STA305 Design and Analysis of Experiments Lecture Notes (2021 Summer)

Jiaqi Bi, University of Toronto

August 1, 2021

Preface

This typesetting note is intended for academic communication and not-for-profit knowledge sharing. Any institutions or individuals **SHOULD NOT** reproduce or use this note in any form for operational or commercial purposes. This note contains some of the lecture notes instructed by Prof. Ramya Thinniyam and personal research notes. The author holds the right of the prosecution for any above behaviors. Moreover, this note should not be used as a substitution for the current user's attending lecture or textbook. I strongly recommend the user combine the note with your textbook and lecture to understand the content thoroughly. I hope anyone using this note can have a better understanding of the Design and Analysis of Experiments topic. Students using this copy should not conduct any academic infractions, including but not limited to reproducing, failure of citations, and using this copy in any forms of academic dishonesty mentioned by the University of Toronto. For specific details of academic integrity, please visit https://www.academicintegrity.utoronto.ca/.
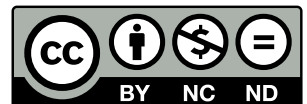
Contents

1 Lec 1 Introduction to Experiments and Types of Data

1.1 General Outline of an Experiment

- An experiment is used to prove a scientific claim

- Simplest Setup: 2 groups

  - Apply a different Treatment to each group, i.e. Treatment and Control
  - Measure a variable of interest, i.e. Response

- Ideal situation: the groups are identical in every possible way, except for the treatment

- If the groups differ in the response, it must be due to the treatment variable.

A flowchart that generally explain this:

```
                    ┌──────────────┐
                    │   2 Groups   │
                    └──────────────┘
               ↙                        ↘
    ┌──────────────┐            ┌──────────────┐
    │  Treatment   │            │   Control    │
    └──────────────┘            └──────────────┘
            ↓                            ↓
    ┌──────────────┐            ┌──────────────┐
    │  Response 1  │            │  Response 2  │
    └──────────────┘            └──────────────┘
               ↘                        ↙
                  ┌──────────────────┐
                  │  Responses defer │
                  │  due to the treat-│
                  │  ment variable   │
                  └──────────────────┘
```
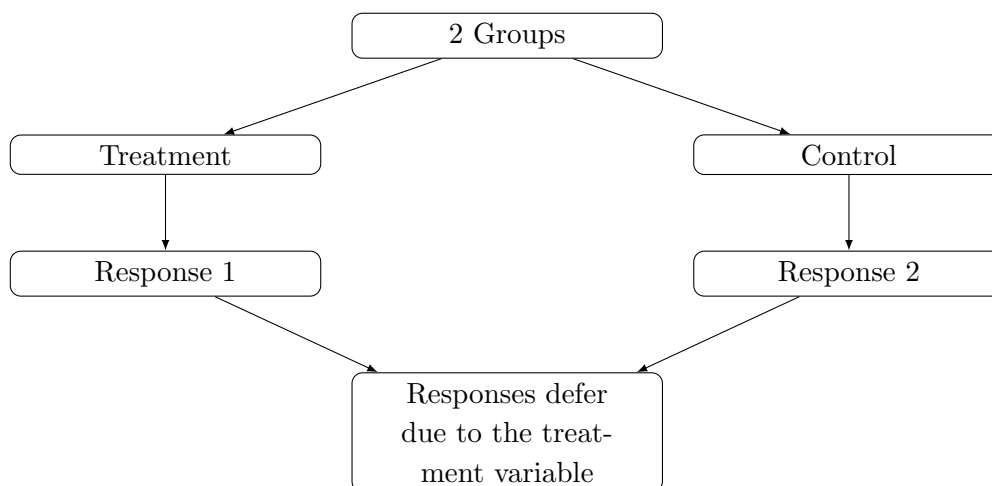
Example of different treatments: Medicine & Placebo

Example 1.1. A research team wants to test their effect of a new medicine on the Migraine. They have the treatment group of New Pill, and they have the control group of Placebo. Then they will test migraine sufferers. The response will be the pain level from 1 to 10 subjectively measured by patients themselves. The duration will be 2 hours after taking the pill. In this experiment, the nuisance variables are: Age, Sleep hours, Diet, Severity of migraines... The treatment has two levels: pill and placebo

1.2 Steps of an Experimental Design

This would be useful for Final Project
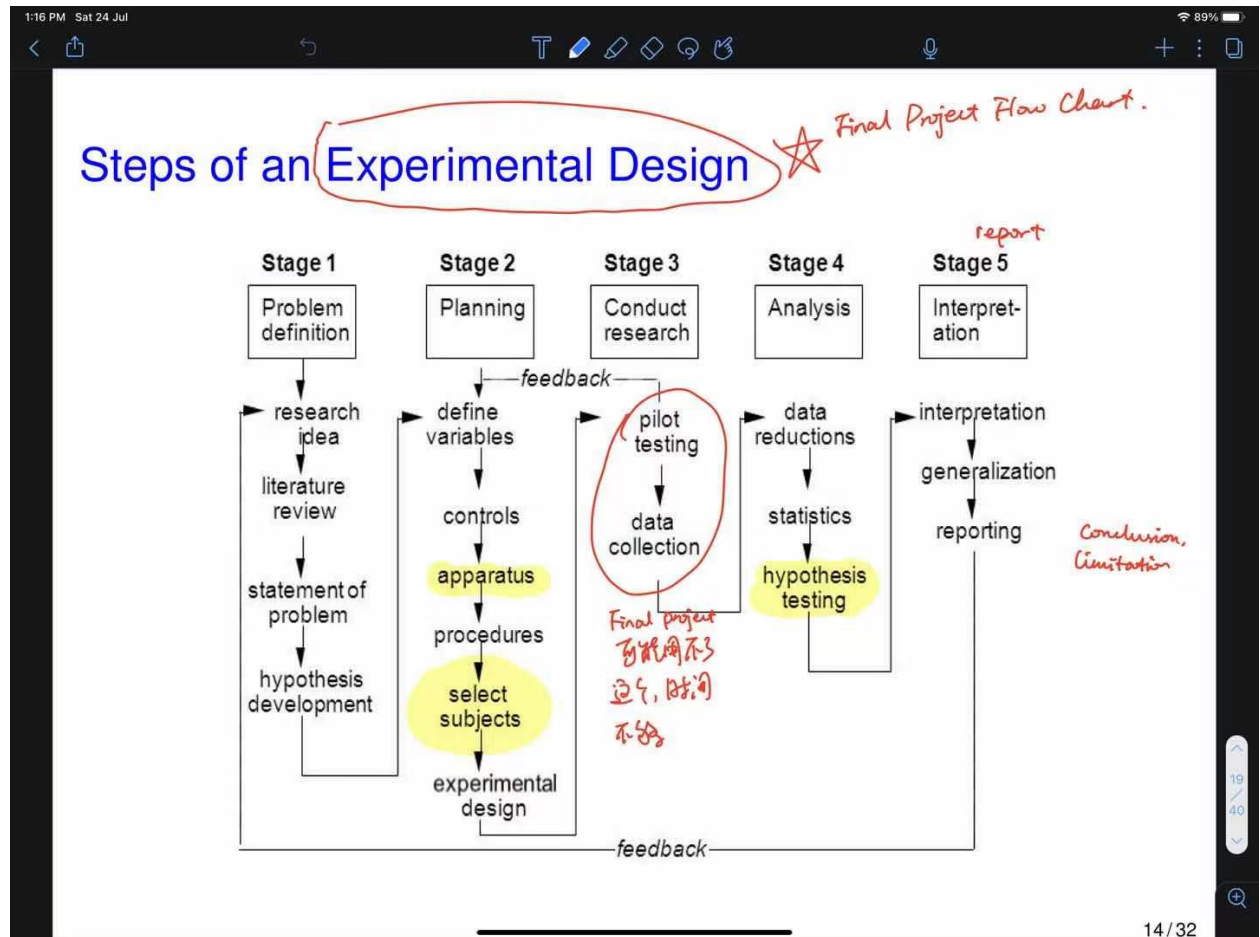
1.3 Definitions used in Statistical Inference

- Population: Set of units that we are interested in studying. E.g. Group of people, objects...

- Experimental Unit: The person or object on which the treatment is applied. Also called "case" or "Element" or "Subject" (When human unit).

- Sample: Subset of the population.

- Variable: A measured characteristic of a population unit. E.g., Age, weight, pain scale...

- Statistical Inference: Estimate, prediction, or generalization about population based information from a sample.

6

1.4 Experimental Study vs. Observational Study

Note this course we mainly focus on the experimental study. If you are interested in observational study you are recommended to take STA304.

Experimental Study

- To assess the effect of a certain treatment or condition

- First randomly assign subjects to treatment, then control the rest. Randomization is important. Experimental data comes from experiments, so there are eligible reasons to conclude causation.

- Common in scientific and psychological research

- Between Subjects Design: Each experimental unit is assigned only one treatment

- Within Subjects Design: Each experimental unit is given all treatments

Observational Study

- No randomization or controlling

- Data is collected as it comes (survey data), results from observational studies.

- Causation cannot be concluded from observational data since there may be other "confounding/lurking variables" (that were not controlled for) that is causing the relationship.

- Common in data mining, economic, and sociological research...

Note that an experiment may be hard to conduct due to ethical reasons, costs...

1.5 Types of Variables

There are two types of variables, which are X and Y such that Y denotes the response, that is measured by the researcher, and influenced by other variables. X is known as predictor or explanatory variable that affect the response variable, which is manipulated by researchers.

Variables can be:

- Quantitative: A variable is quantitative if it takes on numerical values for which arithmetic operations make sense. Ex: Height, marks, incomes... This type of variables can be discrete or continuous.

- Discrete: Variable can take on any one of a finite or countable list of values ($\mathbb{Z}$) (there are gaps in the possible values). Ex: Number of heads in 10 tosses of a coin.

- Continuous: Variable can take on any value possible in an interval (uncountable). Ex: Time to complete a marathon... (There are infinitely many numbers in $\mathbb{R}$)

- Qualitative/Categorical variable: Categories, cannot be measured on numerical scales. Ex: Gender, smoking status (No, Used to, Yes)...

Variables can be measured as either qualitative or quantitative and further categorized as follows:
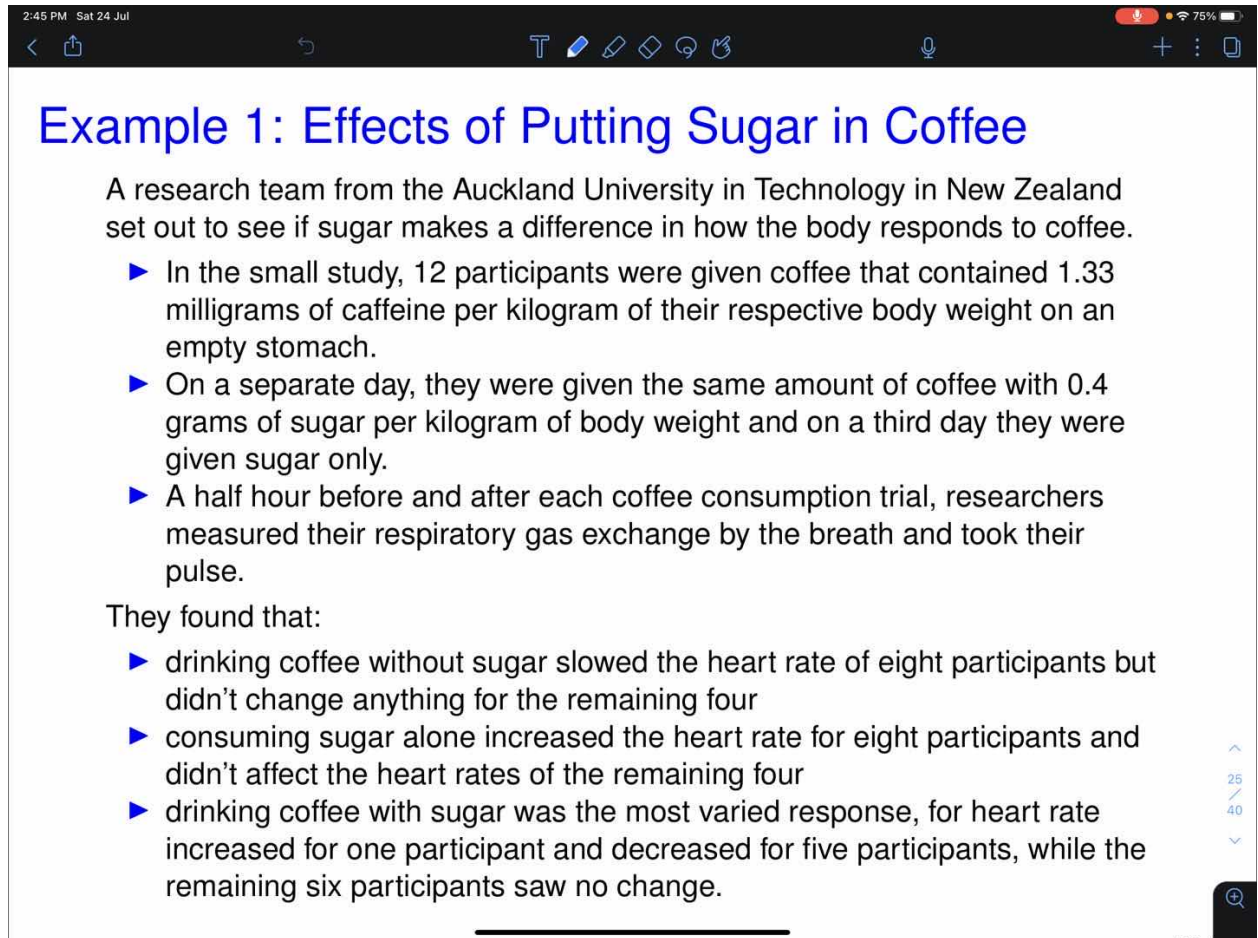
Categorical:

1. Nominal: Categorize units into distinct classes.

   - Unordered categories
   - Numerical computations are not applicable
   - Ex: Gender, POSt, Favorite colors...
   - When controlled by researcher, called Factor

2. Ordinal: Ordered categories without natural units/distance metric

   - Natural ordering to the categories; not just the names of the categories differ
   - Professional rank, grades, satisfactions...

Quantitative:

3. Interval: Numerical measurements which allow for degree of difference between values.

   - Distance is consistent but ratios are meaningless
   - Does not have a true 0 measure (0 is arbitrary)
   - Ex: Temperature ($^\circ$C), dates...

4. Ratio: Numerical measurements on which a unique and non-arbitrary zero value exists

   - Ratios are meaningful, sensible to carry our multiplication/division
   - Ex: Temperature (K), length, time duration...

When an experiment has controls, call X a factor and its categories the factor levels. Single-Factor Experiment has one factor with levels $\{a_1, a_2, a_3, ...\}$. Two-Factor Experiment has Two Factors (Factor A and B) with their own levels. A treatment is a combination of factor levels.

Example 1.2. Example of Effects of Putting Sugar in Coffee

Example 1: Effects of Putting Sugar in Coffee

A research team from the Auckland University in Technology in New Zealand set out to see if sugar makes a difference in how the body responds to coffee.

- ▶ In the small study, 12 participants were given coffee that contained 1.33 milligrams of caffeine per kilogram of their respective body weight on an empty stomach.
- ▶ On a separate day, they were given the same amount of coffee with 0.4 grams of sugar per kilogram of body weight and on a third day they were given sugar only.
- ▶ A half hour before and after each coffee consumption trial, researchers measured their respiratory gas exchange by the breath and took their pulse.

They found that:

- ▶ drinking coffee without sugar slowed the heart rate of eight participants but didn't change anything for the remaining four
- ▶ consuming sugar alone increased the heart rate for eight participants and didn't affect the heart rates of the remaining four
- ▶ drinking coffee with sugar was the most varied response, for heart rate increased for one participant and decreased for five participants, while the remaining six participants saw no change.

1. For the above study, identify the following:

   - Population- All coffee drinkers
   - Experimental unit- A human participant
   - Variables (Which is response/predictor, quantitative/qualitative?)- X: Drink with sugar or not, Y: Respiratory gas exchange and their pulse (Quantitative)
   - Sample- 12 participants
   - A possible Statistical Inference- Generalization on effects of drinking coffee with sugar

2. Is the study observational or experimental? Justify. The data were collected by an experiment, it has controls of giving specific amount of coffee and sugar. It is a within subjects design. Treatments are assigned.

3. What are the factors and levels? Describe the treatments and how many there are? Factor: Type of drink. Factor Levels (3 levels/treatments): Coffee, Coffee with sugar, sugar only) (categorical)

9

4. Name a confounding variable that could be present in this study. The order of treatments applied

Example 1.3. Effects of Teaching Style Suppose you wish to conduct an experiment to see if teaching style (lecturer, facilitator, online technology) and the amount of lecture time (2 hours, 3 hours/week) affects the learning outcome of students in a large university course.

1. What are the factors and levels? Teaching styles (Lecture, Facilitator, Online Tech); The amount of lecture time (2 hours, 3 hours)

2. Treatments? (Lecture & 2 hours, Lecture & 3 hours, Facilitator & 2 hours, Facilitator & 3 hours, Tech & 2 hours, Tech & 3 hours) Total 6 treatments

1.6 Crossed vs. Nested Factors

When all combinations of factors are possible, the experiment is said to be fully crossed. Easier to analyze.

When each level of one factor occurs with a unique set of levels of the other factor, the design is called nested. Hierarchical Design.
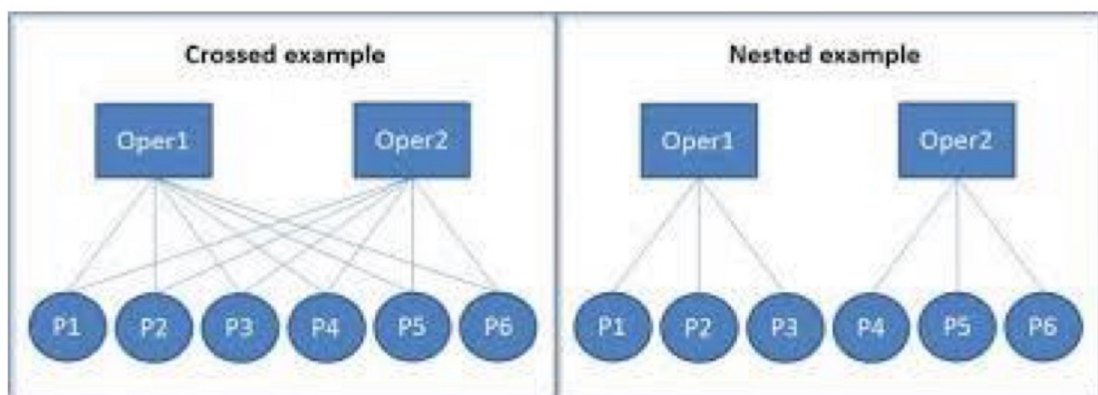
1.7 Nuisance Variables

Variables other than the treatment condition that influence the response variable.

Ex: Teaching Styles:

- Different lecturers for each section

- Age of students

- Time of day for lecture section, etc...

We use Control, Blocking, Randomization, Replication to deal with nuisance variables.

Control: Keep the nuisance variable constant throughout all treatment conditions (no longer a variable)

Ex: Teaching Styles: Same lecturer for each style; Same/Similar lecture times for each style...

Blocking: If the nuisance variable cannot be controlled, but can be observed, use Blocking to ensure each treatment has an equal amount of the variable.

Ex: Assign students to styles so each style has same age distribution.

18 and under , 19-24 , mature
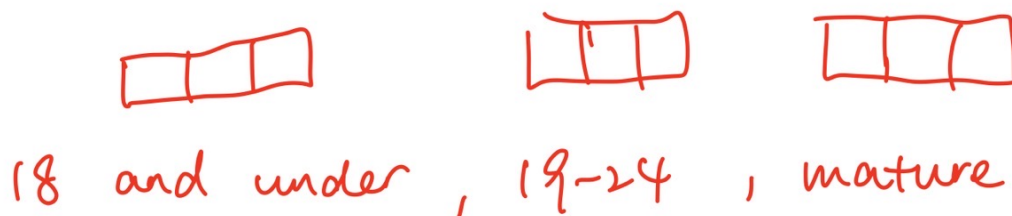
Randomization: If the nuisance variable cannot be controlled or observed, use Randomization to spread out the variables and reduce the chance of confounding.

Ex: Teaching Styles: After blocking, randomly assign students to styles/sections.

Replication: Should not conduct the experiment with only one observation in each treatment. The resulting effect may be due to the experimental unit, and not the treatment. Cannot estimate variance within a specific treatment if it has only one observation. Use of replicates allows us to estimate experimental error, which is used to determine if we have statistically significant results.

Replication can occur at two different levels:

- **Treatment Level**: We take more than one observation in each treatment.

- **Experiment Level**: We wish to replicate the entire experiment to ensure results were not due other features inherent in the experiment, and so we can generalize to other populations.

Ex: Teaching Styles: Many students with similar characteristics (treatments) in each class; Another university could replicate the study.

1.8   Confounding

Nuisance variables that are systematically related to the treatments. Confounding variables can alter the effect of treatments even if non-systematic relationship to treatments, accidental effects can increase variability of responses and mask treatment effects.

Ex: Teaching Styles: Certain styles may be taught at different times of the day. For example, tech section always taught in night and lecture style during day. Differences cannot be uniquely credited to lecture styles.

Randomization decreases the chance that factors not accounted for in the design of the experiment will be confounded with the treatments.

1.9   Blinding

We use this when participants may be influenced once they know their treatment. **Placebo**.

Double blind: Both subjects and researcher are unaware of treatment assignment.

1.10   Balance

Assign the same number of experimental units to each treatment, call it Balanced Design. Balanced and unbalanced designs dealt with differently (calculations, etc...). Balanced design can become unbalanced easily- missing data, destroyed measurements, etc...

2 Lec 2 Review of Hypothesis Testing and One-Way ANOVA with 2 Levels

2.1 Single Factor Analysis with 2 levels

- 1 Factor with 2 levels (2 different treatments, treatment/control, etc.)

- 2 treatments in total

- Want to determine if there is a "treatment effect" (if response varies by treatment)

- Test for differences in means of the 2 populations: sample means are not enough; account for variability within each treatment group

2.2 Review of Hypothesis Testing

Definition 2.1. A Hypothesis Test is a formal statistical test that is performed to decide whether a statement about a set of parameter(s) is reasonable (to make an inference about the value of a parameter and how it relates to a specified/hypothesized numerical value).

Elements of a Hypothesis Test

1. H_0 is Null Hypothesis and H_a is Alternative Hypothesis.

2. Test Statistic: A statistic (function of the data) that involves the parameter value and has a known distribution under H_0.

3. Distribution under H_0: Distribution of test statistic assuming that H_0 is true

4. P-value: Probability of observing a test statistic as or more extreme (more contradictory to H_0 i.e. favorable to H_a) than already observed if the null hypothesis is true. Gives a measure of strength of evidence against H_0. Small P-value, test statistic value is unlikely if H_0 is true (contradiction). Large p-value, test statistic value is likely if H_0 is true (No contradiction)

5. Conclusion: Reject H_0 (and favour H_a) or Fail to Reject H_0. This leads to a practical conclusion about the population(s)/parameter value(s). When H_0 is rejected, we say the test is statistically significant, or there is a statistically significant difference.

2.3 Review: Two Sample T-Tests

Definition 2.2. Interested in comparing two population means when we have small sample sizes. Suppose we have $X_1, ..., X_{n1}$ a random sample from population 1 and $Y_1, ..., Y_{n2}$ a random sample from population 2.

H_0 is $\mu_1 - \mu_2 = D_0$, note that this D_0 is usually 0 but not always (When treatment means are equal). We have $H_a : \mu_1 - \mu_2 \neq D_0$ OR $\underbrace{Ha : \mu_1 - \mu_2 < D_0 \text{ OR } H_a : \mu_1 - \mu_2 > D_0}_{\text{One-tail test}}$

Assumptions:

1. Two samples are *iid* from Normal populations

2. Two samples are independent from each other

The test statistic is

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{se(\bar{x} - \bar{y})}$$

where *se* is the standard error: estimated standard deviation.

2.4 Pooled T-Test

Definition 2.3. Applicable when population variances can be assumed to be equal. I.e. Population 1: $X_1, ..., X_{n1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$ and Population 2: $Y_1, ..., Y_{n2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$ such that $\sigma_1 = \sigma_2$.

Assumptions:

1. Two samples are random samples from the target populations and the samples are independent from each other.

2. The populations have equal standard deviations ($\sigma_1 = \sigma_2$)

3. Two samples are (approximately) Normal

If we can assume that the variances are equal, we use the pooled estimate of the variance:

$$s_p^2 = \frac{\overbrace{(n_1 - 1)}^{\text{df of } s_1^2} s_1^2 + \overbrace{(n_2 - 1)}^{\text{df of } s_2^2} s_2^2}{\underbrace{n_1 + n_2 - 2}_{\text{total df}}}$$

Hypothesis for Pooled T-Tests:

$H_0 : \mu_1 - \mu_2 = D_0$ vs.

1. $H_a : \mu_1 - \mu_2 \neq D_0$ OR

2. $H_a : \mu_1 - \mu_2 < D_0$ OR

3. $H_a : \mu_1 - \mu_2 > D_0$.

Test Statistic for Pooled T-Tests:

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim t_{n_1 + n_2 - 2} \text{ under } H_0$$

1. $p = 2P(t_{n_1 + n_2 - 2} > |t|)$ OR

2. $p = P(t_{n_1 + n_2 - 2} < t)$ OR

3. $p = P(t_{n_1 + n_2 - 2} > t)$

Conclusion: Reject H_0 is $p < \alpha$ and failed to reject otherwise.

2.5 Welch's T-Test Using Satterthwaite Approximation

Definition 2.4. Used when population variances are not equal or inconclusive evidence about equality of variances.

Assumptions: Two samples are random from the target populations and the samples are independent from each other; Two samples are Normal.

Test Statistic for Welch's T-Test

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\mu \text{ under } H_0$$

where

$$\mu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

The degree of freedom is calculated by Satterthwaite Approximation. Note that μ may not be an integer, round down to the nearest integer.

Example 2.5. Example of Arthritis Clinical Study

There are 293 patients who suffering from RA, they are firstly measured of the initial pain. Then 146 are taking Auranofin and 147 are taking Placebo. After 6 months, we measure their pains and compare. There is a repeated measurement which is before pain and after pain. This is a between subjects design. The factor of the experiment is treatment (medicine). It has 2 levels that are Auranofin and Placebo. The responses are Pain before and Pain after 6 months. The researcher will compare the difference. μ_A = mean difference in pain (before-after) for Auranofin users. μ_P = mean difference in pain (before-after) for placebo users.

Question: Does the Auranofin therapy improve the pain? i.e., Is $\mu_A > \mu_P$? Note that μ is a mean difference, not the pain level!

The R Output is following:

R Code

```
# Read in csv file
> arthdata <- read.csv("arthritisdata.csv")
> arthdata
      Treatment Baseline Month6
1     TAuranofin       3      1
2        Placebo       2      2
3     TAuranofin       3      2
.
.
.
292      Placebo       3      4
293  TAuranofin       3      4

# Attach data to R's search path
> attach(arthdata)

# Create Response Variable: Difference in Pain
> DiffinPain <- Baseline-Month6
> DiffinPain
  [1]  2  0  1  0  0  0 -1 -1 -1   . . .

# Difference in Pain for each treatment
> DiffinPainA <- DiffinPain[Treatment=="TAuranofin"]
> DiffinPainP <- DiffinPain[Treatment=="Placebo"]

# Sample sizes, means, and variances by treatment group

> tapply(DiffinPain, Treatment, length)
TAuranofin   Placebo
       146       147

> tapply(DiffinPain, Treatment, mean)
TAuranofin   Placebo
0.6575342 0.3401361

> tapply(DiffinPain, Treatment, var)
TAuranofin   Placebo
1.1784601 0.9383096
```

(handwritten annotations): group by > tapply(variable, group, stat) ; $\leftarrow$ n_A n_P ; $\bar{y}_A$ $\bar{y}_P$; S_A^2 S_P^2

Boxplots

```
# Side by Side boxplots by treatment
> boxplot(DiffinPain ~ Treatment,
xlab="Treatment", ylab="Difference in Pain")
```
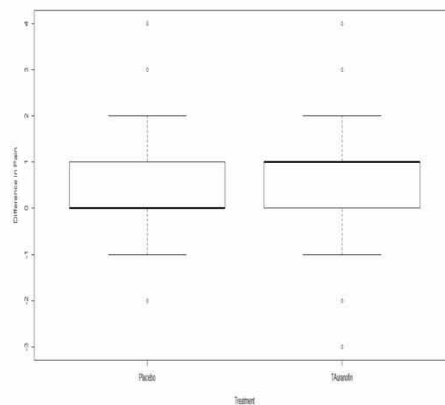
R Code: Pooled t-test

$\sigma_A^2 = \sigma_P^2$

① `> t.test(DiffinPainA,DiffinPainP, var.equal=TRUE)`

```
# OR (same output as above)
```
response ~ factor

② `> t.test(DiffinPain ~ Treatment, var.equal=TRUE)`

alternative = "greater"

```
        Two Sample t-test

data:  DiffinPainA and DiffinPainP

t = 2.641, df = 291, p-value = 0.008713

alternative hypothesis: true difference in means
is not equal to 0
```
two sided

```
95 percent confidence interval:
 0.08086399 0.55393239
```
for $\mu_A - \mu_P$

```
sample estimates:
mean of x mean of y
0.6575342 0.3401361
```
$\bar{y}_A$ $\bar{y}_P$

Using hand calculation:

Pooled T-Test: $H_0 : \mu_A = \mu_P$ vs. $H_a : \mu_A > \mu_P$, we have data:

Treatment	n	$\bar{y}$	s^2
Auranofin	146	0.6575	1.1785
Placebo	147	0.3401	0.9383

With the data above, we have

$$s^2_{\text{pooled}} = \frac{145(1.1785) + 146(0.9383)}{293 - 2} \doteq 1.0660$$

$$t = \frac{(\bar{y}_A - \bar{y}_P) - 0}{\sqrt{s^2_{\text{pooled}}(\frac{1}{n_A} + \frac{1}{n_P})}} = \frac{0.6575 - 0.3401}{\sqrt{1.0660(\frac{1}{146} + \frac{1}{147})}} \doteq 2.64 \sim t_{291} \text{ under } H_0$$

By using T-Table: $p = P(t_{291} > 2.64) = 0.0041 < 0.05$.

19

Since $p < 0.05$, we reject H_0. Strong evidence to conclude that Auranofin therapy helps improve pains.

Using R, we can get the exact p-value that the 2 sided $p = 0.0087$, so the exact p value will be $p = P(t_{291} > 2.64) = \frac{0.0087}{2} = 0.00435$. The conclusion still holds.

2.6   Analysis of Variance (ANOVA)

Definition 2.6. Analysis of Variance (ANOVA) is a collection of statistical models and procedures for comparing factor level means in a factor. We wish to test if there is a statistically significant difference between the factor level/group means.

One-Way ANOVA: One factor; Two-Way ANOVA: Two factors, etc.

It's called "Analysis of Variance" because we compare the **within-group** variance to the **between-group** variance.

Main Idea: If the variation between groups is significantly bigger than the variation within groups, there is a statistically significant difference in the group means.

Note: An ANOVA is equivalent to using a linear regression model with categorical predictors (with indicator variables).

2.7   Recall: Multiple Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}... + \beta_{p-1} X_{p-1,i} + \epsilon_i \text{ for } i = 1, 2, ..., n$$

where

- Y_i: Response for the ith case (Quantitative variable)

- $X_{1,i}...$: Predictors for ith case (Quantitative or categorical)

- ϵ_i: Error term for the ith case, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

- $\beta_0, \beta_1, ..., \beta_{p-1}$: Regression coefficients/parameters, β_0: Intercept

- n: Number of cases/sample size

Matrix form: $Y = X\beta + \epsilon$; where

- Y is an $(n \times 1)$ vector

- X is $(n \times p)$ matrix, each column contains values for each predictor. If model uses intercept, first column of X is all 1s.

- β is a $(p \times 1)$ vector of regression coefficients

- ϵ is a $(n \times 1)$ vector of error terms

Example 2.7. $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$, $X = \begin{pmatrix} 1 & x_{11} & x_{21} & \dots & x_{p-1\ 1} \\ 1 & x_{12} & x_{22} & \dots & x_{p-1\ 2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{p-1\ n} \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$, $\epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$

2.8 Indicator Variables

Definition 2.8. A factor has a levels, when using a model with intercept, use only $(a-1)$ indicators. The level without an indicator is the default or reference level. For $k = 1, 2, ..., a-1$

$$I_{k,i} = \begin{cases} 1, & \text{if } i\text{th case belongs in factor level } k \\ 0, & \text{otherwise} \end{cases}$$

Example 2.9. Linear Regression Model for Arthritis Example

$$Y_i = \beta_0 + \beta_1 I_{A,i} + \epsilon_i \text{ for } i = 1, 2, ..., n$$

where $Y_i =$ difference in pain after 6 months for ith patient.

$$E(Y_i) = \begin{cases} \beta_0 + \beta_1, & \text{if } i\text{th patient received Auranofin (1)} \\ \beta_0, & \text{if } i\text{th patient received placebo (0)} \end{cases}$$

Therefore, β_0 is the mean difference in pain for placebo group, β_1 is the difference in the mean difference of pain (response) between Auranofin and Placebo groups.

$\beta_1 = 0$: No difference

$\beta_1 > 0$: Mean in Auranofin group > Placebo Group

$\beta_1 < 0$: Mean in Auranofin Group < Placebo Group

2.9 One-Way ANOVA with 2 levels

When there are 2 factor levels (comparing 2 means like in Arthritis Example), use t-test to test if $\beta_1 = 0$. Alternatively, use F-Test: Can be used to compare many factor level means. T-Test using linear regression is equivalent to F-Test using ANOVA when comparing 2 means. $t^2 = F$.

Example 2.10. Example of Arthritis Clinical Study- Linear Regression Model

Hand Calculation:

We have $H_0 = \beta_1 = 0$, $H_a = \beta_1 > 0$, according to R output we have $t = 2.64 \sim t_{291}$ under H_0.
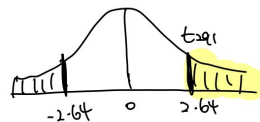
$p = P(t_{291} > 2.64) = \frac{0.00871}{2} \doteq 0.0044 < 0.05$. So, we have strong evidence to conclude that the Auranofin is effective in reducing pain.

2.10 Variation and Sum of Squares

When we apply a treatment, we consider a possible treatment effect. In the two sample case, it is the difference between two groups in the population, i.e., $\mu_1 - \mu_2$. This difference may not be due to the treatment effect alone, but experimental error as well.

Generally,

$H_0 : \mu_1 = \mu_2 = ... = \mu_a$ (No difference in group means, a groups in total)

$H_a : \mu_i \neq \mu_j$ for at least one $i \neq j$ where $i, j = 1, 2, ..., a$ (Difference in group means)

$$X = \frac{\text{variation between groups}}{\text{variation between subjects, within groups}}$$

If $X > 1$, then more evidence we have a difference in group means (more significant).

2.11 One-Way ANOVA Notation

- One factor, Factor A with a levels

- j denotes the group/level and i denotes the position of the subject within that level/group ($j = 1, ..., a$; $i = 1, ..., n_j$)

- y_{ij} denotes the ith observation from the jth group (i is the index, j is the group/level)

- n_j is the number of subjects in the jth group

- $N = \sum_{j=1}^{a} n_j$ is the total sample size

- $\bar{Y}_T$ is the grand sample mean regardless of group

- $\bar{y}_j$ is the sample mean in the jth group

Definition 2.11. Total Sum of Squares

SST measures the total sample variability (total deviation from the grand mean):

$$SST = \underbrace{\sum_{j=1}^{a} \sum_{i=1}^{n_j}}_{\text{All observations}} (y_{ij} - \bar{y}_T)^2$$

- $(N - 1)$ df

- $SST = SSReg + SSE$

Definition 2.12. Sum of Squares for A

Sum of Squares for Factor A (SSA) measures the variability of the factor level means:

$$SSA = \sum_{j=1}^{a}\sum_{i=1}^{n_j}(\hat{y}_{ij} - \bar{y}_T)^2 = \sum_{j=1}^{a} n_j(\bar{y}_j - \bar{y}_T)^2$$

Note that $\hat{y}_{ij}$ is the estimated observation, $\bar{y}_T$ is the grand sample mean. It has $(a-1)$ df, df=number of parameters in the model -1. It's normally called SSA or $SSReg$.

Definition 2.13. Residual/Error Sum of Squares

SSE measures the variability that is left unexplained by the model:

$$SSE = \sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij} - \hat{y}_{ij})^2 = \sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$$

Note that it has $(N-a)$ df, df=number of data - number of parameters in the model. SSE decreases as more predictors are added to the model. We want SSE to be low!

2.12   One-Way ANOVA Table

Note that SS refers to Sum of Squares, MS refers to Mean Square $= \frac{SS}{df}$

ANOVA Table			
Source	df	SS	MS
Group/Regression	$a - 1$	SSA	$\frac{SSA}{a-1} = MSA$
Error	$N - a$	SSE	$\frac{SSE}{N-a} = MSE$
Total	$N - 1$	SST	

$$MSE = \frac{\sum_{j=1}^{a}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2}{N - a} = \frac{\sum_{j=1}^{a}(n_j - 1)s_j^2}{N - a} = s_{\text{pooled}}^2$$

SSA is often called "between group" SS: measures the variability of observations between factor levels

SSE is often called "within group" SS: measures the variability of observations within factor levels.

2.13   F-Test

$$F = \frac{MSA}{MSE} \sim F_{a-1,N-a}$$

Hypothesis Test Using ANOVA

Hypothesis: $H_0 : \mu_1 = \mu_2 = ... = \mu_a$ vs. H_a : at least one of the means μ_j differs from others; for $j = 1, 2, ..., a$.

Test Statistic and its Distribution under H_0:

$$F = \frac{MSA}{MSE} = \frac{\frac{SSA}{a-1}}{\frac{SSE}{N-a}} \sim F_{a-1,N-a} \text{ under } H_0$$

P-value: $p = P(F_{a-1,N-a} > F)$

Example 2.14. Arthiritis Clinical Study-ANOVA

Example: Arthritis Clinical Study -ANOVA

Test the question of interest using ANOVA.

```
# Regression Model with intercept
> arth.regmodel <- lm(DiffinPain ~ Treatment)

# ANOVA table
> anova(arth.regmodel)

Analysis of Variance Table

Response: DiffinPain
            Df  Sum Sq Mean Sq F value   Pr(>F)
Treatment    1   7.379  7.3792  6.9749 0.008713 **
Residuals  291 307.870  1.0580
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Q: What do you notice about your results from the pooled t-test, t-test for slope in linear regression, and from ANOVA F-test?

A:

In this study we have levels $a = 2$, $H_0 : \mu_A = \mu_P$, $H_a : \mu_a \neq \mu_P$.

$$F = \frac{MSA}{MSE} = \frac{MS_{\text{Treatment}}}{MS_{\text{Residual}}} = 6.97 \sim F_{1,291}$$

So the p-value will be $p = P(F_{1,291} > 6.97) = 0.008713 < 0.05$. Therefore, we reject H_0, we have very strong evidence that the Auranofin and Placebo therapy differs. Since $a = 2$, we can have an easy follow up for two sample t-test with $H_a : \mu_A > \mu_P$. Once $\bar{y}_A > \bar{y}_P$ then we cam conclude that Auranofin is effective.

3 Lec 3 One-Way ANOVA with a Levels

3.1 Computational Formulae for SS

- $SST = \sum_{j=1}^{a} \sum_{i=1}^{n_j} y_{ij}^2 - N\bar{y}_T^2 = [Y] - [T]$

- $SSA = \sum_{j=1}^{a} n_j \bar{y}_j^2 - N\bar{y}_T^2 = [A] - [T]$

- $SSE = \sum_{j=1}^{a} \sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^{a} n_j \bar{y}_j^2 = [Y] - [A]$

$[Y]$ denotes the individual observations (n), $[A]$ denotes the actual groups means, $[T]$ denotes the grand means. The actual formula for these are:

- $[Y] = \sum_{j=1}^{a} \sum_{i=1}^{n_j} y_{ij}^2$

- $[A] = \sum_{j=1}^{a} n_j \bar{y}_j^2$

- $[T] = N\bar{y}_T^2$

3.2 Proofs of Computational Formulae for SS

Proof.

$$
\begin{aligned}
SST &= \sum_{j=1}^{a} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_T)^2 \\
&= \sum_i \sum_j y_{ij}^2 - 2\bar{y}_T \sum_i \sum_j y_{ij} + \sum_i \sum_j \bar{y}_T^2 \\
&= [Y] - 2\bar{y}_T N\bar{y}_T + N\bar{y}_T^2 \\
&= [Y] - N\bar{y}_T^2 \\
&= [Y] - [T]
\end{aligned}
$$

$$SSA = \sum_{j=1}^{a} n_j (\bar{y}_j - \bar{y}_T)^2$$

$$= \sum_j n_j \bar{y}_j^2 - 2 \sum_j \bar{y}_T n_j \bar{y}_j + \sum_j n_j y_T^2$$

$$= [A] - 2\bar{y}_T \sum_j n_j \bar{y}_j + N y_T^2$$

$$= [A] - N \bar{y}_T^2$$

$$= [A] - [T]$$

$$SSE = \sum_i \sum_j (y_{ij} - \bar{y}_j)^2$$

$$= \sum_i \sum_j y_{ij}^2 - \sum_i \sum_j 2\bar{y}_j y_{ij} + \sum_i \sum_j \bar{y}_j^2$$

$$= [Y] - 2 \sum_j \bar{y}_j \sum_i y_{ij} + \sum_j n_j \bar{y}_j^2$$

$$= [Y] - \sum_j n_j \bar{y}_j^2$$

$$= [Y] - [A]$$

$$SST = [Y] - [T]$$

$$= [A] - [T] + [Y] - [A]$$

$$= SSA + SSE$$

$\square$

Example 3.1. Example of Fertilizers in Farming

Example: Fertilizers in Farming

There are four different fertilizers (brands A, B, C, D). We wish to determine if there a difference in the crop yield due to the fertilizer.

There are 16 plots of lands available to be tested. Each fertilizer is randomly assigned to four plots. The crop yield (in kg) of each plot is then measured after the fertilizer is used.

▶ Define the experiment: identify the experimental units, factor(s) and levels, response variable(s), etc.
▶ State the appropriate hypotheses to test the question of interest
▶ Is the design balanced or unbalanced?

R output as follows:

R Code - Farming Example

```
> farmdata <- read.csv("farmingfertilizerdata.csv")
> farmdata
    Fertilizer CropYield
1          A        65
2          A        54
.
5          B        55
6          B        58
.
9          C        64
10         C        67
.
13         D        60
.
16         D        70

> attach(farmdata)

> tapply(CropYield, Fertilizer, length)
A B C D
4 4 4 4

> tapply(CropYield, Fertilizer, mean)
    A     B     C     D
58.75 60.00 68.75 65.50

> tapply(CropYield, Fertilizer, sd)
       A        B        C        D
4.856267 4.396969 4.272002 4.434712
```

$a = 4$

n_j (circled: length)

$\bar{y}_j$ (circled: mean)

s_j (circled: sd)

12/16

Boxplots

$y \quad \sim \quad$ Factor

```
> boxplot(CropYield ~ Fertilizer,
main="Farming Fertilizer Experiment",
xlab="Fertilizer",
ylab="Crop Yield (kg)")
```


highest

13/16

Example: Farming - ANOVA

Test the question of interest using ANOVA. Include all the steps.

```
# Regression Model with intercept
> farm.regmodel<- lm(CropYield ~ Fertilizer)

# ANOVA table
> anova(farm.regmodel)

Analysis of Variance Table

Response: CropYield
          Df Sum Sq Mean Sq F value  Pr(>F)
Fertilizer  3  264.5  88.167   4.3629 0.02694 *
Residuals  12  242.5  20.208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(handwritten annotations: linear model, response ~ Factors, $a = 4$, $N = 16$, p-value)

The experimental unit is a plot. The factor is the brand of fertilizer which are A, B, C, D (4 levels/treatments). The response is the crop field (in kg). In the experiment, $N = 16$ plots, $n = 4$, $n_A = n_C = n_C = n_D$ that means it's a balanced design. The hypothesis test is $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ vs. $H_a : \mu_i \neq \mu_j$ for at least one pair of i, j that $i \neq j$, and $i, j =$ A, B, C, D.

$$F = \frac{MS_{\text{Fertilizer}}}{MSE} = 4.3629 \sim F_{3,12} \text{ under } H_0$$

$$p = P(F_{3,12} > 4.3629) \doteq 0.0269$$

Since p value is less than 0.05, reject H_0. Strong evidence to conclude that the brands of fertilizer defer in crop yield.

Question for the next chapter:

Which fertilizer is the best? - We can examine the boxplot or look at all pairwise comparisons between fertilizers.

Is Fertilizer A better than B, etc.? Does using Fertilizer A result in double the crop yield of using B, etc.? - Test Contrasts.

3.3 What happens if we reject H_0?
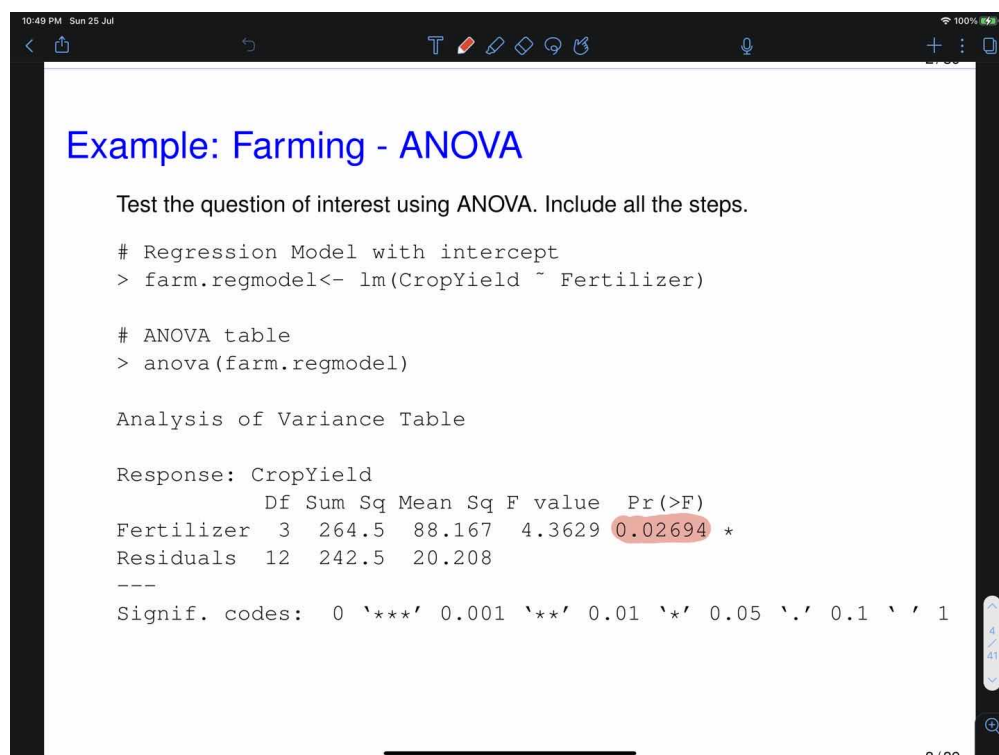
If we reject H_0, ANOVA tells us that there is statistically significant difference between group means. When there are more than 2 groups, it can be because:

- One group has a different mean than others

- Some groups have different means

- All groups have different means from each other

4 Lec 4 Follow-Up Comparisons for One-Way ANOVA: Testing Contrasts and Post-Hoc Analysis

4.1 Recall Example of Fertilizers in Farming- ANOVA

Example 4.1. The R output as follows:

4.2 Pairwise Comparisons

Definition 4.2. Compare one group mean to another group mean (1 df since there are 2 groups). Conducted like a two-sample t-test but use pooled estimate of variance which is calculated using all groups:

$$s_p^2 = \frac{1}{N-a} \sum_{j=1}^{a} (n_j - 1)s_j^2 = MSE$$

where s_j^2 is the sample variance in the jth group and $\sum_{j=1}^{a} n_j = N$. This s_p^2 is a better estimate of the variance than just using observations from 2 groups, which means more power. The idea is to set up (linear) contrasts to make comparisons.

Definition 4.3. Linear Contrasts A linear contrast is a linear combination of the group means: $\psi = \sum_{j=1}^{a} c_j \mu_j$. Restrict sum of the coefficients in the contrast to be 0: $\sum_{j=1}^{a} c_j = 0$.

We use $H_0 : \mu_j = \mu_k$ for $j \neq k$, j, $k = 1, 2, ..., a$. Results in coefficients such as $c = \{1, -1, 0, 0\}$, or $c = \{0, 1, 0, -1\}$, etc. Usually make the sign of ψ positive to reflect direction of the difference we predict. If group 1 is predicted to be higher than group 2, then $c = \{1, -1, 0, 0\}$ rather than $c = \{-1, 1, 0, 0\}$. Test $\psi = 0$ using $\psi = \sum_{j=1}^{a} c_j \bar{Y}_j$.

Advantages of using Contrasts:

- Compare more than 2 groups or more than one set of equalities.

- General procedure that cover many possible research questions.

- Simple to check if contrasts are orthogonal/linearly independent.

Example 4.4. Farming- Pairwise Comparisons Using Contrasts

Set up a table with coefficients to compare these pairwise means:

- Is Fertilizer A different from B?- ψ_1

- Is Fertilizer B different from C?- ψ_2

- Is Fertilizer C different from D?- ψ_3

Group	A	B	C	D
ψ_1	1	-1	0	0
ψ_2	0	1	-1	0
ψ_3	0	0	1	-1

Table Including Means and Contrast Estimates

Group	A	B	C	D	Contrast Estimate
Means	58.75	60	68.75	65.5	$\hat{\psi}$
ψ_1	1	-1	0	0	-1.25
ψ_2	0	1	-1	0	-8.75
ψ_3	0	0	1	-1	3.23

Example of the calculation:

$$\hat{\psi}_1 = \sum_{j=1}^{4} c_j \bar{y}_j = 1(\bar{y}_A) - 1(\bar{y}_B) + 0(\bar{y}_C) + 0(\bar{y}_D)$$

$$= \bar{y}_A - \bar{y}_B$$

$$= 58.75 - 60 = -1.25$$

Example 4.5. Farming- More Complex Comparisons

Set up a table with coefficients to compare the following:

- Compare the two farms: ψ_4

- Compare the two types of element-based fertilizers: ψ_5

Farm	2	2	1	1
Element	P	P	N	P
ψ_4	1	1	-1	-1
ψ_5	1	1	-3	1

Table Including All Comparisons

Group	A	B	C	D	Contrast Estimate
Farm	2	2	1	1	
Element	P	P	N	P	
Means	58.75	60	68.75	65.5	$\hat{\psi}$
ψ_1	1	-1	0	0	-1.25
ψ_2	0	1	-1	0	-8.75
ψ_3	0	0	1	-1	3.23
ψ_4	1	1	-1	-1	-15.5
ψ_5	1	1	-3	1	-22

Note that those numbers are table of C_j's.

4.3 Testing Using Contrasts

Under certain conditions, $\hat{\psi} \sim N\left(\sum_{i=1}^{a} c_j \mu_j, \sigma^2 \sum_{i=1}^{a} \frac{c_j^2}{n_j}\right)$

Definition 4.6. Contrast Sum of Squares

The Contrast Sum of Squares is explained variation by the contrast:

$$SS_\psi = \frac{\hat{\psi}^2}{\sum_{j=1}^{a}(c_j^2/n_j)}$$

- Since 1 df, $MS_\psi = SS_\psi$

- Compare to MSE to see if contrast explains more of the variation than noise:

$$F = \frac{MS_\psi}{MSE}$$

- Equivalent to a t-test: $t = \frac{\hat{\psi}}{\sqrt{MSE\sum_{j=1}^{a}(c_j^2/n_j)}} \sim t_{N-a}$ under H_0

- t Confidence Intervals for contrasts:

$$\hat{\psi} \pm t_{N-a;\alpha/2}\sqrt{MSE\sum_{j=1}^{a}(c_j^2/n_j)}$$

Example 4.7. Farming R output:

Farming R Output - ANOVA

```
# Regression Model with intercept
> farm.regmodel<- lm(CropYield ~ Fertilizer)

# ANOVA table
> anova(farm.regmodel)

Analysis of Variance Table

Response: CropYield
          Df Sum Sq Mean Sq F value  Pr(>F)
Fertilizer  3  264.5  88.167  4.3629 0.02694 *
Residuals  12  242.5  20.208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

R Code - Contrasts

$$Y_i = \beta_1 \mathbb{I}_{A,i} + \beta_2 \mathbb{I}_{B,i} + \beta_3 \mathbb{I}_{C,i} + \beta_4 \mathbb{I}_{D,i} + \varepsilon$$

No intercept

```
# Model without intercept
> model2 <- lm(CropYield ~ Fertilizer - 1)

# Coefficients of contrasts
> L <-matrix(c(
+ 1, -1, 0, 0,
+ 0, 1, -1, 0,
+ 0, 0, 1, -1,
+ 1, 1, -1, -1,
+ 1, 1, -3, 1), nrow=5, byrow=T)

# Required package
> require(multcomp)
```

$\beta_1 = \mu_A$

$\vdots$

$\beta_4 = \mu_D$

35

R Output - Testing Contrasts   *matrix of coefficient*

```
> summary(glht(model2, L), test=adjusted("none"))

Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = CropYield ~ Fertilizer - 1)

Linear Hypotheses:
          Estimate Std. Error t value Pr(>|t|)
 1 == 0    -1.250      3.179   -0.393  0.70104
 2 == 0    -8.750      3.179   -2.753  0.01751 *
 3 == 0     3.250      3.179    1.022  0.32675
 4 == 0   -15.500      4.495   -3.448  0.00482 **
 5 == 0   -22.000      7.786   -2.826  0.01530 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.
(Adjusted p values reported -- none method)
```

Handwritten annotations: H_0: $\psi_1 = 0$ $\psi_2 = 0$ $\psi_3 = 0$ $\psi_4 = 0$ $\psi_5 = 0$

15/39

State the conclusions based on the contrasts that were tested.

- Compare brands of fertilizers

- Compare farms

- Compare element bases

- Which fertilizer brand would you recommend? Why?

Conclusions:

- Moderate (to strong) evidence that Fertilizer C yields more than Fertilizer B

- Strong evidence that Farm 1 yields more than Farm 2

- Moderate (to strong) evidence that Nitrogen-based fertilizers are better than Phosphorous-based.

4.4 Orthogonal Contrasts

Definition 4.8. This is useful for dividing sums of squares into different components. Any SS can be decomposed into as many independent sums as there are degrees of freedom. Two contrasts are

36

said to be orthogonal if: for ψ_1 and ψ_2 with coefficient vectors $\vec{c_1}$ and $\vec{c_2}$ respectively,

$$\sum_{j=1}^{a} \frac{c_{1j}c_{2j}}{n_j} = 0$$

Note that when sample sizes equal, then orthogonality simplifies to dot product of 0:

$$\vec{c_1} \cdot \vec{c_2} = \sum_{j=1}^{a} c_{1j}c_{2j} = 0$$

Orthogonality is a property of a pair of contrasts. If there are k contrasts, then there are $\binom{k}{2}$ pairs.

Definition 4.9. If all contrasts in a set are orthogonal to one another, then it's called mutually orthogonal. There can be no more mutually orthogonal contrasts than the treatment df (i.e. $a-1$). Any set of $a-1$ mutually orthogonal contrasts can be used to express the SS for the treatment/factor (for balanced designs at least):

$$SSA = \sum_{k=1}^{a-1} SS_{\psi k}$$

If orthogonal, each SS is independent and each contrast test can be interpreted additively (no problems with multi-collinearity).

Example 4.10. Example: Farming - Orthogonal Contrasts

How many contrasts are orthogonal amongst the 5 contrasts we tested?

- df for Fertilizer is 3. We are limited to 3 orthogonal contrasts

- Can we find a set of 3? Which contrasts?

In this example, we can easily find that $a = 4$, df$= 3$ (calculated from $a - 1$)

Table to check Orthogonality					
Group	A	B	C	D	
ψ_1	1	-1	0	0	
ψ_2	0	1	-1	0	
ψ_3	0	0	1	-1	
ψ_4	1	1	-1	-1	
ψ_5	1	1	-3	1	Sum
ψ_1 vs ψ_2	0	-1	0	0	-1
ψ_1 vs ψ_3	0	0	0	0	0
ψ_1 vs ψ_4	1	-1	0	0	0
ψ_1 vs ψ_5	1	-1	0	0	0
ψ_2 vs ψ_3	0	0	-1	0	-1
ψ_2 vs ψ_4	0	1	1	1	2
ψ_2 vs ψ_5	0	1	3	0	4
ψ_3 vs ψ_4	0	0	-1	1	0
ψ_3 vs ψ_5	0	0	-3	-1	-4
ψ_4 vs ψ_5	1	1	3	-1	4

We have $\binom{5}{2} = 10$ pairs of contrast. In this example, we have $a = 4$, so there are $a - 1 = 3$ mutually orthogonal contrasts. We will choose ψ_1, ψ_3, ψ_4 to be mutually orthogonal.

4.5 Method for Constructing Mutually Orthogonal Set

We have a groups, we will compare 1st group to average of all other groups. For all other contrasts, make 1st group's coefficient 0. Compare 2nd group to average of the latter groups. Continue this pattern. Make 2nd group's coefficient 0. Compare 3rd group to average of the latter groups, and so on...

Constructing Helmert Contrasts

Start with $c_1 = 1$ and $c_j = -\frac{1}{a-1}$ for $j > 1$. For $j = 2, ..., a - 1$, let $c_k = 0$ when $k \leq j - 1$ and $c_j = 1$ and $c_k = -\frac{1}{a-j}$ when $k \geq j + 1$. These contrasts are known as Helmert Contrasts.

Example 4.11. Farming: Helmert Contrasts

Construct the Helmert Contrasts. Verify that their contrasts are mutually orthogonal.

In this example, we have group $a = 4$, it implies to there are 3 contrasts are mutually orthogonal. Note that the following vector has the order of (A, B, C, D):

$$\psi_1 : \vec{c_1} = (1, -1/3, -1/3, -1/3) \implies A \text{ to average of others}$$

$$\psi_2 : \vec{c_2} = (0, 1, -1/2, -1/2) \implies B \text{ to average of latter groups}$$

$$\psi_3 : \vec{c_3} = (0, 0, 1, -1) \implies C \text{ to average of latter groups}$$

$$\vec{c_1} \cdot \vec{c_2} = \vec{c_1} \cdot \vec{c_3} = \vec{c_2} \cdot \vec{c_3} = 0 \implies \psi_1, \psi_2, \psi_3 \text{ are mutually orthogonal.}$$

4.6   Balanced Designs

ANOVA is the best for balanced designs, there are potential problems with unbalanced designs. MSE in unbalanced designs:

- MSE is a pooled variance, weighted by size of each group. Larger groups dominate MSE towards their variance.

- MSE is still unbiased for σ^2.

In balanced designs:

- MSE gets an equal contribution from each group.

- SSA can be decomposed into the sum of the contrast SS.

- Usually more powerful.

ANOVA can still be used for unbalanced designs. The more unequal the group sizes, the poorer the performance. So, we need to attempt to design balanced experiment. Sometimes necessary to use unbalanced designs (e.g. Larger control group).

4.7   Problem with Multiple Comparisons

Inflation of Type I Error rate: Chance of making a Type I Error is very high when conducting many simultaneous tests.

We are often interested in a family/group of different hypothesis tests or confidence intervals

- All pairwise differences in group means, there are $\binom{a}{2}$ pairwise comparisons.

- Meaningful contrasts

The familywise error rate, α_{FW} is the probability of making at least one Type I Error in the set of tests. Suppose we conduct k independent tests, each at significance level α, then:

$$\alpha_{FW} = P(\text{reject at least one } H_0 | \text{all } H_0 \text{ are true})$$
$$= 1 - (1 - \alpha)^k$$

Example 4.12. Familywise Type I Error Rate

If we conduct $k = 5$ tests (e.g., 5 contrasts such as in the Farming example) with $\alpha = 0.05$, then

$$\alpha_{FW} = P(\text{at least one Type I Error}) = 1 - (1 - 0.05)^5 = 0.2262$$

The probability of committing a Type I Error somewhere in these tests is now 22.62%. Inflated Type I Error: will increase even more as we increase number of tests. If tests are dependent, α_{FW} will be lower, but not lower than the Type I Error rate for any individual test. We need to control for this.

4.8   Pre-planned Tests/Primary Research Questions

Most often, an experiment is set up to answer one primary research question. Could also have follow-up questions.

- If research question/tests are pre-planned/ad-hoc (before looking at the results of the analysis), it can be tested without correcting Type I Error

- Why before and not after? *Fishing for significant results!*

- If question is important, do not conduct it in a family of less powerful tests

- Need to correct if we test secondary research questions or make comparisons after seeing results of analysis (post-hoc)

4.9   Simple Solution to Correct Type I Error

- Decrease the family wise Type I Error rate (α_{FW}) by decreasing the Type I Error rate for each individual test (α)

- Adjust Type I Error for the worst case (independent tests)

- This adjustment results in less power/higher chance of making Type II Error

Are there other solutions? Yes:

1. Bonferroni

2. Sidak-Bonferroni

3. Tukey's HSD

4. Scheffe

4.10 Bonferroni Method

Definition 4.13. Bonferroni inequality: $P(A \cup B) \leq P(A) + P(B)$.

Let A_i be the event of making a Type I Error on ith test, then: $\alpha_{FW} = P(\bigcup_{i=1}^{k} A_i) \leq \sum_{i=1}^{k} P(A_i)$. Note that $\alpha_i = P(A_i)$ for $i = 1, ..., k$.

Bonferroni Method

For each of the k tests, use a significance level of α_{FW}/k. Then CI coverage rate or Type I Error rate is at most $100(1 - \alpha_{FW})\%$ or α_{FW}.

- Conservative procedure

- Since the chance of making at least one Type I Error can be much lower than α, it is likely that the k tests may result in Type II Error (less powerful)

- Type I Errors are not always more serious than Type II (depends on the context)

Example 4.14.

```
3:53 PM   Sat 31 Jul

R Code: Bonferroni - Contrasts

    # Model without intercept
    > model <- lm(CropYield ~ Fertilizer - 1)

    # Coefficients of contrasts
    > L <-matrix(c(
    + 1, -1, 0, 0,
    + 0, 1, -1, 0,
    + 0, 0, 1, -1,
    + 1, 1, -1, -1,
    + 1, 1, -3, 1), nrow=5, byrow=T)

    # Required package
    > require(multcomp)

    > summary(glht(model, L), test=adjusted("bonf"))

    Simultaneous Tests for General Linear Hypotheses

    Fit: lm(formula = CropYield ~ Fertilizer - 1)

    Linear Hypotheses:
            Estimate Std. Error t value Pr(>|t|)
    1 == 0    -1.250      3.179  -0.393   1.0000
    2 == 0    -8.750      3.179  -2.753   0.0876 .
    3 == 0     3.250      3.179   1.022   1.0000
    4 == 0   -15.500      4.495  -3.448   0.0241 *
    5 == 0   -22.000      7.786  -2.826   0.0765 .
    ---
    Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    (Adjusted p values reported -- bonferroni method)
```

Handwritten annotations: $\Psi = \sum\limits_{j=1}^{a} c_j \mu_j$, $k = 5$ tests/contrasts

4.11 Sidak-Bonferroni Procedure

Bonderroni procedure is simple to use (by hand). We can improve procedure if using computer for calculations by directly working with: $\alpha_{FW} = 1 - (1 - \alpha)^k$

Definition 4.15. Sidak Bonderroni Method

Adjust the individual Type I Error as follows:

$$\alpha = 1 - (1 - \alpha_{FW})^{1/k}$$

For example, $k = 5$ and $\alpha_{FW} = 0.05$ compare Bonferroni and Sidak-Bonferroni:

- $\alpha_B = \frac{0.05}{5} = 0.01$

- $\alpha = 1 - (1 - 0.05)^{1/5} = 0.0102$ This is slight higher due to lower individual Type I Error rate.

Still have same family-wise rate, but more power for individual tests.

4.12 Confidence Intervals

For tests, we adjust the p-values or individual significance levels. For CIs, adjust the critical value. Suppose we want to construct k confidence intervals with familywise coverage of $100(1 - \alpha_{FW})\%$, then use

$$\hat{\theta} \pm t_{df, \frac{\alpha_{FW}}{2k}} SE(\hat{\theta})$$

4.13 Tukey's HSD Procedure

- Based on "Studentized Range Distribution" which is based on $\max_{j,k \in \{1,2,\dots,a\}} \{\bar{y}_j - \bar{y}_k\}$

- Gives a simultaneous Type I Error rate of α_{FW} or confidence level of $100(1 - \alpha_{FW})\%$

- Conservative if unbalanced design

- Less conservative than Bonferroni method, especially if group sample sizes are equal (approx.)

Definition 4.16. Tukey's Honestly Significant Difference (for differences between pairs of groups)

Confidence interval for $\mu_k - \mu_j$ is:

$$(\bar{y}_j - \bar{y}_k) \pm \frac{q_{\alpha_{FW};a,N-a}}{\sqrt{2}} \sqrt{MSE(\frac{1}{n_j} + \frac{1}{n_k})}$$

where $q_{\alpha;a,N-a}$ is the appropriate critical value of the Studentized Range Distribution.

To avoid redundancy, compare the largest and smallest means and if we failed to reject H_0, then stop. If H_0 is rejected, compare next largest to next smallest means and continue on...

Code:

```
> TukeyHSD(aov(model_name), factor="factor_name")
```

'R' Gives CIs for all pairwise means. Reject H_0 for the given pair if the adjusted p-value $< \alpha$ or CI does not contain 0.

Example 4.17. Here is the R Code and output for Tukey on the example of Farming:

R Code: Tukey

```
> tukeyCIs = TukeyHSD(aov(farm.regmodel),factor=Fertilizer,
conf.level = 0.95)

   Tukey multiple comparisons of means
     95% family-wise confidence level

Fit: aov(formula = farm.regmodel)

$Fertilizer
      diff        lwr       upr       p adj
B-A   1.25  -8.1872614 10.687261 0.9784249
C-A  10.00   0.5627386 19.437261 0.0367783
D-A   6.75  -2.6872614 16.187261 0.2005978
C-B   8.75  -0.6872614 18.187261 0.0724032
D-B   5.50  -3.9372614 14.937261 0.3510946
D-C  -3.25 -12.6872614  6.187261 0.7399725
```
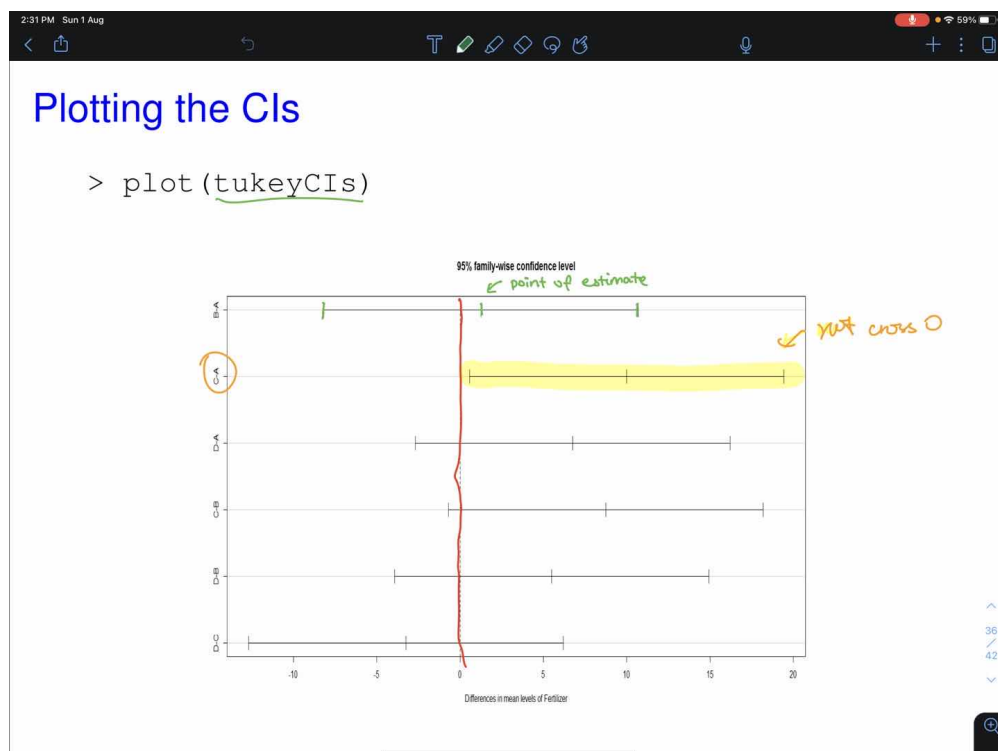
$\alpha_{FW} = 0.05$

$\binom{4}{2} = 6$ pairs

Plotting the CIs

```
> plot(tukeyCIs)
```


not cross 0

point of estimate

Can we control Type I Error over all possible contrasts?

- Using Bonferroni methods, α for individual tests will go to 0 as we increase number of tests.

- Tukey covers pairwise differences only

- If there is some other way, we cam test for any number of questions without worrying i.e. Scheffe.

4.14 Scheffe's Procedure

Consequence for ANOVA: If we failed to reject H_0 in the ANOVA F-Test, no follow-up test will be significant using Scheffe's procedure. If the ANOVA F-Test is significant, there is at least one contrast that is significant using Scheffe's procedure. This is not necessarily true for other multiple comparison procedures.

Apply for contrasts using F or t-test. Contrasts do not have to be specified beforehand.

Definition 4.18. Scheffe's Procedure using F tests

Scheffe's Critical Value, F_S: Multiply the F critical value for the contrast by $(a-1)$:

$$F_S = (a-1)F_{a-1,N-a,\alpha_{FW}}$$

F_S is larger than the uncorrected critical value (accounts for many contrasts that could be tested). Calculate the observed F test statistic and compare it to the Scheffe critical value.

Definition 4.19. Scheffe's Procedure using t tests

Scheffe's Critical Value, t_S:

Recall: $t_v^2 = F_{1,v}$:

$$t_S = \sqrt{F_S} = \sqrt{(a-1)F_{a-1,N-a,\alpha_{FW}}}$$

Calculate the observed t test statistic and compare it to the Scheffe critical value.

Example 4.20. Farming Example

4.15 Pros and Cons of each Method

Which method to choose?

- If design is balanced (approx.) and we are interested in all pairwise differences between groups, Tukey is most powerful.

- If we are interested in follow-up contrasts (beyond pairwise differences) and all comparisons can be specified in advance:
  - Bonfferoni is the most of the time more powerful than Tukey
  - Sidak-Bonferroni is even better

- If we are interested in several follow-up contrasts that connot be specified in advance, Scheffe should be used

What to do if a result is significant before correction but not significant after the correction?

- Exploratory Findings!

- If it is of importance, best to design another experiment to directly test it (pre-planned test)

5 Lec 5 Linear Model Assumptions and ANOVA Coding Schemes

5.1 Statistical Models for ANOVA

- Many different models that can be assumed for ANOVA

- Discussed some such as: Dummy variable/reference, cell means model, etc.

- Assumptions: Gauss-Markov conditions, normality, etc.

- Now develop formal linear models