STAT 332: Sampling and Experimental Design

Professor Riley Metzger
LaTeXer Iris Jiang

Spring 2020

Contents

1 PPDAC

Problem, Plan, Data, Analysis, Conclusion

1.1 Problem

Define the proble:

- 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 unites 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.

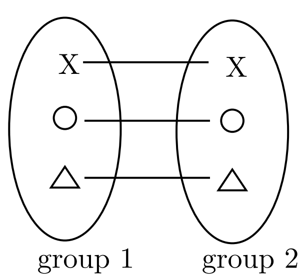
1. *R_j's are always independent.*
2. *Gaus's Theorem: Any Linear combination of normal R.V.s is normal*
3. *$Y_j \sim N(\mu, \sigma^2)$,*

$$E(Y_j) = E(\mu + R_j) = E(\mu) + \mu + 0 = \mu$$
$$V(Y_j) = V(\mu + R_j) = V(R_j) = \sigma^2$$

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

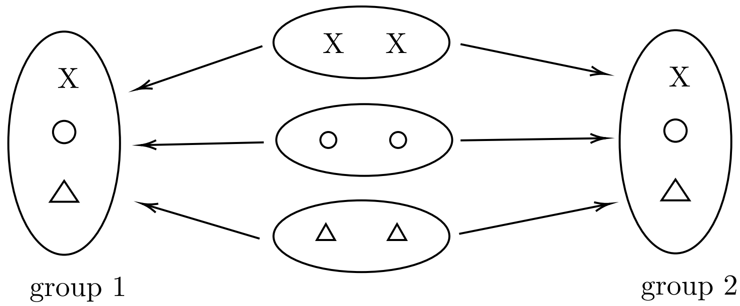
2.2 Independent vs. Dependent Groups

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.

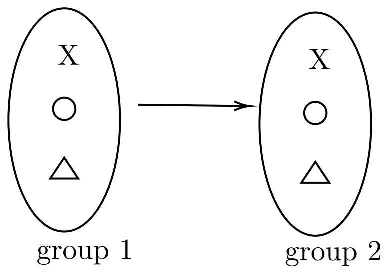


2.2.1 Ways of Creating Dependency

- Twins



- Reuse



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

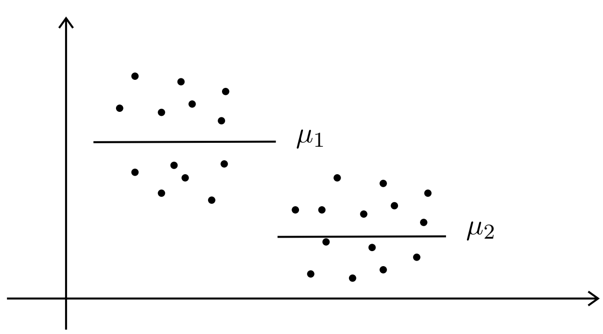
- No relationship between chosen groups Broken parts and non-broken parts

2.3 Model 2A

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

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

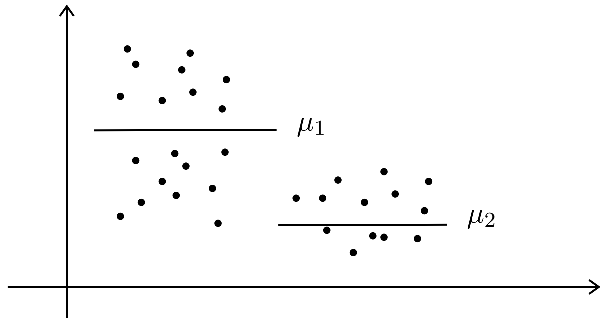
- Y_{ij}: Response of unit j in group i
- μ_i: Mean for group i; not random; unknown
- R_{ij}: The distribution of responses about μ_i



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 $y_{dj} = y_{1j} - y2j$, $\mu_d = \mu_1 - \mu_2$ 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 $y_1, y_2, \ldots, y_n$

3.3 What is the process?

1. Define likelihood function

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

   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 $\frac{\partial l}{\partial \hat{\theta}} = 0$ 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

$$\begin{aligned}
\frac{\partial l}{\partial \hat{\mu_1}} &= \frac{2\sum_{j=1}^{n_1}(y_{1j}-\hat{\mu_1})}{2\hat{\sigma^2}} = 0 \\
\implies & \sum_{j=1}^{n_1}(y_{1j}-\hat{\mu_1}) = 0 \\
& n_1\bar{y_1} - n_1\hat{\mu_1} = 0 \\
\implies & \hat{\mu_1} = \bar{y_1}
\end{aligned}$$

The estimate of population average is the sample average

By symmetry, $\hat{\mu_2} = \bar{y_2}$

$$\begin{aligned}
\frac{\partial l}{\partial \hat{\sigma}} &= -\frac{n}{\hat{\sigma}} - \frac{\sum_{j=1}^{n_1}(y_{1j}-\hat{\mu_1})^2}{2}(-2\hat{\sigma}^{-3}) - \frac{\sum_{j=1}^{n_2}(y_{2j}-\hat{\mu_2})^2}{2}(-2\hat{\sigma}^{-3}) = 0 \\
\implies & -n\hat{\sigma}^2 + \sum_{j=1}^{n_1}(y_{1j}-\hat{\mu_1})^2 + \sum_{j=1}^{n_2}(y_{2j}-\hat{\mu_2})^2 = 0 \\
& \hat{\sigma}^2 = \frac{\sum_{j=1}^{n_1}(y_{1j}-\hat{\mu_1})^2 + \sum_{j=1}^{n_2}(y_{2j}-\hat{\mu_2})^2}{n}
\end{aligned}$$

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

$$\text{Response} = \text{Deterministic Part} + \text{Random Part}$$

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

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 replaces 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 $n = n_1 + n_2$

$$\begin{aligned}
w &= \sum_{ij} r_{ij}^2 = \sum_{ij} (y_{ij} - \hat{\mu_i})^2 \\
&= \sum_{j=1} \sum_{i=1}^{2} (y_{ij} - \hat{\mu_i})^2 \\
&= \sum_{j}^{n_1} (y_{1j} - \hat{\mu_1})^2 + \sum_{j}^{n_2} (y_{2j} - \hat{\mu_2})^2 \\
\frac{\partial w}{\partial \hat{\mu_1}} &= \sum_{j}^{n_1} (y_{1j} - \hat{\mu_1})(-2) = 0 \\
\implies & \quad \hat{\mu_1} = \bar{y_1}
\end{aligned}$$

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 distrubuted*
   - *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 $y_1, y_2, \ldots, y_n$. It is not random. It is a realization of a r.v. $Y_1, Y_2, \ldots, Y_n$.
A statistic is a function of the sample data; $\hat{\theta}$, is not random. But if $y_1, y_2, \ldots, y_n$ 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 captalize our y_i's.

Example 5.1. Model 2A:
$$\underbrace{\hat{\mu} = \bar{y}_1}_{\text{statistic}} \to \underbrace{\tilde{\mu_1} = \bar{Y_1}}_{estimator}$$

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 $Y_1, \ldots, Y_n$ 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

$$\tilde{\mu_1} = \bar{Y_1} = \frac{\sum_{j=1}^{n} Y_{ij}}{n_1}$$

Thus by Gaus theorem, it is normal.

$$\begin{aligned} E(\tilde{\mu_1}) &= E(\bar{Y_1}) = E\left(\frac{\sum_{j=1}^{n} Y_{1j}}{n_1}\right) \\ &= \sum_{j=1}^{n} \frac{E(Y_{1j})}{n_1} \text{ (sum rule)} \\ &= \sum_{j=1}^{n} \frac{E(\mu_i + R_{1j})}{n_1} \\ &= \sum_{j=1}^{n} \frac{\mu_i + E(R_{1j})}{n_1} \text{ (sum rule)} \\ &= \mu_1 \end{aligned}$$

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

$$\begin{aligned} \text{Var}(\tilde{\mu_1}) &= \text{Var}\big(\bar{Y_1}\big) = \text{Var}\left(\frac{\sum_{j=1}^{n} Y_{ij}}{n_1}\right) \\ &= \frac{1}{n_1^2}\text{Var}\left(\sum_{j=1}^{n} Y_{1j}\right) \\ &= \frac{1}{n_1^2}\sum_{j=1}^{n} \text{Var}(Y_{1j}) \text{ (since } Y_{1j} \perp Y_{1i}, \forall i \neq j) \\ &= \frac{1}{n_1^2}\sum_{j=1}^{n} \text{Var}(\mu_i + R_{1j}) \\ &= \frac{1}{n_1^2}\sum_{j=1}^{n} \text{Var}(R_{1j}) = \frac{\sigma^2}{n_1} \end{aligned}$$

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 $Z^2 \sim \chi^2_1$

Theorem 6.2. Let $X \sim \chi^2_m$; let $Y \sim \chi^2_n$; let $X \perp Y$, then $X + Y \sim \chi^2_{n+m}$

Theorem 6.3. Let $Z \sim N(0,1)$, let $X \sim \chi^2_m$, 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 $Y \sim \chi^2_{n-q+c}$

- 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

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

By theorem 3 we know

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

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