STAT 332 - Sampling and Experimental Design

Cameron Roopnarine

Last updated: February 12, 2021

Contents

Contents 1

1 Assignment 1 2
1.1 Lecture 1.00 - PPDAC + Example 2
1.2 Lecture 2.00 - Models, Model 1 3
1.3 Lecture 3.00 - Independent Groups 3
1.4 Lecture 4.00 - Models 2A and 2B 4
1.5 Lecture 5.00 - Model 3 4
1.6 Lecture 6.00 - Model 4 4
1.7 Lecture 7.00 - MLE 5
1.8 Lecture 8.00 - LS 7
1.9 Lecture 9.00 - LS Example 7
1.10 Lecture 10.00 - Estimators 8

Chapter 1

Assignment 1

1.1 Lecture 1.00 - PPDAC + Example

PPDAC: Problem, Plan, Data, Analysis, Conclusion.

- Problem: Define the problem.
  - **Target population** (TP): The group of units referred to in the problem step.
  - **Response**: The answer provided by the TP to the problem.
  - **Attribute**: Statistic of the response.

    EXAMPLE 1.1.1

    What is the average grade of the students in STAT 101?
    - Target population: All STAT 101 students
    - Response: Grade of a STAT 101 student.
    - Attribute: Average grade.
- Plan:
  - **Study population** (SP): The set of units you *can* study

    EXAMPLE 1.1.2

    Does a drug reduce hair loss?
    - Target population: People.
    - Study population: Mice.
  - **Sample**: A subset of the study population.
- Analysis: We analyze the data.
- Conclusion: Refers back to the problem. We also note some common *errors*.
  - **Study error**: The attribute of the population the target population differs from the parameter of the study population.

    EXAMPLE 1.1.3

    Mathematically we can write it down as $a(\text{TP}) - \mu$, however this error is qualitative. Therefore, we cannot actually calculate it.

- **Sample error**: The parameter differs from the sample statistic (estimate).

EXAMPLE 1.1.4

Mathematically we can write it down as $\mu - \bar{x}$, however this error is qualitative. Therefore, we cannot actually calculate it.

- **Measurement error**: The difference between what *we want* to calculate and what *we do* calculate.

1.2 Lecture 2.00 - Models, Model 1

DEFINITION 1.2.1: Model

A **model** relates a parameter to a response.

DEFINITION 1.2.2: Model 1

Model 1 is defined as

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

where

- Y_j: random parameter that is the response of unit j.
- μ: non-random unknown parameter that is the study population mean.
- R_j: the distribution of responses about μ.

REMARK 1.2.3

- R_j's are always independent.
- **Gauss**' Theorem: Any linear combination of normal random variables is normal.
- $Y_j \sim \mathcal{N}(\mu, \sigma^2)$ since

$$\mathbb{E}[Y_j] = \mathbb{E}[\mu + R_j] = \mathbb{E}[\mu] + \mathbb{E}[R_j] = \mu + 0 = \mu$$
$$\mathbb{V}(Y_j) = \mathbb{V}(\mu + R_j) = \mathbb{V}(R_j) = \sigma^2$$

EXAMPLE 1.2.4

Average grade of STAT 101 students.

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

1.3 Lecture 3.00 - Independent Groups

- Dependent: we randomly select one group and we find a match, having the same explanatory variates, for each unit of the first group. For example, twins, reusing members of a group, or matching.
- Independent: are formed when we select units at random from mutually exclusive groups. For example, broken parts and non-broken parts.

1.4 Lecture 4.00 - Models 2A and 2B

DEFINITION 1.4.1: Model 2A

Model 2A is used when we assume the groups have the same standard deviation and is defined as

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

where

- Y_{ij}: response of unit j in group i.
- μ_i: mean for group i.
- R_{ij}: the distribution of responses about μ_i.

DEFINITION 1.4.2: Model 2B

Model 2B is used when $\sigma_1 \neq \sigma_2$ and is defined as

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

1.5 Lecture 5.00 - Model 3

We subtract Model 2A from Model 2B to model a difference between two groups, and we get *Model 3*.

$$\begin{array}{rcccccc} & Y_{1j} & = & \mu_1 & + & R_{1j} \\ - & Y_{2j} & = & \mu_2 & + & R_{2j} \\ \hline & Y_{1j} - Y_{2j} & = & \mu_1 - \mu_2 & + & R_{1j} - R_{2j} \end{array}$$

Let

- $Y_{1j} - Y_{2j} = Y_{dj}$
- $\mu_1 - \mu_2 = \mu_d$
- $R_{1j} - R_{2j} = R_{dj}$

DEFINITION 1.5.1: Model 3

Model 3 is defined as

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

EXAMPLE 1.5.2: Model 3

Heart Rate Before Exercise	Heart Rate After Exercise	d
70	80	10
80	100	20
90	90	0

We could use Model 3.

1.6 Lecture 6.00 - Model 4

Suppose $Y \sim \text{Binomial}(n, p)$; that is, we have n outcomes where each outcome is binary.

$$\mathbb{E}[Y] = np$$

$$\mathbb{V}(Y) = np(1 - p)$$

By the Central Limit Theorem, $Y \dot{\sim} \mathcal{N}(np, np(1-p))$. The proportion is

$$\frac{Y}{n} \dot{\sim} \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

Let's find the expected value and variance of Y/n.

$$\mathbb{E}\left[\frac{Y}{n}\right] = \frac{\mathbb{E}[Y]}{n} = \frac{np}{n} = p$$

$$\mathbb{V}\left(\frac{Y}{n}\right) = \frac{\mathbb{V}(Y)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

DEFINITION 1.6.1: Model 4

Model 4 is defined as

$$\frac{Y}{n} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$$

1.7 Lecture 7.00 - MLE

- What is MLE? It connects the population parameter θ to your sample statistic $\hat{\theta}$.
- How? It chooses the most probable value of θ given our data $y_1, \dots, y_n$.

Process:

(1) Define the **likelihood function**.

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

We assume $Y_i \perp Y_j$ for all $i \neq j$. Therefore,

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

(2) Define the **log-likelihood function** and use log rules to clean it up!

(3) Find $\frac{\partial \ell}{\partial \theta}$.

(4) Set $\frac{\partial \ell}{\partial \theta} = 0$, put hat on all θ's.

(5) Solve for $\hat{\theta}$.

EXAMPLE 1.7.1

Let $Y_{ij} = \mu_i + R_{ij}$ where $R_{ij} \sim \mathcal{N}(0, \sigma^2)$.

$$\begin{aligned} L &= f(Y_{11} = y_{11}, \dots, Y_{2n_2} = y_{2n_2}) \\ &= \prod_{j=1}^{n_1} f(y_{1j}) \prod_{j=1}^{n_2} f(y_{2j}) \\ &= \prod_{j=1}^{n_1} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_{1j}-\mu_1)^2}{2\sigma^2}\right\} \prod_{j=1}^{n_2} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_{2j}-\mu_2)^2}{2\sigma^2}\right\} \end{aligned}$$

Let $n_1 + n_2 = n$, then

$$L = (2\pi)^{-n/2} \sigma^{-n} \exp\left\{-\frac{\sum_{j=1}^{n_1} (y_{1j}-\mu_1)^2}{2\sigma^2}\right\} \exp\left\{-\frac{\sum_{j=1}^{n_2} (y_{2j}-\mu_2)^2}{2\sigma^2}\right\}$$

The log-likelihood is given by

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

Now,

$$\frac{\partial \ell}{\partial \hat{\mu}_1} = 0 + 0 - \frac{\sum_{j=1}^{n_1} 2(y_{1j}-\hat{\mu})(-1)}{2\hat{\sigma}^2} + 0 = 0$$

Hence,

$$0 = \sum_{j=1}^{n_1}(y_{1j}-\hat{\mu}) \implies \sum_{j=1}^{n_1} y_{1j} = \sum_{j=1}^{n_1}\hat{\mu}$$

Note that

$$\sum_{j=1}^{n_1} y_{1j} = \frac{n_1}{n_1}\sum_{j=1}^{n_1} y_{1j} = n_1\bar{y}_{1+}$$

Therefore,

$$n_1\bar{y}_{1+} = n_1\hat{\mu} \implies \bar{y}_{1+} = \hat{\mu}_1$$

By symmetry,

$$\bar{y}_{2+} = \hat{\mu}_2$$

The second partial is given by

$$\frac{\partial \ell}{\partial \sigma} = 0 + \frac{(-n)}{\hat{\sigma}} - \frac{\sum_{j=1}^{n_1}(y_{1j}-\hat{\mu}_1)^2}{2}(-2\hat{\sigma}^{-3}) - -\frac{\sum_{j=1}^{n_2}(y_{2j}-\hat{\mu}_2)^2}{2}(-2\hat{\sigma}^{-3})$$

Multiply both sizes by $\hat{\sigma}^3$, yields

$$0 = -n\hat{\sigma}^2 + \sum_{j=1}^{n_1}(y_{1j}-\hat{\mu}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\hat{\mu}_2)^2$$

Divide both sizes by n and rearrange to get

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

Recall that

$$s^2 = \sum_{i=1}^{n}\frac{(y_i-\bar{y})^2}{n-1}$$

$$s_1^2 = \sum_{j=1}^{n_1}\frac{(y_{1j}-\bar{y}_{1+})^2}{n_1-1}$$

$$s_2^2 = \sum_{j=1}^{n_2}\frac{(y_{2j}-\bar{y}_{2+})^2}{n_2-1}$$

Therefore,

$$\hat{\sigma}^2 = s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

1.8 Lecture 8.00 - LS

- What is LS? Another technique to find $\hat{\theta}$.
- How? It minimizes the "residuals."
- Models:
$$\text{Response} = \text{Deterministic Part} + \text{Random Part}$$
$$Y = f(\theta) + R$$
Let $y_1, y_2, \ldots, y_n$ 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."

DEFINITION 1.8.1: Residual

A **residual** is
$$r_i = y_i - f(\hat{\theta}) = y_i - \hat{y}_i$$

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 $\hat{\theta}$.

1.9 Lecture 9.00 - LS Example

Let's determine the LS of Model 2A.
$$Y_{ij} = \mu_i + R_{ij}$$
Also, let $n = n_1 + n_2$.
$$\begin{aligned} W &= \sum_{ij} r_{ij}^2 = \sum_{ij} (y_{ij} - \hat{\mu}_i)^2 \\ &= \sum_{j=1}^{n} \sum_{i=1}^{2} (y_{ij} - \hat{\mu}_i)^2 \\ &= \sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)^2 + \sum_{j=1}^{n_2} (y_{2j} - \hat{\mu}_2)^2 \end{aligned}$$

$$\begin{aligned} 0 &= \frac{\partial W}{\partial \hat{\mu}_1} \\ &= \sum_{j=1}^{n_1} (y_{1j} - \hat{\mu}_1)(-2) \\ &= \frac{n_1}{n_1} \sum_{j=1}^{n_1} y_{ij} - \sum_{j=1}^{n_1} \hat{\mu}_1 \\ &= n_1 \bar{y}_{1+} - n\hat{\mu}_1 \end{aligned}$$

Therefore, $\hat{\mu}_1 = \bar{y}_{1+}$ and by symmetry $\hat{\mu}_2 = \bar{y}_{2+}$.

REMARK 1.9.1

For LS, $\hat{\sigma}^2$ is always of the form

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

where

- $n =$ number of units
- $q =$ number of non-σ parameters
- $c =$ number of constraints

Note that $\hat{\sigma}^2 = s_p^2$.

REMARK 1.9.2: MLE versus LS

- LS is from 1860's. Unbiased provided R_j is normal.
- MLE is a recent technique and it is much more flexible since it does not require R_j to be normal.
- Minimum? You need to calculate the second derivative, but we're too lazy and unrigorous in this course. No thanks.

1.10 Lecture 10.00 - Estimators