Design of Experiments

1. Introduction

1.1. Terminology

1.2. Types of Experimental Designs

Simple designs:

- Vary one factor at a time.
- Not statistically efficient.
- Wrong conclusions if the factors have interaction

Full factorial design:

- All combinations
- Can find the effect of all factors.
- Too much time and money. •

> E.g., 2 factorial where each of n factors has 2 levels. # of experiments = 2^n.

Fractional factorial designs:

- Less than full factorial design.

- Saves time and expenses.
- Less information.
- May not get all interactions.
- Not a problem if negligible interactions.

2. One Factorial Design and ANOVA

2.1. Assumptions

（1）每一总体均为正态总体，记为 $N(\mu_i,\sigma_i^2)$，$i=1,\cdots,r$.

（2）各总体的方差相同，记为 $\sigma_1^2=\sigma_2^2=\cdots=\sigma_r^2=\sigma^2$.

（3）从每一总体中抽取的样本是相互独立的，即所有的试验结果 y_{ij} 都相互独立.

这三个假定都可以用统计方法进行验证. 譬如，利用正态性检验（§7.5 节）验证（1）成立，利用后面§8.3 的方差齐性检验验证（2）成立，而试验结果 y_{ij} 的独立性可由随机化实现，这里的随机化是指所有试验按随机次序进行.

在水平 A_i 下的试验结果 y_{ij} 与该水平下的指标均值 μ_i 一般总是有差距的，记 $\varepsilon_{ij}=y_{ij}-\mu_i$，$\varepsilon_{ij}$ 称为随机误差. 于是有

$$y_{ij}=\mu_i+\varepsilon_{ij}. \qquad (8.1.2)$$

（8.1.2）式称为试验结果 y_{ij} 的**数据结构式**. 把三个假定用于数据结构式就可以写出单因子方差分析的统计模型：

$$\begin{cases} y_{ij}=\mu_i+\varepsilon_{ij}, \quad i=1,2,\cdots,r, \quad j=1,2,\cdots,m, \\ \text{诸 } \varepsilon_{ij} \text{ 相互独立，且都服从 } N(0,\sigma^2). \end{cases} \qquad (8.1.3)$$

2.2. Target

我们要做的工作是比较各水平下的均值是否相同，即要对如下的一个假设进行检验，

$$H_0:\mu_1=\mu_2=\cdots=\mu_r, \qquad (8.1.1)$$

其备择假设为

$$H_1:\mu_1,\mu_2,\cdots,\mu_r \text{ 不全相等},$$

在不会引起误解的情况下，H_1 通常可省略不写.

如果 H_0 成立,因子 A 的 r 个水平均值相同,称因子 A 的 r 个水平间没有显著差异,简称因子 A **不显著**;反之,当 H_0 不成立时,因子 A 的 r 个水平均值不全相同,这时称因子 A 的不同水平间有显著差异,简称因子 A **显著**.

2.3. Lemma

2.4. Parameter Evaluation

2.4.1. Point-Based

$$\hat{\mu} = \bar{y},$$

$$\hat{a}_i = \bar{y}_{i\cdot} - \bar{y}, \quad i = 1, \cdots, r,$$

$$\hat{\sigma}_M^2 = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{m} (y_{ij} - \bar{y}_{i\cdot})^2 = \frac{S_e}{n}.$$

由最大似然估计的不变性,各水平均值 μ_i 的最大似然估计为

$$\hat{\mu}_i = \bar{y}_{i\cdot}, \quad (8.$$

由于 $\hat{\sigma}_M^2$ 不是 σ^2 的无偏估计,实用中通常采用如下误差方差的无偏估计

$$\hat{\sigma}^2 = MS_e. \quad (8.$$

2.4.2. Confidence Interval

以下讨论各水平均值 μ_i 的置信区间. 由定理 8.1.2 知,$\bar{y}_{i\cdot} \sim N(\mu_i, \sigma^2/m)$,$S_e/\sigma^2 \sim \chi^2(f_e)$,且两者独立,故

$$\frac{\sqrt{m}\,(\bar{y}_{i\cdot} - \mu_i)}{\sqrt{S_e/f_e}} \sim t(f_e),$$

由此给出 A_i 的水平均值 μ_i 的 $1-\alpha$ 的置信区间为

$$\bar{y}_{i\cdot} \pm \hat{\sigma} \cdot t_{1-\alpha/2}(f_e)/\sqrt{m}, \quad (8.1.23)$$

Parameter	Estimate	Variance
μ	$\bar{y}_{..}$	$\boxed{s_e^2/ar}$
α_j	$\bar{y}_{.j}$-$\bar{y}_{..}$	$s_e^2(a-1)/ar$
s_e^2	$\boxed{\frac{\sum e_{ij}^2}{a(r-1)}}$	

Q: what is std

2.5. Hypothesis Check

偏差平方和 Q 的大小与数据个数(或自由度)有关，一般说来，数据越多，其偏差平方和越大. 为了便于在偏差平方和间进行比较，统计上引入了**均方**的概念，它定义为

$$MS=\frac{Q}{f_Q},$$

其意为平均每个自由度上有多少平方和.

如今要对因子平方和 S_A 与误差平方和 S_e 之间进行比较，用其均方

$$MS_A=\frac{S_A}{f_A},\quad MS_e=\frac{S_e}{f_e}$$

进行比较更为合理，因为均方排除了自由度不同所产生的干扰. 故用

$$F=\frac{MS_A}{MS_e}=\frac{S_A/f_A}{S_e/f_e} \tag{8.1.17}$$

对给定的 α，可作如下判断：

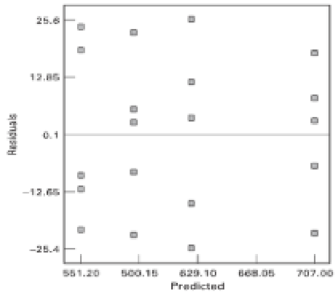
- 如果 $F\geqslant F_{1-\alpha}(f_A,f_e)$，则认为因子 A 显著.
- 若 $F<F_{1-\alpha}(f_A,f_e)$，则说明因子 A 不显著.

2.6. Assumptions Check

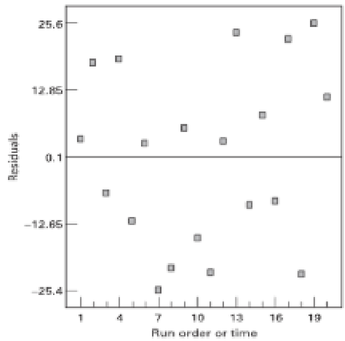
2.6.1. Errors Independent

1. Independent errors

a) Scatter plot of residuals versus the predicted response



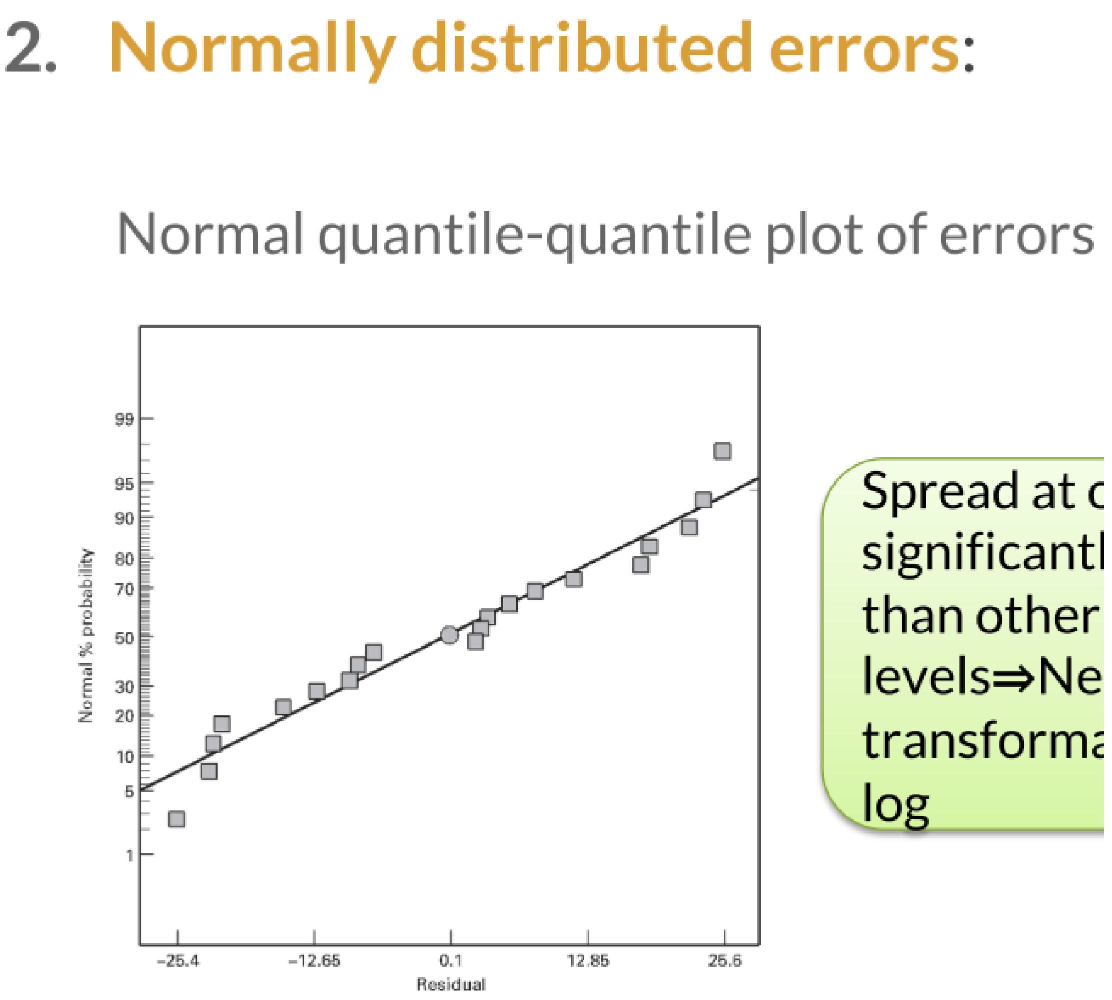
a) Plot the residuals as a function of the experiment number



Trend up or down ⇒ other factors or side effects.

2.6.2. Normally distributed errors

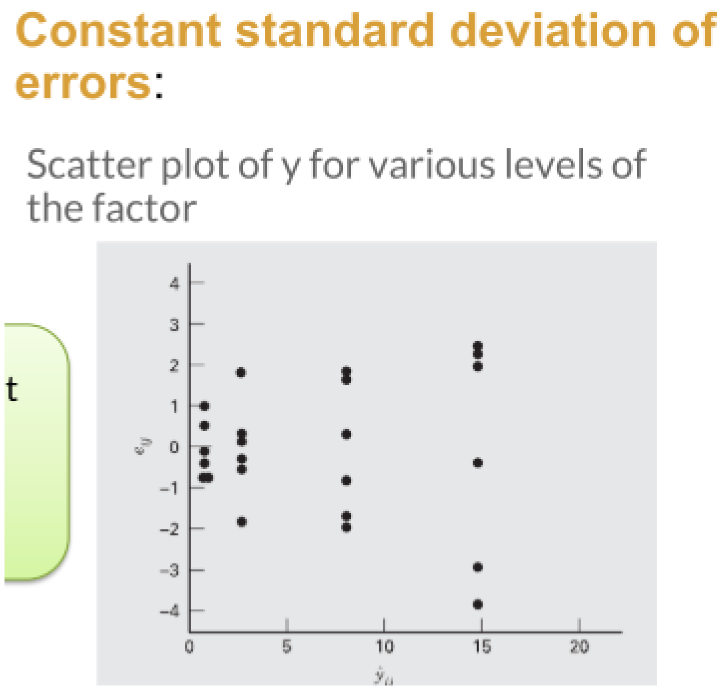
Using Q-Q graph



> 统计学里Q-Q图（Q代表分位数）是一个概率图，用图形的方式比较两个概率分布，把他们的两个分位数放在一起比较。首先选好分位数间隔。图上的点（x,y）**反映出其中一个第二个分布（y坐标）的分位数和与之对应的第一分布（x坐标）的相同分位数**。因此，这条线是一条以分位数间隔为参数的曲线。
> **如果两个分布相似，则该Q-Q图趋近于落在y=x线上**。如果两分布线性相关，则点在Q-Q图上趋近于落在一条直线上，但不一定在y=x线上

2.6.3. Constant Standard Devaiation of Errors

This part is to test errors are i.i.d, especially focus on the standard deviation



3. Two Factor Design

3.1. About Log Transformation

When the max and min values in the exprements is very large, we always do a log transformation

3.2. Assumptions

并设：$X_{ijk} \sim N(\mu_{ij}, \sigma^2)$，$i=1, 2, \ldots, r$，$j=1, 2, \ldots, s$，$k=1, 2, \ldots, t$，各$X_{ijk}$独立。这里，$\mu_{ij}$和$\sigma^2$均为未知参数，或写成

$$\left.\begin{array}{l} X_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \\ \varepsilon_{ijk} \sim \int N(0, \sigma^2)\text{，各}\varepsilon_{ijk}\text{独立，} \\ i=1,2,\cdots,r, j=1,2,\cdots, s, k=1,2,\cdots, t\text{。} \end{array}\right\}$$

3.3. Target

$$\begin{cases} H_{01}: \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0 \\ H_{11}: \alpha_1, \alpha_2, \cdots, \alpha_r \text{不全为零} \end{cases} \tag{7.17}$$

$$\begin{cases} H_{02}: \beta_1 = \beta_2 = \cdots = \beta_s = 0 \\ H_{12}: \beta_1, \beta_2, \cdots, \beta_s \text{不全为零} \end{cases} \tag{7.18}$$

$$\begin{cases} H_{03}: \gamma_{11} = \gamma_{12} = \cdots = \gamma_{rs} = 0 \\ H_{13}: \gamma_{11}, \gamma_{12}, \cdots, \gamma_{rs} \text{不全为零} \end{cases} \tag{7.19}$$

3.4. Models

1. Model: With *r* replications-

α_j : Effect of factor A
β_j :Effect of factor B
γ_{ij} :Effect of interaction A&B.

$$y_{ijk} = \mu + \boxed{\alpha_j} + \boxed{\beta_i} + \boxed{\gamma_{ij}} + e_{ijk}$$

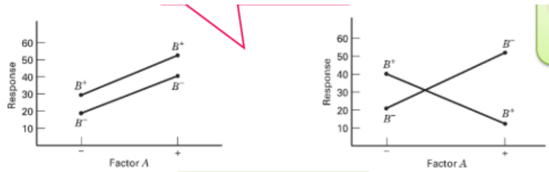
$$\begin{cases} \sum_{j=1}^{a} \alpha_j = 0; \sum_{i=1}^{b} \beta_i = 0; \\ \sum_{j=1}^{a} \gamma_{1j} = \sum_{j=1}^{a} \gamma_{2j} = \cdots = \sum_{j=1}^{a} \gamma_{bj} = 0 \\ \sum_{i=1}^{b} \gamma_{i1} = \sum_{i=1}^{b} \gamma_{i2} = \cdots = \sum_{i=1}^{b} \gamma_{ia} = 0 \\ \sum_{k=1}^{r} e_{ijk} = 0 \quad \forall i, j \end{cases}$$

Q: What about distribution of error?

$$e_{ijk} \sim N(0, \sigma)$$

in which

γ_{ij} is the interaction of A&B



Right figure has interaction

3.5. Parameter evaluation

3.5.1. Point-Based Evaluation

How to estimate the parameters

$$\bar{y}_{ij.} = \mu + \alpha_j + \beta_i + \gamma_{ij}$$
$$\mu = \bar{y}_{...}$$
$$\alpha_j = \bar{y}_{.j.} - \bar{y}_{...}$$
$$\beta_i = \bar{y}_{i..} - \bar{y}_{...}$$
$$\gamma_{ij} = \bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$$

Check the textbook for derivation.

3.5.2. Confidence Interval

Parameter Estimation

Be careful

Parameter	Estimate	Variance
μ	$\bar{y}_{...}$	s_e^2/abr
α_j	$\bar{y}_{i..}$-$\bar{y}_{...}$	$s_e^2(a-1)/abr$
β_i	$\bar{y}_{.j.}$-$\bar{y}_{...}$	$s_e^2(b-1)/abr$
γ_{ij}	$\bar{y}_{ij.}$-$\bar{y}_{i..}$-$\bar{y}_{.j.}$+$\bar{y}_{...}$	$s_e^2(a-1)(b-1)/abr$
s_e^2	$\Sigma\ e_{ijk}^2/\{ab(r-1)\}$	

Degrees of freedom for errors = ab(r-1)

Same to Single Factor Situation

3.5.3. T test approximation

When freedom degree is larger than 32, T distribution is similar to Normal Distribution

3.6. Hypothesis Check

$$SST = SSY - SS0 = SSA + SSB + SSAB + SSE$$

$$\sum_{ijk}(y_{ijk} - \bar{y}_{...})^2 = \sum_{ijk} y_{ijk}^2 - \sum_{ijk} \bar{y}_{...} = \sum_{ijk} \alpha_i^2 + \sum_{ijk} \beta_j^2 + \sum_{ijk} \gamma_{ij}^2 + \sum_{ijk} e_{ijk}^2$$

Source	SSY	SS0	SSA	SSB	SSAB	SSE
Degree of freedom (V)	abr	1	a-1	b-1	(a-1)(b-1)	ab(r-1)

- $\frac{SSA/v_A}{SSE/v_e} \sim F\ [a-1, ab(r-1)\ ]$
- $\frac{SSB/v_B}{SSE/v_e} \sim F\ [b-1, ab(r-1)\ ]$
- $\frac{SSAB/v_{AB}}{SSE/v_e} \sim F\ [(a-1)(b-1), ab(r-1)\ ]$

4. 2^k Factorial Designs

4.1. Background

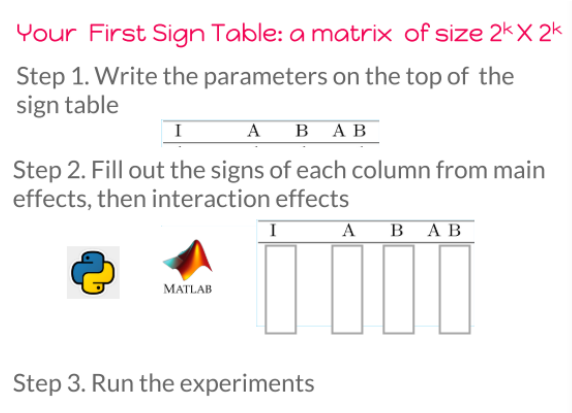
- **K factors**, each at two levels which can be quantitative or qualitative

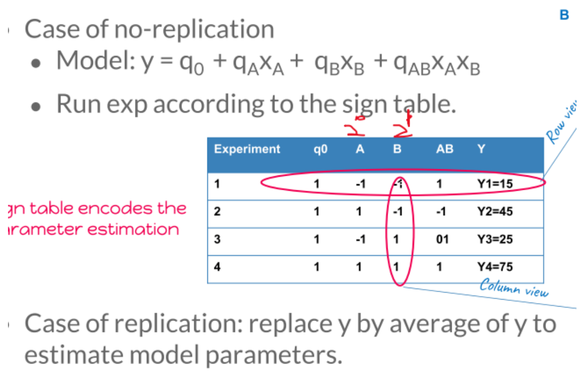
4.2. Coding Scheme and Sign table

Coding Scheme

- We use **(+1,-1)** as the coded variables to denote high and low values of factor
- For a 2^k design
  - The sign table has 2^k columns corresponding to parameters including **intercept, main effects, two factor interactions...up to kfactor interactions**.
  - It has 2^k rows, corresponding to experiment types
  - It can use to estimate 2^k parameters

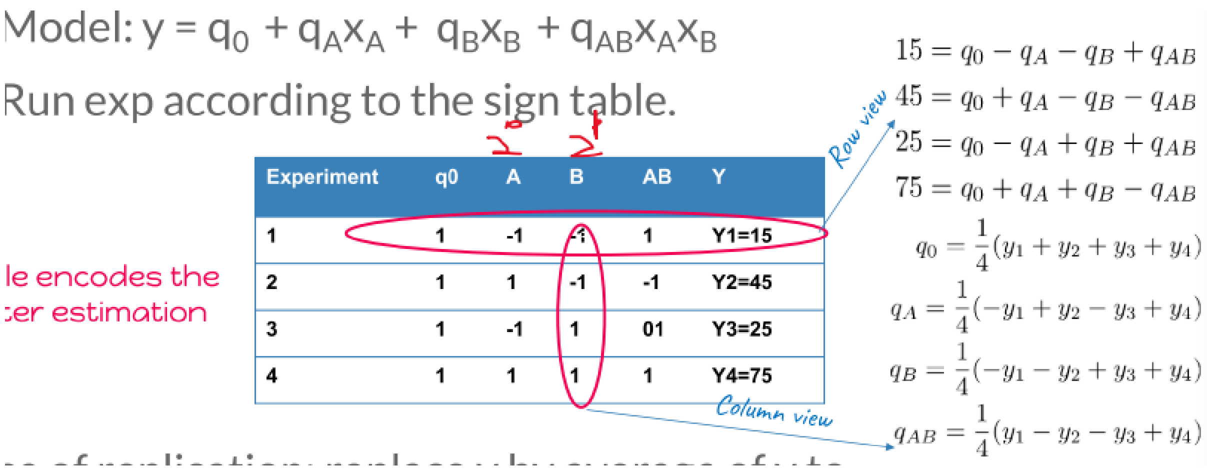
Sign Table





- When First Building the Sign
- When building the Sign Table, first column is parameter of q_0, that is always 1
- Each Line after use binary encoding way, first {+1,-1,+1,-1}, next{+1,+1,-1,-1,+1,+1,-1,-1}... larger the cycles

Parameter Estimation



Property

- Fractional factorial designs also use orthogonal vectors.
- The sum of the products of any two columns is zero
- The sum of the squares of each column is 22.

Back to Natrual variables

$$X_{Ai}=\frac{X_{\text{natural}}-(\text{ low }+\text{ high })/2}{(\text{ high }-\text{ low })/2}$$

4.3. ANOVA for 2^2 design

4.3.1. Hypothesis Check

- Error = Measured value- estimated values

$$\begin{aligned} e_{ij} &= y_{ij}-\hat{y}_i \\ &= y_{ij}-q_0-q_A x_{Ai}-q_B x_{Bi}-q_{AB} x_{Ai} x_{Bi} \end{aligned}$$

- Total Variation

$$SST=\sum_{ij}(y_{ij}-\bar{y}_{..})=SSA+SSB+SSAB+SSE$$

- F-Test
  - Hypothesis: Tested factors have no effect on the response y
  - Compute F0, and compare it with F tables with significance α

$$\frac{SSA/v_A}{SSE/v_e},\ \frac{SSB/v_B}{SSE/v_e},\ \frac{SSAB/v_{AB}}{SSE/v_e}$$

Q: Which factors could you test and how?

$$y_{i,j}=q_0+q_A x_{A_i}+q_B x_{B_i}+q_{AB} x_{A_i} x_{B_i}+e_{ij}$$

- Total variation or total sum of squares:

$$SST=\sum_{ij}(y_{ij}-\bar{y}_{..})^2=\sum_{ij}y_{ij}^2-\sum_{ij}\bar{y}_{..}^2$$

$$SST=\sum_{ij}(y_{ij}-\bar{y}_{..})^2=2^2rq_A^2+2^2rq_B^2+2^2rq_{AB}^2+\sum_{ij}e_{ij}^2$$

→SST=SSA+SSB+SSAB+SSE

Derivation (II)

$$\sum_{ij}y_{ij}^2=\sum_{ij}q_0^2+\sum_{i,j}q_A^2x_{A_i}^2+\sum_{i,j}q_B^2x_{B_i}^2+\sum_{i,j}q_{AB}^2x_{A_i}^2x_{B_i}^2+\sum_{ij}e_{ij}^2$$

Because of row/column sum zero and the orthogonal, each cross product term is zero

$$\boxed{\begin{aligned} \mathrm{SSY} &= \mathrm{SS0}+\mathrm{SSA}+\mathrm{SSB} \\ &\quad +\mathrm{SSAB}+\mathrm{SSE} \end{aligned}}$$

20

For 2 factors, 3 replication, we have freedom shown in right place

- Degrees of freedom of SSA, SSB, SSAB, SSE, and SST

1 , 1, 1, 8, 11

- Compute F0 for all factors

$$\frac{SSA/1}{SSE/8}, \frac{SSB/1}{SSE/8}, \frac{SSAB/1}{SSE/8}$$

4.3.2. Confidence Interval

$$\boxed{q_i \mp t_{[1-\alpha/2;2^2(r-1)]} s_{q_i}}$$

Q: The difference betwe
F-test?

vation

- Assumption

$$\text{error} \sim \mathcal{N}(0, \sigma_e) \qquad \text{y} \sim \mathcal{N}(\bar{y}, \sigma_e)$$

It's for each parameter, where
is for the significance of overa
factors

- Estimate of variance

$$\hat{\sigma_e}^2 = s_e^2 = \frac{1}{2^2(r-1)} \sum_{ij} e_{ij}^2 = \frac{SSE}{2^2(r-1)}$$

  - Denominator = 2^2(r-1) = # of independent terms in SSE,SSE has 2^2(r-1) degrees of freedom.
  - Similarly, $s_{q_A} = s_{q_B} = s_{q_{AB}} = \frac{s_e}{\sqrt{2^2 r}}$

4.4. General 2^k Factorial Design

- Model:

are
and j?

$$y_{ij} = q_0 + q_A x_{Ai} + q_B x_{Bi} + q_{AB} x_{Ai} x_{Bi} + \cdots + e_{ij}$$

- Parameter estimation:

$$q_j = \frac{1}{2^k} \sum_{i=1}^{2^k} S_{ij} \bar{y}_i$$

MATLAB

S_{ij} = (i,j)th entry in the sign table

Q: Example of 3 factors
model and SST
decomposition ?

- ANOVA

$$\text{SSY} = \sum_{i=1}^{2^k} \sum_{j=1}^{r} y_{ij}^2$$
$$\text{SS0} = 2^k r q_0^2$$
$$\text{SST} = \text{SSY} - \text{SS0}$$
$$\text{SS}j = 2^k r q_j^2 \text{j} = 1, 2, \ldots, 2^k - 1$$
$$SST = SSY - SSO = \sum_j SSj + SSE$$

$$y_{i,j} = q_0 + q_A x_{A_i} + q_B x_{B_i} + q_B x_{C_i} + q_{AB} x_{A_i} x_{B_i} + q_{AC} x_{A_i} x_{C_i} + q_{BC} x_{B_i} x_{C_i} + q_{ABC} x_{A_i} x_{B_i} + x_{A_i} + e_{ij}$$

$$SST = SSA + SSB + SSC + SSAB + SSAC + SSBC + SSABC$$

26

- Percentage of y's variation explained by *j*th effect =

$$(\mathrm{SS}j/\mathrm{SST}) \times 100\%$$

- Standard deviation of errors:

$$s_e = \sqrt{\frac{\mathrm{SSE}}{2^k(r-1)}}$$

- Standard deviation of effects:

$$s_{q_0} = s_{q_A} = s_{q_B} = s_{q_{AB}} = s_e/\sqrt{2^k r}$$

5. 2^{k-p} Fractional Factorial Designs

5.1. Background

Because for sign table, k factors we need 2^k experiements, in another word, 2^k experiements can calculate at most 2^k parameters

Large number of factors

- large number of experiments
- full factorial design too expensive
- Use a fractional factorial design

So we want to use something to analyze k factors with only 2^{k-p} experiments:

- 2^{k-1} design requires only 1/2 as many experiments
- 2^{k-2} design requires only 1/4 of the experiments

- Design Table of 2^{7-4} generated by by DoE
- Study 7 factors with only 8 experiments!

Expt No.	A	B	C	D	E	F	G
1	-1	-1	-1	1	1	1	-1
2	1	-1	-1	-1	-1	1	1
3	-1	1	-1	-1	1	-1	1
4	1	1	-1	1	-1	-1	-1
5	-1	-1	1	1	-1	-1	1
6	1	-1	1	-1	1	-1	-1
7	-1	1	1	-1	-1	1	-1
8	1	1	1	1	1	1	1

Q: What's the maximum numb of parameter can we learn frc these 8 experiments?

Q: Which parameters are included in the model

Q: How to decide?

5.2. Property of Fractional Design Features

- Fractional factorial designs also use orthogonal vectors. The sum of each column is zero.

$$\sum_i x_{ij} = 0; \forall j$$

- The sum of the products of any two columns is zero.

$$\sum_i x_{ij} x_{il} = 0, \forall j \neq l$$

- The sum of the squares of each column is 2^{7-4}.

$$\sum_i x_{ij}^2 = 2^{7-4}, \forall j$$

Expt No.	A	B	C	D	E	F	G
1	-1	-1	-1	1	1	1	-1
2	1	-1	-1	-1	-1	1	1
3	-1	1	-1	-1	1	-1	1
4	1	1	-1	1	-1	-1	-1
5	-1	-1	1	1	-1	-1	1
6	1	-1	1	-1	1	-1	-1
7	-1	1	1	-1	-1	1	-1
8	1	1	1	1	1	1	1

5.3. Confounding

Only the combined influence of two or more effects can be computed.

be computed. For example

$$
\begin{aligned}
q_A &= \boxed{\begin{aligned}&\sum_i y_i x_{Ai} \\ =\ &\frac{-y_1+y_2-y_3+y_4-y_5+y_6-y_7+y_8}{8}\end{aligned}} \\
q_D &= \sum_i y_i x_{Di} \\
&= \frac{-y_1+y_2+y_3-y_4+y_5-y_6-y_7+y_8}{8} \\
q_{ABC} &= \sum_i y_i x_{Ai} x_{Bi} x_{Ci} \\
&= \frac{-y_1+y_2+y_3-y_4+y_5-y_6-y_7+y_8}{8}
\end{aligned}
$$

Expt No.	A	B	C	AB	AC	BC	D	
1	-1	-1	-1	1	1	1	-1	y_1
2	1	-1	-1	-1	-1	1	1	y_2
3	-1	1	-1	-1	1	-1	1	y_3
4	1	1	-1	1	-1	-1	-1	y_4
5	-1	-1	1	1	-1	-1	1	y_5
6	1	-1	1	-1	1	-1	-1	y_6
7	-1	1	1	-1	-1	1	-1	y_7
8	1	1	1	1	1	1	1	y_8

$$q_{ABC} = q_D$$

Q: ABC is confounded with D, Problem?

Q:More confounding? How to find?

Algebra of Confounding (How to find?)

Rules

- I is treated as unity
- Any term with a power of 2 is erased

Example. I=ABCD, multiplying A on both sides
→IA = A^2BCD
→A = BCD

Steps

- Selecte the Base, for example I=ABCD
- multiply all combination on both side

Properties

- 2^{k-p} experiment, 2^p different design
- A Fractional Factorial Design is Not Unique

Selection of Confounding Plans

Rersolution of design:

- Number of different terms on both side
- The order of confounding **never change** in a Plan no matter what

Order of an effect = Number of terms. E.g., Order of I=***ABCD*** = 4 becaus
order of I = 0
Order of a confounding = Sum of order of two terms. E.g., AB=CDE is of
AB=ADE
order ??? 5

you multiplied on both side

<u>Design Resolution</u>

The **larger** Resolution, the **better,**

That potentially because, larger resolution means we take more lower-order interaction/factors into consideration