Completely Randomized Designs (CRD) with One Factor

UBCO MDS — DATA 543

Office Hours until 17^{00} today. (13^{00} - 15^{00})

FIP 309 or keep going down the hall into the break room.

Key Concepts for today:

- Three basic principles of experimental design
- Definitions
- CRD with one factor
- Balanced vs. unbalanced
- $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$
- Pooled variance
- t-test + confidence interval comparing two treatments

Experimental Design

There are many types of experimental designs. The appropriate one to use depends upon the objectives of the experimentation.

While there are many factors that go into designing a good experiment, the three basic principles of experimental design are:

- Randomization
- Replication
- Reducing Noise (blocking)

i.e. the "Three Rs"

Randomization

Randomization is the random allocation of treatments to the experimental units.

The anti-typhoid inoculation example:

> *Sir Almroth Wright, a famous immunologist, compared the incidence of typhoid between those who volunteer and those who refused the inoculation. Karl Pearson (the other Father of Statistics) noted that a volunteer may be more particular about their own health, thus more likely to run a lower risk.*

The above example, does *not* randomize properly. Experiments that properly randomize assist in "averaging out" the effects of extraneous factors that may be present.

Replication

Replication: the repetition of a treatment within an experiment.

- Repetition of experimental conditions increases the accuracy of estimates which allows for estimates of the natural variation between experimental units.
- Knowledge of this variation will help generalize any relevant conclusions to similar subjects.

Reducing Noise (Blocking)

Reduce noise: control the conditions in the experiment as much as possible

- Blocking can often be used to control and adjust for some of the variation in experimental units.
- *Blocking* divides experimental units into subsets (blocks) so that units within the same block are more similar than units from different subsets or blocks.

So far we have discussed *survey sampling* which has some particular characteristics:

- Predominantly observational: record factors of explanatory variables and response variables and consider associations between them.
- We seek to learn about some characteristics of the finite population from the sample.

In contrast we will now consider *experimental design*

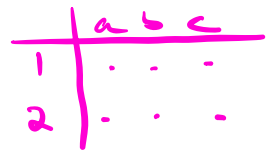
- Experiments actively manipulate the factors and evaluate their effects on the response variables. → only way to establish a causal relationship.
  - the factors whose effects on the response variable are of primary interest are referred to as *treatment/design factors*
- A population is (at least conceptually) infinite.

Experimental Plans

- Typically observational plans are cheaper and easier to conduct than experimental plans.
- However, we must use experimental plans (rather than observational plans) to investigate causal relationships.
- In other words, we use experimental plans to better understand the relationship between the manipulated inputs and the output.
- Note that there will always be varying inputs (not controlled in the experiment) that will vary as the experiment is conducted.

Experimental Design Recap

- The output variable in an experiment is also called the *response*.
- The input variables are referred to as *factors*, with different *levels* that can be controlled or set by the experimenter.



- A *treatment* is a combination of factor levels. **When there is only one factor, its levels are the treatments.**
- An *experimental unit* is a generic term that refers to a basic unit such as material, animal, person, plant, time period, or location, to which a treatment is applied.
- The process of choosing a treatment, applying it to an experiment unit, and obtaining the response is called an *experimental run*.

- Factors can be:

  quantitative e.g. dose of insulin in mg, price change %;
  qualitative e.g. colour, design 1 or design 2, . . .

- Response variates can be

  continuous e.g. profit at a store over a specified period.
  ordinal e.g. ranking on scale of 1 to 5
  count e.g. number of flaws in a roll of paper
  binary e.g. type I diabetes occurs within 5 years or not
  other (categorical, image, map, scatter plot . . .)

- For this module, we restrict attention to experimental plans where we treat the response variate (output) as continuous.

Five broad categories of experimental problems

See section 6.1 in Wu & Jiahua

1. **Treatment comparisons** Compare several treatments and select the best ones.
2. **Variable screening** Determining which factors can be dropped/crucial in influencing the output.
3. **Response surface exploration** find mathematical relationship between the values of these crucial factors and the output.
4. **System optimization** find the best possible setting of the input variables to achieve the desired output
5. **System robustness** if it is difficult to control input variables precisely we may be interested in learning which systems are more robust to deviations from these settings.

Amblyopia motivating example Researchers studying a common eye condition in children (amblyopia[1]) want to investigate the efficacy of 'occlusion therapy' (wearing an eye patch). In particular they are interested in whether prescribing 6 hours of occlusion therapy a day is more effective than prescribing 2 hours of occlusion a day.

How might we go about investigating this?

[1] AKA 'lazy eye': a disorder of sight in which the brain fails to process inputs from one eye, and over time favours the other eye

Experimental Plan Example

The most important first step is establishing *exactly* what question the researchers are trying to answer, and what limitations (if any) there are. Some questions we might ask:

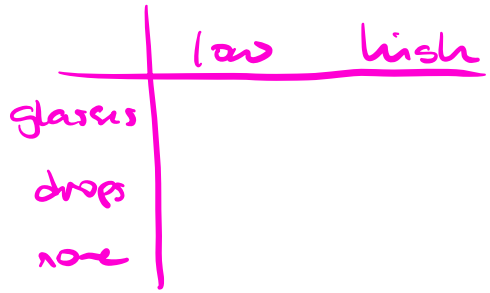
- What does 'more effective' mean? Response ?
- What population are we interested in?
- Are there any ethical concerns?
- What about non-compliance?

The list goes on, and is often a very difficult stage of any investigation.

Experimental Plan Example

- Suppose we recruit 120 patients for the study.
- Here we only have one factor: the prescribed dose having two levels (2 hours or 6 hours).
  - For convenience refer to these as the 'low dose regime' and the 'high dose regime'.
  - In this case we might randomize 60 patients to each treatment group.
  - Since we only have one factor, the levels are the treatments.

- N.B. if we had another factor in the form of additional therapy with levels: glasses, atropine drops or none, then we would have six combinations of the two variates, and so we might randomize the patients into six groups of 20.



Fundamental Principles

When designing an experimental plan, we should consider:

- *Random assignment*: assigning treatments to units using a probabilistic mechanism.
- *Replication*: applying each treatment to more than one unit.
- *Blocking (reducing noise)*: keeping one (or more) explanatory variates held fixed while different treatments are applied to units within that group. (We will discuss this next class)

Randomization

Randomization is applied to the assignments of treatments to reduce (balance out) the influence of unknown (lurking) variables.

- e.g. randomly assigning subjects to treatments assumes, on average, lurking variables will affect each treatment condition equally; so any significant differences between conditions can fairly be attributed to the independent variable.

Random assignment helps us:

- Reduce the risk of confounding by unknown explanatory variates.
- Generates an analysis method (we can use knowledge of the probabilistic mechanism to do so).

Randomization

Random assignment (also random allocation) is the process of applying treatments to experimental units using a probabilistic mechanism.

- e.g. in the amblyopia example, we randomly divided the patients into treatment groups.

Non-random assignment

- would be asking patients which treatment they'd prefer, or
- giving older patients the higher dose of 6 hours because they're more suited to longer doses.

Replication

- **Replication** involves applying each treatment to more than one unit in the sample.
  - e.g. in the amblyopia example, each of the two treatments were applied to 60 patients.

- Replication helps us to estimate the precision of our conclusions.

- You will usually encounter replication in studies. Exceptions occur in very small studies (where there may not be enough units available for all possible treatments), or if you find one treatment leads to high rates of drop out.

Replication

- In general, the outcomes of the response variable using replication will differ.
- This variation reflects the magnitude of *experimental error*.
- Remember, if you apply a treatment to one experimental unit, but measure the response variable 5 times, you do not have 5 replicates, you have 5 duplicates (aka repeated measurements).
- Duplication helps to reduce the *measurement error*, not experimental error.
- It is important to increase the number of replicates, if we intend to detect small treatment effects.

Fundamentals of Experimental Plans

- In the following lectures we will review some commonly used experimental designs.

- We'll consider two types of experimental design
  1. Completely randomized design (CRD) - no blocking is used.
  2. Randomized block design (RBD) - each treatment is repeated once in each block.

- For this lecture we will concentrate on experiments with one factor and two levels (i.e. treatments).

- We will make the necessary extensions in future lectures.

Lawson (2014)'s Chapter 2 bread example

In an experiment to determine the effect of **time to rise** on the **height of bread dough**, one homogeneous batch of bread dough would be divided into n loaf pans with an equal amount of dough in each. The pans of dough would then be divided randomly into t groups. Each group would be allowed to rise for a unique time, and the height of the risen dough would be measured and recorded for each loaf.

- Treatment factor: rise time
- Experimental unit: each loaf.
- Response: height.

Lawson (2014)'s Chapter 2 bread example

- Treatment factor: rise time,
- Experimental unit: individual loaf of bread,
- Response: the measured height.

Although other factors, such as temperature, are known to affect the height of the risen bread dough, they would be held constant and each loaf would be allowed to rise under the same conditions except for the differing rise times.

- **Replication:** r bread loaves are tested at each of the t rise times rather than a single loaf at each rise time.
- **Randomization:** Randomization would prevent lurking variables, such as variability in the yeast from loaf to loaf and trends in the measurement technique over time, from biasing the effect of the rise time.

In R

- Dividing n experimental units into t treatment groups, can be accomplished using base R commands.
- For example, in the bread rise experiment, if the experimenter wants to examine three different rise times (35 minutes, 40 minutes, and 45 minutes) and test four replicate loaves of bread at each rise time, the following code will create the list ...

```
> set.seed(7638)
> f <- factor( rep( c(35, 40, 45 ), each = 4))
> fac <- sample( f, 12 )
> eu <- 1:12
> (plan <- data.frame( loaf=eu, time=fac))
   loaf time
1     1   40
2     2   40
3     3   35
4     4   45
5     5   35
6     6   45
7     7   35
8     8   45
9     9   40
10   10   40
11   11   35
12   12   45
```

Completely Randomized Design with one factor

- The simplest experimental design, is a completely randomized design (CRD).
- In CRD with one treatment factor, n experimental units are divided randomly into t groups.
- Each group is then subject to one of the unique levels or values of the treatment factor.
- If n is a multiple of t, i.e. $n = tr$, then each level of the factor will be applied to r unique experimental units (r replicates)
- All other known independent variables are held constant so that they will not bias the effects.

Completely Randomized Design

- CRDs should be used when there is only one factor under study and the experimental units are homogeneous.

- We say the design is *balanced* if there is an equal number of units receiving each treatment.

- We usually have two main questions of interest:
  1. Are there any differences between treatments?
  2. Which treatment is the best?

A Balanced Completely Randomized Design

Assuming a **balanced plan**, we observe a response y_{ij} for the j^{th} unit receiving treatment i, and write

baseline mean

$$Y_{ij} = \mu + \tau_i + R_{ij}$$

treatment effect

where

- $i = 1, 2, \ldots, t$; (# of treatments)
- $j = 1, 2, \ldots, r$, (# of replicates)
- r is the number of replicates for each treatment, and
- R_{ij} are the experimental errors.

The distribution of the experimental errors, R_{ij}, are mutually independent due to the randomization and assumed to be normally distributed with mean of 0 and variance σ^2. $R_{ij} \sim N(0, \sigma^2)$

An unbalanced Completely Randomized Design

Assuming an **unbalanced plan**, we observe a response y_{ij} for the j^{th} unit receiving treatment i, and write

$$Y_{ij} = \mu + \tau_i + R_{ij} \tag{1}$$

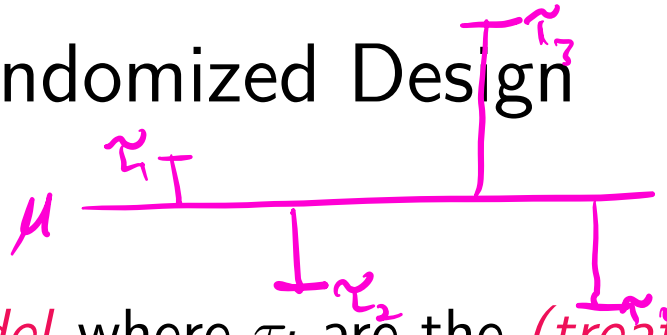
for

- $i = 1, 2, \ldots, t$;
- $j = 1, 2, \ldots, n_i$, and
- $R_{ij} \sim N(0, \sigma^2)$;

n_i: # of replicates in treatment i

A Balanced Completely Randomized Design

$$Y_{ij} = \mu + \tau_i + R_{ij}$$



(1) referred to as the *effects model* where τ_i are the *(treatment) effects* representing the difference between the long-run average of all possible experiments at the ith level of the treatment factor and the overall average.

- μ: mean response across all treatments.
- τ_i: the i^{th} treatment effect.
- The balanced design constraint is $\sum_{i=1}^{t} \tau_i = 0$

.

N.B. the unbalanced constraint is $\sum_{i=1}^{t} n_i \tau_i = 0$ (more on this later)

ABC drug example

Example: suppose we're interested in comparing three drugs—suppose they're named Amazalin (A), Booyahathol (B), and Coolocadine (C)—to treat a disease, and we have 60 patients with which to investigate.

- We might consider dividing our 60 patients into three groups of 20, and giving one drug to patients in each group.
- We could then compare responses (e.g., severity of symptoms) from each group.

Treatment		
Amazalin (A)	Booyahathol (B)	Coolocadine (C)
20	20	20

Completely Randomized Design

In our ABC drug example we have $t = 3$, $r = 20$, and

- μ: mean response across all treatments.
- $\mu + \tau_1$: mean response for drug A. $\mu_A = \mu + \tau_A$
- $\mu + \tau_2$: mean response for drug B.
- $\mu + \tau_3$: mean response for drug C.

Important: τ_i is the *effect* of treatment i *relative* to the overall mean μ.

- e.g τ_1 represents the increase (or decrease) from the overall average response if we use drug A

Alternative representation of CRD model

An alternate way of writing a model for the data is

$$Y_{ij} = \mu_i + R_{ij} \tag{2}$$

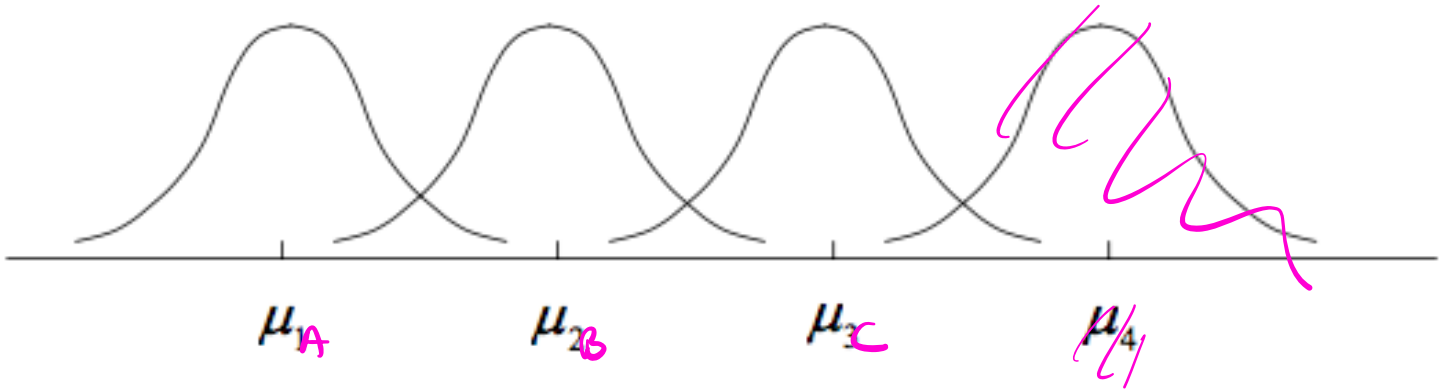
where Y_{ij} is the response for the jth experimental unit subject to the ith level of the treatment factor, $i = 1, \ldots, t$, $j = 1, \ldots, n_i$, and n_i is the number of experimental units or replications in ith level of the treatment factor.

This is sometimes called the cell *means model* or *normal model* with a different mean, μ_i, for each level of the treatment factor.

Graphical representation of CRD model

Lawson (2014) Figure 2.2

Figure 2.2 *Cell Means Model*



Comparing Two Treatments (Without Blocking)

Let's revisit the Amblyopia Example on page 11. We focus only on whether a low or high dose of patching is prescribed, i.e. we have one factor with two levels (in this case our levels=treatments).

Let's divide our 120 patients into two groups of 60:

- low dose (2 hours) and
- high dose (6 hours).

After treatment, we discover that

- the low dose group has an average improvement of 0.061 with a standard deviation of 0.086,
- the high dose group has an average improvement of 0.172 with a standard deviation of 0.183.

- μ: is the mean visual acuity across our two treatment groups (low and high dose).
- τ_1, τ_2: is the increase (or decrease) from the overall average response when using a low or high dose of patching, respectively.
- Our experiment aims to answer *Is there a significant difference between the two treatments?*
- This is equivalent to testing the following statistical hypothesis:

$$H_0 : \mu_1 = \mu_2 \quad \mu_1 - \mu_2 = 0 \quad H_1 : \mu_1 \neq \mu_2$$
$$H_0 : \tau_1 - \tau_2 = 0 \qquad\qquad H_1 : \tau_1 - \tau_2 \neq 0$$
$$\tau_1 = \tau_2$$

- Follow-up: if the answer is yes, which treatment is preferable?
- To test this let us first go over some notation ...

We will use the notation:

- $\overline{Y}_{i\cdot}$ to denote the sample mean of the ith treatment
- $\overline{Y}_{\cdot\cdot}$ to denote the overall sample mean (some sources may use $\overline{Y}_{++}$ for $\overline{Y}_{\cdot\cdot}$ instead).

$$\overline{Y}_{i\cdot} = \frac{1}{n_i}\sum_{i=1}^{n_i} Y_{ij}$$

$$\overline{Y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{t}\sum_{j=1}^{n_i} Y_{ij} \qquad \text{where } n = \sum_{i=1}^{t} n_i$$

For equal number of replicates,

$$\hat{\mu} = \overline{y}_{\cdot\cdot} = \frac{1}{t}\sum_{i=1}^{t}\overline{y}_{i\cdot}$$

A Model for Comparing Two Treatments

For this model, if we consider the group averages

$$E\left[\overline{Y}_{1.}\right] = E\left[\frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j}\right] = \mu + \tau_1$$

$$E\left[\overline{Y}_{2.}\right] = E\left[\frac{1}{n_2}\sum_{j=1}^{n_1} Y_{1j}\right] = \mu + \tau_2$$

- $\mu + \tau_1$: the mean in visual acuity using a low dose regime.
- $\mu + \tau_2$: the mean in visual acuity using a high dose regime.

If we're interested in comparing the low dose regime with the high dose regime, the parameter of interest is

$$(\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$$

hypothesis

Note that we don't care about μ!

For this model, if we consider the overall average

$$\begin{aligned}
\overline{Y}_{..} &= \frac{1}{n_1+n_2}\sum_{i=1}^{2}\sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_1+n_2}\left[\sum_{j=1}^{n_1} Y_{1j} + \sum_{j=1}^{n_2} Y_{2j}\right] \\
&= \frac{1}{n_1+n_2}\left[\frac{n_1}{n_1}\sum_{j=1}^{n_1} Y_{1j} + \frac{n_2}{n_2}\sum_{j=1}^{n_2} Y_{2j}\right] \\
&= \frac{n_1}{n_1+n_2}\overline{Y}_{1j} + \frac{n_2}{n_1+n_2}\overline{Y}_{2j}
\end{aligned}$$

$$\begin{aligned}
E\left[\overline{Y}_{..}\right] &= \frac{n_1}{n_1+n_2}(\mu+\tau_1) + \frac{n_2}{n_1+n_2}(\mu+\tau_2) \\
&= \mu + \frac{n_1}{n_1+n_2}\tau_1 + \frac{n_2}{n_1+n_2}\tau_2
\end{aligned}$$

A Model for Comparing Two Treatments

For μ to represent the overall mean we require

$$\frac{n_1}{n_1+n_2}\tau_1 + \frac{n_2}{n_1+n_2}\tau_2 = 0$$

or equivalently

$$n_1\tau_1 + n_2\tau_2 = 0 \qquad \sum n_i \tau_i = 0$$

We have written the model for two treatments in an unnecessarily complicated way with a constraint because we will need to do so when we consider models for experimental plans with many treatments.

A Model for Comparing Two Treatments

Some final things to note about the model:

$$Y_{ij} = \mu + \tau_i + R_{ij} \qquad i = 1, 2, \quad j = 1, ..., n_i$$

- Inference about μ is not of interest to us.
- Inference about the treatments effects typically involves comparing τ_i to zero.
- We also apply the constraint $n_1\tau_1 + n_2\tau_2 = 0$ for proper interpretation of the parameters.

Comparing Two Treatments Estimation

We observe responses (improvements in visual acuity)

- $y_{11}, y_{12}, \ldots, y_{1n_1}$ in the low dose group, and
- $y_{21}, y_{22}, \ldots, y_{2n_2}$ in the high dose group.

We use least squares to obtain point estimates by solving the following using a Lagrange Multiplier:

$$\min_{\mu,\tau_1,\tau_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \mu - \tau_i)^2 \quad \text{such that } n_1\tau_1 + n_2\tau_2 = 0$$

Comparing Two Treatments Estimation

This gives us:

- the overall sample average:

$$\hat{\mu} = \overline{y}_{..} = \frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij}$$

- the sample average for treatment i minus the overall average:

$$\hat{\tau}_i = \overline{y}_{i.} - \overline{y}_{..} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - \frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} y_{ij}$$

- the estimated difference in treatment effects:

$$\hat{\tau}_1 - \hat{\tau}_2 = \overline{y}_{1.} - \overline{y}_{2.}$$

For our example, we were told that $n_1 = n_2 = 60$,

- the sample average for the low dose group was 0.061 and the sample average for the high dose group was 0.172.
- The overall sample average improvement in visual acuity is

$$\hat{\mu} = \overline{y}_{..} = (\overline{y}_{1.} + \overline{y}_{2.})/2 = (0.061 + 0.172)/2 = 0.1165$$

0.117

The estimated treatment effect for the low dose group is

$$\hat{\tau}_1 = \overline{y}_{1.} - \overline{y}_{..} = 0.061 - 0.117 = -0.0555$$

high

The estimated treatment effect for the ~~low~~ dose group is

$$\hat{\tau}_2 = \overline{y}_{2.} - \overline{y}_{..} = 0.172 - 0.117 = 0.055$$

The estimated difference in treatment effects is

$$\hat{\tau}_1 - \hat{\tau}_2 = \overline{y}_{1.} - \overline{y}_{2.} = 0.061 - 0.172 = -0.111$$

Comparing Two Treatments Estimators

The estimators corresponding to these various estimates have the following distributional properties:

$$Y_{ij} \sim N(\mu + \tau_i, \sigma^2)$$
$$\overline{Y}_{i\cdot} \sim N(\mu + \tau_i, \sigma^2/n_i)$$

and so

$$\tilde{\tau}_1 - \tilde{\tau}_2 = \overline{Y}_{1\cdot} - \overline{Y}_{2\cdot} \sim N\left(\tau_1 - \tau_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$$

Note

- From the properties on the previous slide, it is clear that the larger the sample size n_i, the smaller the variance of the point estimator $\hat{\mu}_i$.

- In other words, replications reduce the experimental error and ensure better point estimates for the unknown parameters and consequently more reliable test for the hypothesis.

- As we've seen before, these distributions rely on σ which is typically unknown. Therefore we need a means of estimating it...

Variance Estimate

It can be shown that the sample variance within treatment i is:

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{i\cdot})^2$$

The so-called pooled overall sample variance is:

$$\hat{\sigma}^2 = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2$$

$$\hat{\sigma}^2_{pooled} = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

- Note that the pooled variance (sometimes denoted $\hat{\sigma}_p^2$ or s_p^2) is not the same as the usual sample variance.
- The pooled variance assumes $\sigma_1 = \sigma_2$ and when this is true, it leads to a more precise estimate of σ^2.

Comparing Two Treatments Estimators

- It can be shown that $\frac{(n_i - 1)\tilde{\sigma}_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}$
- Based on that result we have:

$$\frac{(n_1 - 1)\tilde{\sigma}_1^2}{\sigma^2} + \frac{(n_2 - 1)\tilde{\sigma}_2^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}$$

- Lastly: $\hat{\sigma}^2 = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$, and so

$$\frac{(n_1 + n_2 - 2)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}$$

Comparing Two Treatments

Putting all this together, we have our test statistic:

$$\frac{(\hat{\tau}_1 - \hat{\tau}_2) - (\tau_1 - \tau_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \tag{3}$$

where $\hat{\sigma}$ is the square root of the pooled variance estimator

Leading to the familiar form of the $(1-\alpha) \times 100\%$ confidence interval as

$$\hat{\tau}_1 - \hat{\tau}_2 \pm c \times \text{s.e.}(\hat{\tau}_1 - \hat{\tau}_2)$$

where c is chosen such that $P(|T_{n_1+n_2-2}| \leq c) = 1 - \alpha$, and $\text{s.e.}(\hat{\tau}_1 - \hat{\tau}_2) = \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

Question 1

What is the estimated size of the difference between the two treatments? How precisely have we estimated this difference?

Comparing Two Treatments Example CI

From the amblyopia example, we have

- $n_1 = n_2 = 60$, $\hat{\tau}_1 - \hat{\tau}_2 = -0.112$, $\hat{\sigma}_1 = 0.086$, $\hat{\sigma}_2 = 0.183$
- $\hat{\sigma} = 0.143$ (the $\sqrt{\text{pooled variance}}$)
- $P(|T_{118}| < 1.980) = 0.95$, so $c = 1.980$

vs. $z^* = 1.960$

$$\hat{\tau}_1 - \hat{\tau}_2 \pm c \times \hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = -0.112 \pm 0.0052$$

$$= [-0.163, -0.060]$$

-0.163 -0.060 0

negative: $\mu_1 < \mu_2$

This 95% confidence interval for the *difference* in improvement in visual acuity between the two treatments suggests the high dose group gets better results.

Question 2

Is there any evidence of a difference between the two treatments?

To answer this, we need can carry out a formal test of significance (also called a hypothesis test).

A formal Hypothesis test for comparing two treatments

1. State the null and alternative hypothesis:

$$H_0 : \mu_1 = \mu_2 \qquad H_1 : \mu_1 \neq \mu_2$$
$$H_0 : \tau_1 - \tau_2 = 0 \qquad H_1 : \tau_1 - \tau_2 \neq 0$$

2. Calculated the observed test statistic.

$$t_{obs} = \frac{|\hat{\tau}_1 - \hat{\tau}_2|}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 4.282$$

From (3), we know that $\frac{(\hat{\tau}_1 - \hat{\tau}_2) - (\tau_1 - \tau_2)}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$. We can use this knowledge to calculate a p-value ...

A formal Hypothesis test for comparing two treatments

3. Calculate a p-value

Recall (from DATA 570) that the p-value is the probability of observing a test statistic as extreme or more extreme than that which we observed, assuming the null hypothesis is true.

$$\begin{aligned} \text{Hence the } p\text{-value} = p &= 2 \times P(|t_{118}| \geq t_{obs}) \\ &= 2 \times P(|t_{118}| \geq 4.282) \\ &= 3.79 \times 10^{-5} \end{aligned}$$

Therefore there is significant statistical evidence of a difference between the two treatments.

A formal Hypothesis test for comparing two treatments

- We can carry out the previous calculations using the `pt` function:

```
> print(c(t,n1,n2)) # t = t_obs
[1]  4.282223 60.000000 60.000000
> (p.val <- 2*(1 - pt(t,n1 + n2 - 2))) # p-value
[1] 3.790119e-05
```

- However, it's much easier to just use the *t*-test command in R.
- Suppose that the observations are stored in variables named `Low` and `High` for the low dose and high dose groups, respectively.

A formal Hypothesis test for comparing two treatments

```
# var.equal = FALSE by default
> t.test(Low, High, var.equal = TRUE)
Two Sample t-test

data:  Low and High
t = -4.2822, df = 118, p-value = 3.79e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.16320751 -0.05999137
sample estimates:
 mean of x  mean of y
0.06076164 0.17236108
```

A formal Hypothesis test for comparing two treatments

4. Interpret the result.

- As we did in Data 570, we compare the *p*-value with our significance level α.
- If $p \leq \alpha$, we reject the null hypothesis, otherwise, we *fail to reject* the H_0.
- The default value for α is 0.05 (corresponding to a default 95% confidence interval)
- Since the *p*-value is so small, there is strong evidence against the null hypothesis.
- That is to say, at a significant level of 0.05, there is strong evidence to suggest that $\tau_1 - \tau_2 \neq 0$.

Comparing Two Treatments

Two assumptions required for validity of the analysis based on the linear model:

1. constant variance of the experimental error, σ^2, across all levels of the treatment factor, and
2. normality of the experimental errors.

Verifying these assumptions will be similar to how we did it in Data 570:

- Equality of variances.
  - informal: plot the data, or
  - formal: perform a hypothesis test of $H_0 : \sigma_1^2 = \sigma_2^2$ (e.g., Levene's Test).
- Normality of the errors.
  - check this with a QQ-plot of $r_{ij} = y_{ij} - \overline{y}_{i\cdot}$.

Extra: Welch's t-test

If it is **not** reasonable to assume equal variance we may use the Welch's t-test which has the test statistic:

$$\frac{(\tilde{\tau}_1 - \tilde{\tau}_2) - (\tau_1 - \tau_2)}{\sqrt{\frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_2^2}{n_2}}} \underset{\sim}{\text{approx}} t_k$$

where

$$k = \frac{\left(\frac{\tilde{\sigma}_1^2}{n_1} + \frac{\tilde{\sigma}_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{\tilde{\sigma}_1^2}{n_1}\right) + \frac{1}{n_2 - 1}\left(\frac{\tilde{\sigma}_2^2}{n_2}\right)}$$

R command: `t.test(..., var.equal = FALSE)`

References I

Lawson, J. (2014), *Design and Analysis of Experiments with R*, Vol. 115.