Factorial Designs I

Sean Hellingman ©

Design for Data Science (ADSC2030)

shellingman@tru.ca

Winter 2025

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- We have covered the case examining changes in a single factor.

- In practice, we may want to study the effects of multiple factors.

- Classical approach: *study each factor separately while holding others constant.*
    - Inefficient, and can cause the experimenter to miss interactions.

- **Factorial designs** accentuate the factor effects, allow for estimation of inter-dependency of effects (or interactions).
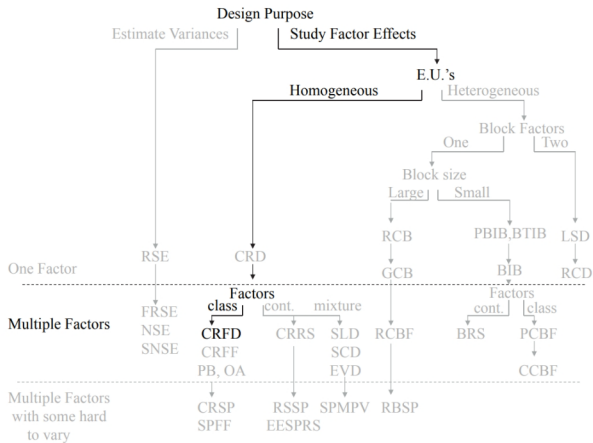   - They are the first technique in the category of what is called treatment design.

Figure: Source: (1)

Factorial Designs

- Factorial Designs examine all possible combinations of factor levels.

- The number of replicates of a specific level of one factor is increased by the product of the number of levels of all other factors in the design.

- The same power or precision can be obtained with fewer replicates.

Illustrative Example 1 I

- Suppose we would like to create paper helicopters from a fixed sheet of paper.

- We then drop the helicopters from a fixed height and time how long it takes for them to hit the ground.

- We can trim paper off of the wings to reduce weight, but this will also reduce wing surface area.

- **We would like to conduct an experiment to determine if changing the wing length changes the flight time**

Paper Helicopter

CUT

FOLD

CUT

FOLD

4"

4.25"
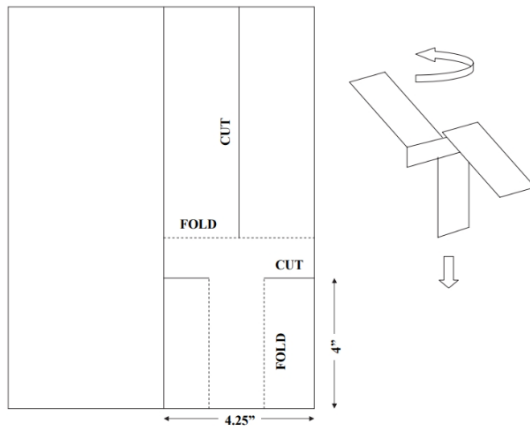
Illustrative Example 1 II

- Identify the following:

  - Experimental unit
  - Difference between replicates and duplicates for this design
  - Treatment factor
  - Possible lurking variables
  - Why randomize

Illustrative Example 2 I

- Now, let us assume that we can also vary the body width of the helicopters, thus giving us two factors.

- Assume that we examine four levels for each of the two factors:
  - Wing Length: 4, 4.75, 5.5, 6 (inches)
  - Body Width: 3.25, 3.75, 4, 4.25 (inches)

- How does the *classical* design approach differ from a *factorial* design?
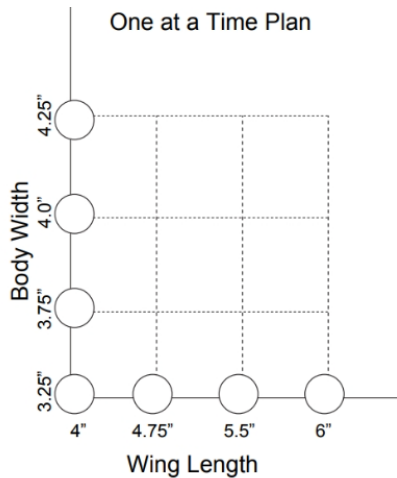
Classic Approach

Figure: Source: (1)

Factorial Design

Figure: Source: (1)
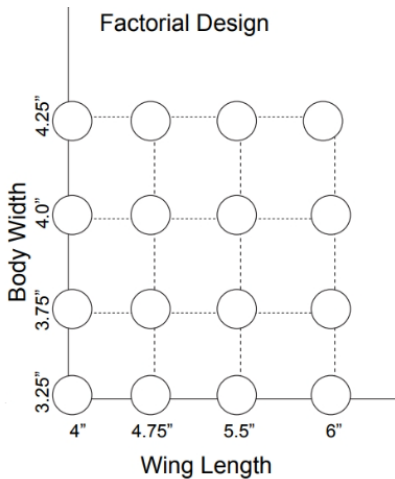
Illustrative Example 2 II

- We can see that experiments are ran at all possible combinations in a factorial design.

- By running two replicates at each combination we get 32 total runs (8 replicates at each individual level).

- We would need to make 56 runs if we took a *classic* approach to achieve the same number of replicates.

- Sometimes referred to as **hidden replication**.

Interpreting Interactions

- If an interaction exists, the effect of one factor on the response will differ depending on the level of the other factor.

- We can examine possible interactions using visualizations using grouped scatterplots.

- Interactions are common in factorial experiments, so it is important to be able to clearly explain the results.

Scatterplots
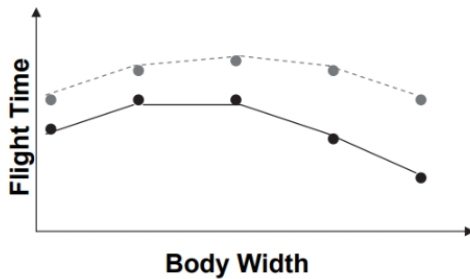
Figure: Source: (1)

Interaction Terms

- **Interaction terms** occur when one explanatory variable modulates the effect of another on the response variable.

- It does **NOT** refer to a relationship between two variables.

- You just have to remember to multiply the coefficient by the product of all the variables in the term.

Example 1

- Assuming the bread experiment was conducted again using a factor of two different temperatures (15C and 20C), import the *Bread.csv* data into R.

- Create a plot of the results to see if an interaction may be present.

Two-Factor Designs in R I

- *There are many ways to create this kind of design in R.*

- We can use the expand.grid() function to create a data frame of all possible combinations:
  - D <- expand.grid(Factor.1 = as.factor(c(level.1,...,level.i)), Factor.2 = as.factor(c(level.1,...,level.j)))

- To replicate you can use the rbind() function:
  - D <- rbind(D,D)

Two-Factor Designs in R II

- Randomization:
  - `set.seed(2030) # Reproducible`
  - `D <-D[order(sample(1:nrow(D))), ] # Randomization`
  - `write.csv(D, file="Design.csv") # save optional`

Example 2

- Assume that we examine four levels for each of the two factors from the helicopter experiment:
  - Wing Length: 4, 4.75, 5.5, 6 (inches)
  - Body Width: 3.25, 3.75, 4, 4.25 (inches)

- Use R to come up with a randomized two-factor design with two replications for each combination.

Analysis

Mathematical Model (Two Factors)

- The mathematical model for a completely randomized two-factor factorial design can be written as:

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

- i represents the level of the first factor.
- j represents the level of the second factor.
- k represents the replicate number.

- This model is called a *cell means model* and μ_{ij} represents the expected response in the ij^{th} cell.

Alternative Mathematical Model

- The effects model:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk}. \tag{2}$$

- α_i and β_j are the main effects:
    - α_i represents the difference between the marginal average of all experiments at the ith level of the first factor and the overall average.
    - β_j represents the difference between the marginal average at the jth level of the second factor and the overall average.
- The interaction effects $\alpha\beta_{ij}$ represents the difference in the cell mean μ_{ij} and $\mu + \alpha_i + \beta_j$.

Effects Model Definitions and Assumptions

- Definitions:
  - $\sum_i \alpha_i = 0$
  - $\sum_j \beta_j = 0$
  - $\sum_i \alpha\beta_{ij} = 0$
  - $\sum_j \alpha\beta_{ij} = 0$

- Assumptions:
  - Experimental errors are independent (randomization)
  - $\epsilon_{ijk} \sim N(0, \sigma^2)$

Examining Results in R

- To examine the cell means in R:
  lm1 <- lm(response $\sim$ Factor1:Factor2, data = data)

- To examine the effects model in R:
  lm2 <- lm(response $\sim$ Factor1 + Factor2 +
  Factor1:Factor2, data = data)

Example 3 Preliminaries

- In R, import the *COdata* dataset into your environment.

- This data comes from an experiment that examined the amount of CO emissions released from burning fuel.

- The response variable (y) is the CO emissions concentrated in grams/metre^3.

- There are two factors:
  - Eth: ethanol additions with three levels (0.1, 0.2, 0.3)
  - Ratio: air/fuel ratio with three levels (14, 15, 16)

- There were ($r = 2$) replicate runs for each combination of factor levels.

Example 3

- Use R to examine the cell means from the CO experiment.

- Use R to examine the effects model from the CO experiment.

Analysis of Variance Table (ANOVA)

- When there are a levels for factor A, b levels for factor B, and r replications for each group, the ANOVA table for a two-factor design:

Source	df	Sum of Squares (ss)	Mean Squares	F-ratio
A	$a-1$	$R(\alpha|\mu)$	$\frac{ssA}{(a-1)}$	$F = \frac{msA}{msE}$
B	$b-1$	$R(\beta|\alpha,\mu)$	$\frac{ssB}{(b-1)}$	$F = \frac{msB}{msE}$
AB	$(a-1)(b-1)$	$R(\alpha\beta|\beta,\alpha,\mu)$	$\frac{ssAB}{(a-1)(b-1)}$	$F = \frac{msAB}{msE}$
Error	$ab(r-1)$	ssE	$\frac{ssE}{ab(r-1)}$	

- Where:
  - $R(\alpha|\mu)$: sum of squares for a model only including A
  - $R(\beta|\alpha,\mu)$: ssE_A - ssE_B where ssE_B is from a reduced model.
  - $R(\alpha\beta|\beta,\alpha,\mu)$: ssE_B - ssE

Two Factor Design Analysis R

- *Is there a significant interaction of the explanatory variables?*

- model <- aov(response $\sim$ Factor1*Factor2, data = data)
- summary(model)

Example 4

- Use R to construct an ANOVA table for the results of the CO experiment.

- Are there significant differences in the levels of the factors and their interaction?

`model.tables()`

- The model.tables() function breaks down the results from the ANOVA.

- Usage: model.tables(model, type = "means", se = T)

Example 5

- Use the model.tables() function to examine *tables of means* from the CO experiment.

- Are any of these results similar to what was found by estimating regression models?

Main Effects Contrasts I

- To estimate contrasts of the main effects, the estimable() function from the *gmodels* package can be used:

  - c1 <- c(-1/2, 0, 1/2) # first vs third
  - c1 <- c(1/2, -1, 1/2) # orthogonal vector

  - mod2 <- aov(CO $\sim$ Eth * Ratio, contrasts = list(Eth = cm, Ratio = cm), data = COdata) # restricted model
  - c <- rbind("Ethanol 0.3 vs 0.1" = c(0,1,0,0,0,0,0,0,0), "Ratio 16 vs 14" = c(0,0,0,1,0,0,0,0,0)) # specific contrasts
  - estimable(mod2,c) # hypothesis tests

Main Effects Contrasts II

- We are testing the null hypothesis that there are no differences in the main effects at the specific factor levels.

- **The estimates are only meaningful if there are no significant interactions between the two factors.**

Example 6

- Adjust the cm and c matrices from the example code to compare the following fixed-effects contrasts:

  - Ethanol 0.3 vs 0.2
  - Ratio 15 vs 16

Plotting Interactions in R I

- We can use visualizations to examine possible interactions in R:
- From our example:

  - ```
    with(COdata, (interaction.plot(Eth, Ratio, CO, type =
        "b", pch = c(18,24,22), leg.bty = "o", main =
        "Interaction Plot of Ethanol and air/fuel ratio",
        xlab = "Ethanol",ylab = "CO emissions")))
    ```

Example 7

- Adjust the example code to visualize how the effects of the Ratio on the CO emissions are modulated by the Ethanol levels.

Number of Replicates

- Replicates are needed to estimate the variance of the error in order to use the ANOVA *F*-test.

- The power of the test will increase as the number of replicates increases (could be expensive).

- Determining the number of replicates needed to achieve a desired power of a test falls into the category of *error control*.

- **The experimenter must select a value of practical difference (Δ).**
    - Usually obtained through the previous/pilot studies.

Number of Replicates in R (cell means)

- We can use the Fpower1() function from the *daewr* package:

  - library(daewr)
  - rmin <- s # smallest number of replicates considered
  - rmax <- m # largest number of replicates considered
  - alpha <- 0.05
  - sigma <- sqrt(sigma_squared) # from pilot study
  - nlev <- levels1*levels2 # **total number of factor combinations**
  - nreps <- c(rmin:rmax)
  - Delta <- delta # practical difference
  - power <- Fpower1(alpha,nlev,nreps,Delta,sigma)
  - power # results

- Rule of thumb: *The number of replicates that result in a power between 0.80 and 0.90 is usually sufficient for most designs.*

Number of Replicates in R (marginal means)

- We can use the Fpower2() function from the *daewr* package:

  - ```
    library(daewr)
    ```
  - ```
    rmin <- s # smallest number of replicates considered
    ```
  - ```
    rmax <- m # largest number of replicates considered
    ```
  - ```
    alpha <- 0.05
    ```
  - ```
    sigma <- sqrt(sigma_squared) # from pilot study
    ```
  - ```
    nlev <- c(levels1,levels2) # numbers of factor levels
    ```
  - ```
    nreps <- c(rmin:rmax)
    ```
  - ```
    Delta <- delta # practical difference
    ```
  - ```
    power <- Fpower2(alpha,nlev,nreps,Delta,sigma)
    ```
  - ```
    power # results
    ```

- **Hidden replication increases the efficiency of the experiments.**

Example 8

- Use both methods (cell means, marginal means) to determine how many replicates are needed for the CO emissions experiment.

- We can make the following assumptions:
    - fewest: 2
    - most: 8
    - $\hat{\sigma}^2 = 32$
    - $\Delta = 10$

- Comment on your results.

Unequal Number of Replicates

Unequal Number of Replicates

- It is uncommon to plan a factorial experiment with an unequal number of replicates per cell.

- In practice, there may be problems with the data collection.

- **As long as the chance of missing observations is not related to the treatment factor levels** the data can still be analysed.

- We need to slightly modify the code we use to perform this analysis.

Modified Code in R

- Assume that there is now an unequal number of replicates in the data.

- We can use the contr.sum argument and the *car* package in R to obtain meaningful results:
    - library(car)
    - model <- lm(response $\sim$ Factor1*Factor2, data = data, contrasts = list(Factor1 = contr.sum, Factor2 = contr.sum))
    - Anova(model,type = "III")

- *Sum contrasts are similar to mean-centering a particular continuous explanatory variable.*

- type = "III" indicates a Partial Sum of Squares (used when observations are missing)

Main Effects Contrasts Unequal Number of Replicates

- We can use estimate the marginal adjusted means for the main effects.

- We can use the *lsmeans* package in R:
  - library(lsmeans)
  - lsmeans(model, "Factor")

- **Again, the estimates are only meaningful if there are no significant interactions between the two factors.**

Example 9 from slides

- Assume that we are missing one observation from the CO emissions experiment (see code).

- In R conduct the following analyses:
    1. Construct an ANOVA table for the data with an unequal number of replicates.
        - Does the missing observation impact our overall results?

    2. Use the lsmeans() to examine the marginal adjusted means for both of the main effects factors.
        - Are these results reliable?

Thoughts on Unequal Number of Replicates

- The aov() function in R **will not** test the hypotheses in the same way as it does with equal replicates.

- When there is an equal number of replications per cell, the sums of squares computed by the aov() function are identical to the type III sums of squares.

- **In general, you can use the methods applied in Example 9 for both cases** (equal and unequal replicates).

One Replicate per Cell

One Replicate per Cell I

- When there is enough power, experiments may be ran with one replicate ($r = 1$) per cell.

- With only one replicate per cell, we cannot calculate the ANOVA *ssE*
  - Cannot conduct *F*-tests on the main effects and interactions in the traditional way.

- If there is no interaction, we can use an additive model:
  $y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$
  - **Dangerous if an interaction actually exists!**

One Replicate per Cell II

- If both factors are numeric and have equally spaced levels then the sum of squares for the interaction can be partitioned.
  - We can estimate an error term and test the significance of the linear part of the interaction.

- We will order the factors and then examine any significance in a linear interaction.

One Replicate per Cell Method I

- In R we can order our data, assume a linear interaction, and then test for the significance of the interaction:
    - Factor1.Order <- as.ordered(data$Factor1)
    - Factor2.Order <- as.ordered(data$Factor2)

    - Factor1.Lin <-
      contr.poly(Factor1.Order)[Factor1.Order,".L"]
    - Factor2.Lin <-
      contr.poly(Factor2.Order)[Factor2.Order,".L"]
    - model <- lm(Response $\sim$ Factor1.Order + Factor2.Order + Factor1.Lin:Factor2.Lin, data = data)
    - anova(model)

One Replicate per Cell Method II

- In R we can directly test for significance of the interaction using the Tukey1df() function:
    - Tukey1df(data)

- Note: The response variable needs to be in the first column of the data frame.
    - data <- data %>%
        relocate(response)

- **Much easier to code but slightly different to read than we are used to.**
    - If the NonAdditivity term is significant then we may assume that there is an interaction.

Example 10

- Assuming that we only received one observation from the CO experiment.

- Use Method I and Method II to determine if there is a significant interaction between the Ethonol and the Ratio variables.

- *Note: The data has been adjusted to assume that the group means are our only observation.*

Example 11

- Import the *virus* dataset into R.

- Take some time to get to know the data.

- Use Method I and Method II to determine if there is a significant interaction between the `Sample` and the `Dilution` variables.

- How would you obtain the coefficient estimates for an additive model if the interactions are not significant?

Exercise 1

- Using the *Bread.csv* data conduct the following analysis:
  1. Use visualizations and an appropriate statistical method to determine if an interaction between the two factors exists.

  2. Examine the main effects using whichever method you prefer.
     - Are these estimates meaningful?

  3. How many replicates are needed for this experiment assuming the following information:
     - $\hat{\sigma}^2 = 2.1$ (from pilot)
     - $\Delta = 3$

Exercise 2

- Remove one observation from the *Bread.csv* data and repeat (1) and (2) from Exercise 1.

- Rework the code from Example 10 to generate a data frame of the cell means and use the appropriate methods to repeat (1) and (2) from Exercise 1.
    - Are the results much different?

References & Resources

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

- model.tables
- gmodels
- Fpower2
- estimable
- Anova
- lsmeans