Chapter 3 Designs with One Source of Variation

3.1 Introduction and Outline

A completely randomized design: Experiments units are assigned to the treatments at random.

Randomization: Implementation on computers

Model

Estimation and testing

Sample size

Using SAS
Randomization Using Computer Programs

Can be done using any software that is capable of generating uniformly distributed random variables.

A SAS program to randomly assign 3 treatments to 13 units. Suppose that treatment 3 shall have 5 units and treatments 1 and 2 each have 4 units.

```sas
*Randomization;
data design1;
input treatment @@;
randomv=ranuni(1234);
lines;
1 1 1 1 2 2 2 2 3 3 3 3 3
;
proc print;
proc sort data=design1;
by randomv;
proc print;
run;
```

Note the number 1234 in ranuni(1234) is the random seed. By specifying a random seed, you are able to replicate the random numbers. The random seed could be any positive integer. ranuni(0) shall generate different random variable at each run. SAS output is:

<table>
<thead>
<tr>
<th>Obs</th>
<th>treatment</th>
<th>randomv</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.03630</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.03912</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.08825</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.08858</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.08948</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.09793</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.10759</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.14312</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.24381</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.25758</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0.38319</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0.44583</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>0.45622</td>
</tr>
</tbody>
</table>

Therefore, units 5, 6, 9 and 11 shall be assigned to treatment 1, units 1, 3, 7 and 10 to treatment 2 and the rest to treatment 3.
3.2. Model for CRD

Let the experiment have \( v \) treatments and the \( it \)th treatment has \( r_i \) replicates. Let \( Y_{it} \) denote the random variable that represents the response obtained on the \( t \)th observation of the \( i \)th treatment.

\[
Y_{it} = \mu_i + \epsilon_{it}, \ t = 1, \ldots, r_i, \ i = 1, \ldots, v
\]

Where \( \epsilon_{it} \sim N(0, \sigma^2) \) are all independent.

Let

\[
\mu = \frac{1}{v} \sum_{i=1}^{v} \mu_i, \tau_i = \mu_i - \mu,
\]

then

\[
\mu_i = \mu + \tau_i, \sum \tau_i = 0.
\]

However, this is only one of many ways to write \( \mu_i = \mu + \tau_i \). But it is always true that

\[
\mu_i - \mu_j = \tau_i - \tau_j.
\]

Another way is to let \( \tau_v = 0, \mu = \mu_v \) and \( \tau_t = \mu_t - \mu_v \). The latter is what SAS implements.

The model can then be written

\[
Y_{it} = \mu + \tau_i + \epsilon_{it}, \ t = 1, \ldots, r_i, \ i = 1, \ldots, v.
\]
3.4 Estimation of Parameters

3.4.1. Notations

The $i$th treatment sample mean is

$$ \overline{Y}_i = \frac{1}{r_i} \sum_{t=1}^{r_i} Y_{it}. $$

The overall mean is

$$ \overline{Y} = \frac{1}{n} \sum_{i=1}^{v} \sum_{t=1}^{r_i} Y_{it}, $$

where $n = r_1 + \cdots + r_v$.

3.4.3 Least Squares Estimation

The least squares estimates of the treatment means are those values that minimize the sum of squared errors:

$$ \sum_{i=1}^{v} \sum_{t=1}^{r_i} (Y_{it} - \mu_i)^2. $$

Hence the LS estimate of $\mu_i$ minimizes $\sum_{t=1}^{r_i} (Y_{it} - \mu_i)^2$.

Solutions:

$$ \hat{\mu}_i = \overline{Y}_i, \text{ for } i = 1, \ldots, v. $$

Hence, the treatment means are estimated by the sample means.

Consequently, a linear combination $\sum_{i=1}^{v} c_i \mu_i$ is estimated by $\sum_{i=1}^{v} c_i \overline{Y}_i$. For example, $\mu_1 - \mu_3$ is estimated by $\overline{Y}_1 - \overline{Y}_3$.

3.4.4 Properties of LSE

1) The LSE is the best linear unbiased estimator (BLUE).

2) $E(\overline{Y}_{i.}) = \mu_i$ (unbiased);

3) $Var(\overline{Y}_{i.}) \leq Var(\sum_{t=1}^{r_i} a_i Y_{it})$ if $\sum_{t=1}^{r_i} a_i Y_{it}$ is unbiased.

For any constants $c_i$, $\sum_{i=1}^{v} c_i \overline{Y}_i \sim N \left( \sum_{i=1}^{v} c_i \mu_i, \sigma^2 \sum_{i=1}^{v} \frac{c_i^2}{r_i} \right)$. 

3.4.5 Estimation of $\sigma^2$

Let us recall an important property: If $Y_1, \cdots, Y_n$ are i.i.d. $N(\mu, \sigma^2)$ and

$$s^2 = \frac{1}{n-1} \sum (Y_i - \bar{Y})^2,$$

then

1) $E(s^2) = \sigma^2$.

2) $(n-1)s^2 / \sigma^2$ has a chi-square distribution with $(n - 1)$ degrees of freedom.

Note the sum of squares for errors

$$SSE = \sum_{i=1}^{v} \sum_{t=1}^{r_i} (Y_{it} - \bar{Y}_i)^2 = \sum_{i=1}^{v} (r_i - 1)S_i^2$$

where

$$S_i^2 = \frac{1}{r_i - 1} \sum_{t=1}^{r_i} (Y_{it} - \bar{Y}_i)^2$$

is the sample variance of the $i$th treatment. Then each $S_i^2$ is an unbiased estimator of $\sigma^2$. Hence

$$E(SSE) = \sum_{i=1}^{v} (r_i - 1)\sigma^2 = (n - v)\sigma^2.$$  

Then the mean squared error

$$MSE = \frac{SSE}{n - v} = \frac{\sum_{i=1}^{v} (r_i - 1)S_i^2}{n - v}$$

is an unbiased estimator of $\sigma^2$. Furthermore, $\frac{SSE}{\sigma^2}$ has a $\chi^2$-distribution with $(n - v)$ degrees of freedom.

Hence, the 95% upper confidence limit for $\sigma^2$

$$\frac{SSE}{\chi_{n-v,0.95}^2}$$

where

$\chi_{n-v,0.95}^2$ is the $5^{th}$ percentile of the chi-square distribution with $(n - v)$ degrees of freedom. These percentiles can be found in SAS and Table A.5 in the book. Please note the discrepancy of notations between p.43 and Table A.5. The SAS program to calculate the 95% percentile of a chi-square distribution with 13 degrees of freedom is given below.
data chisq;
input prob df;
percentile=cinv(prob, df);
lines;
0.05 13;
proc print data=chisq;
run;

Example. In an experiment to study the effects of drivers on the mpg of Toyota Prius, 12 new Prius were randomly assigned to three drivers so that each driver drove four cars and obtained the mpgs. This is a completely randomized design. The data and sample means and sample variances are summarized in the table below.

<table>
<thead>
<tr>
<th>d1</th>
<th>d2</th>
<th>d3</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.33</td>
<td>48.11</td>
<td>49.08</td>
</tr>
<tr>
<td>46.83</td>
<td>50.14</td>
<td>48.89</td>
</tr>
<tr>
<td>51.57</td>
<td>43.22</td>
<td>49.96</td>
</tr>
<tr>
<td>45.33</td>
<td>47.26</td>
<td>49.70</td>
</tr>
<tr>
<td>\bar{y}_i</td>
<td>48.52</td>
<td>47.18</td>
</tr>
<tr>
<td>s_i^2</td>
<td>8.54</td>
<td>8.44</td>
</tr>
</tbody>
</table>

1) Find the least squares estimates for the mean mpg for the three drivers.

2) Find an estimate of the variance \( \sigma^2 \).
The estimate is given by the MSE:

\[
MSE = \frac{1}{n - v} \sum_{i=1}^{3} (\bar{y}_i - 1)s_i^2 = \frac{1}{12 - 3} (3 \times 8.54 + 3 \times 8.44 + 3 \times 0.26) = 5.75.
\]

3) Find the 95% confidence upper limit for \( \sigma^2 \).
First we need to find the 5th percentile for the \( \chi^2 \)-distribution with \( n - v = 9 \) degrees of freedom, which equals \( \chi^2_{0.95} = 16.92 \). The 95% confidence upper limit is given by

\[
\frac{SSE}{\chi^2_{0.95}} = \frac{9 \times 5.75}{3.325} = 15.55.
\]