Generate a resolution III design for $2^k - 1$ variables by saturating this design with variables.
3. Add a further variable as a column of plus signs.
4. Repeat the design with all signs reversed to give a resolution IV design for 2^k variables in 2^{k+1} runs.

An alternative general method is given in Appendix 12A.

12.7. ELIMINATION OF BLOCK EFFECTS IN FRACTIONAL DESIGNS

Fractional designs may be run in blocks, with suitable contrasts used as "block variables." A design in 2^q blocks is defined by q independent contrasts. All effects (including aliases) associated with these basic contrasts *and all their interactions* are confounded with blocks.

Example 2_{V}^{5-1} *Design in Two Blocks of Eight*

Consider again the 2^{5-1} design of Table 12.2. Suppose the investigator decided that interaction between feed rate and catalyst concentration was likely to be negligible. This interaction *13* could then be used for blocking. The eight runs 2, 20, 5, . . . , 15, having a minus sign in the **13** column, would be run in one block, and the eight runs 17, 3, 22, . . . , 32 in the other. Notice that in this design the alias *245* (here assumed negligible) of *13* is also confounded with blocks.

Example: 2_{V}^{5-1} *Design in Four Blocks of Four Runs*

Suppose that, in the 2^{5-1} design of Table 12.2, columns **13** and **23** are confounded with blocks. Then the interaction between these columns **13** × **23** = $\mathbf{123^2} = \mathbf{12}$ is also confounded. The design would thus be appropriate if we were prepared to confound with blocks all two-factor interactions between variables **1**, **2**, and **3** and their aliases. To achieve this arrangement, runs 20, 5, 12, and 29, for which the **13** and **23** columns have signs (− −), could be put in the first block, runs 2, 23, 26, and 15, for which columns **13** and **23** have signs (− +) in the second block, and so on. Thus in terms of a two-way table the arrangement would be as follows:

13	+	III + −	IV + +
	−	I − −	II − +
		−	+
		23	

The Resolution IV Designs as Main Effect Plans in Blocks of Two

It occasionally happens that we must work with very small block sizes. A remarkable class of such designs based on the resolution IV arrangement provides economical main effect plans with a block size of only two. In one investigation the subject of study was an effluent impurity that tended to vary slowly with time. Runs made consecutively were thus much more comparable than those made further apart. It was possible to run the design in blocks of 2-hour periods, one experimental condition being run in the first hour and one in the second. At one stage of the investigation a 16-run main effect plan was used to study the main effects of eight variables based on a blocked 2_{IV}^{8-4} design. The plan is shown in Table 12.13. To see how this is derived, consider the original design given in Table 12.11 and the aliasing strings in Table 12.12. For the blocking scheme suppose that we use any two-factor interaction contrast, say l_{12}, to accommodate $\mathbf{B_1}$, and a second, say l_{13}, to accommodate $\mathbf{B_2}$; then l_{17} cannot be used for $\mathbf{B_3}$ since it can be obtained by multiplying the signs of **12** and **13**. Suppose, therefore, we use l_{14} for $\mathbf{B_3}$. (The reader may confirm that any other remaining two-factor interaction contrast can equally well be employed.) Then the seven columns of signs obtained for $\mathbf{B_1}$, $\mathbf{B_2}$, $\mathbf{B_3}$, $\mathbf{B_1B_2}$, $\mathbf{B_1B_3}$, $\mathbf{B_2B_3}$, $\mathbf{B_1B_2B_3}$ exactly correspond to the contrasts $l_{12}, l_{13}, l_{14}, l_{15}, l_{16}, l_{17}, l_{18}$, in some order. They thus involve only the strings of interactions and not the main effects. When the design is rearranged in the eight blocks as on the right of Table 12.13, it is seen that the second run in each block is the mirror image or "fold-over" of the first run, that is, the signs in one run are exactly reversed in the other.

In designs of this kind, both the ordering within pairs and the sequence in which the pairs (blocks) are run should be random.

Rather than regard all between-block information as lost, the design can be analyzed on the basis that there are two different error variances. The within-block variance is appropriate for inferences about main effects, and the between-block variance for inferences about the strings of two-factor

TABLE 12.13. 2_{IV}^{8-4} **design in eight blocks of size two**

	2_{IV}^{8-4} design								block variable		
run	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	$\mathbf{B_1}$ 12	$\mathbf{B_2}$ 13	$\mathbf{B_3}$ 14
1	−	−	−	+	+	+	−	+	+	+	−
2	+	−	−	−	−	+	+	+	−	−	−
3	−	+	−	−	+	−	+	+	−	+	+
4	+	+	−	+	−	−	−	+	+	−	+
5	−	−	+	+	−	−	+	+	+	−	−
6	+	−	+	−	+	−	−	+	−	+	−
7	−	+	+	−	−	+	−	+	−	−	+
8	+	+	+	+	+	+	+	+	+	+	+
9	+	+	+	−	−	−	+	−	+	+	−
10	−	+	+	+	+	−	−	−	−	−	−
11	+	−	+	+	−	+	−	−	−	+	+
12	−	−	+	−	+	+	+	−	+	−	+
13	+	+	−	−	+	+	−	−	+	−	−
14	−	+	−	+	−	+	+	−	−	+	−
15	+	−	−	+	+	−	+	−	−	−	+
16	−	−	−	−	−	−	−	−	+	+	+

	design rearranged, in eight blocks											
block	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	$\mathbf{B_1}$	$\mathbf{B_2}$	$\mathbf{B_3}$	run
1	+	−	−	−	−	+	+	+	−	−	−	2
	−	+	+	+	+	−	−	−				10
2	−	−	+	+	−	−	+	+	+	−	−	5
	+	+	−	−	+	+	−	−				13
3	+	−	+	−	+	−	−	+	−	+	−	6
	−	+	−	+	−	+	+	−				14
4	−	−	−	+	+	+	−	+	+	+	−	1
	+	+	+	−	−	−	+	−				9
5	−	+	+	−	−	+	−	+	−	−	+	7
	+	−	−	+	+	−	+	−				15
6	+	+	−	+	−	−	−	+	+	−	+	4
	−	−	+	−	+	+	+	−				12
7	−	+	−	−	+	−	+	+	−	+	+	3
	+	−	+	+	−	+	−	−				11
8	+	+	+	+	+	+	+	+	+	+	+	8
	−	−	−	−	−	−	−	−				16

interactions. For large designs two *separate* plots can be made on normal probability paper.

Exercise 12.11. Above we used l_{12}, l_{13}, and l_{14} as three independent interaction contrasts. Confirm that the same final blocking plan is obtained whichever three independent interactions are used.

Exercise 12.12. Suppose that $k = 2^q$ ($q = 1, 2, 3, \ldots$). Show (*a*) that a 2_{IV}^{k-p} design may be obtained with $p = k - q - 1$, and (*b*) that the design can always be run in blocks of two as main effect plans in k variables, and that the "mirrored pair" property always holds.

12.8. DESIGNS OF RESOLUTION V AND HIGHER

At the beginning of this chapter we introduced the 2_V^{5-1} design. Like other resolution V designs, this arrangement has the property that no main effect or two-factor interaction is confounded with any other main effect or two-factor interaction. Table 12.14 lists some other designs of resolution V and higher and shows how they may be blocked so that *no main effect or two-factor interaction is confounded with any other main effect or two-factor interaction.*

For illustration consider the 2_V^{11-4} designs for studying 11 variables in 128 runs. To obtain the design, write a complete factorial in the $11 - 4 = 7$ variables **1, 2, 3, . . . , 7**; then associate new variables **8, 9, $\overline{\mathbf{10}}$, $\overline{\mathbf{11}}$** with interactions as shown in column (5) of the table. To arrange in eight blocks of 16 runs write columns $\mathbf{B_1 = 149}$, $\mathbf{B_2 = 12\overline{10}}$, $\mathbf{B_3 = 89\overline{10}}$. The 16 runs for which $(\mathbf{B_1 B_2 B_3})$ are $(- - -)$ are put in one block, the 16 runs for which $(\mathbf{B_1 B_2 B_3})$ are $(+ - -)$ in another, and so on.

Application of Yates's Algorithm to Fractional Designs

Yates's algorithm can be used in analyzing data from any 2^{k-p} fractional factorial design. The algorithm is applied in the usual way to any embedded complete factorial in k-p factors. For example one way to compute and identify the effects for the 2_{IV}^{8-4} design of Table 12.11 is as follows:

(i) rearrange the 16 runs in Yates order as a complete factorial in variables **1, 2, 3** and **8**
(ii) calculate the effects using Yates algorithm
(iii) associate the calculated effects with their appropriate aliases.

Exercise 12.13. Make an analysis of the data in Table 12.11, using Yates's algorithm.

TABLE 12.14. Construction and blocking of some designs of resolution V and higher so that no main effect or interaction is confounded with any other main effect or interaction

(1) number of variables	(2) number of runs	(3) degree of fractionation	(4) type of design	(5) method of introducing "new" factors	(6) blocking (with no main effect or interaction confounded)	(7) method of introducing blocks
5	16	$\frac{1}{2}$	2_{V}^{5-1}	$\pm 5 = \mathbf{1234}$	not available	
6	32	$\frac{1}{2}$	2_{VI}^{6-1}	$\pm 6 = \mathbf{12345}$	two blocks of 16 runs	$B_1 = \mathbf{123}$
7	64	$\frac{1}{2}$	2_{VII}^{7-1}	$\pm 7 = \mathbf{123456}$	eight blocks of 8 runs	$B_1 = \mathbf{1357}$ $B_2 = \mathbf{1256}$ $B_3 = \mathbf{1234}$
8	64	$\frac{1}{4}$	2_{V}^{8-2}	$\pm 7 = \mathbf{1234}$ $\pm 8 = \mathbf{1256}$	four blocks of 16 runs	$B_1 = \mathbf{135}$ $B_2 = \mathbf{348}$
9	128	$\frac{1}{4}$	2_{VI}^{9-2}	$\pm 8 = \mathbf{13467}$ $\pm 9 = \mathbf{23567}$	eight blocks of 16 runs	$B_1 = \mathbf{138}$ $B_2 = \mathbf{129}$ $B_3 = \mathbf{789}$
10	128	$\frac{1}{8}$	2_{V}^{10-3}	$\pm 8 = \mathbf{1237}$ $\pm 9 = \mathbf{2345}$ $\pm \overline{10} = \mathbf{1346}$	eight blocks of 16 runs	$B_1 = \mathbf{149}$ $B_2 = \mathbf{12\overline{10}}$ $B_3 = \mathbf{89\overline{10}}$
11	128	$\frac{1}{16}$	2_{V}^{11-4}	$\pm 8 = \mathbf{1237}$ $\pm 9 = \mathbf{2345}$ $\pm \overline{10} = \mathbf{1346}$ $\pm \overline{11} = \mathbf{1234567}$	eight blocks of 16 runs	$B_1 = \mathbf{149}$ $B_2 = \mathbf{12\overline{10}}$ $B_3 = \mathbf{89\overline{10}}$

12.9. SUMMARY

Estimation redundancy often occurs in data from 2^k factorials. Many higher order interactions may be negligible, and some of the factors may be without detectable effects of *any* kind. Utilization of fractional factorials can then reduce experimental effort. In general, increase in the degree of fractionation lowers the resolution of the best fraction and increases confounding between effects of various orders. Fractional designs may be employed as building blocks in the iterative acquisition of knowledge. In this evolution, designs can be augmented so that ambiguities revealed at one stage of experimentation can be resolved in the next. A summary of some useful fractional designs is given in Table 12.15.

APPENDIX 12A. STRUCTURE OF THE FRACTIONAL DESIGNS*

Confounding Patterns for Resolution III Designs

The 2^{7-4} design of Table 12.5 was obtained by setting

$$\mathbf{4} = \mathbf{12}, \quad \mathbf{5} = \mathbf{13}, \quad \mathbf{6} = \mathbf{23}, \quad \mathbf{7} = \mathbf{123} \tag{12A.1}$$

Multiplying both sides of each of these identities by **4**, **5**, **6**, and **7**, respectively, provides the four *generating relations* in the form

$$\mathbf{I} = \mathbf{124}, \quad \mathbf{I} = \mathbf{135}, \quad \mathbf{I} = \mathbf{236}, \quad \mathbf{I} = \mathbf{1237} \tag{12A.2}$$

Combinations such as **124** and **135** may be referred to as "words." The *defining relation* includes *all* words that are equal to the identity **I**. These are the *generators* **124**, **135**, **236**, **1237** themselves and all other words that can be obtained by multiplying these generators together. Multiplying two at a time gives

$$\mathbf{I} = \mathbf{2345} = \mathbf{1346} = \mathbf{347} = \mathbf{1256} = \mathbf{257} = \mathbf{167} \tag{12A.3}$$

three at a time† gives

$$\mathbf{I} = \mathbf{456} = \mathbf{1457} = \mathbf{2467} = \mathbf{3567} \tag{12A.4}$$

and four at a time gives

$$\mathbf{I} = \mathbf{1234567} \tag{12A.5}$$

The complete defining relation is therefore

$$\begin{aligned} \mathbf{I} &= \mathbf{124} = \mathbf{135} = \mathbf{236} = \mathbf{1237} = \mathbf{2345} = \mathbf{1346} = \mathbf{347} \\ &= \mathbf{1256} = \mathbf{257} = \mathbf{167} = \mathbf{456} = \mathbf{1457} = \mathbf{2467} = \mathbf{3567} \\ &= \mathbf{1234567} \end{aligned} \tag{12A.6}$$

---

* Further discussion will be found in Box and Hunter (1961).

† For example, $\mathbf{124} \times \mathbf{135} \times \mathbf{236} = \mathbf{1^2 2^2 3^2 456} = \mathbf{456}$.

TABLE 12.15. Two-level fractional factorial designs for k variables and N runs (numbers in parentheses represent replication)

number of runs N \ number of variables k	3	4	5	6	7	8	9	10	11	
4	2^{3-1}_{III} $\pm 3 = 12$									
8	2^3	2^{4-1}_{IV} $\pm 4 = 123$	2^{5-2}_{III} $\pm 4 = 12$ $\pm 5 = 13$	2^{6-3}_{III} $\pm 4 = 12$ $\pm 5 = 13$ $\pm 6 = 23$	2^{7-4}_{III} $\pm 4 = 12$ $\pm 5 = 13$ $\pm 6 = 23$ $\pm 7 = 123$					
16	2^3 2 times	2^4	2^{5-1}_{V} $\pm 5 = 1234$	2^{6-2}_{IV} $\pm 5 = 123$ $\pm 6 = 234$	2^{7-3}_{IV} $\pm 5 = 123$ $\pm 6 = 234$ $\pm 7 = 134$	2^{8-4}_{IV} $\pm 5 = 234$ $\pm 6 = 134$ $\pm 7 = 123$ $\pm 8 = 124$	2^{9-5}_{III} $\pm 5 = 123$ $\pm 6 = 234$ $\pm 7 = 134$ $\pm 8 = 124$ $\pm 9 = 1234$	2^{10-6}_{III} $\pm 5 = 123$ $\pm 6 = 234$ $\pm 7 = 134$ $\pm 8 = 124$ $\pm 9 = 1234$ $\pm \overline{10} = 12$	2^{11-7}_{III} $\pm 5 = 123$ $\pm 6 = 234$ $\pm 7 = 134$ $\pm 8 = 124$ $\pm 9 = 1234$ $\pm \overline{10} = 12$ $\pm \overline{11} = 13$	
32	2^3 4 times	2^4 2 times	2^5	2^{6-1}_{VI} $\pm 6 = 12345$	2^{7-2}_{IV} $\pm 6 = 1234$ $\pm 7 = 1245$	2^{8-3}_{IV} $\pm 6 = 123$ $\pm 7 = 124$ $\pm 8 = 2345$	2^{9-4}_{IV} $\pm 6 = 2345$ $\pm 7 = 1345$ $\pm 8 = 1245$ $\pm 9 = 1235$	2^{10-5}_{IV} $\pm 6 = 1234$ $\pm 7 = 1235$ $\pm 8 = 1245$ $\pm 9 = 1345$ $\pm \overline{10} = 2345$	2^{11-6}_{IV} $\pm 6 = 123$ $\pm 7 = 234$ $\pm 8 = 345$ $\pm 9 = 134$ $\pm \overline{10} = 145$ $\pm \overline{11} = 245$	$(\frac{1}{128})$
64	2^3 8 times	2^4 4 times	2^5 2 times	2^6	2^{7-1}_{VII} $\pm 7 = 123456$	2^{8-2}_{V} $\pm 7 = 1234$ $\pm 8 = 1256$	2^{9-3}_{IV} $\pm 7 = 1234$ $\pm 8 = 1356$ $\pm 9 = 3456$	2^{10-4}_{IV} $\pm 7 = 2346$ $\pm 8 = 1346$ $\pm 9 = 1245$ $\pm \overline{10} = 1235$	2^{11-5}_{IV} $\pm 7 = 345$ $\pm 8 = 1234$ $\pm 9 = 126$ $\pm \overline{10} = 2456$ $\pm \overline{11} = 1456$	$(\frac{1}{64})$
128	2^3 16 times	2^4 8 times	2^5 4 times	2^6 2 times	2^7	2^{8-1}_{VIII} $\pm 8 = 1234567$	2^{9-2}_{VI} $\pm 8 = 13467$ $\pm 9 = 23567$	2^{10-3}_{V} $\pm 8 = 1237$ $\pm 9 = 2345$ $\pm \overline{10} = 1346$	2^{11-4}_{V} $\pm 8 = 1237$ $\pm 9 = 2345$ $\pm \overline{10} = 1346$ $\pm \overline{11} = 1234567$	$(\frac{1}{32})$
	(16)	(8)	(4)	(2)	(1)	$(\frac{1}{2})$	$(\frac{1}{4})$	$(\frac{1}{8})$	$(\frac{1}{16})$	

This defining relation provides the confounding pattern for the whole design. For example, multiplying through by **1** gives

$$\mathbf{1} = \mathbf{24} = \mathbf{35} = \mathbf{1236} = \mathbf{237} = \mathbf{12345} = \mathbf{346} = \mathbf{1347} = \mathbf{256}$$
$$= \mathbf{1257} = \mathbf{67} = \mathbf{1456} = \mathbf{457} = \mathbf{12467} = \mathbf{13567} = \mathbf{234567} \qquad (12A.7)$$

Thus interactions *24*, *35*, *1236*, etc., are confounded with (are aliases of) main effect *1*. By repeatedly using the defining relation, and omitting words with three or more letters, we obtain the abbreviated version of the confounding pattern of Table 12.6, which is appropriate on the assumption that all interactions among three or more variables are negligible.

Note that the defining relation for the 2^{7-4} design contains 16 words and each main effect and interaction has 15 aliases. In general, a 2^{k-p} design is produced by p generators and has a defining relation containing 2^p words.

The 16 possible combinations of ± signs for the four generators

$$\mathbf{I} = \pm\mathbf{124}, \quad \mathbf{I} = \pm\mathbf{135}, \quad \mathbf{I} = \pm\mathbf{236}, \quad \mathbf{I} = \pm\mathbf{1237} \qquad (12A.8)$$

determine the 16 separate fractions. In composing the defining relations for each of these different fractions, we employ the usual rules of algebraic multiplication to determine the signs in the defining relation and hence in the confounding pattern. For example, if the individual generators had been $-\mathbf{124}$, $+\mathbf{135}$, $+\mathbf{236}$ and $-\mathbf{1237}$ the complete defining relation would be $\mathbf{I} = -\mathbf{124} = \mathbf{135} = \mathbf{236} = -\mathbf{1237} = -\mathbf{2345} = -\mathbf{1346} = \mathbf{347} = \mathbf{1256} = -\mathbf{257} = -\mathbf{167} = -\mathbf{456} = \mathbf{1457} = \mathbf{2467} = -\mathbf{3567} = \mathbf{1234567}$.

Exercise 12A.1. Obtain the generators and defining relation for the 2^{7-4}_{IV} fraction obtained by setting $\mathbf{4} = -\mathbf{12}$, $\mathbf{5} = \mathbf{13}$, $\mathbf{6} = \mathbf{23}$, $\mathbf{7} = \mathbf{123}$. Use the defining relation to obtain the aliases of all main effects. By omitting interactions between more than three factors, confirm the entries in Table 12.8.

Resolution

The resolution R of a fractional design is the *length of the shortest word in the defining relation.* It should be clear from this definition that the saturation method for generating fractionals must always give designs of resolution III.

Confounding Patterns for Resolution IV Designs

In Section 12.6 we described the generation of a 2^{8-4}_{IV} design by "fold-over." We now consider more carefully the confounding pattern for this design. The two component groups of eight runs can be regarded as separate 2^{8-5} designs with generating relations

$$\mathbf{I}_8 = \mathbf{8} = \mathbf{124} = \mathbf{135} = \mathbf{236} = \mathbf{1237} \qquad (12A.9)$$

and

$$\mathbf{I}_8 = -\mathbf{8} = -\mathbf{124} = -\mathbf{135} = -\mathbf{236} = \mathbf{1237} \qquad (12A.10)$$

respectively, where the notation $\mathbf{I}_8$ refers to a column of eight plus signs. A set of four generators, and hence the defining relation, for the combined 2^{8-4} design can now be obtained from these expressions as follows. The relation $\mathbf{I}_8 = \mathbf{1237}$ holds for both sets of eight runs. Consequently, if $\mathbf{I}_{16}$ represents the column of 16 plus signs associated with the combined design,

$$\mathbf{I}_{16} = \mathbf{1237}$$

and **1237** is a generator for the combined design. Also, for the first part of the design $\mathbf{I}_8 = (\mathbf{8})(\mathbf{124}) = \mathbf{1248}$, and for the second part $\mathbf{I}_8 = (-\mathbf{8})(-\mathbf{124}) = \mathbf{1248}$. Thus for the complete design $\mathbf{I}_{16} = \mathbf{1248}$.

By a similar argument $\mathbf{I}_{16} = \mathbf{1358} = \mathbf{2368}$. Thus the complete set of four generators is

$$\mathbf{I}_{16} = \mathbf{1237}, \qquad \mathbf{I}_{16} = \mathbf{1248}, \qquad \mathbf{I}_{16} = \mathbf{1358}, \qquad \mathbf{I}_{16} = \mathbf{2368} \tag{12A.11}$$

By multiplication as before we obtain the defining relation:

$$\begin{aligned}\mathbf{I}_{16} &= \mathbf{1237} = \mathbf{1248} = \mathbf{1358} = \mathbf{2368}\\ &= \mathbf{3478} = \mathbf{2578} = \mathbf{1678} = \mathbf{2345} = \mathbf{1346} = \mathbf{1256}\\ &= \mathbf{1457} = \mathbf{2467} = \mathbf{3567} = \mathbf{4568}\\ &= \mathbf{12345678}\end{aligned} \tag{12A.12}$$

Since the shortest word is of length four, the combined design is of resolution IV as required.

Exercise 12A.2. Verify the confounding pattern shown in Table 12.13, and extend it to include aliases of all orders.

Exercise 12A.3. Although the defining relation for any fractional factorial is unique, the generators for a 2^{k-p} factorial design with $p > 1$ are not unique. Any set of $k - p$ words that generate the defining relation is a set of generators. Show that an alternative set of generators for the 2_{IV}^{8-4} design of Table 12.11 is **3478**, **2578**, **1457**, and **2467**.

An Alternative Method for Generating Resolution IV Designs

As an alternative to the fold-over method, any 2_{IV}^{k-p} design may be constructed as follows:

1. Write the design matrix for a full factorial design for $k - p$ variables.
2. Associate extra variables with all interaction columns containing an *odd* number of numerals.

We demonstrate by obtaining anew the 2_{IV}^{8-4} design of Table 12.11, whose confounding was discussed above. To do this, write down a 2^4 factorial for the variables **1**, **2**, **3**, and **8**. The four three-factor interaction columns are then **128**, **138**, **238**, and **123**. Now associate these with the four "new" variables **4**, **5**, **6**, and **7** to obtain a set of four generators:

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

The design thus constructed is identical to that given in Table 12.11. The only reason for starting with variables termed **1**, **2**, **3**, and **8** instead of **1**, **2**, **3**, and **4** is to make it easy to see the identity between this method and the preceding one.

The defining relation for this design is $\mathbf{I} = \mathbf{1248} = \mathbf{1358} = \mathbf{2368} = \mathbf{1237} = \mathbf{2345} = \mathbf{1346} = \mathbf{3478} = \mathbf{1256} = \mathbf{2578} = \mathbf{1678} = \mathbf{4568} = \mathbf{1457} = \mathbf{2467} = \mathbf{3567} = \mathbf{1234567}$. The identical design could be obtained, for example, using the generators **1237**, **1248**, **1346**, **2345**, that is, by first writing down the 16 runs of a 2^4 factorial in variables **1**, **2**, **3** and **4**, and then associating the extra variables with the three factor interactions. This method for constructing the 2_{IV}^{8-4} design is that presented in Table 12.15.

APPENDIX 12B. CHOOSING ADDITIONAL RUNS TO RESOLVE AMBIGUITIES FROM FRACTIONAL FACTORIALS

In the injection molding example of Section 12.6 the linear contrast $l_{15} = 4.6$ is not easily explained by system noise. But there is ambiguity in its interpretation, since it estimates the sum of the effects *15* + *26* + *38* + *47*. Three is the smallest number of additional runs that could allow separate estimation of the true values of these two-factor interactions. However, since the additional runs have to be made at a different time from the first 16 runs, we should also allow for a general change in level (a block effect). Thus the minimum number of additional runs we need is 4. We now consider how 4 such runs might be chosen.

For the runs made so far, the columns of signs corresponding to the interactions *15*, *26*, *38*, and *47* are identical. We need 4 additional runs that will permit separate estimation of the mean values of these interactions and will also allow for a possible block effect (a change in level between the first 16 runs and the additional 4 runs). One sensible possibility is to employ additional runs that yield signs for the interaction columns as follows:

15	**26**	**47**	**38**
+	−	−	+
+	+	−	−
+	−	+	−
+	+	+	+

Four additional runs that will do this are:

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
+	+	+	+	+	−	−	+
+	+	+	+	+	+	−	−
+	+	+	+	+	−	+	−
+	+	+	+	+	+	+	+

This four-run design is obtained by writing a plus sign for each element in the **1**, **2**, **3**, **4** columns and then choosing the remaining signs to satisfy the requirements of the previous table. The choice is not unique. In particular, the signs in one or more rows and/or in one or more columns may be switched, and the resulting design will still possess the desired characteristics. The experimenter has many choices. Suppose that he decides to perform a design in which variables **3** and **5**, which apparently have the largest main effects, are not maintained at the same levels but are varied in a 2^2 factorial design. Such a sequence is easily obtained as before by assigning the necessary signs to columns **3** and **5** and then arranging the other columns to satisfy the required condition. One such arrangement for the injection molding example is given in Table 12B.1, which also contains new data.

TABLE 12B.1. Four additional runs with data, injection molding example

run	1	2	3	4	5	6	7	8	shrinkage
17	−	+	+	+	−	−	−	+	29.4
18	−	+	−	−	−	+	+	+	19.7
19	+	+	−	−	+	−	−	+	13.6
20	+	+	+	+	+	+	+	+	24.7

Incorporating the New Data

In general, the incorporation of design fragments can always be achieved by use of the method of least squares (see Chapter 14 and references given there). For designs of the kind here considered, where the number of extra constants is exactly equal to the number of additional runs, the least squares analysis simplifies. It then corresponds exactly to a commonsense analysis that is illustrated below for the data of Tables 12.11 and 12.B.1.

From the analysis of the first 16 runs it appears that, apart from the effect of noise, the data may be explained by a mean level plus the main effects of variables **3** and **5** and the effect of one more of interactions *15*, *26*, *47*, and *38*. The main effect of variable **8** is the next largest effect and has also been treated as real in this analysis. We denote the mean level in the second block by M. Now the response expected when an effect (main effect or interaction) is at the plus level is the mean plus one half of that effect. Similarly the response expected at the minus level is the mean minus one half of that effect. Thus we can make a table of the following kind, which summarizes all of the information available about the four interactions both from the new runs and from the previous 16 runs.

run	M	$\frac{1}{2}l_3 = 2.75$	$\frac{1}{2}l_5 = -1.9$	$\frac{1}{2}l_8 = 0.6$	$\frac{1}{2}(15)$	$\frac{1}{2}(26)$	$\frac{1}{2}(47)$	$\frac{1}{2}(38)$	
17	+	+	−	+	+	−	−	+	$29.4 = y_{17}$
18	+	−	−	+	+	+	−	−	$19.7 = y_{18}$
19	+	−	+	+	+	−	+	−	$13.6 = y_{19}$
20	+	+	+	+	+	+	+	+	$24.7 = y_{20}$
					+	+	+	+	$2.3 = \frac{1}{2}l_{15}$

In the first row of this table, for example, $\frac{1}{2}l_3 = 2.75$ is the best available estimate (taken from Table 12.12) of one half of the main effect of factor **3**. The second row of the table tells us that the result $y_{17} = 29.4$ should be explained by the equation

$$M + 2.75 - (-1.9) + 0.6 + \tfrac{1}{2}(15 - 26 - 47 + 38) = 29.4 \tag{12B.1}$$

This yields

$$M + \tfrac{1}{2}(15 - 26 - 47 + 38) = 24.15 \tag{12B.2}$$

The last row of the table presents all the information provided by the first 16 runs about the interaction effects, that is, that half their sum is equal to $\frac{1}{2}l_{15} = 2.3$. Putting the reduced equations together, we have

M	$\frac{1}{2}(15)$	$\frac{1}{2}(26)$	$\frac{1}{2}(47)$	$\frac{1}{2}(38)$	
+	+	−	−	+	24.15
+	+	+	−	−	19.95
+	+	−	+	−	17.65
+	+	+	+	+	23.25
	+	+	+	+	2.3

From the last two equations M is estimated as 20.95. Substituting this value in each of the preceding equations, we obtain

$\frac{1}{2}(15)$	$\frac{1}{2}(26)$	$\frac{1}{2}(47)$	$\frac{1}{2}(38)$	
+	−	−	+	3.2
+	+	−	−	−1.0
+	−	+	−	−3.3
+	+	+	+	2.3

From this table the effect of interest seems mainly associated with interaction *38*. More precisely, by solving these last four equations we obtain

$$0.6 \rightarrow 15, \quad 0.7 \rightarrow 26, \quad -1.6 \rightarrow 47, \quad 4.9 \rightarrow 38$$

It seems very likely, therefore, that a considerable interaction between holding pressure and screw speed accounts for the majority, if not all, of the interaction effects found.

The nature of this interaction may be comprehended from the two-way table below, which shows the average shrinkage at all combinations of low and high values of holding pressure and screw speed using data from the original 16 runs.

screw speed **(8)**	+	15.3	25.4
	−	18.7	19.6
		−	+
		holding pressure **(3)**	

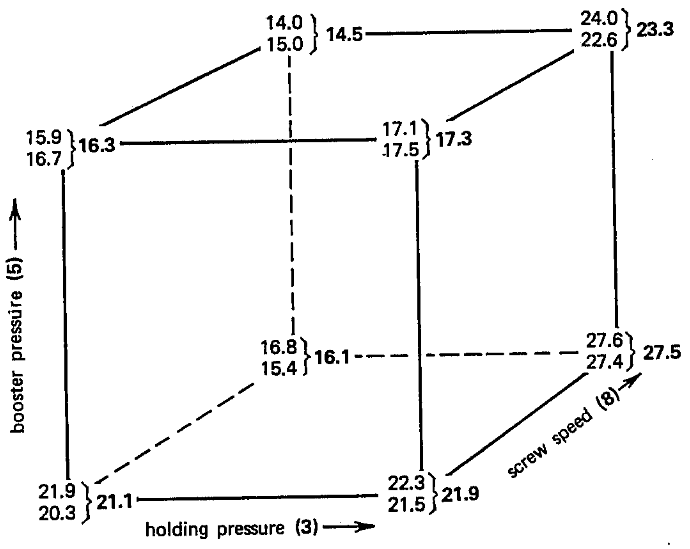


FIGURE 12B.1. The 2_{IV}^{8-4} design as a replicated 2^3 factorial in variables 3, 5, and 8.

We see that, whereas at low holding pressure increasing screw speed reduces shrinkage from 19 to 15 units, at high holding pressure increasing screw speed increases shrinkage from 20 to 25 units.

A Replicated 2^3 *Design Encapsulated in the* 2_{IV}^{8-4} *Fractional Factorial*

The results appear to be explained by main effects and interactions of factors **3**, **5**, and **8**. Supposing the remaining factors to be essentially inert, we rearrange the data as a replicated 2^3 factorial in factors **3**, **5**, and **8** in Table 12B.2 and Figure 12B.1, taking advantage of the property that the design provides duplicated 2^3 factorials in all possible subsets of three variables.

TABLE 12B.2. **The 2_{IV}^{8-4} design as a replicated 2^3 factorial in variables 3, 5, and 8**

runs	3	5	8	shrinkage		average from first 16 runs
14,16	−	−	−	21.9	20.3	21.1
9, 11	+	−	−	22.3	21.5	21.9
13, 15	−	+	−	15.9	16.7	16.3
10, 12	+	+	−	17.1	17.5	17.3
2, 4 (18*)	−	−	+	16.8	15.4 (18.5*)	16.1
5, 7 (17*)	+	−	+	27.6	27.4 (28.2*)	27.5
1, 3 (19*)	−	+	+	14.0	15.0 (12.4*)	14.5
6, 8 (20*)	+	+	+	24.0	22.6 (23.5*)	23.3

In Table 12B.2 the results from the additional runs (17, 18, 19, 20), indicated by asterisks, have been adjusted for differences associated with blocks, that is, 20.95 − 19.75 = 1.2 has been subtracted from each value. These adjusted values are seen to agree quite well with the original shrinkage values *on the assumption that* **3**, **5**, *and* **8** *are the important factors.*

REFERENCES AND FURTHER READING

For further discussion of fractional factorial designs see the following and the references listed therein:

Daniel, C. (1976). *Applications of Statistics to Industrial Experimentation*, Wiley.

Plackett, R. L., and J. P. Burman (1946). The design of optimum multifactorial experiments, *Biometrika*, 33, 305.

Box, G. E. P., and J. S. Hunter (1961). The 2^{k-p} fractional factorial designs, *Technometrics*, **3**, 311, 449.

For augmentation of designs see Daniel's book and the following articles:

Davies, O. L., and W. A. Hay (1950). Construction and uses of fractional factorial designs in industrial research, *Biometrics*, **6**, 233.

Box, G. E. P., and K. B. Wilson (1951). On the experimental attainment of optimum conditions, *Roy. Stat. Soc., Ser. B*, **13**, 1.

Daniel, C. (1962). Sequences of fractional replicates in the 2^{p-q} series, *J. Am. Stat. Soc.*, **58**, 403.

Box, G. E. P., (1966). A note on augmented designs, *Technometrics*, **8**, 184.

QUESTIONS FOR CHAPTER 12

1. What is a fractional factorial design?
2. What is a half-fraction, and how can you construct such a design?
3. What is a saturated design, and how can you construct such a design?
4. Discuss the sequential use of fractional designs.
5. A 2^{8-3} design has how many runs? How many variables? How many levels for each variable? Answer the same questions for a 2^{k-p} design.
6. All other things being equal, why would a resolution IV design be preferred to a resolution III design?
7. Is it possible to construct a 2^{8-3} design of resolution III? Resolution IV? Resolution V?
8. In what situations is it useful to employ fractional factorial designs?

9. How can fractional factorial designs be blocked? How should they be randomized?
10. How might you analyze data from a 2^{7-1} design? A 2^{7-4} design?
11. What is a defining relation? A generator? A confounding pattern? Why is it necessary to know the confounding pattern for a fractional factorial design?
12. Construct, starting with the 12-run Plackett and Burman design, a 12-variable resolution IV design.
13. Design a 2^{5-1} design in eight blocks of size two so that main effects are clear of block effects.

CHAPTER 13

More Applications of Fractional Factorial Designs

The purpose of this chapter is to give further examples of the use of fractional factorial designs. It is a companion to Chapter 11, which served the same purpose for factorial designs.

13.1. EXAMPLE 1: EFFECTS OF FIVE VARIABLES ON SOME PROPERTIES OF CAST FILMS

Full design = 2^5 = 32 runs

In this example five variables were studied: the catalyst concentration, the amount of a certain additive, and the amounts of three emulsifiers *A*, *B*, and *C*. In an initial 2^{5-2} fraction eight polymer solutions were prepared, each was spread as a film on a microscope slide, and the properties of the films were recorded after they dried. The results for six different responses are shown in Table 13.1.

Surprisingly many conclusions can be drawn from these results by mere visual inspection. On the assumption (reasonable for this particular application) of dominant main effects, the important variable affecting haziness is emulsifier *A* (variable **3**). The important variable affecting adhesion is catalyst concentration (variable **1**). The important variables affecting the remaining responses of grease on top of the film, grease under the film, dullness of the film when the pH was adjusted, and dullness of the film when the original pH was used are **4**, **5**, **4**, and **4**, respectively.

Commentary

If experiments are carefully planned, a great deal of information can sometimes be obtained without much mathematical analysis (no computation was needed in this case). The converse of this statement, however, is not

TABLE 13.1. 2^{5-2} **fraction (4 = 23, 5 = 123) with results for six responses of interest, Example 1**

	variable					response					
run	1	2	3	4	5	③ hazy?	① adheres?	④ grease on top of film?	⑤ grease under film?	④ ~~dull~~ (adjusted pH)?	(~4) dull (original pH)?
1	−	−	−	+	−	no	no	yes	no	slightly	yes
2	+	−	−	+	+	no	yes	yes	yes	slightly	yes
3	−	+	−	−	+	no	no	no	yes	no	no
4	+	+	−	−	−	no	yes	no	no	no	no
5	−	−	+	−	+	yes	no	no	yes	no	slightly
6	+	−	+	−	−	yes	yes	no	no	no	no
7	−	+	+	+	−	yes	no	yes	no	slightly	yes
8	+	+	+	+	+	yes	yes	yes	yes	slightly	yes

variable	−	+
1 catalyst (%)	1	$1\frac{1}{2}$
2 additive (%)	$\frac{1}{4}$	$\frac{1}{2}$
3 emulsifier *A* (%)	2	3
4 emulsifier *B* (%)	1	2
5 emulsifier *C* (%)	1	2

true. If experiments are not carefully planned, it may be impossible to obtain much useful information even with extensive and careful analysis. This is the reason why design is more important than analysis. In this case, after the initial set of eight runs further fractions could have been run to confirm the tentative findings, but the experimenter judged this unnecessary and moved successfully to the next part of his investigation.

$4 = 23 \quad 5 = 123 \quad p = 2 \Rightarrow 2^p$ words

Exercise 13.1. Show that the generating relations for the design of Table 13.1 are **I** = **234** and **I** = **1235**. Use these relations to obtain the defining relation and the alias structure. *Partial answer*: **I** = **234** = **1235** = **145**, l_1 → *1234* + *235* + *45*. ✓

$I = 234 = 1235 = 145 \quad l_1 \Rightarrow 1234 + 235 + 45 + 1$

Exercise 13.2. Show that the alias structure may be obtained alternatively by thinking of the design as a saturated 2_{III}^{7-4} design in which two variables have been omitted.

$2_{\text{III}}^{5-2} \equiv 2_{\text{III}}^{7-4} \overset{?}{\Rightarrow} 2_{\text{III}}^{8-5}$ 3 variables omitted ?

The reader may ask how the experimenter could be certain that the effects were due to main effects and not to interactions in the alias strings. The answer is that he was not certain. From his knowledge of these variables, however, he thought that to press on was the best bet. In doing so he took a calculated risk of having to turn back at a later stage, if his bet did not come off. Decisions of this kind by experimenters are not peculiar to the use of fractional factorial designs. The truth is that the experimenter, in deciding what to do next, is never *certain* of what is best. He always guesses or, to use a more palatable term, uses judgment, and he always runs the risk of being wrong. This apparently dangerous mode of life may come as a surprise, but it is easy to see that efficiency requires it.

Imagine three experimenters, one who was very cautious, one who took carefully calculated risks, and one who jumped to conclusions on very thin evidence. Bear in mind that time and resources are really *always* limited. Consequently to repeat a set of experiments unnecessarily is to deny that effort to the investigation of other variables; on the other hand, frequent doubling back, because false trails are followed, wastes effort. It is clear, then, that the experimenter who takes some suitable intermediate position between the ultraconservative and the foolhardy is likely to be most successful.

"But," some may say, "I thought statistics made everything objective." If by that they mean that statistics ought to lead along a unique, painless route to the truth, they are mistaken. What statistics does, for a given amount of effort, is to lessen the risks of being wrong, either through missing important facts or through giving credence to phenomena that have no reality. Statistics allows the investigator to play the calculated-risk game most efficiently.

13.2. EXAMPLE 2: STABILITY OF NEW PRODUCT

A chemist in an industrial development laboratory was trying to formulate a household liquid product using a completely new process. He was able to demonstrate fairly quickly that the product could be manufactured and that it possessed a number of attractive properties. Unfortunately it could not be marketed because it was unstable.

When the statistician first met him, the chemist had for months been trying many different ways of synthesizing the product in the hopes of hitting on conditions that would give stability, but without success. He had succeeded, however, in identifying four variables that had important influences on stability: **(1)** acid concentration, **(2)** catalyst concentration, **(3)** temperature, and **(4)** monomer concentration. Full design = 16 expts

With his budget almost expended, he agreed somewhat reluctantly to perform his first statistically planned experiment, the 2^{4-1} fractional factorial design shown in Table 13.2. In these tests, which were performed in random order, the chemist was trying to achieve a stability value of at least 25. His initial reaction to the data was disgust, since none of the individual observations reached the desired stability level.

Using the analysis shown in Table 13.3, the statistician suggested the simplest explanation of the data was that only two variables, **1** and **2**, influenced stability, the other two being inert. If this hypothesis were true, the design could be viewed as a duplicated 2^2 factorial design in acid concentration **(1)** and catalyst concentration **(2)** as shown in Figure 13.1. The chemist was asked, therefore, whether he would have been surprised if he had obtained discrepancies in duplicate runs similar to those in Figure 13.1. He said that he would not.

TABLE 13.2. Results for Example 2

4 = 123.

	variable				
test	**1**	**2**	**3**	**4**	stability (R)
1	−	−	−	−	20
2	+	−	−	+	14
3	−	+	−	+	17
4	+	+	−	−	10
5	−	−	+	+	19
6	+	−	+	−	13
7	−	+	+	−	14
8	+	+	+	+	10

variable	−	+
1 acid concentration (%)	20	30
2 catalyst concentration (%)	1	2
3 temperature (°C)	100	150
4 monomer concentration (%)	25	50

$\beta_1 = -2{,}875$
$\beta_2 = -3{,}75$
$\beta_3 = -0{,}625$
$\beta_4 = 0{,}375$

$I = 1234 \Rightarrow$ resolution IV design

A saturated 4 variable design is a 2^3 design (Resolution = IV)

TABLE 13.3. Analysis of data for Example 2

main effects and three-factor interactions

$l_1 = -5.8 \rightarrow 1 + 234$
$l_2 = -3.8 \rightarrow 2 + 134$
$l_3 = -1.2 \rightarrow 3 + 124$
$l_4 = 0.8 \rightarrow 4 + 123$

two-factor interactions

$l_{12} = 0.2 \rightarrow 12 + 34$
$l_{13} = 0.8 \rightarrow 13 + 24$
$l_{14} = -0.2 \rightarrow 14 + 23$

average and four-factor interaction

$l_{1234} = 14.6 \rightarrow \text{average} + \frac{1}{2}(1234)$

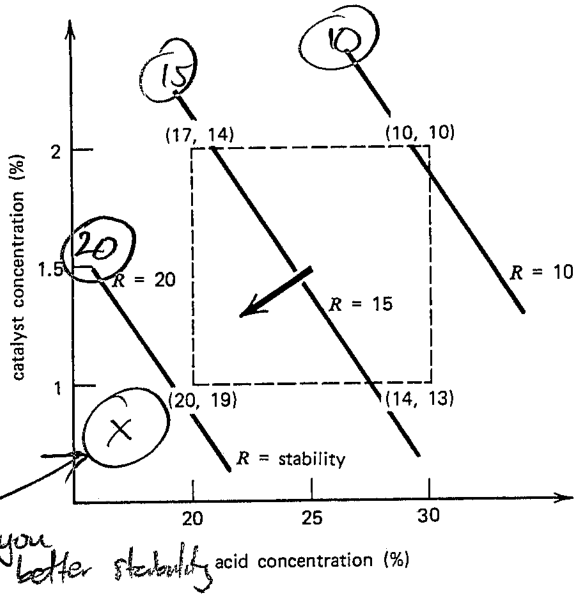
$\Rightarrow \beta_3 \sim \beta_4 \sim 0$

The possible implications of the results if the statistician's hypothesis was true were demonstrated by roughly sketching in by eye contour lines of a "stability plane" as shown in Figure 13.1. This picture suggested that experiments should be performed in the direction of the arrow. The contour lines shown are actually those obtained by the method of least squares (see Section 14.1). The direction at right angles to these contours is called the *direction of steepest ascent* (see Section 15.2). Refinement is, however, unnecessary in this example; "eyeball analysis" is all that is needed. A few exploratory runs performed in the general direction indicated produced, for the first time since the beginning of the investigation, a product with stability greater than the goal of 25.

Commentary

This example illustrates the following:

1. How a fractional factorial design was used for screening purposes to isolate two important variables from the original set of four proposed by the experimenter.
2. How a desirable *direction* in which to carry out further experiments was discovered.



this where you should obtain better stability

FIGURE 13.1. Results of 2^{4-1} fractional factorial design viewed as a duplicated 2^2 factorial design with approximate stability contours, Example 2.

An additional result of this investigation was that the experimenter became an evangelist for experimental design.

13.3. EXAMPLE 3: BOTTLENECK AT THE FILTRATION STAGE OF AN INDUSTRIAL PLANT

A number of similar chemical plants had been successfully operating for several years in different locations. In the older plants the time to complete a particular filtration cycle was about 40 minutes, but in a newly constructed plant filtration took almost twice as long, resulting in serious delays. What was the cause of the difficulty?

Seven Variables

To begin to solve this problem a meeting was called, at which a number of possibilities were considered.

1. Source of water supply. The plant engineer explained that the water for the new plant, which came from the city reservoir some 30 miles away, differed somewhat in mineral content from that available at the other locations. Some well water was available at the new site that more closely resembled the water supply at the older plants. The engineer suggested, therefore, that some tests be run with well water.
2. Origin of raw material. The process superintendant pointed out that the raw material, which was manufactured on site, was not identical in all respects to that used in the older plants. Consequently he proposed that some of the raw material from one of the older plants be shipped to the new plant and used in some tests.
3. Level of temperature. The temperature of filtration in the new plant was slightly lower than in the older plants. The plant chemist thought that this might be the cause of the problem.
4. Presence of recycle. A major difference between the new plant and the older ones was a recycle device absent in the latter. It was suggested that the inclusion of this device could increase filtration time.
5. Rate of addition of caustic soda. The rate of caustic soda addition in the new plant was higher than in the older plants. The process foreman suggested that the rate be decreased in the new plant.
6. Type of filter cloth. A new type of filter cloth was being used in the new plant. The process superintendent pointed out that it would be a relatively simple matter to get some filter cloths from the older plants and make some test runs with them.
7. Length of holdup time. In the new plant the holdup time was lower than in the older plants, and the quality control engineer gave reasons for believing that this might be the cause of the problem.

Much disagreement was expressed at the meeting about these factors. Some participants even argued that changes proposed by others were ridiculous.

The Design and the Results

Full = 128 runs → Saturated Res = III design

To sort out these ideas the 2^{7-4} screening design shown in Table 13.4 was performed on the plant. At the outset the attitude of the person responsible for the investigation was that out of these seven factors perhaps one or two would be found to be important. The chance was small, he thought, that there would be as many as three or more important variables; in fact, it was judged quite possible that none of those selected for investigation would have any effect at all. The order of the eight tests was randomized, and the data shown in Table 13.4 were obtained.

TABLE 13.4. Results of Example 3

variable	−	+
1 water supply	town reservoir	well
2 raw material	on site	other
3 temperature	low	high
4 recycle	yes	no
5 caustic soda	fast	slow
6 filter cloth	new	old
7 holdup time	low	high

test	1	2	3	12 4	13 5	23 6	123 7	filtration time (min) y
1	−	−	−	+	+	+	−	68.4
2	+	−	−	−	−	+	+	77.7
3	−	+	−	−	+	−	+	66.4
4	+	+	−	+	−	−	−	81.0
5	−	−	+	+	−	−	+	78.6
6	+	−	+	−	+	−	−	41.2
7	−	+	+	−	−	+	−	68.7
8	+	+	+	+	+	+	+	38.7

2² design

2² design

4 generators

2⁴ words/defining relationships

duplicates essentially

duplicates

Four Tentative Interpretations of Results

In Table 13.5 three of the calculated effects (l_1, l_3, and l_5) are large in absolute value and have been circled. There are several possible interpretations. Four of the most likely are:

1. Main effects *1*, *3*, and *5* are producing the effects.
2. Main effects *1* and *3* and interaction *13* are producing the effects.
3. Main effects *1* and *5* and interaction *15* are producing the effects.
4. Main effects *3* and *5* and interaction *35* are producing the effects.

The Second Design

To reduce these ambiguities a selected set of eight additional tests (see Table 13.6) was run, converting the original resolution III design to one of resolution IV. This was done by "fold-over" (Chapter 12), that is, by arranging that the added fraction had signs opposite to those in the original design.

TABLE 13.5. Calculated values and abbreviated confounding pattern for eight-run filtration experiment, Example 3

$l_1 = \boxed{-10.9} \rightarrow 1 + 24 + 35 + 67$

$l_2 = -2.8 \rightarrow 2 + 14 + 36 + 57$

$l_3 = \boxed{-16.6} \rightarrow 3 + 15 + 26 + 47$

$l_4 = 3.2 \rightarrow 4 + 12 + 37 + 56$

$l_5 = \boxed{-22.8} \rightarrow 5 + 13 + 27 + 46$

$l_6 = -3.4 \rightarrow 6 + 17 + 23 + 45$

$l_7 = 0.5 \rightarrow 7 + 16 + 25 + 34$

TABLE 13.6. Results of second filtration experiment, Example 3

called fold ove

test	1	2	3	−12 4	−13 5	−23 6	123 7	filtration time (min) y
9	+	+	+	−	−	−	+	66.7
10	−	+	+	+	+	−	−	65.0
11	+	−	+	+	−	+	−	86.4
12	−	−	+	−	+	+	+	61.9
13	+	+	−	−	+	+	−	47.8
14	−	+	−	+	−	+	+	59.0
15	+	−	−	+	+	−	+	42.6
16	−	−	−	−	−	−	−	67.6

Analysis of Sixteen Results

Combining the data from both eight-run designs yields the estimates given in Table 13.7. The three largest effects in absolute value are l_1, l_5, and l_{15}. It now seems likely that **1** and **5** are the two most important variables, with not only large main effects but also a large interaction. On this interpretation, **2**, **3**, **4**, **6**, and **7** are inert variables and the 16 tests are essentially a 2^2 factorial design in variables **1** and **5** replicated four times (see Figure 13.2).

Similar to previous example where it was essentially just a duplicate of a 2^2 design

MULTIVARIATE METHODS for PROCESS ANALYSIS, MONITORING & OPTIMIZATION

John MacGregor
Dept. of Chemical Engineering
McMaster University
Hamilton, ON, Canada

INTRODUCTION

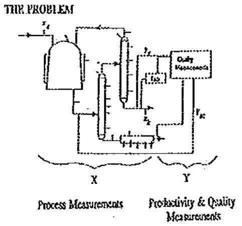
- Collect large amounts of data on industrial processes
- Need to use these data to
  - Trouble-shoot problems
  - Monitor process operation
  - Build inferential models
  - Improve processes
- Need efficient methods

OUTLINE

- Nature of Process Data
- Latent Variable Models
  - PCA & PLS Estimation
- Exploration & Analysis of Process Databases
  - Monomer recovery unit
- Process Monitoring
  - Batch polymerization process
- Product Design

NATURE OF PROCESS DATA

- Process data are messy
- Large # of process variables (e.g. 1000). Many quality variables.
- Extreme colinearity. Not 1000 things happening! Only a few underlying events drive the process.
- Signal/Noise ratio low
- Missing Data



eg: Distillation Column

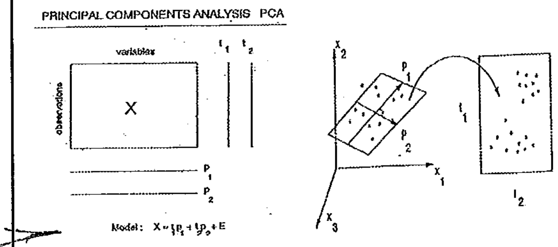
In control one aims for zero S:N ratio
so we need to magnify S:N to get signal

MULTIVARIATE STATISTICAL METHODS and the LATENT VARIABLE MODEL

- Linear LV Model:
  - $\underset{(N^*K)}{X} = \underset{(N^*A)}{T_A}\ \underset{(A^*K)}{P_A^T} + E = t_1p_1^T + t_2p_2^T + \ldots$
  - $Y = T_A\ Q_A^T + F$

  - True dimension of operating space is A<<K
  - All variables have error
  - Model for X space as well as Y (Key point)
  - PCA, PLS

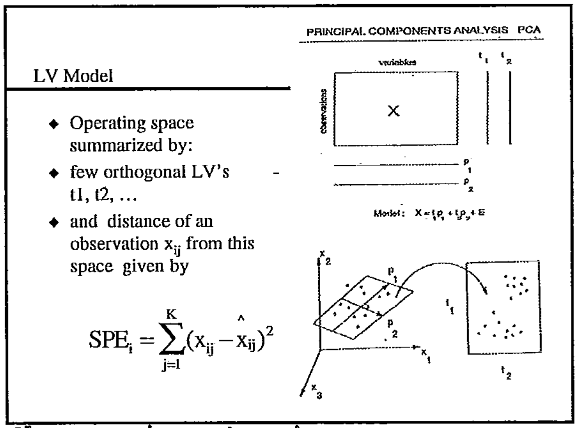
Multivariate Statistical Methods:
Principal Component Analysis (PCA)



Like regression $t_i \equiv x_i$
loadings $\longrightarrow p_i \equiv \hat{\beta}_i$

extract max variance from X and summarize it in <u>one</u> variable t_1 that can explain more variance than any other variable

If test: After t_2 extracting t_3 will be insignificant
$\Rightarrow t_2$ & t_1 describe all info & the rest is just noise

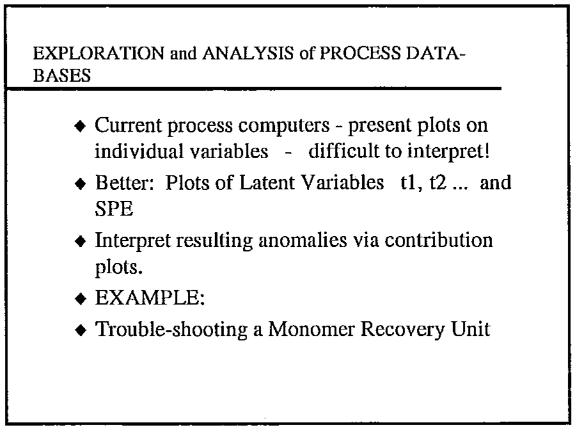


LV Model

- Operating space summarized by:
- few orthogonal LV's - t1, t2, ...
- and distance of an observation x_{ij} from this space given by

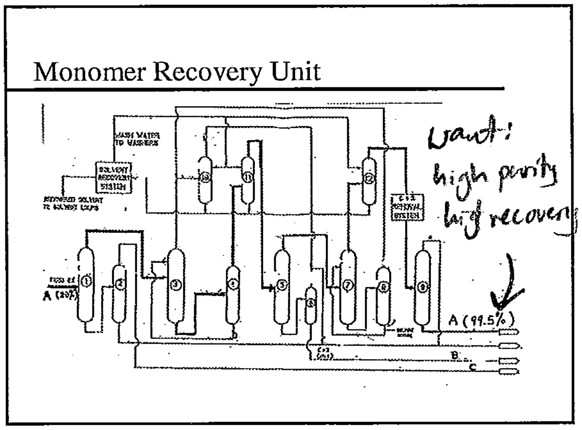
$$SPE_i = \sum_{j=1}^{K}(x_{ij} - \hat{x}_{ij})^2$$

if an upset enters the new points move off the plane and SPE (DModX) will start to increase



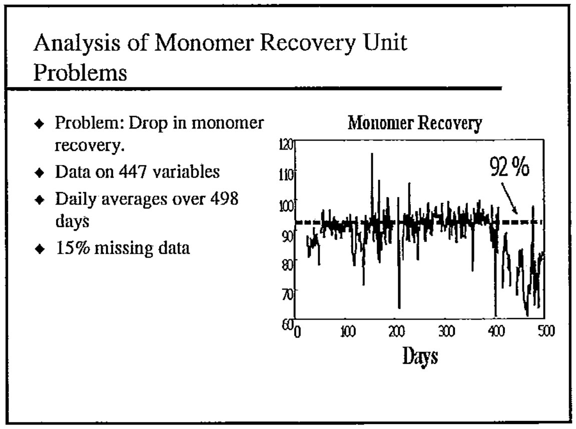
EXPLORATION and ANALYSIS of PROCESS DATA-BASES

- Current process computers - present plots on individual variables - difficult to interpret!
- Better: Plots of Latent Variables t1, t2 ... and SPE
- Interpret resulting anomalies via contribution plots.
- EXAMPLE:
- Trouble-shooting a Monomer Recovery Unit



Monomer Recovery Unit

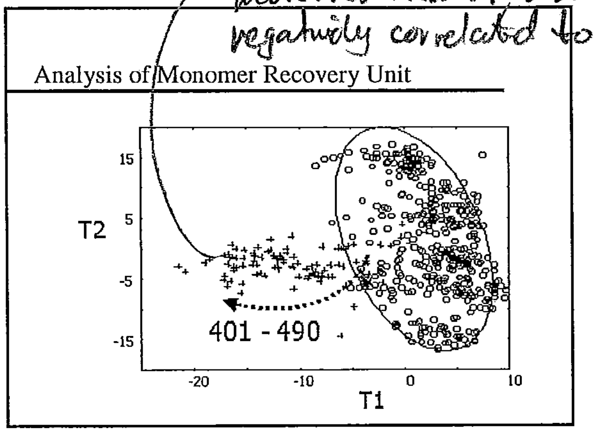
Want: high purity high recovery



Analysis of Monomer Recovery Unit Problems

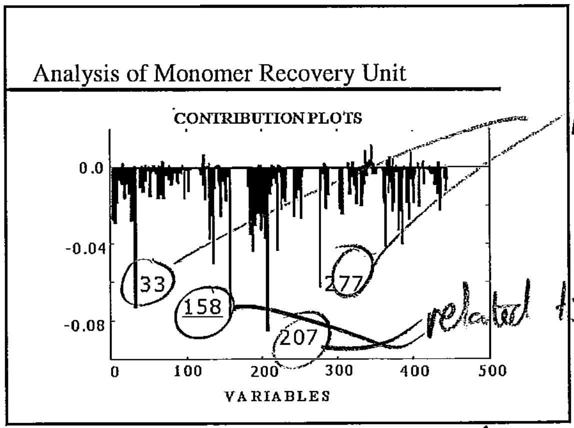
- Problem: Drop in monomer recovery.
- Data on 447 variables
- Daily averages over 498 days
- 15% missing data

What causes this negative movement in T1 which are negatively correlated to T1



Analysis of Monomer Recovery Unit

something has happened for it to diverge off the plane



Analysis of Monomer Recovery Unit

Monomer Δ feed related

related to column 3 temp top

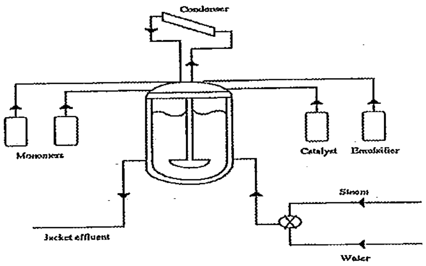
which variables contributed to this movement off the space, Contribution of each column variable to the particular t_i

Analysis of Monomer Recovery Unit

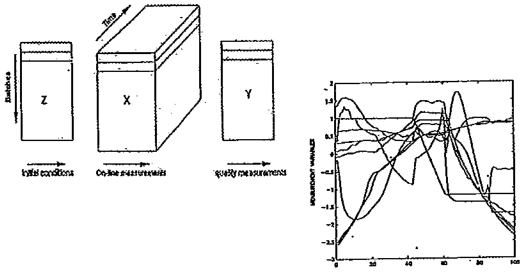
- Contribution plots show:
  1. Recovery low when concentration of A in feed is high (variables 33 & 277)
  2. Problem appears in column number 3 (variables 158 & 207)
  3. Tray #129 temperature has largest contribution (variable 209)
  4. Suggestion: Put controller on tray 129 temperature
  5. It worked. Recovery back to 91%

Even if it just isolates the column that causes problems, then do a DOE about that column, unit etc

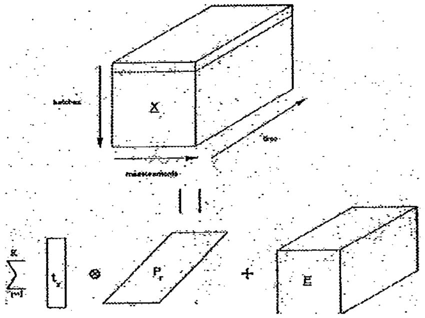
ANALYSIS OF BATCH PROCESS DATA



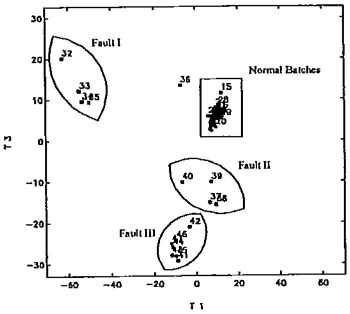
Nature of Batch Data



Multiway PCA



55 Batches from a Nylon Polymerization Process



Analysis of Nylon Polymerization Process:

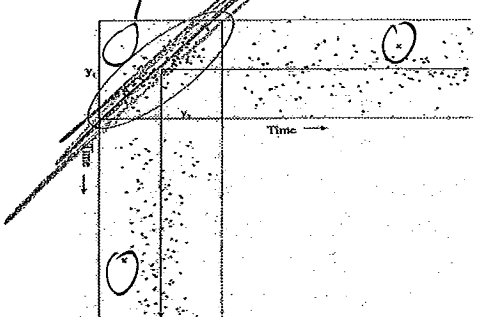
- 3 major groups of bad batches observed.
- Contribution plots showed possible reasons for the different groups of bad batches.
- These were fixed through mechanical and operating policy changes
- Improvements implemented world-wide

PROCESS MONITORING

- Want to monitor processes in real-time.
- Traditional SPC charts consider only one variable at a time and usually only the quality variables Y.
- Lose most of information if ignore the process variables (X)
  - Many more X's ; - X's more precise & more frequent
  - Fingerprints of faults also in X;
  - Need X's to diagnose the problem
- Problem when have many nearly collinear variables. *

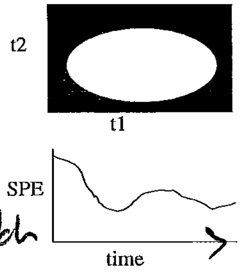
Need for Multivariate SPC

this has shifted outside your normal operating plane



Multivariate SPC Charts

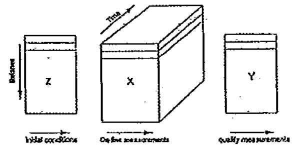
- Represent all information in few new variables:
- Latent variables: t1,t2
- Squared Prediction Error: SPE



SPE limits vary with time for batch processes

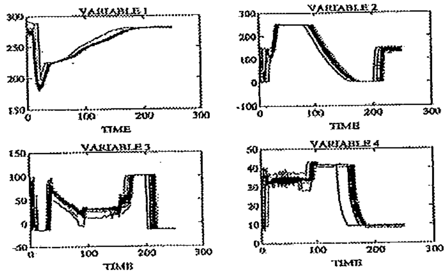
Industrial Polycondensation Reactor

- 61 Batches
- 14 Initial Set-up variables (Z)
- 10 Process variables (X) at 250 time intervals
- 4 Quality variables (Y)

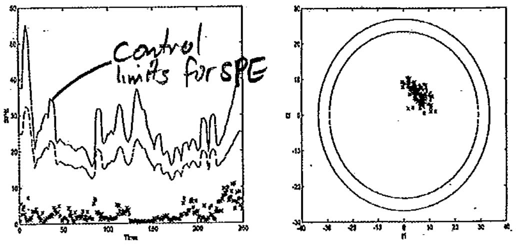


The model; data used in the model (to build it) MUST have a small SPE (No t_1/t_2 off the plane)

Profiles for first 4 variables

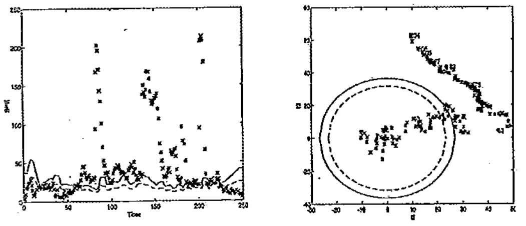


On-line Monitoring of a Good Batch

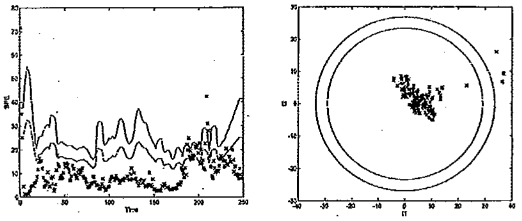


control limits for SPE

On-line Monitoring of a Bad Batch-1



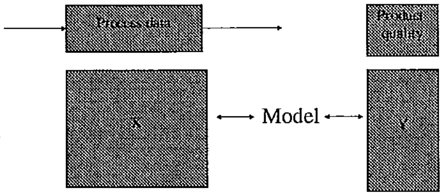
On-line Monitoring of a Bad Batch-2



Product design

- ◆ PROBLEM:
  - Given new product specifications y_{des} find process conditions x_{pred} which will produce this product.
- ◆ Possible Approaches:
  - Theoretical models and optimization
  - Response surface methods -- DOE
  - Statistical methods based on historical data

What do we have?



????????????? ******

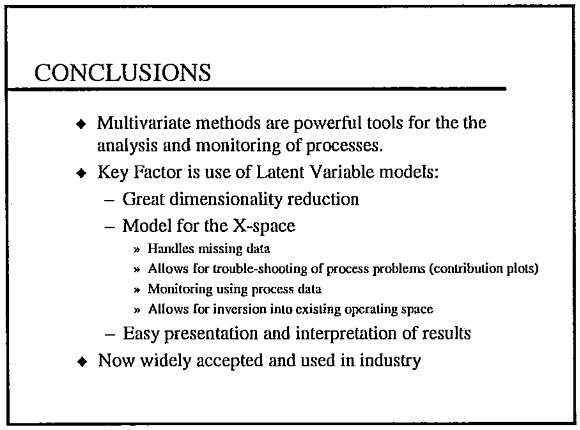
- ◆ Design is an inversion problem

Inversion of Empirical Models

- ◆ Model: $Y = X B$
- ◆ $y^T_{des} = x^T_{new} B$
- ◆ Inversion: $x^T_{new} = y^T_{des} B^{-1}$
- ◆ Problem:
  - More x's than y's
  - Infinite # solutions by inverting MLR, NN models
  - No respect for existing process operating procedures & constraints
- ◆ Latent variable models -- model X as well as Y
  - Inversion gives window of solutions that lie in X space of existing operation
  - Respects past operation and constraints

Several Related Problems

1. Transfer production of a product from one plant to another.
2. Alignment of operating conditions among plants that produce same product.
3. Scale-up from pilot plant to plant reactor.



CONCLUSIONS

- Multivariate methods are powerful tools for the the analysis and monitoring of processes.
- Key Factor is use of Latent Variable models:
  - Great dimensionality reduction
  - Model for the X-space
    - Handles missing data
    - Allows for trouble-shooting of process problems (contribution plots)
    - Monitoring using process data
    - Allows for inversion into existing operating space
  - Easy presentation and interpretation of results
- Now widely accepted and used in industry

$N = 20$ $X^T X = \overset{20\times20}{\cancel{1000}} \Rightarrow$ can only
eg $K=1000$ estimate 20 parameters

ChE 4C3/6C3
Overview of Principal Component Analysis

Professor John MacGregor
McMaster University
Canada

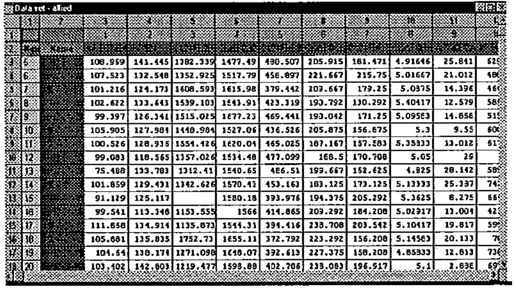
Starting point: Problem ⇒ Data Table X (N x K) Training set

- Data set = table (matrix)
  - N rows (objects, samples, ...)
  - K variables (properties, tests, ...)
- Often many variables -- large K
- Often few observations (K >> N)
- or many of both (N and K large)
- Missing data
- Clusters and collinearity

Objects (cases, samples, rows, ...):
- Analytical samples
- Process time points
- Trials (experim. runs)
- Chemical compounds, ...

Variables (tags, properties, columns):
- Sensors (T, P, flow, pH, conc., ...)
- Spectral amplitudes (NMR, NIR, Raman, UV-Vis, XRF, ...)
- Chromatographic Peaks (HPLC, GC, Electroforesis, ...)

Example of a multivariate data set of A polymerization, N=820 observations, K=160 variables.

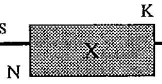


	1	2	3	4	5	6	7	8	9	10	11
1			1	2	3	4	5	6	7	8	9
2	Num	Name	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
3	5		108.969	141.445	1382.339	1477.49	430.507	205.915	181.471	4.91646	25.841
4	6		107.523	132.548	1352.925	1517.79	456.897	221.667	215.75	5.01667	21.012
5	7		101.216	124.173	1608.593	1615.98	379.442	203.667	179.25	5.0375	14.396
6	8		102.622	133.643	1539.103	1543.91	423.319	193.792	130.292	5.40417	12.579
7	9		99.397	126.341	1515.025	1677.23	469.441	193.042	171.25	5.09583	14.656
8	10		105.905	127.984	1440.984	1527.06	436.526	205.875	156.675	5.3	9.55
9	11		100.526	128.936	1554.426	1620.04	465.025	187.167	157.583	5.35333	13.012
10	12		99.083	118.565	1357.026	1534.48	477.099	168.5	170.708	5.05	29
11	13		75.488	133.783	1312.41	1540.65	486.51	193.667	152.625	4.825	28.142
12	14		101.859	129.431	1342.626	1570.43	453.163	183.125	173.125	5.13333	25.337
13	15		91.129	125.117		1580.18	393.976	194.375	205.292	5.3625	8.275
14	16		99.541	113.348	1153.555	1566	414.865	209.292	184.208	5.02917	13.004
15	17		111.658	134.914	1135.873	1544.31	394.416	238.708	203.542	5.10417	19.817
16	18		105.681	135.835	1752.73	1655.11	372.792	223.292	156.208	5.14583	20.133
17	19		104.64	138.174	1271.098	1648.07	392.613	227.375	158.208	4.85333	12.813
18	20		103.402	142.803	1219.477	1598.88	402.706	233.083	196.517	5.1	3.636

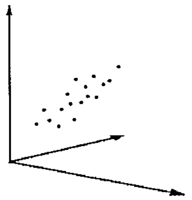
Why Multivariate Analysis by Projection ?
PCA PLS

- Deal with the Dimensionality Problem
- Handles all Types of data tables
  - Short: N<<K Square: N =. K Long: N >> K
- Handle collinearities
- Handle missing data
- Robust to noise in both X and Y
  - noise can be non-random
- Separates regularities from noise:
  - Models X & models Y
  - Models relation between X and Y
- Extracts information from all data simultaneously: MVSPC
- Results displayed graphically

Data table = K-space & N points



Transformation (log, square root, ...) may first be needed

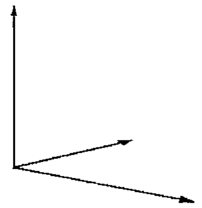


- K-dimensional space (K variables)
- Each variable defines a coordinate axis, length s_k
- Object/observation = point in K-space
- data table = swarm of points in K-space

Data tables, matrices, rows = obs, columns = variables

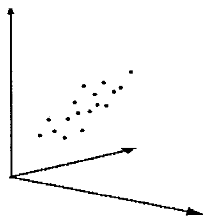
- Data table = matrix
  = array with N rows and K columns
- Centering = subtracting column averages,
  → columns that vary around zero
- Scaling; usually dividing columns by their SD (unit variance)
- PCA
  - Summarizing rows → linear combinations row-wise of X (scores, t_a)
  - Summarizing col.s → linear combinations col.-wise of X (loadings, p_a)

PCA -- Geometric Interpretation, 1 X-space



- For the matrix of the *training set*, X, we construct a space, with K dimensions
- This is called X-space
- Each variable has one coordinate axis, with the length determined by its scaling, usually unit variance (UV, $s_k^2 = 1$)

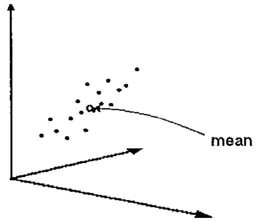
PCA -- Geometric Interpretation, 2 points



- Each training set object is represented by one point in X-space.
- The "training" data matrix X hence is as swarm of points in this space

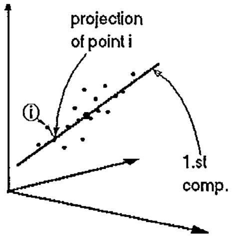
---

PCA -- Geometric Interpretation, 3 average



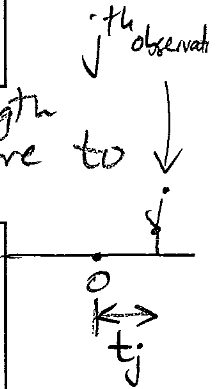
- First, we calculate the average of each variable. The vector of variable averages is a point in X-space.
- This average is subtracted from the the data matrix, corresponding to moving the origin of the coordinate system to the middle of the data swarm.

PCA -- Geometric Interpretation, 4 1.st PC

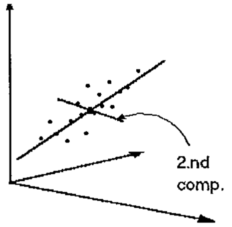


- The first principal component (PC) is a line in X-space that best approximates the data (least squares).
- The line goes through the average point.
- The direction of the line is determined by the loading vector p_1 (elements p_{1k})
- The coordinates of the points (i) are t_{i1}

Values of t represent projection length ~~to the line~~ from the centre to the line

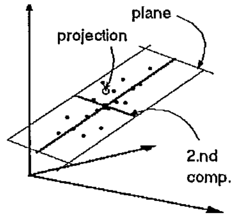


PCA -- Geometric Interpretation, 5 2.nd PC

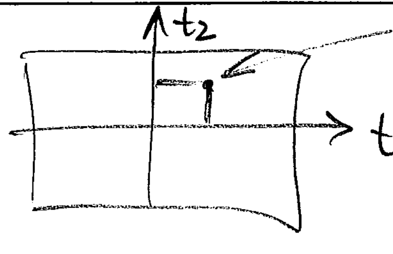


- The second PC is a line in X-space orthogonal to the line of the 1.st component
- Also goes through the average point.
- This line improves the approximation of the data points as much as possible.

PCA -- Geometric Interpretation, 6 PC-plane



- The principal components (PC.s) together form a plane (or hyper-plane) in X-space.
- The variability around the X-(hyper)plane is used to calculate a tolerance interval within which new objects similar to the training set will be situated.
- Projection = window in K-space



projected point onto the plane defined by p_1 & p_2

t_1 & t_2 represent the ⊥ projected point to the plane

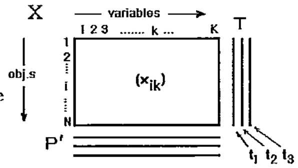
Projection -- PCA

The **scores t_{ia}** (comp. **a**, object **i**) are the places along the lines where the objects are projected

The scores, $\mathbf{t_{ia}}$, are *new variables* that best **summarize** the old ones; that are **combinations** of the old ones with coefficients $\mathbf{p_{ak}}$

Sorted on importance, $\mathbf{t_1}$, $\mathbf{t_2}$, $\mathbf{t_3}$,...

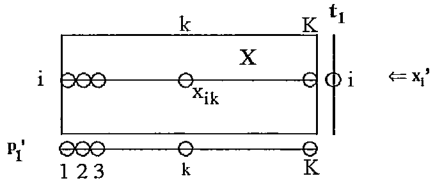
$$\boxed{X = T P' + E}$$



---

a score; a linear combination side-wise (row-wise) = weighted average of all values for the row

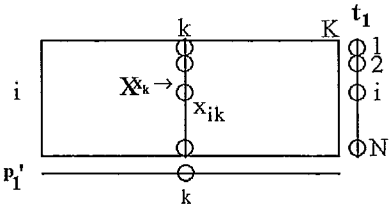
$$t_{1i} = x_{i1}\, p_{11} + x_{i2} p_{12} + x_{i3} p_{13} + \ldots + x_{ik} p_{1k} + \ldots + x_{iK} p_{1K} = x_i'\, p_1$$



---

a loading (and PLS-weight) = weighted col. average = a linear combination vertically, column-wise

$$p_{1k} = t_{11} x_{1k} + t_{12} x_{2k} + t_{13} x_{3k} + \ldots + t_{1i} x_{ik} + \ldots + t_{1N} x_{Nk} = t_1'\, x_k$$



Loading p_{ik} = importance of variable k in first PC

How to estimate the loadings (p) & scores (t)

- p_i's are the eigenvectors of X^TX matrix
- Eigenvalues λ_i's are variances of t_i's $\lambda_1 > \lambda_2 > \lambda_3$
- Can calculate them recursively using NIPALS-algorithm

---

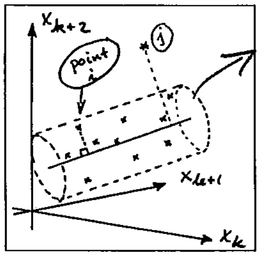
for one principal compent

Observation (row) residuals,

K

N $\boxed{E = X - TP'}$

- Row Residuals = "left-overs"
- How well model fits an observation or row in X
- Row SD, distance to the model (DModX) = $s_i = [\Sigma_k e_{ik}^2/(K\text{-}A\text{-}1)]^{1/2}$
- Critical Distance: Dcrit $= s_0 * \text{sqrt } F$



confidence interval

$$s_i = \sqrt{\frac{\Sigma\, e_{ik}^2}{K-A-1}}$$

Column (variable) residuals

Measures of size of residuals:

- R_k^2 measures how well the models describes the variable (k)
- Q_k^2 measures the predictive power (cross-validation) of var. k
- R^2 and Q^2 -- same, but over all variables, whole matrix

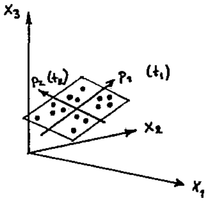
$R^2 = 1 - [SS_{resid} / SS_{data}]$ SS = sum of squares

$Q^2 = 1 - [SS_{predictive\ resid} / SS_{data}] = 1 - [PRESS / SS_{data}]$

t_i = coordinates in plane
p_i = define the plane's axes

PCA, overview of a data table (data set):

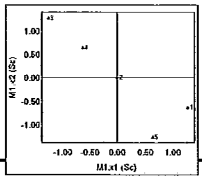
1. Transformation (optional)
2. Centering -- subtract column averages
3. Scaling -- usually, divide by col. SD.s
4. PCA = least squares projection of data onto (hyper)-plane
5. scores, t , are coordinates in the (hyper)-plane
6. loadings, **p**, define the direction of the (hyper)-plane



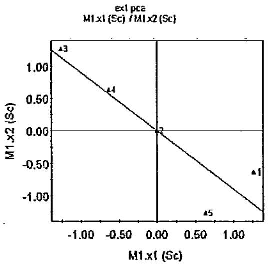
$$X = T P' + E$$

Ex.1: Two variables x1 ,x2
Data with Centering and scaling the variables

	X1	x2	x1c	x2c	x1cs	x2cs
•	98	45	2	-1	1.26	-0.63
•	96	46	0	0	0	0
•	94	48	-2	2	-1.26	1.26
•	95	47	-1	1	-0.63	0.63
•	97	44	1	-2	0.63	-1.26
Avg	96	46				
SD	1.58	1.58				
ws	0.63	0.63				



The first PC score (t_1)
Combination of x1 and x2 with equal (but opposite) weights



x1 and x2 well correlated:

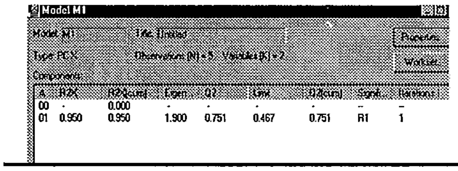
When x1 decreases, x2 increases, and vice versa

t_1 = projection onto line

(position along the line, relative to the center)

Ex.1: Weights (loadings) of x1 and x2 in the first PC
(how x1 and x2 combine to form the new variable t1)

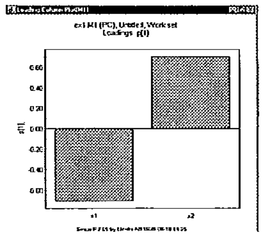
- p11 = weight of x1 = - 0.71
- p12 = weight of x2 = 0.71
- t1 = -0.71 x1 + 0.71 x2



Model: M1 Title: Untitled
Type: PCX Observations (N) = 5 Variables (K) = 2
Components:

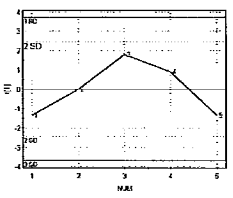
A	R2X	R2X(cum)	Eigen	Q2	Limit	Q2(cum)	Signif.	Iterations
00	.	0.000	.	.	.	.	–	–
01	0.950	0.950	1.900	0.751	0.467	0.751	R1	1

Loadings p and score t $t = x*p$

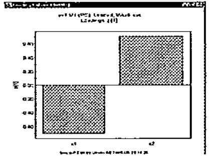


x1	x2	t1
1.26	-0.63	-1.34
0	0	0
-1.26	1.26	1.79
-0.63	0.63	0.89
0.63	-1.26	-1.34
p1	p2	
-0.707	0.707	

Interpreting score plot t1 using the loadings
scores (t): $t = X*p$ $t_i = \Sigma_k\ p_k\ x_{ik}$



x1sc	x2sc	t1
1.26	-0.63	-1.34
0	0	0
-1.26	1.26	1.79
-0.63	0.63	0.89
0.63	-1.26	-1.34

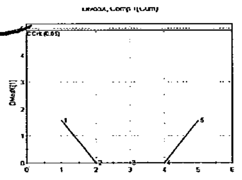


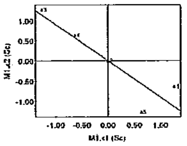
$\Sigma_k\ p_k^2 = 1.0$ (normalization)

-0.71	0.71	(p_1)

→ 95% CI

Residuals: $E = X - t p'$ $e_{ik} = x_{ik} - t_i p_k$
Distance to the model (row residual SD)





e1	e2	DModX
0.32	0.32	0.20
0	0	0
0	0	0
0	0	0
-0.32	-0.32	0.20

How many components (A) in PCA ?

- Cross-validation, continue until Q^2 does not improve (check one component beyond)
- Eigen-values larger than 1 or 2 (eigen-value = % SS explained * K)
- Plot of eigen-values ("scree plot") (subjective)
- To make $R^2 > 0.9$ or 0.95 (risky)
- Others

- For graphical display, use two or three components

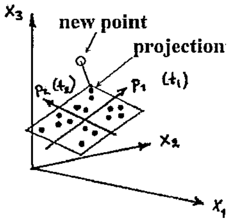
Cross-validation (CV)
a way to assess the predictive power of a model

- A number of rounds of model fitting, with parts of data matrix kept out.
- In each round, the kept out part is then predicted from the model.
- PRESS is sum of squared differences between predicted and observed x-elements, over all rounds of model fitting

$$PRESS = \Sigma (x_{ik} - x_{pred,ik})^2$$

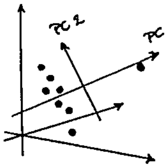
$$Q^2 = 1 - [PRESS / SS_{data}]$$

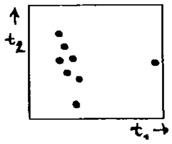
PCA, predictions (new objects = prediction set)

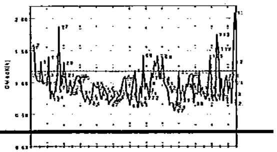


- New point (x_j') row vector [x_{j1} x_{j2} .... x_{jK}]
- Coordinates on PC plane: [t_{1j} t_{2j} t_{3j} ... t_{Aj}] = t_j'
- Residuals ($e_j' = x_j' - t_j' P$) row vector [e_{j1} e_{j2} .... e_{jK}] summarized as $SD_j = s_j$ = $DModX_j$ (distance to model)

Outliers: Easily seen in score plots, A < 5 (serious ones)





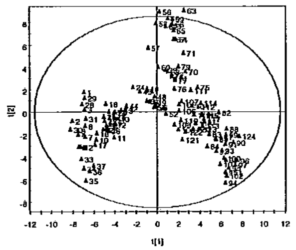


And in DModX (moderate outliers)

Other inhomogeneities (strong groups, clusters)

- Also seen in score plots

- Groups/clusters
- Similaritities
- Classification

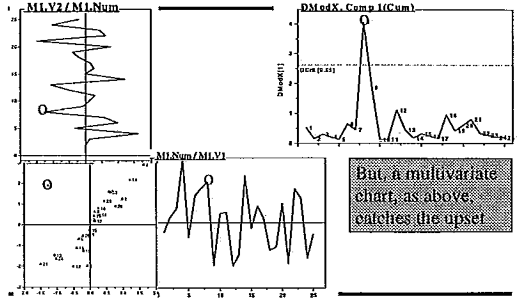


Multivariate Statistical Process Control (MSPC)

- Use all process and quality variables simultaneously
- Monitor the process
- Detect changes
- Find assignable causes

- Multivariate charts offer tremendous advantage over traditional univariate charts
  - Look at more than just magnitude of each variable
  - Look also at relationships among all variables

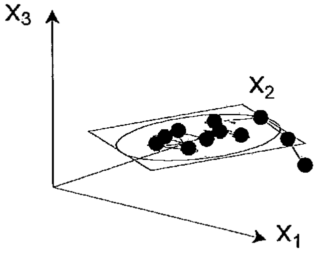
Separate charts of correlated variables may miss serious upsets, #8 This risk increases with the number of variables, K.



But, a multivariate chart, as above, catches the upset

MSPC

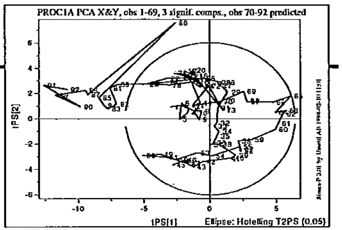
Process is "OK" if data is close to the model plane)



Deviations:

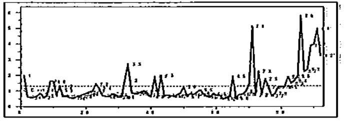
- Away from plane if DModX is large
- In plane: Scores outside Hotelling ellipse (t_a normally distributed)
- Displayed as ordinary control charts

M-SPC



A projection of the 33 dim. space onto a 2-dim. plane

shows evolution in time of ALL the variables



and a summary of the deviations (the distance between the data and model)

The "contribution plots": Show what has happened in the individual observation

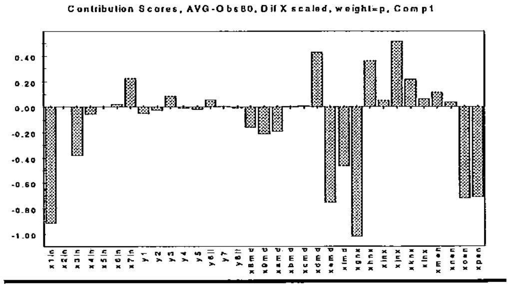
- A large residual (DModX), e.g., point 86, is suspect

  we look at the residuals $z_k = e_{86,k}$
  times a weight ($v_k = sqrt(R_k^2)$)
- A score value (e.g. point 80) is suspect

  we look at the scaled data $z_k = (x_{80,k} - xavg_k)*ws_k$
  times a weight ($v_k = p_k$, or, $v_k = sqrt(R_k^2)$)
- Contributions = $z_k\ v_k$; for k=1,2,…,K (each variable)
- The "contribution" plots identify "culprit" variables

Why is obs. 80 outside the "normal" area in t-space?

Contribution plot of data (here X&Y) row-wise
(observation - mean vector)* weights [w = p1]



Use of PCA; one table only (called X)

- Overview, summary of X
  - Graphics: -scores, loadings, contribution, DModX
  - use T as descriptors of the objects instead of X
  - Great reduction in dimension
- similarities, groups,
- MSPC
- classification

AREAS OF APPLICATION FOR MULTIVARIATE METHODS

1. Analysis of historical data (Process trouble-shooting)
   Both batch and continuous processes
2. Multivariate Statistical Process Control (MSPC)
3. Soft sensors / Inferential models
4. Multivariate calibration (e.g. NIR spectrometers)
5. Classification / Pattern recognition

AREAS OF APPLICATION FOR MULTIVARIATE METHODS

6. Multivariate time series analysis
7. Quantitative Structure Activity Relations (QSAR) and
   Quantative Structural Property Relations (QSPR)
8. Multi-Spectral Image Analysis
9. Other Areas:
   1. Product design / Model inversion problems
   2. Multivariate specifications