Multi-way ANOVA

CONCEPTS FROM EXPERIMENTAL DESIGN

NOTATION FOR TWO-WAY ANOVA

PROFILE PLOTS

Motivation

Most statistical methods can be viewed as special cases of more general methods

Multi-way ANOVA is a prime example, where it can be phrased as a multiple regression with categorical explanatory variables

(a categorical explanatory variable is sometimes called a **FACTOR**)

(Can you think of any other examples of special cases of more general methods?)

Reasons for considering the special case:

- Assumptions are easier to check/more relevant
- Relationship from science to statistics is easier to see/explain
- Usually more straightforward to get more nuanced information

(In particular: decomposition into different sums of squares)

Some Important Ideas

Know what the “experimental design” is (or isn’t)

(here, I’m writing “experimental design” to indicate that it isn’t necessarily an experiment nor designed e.g. observational study)

The experimental design drives the analysis (e.g. what kind of model to use)

(Not the other way around…)

Think about the nature of the experiment and what is truly of interest

(why are you looking at the data at all?)

How was the data generated?

Some Concepts from Experimental Design

A Necessary Discussion

Some experimental design concepts are crucial for understanding multi-way ANOVA

Some important concepts along with heuristic definitions

Term	Heuristic Definition
Experimental Unit	"Object that gets assigned the treatment"
Replication	"Repeating the treatment"
Block	"Remove sources of variation or confounding"
Split-plot	"Due to practical constraints, different factors can have different experimental units"
Crossed vs. Nested design	"Crossed is all combinations, nested is more limited"

Let's define and discuss these..

(The agricultural/experiment history of statistics has influenced the terminology)

Experimental Unit

Heuristic Definition: "Object that gets assigned the treatment"

Formal Definition: For a factor X, the EXPERIMENTAL UNIT can be assigned to different levels of X

(Hence, experimental units are with respect to a specific factor X)

The most common scenario in practice is when X is a treatment

Example: Define X be the treatment: drug or placebo, which gets assigned to subjects randomly. Then the EXPERIMENTAL UNIT is the subject (with respect to this X)

Crucial: different experimental units must be capable of being assigned different treatments

When the object actually being measured is not the experimental unit, we sometimes call it the SAMPLING UNIT

Replication

Heuristic Definition: "Repeating the treatment"

Formal Definition: **REPLICATION** is when the full set of levels for a factor X occur multiple times to the experimental units (with respect to that factor)

Ultimately, the experimental unit is what is replicated to increase the degrees of freedom

Warning: e A common mistake is adding replication to something that isn't the experimental unit. This doesn't contribute to the analysis (and hence wastes resources). It is sometimes referred to as **PSEUDO-REPLICATION**

Experimental Unit: Example

Example: Examine the relationship between pollution and fish lesions

Set up 2 aquariums, each with 10 fish

- Randomly assign a water treatment (polluted vs. control) to the aquariums
- Wait 30 days. Catch 5 fish from each aquarium and count the number of lesions
- What is the experimental unit?
- Is there replication?

(Important: increasing the number of fish is not increasing replication! We will see that more fish really has no effect)

Experimental Unit: Example

Example: Examine the relationship between pollution and fish lesions

Set up 10 aquariums, each with 1 fish

- Randomly assign a water treatment (polluted vs. control) to the aquariums
- Wait 30 days. Catch the fish from each aquarium and count the number of lesions
- What is the experimental unit (with respect to the treatment)?
- Is there replication?

Experimental Unit: Example

Example: Examine the relationship between pollution and fish lesions

Set up 10 aquariums, each with 10 fish

- Randomly assign a water treatment (polluted vs. control) to the aquariums
- Wait 30 days. Catch 2 fish from each aquarium and count the number of lesions
- What is the experimental unit?
- Is there replication?

(A two-factor design (aquarium and treatment), we will call it a 2-way ANOVA)

Block

Heuristic Definition: “Remove sources of variation or confounding”

Sometimes experimental units come in “groups” that are similar, but this grouping structure isn’t of direct interest.

Formal Definition: The inclusion of this “group” structure via an explanatory variable is referred to as a **BLOCK**

Example: We have a field of plots for experimenting between two species of corn. The field is broken up into 8 plots, but due to a slight slope in the field, 4 of the plots are “up slope” and 4 of the plots are “down slope”. This up vs. down slope factor would be a blocking explanatory variable

Example: We are testing for two ads on a website. Sometimes people view the website with Chrome, Safari, Firefox, Internet Explorer, or other. This browser variable would be a blocking explanatory variable

Split-plot

Heuristic Definition: "Due to practical constraints, different factors can have different experimental units"

Suppose there are two or more factors

Formal Definition: A **SPLIT-PLOT** design is when the experimental unit is different for different explanatory variables

Example: We have a field of 8 plots allocated for testing two different species of corn. Irrigation timing has an affect that we would like to control for. We have to irrigate groups of 4 plots at the same time and assign treatments to an entire plot

So we have an $x_1 =$ irrigation, $x_2 =$ treatment

What are the experiment units?

- irrigation: group of 4 plots (known as the **WHOLE PLOT**)
- treatment: individual plot (known as the **SPLIT PLOT**)

Back to Fish Example

Example: Examine the relationship between polluted and fish lesions

- Set up 10 aquariums, each with 5 fish
- Wait 30 days and count the number of lesions per fish

Set up an experimental design given there is a single heater on one side of the room and answer:

- What are the explanatory variables?
- What are the experimental units?
- What are the blocking variables?
- Is this a split-plot design? If so, what is the whole plot and split plot?

Crossed vs. Nested Design

Heuristic Definition: “Crossed is all combinations, nested is more limited”

Are explanatory variables **CROSSED** or **NESTED**?

Formal Definition: If the levels of one of the factors occur at all of the levels of the other factors, then they are **CROSSED** (sometimes referred to as **FACTORIAL** design).

If factor B is **NESTED** in a factor A if every level of B occurs within exactly one level of A

Crossed vs. Nested Design

"If factor B is **NESTED** in a factor A if each level of B occurs within only one level of A"

Example: Two levels of a treatment correspond to a hormone therapy involving estrogen.

Due to differences in genders, standards of care demand different amounts of estrogen are administered to females than males.

In this case the treatment is nested in gender.

Example: I have 4 fertilizers and 2 plant species. I apply all 4 fertilizers to both plant species

In this case the these factors are crossed.

Example: I have 4 fertilizers and 2 plant species. I apply 2 fertilizers to one species and I apply the other 2 fertilizers to the other species

In this case the fertilizer is nested in species

Back to ANOVA

One-way ANOVA notation

- $Y_{ij} = \mu_j + \varepsilon_{ij}$

or

- $Y_{ij} = \mu + \mu_j + \varepsilon_{ij}$

(Often called the ANOVA model due to the overall mean μ and each group mean μ_j being included)

or

- $\mu\{Y|X = j\} = \mu_j$

Or even

$\mu\{Y|Factor\} = Factor$

Where,

- There are n_j observations in group j
- $n = \sum_{j=1}^{J} n_j$
- $\varepsilon_{ij} \sim N(0, \sigma^2)$ and are independent

Two-way ANOVA

No matter the number of levels (J), in one-way ANOVA there was only one categorical explanatory variable (X):

Let's generalize this to having two explanatory variables: A and B

The notation is a bit more cumbersome: for $j = 1, \dots, J$ k$= 1, \dots, K$

- The interaction (non additive) model is written as

$Y_{ijk} = \mu + \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk} \leftrightarrow \mu\{Y|$ A, B} = A + B + A*B

- The additive (non interaction) model is written as

$Y_{ijk} = \mu + \mu_j + \mu_k + \varepsilon_{ijk} \leftrightarrow \mu\{Y|$ A,B} = A + B

J and K are the number of levels of the first and second factor, respectively

(This is a special case of a multiple regression with two explanatory variables that take J and K values, respectively)

One-way ANOVA

Back to aquarium/pollution example:

Suppose we have

- 3 different treatments: no, low, and high pollution
- Each treatment is applied to 4 aquariums
- Each aquarium has 1 fish (so fish is the experiment unit)

(here, increasing fish directly affects the df of the relevant test)

Source	DF	SS	MS	F	Pr > F
Pollution	2	SS(pollution)	MS(pollution)	MS(pollution)/MS(error)	p-value
Error	$12 - 3$	SS(error)	MS(error)		
Total	$12 - 1$	SS(total)			

Crossed Designs

Crossed or Factorial Design: Interaction

Suppose we have two factors A and B
(Say, A is types of fertilizer (J = 2) and B is species of plant (K = 2))

- We repeat the experiment 5 times
- The response is the total dried weight of a plant
- We want to look at all combinations of fertilizer and plant, and allow for the fertilizer to have a different effect for different species J $\rightarrow$ interaction model

Crossed or Factorial Design: Interaction

Suppose we have two factors A and B
(Say, A is types of fertilizer (J = 2) and B is species of plant (K = 2))

The interaction model would be:

$$Y_{ijk} = \mu_j + \mu_k + \mu_{jk} + \varepsilon_{ijk} \leftrightarrow \mu\{Y|\text{ A, B}\} = \text{ A + B + A*B}$$

Source	DF	SS	MS	F	Pr > F
Fertilizer	J-1 = 1				
Species	K-1 = 1				
Fertilizer*Species	(J-1)*(K-1) = 1				
Error	n - J*K = 16				
Total	n - 1 = 19				

Crossed or Factorial Design: Additive

Suppose we have two factors A and B
(Say, A is types of fertilizer (J = 2) and B is species of plant (K = 2))

The additive model would be:

$$Y_{ijk} = \mu_j + \mu_k + \varepsilon_{ijk} \leftrightarrow \mu\{Y|\text{ A, B}\} = \text{ A + B}$$

Source	DF	SS	MS	F	Pr > F
Fertilizer	J-1 = 1				
Species	K-1 = 1				
Error	$n-(J-1+K-1)-1 = 17$				
Total	n - 1 = 19				

Profile Plots

Profile Plots

Some pictures (known as **PROFILE PLOTS**) will help us develop some intuition about interaction models

For two factors A and B, there will be two (piece-wise linear) lines describing the relationship between the factors

We can visually see both the additive and interaction effects

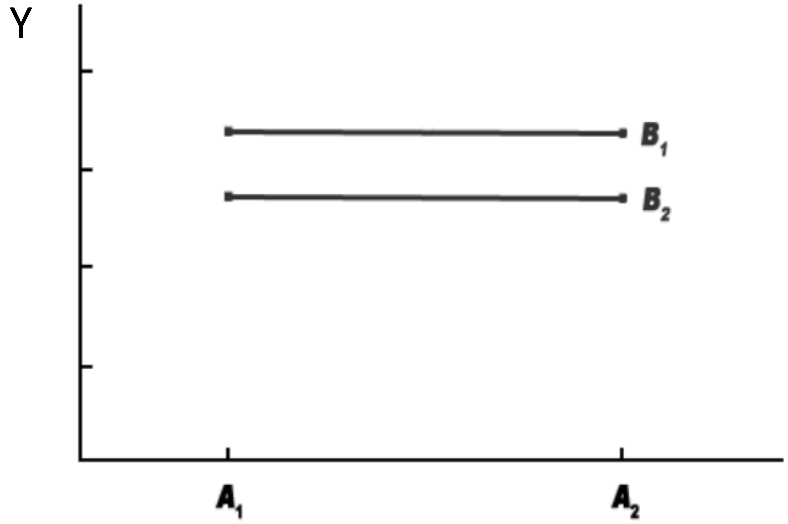
The ANOVA table helps us to formally test which situation is happening

Two-Level, Two-Way ANOVA

Note: To Check for A look at avg. Y for A1 vs. A2

To Check for B, rewrite plot with B on horizontal and look at avg. Y for B1 vs. B2

To Check for interaction, look for parallel vs. not parallel lines

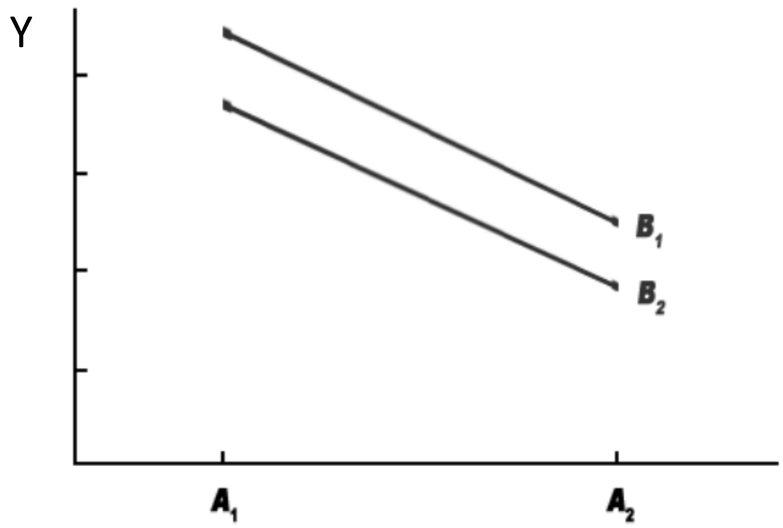


No effect of Factor A

Small effect of Factor B

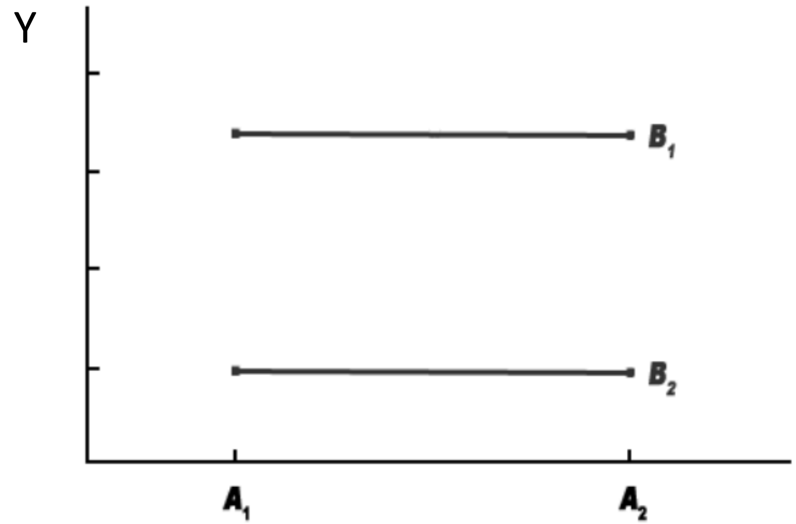
No interaction between Factor A and Factor B.

Two-Level, Two-Way ANOVA



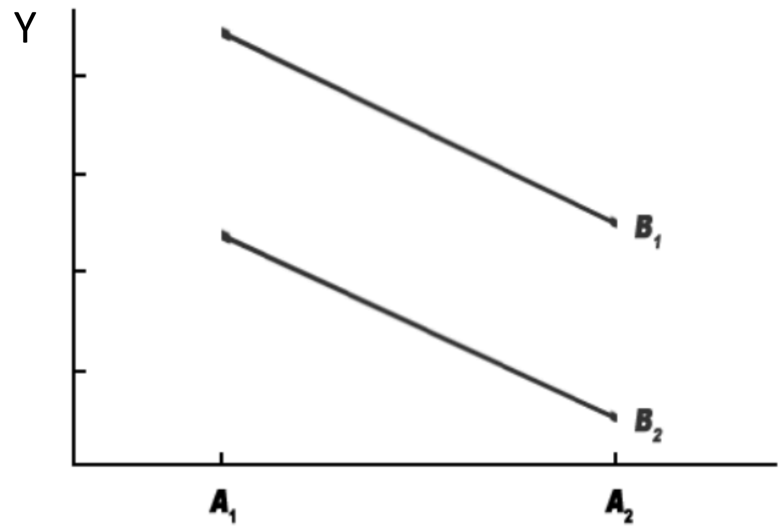
Large effect of Factor A
Small effect of Factor B
No interaction between Factor A and Factor B.

Two-Level, Two-Way ANOVA



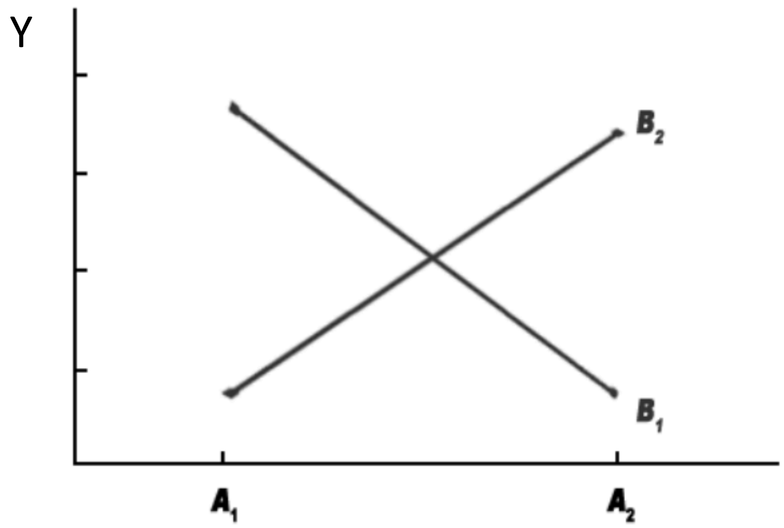
No effect of Factor A
Large effect of Factor B
No interaction between Factor A and Factor B.

Two-Level, Two-Way ANOVA



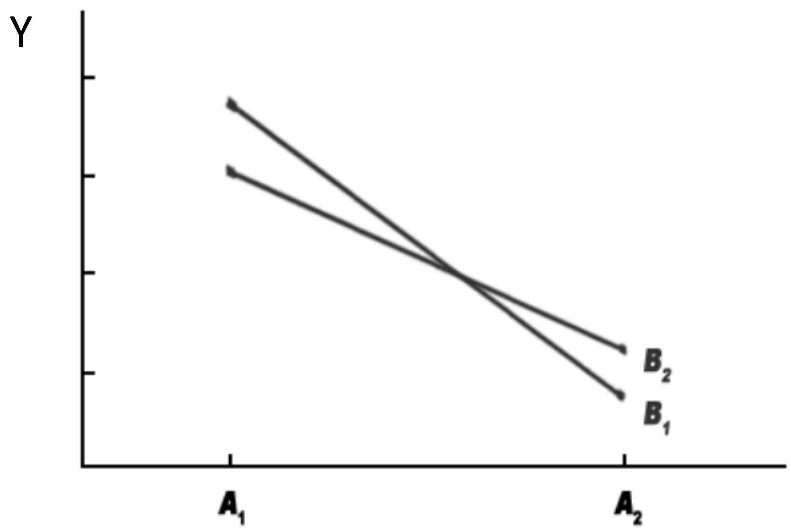
Large effect of Factor A
Large effect of Factor B
No interaction between Factor A and Factor B.

Two-Level, Two-Way ANOVA



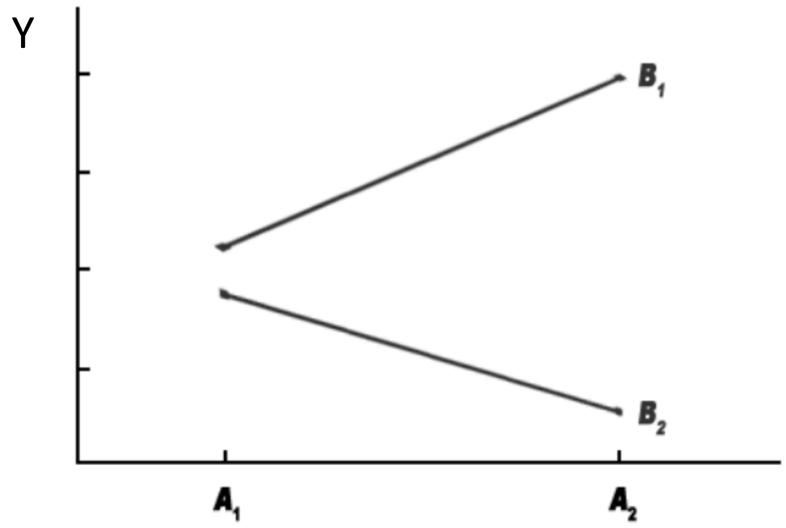
No effect of Factor A
No effect of Factor B
Large interaction between Factor A and Factor B.

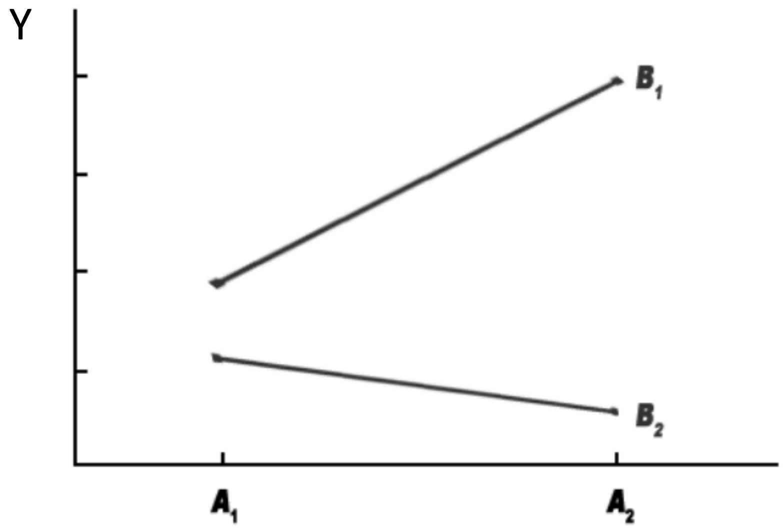
Two-Level, Two-Way ANOVA



Large effect of Factor A
No effect of Factor B
Slight interaction between Factor A and Factor B.

Two-Level, Two-Way ANOVA





No effect of Factor A
Large effect of Factor B
Very large interaction

Yes effect of Factor A
Large effect of Factor B
Large interaction