Designs with One Source of Variation

Ex) only one treatment factor

Hao Zhang

Completely Randomized Design

- Used when we want to compare several treatments and the experimental units are homogeneous.
- No blocking or stratification.

Subjects

★Definition

The experimental units are assigned to the treatments completely at random, subject to the number of replicates for each treatment. The number of $\boxed{\text{replicates}}$ is the number of units assigned to the

★Definition

treatment and equals the number of observations to be taken for the treatment.

How to randomize

Ex) 3 treatments: A (1), B (2), C (3)
Replicates: 4 per treatment → 4×3=12 total # of obs.

Notation: r_i to denote the number of replicates for ith treatment, and $n = \Sigma r_i$. Code the treatments from 1 to v and label the experimental units 1 to n.

v: # of treatments

n: total # of obs.

- Step 1: Enter into one column r_1 1's, then r_2 2's, ..., and finally r_v v's, giving a total of $n = \Sigma r_i$ entries. These represent the treatment labels. 1,1,1,1,2,2,2,2,3,3,3,3

  repeated treatment labels

- Step 2: Enter into another column $n = \Sigma r_i$ random numbers, including enough digits to avoid ties. 12 random #'s
- Step 3: Reorder both columns so that the random numbers are put in ascending order. This arranges the treatment labels into a random order.
- Step 4: Assign experimental unit t to the treatment whose label is in row t.

This ensures treatments are randomly distributed to units.

Model for CRD

Let the experiment have v treatments and the ith treatment has r_i replicates. Let Y_{it} denote the response obtained on the t th observation of the i th treatment.

μ_i = Mean effect of treatment i

$$Y_{it} = \mu_i + \underset{\text{random noise}}{\epsilon_{it}},\ t = 1, \ldots, r_i, i = 1, \ldots, v$$

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

Model parameters: $\mu_i, i = 1, 2, \ldots, v$ and σ^2.

Equivalent treatment effects model:

τ_i = treatment effect

$$Y_{it} = \mu + \tau_i + \epsilon_{it},\ t = 1, \ldots, r_i, i = 1, \ldots, v$$

There are numerous ways to write $\mu_i = \mu + \tau_i$. Hence, τ_i's are not uniquely specified. Unless we impose constraint

- $\mu = (1/v) \sum_i^v \mu_i,\ \tau_i = \mu_i - \mu$.
- $\tau_v = 0,\ \mu = \mu_v,\ \tau_i = \mu_i - \mu_v$. SAS implements this one.

Estimability

Because τ_i's are not unique, they are not estimable. However, if $\sum_i \overset{\text{contrast}}{c_i} = 0$, $\sum c_i\tau_i = \sum_i c_i\mu_i$ as long as $\mu_i = \mu + \tau_i$ for all i. It is usually called a contrast that is estimable. Any linear combination $\sum c_i \mu_i$ such that $\sum c_i = 0$.

Examples:

- $\tau_1 - \tau_2$ "differences between treatment 1 and 2."
- $(\tau_1 + \tau_2)/2 - \tau_3$ "is the average of 1 and 2 better than 3?"
- $(\tau_1 + \tau_2)/2 - (\tau_3 + \tau_4)/2$ "Comparing two groups of treatments."

Everything expressed in terms of treatment means are estimable!

Parameter Estimation

Notations

- Y_{it}: the random variable representing the tth outcome of treatment i.
- y_{it}: the observed value of Y_{it}.
- I might suppress the differences in my notes.
- $$\underbrace{\boxed{\bar{y}_{i\cdot} = (1/r_i)\sum_{i=1}^{r_i} y_{it}}}_{\text{Sample mean for treatment } i},\ \underbrace{\bar{y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{v}\sum_{t=1}^{r_i} y_{it} = \frac{1}{n}\sum_{i=1}^{v} y_{i\cdot} = \frac{1}{n} y_{\cdot\cdot} = \bar{y}_{\cdot\cdot}.}_{\text{overall mean}}$$

  where $n = \sum_{i=1}^{v} r_i$.

Least Squares Estimation: (LSE)

Objective:

Minimize the sum of squared errors (SSE)

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

Solution:

$$\hat{\mu}_i = \bar{y}_{i\cdot}.$$

☆ Sample mean of each treatment is the best estimate of its true mean.

Properties:

- The LSE is the best linear unbiased estimator (BLUE). $E\left(\bar{Y}_{i.}\right) = \mu_i$ (unbiased).
- $\bar{Y}_{i.} \sim N\left(\mu_i, \sigma^2 / r_i\right)$. Normally distributed
- For any constants

$$\sum_{i=1}^{v} c_i \bar{Y}_{i.} \sim N\left(\sum_{i=1}^{v} c_i \mu_i, \sigma^2 \sum_{i=1}^{v} \frac{c_i^2}{r_i}\right)$$

supports hypothesis testing + confidence intervals.

Estimation of σ^2: Residual Variance

Sample variance

$$SSE = \sum_i \sum_t \left(Y_{it} - \bar{Y}_{i.}\right)^2 = \sum_i (r_i - 1) S_i^2$$

$$\hat{\sigma}^2 = SSE/(n - v) = MSE$$ = Variance within treatments

$$E(\hat{\sigma}^2) = \sigma^2.$$

unbiased estimator of σ^2

Upper Confidence Limit of σ^2

Upper Confidence Limit of σ^2

SSE/σ^2 has a χ^2-distribution with $(n-v)$ degrees of freedom. Hence, the 95% upper confidence limit for σ^2

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

where $\chi^2_{n-v,0.95}$ is the 5^{th} percentile of the chi-square distribution with $(n-v)$ degrees of freedom.

SAS Code

Percentiles of distribution

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;
```

The Prius Experiment

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 are given below.

$d1$	$d2$	$d3$
50.33	48.11	49.08
46.83	50.14	48.89
51.57	43.22	49.96
45.33	47.26	49.70

The SAS program to get the sample means and sample standard deviations:

SAS Code

```
data prius;
input driver mpg;
lines;
1 50.33
1 46.83
1 51.57
1 45.33
2 48.11
2 50.14
2 43.22
2 47.26
3 49.08
3 48.89
3 49.96
3 49.70
;
run;
proc print data=prius;
run;
```

SAS Code

```
proc means data=prius;
by driver;
run;
```

Estimation

1. Find an estimate of the variance σ^2. The estimate is given by the MSE:

$$
\begin{aligned}
MSE &= \frac{1}{n-v}\sum_{i=1}^{3}(r_i-1)\,s_i^2 \qquad (1)\\
&= \frac{1}{{}_{n}\,12-3}(\overset{v}{3}*\underset{s_1^2}{2.92^2}+\overset{v}{3}*\underset{s_2^2}{2.90^2}+\overset{v}{3}*\underset{s_3^2}{0.51^2}) \qquad (2)\\
&= 5.75. \qquad (3)
\end{aligned}
$$

2. Find the 95% confidence upper limit for σ^2. First we need to find the 5$^{\text{th}}$ percentile for the χ^2-distribution with $\overset{12}{n}-\overset{3}{v}=9$ degrees of freedom, which equals $\chi^2_{9,0.95}=3.325$. The 95% confidence upper limit is given by $\frac{SSE}{\chi^2_{9,0.95}}=\frac{\overset{df}{9}*\overset{MSE}{5.75}}{3.325}=15.55$.