

STAT 332: Sampling and Experimental Design

Professor Riley Metzger
L^AT_EXer Iris Jiang

Spring 2020

Contents

1	PPDAC	2
1.1	Problem	2
1.2	Plan	2
1.3	Data	2
1.4	Analysis	2
1.5	Conclusion	2
1.6	Errors	3
2	Models	3
2.1	Model I	3
2.2	Independent vs. Dependent Groups	3
2.2.1	Ways of Creating Dependency	3
2.3	Model 2A	4
2.4	Model 2B	4
2.5	Model 3	5
2.6	Model 4	5
3	Maximum Likelihood Estimation (MLE)	6
3.1	What is it?	6
3.2	How does it work?	6
3.3	What is the process?	6
3.4	Example	6
4	Least Squares	7
4.1	What is it?	7
4.2	How?	8
4.3	Models	8
4.4	Process	8
4.5	Example	8
5	Estimators	9
5.1	Example	10
6	Sigma	11
6.1	Example	11
7	Confidence Interval	11
7.1	Model 1 Example	12
7.2	Model 2A Example	12
7.3	Model 2B Example	13
7.4	Model 3 Example	13
7.5	Model 4 Example	13
8	Hypothesis Testing	13

9 Models Cont.	14
9.1 Model 5: Completely Randomized Design (CRD)	14
9.1.1 Example	14
9.2 Least Square	15
9.3 Estimators	15
9.4 Confidence Interval	16
10 Analysis of Variance (ANOVA)	16
11 Model 6	17
11.1 Least Square	17
12 Model 7	17
12.1 Least Square	17

1 PPDAC

Problem, Plan, Data, Analysis, Conclusion

1.1 Problem

Define the problem:

- Target Population (T.P.): The group of units referred to in the problem step
- Response: The answer provided by the T.P. to the problem
- Attribute: statistic of the response

What is the average grade of students in STAT 101?

Solution.

- T.P.: All STAT 101 students
- Response: Grade of a STAT 101 student
- Attribute: Average grade

□

1.2 Plan

How?

- Study population (S.P.): The set of units you **can** study Problem: Does a drug reduce hair loss

Solution. You can not use untested drug directly on people out of ethical concerns

T.P.: People

S.P.: Mice

□

- Sample: A subset of the study population

1.3 Data

Collect the data, according to the plan.

1.4 Analysis

Analyse the data.

1.5 Conclusion

Refers back to the problem.

1.6 Errors

- Study Error: The attribute of the T.P. differs from the parameter of the S.P. $a(T.P.) - \mu$
- Sample Error: The parameter differs from the sample statistic (estimate). $\mu - \bar{x}$
- Measurement Error: The difference between what we want to calculate and what we do calculate.

2 Models

Definition 2.1 (Model). A model relates a parameter to a response.

2.1 Model I

$$Y_j = \mu + R_j, R_j \sim N(0, \sigma^2)$$

- y_j : The response of unit j , it is random.
- μ : S.P. mean, it is not random and it is unknown
- R_j : The distribution of responses about μ

Note.

1. R_j 's are always independent.
2. Gaus's Theorem: Any Linear combination of normal R.V.s is normal
3. $Y_j \sim N(\mu, \sigma^2)$,

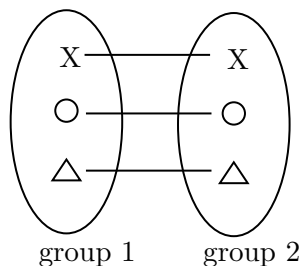
$$E(Y_j) = E(\mu + R_j) = E(\mu) + \mu + 0 = \mu$$

$$V(Y_j) = V(\mu + R_j) = V(R_j) = \sigma^2$$

Average grade of STAT 101: $Y_j = \mu + R_j, R_j \sim N(0, \sigma^2)$

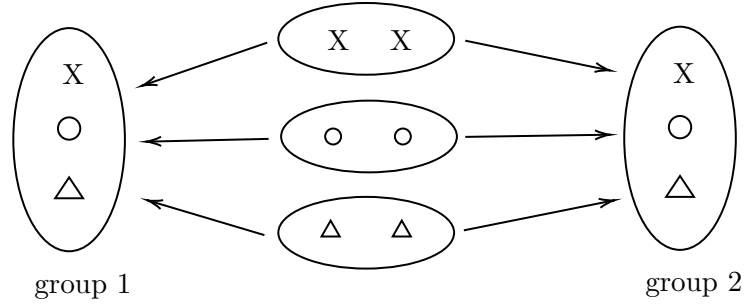
2.2 Independent vs. Dependent Groups

Definition 2.2 (Dependent). We randomly select one group and we find a match, having the same explanatory variates, for each unit of the first group.

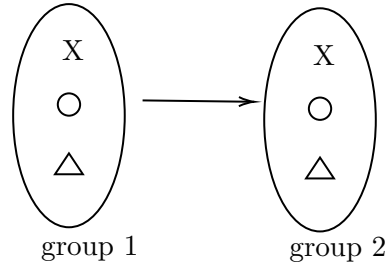


2.2.1 Ways of Creating Dependency

- Twins



- Reuse



Definition 2.3 (Independent). Are formed when we select units at random from mutually exclusive groups.

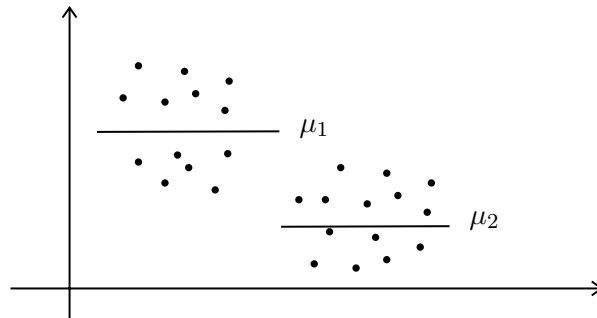
- No relationship between chosen groups Broken parts and non-broken parts

2.3 Model 2A

Independent groups where we assume the groups have the same standard deviation.

$$Y_{ij} = \mu_i + R_{ij}, R_{ij} \sim (0, \sigma^2)$$

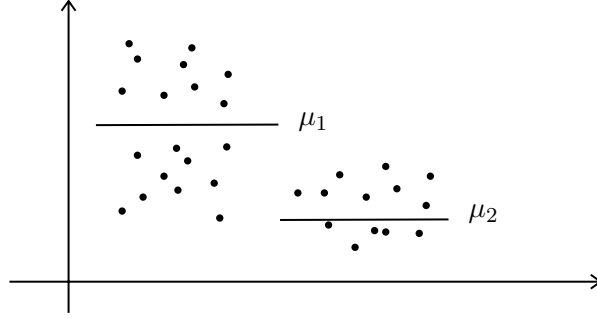
- Y_{ij} : Response of unit j in group i
- μ_i : Mean for group i ; not random; unknown
- R_{ij} : The distribution of responses about μ_i



2.4 Model 2B

Independent groups but $\sigma_1 \neq \sigma_2$

$$Y_{ij} = \mu_i + R_{ij}, R_{ij} \sim N(0, \sigma_i^2)$$



2.5 Model 3

Lets construct two groups using twins and get two groups. Set group 1:

$$y_{1j} = \mu_1 + R_{1j}$$

and group 2:

$$y_{2j} = \mu_2 + R_{2j}$$

and we subtract them:

$$y_{1j} - y_{2j} = \mu_1 - \mu_2 + R_{1j} - R_{2j}$$

Let $y_{dj} = y_{1j} - y_{2j}$, $\mu_d = \mu_1 - \mu_2$ and $R_{dj} = R_{1j} - R_{2j}$. Then we get a new model:

$$y_{dj} = \mu_d + R_{dj}, R_{dj} \sim N(0, \sigma_d^2)$$

heart rate before exercise	heart rate after exercise	difference (d)
70	80	10
80	100	20
90	90	0

$y_{dj} = \mu_d + R_{dj}$, $R_{dj} \sim N(0, \sigma_d^2)$ studies the difference.

2.6 Model 4

Recall:

$$Y \sim \text{Bin}(n, \pi)$$

- n outcomes
- each outcome is binary

$$E(Y) = n\pi, \text{Var}((Y)) = n\pi(1 - \pi)$$

By the Central Limit Theorem

$$Y \sim N(n\pi, n\pi(1 - \pi))$$

The proportion is $\frac{Y}{n} \sim N(\pi, \frac{\pi(1-\pi)}{n})$

$$E(\frac{Y}{n} = \frac{E(Y)}{n}) = \pi, \text{Var}((\frac{Y}{n})) = \frac{\text{Var}((Y))}{n^2} = \frac{\pi(1 - \pi)}{n}$$

3 Maximum Likelihood Estimation (MLE)

3.1 What is it?

It connects the population parameter (θ) to the sample statistic ($\hat{\theta}$).

3.2 How does it work?

It choose the most probable value of θ given our data y_1, y_2, \dots, y_n

3.3 What is the process?

1. Define likelihood function

$$L = f(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n)$$

We assume that $Y_i \perp Y_j, \forall i \neq j$

$$L = f(Y_1 = y_1)f(Y_2 = y_2) \cdots f(Y_n = y_n)$$

2. Define log likelihood function

$$l = \ln(L)$$

use log rules to clean it up

3. Find $\frac{\partial l}{\partial \theta}$ for all θ
4. Set $\frac{\partial l}{\partial \theta} = 0$ and solve for $\hat{\theta}$

3.4 Example

Consider $Y_{ij} = \mu_i + R_{ij}$ (Model 2A), Estimate using MLE, μ_1, μ_2, σ , assuming our group sizes are n_1 and n_2 ; $n = n_1 + n_2$.

Note the fact $R_{ij} \sim N(0, \sigma^2)$, hence $Y_{ij} \sim N(\mu_i, \sigma^2)$

Recall the pdf of a normal distribution:

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right)$$

1. Define likelihood function

$$\begin{aligned} L &= \prod_{ij} f(j_{ij}) = \prod_{j=1}^{n_1} f(y_{1j}) \prod_{j=1}^{n_2} f(y_{2j}) \\ &= \prod_{ij}^{n_1} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_{1j} - \mu_1)^2}{2\sigma^2}\right) \prod_{ij}^{n_2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_{2j} - \mu_2)^2}{2\sigma^2}\right) \\ &= (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left(-\frac{\sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2}{2\sigma^2}\right) \exp\left(-\frac{\sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2}{2\sigma^2}\right) \end{aligned}$$

2. Define log likelihood function

$$l = -\frac{n}{2} \ln(2\pi) - n \ln(\sigma) - \frac{\sum_{j=1}^{n_1} (y_{1j} - \mu_1)^2}{2\sigma^2} - \frac{\sum_{j=1}^{n_2} (y_{2j} - \mu_2)^2}{2\sigma^2}$$

3. Find $\frac{\partial l}{\partial \mu_1}$, $\frac{\partial l}{\partial \mu_2}$ and $\frac{\partial l}{\partial \sigma}$. And set them to be 0

$$\begin{aligned} \frac{\partial l}{\partial \hat{\mu}_1} &= \frac{2 \sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)}{2\hat{\sigma}^2} = 0 \\ \Rightarrow \sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1) &= 0 \\ n_1 \bar{y}_1 - n_1 \hat{\mu}_1 &= 0 \\ \Rightarrow \hat{\mu}_1 &= \bar{y}_1 \end{aligned}$$

The estimate of population average is the sample average

By symmetry, $\hat{\mu}_2 = \bar{y}_2$

$$\begin{aligned} \frac{\partial l}{\partial \hat{\sigma}} &= -\frac{n}{\hat{\sigma}} - \frac{\sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)^2}{2} (-2\hat{\sigma}^{-3}) - \frac{\sum_{j=1}^{n_2} (y_{2j} - \hat{\mu}_2)^2}{2} (-2\hat{\sigma}^{-3}) = 0 \\ \Rightarrow -n\hat{\sigma}^2 + \sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \hat{\mu}_2)^2 &= 0 \\ \hat{\sigma}^2 &= \frac{\sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \hat{\mu}_2)^2}{n} \end{aligned}$$

MLE doesn't necessarily give you something unbiased, LSM however is generally unbiased if the error term is normal.

The above $\hat{\sigma}^2$ is biased, we will need some twit to make it unbiased.

Let

$$\hat{\sigma}^2 = \frac{\sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \hat{\mu}_2)^2}{n_1 + n_2 - 2}$$

Recall: An estimator for θ is unbiased if $E(\tilde{\theta}) = \theta$.

We can rewrite it another way:

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\frac{n_1-1}{n_1-1} \sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)^2 + \frac{n_2-1}{n_2-1} \sum_{j=1}^{n_2} (y_{2j} - \hat{\mu}_2)^2}{n_1 + n_2 - 2} \\ &= \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = s_p^2 \end{aligned}$$

4 Least Squares

4.1 What is it?

Another technique to find $\hat{\theta}$

4.2 How?

It minimizes the residuals.

4.3 Models

Response = Deterministic Part + Random Part

$$y = f(\theta) + R$$

Let y_1, y_2, \dots, y_n be realizations of y . Let $\hat{y}_i = f(\hat{\theta})$, where $f(\hat{\theta})$ is simply $f(\theta)$ with θ replaced by $\hat{\theta}$. We call \hat{y}_i our “prediction”.

A residual is

$$r_i = y_i - f(\hat{\theta}) = y_i - \hat{y}_i$$

4.4 Process

1. Define the w function $w = \sum r^2$
2. Calculate $\frac{\partial w}{\partial \theta}$ for all non- σ parameters
3. Set $\frac{\partial w}{\partial \theta} = 0$ and replace θ by $\hat{\theta}$
4. Solve for $\hat{\theta}$

4.5 Example

Consider Model 2A, $y_{ij} = \mu_i + R_{ij}$

- y_{ij} : response
- μ_i : deterministic part
- R_{ij} : random part

Let $n = n_1 + n_2$

$$\begin{aligned} w &= \sum_{ij} r_{ij}^2 = \sum_{ij} (y_{ij} - \hat{\mu}_i)^2 \\ &= \sum_{j=1}^2 \sum_{i=1}^2 (y_{ij} - \hat{\mu}_i)^2 \\ &= \sum_j^{n_1} (y_{1j} - \hat{\mu}_1)^2 + \sum_j^{n_2} (y_{2j} - \hat{\mu}_2)^2 \\ \frac{\partial w}{\partial \hat{\mu}_1} &= \sum_j^{n_1} (y_{1j} - \hat{\mu}_1)(-2) = 0 \\ \implies \hat{\mu}_1 &= \bar{y}_1 \end{aligned}$$

By symmetry, $\hat{\mu}_2 = \bar{y}_2$

Note.

1. $\hat{\sigma}^2$ is always of the form

$$\hat{\sigma}^2 = \frac{w}{n - q + c}$$

- n : number of units (sample size)
- q : number of non- σ parameters
- c : number of constraints

In the example, $\hat{\sigma}^2 = \frac{\sum_j^{n_1} (y_{1j} - \hat{\mu}_1)^2 + \sum_j^{n_2} (y_{2j} - \hat{\mu}_2)^2}{n_1 + n_2 - 2}$, we can further show $s_p^2 = \frac{s_1^2(n_1-1) + s_2^2(n_2-1)}{n_1 + n_2 - 2}$

2. MLE vs. LS

- LS:
 - is from 1860's (older technique)
 - LS is unbiased provided R_j is normally distributed
- MLE:
 - recent technique
 - much more flexible - it does NOT need R_j to be normal

3. Minimum? We can assume LS provides a minimum second derivative.

5 Estimators

Our sample data is y_1, y_2, \dots, y_n . It is not random. It is a realization of a r.v. Y_1, Y_2, \dots, Y_n . A statistic is a function of the sample data; $\hat{\theta}$, is not random. But if y_1, y_2, \dots, y_n changes, so does $\hat{\theta}$.

For that reason you can think of $\hat{\theta}$ as the realization of a r.v. $\tilde{\theta}$, called an **estimator**. To move from $\hat{\theta}$ to $\tilde{\theta}$, we capitalize our y_i 's.

Example 5.1. Model 2A:

$$\underbrace{\hat{\mu} = \bar{y}_1}_{\text{statistic}} \rightarrow \underbrace{\tilde{\mu}_1 = \bar{Y}_1}_{\text{estimator}}$$

Theorem 5.1 (Gaus). Any linear combination of normal r.v.'s is still normal.

Let $X \sim N(\mu_x, \sigma_x^2)$

Let $Y \sim N(\mu_y, \sigma_y^2)$

Let $X \perp Y$

Let a, b, c be constants, $a, b \neq 0$.

Let $L = ax + by = c$.

Then $L \sim N(E(L), \text{Var}(L))$

Theorem 5.2 (Central Limit Theorem). Let Y_1, \dots, Y_n be a sequence of r.v.'s.

Let $E(Y_i) = \mu, \forall i$.

Let $\text{Var}(Y_i) = \sigma^2 < \infty, \forall i$

Let $Y_i \perp Y_j, \forall i \neq j$

Then $\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$

5.1 Example

Model 2A: $Y_{ij} = \mu_i + R_{ij}, R_{ij} \sim N(0, \sigma^2)$. What is the distribution of $\tilde{\mu}_1$?
Using LS or MLE we get

$$\hat{\mu}_1 = \bar{y}_1$$

Our corresponding estimator is

$$\tilde{\mu}_1 = \bar{Y}_1 = \frac{\sum_{j=1}^n Y_{1j}}{n_1}$$

Thus by Gaus theorem, it is normal.

$$\begin{aligned} E(\tilde{\mu}_1) &= E(\bar{Y}_1) = E\left(\frac{\sum_{j=1}^n Y_{1j}}{n_1}\right) \\ &= \sum_{j=1}^n \frac{E(Y_{1j})}{n_1} \text{ (sum rule)} \\ &= \sum_{j=1}^n \frac{E(\mu_1 + R_{1j})}{n_1} \\ &= \sum_{j=1}^n \frac{\mu_1 + E(R_{1j})}{n_1} \text{ (sum rule)} \\ &= \mu_1 \end{aligned}$$

This is an unbiased estimator: $E(\tilde{\theta}) = \theta \implies \tilde{\theta}$ is an unbiased estimator of θ

$$\begin{aligned} \text{Var}(\tilde{\mu}_1) &= \text{Var}(\bar{Y}_1) = \text{Var}\left(\frac{\sum_{j=1}^n Y_{1j}}{n_1}\right) \\ &= \frac{1}{n_1^2} \text{Var}\left(\sum_{j=1}^n Y_{1j}\right) \\ &= \frac{1}{n_1^2} \sum_{j=1}^n \text{Var}(Y_{1j}) \text{ (since } Y_{1j} \perp Y_{1i}, \forall i \neq j) \\ &= \frac{1}{n_1^2} \sum_{j=1}^n \text{Var}(\mu_1 + R_{1j}) \\ &= \frac{1}{n_1^2} \sum_{j=1}^n \text{Var}(R_{1j}) = \frac{\sigma^2}{n_1} \end{aligned}$$

Therefore, $\tilde{\mu}_1 \sim N(\mu_1, \frac{\sigma^2}{n_1})$, and by symmetry $\tilde{\mu}_2 \sim N(\mu_2, \frac{\sigma^2}{n_2})$

6 Sigma

Theorem 6.1. Let $Z \sim N(0, 1)$, then $Z^2 \sim \chi_1^2$

Theorem 6.2. Let $X \sim \chi_m^2$; let $Y \sim \chi_n^2$; let $X \perp Y$, then $X + Y \sim \chi_{n+m}^2$

Theorem 6.3. Let $Z \sim N(0, 1)$, let $X \sim \chi_m^2$, then $\frac{Z}{\sqrt{\frac{X}{m}}} \sim t_m$

Theorem 6.4. Let $Y = \frac{(n-q+c)\tilde{\sigma}^2}{\sigma^2}$, then $Y \sim \chi_{n-q+c}^2$

- n : number of units (sample size)
- q : number of non- σ parameters
- c : number of constraints

6.1 Example

Model 1: $Y_j = \mu + R_j, R_j \sim N(0, \sigma^2)$. What is the distribution of $\frac{\tilde{\mu} - \mu}{\frac{\tilde{\sigma}}{\sqrt{n}}}$?

We know by LS or MLE that

$$\hat{\mu} = \bar{y}$$

We know

$$\tilde{\mu} = \bar{Y}$$

Therefore, we know $\tilde{\mu} \sim N(\mu, \frac{\sigma^2}{n})$.

We standardise

$$Z = \frac{\tilde{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

By theorem 4 we know

$$X = \frac{(n-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi_{n-1}^2$$

By theorem 3 we know

$$\frac{Z}{\sqrt{\frac{X}{n-1}}} = \frac{\frac{\tilde{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)\tilde{\sigma}^2}{\sigma^2 \frac{1}{n-1}}}} = \frac{\tilde{\mu} - \mu}{\frac{\tilde{\sigma}}{\sqrt{n}}} \sim t_{n-1}$$

Recall:

$$\frac{\tilde{\mu} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

By replacing σ by $\tilde{\sigma}$, we end up using a t distribution instead of a normal

7 Confidence Interval

We assume our estimator is

$$\tilde{\theta} \sim N(\theta, \text{Var}(\tilde{\theta}))$$

The CI:

$$\theta : \text{estimate} \pm c \times \text{SE} = \hat{\theta} \pm c \sqrt{\text{Var}(\tilde{\theta})}$$

If we don't know σ we replace it by $\hat{\sigma}$.

7.1 Model 1 Example

$$Y_j = \mu + R_j, R_j \sim N(0, \sigma^2)$$

By LS, we know

$$\hat{\mu} = \bar{y}$$

The estimator

$$\tilde{\mu} - \bar{Y}$$

The distribution of our estimator is

$$\tilde{\mu} = N(\mu, \frac{\sigma^2}{n})$$

Our CI:

$$\text{estimate} \pm c \times \text{SE} = \hat{\mu} \pm c \frac{\sigma}{\sqrt{b}} = \bar{y} \pm c \frac{\sigma}{\sqrt{n}}$$

where $c \sim N(0, 1)$.

If σ is unknown:

$$\mu : \bar{y} \pm c \frac{s}{\sqrt{n}}, \quad c \sim t_{n-1}$$

Recall: $s = \frac{\sum (y_i - \bar{y})^2}{n-1}$

7.2 Model 2A Example

$$Y_{ij} = \mu_i + R_{ij}, \quad R_{ij} \sim N(0, 1)$$

By LS, $\hat{\mu}_1 = \bar{y}_1$; $\hat{\mu}_2 = \bar{y}_2$

The estimators $\tilde{\mu}_1 = \bar{Y}_1$; $\tilde{\mu}_2 = \bar{Y}_2$

The distributions are

$$\tilde{\mu}_1 \sim N(\mu_1, \frac{\sigma^2}{n_1})$$

$$\tilde{\mu}_2 \sim N(\mu_2, \frac{\sigma^2}{n_2})$$

$$\tilde{\mu}_1 - \tilde{\mu}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2})$$

Our CI is

$$\mu_1 - \mu_2 : \hat{\mu}_1 - \hat{\mu}_2 \pm c\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad c \sim N(0, 1)$$

Most of time we don't know σ , we will need to estimate it;

$$\mu_1 - \mu_2 : \hat{\mu}_1 - \hat{\mu}_2 \pm cs_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \quad c \sim t_{n_1+n_2-2}$$

7.3 Model 2B Example

$$Y_{ij} = \mu_i + R_{ij}, \quad R_{ij} \sim N(0, \sigma_i^2)$$

By LS, $\hat{\mu}_1 = \bar{y}_1$; $\hat{\mu}_2 = \bar{y}_2$

The estimators $\tilde{\mu}_1 = \bar{Y}_1$; $\tilde{\mu}_2 = \bar{Y}_2$

The distributions are

$$\tilde{\mu}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$$

$$\tilde{\mu}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$$

$$\tilde{\mu}_1 - \tilde{\mu}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$$

Our CI is

$$\mu_1 - \mu_2 : \hat{\mu}_1 - \hat{\mu}_2 \pm c \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \quad c \sim N(0, 1)$$

Most of time we don't know σ , we will need to estimate it;

$$\mu_1 - \mu_2 : \hat{\mu}_1 - \hat{\mu}_2 \pm c \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \quad c \sim t_{n_1+n_2-2}$$

7.4 Model 3 Example

$$Y_{dj} = \mu_d + R_{dj}, \quad R_{dj} \sim N(0, \sigma_d^2)$$

This is the **same** as model 1.

$$\mu_d : \bar{y}_d \pm c \frac{\sigma_d}{\sqrt{n_d}}; \quad c \sim N(0, 1)$$

$$\mu_d : \bar{y}_d \pm c \frac{s_d}{\sqrt{n_d}}; \quad c \sim t_{n_d-1}$$

7.5 Model 4 Example

$$\tilde{\pi} \sim N(\pi, \frac{\pi(1-\pi)}{n})$$

CI:

$$\hat{\pi} \pm c \frac{\hat{\pi}(1-\hat{\pi})}{n}; \quad c \sim N(0, 1)$$

8 Hypothesis Testing

1. Define the hypothesis

H_0	H_a
$\theta = \theta_0$	$\theta \neq \theta_0$
$\theta \geq \theta_0$	$\theta < \theta_0$
$\theta \leq \theta_0$	$\theta > \theta_0$

2. Discrepancy

$$d = \frac{\text{EST} - H_0 \text{ value}}{\text{SE}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\tilde{\theta})}}$$

Given $\tilde{\theta} \sim N(\theta, \text{Var}(\tilde{\theta}))$ where $D \sim N(0, 1)$ when σ is known or $D \sim t_{n-q+c}$ when σ is unknown.

3. p -value

H_a	p -value
$\theta \neq \theta_0$	$2P(D > d)$
$\theta > \theta_0$	$P(D > d)$
$\theta < \theta_0$	$P(D < d)$

4. Conclusion

$p\text{-value} > 0.1$	No evidence to reject H_0
$0.1 > p\text{-value} > 0.05$	There is evidence to reject H_0
$0.05 > p\text{-value} > 0.01$	There is some evidence to reject H_0
$p\text{-value} < 0.01$	There is tons of evidence to reject H_0

9 Models Cont.

9.1 Model 5: Completely Randomized Design (CRD)

$$Y_{ij} = \mu + \tau_i + R_{ij}, R_{ij} \sim N(0, \sigma^2)$$

$i = 1, 2, \dots, t$ (number of treatments)

$j = 1, 2, \dots, r$ (number of replicates/treatments)

Number of units = tr .

μ is the S.P. mean; $\mu + \tau_i$ is the group mean; τ_i is the treatment effect of group i .

R_{ij} is the distribution of values about the deterministic part of the model.

Constraint: $\sum \tau_i = 0$

9.1.1 Example

Group 1	Group 2
60	75
65	75
70	80

$$\hat{\mu} = \frac{60 + 65 + \dots + 80}{6} = 70$$

$$\hat{\mu} + \hat{\tau}_1 = \frac{60 + 65 + 70}{3} = 65; \quad \hat{\mu} + \hat{\tau}_2 = \frac{75 + 75 + 80}{3} = 75$$

$$\hat{\tau}_1 = -5; \quad \hat{\tau}_2 = +5$$

9.2 Least Square

$$w = \sum_{i,j} r_{ij}^2 + \lambda(\tau_1 + \cdots + \tau_t) = \sum_{i,j} (y_{ij} - \mu - \tau_i)^2 + \lambda(\tau_1 + \cdots + \tau_t)$$

Find $\frac{\partial w}{\partial \mu}, \frac{\partial w}{\partial \tau_1}, \frac{\partial w}{\partial \tau_2}, \dots, \frac{\partial w}{\partial \tau_t}, \frac{\partial w}{\partial \lambda}$

Set to zero and solve:

$$\hat{\mu} = \bar{y}_{++}; \quad \hat{\tau}_i = \bar{y}_{i+} - \bar{y}_{++}; \quad \sigma^2 = \frac{w}{n - q + c} = \frac{w}{tr - 1 - t + 1} = \frac{w}{tr - t}$$

9.3 Estimators

$$Y_{ij} = \mu + \tau_i + R_{ij}, R_{ij} \sim N(0, \sigma^2)$$

Today we consider $i = 1, 2; j = 1, 2, \dots, r$, number of units = $2r$.

Recall: $\hat{\mu} = \bar{y}_{++}$.

The estimator is $\tilde{\mu} = \bar{Y}_{++}$

$$\begin{aligned} E(\bar{Y}_{++}) &= E\left(\frac{\sum_{i=1}^2 \sum_{j=1}^r Y_{ij}}{2r}\right) \\ &= E\left(\frac{\sum_{i=1}^2 \sum_{j=1}^r (\mu + \tau_i + R_{ij})}{2r}\right) \\ &= \frac{\sum_{i=1}^2 \sum_{j=1}^r \mu + \tau_i + E(R_{ij})}{2r} \\ &= \frac{\sum_{i=1}^2 \sum_{j=1}^r \mu + \tau_i}{2r} \\ &= \frac{2ru + \sum_{j=1}^r (\tau_1 + \tau_2)}{2r} \\ &= \mu \end{aligned}$$

Hence, $E(\tilde{\mu}) = \mu$, unbiased!

$$\begin{aligned} \text{Var}(\bar{Y}_{++}) &= \text{Var}\left(\frac{\sum_{i=1}^2 \sum_{j=1}^r Y_{ij}}{2r}\right) \\ &= \frac{\sum_{i=1}^2 \sum_{j=1}^r \text{Var}(Y_{ij})}{(2r)^2} \\ &= \frac{2r\sigma^2}{(2r)^2} \text{ (by CLT)} \\ &= \frac{\sigma^2}{2r} \end{aligned}$$

Recall: $\hat{\tau}_1 = \bar{y}_{1+} - \bar{y}_{++}$.

The estimator is $\tilde{\tau}_1 = \bar{Y}_{1+} - \bar{Y}_{++}$

$$\begin{aligned}
E(\tilde{\tau}_1) &= E(\bar{Y}_{1+} - \bar{Y}_{++}) \\
&= E\bar{Y}_{1+} - \mu \\
&= E\left(\frac{\sum_{j=1}^2 Y_{ij}}{r}\right) - \mu \\
&= \frac{\sum_{j=1}^2 \mu + \tau_1}{r} - \mu \\
&= \frac{r\mu + \tau_1}{r} - \mu \\
&= \tau_1
\end{aligned}$$

Hence, $E(\tilde{\tau}_1) = \tau_1$, unbiased!

$$\begin{aligned}
\text{Var}(\tilde{\tau}_1) &= \text{Var}(\bar{Y}_{1+} - \bar{Y}_{++}) \\
&= \text{Var}\left(\bar{Y}_{1+} - \left(\frac{\bar{Y}_{1+} + \bar{Y}_{2+}}{2}\right)\right) \\
&= \text{Var}\left(\frac{1}{2}\bar{Y}_{1+} - \frac{1}{2}\bar{Y}_{2+}\right) \\
&= \frac{1}{4}\text{Var}(\bar{Y}_{1+}) + \frac{1}{4}\text{Var}(\bar{Y}_{2+}) \quad (\text{since indep.}) \\
&= \frac{1}{4} \frac{\sigma^2}{r} + \frac{1}{4} \frac{\sigma^2}{r} \\
&= \frac{1}{2r} \sigma^2 \implies \begin{cases} \text{C.I. for } \tau_1: \hat{\tau}_1 \pm c\sqrt{\frac{\hat{\sigma}^2}{2r}} \\ d = \frac{\tilde{\tau}_1 - \tau_0}{\sqrt{\frac{\hat{\sigma}^2}{2r}}}, \sim t_{n-1+c} \end{cases}
\end{aligned}$$

9.4 Confidence Interval

$$\text{C.I. for } \tau_1: \hat{\tau}_1 \pm c\sqrt{\frac{\hat{\sigma}^2}{2r}}; \quad c \sim t_{n-1+c}$$

$$\text{C.I. for } \mu: \hat{\mu} \pm c\sqrt{\frac{\hat{\sigma}^2}{2r}}; \quad c \sim t_{n-1+c}$$

10 Analysis of Variance (ANOVA)

$$\bar{Y}_{ij} = \mu + \tau_i + R_{ij}, \quad R_{ij} \sim N(0, \sigma^2)$$

Recall

$$\begin{aligned}
w &= \sum_{ij} r_{ij}^2 = \sum_{ij} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2 \\
&= \cdots = \sum_{ij} (y_{ij} - \hat{\mu})^2 + (-r) \sum_j (\bar{y}_i - \bar{y})^2
\end{aligned}$$

Rearranging:

$$\underbrace{\sum_{ij} (y_{ij} - \bar{y})^2}_{\text{Sum Squares of Total (SS}_{\text{tot}})} = \underbrace{r \sum_i (y_i - \bar{y})^2}_{\text{Sum Squares of Treatment (SS}_{\text{trt}})} + \underbrace{\sum_{ij} (y_{ij} - \hat{\mu} - \hat{\tau}_i)^2}_{\text{Sum Squares of Residual (SS}_{\text{res}})}$$

11 Model 6

Unbalanced Completely Random Design.

$$Y_{ij} = \mu + \tau_i + R_{ij}, \quad R_{ij} \sim N(0, \sigma^2)$$

where $i = 1, 2, \dots, t$; $j = 1, 2, \dots, r_i$

constraint $\sum_{i=1}^t r_i \tau_i = 0$

11.1 Least Square

$$w = \sum r_{ij}^2 + \lambda \left(\sum_{i=1}^t r_i \tau_i \right)$$

Results in

$$\hat{\mu} = \bar{y}; \quad \tau_i \bar{y}_i - \bar{y}; \quad \hat{\sigma}^2 = \frac{w}{\sum_{i=1}^t r_i - t - 1 + 1}$$

12 Model 7

Randomized Block Design (RBD)

$$Y_{ij} = \mu + \tau_i + \beta_j + R_{ij}, \quad R_{ij} \sim N(0, \sigma^2)$$

β_j is the j -th block effect.

Note: $i = 1, 2, \dots, t$; $\sum \tau_i = 0$; $j = 1, 2, \dots, r$; $\sum \beta_j = 0$

12.1 Least Square

$$w = \sum r_{ij}^2 + \lambda_1 \left(\sum \tau_i \right) + \lambda_2 \left(\sum \beta_j \right)$$

Solving

$$\hat{\mu} = \bar{y}; \quad \hat{\tau}_i = \bar{y}_i - \bar{y}; \quad \hat{\beta}_j = \bar{y}_j - \bar{y}; \quad \hat{\sigma}^2 = \frac{w}{rt - t - r - 1 + 2}$$