Completely Randomized Designs with One Factor I

Sean Hellingman ©

Design for Data Science (ADSC2030)

shellingman@tru.ca

Winter 2025

Topics

2. Introduction
3. Replication and Randomization
4. Linear Model for CRD
5. Estimation of the Variance of the Experimental Error
6. Contrast Effects
7. Hypothesis Test of No Treatment Effects
8. Word of Caution
9. Exercises and References

Introduction

- **Completely Randomized Designs** (CRD) are used when there is only one factor under study and the experimental units are homogeneous.

- In other words, we are concerned with only one factor and all of our experimental units/subjects are exactly the same.
  - **All other known independent variables are held constant to not bias any effects.**

- Generally used to examine the effects of the changes in the one factor on the response.

Completely Randomized Designs (CRD)

- With one treatment factor, n experimental units are divided randomly into t groups.
- Each of the t groups are then subject to one of the unique levels/values of the treatment factor.
- If $n = tr$ is a multiple of t:
  - Each level of the factor will be applied to r unique experimental units.
  - There will be r replicates of each run (same level of treatment factor).
- If n is not a multiple of t, then there will be an unequal number of replicates.

Illustrative Example 1A

- In an experiment to determine the effect of time to rise on the height of bread dough:
  - Make one **homogeneous** batch of bread dough.
  - The dough would be divided into n pans with an equal amount of dough.
  - The pans of dough would be **randomly** divided into t groups.
  - Each group would rise for a unique amount of time and the height would be recorded for each loaf.

Illustrative Example 1B

- From the bread dough experiment, identify the following:
  - Treatment factor:
  - Experimental unit:
  - Response:
  - Other factors to be held constant:

Replication and Randomization

- Replication and Randomization are techniques used for error control.
- Recall: *Experimental Error* is the difference between the observed response for a specific experiment and the long run average of the results of all experiments conducted using the exact same combination of variables and factors.
- There are broadly two types of errors:
  1. Bias error: tends to remain constant or change in a consistent pattern over the runs in an experimental design.
  2. Random error: changes from one experiment to another in an unpredictable manner and average to be zero.

Replication

- **Replication** of experimental units allows for the variance of the experimental error to be calculated.
- **If the variability among the treatment means is not as large as the experimental error variance then the differences in the results are probably due to the experimental units and not the factors.**
- Impossible to tell if the differences are real or not without replication.
- Replication dictates that there r bread loaves tested for each of the t rise times.

Randomization

- **Randomization** is the random division of experimental units into groups.
- Randomization helps to ensure the experiment is free of biases caused by other lurking variables.
- When the experimental units are randomized, the hypothesis that the treatment effect is 0 can be tested.
- Randomization of the bread dough would prevent lurking variables such as yeast distribution and any trends in the measurement technique from biasing the effect of the time factor.

Randomized Data Collection in R

- We can use base R to construct a randomized design with one factor:
  - `set.seed(2030)` *Reproducibility*
  - `f <- factor(rep( c(level.1, level.2, ..., level.t), each = r))` *Repeat the factor levels for r replications*
  - `fac <- sample(f, t*r)` *Randomization*
  - `eu <- 1:n` *index for each experimental unit*
  - `plan <- data.frame(Unit = eu, FactorLevel = fac)` *Create a data frame of your design*
  - `write.csv(plan, file = "Plan.csv", row.names = FALSE)` *If you want to save your plan*
- *Note:* Only works when n is a multiple of t.

Example 1

- From the bread dough experiment, assume the following:
  - Three rise times (35, 40, and 45 minutes).
  - Four replicate (r) loaves for each rise time.
- In R, set your seed to 7638 and construct a randomized design for the rise time of the bread.

Illustrative Example 2

- Recall the checklist we covered in Introduction & Definitions and apply it to the bread rising experiment.
  1. Objectives
  2. Experimental Units
  3. Response/Dependent Variable
  4. Independent and Lurking Variables
  5. Pilot Tests (hypothetical)
  6. Choose Experimental Design
  7. Determine the number of Replicates
  8. Randomize Experimental Units to Treatment Levels

CRD Mathematical Model

- The mathematical model for the data obtained from a CRD with an unequal number of replicates for each factor level can be written:

$$Y_{ij} = \mu_i + \epsilon_{ij}. \tag{1}$$

- $Y_{i,j}$ is the response for the j^{th} experimental unit subject to the i^{th} level of the treatment factor. $(i = 1, ..., t,\ j = 1, ..., r_i)$
- r_i is the number of replications (EUs) in the i^{th} level of the treatment factor.
- There is a different mean μ_j for each level of the treatment factor.
- The experimental errors ϵ_{ij} are assumed to be independent (randomization) and normally distributed.

CRD Alternative Mathematical Model

- The effects model:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}. \tag{2}$$

- τ_i (effects) represents the difference in the long run average (all possible experiments) and the average at the i^{th} treatment level.
- With the normality assumption:
  - $Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$
  - $\epsilon_{ij} \sim N(0, \sigma^2)$

Estimating μ_i

- Recall that least squares estimation minimizes the error sum of squares.
- It can be shown using least squares:

$$\hat{\mu}_i = \bar{y}_{i\cdot} \ . \tag{3}$$

- Equivalent to maximum likelihood under these assumptions.
- *Does this model look familiar?*

Example 2

- Use the *daewr* package to import the results of the bread CRD experiment (*bread*). (If the package isn't installed the results are included as a vector).
- Estimate the appropriate model to estimate the effects of the rising time factor on the response variable height.
- Interpret your results.

Estimation of σ^2

- The estimate of the variance of the experimental error σ^2 is $ssE/(n-t)$.
- **It is only possible to estimate this variance when replicate experiments at the each factor level are used.**
- From theory: $ssE/\sigma^2 \sim \chi^2_{n-1}$

Estimation of Effects Model

- If we want to estimate the *effects model* the functions are a bit more complicated.

- In R we can use the *gmodels* package to estimate the average difference in the means of the treatment factors:
  - `library(gmodels)`
  - `fit.contrast(model, "factor", coeff)`
    - Where `coeff` is a vector/matrix expressing which contrasts.

- Exmples:

```
cmat <- rbind( "1 vs 4" =c(-1, 0, 0, 1),
    "1+2 vs 3+4"=c(-1/2,-1/2, 1/2, 1/2),
    "1 vs 2+3+4"=c(-3/3, 1/3, 1/3, 1/3))
```

Example 3

- Use the `fit.contrast()` function to estimate the average difference in the means for the first and second levels of the treatment factor from the bread example.
  - You will need to use your estimated model from Example 2.

- **Note:** Theoretically we are estimating $(\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$

Hypothesis Test

- The statistical hypothesis of interest in the CRD model:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_t \quad \text{OR} \quad \tau_1 = \tau_2 = ... = \tau_t$$

$$H_1 : \text{At least two of the values differ.}$$

Sum of Squares about the Mean

- Sum of squares about the mean (total variability):

$$ssTotal = \sum_{i=1}^{t} \sum_{j=i}^{r_i} (y_{ij} - \bar{y}_{..})^2 \tag{4}$$

- Partitioning the sum of squares:

$$ssTotal = ssT + ssE \tag{5}$$

- ssT and ssE both follow a chi-squared distribution

Analysis of Variance Table (ANOVA)

Source	df	Sum of Squares	Mean Squares	F-ratio
Treatment	$t-1$	ssT	msT	$F = msT/msE$
Error	$n-t$	ssE	msE	
Total	$n-1$	$ssTotal$	$msTotal$	

- Where:
  - $msT = ssT/(t-1)$
  - $msE = ssE/(n-t)$
  - $msTotal = ssTotal/(n-1)$

ANOVA

- Under the null hypothesis F follows the F-distribution with $t-1$ and $n-t$ degrees of freedom.
- In R:
  - `model <- aov(response ~ factor, data = data)`
  - `summary(model)`
- The output gives you an ANOVA table and significance of the test.

Example 4

- In R, perform an ANOVA test to determine if there is any significant differences in the means of the response for each of the three factors in the bread example.

- Comment on your results.

Warning

- Statistical software makes it easy to estimate the models and perform the ANOVA test required for CRDs.
- **If the experiment was not properly conducted, the analysis of the data could be completely useless.**
- This may occur when:
  - Replicates are substituted by sub-samples.
  - Experimental units are not properly randomized.

Illustrative Example 3

- Let's say a professor would like to compare two teaching styles on exam scores.
- They teach two classes (morning and afternoon) using two different teaching styles.
- If they treat the individual students as replicates, the results could be totally wrong.
  - **The experimental unit in this case is the entire class** because the teaching method is being applied simultaneously to the entire class.
- They would only have one observation (class average) per class and a proper statistical test cannot be completed.
- Also, this example appears to be without randomization as students may have different reasons for being in a morning or afternoon class.

Exercise 1

- Use R to construct a CRD with one factor for the following situation:
  - Set your seed to 2030.
  - Want to design an experiment to determine the tensile strength of different cotton/synthetic blends.
    - Four cotton percentages (25%, 30%, 35%, and 40%)
    - Five replications (r) for each cotton percentage

Exercise 2

- Assume the following results vector (from Exercise 1):
  - Results = (14,19,25,8,7,25,22,10,10,18,18,19,10,11,19,18,15,23,12,11)
- Estimate the appropriate model to estimate the effects of the cotton content on the tensile strength response variable.
- Use the `fit.contrast()` function to estimate the average difference in the means for the second and fourth levels of the treatment factor.

References & Resources

1. Lawson, J. (2014). *Design and Analysis of Experiments with R (Vol. 115)*. CRC press.

- gmodels
- Box-Cox