Chemometrics

MA4605

Week 9. Lecture 17. Design of Experiments

November 1, 2011

Design of Experiments terminology

- Factor is any aspect of the experimental conditions which may affect the result obtained form an experiment.
- Controlled Factor is any factor that can be altered by the experimenter at will.
- Uncontrolled Factor is any factor that can not be freely altered.
- Factor Levels are the discretized values of indicating the degree of presence of a given factor (for example, *high* and *low*).

Stages of the Experimental Design

- identifying the factors which may affect the result of an experiment
- designing the experiment so that the effects of uncontrolled factors are minimized
- using statistical analysis to separate and evaluate the effects of the various factors involved

Blocking and Randomization

- *Blocking* is a fundamental concept in good experimental design and it is employed when an investigator is aware of the presence of extra sources of variation in addition to the treatments: the day in which an experiment is carried out.
- The blocking process groups the experimental units into clusters in an attempt to improve the comparison of treatments
- A cluster of results that contains one measurement for each treatment(measurement on each day) is known as block
- *Randomization* can then be used to reduce the variability from the remaining sources, such as the order in which the experiments are carried each day
- Its purpose is to ensure the layout of the experiment does not consistently favor one or the other treatment

Two-way ANOVA

We can extend ANOVA to include more than one factor(one-way ANOVA).

- Two factors.
- One factor **A** with k levels (called treatments), another factor **B** with b levels (called blocks).
- Three sources of variation: between treatments, between blocks, experimental variation.
- No interaction between the two factors.

Two-way ANOVA sum of squares

The total variability is partitioned into three components:

- the variability due to the different treatments (k)
- the variability due to the different blocks(b)
- the error variability (residuals)

$$
\begin{aligned}
SS_{Total} &= SS_A + SS_B + SS_{Residuals} \\
SS_{Total} &= \sum_{i=1}^{b}\sum_{j=1}^{k}(y_{ij} - \bar{\bar{y}})^2 \\
SS_A &= b\sum_{j=1}^{k}(\bar{y}_{.j} - \bar{\bar{y}})^2 \\
SS_B &= k\sum_{i=1}^{b}(\bar{y}_{i.} - \bar{\bar{y}})^2 \\
SS_{Residual} &= SS_{Total} - SS_A - SS_B
\end{aligned}
$$

Example of a randomized block design

Samples from five different suspensions of bacteria A,B,C,D,E, were examined under a microscope by four different observers I,II,II,IV; the order in which each observer dealt with the samples was randomized to reduce errors due to fatigue, and the number of organisms recorded from the samples are summarized below.

Observer number	A	B	C	D	E	Means
I	68	71	54	95	73	72.2
II	82	78	67	116	85	85.6
III	77	74	65	103	88	81.4
IV	59	70	54	90	76	69.8
Means	71.5	73.25	60	101	80.5	77.25

Computations

Compute the total sum of squares

$$
\begin{aligned}
SS_{Total} &= \sum_{i=1}^{b}\sum_{j=1}^{k}(y_{ij} - \overline{\overline{y}})^2 \\
&= (68 - 77.25)^2 + (71 - 77.25)^2 + ... + (76 - 77.25)^2 \\
&= 4717.75
\end{aligned}
$$

Compute the sum of squares between treatments

$$
\begin{aligned}
SS_A &= b\sum_{j=1}^{k}(\overline{y}_{\cdot j} - \overline{\overline{y}})^2 \\
&= 4(71.5 - 77.25)^2 + 4(73.25 - 77.25)^2 + 4(60 - 77.25)^2 \\
&\quad +4(101 - 77.25)^2 + 4(80.5 - 77.25)^2 = 3685
\end{aligned}
$$

Compute the sum of squares between blocks

$$
\begin{aligned}
SS_B &= k\sum_{i=1}^{b}(\overline{y}_{i.} - \overline{\overline{y}})^2 \\
&= 5[(72.2-77.25)^2 + (85.6-77.25)^2 + (81.4-77.25)^2 + (69.9-77.25)^2 \\
&= 839.75
\end{aligned}
$$

Compute the residual sum of squares

$$
\begin{aligned}
SS_{Residual} &= SS_{Total} - SS_A - SS_B \\
&= 4717.75 - 3685 - 839.75 = 193
\end{aligned}
$$

Degrees of freedom

The associated degrees of freedom: for between the blocks variation **b-1**, for between the treatments variation **k-1.**
The number of degrees of freedom associated with residuals:
kb-1-(b-1)-(k-1) = **kb-b-k +1**.
The total number of degrees of freedom: **kb-1** = **N-1.**

Analysis of Variance Table
Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	4	3685.0	921.25	57.280	1.018e-07
B	3	839.8	279.92	17.404	0.0001145
Residuals	12	193.0	16.08		

Q1: are the differences between treatment means significant?
Q2: are the differences between block means significant?
A: Both of the F-ratios are highly significant.
The implications are

- there is a significant variation between suspension types
- there is a significant variation between observers

Two Way ANOVA. Example 2

In an experiment to compare the percentage efficiency of different chelating agents (A,B,C,D) in extracting a metal ion from aqueous solution, the following results were obtained.

	Chelating agent			
Day	**A**	**B**	**C**	**D**
1	84	80	83	79
2	79	77	80	79
3	83	78	80	78

On each day a fresh solution of metal ion was prepared and the extraction performed with each chelating agents taken in random order.

In this experiment

- Controlled factor = chelating agent
  - chosen by the experimenter
  - 4 levels
- Uncontrolled factor = day on which the experiment is performed
  - differences in lab temperature, pressure on different days can not be freely altered
  - 3 levels

With two-way ANOVA we can both test for a significant effect due to the controlled factor, and to estimate the variance due to the uncontrolled(random) factor.

Two-way ANOVA model

The observations(percentage efficiency) are assumed to follow the mathematical model:

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$

where

- μ is the overall mean efficiency regardless of the chelating agent or the day in which the experiment is carried out
- τ_i is the chelating agent effect
- β_j is the effect of the day
- ϵ_{ij} is the error term

i= 1...b = 1...3 and j=1...k = 1...4

Test the significance of two factors

The two-way ANOVA tests two sets of hypotheses:
The chelating agents effects

- $H_0 : \tau_A = \tau_B = \tau_C = \tau_D$
- $H_a : \tau_A \neq \tau_B \neq \tau_C \neq \tau_D$(at least one τ_i is different)

The effect of day

- $H_0 : \beta_1 = \beta_2 = \beta_3$
- $H_a : \beta_1 \neq \beta_2 \neq \beta_3$(at least one β_i is different)

Two-way ANOVA output in R

```
y<-c(84, 80, 83, 79, 79, 77 ,80, 79, 83, 78, 80, 78)
A<- rep(1:4,3)
[1] 1 2 3 4 1 2 3 4 1 2 3 4
B<-rep(1:3,each=4)
[1] 1 1 1 1 2 2 2 2 3 3 3 3
model <- lm(y ~ A+B)
anova(model)
```

summary(lm(y $\sim$ A+B))
Analysis of Variance Table
Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	k-1=3	28.6667	9.5556	5.8305	0.03276
B	b-1=2	15.5000	7.7500	4.7288	0.05848
Residuals	(b-1)(k-1)=6	9.8333	1.6389		

— — —

The effect of treatments is significant since the p-value for treatments is $0.03276 < 0.05$.
If we ignore the blocking effect (not consider the different days):
summary(lm(y $\sim$ A))
Analysis of Variance Table
Response: y

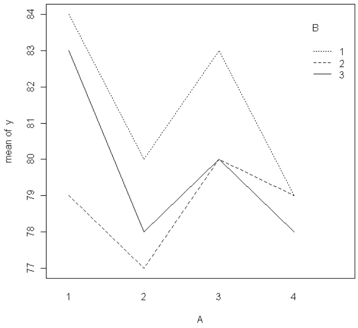
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
A	k-1= 3	28.6667	9.5556	3.0175	0.09405
Residuals	n-k= 8	25.333	3.1667		

— — —

No effect of treatments has been detected without blocking since the p-value for treatments is $0.09405 > 0.05$

Interaction plot

The two-way ANOVA model we used assumes there are no interactions between the two factors: chelating agent and day. We can visually inspect this claim by plotting the results grouped by the two factors.



- The lines are not parallel, indicating the presence of interactions between the two factors: chelating agent and day.
    - the effect of the chelating agent on the efficiency of metal extraction is dependent on the day in which the experiment is done.
- The lines are not quite horizontal, indicating that the efficiency of extraction the metal ion from aqueous solution is dependent on the chelating agent.
- The lines are at different heights on the graph, indicating that the efficiency varies from day to day.

We must consider a more complex model in which we can account for interactions between factors.

Interaction example

No interactions:

	A	B
Day 1	80	82
Day 2	77	79

Day and chelating agent are independent.

Interactions:

	A	B
Day 1	80	82
Day 2	77	83

The difference between the two agents depends on the day of the measurement. The results on the two days depends on the chelating agent used.

Interactions require replicates

- The presence of interactions involves more parameters to be estimated which can be only done if there is a sufficient number of data.
- It requires to replicate the observations with all other factors that possibly affect the experiment randomized.