STAT 423/523 Statistical Methods for Engineers and Scientists

Lecture 8: Multifactor Analysis of Variance I

Chencheng Cai

Washington State University

Multifactor ANOVA

- In the previous lectures, we have discussed one-way ANOVA, which is used to compare the means of two or more groups.
- In this lecture, we will discuss multifactor ANOVA, which is used to compare the means of two or more groups when there are two or more factors.

Multifactor ANOVA

- In the previous lectures, we have discussed one-way ANOVA, which is used to compare the means of two or more groups.
- In this lecture, we will discuss multifactor ANOVA, which is used to compare the means of two or more groups when there are two or more factors.

Topics to be covered:

- Two-factor ANOVA without replication
- Two-factor ANOVA with replication

Multifactor ANOVA

- In the previous lectures, we have discussed one-way ANOVA, which is used to compare the means of two or more groups.
- In this lecture, we will discuss multifactor ANOVA, which is used to compare the means of two or more groups when there are two or more factors.

Topics to be covered:

- Two-factor ANOVA without replication
- Two-factor ANOVA with replication

We will focus on:

- Sum of squares and mean squares
- F-test and multiple comparisons
- Fixed and random effects

Two-Factor ANOVA without Replication

Suppose we have two factors:

- Factor A: I levels, $i = 1, 2, \ldots, I$
- Factor B: J levels, $j = 1, 2, \ldots, J$

The possible number of treatments is $I \times J$.
Remark: different index notations from one-way ANOVA.

Two-Factor ANOVA without Replication

Suppose we have two factors:

- Factor A: I levels, $i = 1, 2, \ldots, I$
- Factor B: J levels, $j = 1, 2, \ldots, J$

The possible number of treatments is $I \times J$.
Remark: different index notations from one-way ANOVA.

Furthermore, we assume we have **one** observation for each treatment:

- X_{ij} is the observation for the i-th level of factor A and the j-th level of factor B.
- x_{ij} the observed value of X_{ij}.

Two-Factor ANOVA without Replication

Suppose we have two factors:

- Factor A: I levels, $i = 1, 2, \ldots, I$
- Factor B: J levels, $j = 1, 2, \ldots, J$

The possible number of treatments is $I \times J$.
Remark: different index notations from one-way ANOVA.

Furthermore, we assume we have **one** observation for each treatment:

- X_{ij} is the observation for the i-th level of factor A and the j-th level of factor B.
- x_{ij} the observed value of X_{ij}.

The analysis of variance for such a model is called a **two-factor ANOVA without replication**.

Example

A research on the erabsiablity of stains on a fabric from three brands of pen and four different washing treatments. The observation is quantitaive measurement of color change.

		Washing Treatment					
		1	**2**	**3**	**4**	**Total**	**Average**
Brand of Pen	**1**	.97	.48	.48	.46	2.39	.598
	2	.77	.14	.22	.25	1.38	.345
	3	.67	.39	.57	.19	1.82	.455
	Total	2.41	1.01	1.27	.90	5.59	
	Average	.803	.337	.423	.300		.466

The Means

Similar to one-way ANOVA, we can define the following means:

- The sample mean of the i-th level of factor A:

$$\bar{X}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} X_{ij}$$

- The sample mean of the j-th level of factor B:

$$\bar{X}_{\cdot j} = \frac{1}{I}\sum_{i=1}^{I} X_{ij}$$

- The grand sample mean:

$$\bar{X}_{\cdot\cdot} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}$$

Fixed Effect Model

In a fixed effect model, we assume

$$X_{ij} = \mu_{ij} + \varepsilon_{ij} \quad \text{for } i = 1, 2, \ldots, I, j = 1, 2, \ldots, J$$

with $\varepsilon_{ij} \sim N(0, \sigma^2)$.

Fixed Effect Model

In a fixed effect model, we assume

$$X_{ij} = \mu_{ij} + \varepsilon_{ij} \quad \text{for } i = 1, 2, \ldots, I, j = 1, 2, \ldots, J$$

with $\varepsilon_{ij} \sim N(0, \sigma^2)$.

- Total number of observations: IJ
- Total number of parameters: $IJ + 1$
- The model is not estimable until we impose extra constraints.

Fixed Effect Model

We consider the following **additive model**:

$$X_{ij} = \alpha_i + \beta_j + \varepsilon_{ij},$$

that is, we assume $\mu_{ij} = \alpha_i + \beta_j$.

Fixed Effect Model

We consider the following **additive model**:

$$X_{ij} = \alpha_i + \beta_j + \varepsilon_{ij},$$

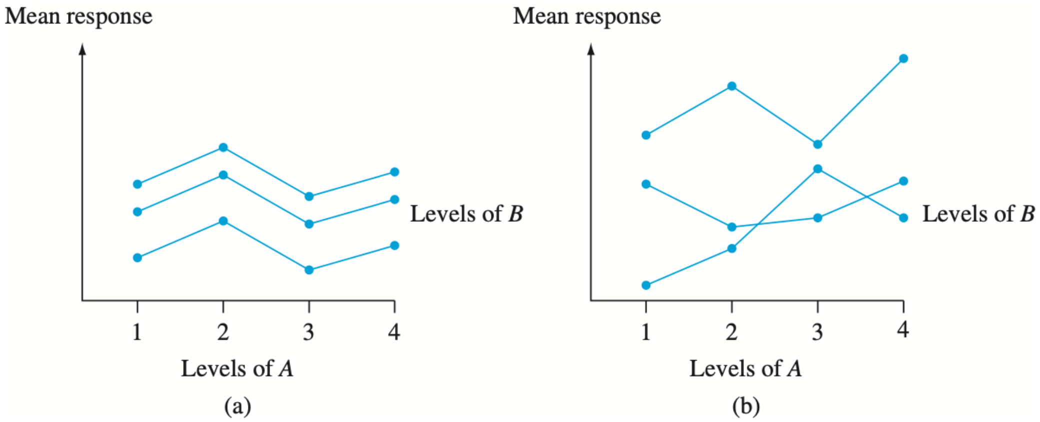
that is, we assume $\mu_{ij} = \alpha_i + \beta_j$.

The difference in responeses between any two treatment levels can be decomposed into the sum of the differences of the corresponding factor levels:

$$\mu_{ij} - \mu_{i'j'} = (\alpha_i - \alpha_{i'}) + (\beta_j - \beta_{j'})$$

Fixed Effect Model

The additivity assumption can be checked visually.



Fixed Effect Model

$$X_{ij} = \alpha_i + \beta_j + \varepsilon_{ij},$$

However, this model still has the **identifiability problem**:
The following transformation of the parameters does not change the model:

$$\alpha_i \to \alpha_i + c, \quad \beta_j \to \beta_j - c$$

for any constant c.
Additional constraints are needed to make the parameters unique.

Fixed Effect Model

We consider the following model:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

with $\sum_{i=1}^{I} \alpha_i = 0$ and $\sum_{j=1}^{J} \beta_j = 0$.

Fixed Effect Model

We consider the following model:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

with $\sum_{i=1}^{I} \alpha_i = 0$ and $\sum_{j=1}^{J} \beta_j = 0$.

The model is now **identifiable**.

Fixed Effect Model

We consider the following model:

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

with $\sum_{i=1}^{I} \alpha_i = 0$ and $\sum_{j=1}^{J} \beta_j = 0$.

The model is now **identifiable**.

Proof: Let $(\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J)$ and $(\mu', \alpha'_1, \ldots, \alpha'_I, \beta'_1, \ldots, \beta'_J)$ be two sets of parameters that give the same model. Then we have

$$\mu + \alpha_i + \beta_j = \mu' + \alpha'_i + \beta'_j \quad \text{for all } i, j.$$

Take the sum over i and j, we have $\mu = \mu'$. Take the sum over j, we have $\alpha_i = \alpha'_i$ for all i. Take the sum over i, we have $\beta_j = \beta'_j$ for all j.

Fixed Effect Model

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

The interpretation of the parameters:

- μ: the grand mean
- α_i: the effect of the i-th level of factor A
- β_j: the effect of the j-th level of factor B

Fixed Effect Model

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

The interpretation of the parameters:

- μ: the grand mean
- α_i: the effect of the i-th level of factor A
- β_j: the effect of the j-th level of factor B

Because:

- $\mu = E[\bar{X}_{..}]$
- $\alpha_i = E[\bar{X}_{i.}] - E[\bar{X}_{..}]$
- $\beta_j = E[\bar{X}_{.j}] - E[\bar{X}_{..}]$

Fixed Effect Model

$$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$$

The parameters can be estimated unbiasedly by:

$$\hat{\mu} = \bar{X}_{..} = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}$$

$$\hat{\alpha}_i = \bar{X}_{i.} - \bar{X}_{..} = \frac{1}{J}\sum_{j=1}^{J} X_{ij} - \bar{X}_{..}$$

$$\hat{\beta}_j = \bar{X}_{.j} - \bar{X}_{..} = \frac{1}{I}\sum_{i=1}^{I} X_{ij} - \bar{X}_{..}$$

Verify that $\sum_{i=1}^{I} \hat{\alpha}_i = 0$ and $\sum_{j=1}^{J} \hat{\beta}_j = 0$.

Sum of Sqaures

We can define the following sum of squares for the two-factor ANOVA:

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{..})^2 \qquad df: IJ - 1$$

$$SSA = J\sum_{i=1}^{I}(\bar{X}_{i.} - \bar{X}_{..})^2 \qquad df: I - 1$$

$$SSB = I\sum_{j=1}^{J}(\bar{X}_{.j} - \bar{X}_{..})^2 \qquad df: J - 1$$

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2 \qquad df: (I-1)(J-1)$$

Verify that

$$SST = SSA + SSB + SSE$$

Mean Sqaures

We can define the following mean squares for the two-factor ANOVA:

$$MSA = \frac{SSA}{I-1}$$

$$MSB = \frac{SSB}{J-1}$$

$$MSE = \frac{SSE}{(I-1)(J-1)}$$

Mean Sqaures

We can define the following mean squares for the two-factor ANOVA:

$$MSA = \frac{SSA}{I-1}$$
$$MSB = \frac{SSB}{J-1}$$
$$MSE = \frac{SSE}{(I-1)(J-1)}$$

The expected mean squares are:

$$E[MSA] = \sigma^2 + J \sum_{i=1}^{I} \alpha_i^2$$
$$E[MSB] = \sigma^2 + I \sum_{j=1}^{J} \beta_j^2$$
$$E[MSE] = \sigma^2$$

Hypothesis Testing

To test the main effects of factor A, we consider the following hypotheses:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$
$$H_a : \text{At least one } \alpha_i \text{ is not zero}$$

Hypothesis Testing

To test the main effects of factor A, we consider the following hypotheses:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$
$$H_a : \text{At least one } \alpha_i \text{ is not zero}$$

Reject null when

$$F = \frac{MSA}{MSE} > F_{\alpha, I-1, (I-1)(J-1)}$$

Hypothesis Testing

To test the main effects of factor A, we consider the following hypotheses:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0$$
$$H_a : \text{At least one } \alpha_i \text{ is not zero}$$

Reject null when

$$F = \frac{MSA}{MSE} > F_{\alpha, I-1, (I-1)(J-1)}$$

Similarly, to test the main effects of factor B, we consider the following hypotheses:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_J = 0$$
$$H_a : \text{At least one } \beta_j \text{ is not zero}$$

Reject null when

$$F = \frac{MSB}{MSE} > F_{\alpha, J-1, (I-1)(J-1)}$$

From the Perspective of Nested Models

The testing of the main effect of factor A is equivalent to testing the following nested models:

Full model:	$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$
Reduced model:	$X_{ij} = \mu + \beta_j + \varepsilon_{ij}$

From the Perspective of Nested Models

The testing of the main effect of factor A is equivalent to testing the following nested models:

Full model:	$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$
Reduced model:	$X_{ij} = \mu + \beta_j + \varepsilon_{ij}$

- The sum of squares error for the full model is SSE.
- The sum of squares error for the reduced model is $SSE + SSA$.

From the Perspective of Nested Models

The testing of the main effect of factor A is equivalent to testing the following nested models:

Full model:	$X_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$
Reduced model:	$X_{ij} = \mu + \beta_j + \varepsilon_{ij}$

- The sum of squares error for the full model is SSE.
- The sum of squares error for the reduced model is $SSE + SSA$.
- The F-test statistic is

$$F = \frac{(SSE + SSA - SSE)/(I-1)}{SSE/[(I-1)(J-1)]} = \frac{MSA}{MSE}$$

The ANOVA Table

The ANOVA table for the washing example is:

Source of Variation	df	Sum of Squares	Mean Square	f
Factor A (brand)	$I - 1 = 2$	SSA = .1282	MSA = .0641	$f_A = 4.43$
Factor B (wash treatment)	$J - 1 = 3$	SSB = .4797	MSB = .1599	$f_B = 11.05$
Error	$(I - 1)(J - 1) = 6$	SSE = .0868	MSE = .01447	
Total	$IJ - 1 = 11$	SST = .6947		

Tukey's Method for Multiple Comparison

The procedure is same as one-way ANOVA except that we have different thresholds for the different means:

$$w_A = Q_{\alpha,I,(I-1)(J-1)}\sqrt{MSE/J}, \quad w_B = Q_{\alpha,J,(I-1)(J-1)}\sqrt{MSE/I}$$

Tukey's Method for Multiple Comparison

The procedure is same as one-way ANOVA except that we have different thresholds for the different means:

$$w_A = Q_{\alpha,I,(I-1)(J-1)}\sqrt{MSE/J}, \quad w_B = Q_{\alpha,J,(I-1)(J-1)}\sqrt{MSE/I}$$

- We use w_A to compare the means of factor A.
- We use w_B to compare the means of factor B.

Example

Recall the washing example.

		Washing Treatment					
		1	2	3	4	Total	Average
Brand of Pen	1	.97	.48	.48	.46	2.39	.598
	2	.77	.14	.22	.25	1.38	.345
	3	.67	.39	.57	.19	1.82	.455
	Total	2.41	1.01	1.27	.90	5.59	
	Average	.803	.337	.423	.300		.466

Example

Recall the washing example.

		Washing Treatment					
		1	2	3	4	Total	Average
Brand of Pen	1	.97	.48	.48	.46	2.39	.598
	2	.77	.14	.22	.25	1.38	.345
	3	.67	.39	.57	.19	1.82	.455
	Total	2.41	1.01	1.27	.90	5.59	
	Average	.803	.337	.423	.300		.466

If we want to compare the different washing treaments, the threshold is

$$w = Q_{\alpha,4,6}\sqrt{MSE/3} = 0.34.$$

Example

Recall the washing example.

		Washing Treatment 1	2	3	4	Total	Average
Brand of Pen	1	.97	.48	.48	.46	2.39	.598
	2	.77	.14	.22	.25	1.38	.345
	3	.67	.39	.57	.19	1.82	.455
	Total	2.41	1.01	1.27	.90	5.59	
	Average	.803	.337	.423	.300		.466

If we want to compare the different washing treaments, the threshold is

$$w = Q_{\alpha,4,6}\sqrt{MSE/3} = 0.34.$$

Therefore, the means of washing treatments 2, 3, 4 are close to each other.

Completely Randomized Design and Randomized Block Design

If we would like to compare the means of different levels of factor A, we can consider the following **completely randomized design**:

1. Sample IJ units randomly from the population.

Completely Randomized Design and Randomized Block Design

If we would like to compare the means of different levels of factor A, we can consider the following **completely randomized design**:

1. Sample IJ units randomly from the population.
2. Randomly choose J units from the IJ units for the first level of factor A.

Completely Randomized Design and Randomized Block Design

If we would like to compare the means of different levels of factor A, we can consider the following **completely randomized design**:

1. Sample IJ units randomly from the population.
2. Randomly choose J units from the IJ units for the first level of factor A.
3. Randomly choose J units from the remaining $IJ - J$ units for the second level of factor A.
4. ...
5. Randomly choose J units from the remaining $2J$ units for the $(I - 1)$-th level of factor A.
6. The remaining J units are for the I-th level of factor A.

Completely Randomized Design and Randomized Block Design

However, there might be other covariates Z that affect the response variable. In this case, we can consider the following **randomized block design** as a generalization as a paired experiments:

Completely Randomized Design and Randomized Block Design

However, there might be other covariates Z that affect the response variable. In this case, we can consider the following **randomized block design** as a generalization as a paired experiments:

1. Divide the population into I blocks based on their Z values. Call it factor B or **blocks**.

Completely Randomized Design and Randomized Block Design

However, there might be other covariates Z that affect the response variable. In this case, we can consider the following **randomized block design** as a generalization as a paired experiments:

1. Divide the population into I blocks based on their Z values. Call it factor B or **blocks**.
2. Sample I units from the population in the first block.
3. Randomly assign the I units to the I levels of factor A.

Completely Randomized Design and Randomized Block Design

However, there might be other covariates Z that affect the response variable. In this case, we can consider the following **randomized block design** as a generalization as a paired experiments:

1. Divide the population into I blocks based on their Z values. Call it factor B or **blocks**.
2. Sample I units from the population in the first block.
3. Randomly assign the I units to the I levels of factor A.
4. Sample I units from the population in the second block.
5. Randomly assign the I units to the I levels of factor A.
6. ...
7. Sample I units from the population in the J-th block.
8. Randomly assign the I units to the I levels of factor A.

Completely Randomized Design and Randomized Block Design

However, there might be other covariates Z that affect the response variable. In this case, we can consider the following **randomized block design** as a generalization as a paired experiments:

1. Divide the population into I blocks based on their Z values. Call it factor B or **blocks**.
2. Sample I units from the population in the first block.
3. Randomly assign the I units to the I levels of factor A.
4. Sample I units from the population in the second block.
5. Randomly assign the I units to the I levels of factor A.
6. ...
7. Sample I units from the population in the J-th block.
8. Randomly assign the I units to the I levels of factor A.

The other way that random samples from each levels of factor A is randomly assigned to the blocks is also possible.

Example

An organization would like to study the annual power consumption of five brands of dehumidifiers.

Example

An organization would like to study the annual power consumption of five brands of dehumidifiers.

- The brand is factor A with five levels.
- The completely randomized design with $J = 4$ is
  - Randomly sample $J = 4$ dehimifiers from each brand and test the power consumption.

Example

An organization would like to study the annual power consumption of five brands of dehumidifiers.

- The brand is factor A with five levels.
- The completely randomized design with $J = 4$ is
  - Randomly sample $J = 4$ dehimifiers from each brand and test the power consumption.

However, the humidity level might affect the power consumption. Therefore, we can consider $J = 4$ different humidity levels as blocks.

Example

An organization would like to study the annual power consumption of five brands of dehumidifiers.

- The brand is factor A with five levels.
- The completely randomized design with $J = 4$ is
  - Randomly sample $J = 4$ dehimifiers from each brand and test the power consumption.

However, the humidity level might affect the power consumption. Therefore, we can consider $J = 4$ different humidity levels as blocks.

- The humidity level is factor B with four levels.
- The randomized block design is
  - Sample $J = 4$ dehumidifiers from each brand.
  - Randomly assign the $J = 4$ dehumidifiers to the $J = 4$ humidity levels and test the power consumption.

Example

Treatments (brands)	Blocks (humidity level) 1	2	3	4	$x_{i\cdot}$	$\bar{x}_{i\cdot}$
1	685	792	838	875	3190	797.50
2	722	806	893	953	3374	843.50
3	733	802	880	941	3356	839.00
4	811	888	952	1005	3656	914.00
5	828	920	978	1023	3749	937.25
$x_{\cdot j}$	3779	4208	4541	4797	17,325	
$\bar{x}_{\cdot j}$	755.80	841.60	908.20	959.40		866.25

Example

Treatments (brands)	Blocks (humidity level) 1	2	3	4	$x_{i\cdot}$	$\bar{x}_{i\cdot}$
1	685	792	838	875	3190	797.50
2	722	806	893	953	3374	843.50
3	733	802	880	941	3356	839.00
4	811	888	952	1005	3656	914.00
5	828	920	978	1023	3749	937.25
$x_{\cdot j}$	3779	4208	4541	4797	17,325	
$\bar{x}_{\cdot j}$	755.80	841.60	908.20	959.40		866.25

Source of Variation	df	Sum of Squares	Mean Square	f
Treatments (brands)	4	53,231.00	13,307.75	$f_A = 95.57$
Blocks	3	116,217.75	38,739.25	$f_B = 278.20$
Error	12	1671.00	139.25	
Total	19	171,119.75		

Randomized Block Design

The purpose of the randomized block design is to offset the effect of any **confounders**.

Randomized Block Design

The purpose of the randomized block design is to offset the effect of any **confounders**.

- The confounder is a variable that is correlated with the factor of interest and affects the response variable.
- Ignoring the confounder might lead to totally wrong conclusions.

Randomized Block Design

The purpose of the randomized block design is to offset the effect of any **confounders**.

- The confounder is a variable that is correlated with the factor of interest and affects the response variable.
- Ignoring the confounder might lead to totally wrong conclusions.

In the previous example, the power consumption is likely an increasing function of the humidity level and the brand number.

Randomized Block Design

The purpose of the randomized block design is to offset the effect of any **confounders**.

- The confounder is a variable that is correlated with the factor of interest and affects the response variable.
- Ignoring the confounder might lead to totally wrong conclusions.

In the previous example, the power consumption is likely an increasing function of the humidity level and the brand number.

However, if in high humidity levels, people tend to use dehumidifiers with a lower brand number to save energy and vice versa, the power consumption might be a decreasing function of the brand number.

Randomized Block Design

The purpose of the randomized block design is to offset the effect of any **confounders**.

- The confounder is a variable that is correlated with the factor of interest and affects the response variable.
- Ignoring the confounder might lead to totally wrong conclusions.

In the previous example, the power consumption is likely an increasing function of the humidity level and the brand number.

However, if in high humidity levels, people tend to use dehumidifiers with a lower brand number to save energy and vice versa, the power consumption might be a decreasing function of the brand number.

See also: **Simpson's paradox.**