

Unit 6

Introduction to Analysis of Variance

“Always graph results of an analysis of variance”
- Gerald van Belle.

Analysis of variance is just a special case of normal theory regression. The response variable is continuous and assumed distributed normal (same as what we learned in Unit 5 – Regression. In an analysis of variance, however, all the predictor variables are categorical variables called factors. The possible values of the categorical predictors are called levels.

One-way analysis of variance: 1 categorical predictor with 3 or more levels

Two-sample t-test: One-way analysis of variance with 2 levels

Two-way analysis of variance: 2 categorical predictors (factors), regardless of the # levels

Three-way analysis of variance: 3 categorical predictors/factors . *And so on.*

So why the fuss? If an analysis of variance model is just a linear regression model with discrete predictors only and no continuous predictors, then why not just call it a day? There's a reason. The framework of analysis of variance (**in a nutshell: the manner in which the total variability in outcomes is separated out into its component portions**) works wonderfully for the analysis of many experimental designs (**this, as opposed to the framework we've been considering to this point which has been a regression analyses of observational data**).

In a **factorial design**, there are observations at every combination of levels of the factors. Factorial designs are good for exploring *interactions* (*effect modification*) between factors. An interaction between factor I and factor II is said to exist when the response to factor II depends on the level of factor I and vice versa.

In a **nested or hierarchical design**, such as a two-level nested design, the analysis is of units (eg-patients) that are clustered by level of factor I (eg- hospital) which are in turn clustered by level of factor II (eg – city). Nested designs are good for controlling for *confounding*.

A special type of nested design is the **longitudinal or repeated measurements design**. Repeated measurements are clustered within subjects and the repeated measurements are made over a meaningful dimension such as time (eg – growth over time in children) or space. The analysis of repeated measurements data is introduced elsewhere.

Unit 6 is an introduction to analysis of variance.

Table of Contents

| Topic | | |
|-------|--|----|
| | Learning Objectives | 3 |
| | 1. The Logic of Analysis of Variance | 5 |
| | 2. Introduction to Analysis of Variance Modeling | 12 |
| | 3. One Way Fixed Effects Analysis of Variance | 17 |
| | 4. Checking Assumptions of the One Way Analysis of Variance..... | 24 |
| | a. Tests for Homogeneity of Variance | 24 |
| | b. Graphical Assessments and Tests of Normality | 27 |
| | 5. Introduction to More Complicated Designs | 29 |
| | a. Balanced versus Unbalanced | 29 |
| | b. Fixed versus Random | 30 |
| | c. Factorial versus Nested | 31 |
| | 6. Some Other Analysis of Variance Designs | 33 |
| | a. Randomized Block Design | 33 |
| | b. Two Way Fixed Effects Analysis of Variance – Equal n | 36 |
| | c. Two Way Hierarchical or Nested Design | 48 |

Learning Objectives

When you have finished this unit, you should be able to:

- **Explain how** *analysis of variance is a special case of normal theory linear regression*.
- **Perform and interpret a** *one way analysis of variance*.
- **Explain what is meant by a** *multi-way analysis of variance*.
- **Explain what is meant by a** *factorial design analysis of variance*.
- **Explain the meaning of** *interaction of two factors*.
- **Explain what is meant by a** *nested design analysis of variance*.
- **Perform and interpret a** *two-way factorial analysis of variance, including an assessment of interaction*.

Introduction to Annoying Notation! Your Roadmap for Keeping Track

- With apologies, we move from Y to X:
Whereas previously, outcomes are represented using the notation Y
In analysis of variance, often, outcomes are represented using the notation X
- Keeping track of subscripts:
Subscripts are needed to identify group and individual

In the ONE WAY anova (X_{ij})

- “i” tells you the group, the level of the ONE categorical predictor (also called factor)
- “j” tells you the individual within that group

In the TWO WAY anova (X_{ijk})

- “i” tells you the level of the FIRST categorical predictor (also called Factor I)
- “j” tells you the level of the SECOND categorical predictor (also called Factor II)
- “k” tells you the individual within the two way classification of groups

- Subscripts that are DOTS are telling you that you are looking at a total
For example:

$X_{ij.}$ = sum of all the observations solely for the group defined by factor I at level=i, factor II at level=j; hence, no “k” subscript

$X_{i..}$ = sum of all the observations for the group defined by factor I at level=i,
So this is the sum taken over all levels of factor II (hence no “k” and no “j” subscript)

$X_{...}$ = sum of all the observations (hence no “k” and no “j” and no “i” subscripts)

- There are a lot of variances floating around and these now have subscripts too. Mostly. Your key:

S^2 = Sample variance of the entire pool all the observations (with no regard to group)

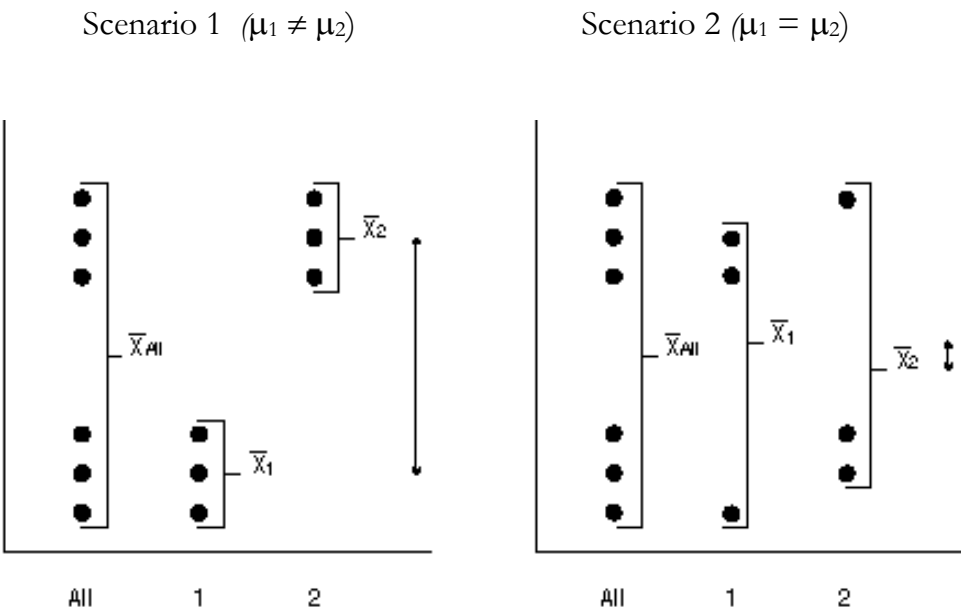
S_i^2 = Sample variance of **ONLY** the observations in the group defined by the categorical predictor at level = “i”

- **And those pesky bars on top:** BARS ON TOP are telling you that you are looking at an **average**

1. The Logic of Analysis of Variance

Analysis of Variance might seem like a misnomer, but it's not...

- Analysis of variance, like normal theory regression introduced in Unit 5, is the modeling and analysis of the variability of means. We ask: is the variability among the group level means statistically significantly greater than the variability of individuals within a group? Think "signal-to-noise".
- Consider the following scenarios.



| Scenario 1 means are <u>different</u> $\mu_1 \neq \mu_2$ | Scenario 2 means are the <u>same</u> $\mu_1 = \mu_2$ |
|---|---|
| S_1^2 and S_2^2 each summarize “noise” controlling for location. | S_1^2 and S_2^2 each summarize “noise” controlling for location. |
| The size of $ \bar{X}_1 - \bar{X}_2 $ is larger than “noise” | $ \bar{X}_1 - \bar{X}_2 $ is within the neighborhood of “noise”. |
| S^2 is <u>larger</u> than S_1^2 and S_2^2 because it is made larger by the extra variability among individuals due to change in location. | S^2 is <u>similar</u> in size to S_1^2 and S_2^2 because it does not have to accommodate an extra source of variability because of location differences between the two groups. |

When the sample size in each group is the same (and equal to n in each group), it’s easier to see how analysis of variance is an analysis of the *variability of the means*. When the samples sizes are not equal, the algebra is not so tidy.



Example - Stress and Heart Rate (... Or all things friends and pets! ...)

source: Gerstman BB. Basic Biostatistics: Statistics for Public Health Practice, pp259-262. The data used by Gerstman are from Allen et al (1991) Presence of human friends and pet dogs as moderators of autonomic responses to stress in women. *J. Personality and Social Psychology* 61(4); 582-589.

Does the companionship of a pet provide psychological relief to its owners when they are experiencing stress? In an experiment to address this question, consenting participants were randomized to one of three conditions: 1- Pet Present, 2-Friend Present, or 3-Neither friend nor pet present. Each participant was then exposed to a stressor (it happened to be mental arithmetic). The response variable is X = heart rate.

Selected Summary Statistics, by Group:

| | Group 1 Pet Present | Group 2 Friend Present | Group 3 Neither Pet nor Friend |
|--|---|---|---|
| n | $n_1 = 15$ | $n_2 = 15$ | $n_3 = 15$ |
| \bar{X} | $\bar{X}_1 = 73.48$ | $\bar{X}_2 = 91.33$ | $\bar{X}_3 = 82.52$ |
| S | $S_1 = 9.97$ | $S_2 = 8.34$ | $S_3 = 9.24$ |
| S^2 | $S_1^2 = 99.40$ | $S_2^2 = 69.57$ | $S_3^2 = 85.41$ |
| $(n-1)S^2$ | $(n_1-1)S_1^2 =$ | $(n_2-1)S_2^2 =$ | $(n_3-1)S_3^2 =$ |
| $= \sum (X - \bar{X})^2$ | $\sum_{j=1}^{n_1} (X_{1j} - \bar{X}_1)^2$ | $\sum_{j=1}^{n_2} (X_{2j} - \bar{X}_2)^2$ | $\sum_{j=1}^{n_3} (X_{3j} - \bar{X}_3)^2$ |
| | $= 1,391.57$ | $= 974.05$ | $= 1,195.70$ |

Analysis of Variance Question:

Do these data provided statistically significant evidence that the means of the stress scores (group 1 v 2 v 3) are different, ie – that μ_1 , μ_2 , and μ_3 are not equal?

The Reasoning in an Analysis of Variance

Proof by Contradiction

Signal-to-Noise

We illustrate with a one way fixed effects analysis of variance.

1. *As we always do in statistical hypothesis testing .. We begin by provisionally entertaining as true the null hypothesis. Here, the null hypothesis says that the means are equal.*

- We'll assume the null hypothesis is true, then apply this model to the observed data, and look to see if its application has led to an unlikely result, warranting rejection of the null hypothesis of equality of μ_1 , μ_2 , and μ_3 .

2. *Just as there are for regression, in analysis of variance some assumptions are required if we want to compute any probabilities (p-values and confidence intervals).*

Assumptions for a One Way Fixed Effects Analysis of Variance:

- $X_{11} \cdots X_{1n_1}$ are distributed Normal (μ_1, σ^2)
- $X_{21} \cdots X_{2n_2}$ are distributed Normal (μ_2, σ^2)
- $X_{31} \cdots X_{3n_3}$ are distributed Normal (μ_3, σ^2)
- The variances are all equal to σ^2 .
- The observations are all independent.

3. *Specify H_0 and H_A .*

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_A : not

4. *“Reason” an appropriate test statistic (Signal-to-Noise).*

NOISE:

“Within Group Variability = Noise” (Answers: What is the variability of individuals about their own group means?)

- In each of groups 1, 2, and 3 we obtain the separate, group-specific, S_1^2 , S_2^2 , and S_3^2
- Under the assumption of a common σ^2 , each of S_1^2 , S_2^2 , and S_3^2 is an estimate of the same common σ^2 .
- So we combine the 3 separate estimates of the common variance σ^2 into a single (better) guess of the common variance σ^2 . To do this, we compute a weighted average. The weights are the degrees of freedom of each of S_1^2 , S_2^2 , and S_3^2

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis

Noise is the within group variability and estimates the variability among individuals, controlling for location.

$$\text{Estimate of } \sigma^2 = \hat{\sigma}_{\text{within}}^2 = \frac{\sum_{i=1}^3 (n_i - 1) S_i^2}{\sum_{i=1}^3 (n_i - 1)}$$

Nice. The expected value of this estimate is the assumed common error variance

$$E \left[\frac{\sum_{i=1}^3 (n_i - 1) S_i^2}{\sum_{i=1}^3 (n_i - 1)} \right] = E \left[\hat{\sigma}_{\text{within}}^2 \right] = \sigma^2$$

SIGNAL:

“Between Group Variability = Signal” (Answers: What is the variability among the group means themselves?)

An appealing (because it is so intuitive) strategy would be to calculate the sample variance of the group specific means \bar{X}_1, \bar{X}_2 and \bar{X}_3 . What we actually do is a slight modification of this intuition. The modification takes into account group specific sample sizes. Specifically:

- We construct a special little data set with sample of size = 3, defined:

$$X_1^* = \sqrt{n_1} \bar{X}_1 = \sqrt{15} (73.4831) = 284.59882$$

$$X_2^* = \sqrt{n_2} \bar{X}_2 = \sqrt{15} (91.3251) = 353.70059$$

$$X_3^* = \sqrt{n_3} \bar{X}_3 = \sqrt{15} (82.5241) = 319.61446$$

The sample mean of these “special data set” means is $\bar{X}^* = 319.3046$

$$\sigma_{BETWEEN}^2$$

The expected value of the sample variance of our “starry” group means, X_1^* , X_2^* , and X_3^* is $\sigma_{BETWEEN}^2$:

$$E \left[\frac{\sum_{i=1}^3 (X_i^* - \bar{X}^*)^2}{(3-1)} \right] = \sigma_{between}^2$$

The expected value of the variability of the “starry” group means will be smaller or larger depending on how different the true means are, that is, depending on whether the null is true or not.

Null true: When the null hypothesis H_0 is true (the means are equal), and only when H_0 is true

$$E \left[\frac{\sum_{i=1}^3 (X_i^* - \bar{X}^*)^2}{(3-1)} \right] = \sigma_{between}^2 = \sigma^2$$

Alternative true: Otherwise (when the alternative hypothesis H_A is true),

the sample variance of X_1^* , X_2^* , and X_3^* is an estimate of a quantity ($\sigma_{between}^2$) that is larger than σ^2 .

$$E \left[\frac{\sum_{i=1}^3 (X_i^* - \bar{X}^*)^2}{(3-1)} \right] = \sigma_{between}^2 = \sigma^2 + \Delta \quad \text{where}$$

this is the amount larger and this is ALWAYS positive!

$$\Delta = \text{function}(\mu_1, \mu_2, \mu_3) > 0$$

Thus, the “signal” is related to the group specific means and, in particular, how different they are:

$$\Delta = \text{function}(\mu_1, \mu_2, \mu_3) > 0$$

In case you are interested: In a one-way analysis of # groups = K and equal sample size = n in each group:

$$\Delta = \frac{n \sum_{i=1}^K (\mu_i - \bar{\mu})^2}{K-1}$$

SIGNAL – to – NOISE:

The “Signal-to-Noise” analysis compares the *between* group means variability to the *within* groups variability. In the analysis of variance application, the comparison that is made is actually

$$\begin{aligned} \frac{\text{Noise} + \text{Signal}}{\text{Noise}} &= \frac{\text{Variability among a function of the group means}}{\text{Variability of individuals within groups}} \\ &= \frac{\text{Var}(X_1^*, X_2^*, X_3^*)}{\text{“ weighted (using df) sum of } S_1^2, S_2^2, \text{ and } S_3^2 \text{”}} \\ &= \left[\frac{\hat{\sigma}_{\text{between}}^2}{\hat{\sigma}_{\text{within}}^2} \right] \\ &= \frac{\left[\sum_{i=1}^3 (X_i^* - \bar{X}^*)^2 / (3-1) \right]}{\left[\left\{ \sum_{i=1}^3 (n_i - 1) S_i^2 \right\} / \left\{ \sum_{i=1}^3 (n_i - 1) \right\} \right]} \end{aligned}$$

5. Perform the calculations.

Using the values in the table on page 6, we have

$$\begin{aligned} \hat{\sigma}_{\text{within}}^2 &= \frac{\sum_{i=1}^3 (n_i - 1) S_i^2}{\sum_{i=1}^3 (n_i - 1)} = \frac{(1391.57 + 974.05 + 1195.70)}{(14 + 14 + 14)} \\ &= 84.79 \end{aligned}$$

Using the values of our starry means X_i^* on page 8, we also have

$$\hat{\sigma}_{\text{between}}^2 = \frac{\sum_{i=1}^3 (X_i^* - \bar{X}^*)^2}{(3-1)} = 1,193.836$$

Overall F Test. We use the F distribution to compare the variability of the means to the variability of individuals about their own means.

- Luckily, these two quantities are independent. When the null hypothesis H_0 is true:

- Overall $F = \frac{\hat{\sigma}_{between}^2 / \sigma^2}{\hat{\sigma}_{within}^2 / \sigma^2}$ is distributed F with

$$\text{Numerator degrees of freedom} = (k - 1) = (3 - 1) = 2$$

$$\text{Denominator degrees of freedom} = \sum (n_i - 1) = (3)(15 - 1) = 42$$

| | $\hat{\sigma}_{within}^2$ | $\hat{\sigma}_{between}^2$ | <u>Expected Value of</u> $\hat{\sigma}_{between}^2 / \hat{\sigma}_{within}^2$ | F | p-value |
|------------------------------|---------------------------|----------------------------|--|-----|---------|
| H_0 true (means equal) | σ^2 | σ^2 | 1 | 1 | Large |
| H_A true (means NOT equal) | σ^2 | $\sigma^2 + \Delta$ | > 1 | > 1 | Small |

For our data, $F = 1193.836 / 84.79 = 14.08$

The accompanying p-value is Probability[$F_{df=2,42} \geq 14.08$] = .00002.

6. “Evaluate” findings and report.

The assumption of the null hypothesis of equal means has led to an *extremely unlikely* result! The null hypothesis chances were approximately, 2 chances in 100,000 of obtaining 3 means of groups that are as different from each other as are 73.48, 91.33, and 82.52. The null hypothesis is rejected.

7. Interpret in the context of biological relevance.

This analysis provides statistically significant evidence of group differences in heart rate, depending on companionship by pet or by friend. But we do not know which, or if both, provides the benefit!! Okay then! I think we should all have lots of friends and lots of pets!

2. Introduction to Analysis of Variance Modeling

Preliminary note - See again page 4. In the pages that follow, I am using the notation X to refer to the *outcome variable* in analysis of variance. Previously, in unit 5, I used the notation Y (Sorry for the annoyance!)

Analysis of variance models, like regression models, have an identifiable basic structure.

Structure of Analysis of Variance Model

Observed = mean + random error

$$X = \mu + \varepsilon$$

\swarrow \uparrow \swarrow
 Observed Expected Random error
 data value is modeled

- μ - This is the expected value of X which we write as $\mu = E[X] = \text{linear model(stuff)}$
- ε - This is the idea of “random error”, “error in measurement”, “noise”
- **Subscripts** –Subscripts keep track of group membership and persons within groups.
For example: X_{ij} = Observed value for j^{th} person in the i^{th} group.

A special feature of analysis of variance models is their use of **subscripts**.

Example – One way fixed effects analysis of variance.

Structure of One-Way Fixed Effects Analysis of Variance Model

Observed = mean + random error

Subscripts:

“ i ” keeps track of the group.

“ j ” keeps track of the individual within the group.

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

\swarrow \uparrow \swarrow
 Observed Expected Random error
 data value value is the for “ j^{th} ” observation
 for “ j^{th} ” person mean of “ i^{th} ” of the “ i^{th} ” group mean.
 in “ i^{th} ” group group

Introduction to Defining an ANOVA Model

The One Way Fixed Effects Anova

In a one way fixed effects analysis of variance (anova) model, $E(X_{ij}) = \mu_i$ is *completely general*. That is, the group means can be anything. We are not modeling them as lying on a line or lying on any sort of functional form for that matter (e.g., polynomial). Recall. The notation $E(\)$ is referring to the expected value.

Example – Heart rate and stress, *continued* -

In this analysis, the null hypothesis was that the means are all the same. The alternative hypothesis said simply “the means are not all the same”

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_A : At least one is different .

- More generally, suppose the number of groups = K, instead of 3 in the heart rate example.
- KEY to subscripts for X_{ij} in a one way fixed effects anova.

Subscript “i” (identifies group)

The first subscript will be “i” and will index the groups $i=1, \dots, K$.

Subscript “j” (identifies individual within the group)

The second subscript will be “j” and will index the jth individual in the ith group $j = 1, \dots, n_i$.

n_i = sample size in the ith group

Let

μ_i = mean for persons in the subpopulation that is the i^{th} group

μ = overall mean, over the entire population, (that is over all subpopulations)

Deviation from means model. This is a nifty *re-write* that rewrites μ_i as a new expression that is equal to itself.

This is done by adding and subtracting μ to μ_i .

$$\begin{array}{ccccccc} \mu_i & = & \mu & + & (\mu_i - \mu) \\ \uparrow & & \uparrow & & \uparrow \\ \text{mean for} & & \text{overall} & & \text{deviation from} \\ \text{group “i”} & & \text{mean} & & \text{mean specific to } i^{\text{th}} \text{ group} \end{array}$$

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis

Same nifty trick to obtain a rewrite of the observed X_{ij} . Notice (below) that we are adding and subtracting two things this time:

$$X_{ij} = \bar{X} + (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i)$$

One more nifty maneuver lets us express the variability of individuals about the overall mean as the sum of two contributions. This is useful for analysis purposes.

$$(X_{ij} - \bar{X}) = (\bar{X}_i - \bar{X}) + (X_{ij} - \bar{X}_i)$$

Each source (between or within) contributes its own share to the total variability via the following (wonderful) result.

$$\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2 = \sum_{i=1}^K \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 + \sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

total variability
variability of means
"Between groups"
variability of individuals
"Within groups"

We keep track of all this in an analysis of variance table.

| Source | df ^a | Sum of Squares | Mean Square | Variance Ratio = F |
|---------------------------------------|--------------------------|---|--|--|
| Between groups (due model) | (K-1) | $\sum_{i=1}^K \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2$ | $\frac{\sum_{i=1}^K \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2}{(K-1)}$ $= \hat{\sigma}_{between}^2$ | $F = \frac{\hat{\sigma}_{between}^2}{\hat{\sigma}_{within}^2}$ |
| Within Groups (due error/residual) | $\sum_{i=1}^K (n_i - 1)$ | $\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$ | $\frac{\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^K (n_i - 1)}$ $= \hat{\sigma}_{within}^2$ | |
| Total | N - 1 | $\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$ | | |

^a degrees of freedom

Note: Just to be clear:

$$N = \text{grand total sample size, taken over all groups} = \sum_{i=1}^K n_i$$

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

Example – continued from page 5: Stress and Heart Rate

(source: Gerstman BB. *Basic Biostatistics: Statistics for Public Health Practice*, pp 259-262).

The data used by Gerstman are from Allen et al (1991) Presence of human friends and pet dogs as moderators of autonomic responses to stress in women. *J. Personality and Social Psychology* 61(4); 582-589.

Here are the data

| | Treatment | | |
|--|------------------------|---------------------------|------------------------------------|
| | 1 = Pet Present | 2 = Friend Present | 3 = Neither Pet, Nor Friend |
| | 69.17 | 99.69 | 84.74 |
| | 68.86 | 91.35 | 87.23 |
| | 70.17 | 83.40 | 84.88 |
| | 64.17 | 100.88 | 80.37 |
| | 58.69 | 102.15 | 91.75 |
| | 79.66 | 89.82 | 87.45 |
| | 69.23 | 80.28 | 87.78 |
| | 75.98 | 98.20 | 73.28 |
| | 86.45 | 101.06 | 84.52 |
| | 97.54 | 76.91 | 77.80 |
| | 85.00 | 97.05 | 70.88 |
| | 69.54 | 88.02 | 90.02 |
| | 70.08 | 81.60 | 99.05 |
| | 72.26 | 86.98 | 75.48 |
| | 65.45 | 92.49 | 62.65 |

Analysis of Variance Table

| Source | df ^a | Sum of Squares | Mean Square | Variance Ratio = F |
|----------------|-----------------|----------------|--------------------------------------|--|
| Between groups | 2 | 2387.69 | $\hat{\sigma}_{between}^2 = 1193.84$ | $F = \frac{\hat{\sigma}_{between}^2}{\hat{\sigma}_{within}^2} = 14.08$ |
| Within Groups | 42 | 3561.30 | $\hat{\sigma}_{within}^2 = 84.79$ | |
| Total | 44 | 5948.99 | | |

^a degrees of freedom

R Illustration

```

Descriptives by group.
library(FSA) # Summarize( ) in package {FSA} is nice for analysis of variance (imho)
load(file="pets.Rdata")

pets$group <- as.factor(pets$group) # group variable must be factor
FSA::Summarize(hrt_rate~group,data=pets,digits=2,na.rm=TRUE)
##          group n mean sd min Q1 median Q3 max
## 1      Pet Present 15 73.48 9.97 58.69 69.02 70.08 77.82 97.54
## 2      Friend Present 15 91.33 8.34 76.91 85.19 91.35 98.95 102.20
## 3 Neither Friend nor Pet 15 82.52 9.24 62.65 76.64 84.74 87.62 99.05

One-Way ANOVA
aov1 <- aov(hrt_rate ~ group, data=pets) # KEY: aov(YVAR ~ GROUPVAR, data= )
anova(aov1)
## Analysis of Variance Table
##
## Response: hrt_rate
##          Df Sum Sq Mean Sq F value    Pr(>F)
## group      2 2387.7 1193.84  14.079 0.00002092 ***
## Residuals 42 3561.3   84.79
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

MATCH! Hand calculations and R results match. Phew.

3. The One Way Fixed Effects Analysis of Variance

Not sure what “fixed” means?

Stay tuned. Fixed versus Random Effects are Introduced in Section 5b

A two independent samples t-test is exactly a one way analysis of variance; # groups = $K = 2$. The stress and heart rate example (see again, page 6) is a one way fixed effects analysis of variance with $K=3$. A two sample t-test is also one way analysis of variance; $K=2$. *If you square the value of your t-test, the result is the value of the anova F-test.*

| 2 independent groups t-test The BIOSTATS 540 “Lense” | 2 independent groups t-test The One-Way Anova “Lense” |
|---|--|
| Is the “signal” $(\bar{X}_1 - \bar{X}_2)$ large relative to “noise” where “noise” = $SE(\bar{X}_1 - \bar{X}_2)$? | Is the variability of (\bar{X}_1, \bar{X}_2) large relative to “noise” where “noise” = weighted average of S_1^2, S_2^2 |
| $t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)}$ | $F = \frac{\text{function of variability of data } \bar{X}_1, \bar{X}_2}{\text{function of variability of "noise" } S_1^2, S_2^2}$ |
| = measure of distance of $(\bar{X}_1 - \bar{X}_2)$ from 0, expressed on SE scale. | = measure of variability among \bar{X}_1, \bar{X}_2 (“signal”) to variability of individuals within groups (“noise”) |

Assumptions

1. Normality. The observed outcomes are distributed normal.

Group 1: $X_{11} \dots X_{1n_1}$ are a simple random sample from a $\text{Normal}(\mu_1, \sigma^2)$

Group 2: $X_{21} \dots X_{2n_2}$ are a simple random sample from a $\text{Normal}(\mu_2, \sigma^2)$

Etc.

Group K: $X_{K1} \dots X_{1n_K}$ are a simple random sample from a $\text{Normal}(\mu_K, \sigma^2)$

2. Constant variance. The K separate variance parameters are equal

3. Independence. The observations are independent

4. Null and Alternative Hypotheses

$$H_O : \mu_1 = \dots = \mu_K$$

H_A : At least one is different

One Way Fixed Effects Analysis of Variance

$$\text{Model of } E[X_{ij}] = \mu_i$$

Deviation from Means Coding

Recall -

“i” is the first subscript. It tells you the group.

“j” is the second subscript. It tells you the individual within the group.

Model for the mean in the ith group -

$$(1) \quad E[X_{ij}] = \mu_i$$

$$= [\mu + \tau_i] \quad \text{where}$$

$$(2) \quad \sum_{i=1}^K \tau_i = 0$$

Key: The “different-ness” of each mean is captured in the τ_1, \dots, τ_K . Notice the following:

Group 1: $\mu_1 = \mu + [\mu_1 - \mu] = \mu + \tau_1$ says that $[\mu_1 - \mu] = \tau_1$

Group K: $\mu_K = \mu + [\mu_K - \mu] = \mu + \tau_K$ says that $[\mu_K - \mu] = \tau_K$



psst This is deviation of means coding

- By definition, $\sum_{i=1}^K \tau_i = 0$

- If the means are NOT EQUAL, then at least one $\tau_i = [\mu_i - \mu] \neq 0$

One Way Analysis of Variance Fixed Effects Model

Setting:

K groups indexed $i = 1, 2, \dots, K$

Group specific sample sizes: n_1, n_2, \dots, n_K

X_{ij} = Observation for the j th individual in the i th group

The one way analysis of variance fixed effects model of X_{ij} is defined as follows:

$$X_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where

μ = grand mean

$$\tau_i = [\mu_i - \mu]$$

$$\sum_{i=1}^K \tau_i = 0$$

and

ε_{ij} is random error distributed $\text{Normal}(0, \sigma^2)$

Estimation

| Parameter | Estimate using Sample Data |
|---|---|
| μ = overall or “grand” mean | $\bar{X}_{..}$ |
| μ_i = mean for group defined by level “i” of predictor | $\bar{X}_{i.}$ |
| $\tau_i = [\mu_i - \mu]$ measures “differentness” of i th group mean relative to overall or “grand” mean | $[\bar{X}_{i.} - \bar{X}]$ |
| σ^2 = the assumed common within group variance | $\hat{\sigma}_{within}^2 = \frac{\sum_{i=1}^K (n_i - 1) S_i^2}{\sum_{i=1}^K (n_i - 1)}$ |

| | Expected Value of | | | | p-value |
|------------------------------|---------------------------|----------------------------|--|-------|---------|
| | $\hat{\sigma}_{within}^2$ | $\hat{\sigma}_{between}^2$ | $\hat{\sigma}_{between}^2 / \hat{\sigma}_{within}^2$ | F | |
| H_0 true (means equal) | σ^2 | σ^2 | 1 | 1 | Large |
| H_A true (means NOT equal) | σ^2 | $\sigma^2 + \Delta$ | > 1 | > 1 | Small |

Where ...
$$\Delta = \frac{n \sum_{i=1}^K (\mu_i - \bar{\mu})^2}{K - 1}$$
 where n = common sample size in each group.

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis

Example

Three groups of physical therapy patients were subjected to different treatment regimens. At the end of a specified period of time each was given a test to measure treatment effectiveness. The sample size was $n=4$ in each group. The following scores were obtained.

| Treatment | | |
|---|---|---|
| 1 | 2 | 3 |
| 75 | 25 | 100 |
| 80 | 75 | 80 |
| 75 | 25 | 100 |
| 50 | 75 | 40 |
| $\bar{X}_1=70$ | $\bar{X}_2=50$ | $\bar{X}_3=80$ |
| $X_1^*=\sqrt{n}[\bar{X}_1]=(2)(70)=140$ | $X_2^*=\sqrt{n}[\bar{X}_2]=(2)(50)=100$ | $X_3^*=\sqrt{n}[\bar{X}_3]=(2)(80)=160$ |

$$\bar{X}^* = \frac{\sum_{i=1}^3 X_i^*}{3} = 133.33$$

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_A: \text{not}$$

Step 1: Test the assumption of equality of variances.

Stay tuned. Tests of the assumption of equality of variances are discussed later (see, Section 4a). These are of limited usefulness for two reasons:

- (1) Tests of equality of variance tend to be sensitive to the assumption of normality.
- (2) Analysis of variance methodology is pretty robust to violations of the assumption of a common variance.

Step 2: Estimate the within group variance (“noise”). This will be a weighted average of the k separate sample variances.

$$\hat{\sigma}_{within}^2 = \frac{\sum_{i=1}^{K=3} (n_i - 1) S_i^2}{\sum_{i=1}^3 (n_i - 1)} = \frac{5450.00}{9} = 605.56$$

Step 3: Estimate the between group variance (“noise + signal”). This will be a sample variance calculation for our starry means data set comprised of $X_1^* = \sqrt{n}\bar{X}_1$, $X_2^* = \sqrt{n}\bar{X}_2$, $X_3^* = \sqrt{n}\bar{X}_3$ ’s
 Forgot what’s meant by “noise”, “signal to noise” and “noise + signal”? Take a look back at page 10.

$$\hat{\sigma}_{between}^2 = \left[\frac{\sum_{i=1}^3 (X_i^* - \bar{X}^*)^2}{(3-1)} \right] = \frac{1866.67}{2} = 933.33$$

Step 4: Summarize in an analysis of variance table. Perform F-test of null hypothesis of equal means. **Note:** This is an overall F Test

| Source | df ^a | Sum of Squares | Mean Square | Variance Ratio = F |
|---------------------------------------|------------------------------|--|---|---|
| Between groups (due model) | (K-1) = 2 | $\sum_{i=1}^K \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2$ = 1866.67 | $\sum_{i=1}^K \sum_{j=1}^{n_i} (\bar{X}_i - \bar{X})^2 / (K-1)$ = $\hat{\sigma}_{between}^2 = 933.33$ | $F = \frac{\hat{\sigma}_{between}^2}{\hat{\sigma}_{within}^2} = 1.54$ |
| Within Groups (due error/residual) | $\sum_{i=1}^K (n_i - 1) = 9$ | $\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$ = 5450.00 | $\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 / \sum_{i=1}^K (n_i - 1)$ = $\hat{\sigma}_{within}^2 = 605.56$ | |
| Total | (N - 1) ^b | $\sum_{i=1}^K \sum_{j=1}^{n_i} (X_{ij} - \bar{X})^2$ = 7316.67 | | |

^a degrees of freedom

^bN = grand total sample size, taken over all groups = $\sum_{i=1}^K n_i$

p-value = Pr [F_{DF=2,9} ≥ 1.54] = .27

Conclusion. The null hypothesis is not rejected. These data do not provide statistically significant evidence of differences in group means, that effectiveness of treatment is different, depending on the type of treatment received (“1” versus “2” versus “3”).

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

R Illustration

```

Input data directly
pt_table=read.table(text="
txgroup yscore
1.00    75.00
1.00    80.00
1.00    75.00
1.00    50.00
2.00    25.00
2.00    75.00
2.00    25.00
2.00    75.00
3.00    100.00
3.00    80.00
3.00    100.00
3.00    40.00",header=TRUE)
ptdata <- as.data.frame.matrix(pt_table)
ptdata$txgroup <- as.factor(ptdata$txgroup)

```

Anova.

```

aov2 <- aov(yscore ~ txgroup, data=ptdata)
anova(aov2)
## Analysis of Variance Table
##
## Response: yscore
##           Df Sum Sq Mean Sq F value Pr(>F)
## txgroup    2 1866.7   933.33   1.5413 0.2657
## Residuals  9 5450.0   605.56

```

Interpretation: This matches the hand calculation. The null hypothesis is not rejected. Conclude these data do not provide statistically significant evidence that effectiveness of treatment is different, depending on the type of treatment received ("1" versus "2" versus "3").

Step 5: Don't forget to look at your data!

R Illustration

```
library(FSA)
```

Numerical Descriptives

Descriptives - Overall

```
FSA::Summarize(yscore ~ 1,data=ptdata,digits=2, na.rm=TRUE) # KEY: (yvar ~ 1) for descriptives overall
```

```
##      n  mean   sd  min   Q1 median   Q3   max
## 12.00 66.67 25.79 25.00 47.50 75.00 80.00 100.00
```

Descriptives - by Treatment Group

KEY: (yvar ~ groupvar) for descriptives, by group

```
FSA::Summarize(yscore ~ txgroup,data=ptdata,digits=2, na.rm=TRUE)
```

```
## txgroup n mean   sd min   Q1 median   Q3 max
## 1      4 66.67 25.79 25.00 47.50 75.00 80.00 100.00
## 2      4 66.67 25.79 25.00 47.50 75.00 80.00 100.00
## 3      4 66.67 25.79 25.00 47.50 75.00 80.00 100.00
```

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

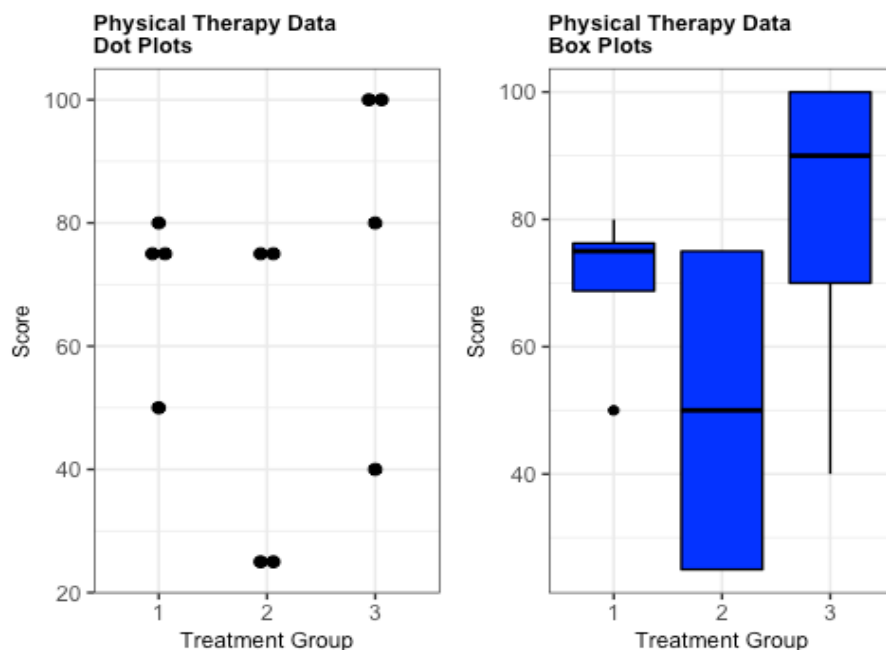
Graphical Descriptives

```
library(ggplot2)
library(gridExtra)

# panel 1 = side-by-side dot plot
p1 <- ggplot(data=ptdata,aes(x=factor(txgroup),y=yscore)) +
  geom_dotplot(dotsize=0.75,binaxis = "y",
    stackdir = "center",binpositions="all") +
  xlab("Treatment Group") +
  ylab("Score") +
  ggtitle("Physical Therapy Data \nDot Plots") +
  theme_bw() +
  theme(axis.text = element_text(size = 9),
    axis.title = element_text(size = 9),
    plot.title = element_text(size = 9, face = "bold"))

# panel 2 = side-by-side box plot
p2 <- ggplot(data=ptdata,aes(x=factor(txgroup),y=yscore)) +
  geom_boxplot(color="black",fill="blue") +
  xlab("Treatment Group") + ylab("Score") +
  ggtitle("Physical Therapy Data \nBox Plots") +
  theme_bw() +
  theme(axis.text = element_text(size = 9),
    axis.title = element_text(size = 9),
    plot.title = element_text(size = 9, face = "bold"))

# Create arrange the two panels (side-by-side) in one graph using grid.arrange( ) in {gridExtra}
gridExtra::grid.arrange(p1, p2, ncol=2)
```



4. Checking Assumptions of the One Way Analysis of Variance

a. Tests for Homogeneity of Variance

Violation of the assumption of homogeneity of variances is sometimes, *but not always*, a problem.

- Recall that the denominator of the overall F statistic in a one way analysis of variance is the “within group mean square.” It is a weighted average of the separate within group variance estimates S^2 and, as such, is an estimate of the assumed common variance.
 - When the within group variance parameters $\sigma_1^2, \sigma_2^2, \dots, \sigma_K^2$ are **at least reasonably similar**, then the within group mean square is a good summary of the within group variability.
- The **overall F** test for equality of means in an analysis of variance is **reasonably robust** to moderate violations of the assumption of homogeneity of variance.
- However, **pairwise t-tests** and hypothesis tests of **contrasts** are **not robust** to violations of homogeneity of variance, as when the $\sigma_1^2, \sigma_2^2, \dots, \sigma_K^2$ are very unequal.

Tests of homogeneity of variance are appropriate for the one-way analysis of variance only.

There are a variety of tests available.

- F Test for Equality of Two Variances** – This was introduced in BIOSTATS 540 Unit 10, Section 2d *Hypothesis Testing*, p 24 (<https://people.umass.edu/biep540w/pdf/10.%20TWO%20Sample%20Inference%202019.pdf>)
- Bartlett’s test** - This test has high statistical power when the assumption of normality is met. However, it is very sensitive to the assumption of normality.
- Levene’s test** – This test has the advantage of being much less sensitive to violations of normality. Its disadvantage is that it has less power than Bartlett’s test.
- Brown-Forsythe test**, also called **Levene (med) test** – Similar to Levene’s Test.

Bartlett's Test

- $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_K^2$
 $H_A: \text{At least one } \sigma_i^2 \text{ is unequal to the others}$
- Obtain the K separate group-specific sample variances S_1^2, \dots, S_K^2
- Obtain $\hat{\sigma}_{within}^2 = \frac{\sum_{i=1}^K (n_i - 1) S_i^2}{\sum_{i=1}^K (n_i - 1)}$ the estimate of the (null hypothesis) common σ^2
- Compute $B = [\ln(\hat{\sigma}_{within}^2)] \left(\sum_{i=1}^K (n_i - 1) \right) - \sum_{i=1}^K (n_i - 1) \ln(S_i^2)$ *note – Some texts use B as the test statistic.*
- Compute $C = 1 + \frac{1}{3(K-1)} \left\{ \sum_{i=1}^K \frac{1}{(n_i - 1)} - \frac{1}{\sum_{i=1}^K (n_i - 1)} \right\}$ *-note- C is a correction factor*
- Compute Bartlett Test Statistic = $\frac{B}{C}$ *note – the distribution of B/C is better approximated by chi square.*
- When the null hypothesis is true, Bartlett Test Statistic is distributed Chi square (df=K-1)
 - Reject null for large values of Bartlett Test

Levene's Test (“Dispersion variable Analysis of Variance”)

The idea of Levene's test and its modification is to create a new random variable, a *dispersion random variable* which we will represent as **d** and that is a measure of how the variances are different. Levene's test (and its modifications) is a one way analysis of variance on a dispersion random variable **d**.

- $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_K^2$
 $H_A: \text{At least one } \sigma_i^2 \text{ is unequal to the others}$
- Compute $d_{ij} = |X_{ij} - \bar{X}_i|$
- Perform a one way analysis of variance of the d_{ij}
- When the null hypothesis is true, the Levene Test One Way Analysis of Variance is distributed F (numerator df = K-1, denominator df = N-K)
 - Reject null for large values of Levene Test One Way Anova F

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis

Brown and Forsythe Modification of Levene's Test (“Dispersion variable Analysis of Variance”)

The Brown and Forsythe modification of Levene's test utilizes as its dispersion random variable d the absolute deviation from the group median.

- $H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_K^2$
 H_A : At least one σ_i^2 is unequal to the others
- Compute $\tilde{d}_{ij} = |X_{ij} - \text{median}(X_{i1}, X_{i2}, \dots, X_{in_i})|$
- Perform a one way analysis of variance of the \tilde{d}_{ij}
- When the null hypothesis is true, the Brown and Forsythe One Way Analysis of Variance is distributed F (numerator $df = K-1$, denominator $df = N-K$)
 - Reject null for large values of Brown and Forsythe Test One Way Anova F

R Illustration

```
library(car) # leveneTest( ) in package {car}

# Bartlett Test
bartlett.test(yscore~txgroup, data=ptdata)
##
## Bartlett test of homogeneity of variances
##
## data: yscore by txgroup
## Bartlett's K-squared = 1.5601, df = 2, p-value = 0.4584

Interpretation: Do NOT reject the null (p-value = .46). Assumption of the null has not led to a
contradiction. Conclude there is NO statistically significant evidence that the variances are unequal.

# Brown and Forsythe Test in R is called a “Modified Levene Test”

car::leveneTest(yscore~txgroup, data=ptdata)
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 2    1.8  0.22
##      9

Interpretation: Same. Do NOT reject the null (p-value = .22). Assumption of the null has not led to a
contradiction. Conclude there is NO statistically significant evidence that the variances are unequal.
```

Statistics in Practice: Guidelines for Assessing Homogeneity of Variance

- Look at the variances (or standard deviations) for each group first!
 - Compare the numeric values of the variances (or standard deviations)
 If the ratio of the standard deviations is less than 3 (or so), it's okay
 not to worry about homogeneity of variances
 - Construct a side-by-side box plot of the data and have a look at the sizes of the boxes.

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

- **Second, assess the reasonableness of the normality assumption.** This is important to the validity of the tests of homogeneity of variances.
- **Levene's test** of equality of variances is least affected by non-normality; it's a good choice.
- **Bartlett's test** should be used with caution, given its sensitivity to violations of normality.

b. Graphical Assessments and Tests of Normality

Graphical assessments and tests of normality were introduced previously. See again BIOSTATS 640 Unit 5, Regression & Correlation, page 47.

Analysis of Variance methods are **reasonably robust** to violation of the assumption of normality in analysis of variance.

- The assumption of normality in a one way analysis of variance is that within each of the K groups the individual X_{ij} for $j = 1, 2, \dots, n_i$ are assumed to be a simple random sample from a $\text{Normal}(\mu_i, \sigma^2)$,
- The analysis of variance is actually relying on normality of the **sampling distribution of the means** \bar{X}_i for $i = 1, 2, \dots, K$. This is great because we can appeal to the central limit theorem.
- Thus, provided that the sample sizes in each group are reasonable (say 20-30 or more), and provided the underlying distributions are not too too different from normality, then the analysis of variance is reasonably robust to violation of the assumption of normality.

As with normal theory regression, assessments of normality are of two types in analysis of variance.

- Preliminary is to calculate the **residuals**:
 - For X_{ij} = observation for "j"th person in group i
 - Residual $r_{ij} = (X_{ij} - \bar{X}_i)$ **difference between observed and mean for group**
- **1. Graphical Assessments of the distribution of the residuals:**
 - Dot plots with overlay normal
 - Quantile-Quantile plots using referent = normal
- **2. Numerical Assessments:**
 - Calculation of skewness and kurtosis statistics
 - Shapiro Wilk test of normality
 - Kolmogorov Smirnov/Lillefors tests of normality
 - Anderson Darling/Cramer von Mises tests of normality

Nature — Population/
Sample — Observation/
Data — Relationships/
Modeling — Analysis/
Synthesis

Statistics in Practice: Guidelines for the Handling of Violations of Normality

- **Small sample size setting (within group sample size $n < 20$, approximately):**
 - **Replace the normal theory one-way analysis of variance with a [Kruskal Wallis](#) nonparametric one way analysis of variance.** You can find a detailed description in Unit 3, [Nonparametrics](#)
- **Large sample size setting:**
If you *really must*, consider a normalizing data transformation. Possible transformations include the following:
 - (1) [Logarithmic Transformation](#): $X^* = \ln(X+1)$ helps positive skewness
 - (2) [Square Root Transformation](#): $X^* = \sqrt{X+0.5}$ helps heteroscedasticity
 - (3) [Arcsine Transformation](#): $p^* = \arcsine \sqrt{p}$ for outcome 0 to 100 percentage
 - (4) If your data are actual proportions of the type X/n and you have X and n consider [Anscombe Arcsine Transformation](#):

$$p^* = \arcsine \sqrt{\frac{X + \frac{3}{8}}{n + \frac{3}{4}}}$$

5. Introduction to More Complicated Designs

So far we have considered just one analysis of variance design: the one-way analysis of variance design.

- 2 or more groups were compared (e.g. – we compared 4 groups)
- However, the groups represented levels of just 1 factor (e.g. race/ethnicity)

Analysis of variance methods can become more complicated for a variety of reasons, including but not limited to the following-

- (1) The model may have more terms, including interactions and/or adjustment for confounding
Fitting and interpretation become more challenging.
- (2) One or more of the terms in the model might be measured with error instead of being fixed –
Estimates of variance, their interpretation and confidence interval construction are more involved.
- (3) The partitioning of total variability might not be as straightforward as what we have seen –
Understanding and working with analysis of variance tables and, especially, knowing which F test to use, can be hard.

a. Balanced versus Unbalanced

The distinction pertains to the partitioning of the total variability and, specifically, the complexity involved in the variance components and their estimation.

BALANCED

The sample size in each cell is the same.

Equality of sample size makes the analysis easier.

Specifically, the partitioning of SSQ is straightforward

A 2 way balanced anova with $n=1$ is called the randomized block design

UNBALANCED

The sample sizes in the cells are different.

The partitioning of SSQ is no longer straightforward.

Here, a regression approach (reference cell coding – this comes later)
is sometimes easier to follow.



b. Fixed versus Random Effects

Hack! “Fixed” versus “random” is more complicated than you might think.

- The distinction has to do with *how the inferences will be used*
- The formal analysis of variance is largely *unchanged*.

There exist a number of definitions of “fixed” versus “random” effects. Among them are the following.

- **Fixed effects** are levels of effects chosen by the investigator, whereas **Random effects** are selected at random from a larger population
- **Fixed effects** are either (1) levels chosen by the investigator or (2) *all* the levels possible
Random effects are a random sample from some universe of all possible levels.
- **Fixed effects** are effects that are interesting in themselves, whereas Effects are investigated as **random** if the underlying population is of interest.

Examples

- **Examples of Fixed Effects** – (1) Treatment (medicine v surgery); (2) Gender (all genders)
- **Examples of Random Effects** – (1) litter of animals (this is an example of a random block); (2) interviewer in a data collection setting where there might be multiple interviewers or raters.

Illustration of “Fixed” versus “Random” Thinking

In “fixed” versus “random”, the ways we think about the null hypothesis are slightly different. Consider a **one way anova** analysis which explores variations in SAT scores, depending on the university affiliation of the students.

- Outcome is X_{ij} = SAT score for “j”th individual at University “i”
- Factor is University with

i = 1 if University is Massachusetts
2 if University is Wisconsin
3 if University is Alaska

- Subscript “j” indexes student within the University

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis

- FIXED effects perspective

Interest pertains only to the 3 Universities (MA, WI, and AK). → Null hypothesis is

$$H_0: \mu_{\text{Massachusetts}} = \mu_{\text{Wisconsin}} = \mu_{\text{Alaska}}$$

- RANDOM effects perspective

Massachusetts, Wisconsin, and Alaska are a random sample from the population of universities in the US. → Null hypothesis is

H_0 : Mean SAT scores are equal at all American Universities

c. Factorial versus Nested

The distinction “factorial” versus “nested” is an important distinction pertaining to the discovery of **interaction or effect modification** (factorial) versus **control for confounding** (nested).

To discover effect modification → Factorial Design

To control for (eliminate) confounding → Nested Design

Consider the context of a two way analysis of variance that explores Factor A at “a” levels and Factor B at “b” levels

FACTORIAL

All combinations of factor A and factor B are investigated.

Factorial design permits investigation of A x B interaction

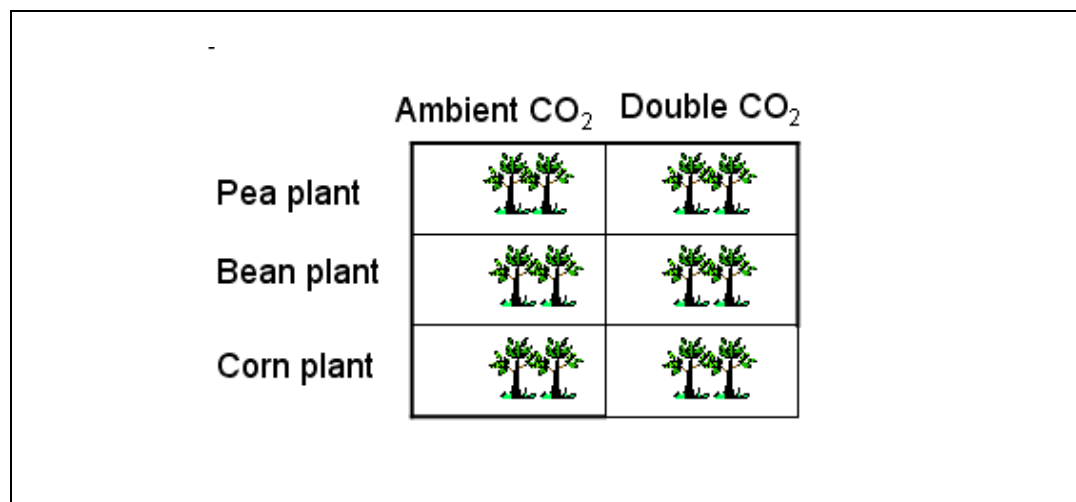
Thus, good for exploration of effect modification, synergism, etc.

Frequently used in public health, observational epidemiology



Example of Factorial Design.

Factor A = Plant at 3 levels (rows) and Factor B = CO₂ at 2 levels (columns)



Note: All $(3)(2) = 6$ combinations of plant x CO₂ are investigated

NESTED

The levels of the second factor are nested in the first.

Confounding (by Factor A) of the Factor B- Outcome relationship is controlled through use of stratification on Factor A

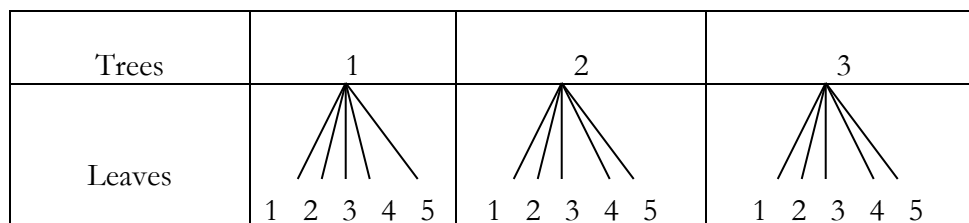
Familiar examples are *hierarchical, split plot, repeated measures, mixed models*.

Nested designs are frequently used in biology, psychology, and complex survey methodologies.

Factor A, the stratifying variable, is sometimes called the “primary sampling unit.”

Example of Nested Design.

Factor A = Trees at 3 levels and Factor B = Leaf at 5 levels, nested within its trees of belonging!



6. Some Other Analysis of Variance Designs

What design should you use?

In brief, the answer depends on (1) the research question (2) knowledge of underlying biology and, specifically, knowledge of external influences that might be effect modifying or confounding or both and (3) availability of sample size.

Briefly, three other analysis of variance designs are introduced here

- The Randomized Complete Block Design
- Two Way Fixed Effects Analysis of Variance – Equal cell numbers
- The Two Way Hierarchical or Nested Design

a. The Randomized Complete Block Design

Example -

An investigator wishes to compare 3 treatments for HIV disease. However, it is suspected that response to treatment might be confounded by baseline cd4 count. The investigator seeks to control (better yet, eliminate) confounding. To accomplish this, consenting subjects are grouped into 8 “homogeneous” blocks according to cd4 count. Within each block, baseline cd4 counts are assumed to be similar.

Within each block, there are 3 subjects, *one per treatment*. Assignment of subject to treatment within each block is *randomized*.

Data Layout –

| Block is Stratum of cd4 count, $i =$ | Treatment, $j =$ | | |
|--------------------------------------|------------------|--------|--------|
| 1 | Drug 3 | Drug 1 | Drug 2 |
| 2 | Drug 2 | Drug 1 | Drug 3 |
| 3 | ... | ... | ... |
| 4 | ... | ... | ... |
| 5 | ... | ... | ... |
| 6 | ... | ... | ... |
| 7 | ... | ... | ... |
| 8 | Drug 1 | Drug 3 | Drug 2 |

While there are two factors ..

- The row factor is called a blocking factor; its influence on outcome is *not* of interest. But we do want to control for its possible confounding effect of baseline CD4 count
- Only the column factor is of interest – Treatment (drug) at 3 levels.

A characteristic of a randomized complete block design is that the **sample size is ONE** in each *Treatment* \times *Block* combination.

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

Randomized Complete Block Design Model

Setting:

I blocks or treatments indexed $i = 1, 2, \dots, I$

J treatments indexed $j = 1, 2, \dots, J$

Sample size is 1 in each block x treatment combination

X_{ij} = Observation for the one individual in the i^{th} block who received the j^{th} group/treatment

The randomized complete block design model of X_{ij} is defined as follows:

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

Where -

μ = grand mean

$$\alpha_i = [\mu_i - \mu] \text{ and } \sum_{i=1}^I \alpha_i = 0$$

$$\beta_j = [\mu_j - \mu] \text{ and } \sum_{j=1}^J \beta_j = 0$$

and -

ε_{ij} is random error distributed $\text{Normal}(0, \sigma^2)$

Randomized Complete Block Design Analysis of Variance

“i” indexes block, $i = 1 \dots I$

“j” indexes treatment, $j = 1 \dots J$

μ_{ij} = Mean [Outcome] for drug “j” in block “i”

$i = 1 \dots I$

In this example, $I = 8$ because there are 8 blocks

$j = 1, 2, \dots, J$

In this example, $J = 3$ because there are 3 treatments

$$E[X_{ij}] = \mu_{ij} = \mu + \alpha_i + \beta_j \rightarrow$$

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

$$(1) \quad \alpha_i = [\mu_{i.} - \mu] \text{ and } \sum_{i=1}^I \alpha_i = 0$$

$$(2) \quad \beta_j = [\mu_{.j} - \mu] \text{ and } \sum_{j=1}^J \beta_j = 0$$

$$X_{ij} = \bar{X}_{..} + [\bar{X}_{i.} - \bar{X}_{..}] + [\bar{X}_{.j} - \bar{X}_{..}] + [X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}] \text{ algebraic identity}$$

- ϵ_{ij} is assumed distributed Normal(0, σ^2)
- Because $n=1$ in each cell, **block x treatment interactions**, if they exist, **cannot be estimated**.
(Bummer – this means we cannot assess affect modification)

Total SSQ and its Partitioning

$$X_{ij} = \bar{X}_{..} + [\bar{X}_{i.} - \bar{X}_{..}] + [\bar{X}_{.j} - \bar{X}_{..}] + [X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}] \rightarrow$$

$$[X_{ij} - \bar{X}_{..}] = [\bar{X}_{i.} - \bar{X}_{..}] + [\bar{X}_{.j} - \bar{X}_{..}] + [X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}].$$

Squaring both sides and summing over all observations yields (because the cross product terms sum to zero!)

$$\begin{aligned} & \sum_{i=1}^I \sum_{j=1}^J [X_{ij} - \bar{X}_{..}]^2 \\ &= \sum_{i=1}^I \sum_{j=1}^J [\bar{X}_{i.} - \bar{X}_{..}]^2 + \sum_{i=1}^I \sum_{j=1}^J [\bar{X}_{.j} - \bar{X}_{..}]^2 + \sum_{i=1}^I \sum_{j=1}^J [X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}]^2 \\ &= J \sum_{i=1}^I [\bar{X}_{i.} - \bar{X}_{..}]^2 + I \sum_{j=1}^J [\bar{X}_{.j} - \bar{X}_{..}]^2 + \sum_{i=1}^I \sum_{j=1}^J [X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..}]^2 \end{aligned}$$

Analysis of Variance Table

| Source | df ^a | Sum of Squares | E (Mean Square) | F |
|-----------------------------|-----------------|---|---|--|
| Due <u>block</u> | (I-1) | $J \sum_{i=1}^I (\bar{X}_{i.} - \bar{X}_{..})^2$ | $\sigma^2 + J \left[\frac{\sum_{i=1}^I \alpha_i^2}{(I-1)} \right]$ | $F = \frac{MSQ_{\text{block}}}{MSQ_{\text{residual}}}$ df = (I-1), (I-1)(J-1) |
| Due <u>treatment</u> | (J-1) | $I \sum_{j=1}^J (\bar{X}_{.j} - \bar{X}_{..})^2$ | $\sigma^2 + I \left[\frac{\sum_{j=1}^J \beta_j^2}{(J-1)} \right]$ | $F = \frac{MSQ_{\text{treatment}}}{MSQ_{\text{residual}}}$ df = (J-1), (I-1)(J-1) |
| Residual | (I-1)(J-1) | $\sum_{i=1}^I \sum_{j=1}^J (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$ | σ^2 | |
| Total | IJ - 1 | $\sum_{i=1}^I \sum_{j=1}^J (X_{ij} - \bar{X}_{..})^2$ | | |

^a degrees of freedom

b. The Two Way Fixed Effects Analysis of Variance

Example -

Fish growth is thought to be influenced by light or by temperature or by both and possibly by both in combination. To explore this, an investigator considered all possible combinations of light at 2 levels and water temperature at 3 levels. Thus the total number of combinations of light and temperature is $2 \times 3 = 6$. This is an example of a *two way factorial analysis of variance*.

Outcome X = fish growth at six weeks

- Factor I is Light at 2 levels (low and high)
- Factor II is Water Temperature at 3 levels (cold, lukewarm, warm)
- **X_{ijk}** = growth at six weeks for the **kth** fish at **light level=i** and **water temperature=j**

Following are the data

| Light | Water Temp | X = Fish Growth |
|--------|------------|-----------------|
| 1=low | 1=cold | 4.55 |
| 1=low | 1=cold | 4.24 |
| 1=low | 2=lukewarm | 4.89 |
| 1=low | 2=lukewarm | 4.88 |
| 1=low | 3=warm | 5.01 |
| 1=low | 3=warm | 5.11 |
| 2=high | 1=cold | 5.55 |
| 2=high | 1=cold | 4.08 |
| 2=high | 2=lukewarm | 6.09 |
| 2=high | 2=lukewarm | 5.01 |
| 2=high | 3=warm | 7.01 |
| 2=high | 3=warm | 6.92 |

There are 3 analysis questions and, thus, three null hypotheses of interest:

- (1) Ho: No effect due to light,
e.g. mean fish length is the same over the two levels of light
- (2) Ho: No effect due to temperature
e.g. mean fish length is the same over the three levels of temperature
- (3) Ho: Not a differential effect
of one treatment over the levels of the other (e.g. no interaction)

Note – As we will see in the next page, the order in which we test these hypotheses matters!

Two Way Factorial Analysis of Variance Fixed Effects Model

Setting:

I levels of Factor #1 and are indexed $i = 1, 2, \dots, I$

J levels of Factor #2 and are indexed $j = 1, 2, \dots, J$

The sample size for the group defined by Factor 1 at level “i” and Factor 2 at level “j” is n_{ij}

“k” indexes the kth observation in the “ij”th group

X_{ijk} = Observation for the kth individual in the jth block of the ith group/treatment

The two way factorial analysis of variance fixed effects model of X_{ijk} is defined as follows:

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

where

μ = grand mean

$$\alpha_i = [\mu_{i.} - \mu] \text{ and } \sum_{i=1}^I \alpha_i = 0$$

$$\beta_j = [\mu_{.j} - \mu] \text{ and } \sum_{j=1}^J \beta_j = 0$$

$$(\alpha\beta)_{ij} = \mu_{ij} - [\mu + \alpha_i + \beta_j] = \mu_{ij} - \mu - \alpha_i - \beta_j = \mu_{ij} - \mu - [\mu_{i.} - \mu] - [\mu_{.j} - \mu]$$

$$\sum_{i=1}^I (\alpha\beta)_{ij} = 0 \text{ and } \sum_{j=1}^J (\alpha\beta)_{ij} = 0$$

and

ε_{ij} is random error distributed $\text{Normal}(0, \sigma^2)$

Two Way Factorial Fixed Effects Analysis of Variance

Let i index light, $i = 1, 2$.

j index water temperature, $j = 1, 2, 3$.

k index individual in the $(i, j)^{\text{th}}$ group, $k = 1, \dots, n_{ij}$.

μ_{ij} = Expected mean growth at 6 weeks for fish raised under conditions of light = “ i ” and water temperature = “ j ”.

$$\mu_{ij} = \mu + [\mu_{i.} - \mu] + [\mu_{.j} - \mu] + [\mu_{ij} - \mu - (\mu_{i.} - \mu) - (\mu_{.j} - \mu)]$$

μ = Overall population mean

$\alpha_i = [\mu_{i.} - \mu]$ is the light effect. It is estimated by $[\bar{X}_{i.} - \bar{X}_{...}]$

$\beta_j = [\mu_{.j} - \mu]$ is the water temperature effect. It is estimated by $[\bar{X}_{.j.} - \bar{X}_{...}]$

$(\alpha\beta)_{ij} = [\mu_{ij} - \mu - (\mu_{i.} - \mu) - (\mu_{.j} - \mu)]$ is the extra, joint, effect of the i^{th} light level and j^{th} water temperature. It is estimated by $[\bar{X}_{ij.} - \bar{X}_{...} - (\bar{X}_{i.} - \bar{X}_{...}) - (\bar{X}_{.j.} - \bar{X}_{...})]$

Thus, an individual response X_{ijk} is modeled

$$\begin{aligned} X_{ijk} &= \mu_{ij} + \varepsilon_{ijk} \\ &= \mu + [\mu_{i.} - \mu] + [\mu_{.j} - \mu] + [\mu_{ij} - \mu - (\mu_{i.} - \mu) - (\mu_{.j} - \mu)] + \varepsilon_{ijk} \\ &= \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk} \end{aligned}$$

Assumptions

$$\begin{aligned} \bullet \sum_{i=1}^I \alpha_i &= 0 & \sum_{j=1}^J \beta_j &= 0 \\ \bullet \sum_{i=1}^I (\alpha\beta)_{ij} &= 0 & \sum_{j=1}^J (\alpha\beta)_{ij} &= 0 \end{aligned}$$

Analysis of Variance

| Source | df ^a | Sum of Squares | Mean Square | F |
|------------------------------|--|---|---|--|
| Due <u>light</u> | (I-1) | $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (\bar{X}_{i..} - \bar{X}_{...})^2$ | $\frac{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (\bar{X}_{i..} - \bar{X}_{...})^2}{(I-1)}$ $= \hat{\sigma}_{light}^2$ | $F = \frac{\hat{\sigma}_{light}^2}{\hat{\sigma}^2}$ |
| Due <u>temperature</u> | (J-1) | $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (\bar{X}_{.j.} - \bar{X}_{...})^2$ | $\frac{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (\bar{X}_{.j.} - \bar{X}_{...})^2}{(J-1)}$ $= \hat{\sigma}_{temp}^2$ | $F = \frac{\hat{\sigma}_{temp}^2}{\hat{\sigma}^2}$ |
| Due <u>interaction</u> | (I-1)(J-1) | $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2$ | $\frac{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2}{(I-1)(J-1)}$ $= \hat{\sigma}_{light*temp}^2$ | $F = \frac{\hat{\sigma}_{light*temp}^2}{\hat{\sigma}^2}$ |
| Within Groups (due error) | $\sum_{i=1}^I \sum_{j=1}^J (n_{ij} - 1)$ | $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X}_{ij.})^2$ | $\frac{\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X}_{ij.})^2}{\sum_{i=1}^I \sum_{j=1}^J (n_{ij} - 1)}$ $= \hat{\sigma}^2$ | |
| Total | N - 1 | $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^{n_{ij}} (X_{ijk} - \bar{X}_{...})^2$ | | |

^a degrees of freedom

A meaningful analysis might proceed in this order.

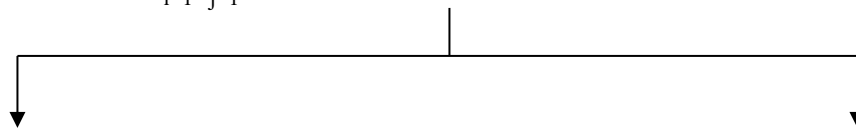
Step 1. Test for no interaction

- If there is interaction, this means that the effect of light on growth depends on the water temperature and vice versa.
- Accordingly, the meaning that can be given to an analysis of effects of light (Factor I) or an analysis of the effects of water temperature (Factor II) depend on an understanding of interaction

- The correct F statistic tests this interaction: $F = \frac{\hat{\sigma}_{\text{light*temperature}}^2}{\hat{\sigma}^2}$

- Numerator df = (I-1)(J-1).

- Denominator df = $\sum_{i=1}^I \sum_{j=1}^J (n_{ij} - 1)$



| If interaction is NOT significant | If interaction is SIGNIFICANT |
|--|---|
| <p>Step 2. Test for main effect of Factor I</p> <p>Use $F = \frac{\hat{\sigma}_{\text{factor I}}^2}{\hat{\sigma}^2}$</p> <p>Numerator df = (I-1).</p> <p>Denominator df = $\sum_{i=1}^I \sum_{j=1}^J (n_{ij} - 1)$</p> <p>Repeat for Factor II.</p> | <p>Step 2. STOP. Report report the biological variability you just discovered.</p> <p>NOTE: It is still possible to assess main effects but take care to understand its meaning in this situation!</p> <p><u>In particular</u>, because the interaction is significant, we have inferred at this point that the response to one level of a factor (eg. Factor II = water temperature) depends on the level of another factor (eg. Factor I = level of light)</p> <p>So then, if the analyst decides to collapse the data to a one-way analysis of variance, then the meaning of the analysis of a main effect is that it yields an estimate of the <u>average</u> main effect, taken over the levels of the other factor. FYI -This is typically not of interest.</p> |

Example - Continued

| Source | df ^a | Sum Squares | Mean Square | F |
|-------------------------------|--|-------------|--|---|
| Due <u>light</u> | (2-1) = 1 | 2.98 | $2.98 = \hat{\sigma}_{\text{light}}^2$ | $F = \frac{\hat{\sigma}_{\text{light}}^2}{\hat{\sigma}_{\text{error}}^2} = 10.39$ P = .018 |
| Due <u>water temp</u> | (3-1) = 2 | 3.984 | $1.992 = \hat{\sigma}_{\text{temp}}^2$ | $F = \frac{\hat{\sigma}_{\text{temp}}^2}{\hat{\sigma}_{\text{error}}^2} = 6.95$ p = .027 |
| Due <u>interaction</u> | (I-1)(J-1) = 2 | 1.268 | $0.634 = \hat{\sigma}_{\text{light*temp}}^2$ | $F = \frac{\hat{\sigma}_{\text{light*temp}}^2}{\hat{\sigma}_{\text{error}}^2} = 2.21$ p = .191 |
| Error | $\sum_{i=1}^I \sum_{j=1}^J (n_{ij} - 1) = 6$ | 1.721 | $0.2868 = \hat{\sigma}_{\text{error}}^2$ | |
| Total | N - 1 = 11 | 9.953 | | |

^a degrees of freedom

Interpretation:

Proceed in the right order:

- **STEP #1:** Assess for evidence of interaction (biology worth discovering)
In the “Due Interaction” row, find F-statistic = 2.21 on df = 2,6. The p-value is .19, suggesting that there is NO statistically significant evidence of interaction. Not surprising, given the small sample size!
- **STEP 2:** With no interaction, you can now go on to assess output for strength of main effects.
 - * Main Effect of TEMP: F=6.95 on df=2,6. p-value = .03.
Conclude statistically significant
 - * Main Effect of LIGHT: F=10.39 on df=1,6. p-value = .02.
Conclude statistically significant.

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis

R Illustration

Load data. Descriptives by group.

```
library(FSA) # Here, you'll see the advantage of using Summarize( ) in package {FSA}
load(file="fishgrowth.Rdata")
FSA::Summarize(growth ~ light + temp, data=fishgrowth, digits=2, na.rm=TRUE)

##   light      temp n mean   sd min   Q1 median   Q3  max
## 1 1=low    1=cold 2 4.39 0.22 4.24 4.32  4.39 4.47 4.55
## 2 2=high    1=cold 2 4.82 1.04 4.08 4.45  4.82 5.18 5.55
## 3 1=low    2=lukewarm 2 4.88 0.01 4.88 4.88  4.88 4.89 4.89
## 4 2=high    2=lukewarm 2 5.55 0.76 5.01 5.28  5.55 5.82 6.09
## 5 1=low     3=warm 2 5.06 0.07 5.01 5.04  5.06 5.08 5.11
## 6 2=high     3=warm 2 6.97 0.06 6.92 6.94  6.96 6.99 7.01
```

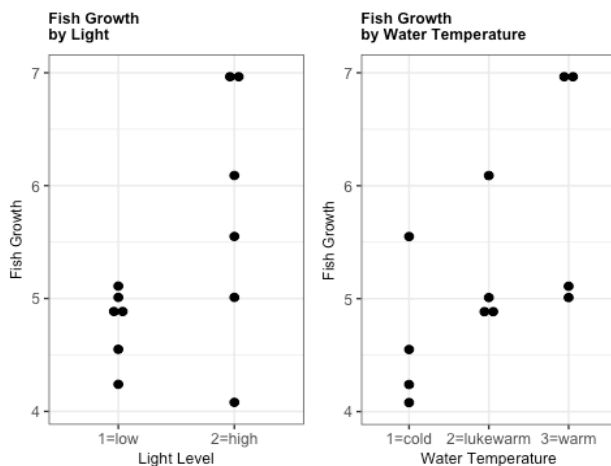
Graphs: 1) side-by-side dotplot 2) side-by-side boxplot

```
library(ggplot2)
library(gridExtra)

# panel 1a = side-by-side dot over Light
p1a <- ggplot(data=fishgrowth, aes(x=factor(light), y=growth)) +
  geom_dotplot(dotsize=0.75, binaxis = "y", stackdir = "center", binpositions="all") +
  xlab("Light Level") + ylab("Fish Growth") +
  ggtitle("Fish Growth \nby Light") +
  theme_bw() +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 9),
        plot.title = element_text(size = 9, face = "bold"))

# panel 1b = side-by-side dot over temp
p1b <- ggplot(data=fishgrowth, aes(x=factor(temp), y=growth)) +
  geom_dotplot(dotsize=0.75, binaxis = "y", stackdir = "center", binpositions="all") +
  xlab("Water Temperature") + ylab("Fish Growth") +
  ggtitle("Fish Growth \nby Water Temperature") +
  theme_bw() +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 9),
        plot.title = element_text(size = 9, face = "bold"))

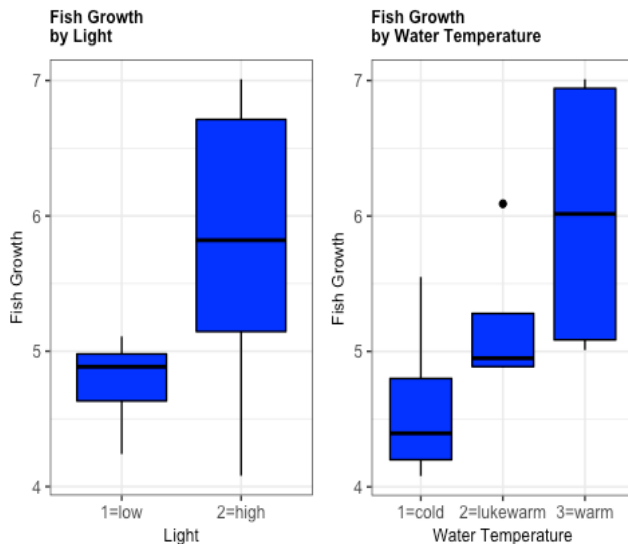
gridExtra::grid.arrange(p1a, p1b, ncol=2)
```



```
# panel 2a = side-by-side box plot over light
p2a <- ggplot(data=fishgrowth,aes(x=factor(light),y=growth)) +
  geom_boxplot(color="black",fill="blue") +
  xlab("Light") + ylab("Fish Growth") +
  ggtitle("Fish Growth \nby Light") +
  theme_bw() +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 9),
        plot.title = element_text(size = 9, face = "bold"))

# panel 2b = side-by-side box plot over temp
p2b <- ggplot(data=fishgrowth,aes(x=factor(temp),y=growth)) +
  geom_boxplot(color="black",fill="blue") +
  xlab("Water Temperature") + ylab("Fish Growth") +
  ggtitle("Fish Growth \nby Water Temperature") +
  theme_bw() +
  theme(axis.text = element_text(size = 9),
        axis.title = element_text(size = 9),
        plot.title = element_text(size = 9, face = "bold"))

gridExtra::grid.arrange(p2a, p2b, ncol=2)
```



Two-Way Anova. Fit of Model.

```
library(car)
# Fit of anova model using R command aov( ). Assign fit to object that I name aov3.
aov3 <- aov(growth ~ temp + light + temp:light, data=fishgrowth)
anova(aov3)
## Analysis of Variance Table
##
## Response: growth
##          Df Sum Sq Mean Sq F value    Pr(>F)
## temp      2  3.9843  1.99216    6.9462 0.02744 *
## light     1  2.9800  2.98003   10.3906 0.01806 *
## temp:light 2  1.2676  0.63381    2.2099 0.19093
## Residuals 6  1.7208  0.28680
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Two-Way Anova. Tests of Homogeneity of Variance

tests of homogeneity of variance

R Modified Levene Test is Brown and Forsythe Test

library(car)

car::leveneTest(growth ~ temp*light, data=fishgrowth)

Warning in anova.lm(lm(resp ~ group)): ANOVA F-tests on an essentially
perfect fit are unreliable

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 5 57841111952702865723844244537344 < 0.0000000000000022 *** Wow. Should have rounded. But forging on..

6

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Post Fit Estimation of Means and Nifty Plot

library(emmeans)

library(ggplot2)

library(gridExtra)

emm1 = emmeans::emmeans(aov3, specs = "light", by="temp")

emm1

temp = 1=cold:

light emmean SE df lower.CL upper.CL

1=low 4.39 0.379 6 3.47 5.32

2=high 4.82 0.379 6 3.89 5.74

##

temp = 2=lukewarm:

light emmean SE df lower.CL upper.CL

1=low 4.88 0.379 6 3.96 5.81

2=high 5.55 0.379 6 4.62 6.48

##

temp = 3=warm:

light emmean SE df lower.CL upper.CL

1=low 5.06 0.379 6 4.13 5.99

2=high 6.96 0.379 6 6.04 7.89

##

Confidence level used: 0.95

Estimated means wrt TEMP separately by LIGHT

emm2 = emmeans::emmeans(aov3, specs = "temp", by="light")

emm2

light = 1=low:

temp emmean SE df lower.CL upper.CL

1=cold 4.39 0.379 6 3.47 5.32

2=lukewarm 4.88 0.379 6 3.96 5.81

3=warm 5.06 0.379 6 4.13 5.99

##

light = 2=high:

temp emmean SE df lower.CL upper.CL

1=cold 4.82 0.379 6 3.89 5.74

2=lukewarm 5.55 0.379 6 4.62 6.48

3=warm 6.96 0.379 6 6.04 7.89

##

Confidence level used: 0.95

Plot - Estimated means (95% CI) wrt LIGHT separately by TEMP

p3a <- plot(emm1) + theme_bw() +

labs(x = "Estimated Mean Fish Growth (95% CI)", y = "Light") +

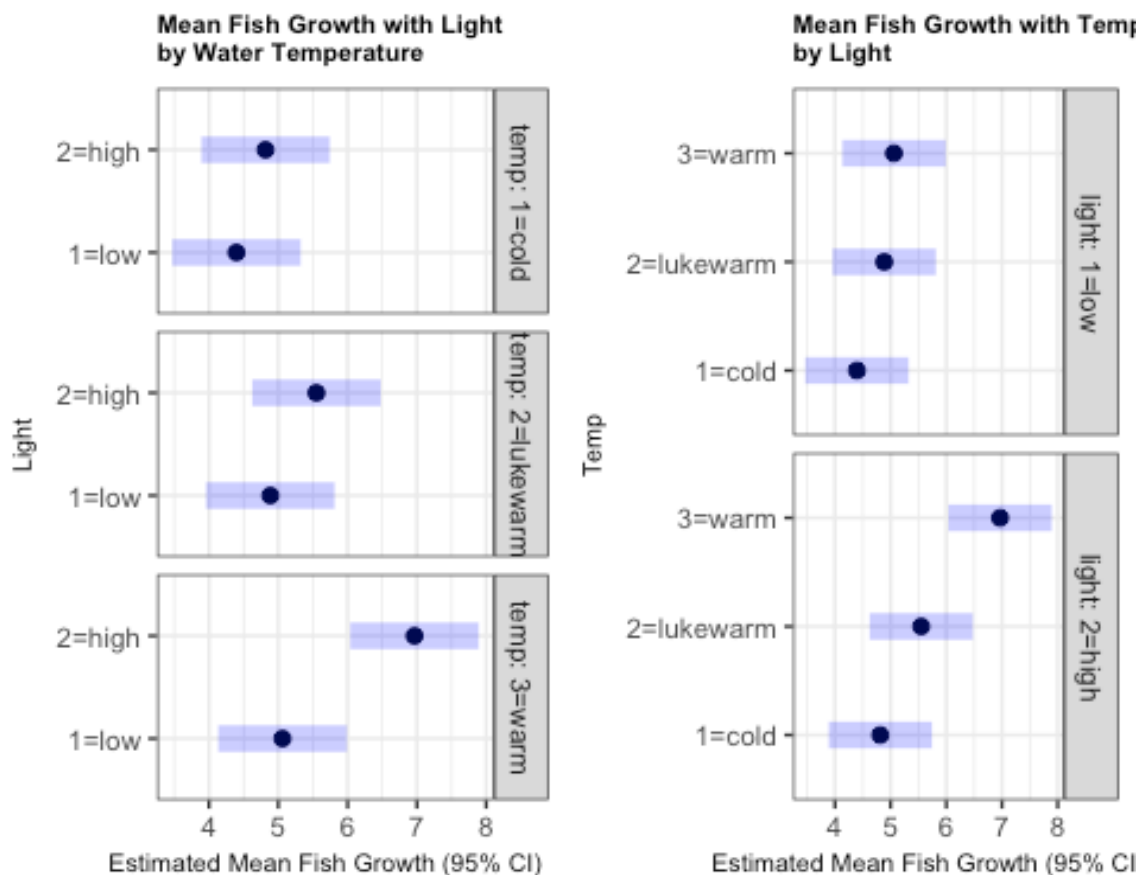
ggtitle("Estimated Mean Fish Growth with Light \nby Water Temperature") +

theme(axis.title = element_text(size = 8),

plot.title = element_text(size = 8, face = "bold"))

```
# Plot Estimated means (95% CI) wrt TEMP separately by LIGHT
p3b <- plot(emm2) + theme_bw() +
  labs(x = "Estimated Mean Fish Growth (95% CI)", y = "Temp") +
  ggtitle("Estimated Mean Fish Growth with Temp \nby Light") +
  theme(axis.title = element_text(size = 8),
        plot.title = element_text(size = 8, face = "bold"))
```

```
gridExtra::grid.arrange(p3a, p3b, ncol=2)
```

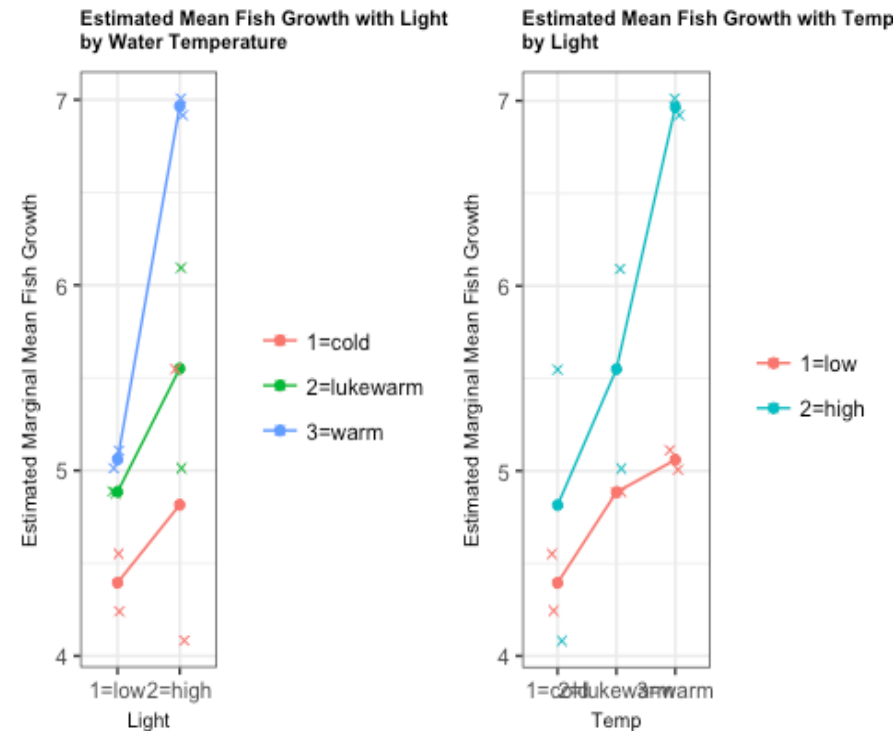


Post Fit Interaction Plots

```
library(emmeans)
library(ggplot2)
library(gridExtra)
# Interaction Plot Y=growth X=LIGHT: separately by TEMP
# emmip(FITOBJECT, STRATIFYVAR ~ XVAR)
p4a <- emmip(aov3, temp ~ light) +
  geom_jitter(aes(x = light, y = growth, colour = temp),
             data = fishgrowth, pch = 4, width = 0.1) +
  labs(y = "Estimated Marginal Mean Fish Growth", x="Light", colour = "") +
  ggtitle("Estimated Mean Fish Growth with Light \nby Water Temperature") +
  theme_bw() +
  theme(axis.title = element_text(size = 8),
        plot.title = element_text(size = 8, face = "bold"))
```

```
# Interaction Plot Y=growth X=TEMP: separately by LIGHT
# emmip(FITOBJECT, STRATIFYVAR ~ XVAR)
p4b <- emmip(aov3, light ~ temp) +
  geom_jitter(aes(x = temp, y = growth, colour = light),
    data = fishgrowth, pch = 4, width = 0.1) +
  labs(y = "Estimated Marginal Mean Fish Growth", x="Temp", colour = "") +
  ggtitle("Estimated Mean Fish Growth with Temp \nby Light") +
  theme_bw() +
  theme(axis.title = element_text(size = 8),
    plot.title = element_text(size = 8, face = "bold"))

gridExtra::grid.arrange(p4a, p4b, ncol=2)
```



Summary - The analysis of variance table tells us the following.

1. Fail to reject the hypothesis of no interaction. Thus we can test for main effects using the MSE.
2. The statistical test of the null hypothesis of no temperature effect is borderline significant.
3. The statistical test of the null hypothesis of no light effect is also borderline significant.
4. Bottom line? NOT ENOUGH SAMPLE SIZE TO DO MUCH HERE...

c. The Two Way Hierarchical (Nested) Design

Example -

An investigator wishes compare 3 sprays applied to leaves on trees. Each treatment is applied to 6 leaves of 4 trees. Thus, the total number of observations is $(3 \text{ treatments})(4 \text{ trees})(6 \text{ leaves}) = 72$

In an example such as this, sampling is done in multiple stages. Here – (1) In stage 1, a random sample of trees is selected (thus “tree” is the primary sampling unit) and (2) in stage 2, within each tree, a random sample of leaves is selected for measurement.

Outcome X = Nitrogen concentration in the leaf

Group variable is tree

Primary sampling unit is leaf (and leaf is nested within tree)

Following are the data

| | | Tree (nested within spray) | | | |
|-------|---|----------------------------|-------|-------|-------|
| | | 1 | 2 | 3 | 4 |
| Spray | 1 | 4.50 | 5.78 | 13.32 | 11.59 |
| | | 7.04 | 7.69 | 15.05 | 8.96 |
| | | 4.98 | 12.68 | 12.67 | 10.95 |
| | | 5.48 | 5.89 | 12.42 | 9.87 |
| | | 6.54 | 4.07 | 10.03 | 10.48 |
| | | 7.20 | 4.08 | 13.50 | 12.79 |

| | | Tree (nested within spray) – a different set of 4 trees! | | | |
|-------|---|--|-------|-------|-------|
| | | 1 | 2 | 3 | 4 |
| Spray | 2 | 15.32 | 14.53 | 10.89 | 15.12 |
| | | 14.97 | 14.51 | 10.27 | 13.79 |
| | | 14.81 | 12.61 | 12.21 | 15.32 |
| | | 14.26 | 16.13 | 12.77 | 11.95 |
| | | 15.88 | 13.65 | 10.45 | 12.56 |
| | | 16.01 | 14.78 | 11.44 | 15.31 |

| | | Tree (nested within spray) - a different set of 4 trees! | | | |
|-------|---|--|------|------|------|
| | | 1 | 2 | 3 | 4 |
| Spray | 3 | 7.18 | 6.70 | 5.94 | 4.08 |
| | | 7.98 | 8.28 | 5.78 | 5.46 |
| | | 5.51 | 6.99 | 7.59 | 5.40 |
| | | 7.48 | 6.40 | 7.21 | 6.85 |
| | | 7.55 | 4.96 | 6.12 | 7.74 |
| | | 5.64 | 7.03 | 7.13 | 6.81 |

In this two-way hierarchical (nested) design,

I = # treatments.

In this example, $I=3$

J = # primary sampling units

In this example, $J=4$

n = # secondary sampling units, nested

In this example, $n=6$

- Only one factor is of interest – Treatment (spray) at 3 levels.
- The effect of **tree** is **random** and is *not* of interest.
- Similarly, the effect of **leaf** is **random** and is *not* of interest.

The Two Way Hierarchical (Nested) Design Model

Setting:

I groups or treatments indexed $i=1, 2, \dots, I$

J primary sampling units nested within each treatments and indexed $j=1, 2, \dots, J$

Sample size is n in each treatment x block combination; these are *secondary sampling* units

k indexes the secondary sampling units and are indexed $k=1, 2, \dots, n$

X_{ijk} = Observation for the k^{th} secondary sampling unit of the j^{th} primary sampling unit in the i^{th} group/treatment

The two way hierarchical (nested) design model of X_{ijk} is defined as follows:

$$X_{ijk} = \mu + \alpha_i + b_{(i)j} + \varepsilon_{(ij)k}$$

where

μ = grand mean

$$\alpha_i = [\mu_i - \mu] \text{ and } \sum_{i=1}^I \alpha_i = 0$$

and

$b_{(ij)}$ is random and distributed $\text{Normal}(0, \sigma_b^2)$

ε_{ij} is random error distributed $\text{Normal}(0, \sigma^2)$

$b_{(ij)}$ and $\varepsilon_{(ij)k}$ are mutually independent

Two Way Hierarchical (Nested) Analysis of Variance

The nested model, *for the reason of having random effects*, looks a little different from a fixed effects model.

$$\begin{aligned}\mu_i &= \text{Mean [Outcome] for spray "i"} \\ &= \mu + \alpha_i\end{aligned}$$

$$X_{ijk} = \mu + \alpha_i + b_{(i)j} + \varepsilon_{(ij)k} \quad \text{where}$$

- The parenthesis notation “(i)j” tells us that tree “j” is nested in spray “i”
- The parenthesis notation “(ij)k” tells us that leaf “k” is nested in the “jth” tree receiving spray = “i”

$$\alpha_i = [\mu_i - \mu] \quad \text{and} \quad \sum_{i=1}^I \alpha_i = 0$$

$$X_{ijk} = \bar{X}_{..} + [\bar{X}_{i..} - \bar{X}_{...}] + [\bar{X}_{ij.} - \bar{X}_{i..}] + [X_{ijk} - \bar{X}_{ij.}] \quad \text{algebraic identity}$$

Assumptions

The $b_{(i)j}$ are independent and distributed Normal(0, σ_b^2)

The $\varepsilon_{(ij)k}$ are independent and distributed Normal(0, σ_e^2)

The $b_{(i)j}$ and $\varepsilon_{(ij)k}$ are mutually independent

Total SSQ and its Partitioning

$$X_{ijk} = \bar{X}_{..} + [\bar{X}_{i..} - \bar{X}_{...}] + [\bar{X}_{ij.} - \bar{X}_{i..}] + [X_{ijk} - \bar{X}_{ij.}] \rightarrow$$

$$[X_{ijk} - \bar{X}_{..}] = [\bar{X}_{i..} - \bar{X}_{...}] + [\bar{X}_{ij.} - \bar{X}_{i..}] + [X_{ijk} - \bar{X}_{ij.}].$$

Squaring both sides and summing over all observations yields (because the cross product terms sum to zero!)

$$\begin{aligned} & \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^n [X_{ijk} - \bar{X}_{...}]^2 \\ &= \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^n [\bar{X}_{i..} - \bar{X}_{...}]^2 + \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^n [\bar{X}_{ij.} - \bar{X}_{i..}]^2 + \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^n [X_{ijk} - \bar{X}_{ij.}]^2 \\ &= Jn \sum_{i=1}^I [\bar{X}_{i..} - \bar{X}_{...}]^2 + n \sum_{i=1}^I \sum_{j=1}^J [\bar{X}_{ij.} - \bar{X}_{i..}]^2 + \sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^n [X_{ijk} - \bar{X}_{ij.}]^2 \end{aligned}$$

Analysis of Variance Table

| Source | df ^a | Sum of Squares | E (Mean Square) | F |
|--------------------------------|-----------------|--|--|--|
| Due treatment | (I-1) | $Jn \sum_{i=1}^I (\bar{X}_{i..} - \bar{X}_{...})^2$ | $\sigma_e^2 + n\sigma_b^2 + Jn \left[\frac{\sum_{i=1}^I \alpha_i^2}{(I-1)} \right]$ | $F = \frac{MSQ_{\text{treatment}}}{MSQ_{\text{within treatment among samples}}}$ df = (I-1), I(J-1) |
| Within treatment Among samples | I(J-1) | $n \sum_{i=1}^I \sum_{j=1}^J (\bar{X}_{ij.} - \bar{X}_{i..})^2$ | $\sigma_e^2 + n\sigma_b^2$ | $F = \frac{MSQ_{\text{within treatment among samples}}}{MSQ_{\text{residual}}}$ df = (J-1), IJ(n-1) |
| Residual | IJ(n-1) | $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^n (\bar{X}_{ijk} - \bar{X}_{ij.})^2$ | σ_e^2 | |
| Total | IJn - 1 | $\sum_{i=1}^I \sum_{j=1}^J \sum_{k=1}^n (X_{ijk} - \bar{X}_{...})^2$ | | |

^a degrees of freedom

Note: The correct F test for treatment has in the denominator the mean square for within treatment among samples. This can be appreciated as the correct definition by looking at the expected mean squares.