

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

Jiaqi Bi, University of Toronto

July 24, 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 Elementary Number Theory 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/>.

This work is licensed under a Creative Commons “Attribution-NonCommercial-NoDerivatives 4.0 International” license.



# Contents

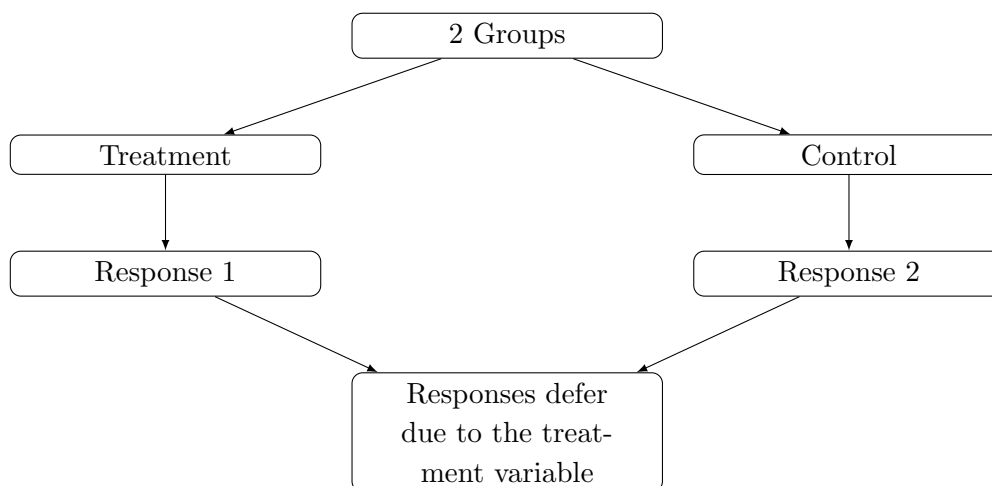
<b>1 Lec 1 Introduction to Experiments and Types of Data</b>	<b>4</b>
1.1 General Outline of an Experiment . . . . .	4
1.2 Steps of an Experimental Design . . . . .	5
1.3 Definitions used in Statistical Inference . . . . .	5
1.4 Experimental Study vs. Observational Study . . . . .	6
1.5 Types of Variables . . . . .	6
1.6 Crossed vs. Nested Factors . . . . .	9
1.7 Nuisance Variables . . . . .	9
1.8 Confounding . . . . .	11
1.9 Blinding . . . . .	11
1.10 Balance . . . . .	11
<b>2 Lec 2 Review of Hypothesis Testing and One-Way ANOVA with 2 Levels</b>	<b>12</b>
2.1 Single Factor Analysis with 2 levels . . . . .	12
2.2 Review of Hypothesis Testing . . . . .	12
2.3 Review: Two Sample T-Tests . . . . .	12
2.4 Pooled T-Test . . . . .	13
2.5 Welch's T-Test Using Satterthwaite Approximation . . . . .	14

# 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:

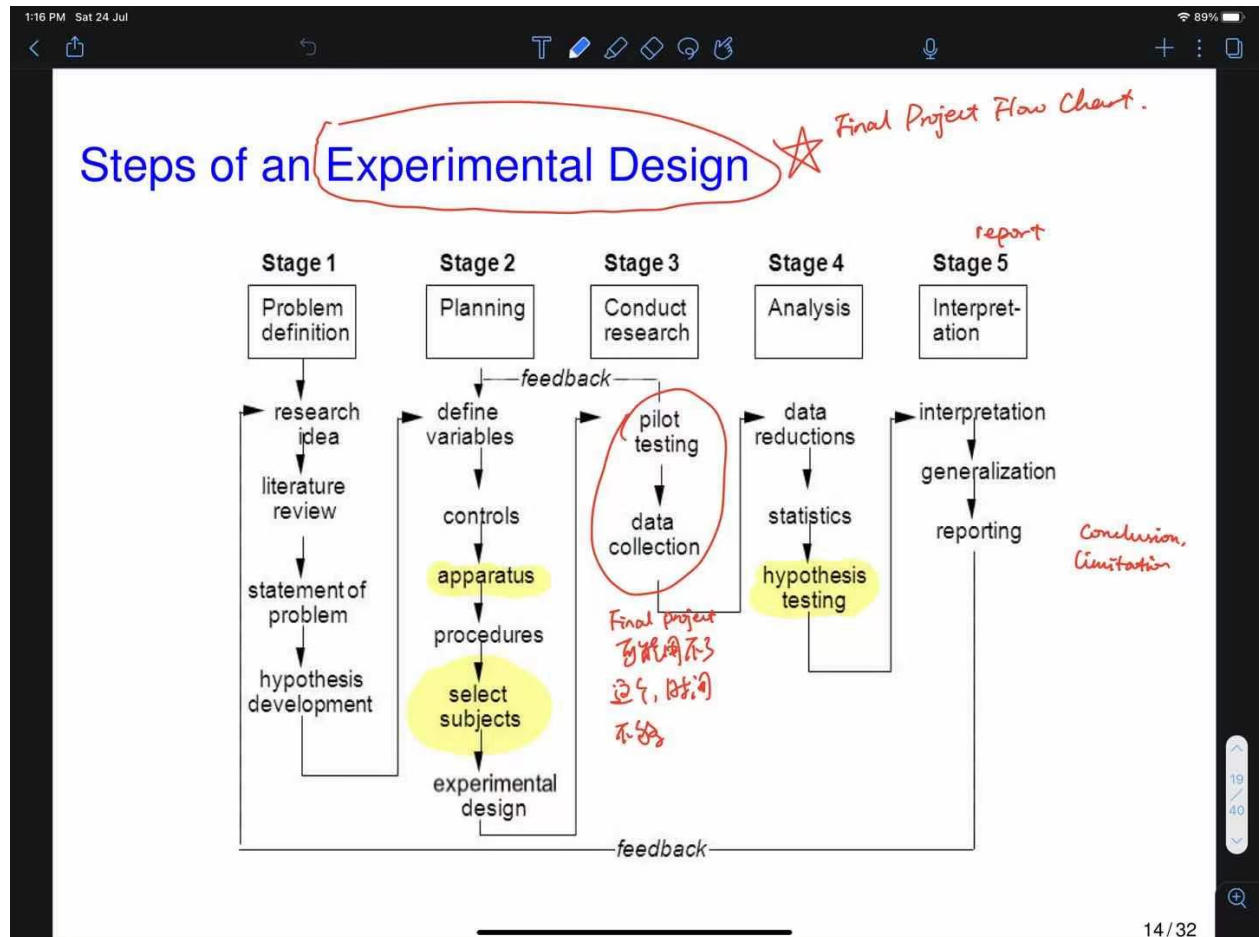


Example of different treatments: Medicine & Placebo

**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.

## 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:

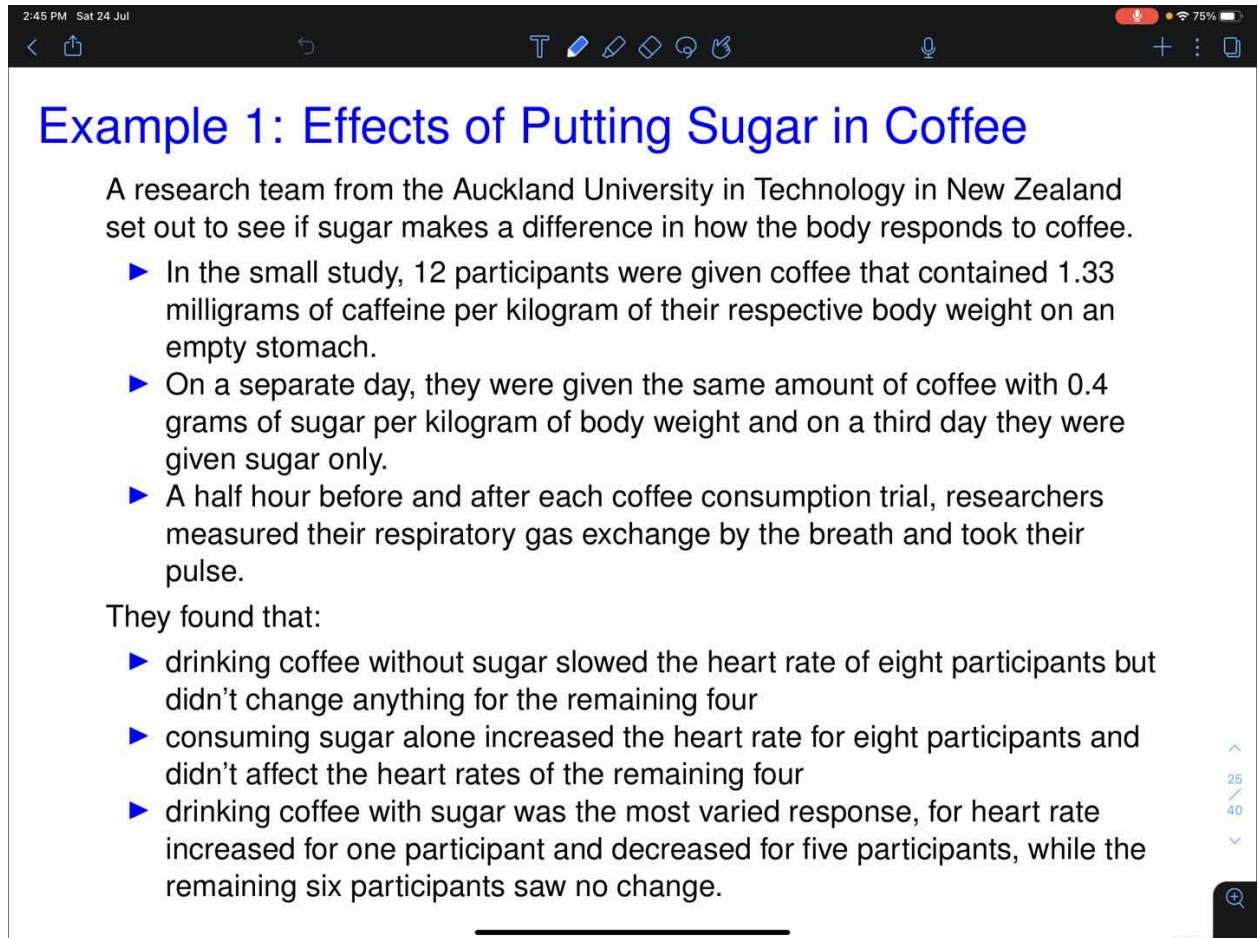
1. Nominal: Categorize units into distinct classes.
  - Unordered categories
  - Numerical computations are not applicable
  - Ex: Gender, POSt, Favorite colors...
  - When controlled by researcher, called **Factor**
2. Ordinal: Ordered categories without natural units/distance metric
  - Natural ordering 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 ( $^{\circ}\text{C}$ ), dates...
4. Ratio: Numerical measurements on which a unique and non-arbitrary zero value exists
  - Ratios are meaningful, sensible to carry out 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  $\{a_1, a_2, a_3, \dots\}$ . 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



The screenshot shows a presentation slide with a dark header bar containing navigation icons and a status bar at the top right showing the time as 2:45 PM on Saturday, July 24, and a battery level of 75%. The slide title is 'Example 1: Effects of Putting Sugar in Coffee' in blue. The text describes a research study from Auckland University of Technology in New Zealand. It details three conditions: coffee with caffeine, coffee with sugar, and sugar alone, and reports findings on heart rate and respiratory gas exchange for 12 participants.

## 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)*



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

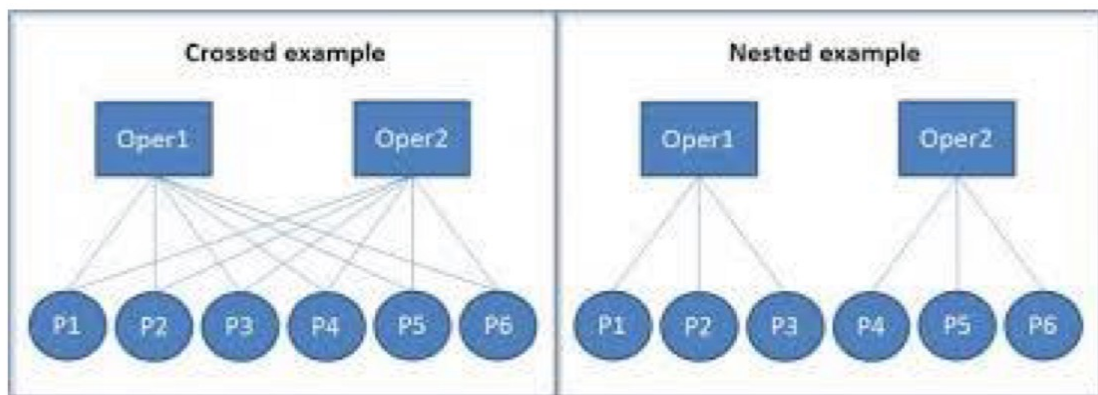
**Example 1.3. Effects of Teaching Style** Suppose you wish to conduct an experiment to see if teaching style (lecturer, facilitator, online technology) and the amount of lecture time (2 hours, 3 hours/week) affects the learning outcome of students in a large university course.

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...

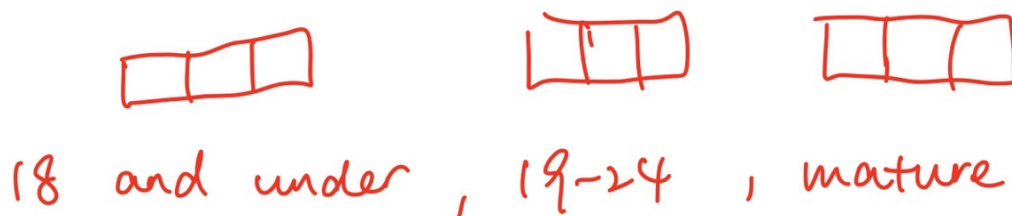
We use **Control, Blocking, Randomization, Replication** to deal with nuisance variables.

**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.



**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  $X_1, \dots, X_{n1}$  a random sample from population 1 and  $Y_1, \dots, Y_{n2}$  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  $\underbrace{H_a : \mu_1 - \mu_2 < D_0 \text{ OR } H_a : \mu_1 - \mu_2 > D_0}_{\text{One-tail test}}$

**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:  $X_1, \dots, X_{n1} \stackrel{iid}{\sim} N(\mu_1, \sigma_1^2)$  and Population 2:  $Y_1, \dots, Y_{n2} \stackrel{iid}{\sim} N(\mu_2, \sigma_2^2)$  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** :

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

**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  $H_0$  if  $p < \alpha$  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  $\mu$  may not be an integer, round down to the nearest integer.

### Example 2.5. Example of Arthritis Clinical Study

11:05 PM Sat 24 Jul

Example: <sup>关节炎</sup> Arthritis Clinical Study

A clinical trial is conducted to determine the effects of auranofin therapy on the treatment of rheumatoid arthritis. The initial level of pain was measured for each of 293 patients who suffer from rheumatoid arthritis and then each was randomly assigned to receive auranofin pills or a placebo pill (3 mg to be consumed twice daily). After six months, the level of pain is once again measured for each patient to see if there is an improvement.

Question of Interest: <sup>one-sided</sup> Does auranofin therapy help to improve arthritis pain? <sup>two-sided. Is there a difference between A or B that relates to ...?</sup>

- ▶ Define the experiment: identify the factor(s) and levels, response variable(s), etc.
- ▶ State the appropriate hypotheses to test the question of interest

7/24

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.  $\mu_A$  = mean difference in pain (before-after) for Auranofin users.  $\mu_P$  = 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  $\mu$  is a mean difference, not the pain level!

The R Output is following:

```
11:21 PM Sat 24 Jul
< >
T
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
< n_A n_P

> tapply(DiffinPain, Treatment, mean)
TAuranofin Placebo
0.6575342 0.3401361
< \bar{y}_A \bar{y}_P

> tapply(DiffinPain, Treatment, var)
TAuranofin Placebo
1.1784601 0.9383096
< s_A^2 s_P^2

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

> tapply(DiffinPain, Treatment, mean)
TAuranofin Placebo
0.6575342 0.3401361
< \bar{y}_A \bar{y}_P

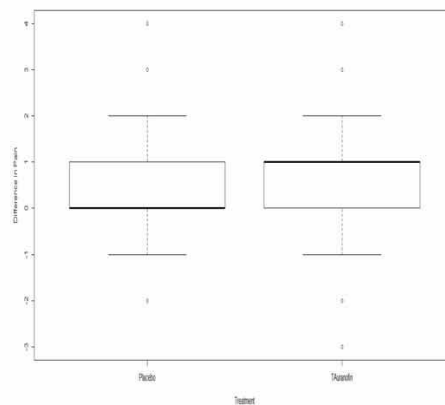
> tapply(DiffinPain, Treatment, var)
TAuranofin Placebo
1.1784601 0.9383096
< s_A^2 s_P^2

8/24
```



## Boxplots

```
# Side by Side boxplots by treatment  
> boxplot(DiffinPain ~ Treatment,  
          xlab="Treatment", ylab="Difference in Pain")
```



11:21 PM Sat 24 Jul 38%

## R Code: Pooled t-test

$\sigma_A^2 = \sigma_P^2$   
 $\text{response} \sim \text{factor}$   
 alternative = "greater"

```

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

# OR (same output as above)
2 > t.test(DiffinPain ~ Treatment, var.equal=TRUE)

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

95 percent confidence interval:
0.08086399 0.55393239

sample estimates:
mean of x mean of y
0.6575342 0.3401361
  
```

$\bar{y}_A$        $\bar{y}_P$

10/24

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

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

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

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

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.