Completely Randomized Design (CRD)

Professor: Hammou El Barmi
Columbia University

Introduction

- The simplest randomized experiment for comparing several treatments is the Completely Randomized Design, or CRD.
- As the name implies, the completely randomized design (CRD) refers to the random assignment of experimental units to a set of treatments.
- It is essential to have more than one experimental unit per treatment to estimate the magnitude of experimental error and to make probability statements concerning treatment effects.

- We will study CRDs and their analysis in some detail, before considering any other designs, because many of the concepts and methods learned in the CRD context can be transferred with little or no modification to more complicated designs.
- Here, we define completely randomized designs and describe the initial analysis of results.
- Structure of a CRD: We have k treatments to compare and n units to use in our experiment.
- To have a completely randomized design:
  1. select sample sizes $n_1, n_2, \ldots, n_k$ with $n_1 + n_2 + \ldots + n_k = n$.
  2. choose at random n_1 units at random to receive treatment 1, n_2 units at random rom the $n - n_1$ remaining to receive treatment 2, and so on
  3. Here, the populations refer to conceptual ones in which there is one population for each of the treatments in the experiment.
- For observational studies, random samples are taken from each of the populations of interest.

- Note that complete randomization only addresses the assignment of treatments to units; selection of treatments, experimental units, and responses is also required.
- Completely randomized designs are the simplest, most easily understood, most easily analyzed designs. For these reasons, we consider the CRD first when designing an experiment.

Assumptions of the CRD:

- Samples are independent
- Homogeneous Variance: we shall assume that the populations of interest all have the same variability, i.e. they all have the same variance
- Normality: we assume that each population is normally distributed

Preliminary Exploratory Date Analysis (EDA)

It is generally advisable to conduct a preliminary exploratory or graphical analysis of the data prior to any formal modeling, testing, or estimation. Preliminary analysis could include:

- Simple descriptive statistics such as means, medians, standard errors, interquartile ranges
- Plots, such as stem and leaf diagrams, box-plots, and scatter-plots

- The basis for the statistical analysis of the data is a statistical model which specifies the the probability distribution of the data up to some unknown parameters.
- The problem of comparing treatments is then tackled by
  - Choosing a model (among several possibilities) on the basis of EDA
  - Estimating the parameters of the model (least squares, mle, others)
- Convinient family of models

$$Y_{ij} = f(i) + \epsilon_{ij}, i = 1, 2, \ldots, k, j = 1, 2, \ldots, n_j$$

  where Y_i is the jth response in the ith group and ϵ_{ij}s are iid $N(0, \sigma^2)$. This implies that $Y_{ij} \sim N(f(i), \sigma^2)$.
- Inference focuses on on the mean structure
- Commonly used mean structures:
  - $f(i) = \mu_i$ (separate means or saturated model)
  - $f(i) = \mu$ (common mean, single mean)
  - $f(i) = \theta_0 + \theta_1 z_i$ (dose-response where z_i is the dose of treatment i)

- The separate mean model also called a one-way analysis of variance (ANOVA) is a generalization of the two sample t-test ($k \geq 2$)
- In this case we assume that the populations of interest have the following (unknown) population means and variances

	population 1	population 2	$\cdots$	population k
mean	μ_1	μ_2	$\cdots$	μ_k
variance	σ_1^2	σ_2^2	$\cdots$	σ_k^2

- Goal: test whether $\mu_1 = \mu_2 = \cdots = \mu_k$

- In this case we sometimes express the group means μ_i as $\mu_i = \mu^* + \alpha_i$.
- When this is the case, the model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, i = 1, 2, \ldots, k, j = 1, 2, \ldots, n_j$$

  where ϵ_{ij}s are iid $N(0, \sigma^2)$. This implies that $y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$.
- The constant μ is called the overall mean, and α_i is called the ith treatment effect.
- In this formulation, the single mean model is the situation where all the α_i values are equal to each other: for example, all zero.
- This introduction of μ and α_i seems like a needless complication, and at this stage of the game it really is. However, the treatment effect formulation will be extremely useful later when we look at factorial treatment structures.

- This model is overparametrized (k+1 parameters for k means). Hence the parameters not uniquely determined.
- We solve this problem by imposing a restriction on the set $\{\mu, \alpha_1, \alpha_2, \ldots, \alpha_k\}$. The restriction must be of the form

$$c_0\mu + \sum_{i=1}^{k} c_i\alpha_i = c \qquad \text{where} \qquad c_0 \neq \sum_{i=1}^{k} c_i$$

- Some examples:
  - $\sum_{i=1}^{k} n_i \alpha_i = 0$. In this case

$$\mu = \frac{\sum_{i=1}^{k} n_i \mu_i}{n} \quad \text{and} \quad \alpha_i = \mu_i - \frac{\sum_{i=1}^{k} n_i \mu_i}{n}$$

  - $\alpha_1 = 0$. In this case

$$\mu_1 = \mu \quad \text{and} \quad \alpha_i = \mu_i - \mu_1$$

- Virtually all inferences are the same regardless of which restriction of the above form is used.
- Since the k treatments effects are not free to vary without one restriction, we say that they have $k - 1$ degrees of freedom

CRD-Example

- Suppose we have five medical treatments and ten subjects on each treatment.
- Goal: Compare the treatments in terms of their effectiveness
- If there were two treatments, what would we use?
- We will compare means among treatment groups.
- In the context of CRD, we say these five treatment make one factor with five levels and each level represents a treatment.

CRD-Example

- The Kenton Food Company wished to test four different package designs for a new breakfast cereal.
- Twenty stores, with approximately equal sales volumes, were selected as the experimental units.
- Each package design was assigned to 5 stores. A fire occurred in one store during the study and so this store had to be dropped from the study (sample size = 20-1=19).
- Hence, one of the designs was tested in only 4 stores. Factors such as location and others which could affect sales were kept the same for all the stores in the experiment.

- To answer this question, random samples from each of the k-populations (each population corresponds to a level of the factor) leading to

	sample 1	sample 2	$\cdots$	sample k
size	n_1	n_2	$\cdots$	n_k
sample	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$	$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$	$\cdots$	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$
sample mean	$\bar{Y}_{1\bullet}$	$\bar{Y}_{2\bullet}$	$\cdots$	$\bar{Y}_{k\bullet}$
sample variance	s_1^2	s_2^2	$\cdots$	s_k^2

- The sample means are $\bar{Y}_{1\bullet}, \bar{Y}_{2\bullet}, \ldots, \bar{Y}_{k\bullet}$ and the average response over all the samples is

$$\bar{Y}_{\bullet\bullet} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{\sum_{i=1}^{k} n_i} = \frac{\sum_{i=1}^{k} n_i \bar{Y}_{i\bullet}}{n}$$

where

$$n = \sum_{i=1}^{k} n_i.$$

- The least square estimate of the μ_i is the separate mean model is $\bar{Y}_{i\bullet}$
- The least square estimate of the μ_i^* and α_i in the separate mean model with the restriction $\sum_{i=1}^{k} n_i \alpha_i = 0$ is $\hat{\mu} = \bar{Y}_{\bullet\bullet}$ and $\hat{\alpha}_i = \bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}$
- The standard errors:
  - $se(\bar{Y}_{i\bullet}) = \sqrt{var(\bar{Y}_{i\bullet})} = \sigma/\sqrt{n_i}$
  - $se(\bar{Y}_{\bullet\bullet}) = \sigma/\sqrt{n}$
  - $se(\hat{\alpha}_i) = \sigma\sqrt{\frac{1}{n_i} - \frac{1}{n}}$
  - An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2}{n-k}$$

Example

Example: A forest manager is responsible for the selection and purchase of chainsaws for her field crew. Her primary interest is worker safety. She is provided with data on chainsaw kickback values (degrees of deflection) for 4 brands of chainsaws (A, B, C, D) with 5 observations each (i.e. $n_1 = n_2 = n_3 = n_4 = n_5 = 5$) . The obvious null hypothesis is:

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$$

against H_a : at least two of these means are not equal. Here μ_j is average angle of deflection

Example

A	B	C	D
42	28	57	29
17	50	45	40
24	44	48	22
39	32	41	34
43	61	54	30

- ANOVA is a method for testing whether any of the treatment means are statistically significantly different from the others, i.e. test of $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against H_a : Not H_0 (that is at least two means are not equal)
- An F test is used to test $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against H_a : Not H_0 (that is at least two means are not equal)
- The assumptions needed for the test are analogous to the pooled two sample t-test
- The F-test is computed from the ANOVA table which breaks the spread in the combined data SST (Total Sum of Squares) into two components (or sums of squares): within sum of squares (SSE) and the between sums of square (SSR)

$$SST = SSE + SSR$$

- The Between SS (often called the model Sum of Squares) measures the spread between the sample means

$$SSR = \sum_{i=1}^{k} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$$

- The within SS (often called Error or Residual Sum of Squares) is

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2$$

- Each SS has its own degrees of freedom (df)

$$df(SST) = n - 1 \quad df(SSR) = k - 1 \quad \text{and} \quad df(SSE) = n - k$$

- it is always the case that

$$df(SST) = df(SSR) + df(SSE)$$

- The mean square error for each source of variation is the corresponding SS divided by its df, that is,

$$MSR = \frac{SSR}{k-1} \quad \text{and} \quad MSE = \frac{SSE}{n-k}$$

- Notice that $\hat{\sigma}^2 = MSE$.

The sums of squares and their dfs are neatly arranged into called the ANOVA table

Source	df	SS	MS	F
Model (Between Groups)	k-1	SSR	MSR= SSR/(k-1)	MSB/MSE
Error (Within Groups)	n-k	SSE	MSE= SSE/(n-k)	
Total	n-1	SST		

- The decision on whether to reject $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ is based on the

$$F = \frac{MSR}{MSE}$$

.

- We have $E(MSE) = \sigma^2$ and

$$E(MSR) = \sigma^2 + \frac{\sum_{i=1}^{k} n_i(\mu_i - \mu)^2}{k-1}$$

where

$$\mu = \frac{\sum_{i=1}^{k} n_i \mu_i}{n}.$$

Therefore when H_0 is true

$$\frac{E(MSR)}{E(MSE)} = 1$$

Sometimes we represent the table as follows

Source	df	SS	MS	F
Treatment	k-1	SSR	MSR= SSR/(k-1)	MSB/MSE
Residual	n-k	SSE	MSE= SSE/(n-k)	
Total	n-1	SST		

- Under $H_0, F = MSR/MSE \sim F(k-1, n-k)$ so a size γ test reject H_0 is $F > F_\gamma(k-1, n-k)$
- The p-value$= P_{H_0}(F(k-1, n-k) >$ observed $MSR/MSE)$. Small p-values are evidence that H_0 may be false.

Example

angle	brand
42	a
17	a
24	a
39	a
43	a
28	b
50	b
44	b
32	b
61	b
57	c
45	c
48	c
41	c
54	c
29	d
40	d
22	d
34	d
30	d

Example

$>$ anova(angle $\sim$ brand)

Analysis of Variance Table

Response: angle

	Df	Sum Sq	Mean Sq	F value	$Pr(>F)$
brand	3	1080	360.00	3.5556	0.03823
Residuals	16	1620	101.25		

- Large values of F indicate large variability among the sample means relative to the spread of the data within the samples. That is, large values of of F suggest that H_0 is false
- We reject H_0 is $F > F_\gamma(k-1, n-k)$ or if $p - value < \gamma$.
- For $k = 2$, the F test is equivalent to the pooled two-sample t-test

In the example if we take $\gamma = 0.05$, since the p-value=0.03823, we would reject H_0

Example

- During cooking, doughnuts absorb fat in various amounts.
- A scientist wished to learn whether the amount absorbed depends on the type of fat.
- For each of 4 fats, 6 batches of 24 doughnuts were prepared. The data are grams of fat absorbed by batch.
- Let μ_i = population mean of fat i absorbed per batch of 24 doughnuts.
- The Scientist wishes to test $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ against H_a : Not H_0.

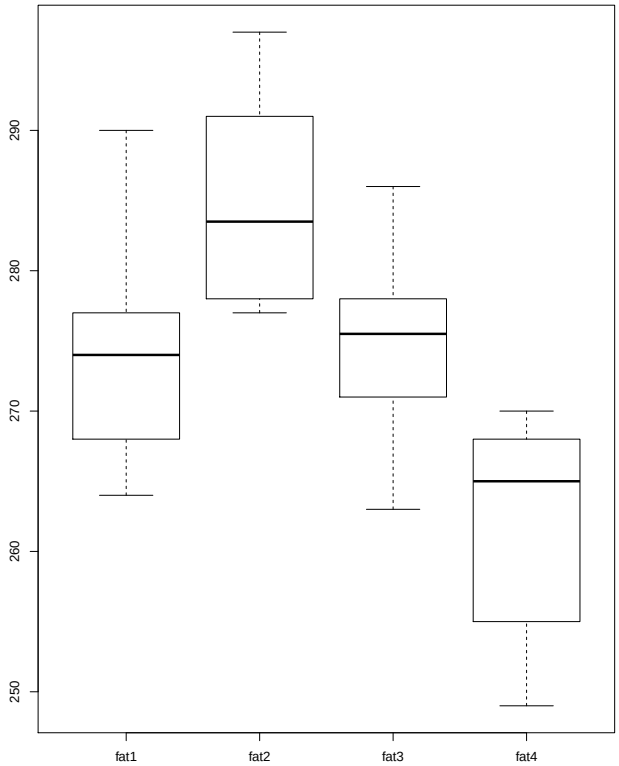
fat 1	fat 2	fat 3	fat 4
264	278	275	255
272	291	286	266
268	297	278	249
277	282	271	264
290	285	263	270
276	277	276	268

Example

```
> fat<-c(rep("fat1",6),rep("fat2",6),rep("fat3",6),rep("fat4",6))
> amount<-c(264,272,268,277,290,276,278,291,297,282,285,277,275,286,278,271,
263,276,255,266,249,264,270,268)
> data<-data.frame(fat,amount)
> summary(data[,2][data[,1]=='fat1'])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  264.0   269.0   274.0   274.5   276.8   290.0

  > summary(data[,2][data[,1]=='fat2'])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  277.0   279.0   283.5   285.0   289.5   297.0

> summary(data[,2][data[,1]=='fat3'])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  263.0   272.0   275.5   274.8   277.5   286.0

> summary(data[,2][data[,1]=='fat4'])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  249.0   257.2   265.0   262.0   267.5   270.0
```

> boxplot(data[,2] data[,1])

Figure: Histogram and Box Plots



Example

```
> summary(fit)
             Df Sum Sq Mean Sq F value Pr(>F)
data[, 1]     3   1596   531.8   7.948 0.0011 **
Residuals    20   1338    66.9
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```

$$SSR = 1596,\ SSE = 1328,\ MSR = 531.8,\ MSE = 66.9, F = 7.95$$

If we take $\alpha = 0.05$, we have $F(0.05, 2, 20) = 3.098391$. since $7.95 > 3.098391$ we reject H_0
Also p-value=0.0011<0.05 we reject H_0

```
> summary(lm(data[,2]~ a,  contrasts = list(a = "contr.sum")))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 274.0833     1.6698 164.143  < 2e-16 ***
a1            0.4167     2.8922   0.144  0.88689
a2           10.9167     2.8922   3.775  0.00119 **
a3            0.7500     2.8922   0.259  0.79804
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 8.18 on 20 degrees of freedom
Multiple R-squared:  0.5438,Adjusted R-squared:  0.4754
F-statistic: 7.948 on 3 and 20 DF,  p-value: 0.001104
```

The fitted model is

$$\hat{Y}_{ij} = 274.0833 + \hat{\alpha}_i$$

where $\hat{\alpha}_1 = 0.4167, \hat{\alpha}_2 = 10.9167, \hat{\alpha}_3 = 0.7500$ and $\hat{\alpha}_4 = -\hat{\alpha}_1 - \hat{\alpha}_2 - \hat{\alpha}_3 = -12.0834$

Reference cell method

- Define $\mu^* \equiv \mu_1$ (reference cell) and $\alpha_i^* = \mu_i - \mu^*$ ($\alpha_1^* = 0$ by definition).
- ANOVA model: $\mu_i = \mu^* + \alpha_i^*, i = 1, 2, \ldots, k$.
- Regression model:

$$Y_{ij} = \mu^* + \alpha_i + \epsilon_{ij}$$

with

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_k \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \mu^* \\ \alpha_2^* \\ \alpha_3^* \\ \vdots \\ \alpha_k^* \end{bmatrix} = X_2\boldsymbol{\beta}$$

- Interesting hypotheses:
  $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \Leftrightarrow H_0 : \alpha_2^* = \alpha_3^* = \ldots = \alpha_k^* = 0$ or $H_0 : C\boldsymbol{\beta} = \mathbf{0}$ where $C = [\mathbf{0}, I_{k-1}]$

```
> lm(data[,2]~factor(data[,1]))

Call:
lm(formula = data[, 2] ~ factor(data[, 1]))

Coefficients:
  (Intercept)      factor(data[, 1])fat2    factor(data[, 1])fat3  factor(data[, 1])fat4
  274.5000                       10.5000                   0.3333               -12.5000
```

The reference mean here is the mean of fat1. The model is

$$\hat{Y} = \begin{cases} 274.5, & \text{if fat 1} \\ 274.5 + 10.5 = 285 & \text{if fat 2} \\ 274.5 + 0.3 = 274.83 & \text{if fat 3} \\ 274.5 - 12.5 = 222 & \text{if fat 4} \end{cases}$$

```
  > summary(lm(data[,2]~factor(data[,1])))

Call:
lm(formula = data[, 2] ~ factor(data[, 1]))

Residuals:
     Min       1Q   Median       3Q      Max
-13.0000  -6.6250   0.6667   4.5000  15.5000

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           274.5000     3.3396  82.196   <2e-16 ***
factor(data[, 1])fat2  10.5000     4.7229   2.223   0.0379 *
factor(data[, 1])fat3   0.3333     4.7229   0.071   0.9444
factor(data[, 1])fat4 -12.5000     4.7229  -2.647   0.0155 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 8.18 on 20 degrees of freedom
Multiple R-squared:  0.5438,Adjusted R-squared:  0.4754
F-statistic: 7.948 on 3 and 20 DF,  p-value: 0.001104
```

Cell mean method (here the cell means are the parameters)

- ANOVA model: $\mu_i = \mu_i$.
- Regression model:

$$Y_{ij} = \mu_i + \epsilon_{ij}$$

with

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_k \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_k \end{bmatrix} = X_3\beta$$

- Interesting hypotheses: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \Leftrightarrow H_0 := \alpha_3^* = \ldots = \alpha_k^* = 0$ or $H_0 : C\boldsymbol{\beta} = \mathbf{0}$ where $C = I_k$

```
> lm(data[,2]~factor(data[,1])-1)

Call:
lm(formula = data[, 2] ~ factor(data[, 1]) - 1)

Coefficients:
factor(data[, 1])fat1  factor(data[, 1])fat2  factor(data[, 1])fat3
                274.5                  285.0                  274.8
factor(data[, 1])fat4
                262.0
```

```
> summary(lm(data[,2]~factor(data[,1])-1))

Call:
lm(formula = data[, 2] ~ factor(data[, 1]) - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-13.0000  -6.6250   0.6667   4.5000  15.5000

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
factor(data[, 1])fat1   274.50       3.34   82.20   <2e-16 ***
factor(data[, 1])fat2   285.00       3.34   85.34   <2e-16 ***
factor(data[, 1])fat3   274.83       3.34   82.30   <2e-16 ***
factor(data[, 1])fat4   262.00       3.34   78.45   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 8.18 on 20 degrees of freedom
Multiple R-squared:  0.9993,Adjusted R-squared:  0.9991
F-statistic:  6742 on 4 and 20 DF,  p-value: < 2.2e-16
```