STAT 332 Notes

Thomas Liu

December 1, 2024

Contents

1 Sample Survey Issues

1.1 Terminology

1.1.1 Surveys

Any activity that collects information in an organized and methodical manner about characteristics of the units of a population and compiles such info into useful summary form
Conduct surveys to learn about a population
In survey sampling, population is finite

1.1.2 Census

A censs is survey where we examine every unit
A sample survey is preferred over census because

- lower cost
- timeliness
- carefully conducted survey provides better quality estimates than sloppy census

1.1.3 Observation Unit

An entity on which measurement may be taken

1.1.4 Target Population

A collection of observation units we want to study

1.1.5 Sampling Unit

A collection of observation units that may be sampled

1.1.6 Sampling Frame

The device which allows access to the sampling units in the population from which sample may be selected

1.1.7 Study Population

The collection of all possible observation units that may have been measured in sample

Example

Do a survey among current UW undergrad, obtain a list of email address of students who volunteered during orientation, pick 100 of them, send the survey to them

- observation units: current UW undergrad
- target population: current UW undergrad
- sampling unit: UW undergrad who volunteered during orientation
- sampling frame: list of email address of UW undergrad
- study population: UW undergrad who volunteered during orientation

1.2 Sampling

1.2.1 Sampling Protocol

The mechanism by which we choose our sample

- Probability sampling protocol: probabilistic method is used to select the sample from the frame
- Non-probability sampling protocol: samples are selected based on subjective judgement of the interviewer

1.2.2 Non-probability Sampling Protocols

- convenience sampling: units are sampled based on whats's easily available
- self-selection sampling: units choose themselves
- quota sampling: units selected so some attributes of the sample match down attributes in the target population
- judgement sampling: units selected so samplers think the sample will be representative of the whole population

Issue with These Protocols

- convenience: does the sample really represent the target population?
- self-selection: does the sample volunteered really represent the whole class?
- quota: there are some other important attributes not take into account
- judgement: judgement might be biased

1.3 Errors

- study error: difference in attributes of interest between target and study population
- sample error: difference in attributes of interest between study population and sample
- measurement error: difference between response measured and true value of attributes of interest for respondents

Thus, we assume that

- sampling frame is complete (contain everyone in the target population)
- there is no non-response
- measurements are accurate

2 Simple Probability Sample

2.1 Sampling Protocol / Sampling Design

Let $\mathcal{D}$ be the set of all possible samples under chosen sampling design

- probability sampling design is determined by assigning to each possible sample s, where $s \subseteq \mathcal{D}$, a probability $P(s)$ $(P(s) > 0)$
- $\sum_{s \in \mathcal{D}} P(s) = 1$

2.2 Inclusion Probabilities

- $p_i = P(i \in S)$: the first order inclusion probability for unit i is probability that unit i is in random sample S
- $p_{ij} = P(i \in S, j \in S)$: the second order inclusion probability for units i and j is probability that both units i and j are in random sample S
- Note $p_{ii} = p_i$

2.3 Sampling Weights

Define it as

$$w_i = \frac{1}{p_i}$$

for any sampling design, inverse of first order inclusion probability

2.4 Common Protocols

2.4.1 Simple Random Sampling Without Replacement

Sampling protocol in which every one of $\binom{N}{n}$ distinct samples has an equal chance of being drawn

$$P(s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{if } s \text{ is of size } n \quad \rightarrow \quad |s| = n, \\ 0 & \text{otherwise.} \end{cases}$$

2.4.2 Stratified Random Sampling

Divide population of size N into H non-overlapping and non-empty sub-populations (called strata). In each stratum take a simple random sample without replacement of size n_h, $h = 1, \cdots, H$ such that the total sample size from H strata is n

2.4.3 Systematic Sampling

Consider fram $U = \{1, 2, \cdots, N\}$
To take a systematic sample of size n from population of size N, choose a number from 1 to $k = \frac{N}{n}$, and call this number r. Then take all units $r, r+k, r+2k, \cdots, r+(n-1)k$ from population, then the systematic sample is $s = \{y_j : j = r, r+k, r+2k, \cdots, r+(n-1)k\}$

2.4.4 Cluster Sampling

Divide population of size N into C non-overlapping and non-empty sub-populations (called clusters). Take a simple random sample of clustes without replacement, and in eawch cluster, take a census

2.4.5 Two-stage Sampling

Divide population of size N into C non-overlapping and non-empty sub-populations (like clusters). Take a simple random sample of clusters without replacement, and in each cluster, take a simple random sample without replacement

2.5 Population Parameters

Suppose out target population is $U = \{1, 2, 3, \cdots, U\}$

- N: population size
- study variable or response of interest is

$$y_i \text{ for individual } i$$

- population average: $\mu = \dfrac{1}{N} \displaystyle\sum_{i=1}^{N} y_i$

- population total: $\tau = \sum_{i=1}^{N} y_i$
- population variance: $\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$, this is unbiased

Population Parameters for Binary Response

Suppose the study variance or response of interest is binary
We use an indicator variable z_i

$$z_i = \begin{cases} 1 & \text{if (condition)} \\ 0 & \text{otherwise} \end{cases}$$

The population total is $\tau = \sum_{i=1}^{N} z_i$ and population average is a proportion

$$\mu_z = \frac{1}{N}\sum_{i=1}^{N} z_i = \frac{\tau}{N} = \pi$$

2.5.1 Variance Properties

- For any response we have

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2 = \frac{1}{N-1}[(\sum_{i=1}^{N} y_i^2) - N\mu^2]$$

- For any binary responses we have

$$\sigma_z^2 \approx \pi(1-\pi)$$

  for large N

We use $\hat{}$ for sample estimates, and $\tilde{}$ for estimator
Estimates are fixed values, but estimators are random variables

- $E[\tilde{\mu}] = \sum_{s \in D} \overline{y}(s) P(S = s)$
- $Var[\tilde{\mu}] = \sum_{s \in D} (\overline{y}(s) - E[\tilde{\mu}]^2) P(S = s)$

2.6 Random Indicator Variables

$$I_i = I(i \in S) = \begin{cases} 1 & \text{if unit } i \text{ is in sample } S \\ 0 & \text{otherwise} \end{cases}$$

2.6.1 Useful Results

- $P(I_i = 1) = p_i$
- $P(I_i = 0) = 1 - p_i$
- $E[I_i] = 0(1 - p_i) + 1p_i = p_i$
- $I_i^2 = I_i$
- $Var(I_i) = p_i(1 - p_i)$
- $Cov(I_i, I_j) = p_{ij} - p_i p_j$

2.7 Simple Random Sampling Without Replacement (SRSWOR)

Select n units from frame of N units so that each sample of size n has the same probability of selection

- sampling frame: $U = \{1, 2, \cdots, N\}$
- sample size: fixed at n
- $\binom{N}{n}$ possible samples of size n
- sampling protocol

$$P(S = s) = \begin{cases} \frac{1}{\binom{N}{n}} & \text{if } s \text{ has } n \text{ distinct elements} \\ 0 & \text{otherwise} \end{cases}$$

Inclusion Probabilities

First order inclusion probabilities: $p_i = \frac{n}{N}$
Second order inclusion probabilities: $p_{ij} = \frac{n(n-1)}{N(N-1)}, i \neq j$

2.8 Estimator Results

- μ and σ^2: our population average and variance
- $\tilde{\mu}$ and $\tilde{\sigma}^2$: corresponding estimator
- Under SRSWOR,

$$E(\tilde{\mu}) = \mu \quad Var(\tilde{\mu}) = (1 - \frac{n}{N})\frac{\sigma^2}{n}$$

- Under SRSWOR, $\tilde{\mu}$ is unbiased estimator for population average μ

Helpful Derivations

When sampling design is SRSWOR, we have

- $Var(I_i) = p_i(1-p_i) = \frac{n}{N}(1-\frac{n}{N})$
- $Cov(I_i, I_j) = p_{ij} - p_i p_j = \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n}{N}\frac{1}{(N-1)}(1-\frac{n}{N})$

Variance Estimator

Under SRSWOR

- $E[\tilde{\sigma}^2] = \sigma^2$, $\tilde{\sigma}^2$ is unbiased
- $Var(\tilde{\mu}) = (1-\frac{n}{N})\frac{\sigma^2}{n}$
- $f = \frac{n}{N}$ is sampling fraction
- $1 - f = 1 - \frac{n}{N}$ is finite population correction factor (fpc)

we can write $Var(\tilde{\mu}) = (1-f)\frac{\sigma^2}{n}$

Standard Error

$$\hat{SD}(\tilde{\mu}) = \sqrt{\hat{Var}(\tilde{\mu})} = \sqrt{(1-\frac{n}{N})\frac{\hat{\sigma}^2}{n}} = \hat{\sigma}\sqrt{\frac{(1-f)}{n}} = s.e.(\hat{\mu})$$

It is not the sample standard deviation, it is smaller.

2.9 Confidence Intervals

Central Limit Theorem (for finite populations) tells if N, n and $N-n$ are large, then

$$\frac{\tilde{\mu} - \mu}{\sqrt{(1-f)\frac{\tilde{\sigma}^2}{n}}} \sim N(0,1)$$

A large sample $100(1-\alpha)\%$ confidence interval for population average μ is

$$\hat{\mu} \pm c \times s.e.(\hat{\mu}) = \hat{\mu} \pm c\sqrt{(1-f)\frac{\hat{\sigma}^2}{n}}$$

where c is chosen such that $P(Z \leq c) = 1 - \alpha/2$ or $P(Z > c) = \alpha/2$ for standard normal distribution $Z \sim N(0,1)$

Estimating Total

- $\tau = N\mu$
- $E(\tilde{\tau}) = \tau$
- $Var(\tilde{\tau}) = N^2 Var(\tilde{\mu})$, $SD(\tilde{\tau})N \times SD(\tilde{\mu})$
- $100(1-\alpha)\%$ CI for τ is

$$\hat{\tau} \pm c \times s.e.(\hat{\tau}) = N(\hat{\mu} \pm c \times s.e.(\hat{\mu}))$$

Estimating a Proportion: Standard Error

We know if $y_i \in \{0, 1\}$, the standard deviation is

$$\sigma = \sqrt{\frac{N}{N-1}\pi(1-\pi)} \approx \sqrt{\pi(1-\pi)}$$

which results in

$$s.e.(\hat{\pi}) = \sqrt{1-\frac{n}{N}}\frac{\hat{\sigma}}{\sqrt{n}} \approx \sqrt{1-\frac{n}{N}}\frac{\sqrt{\hat{\pi}(1-\hat{\pi})}}{\sqrt{n}}$$

$100(1-\alpha)\%$ confidence interval for proportion π is

$$\hat{\pi} \pm c \times s.e.(\hat{\pi}) = \hat{\pi} \pm c\sqrt{1-\frac{n}{N}}\frac{\sqrt{\hat{\pi}(1-\hat{\pi})}}{\sqrt{n}}$$

in which $P(Z \leq c) = 1 - \alpha/2$

2.10 Margin of Error

We want a confidence interval of length $2L$, or

$$\hat{\mu} \pm L$$

$$L = c\sqrt{(1-\frac{n}{N})\frac{\hat{\sigma^2}}{n}}$$

Then we can write n as

$$n = (\frac{1}{N} + \frac{L^2}{c^2\hat{\sigma}^2})^{-1} \approx \frac{c^2\hat{\sigma}^2}{L^2}$$

with approximation hold when N is large

2.11 Pilot Study

To determine sample size, we need the $s.e.$ of estimate, which requires $\hat{\sigma}$
Pilot Study: a small porportion of the sample is collected to estimate the nuisance parameters (such as σ in the estimation of μ). This relatively small sample is called 'pilot'
pilot sample + additional sample = n

3 Ratio and Regression Estimation with SRS

3.1 Ratio Estimation

$$\theta = \frac{\tau_Y}{\tau_X} = \frac{N\mu_Y}{N\mu_X} = \frac{\mu_Y}{\mu_X}$$

is the ration of means
Note that

$$E(\frac{\tilde{\mu}_Y}{\tilde{\mu}_X}) \neq \frac{E(\tilde{\mu}_Y)}{\tilde{\mu}_X}$$

We have

$$\tilde{\theta} = \frac{\tilde{\mu}_Y}{\tilde{\mu}_X} \approx \frac{\mu_Y}{\mu_X} + \frac{1}{\mu_X}(\tilde{\mu}_Y - \mu_Y) - \frac{\mu_Y}{\mu_X^2}(\tilde{\mu}_X - \mu_X)$$

With

$$E(\tilde{\theta}) \approx \frac{\mu_Y}{\mu_X} = \theta$$

$$Var(\tilde{\theta}) = \frac{1}{\mu_X^2} Var(\tilde{\mu}_Y - \theta\tilde{\mu}_X)$$

3.1.1 Residual / Redefined Estimator

We define residual as $\tilde{\mu}_Y - \theta\tilde{\mu}_X$, and

$$\tilde{\mu}_Y - \theta\tilde{\mu}_X = \tilde{\mu}_r$$

3.1.2 Properties of $\tilde{\mu}_r$

- $E[\tilde{\mu}_r] = \mu_r$
- $Var(\tilde{\mu}_r) = (1 - \frac{n}{N})\frac{\sigma_r^2}{n}$ where $\sigma_r^2 = \frac{1}{N-1}\sum_{i\in U}[y_i - \theta x_i - (\mu_Y - \theta\mu_X)]^2$

3.1.3 Ratio Estimation Confidence Interval

If we assume a large enough sample, $\tilde{\theta}$ is approximately Gaussian, and corresponding $100(1-\alpha)\%$ confidence interval for ratio θ can be constructed as

$$\hat{\theta} \pm c \times S.E.(\hat{\theta})$$

where standard error is

$$s.e.(\hat{\theta}) = \sqrt{\hat{Var}(\tilde{\theta})} = \frac{1}{|\hat{\mu}_X|}\sqrt{(1 - \frac{n}{N})\frac{\hat{\sigma}_r^2}{n}}$$

3.2 Ratio Estimation of the Mean

Suppose the relationship between X and Y is

$$\text{response} = \text{signal} + \text{noise}$$

$$y_i = \theta x_i + R_i$$

3.2.1 Ratio Estimation of the Mean

If X and Y are correlated, knowing things about X could help improve estimateion of the mean of Y

3.2.2 Ratio Estimate of μ_Y

$$\hat{\mu}_{ratio} = \hat{\theta}\mu_X = (\frac{\hat{\mu}_Y}{\hat{\mu}_X})\mu_X = \hat{\mu}_Y[\frac{\mu_X}{\hat{\mu}}]$$

We also have

$$E(\tilde{\mu}_{ratio}) \approx \mu_y$$

$$Var(\tilde{\mu}_{ratio}) \approx (1 - \frac{n}{N})\frac{\sigma_r^2}{n}$$

with $\sigma_r^2 = \frac{1}{n-1}\sum_{i\in S}(y_i - \hat{\theta}x_i)^2$

3.2.3 Confidence Interval

$$\hat{\mu}_{ratio} \pm c \times s.e.(\hat{\mu}_{ratio})$$

where

$$s.e.(\hat{\mu}_{ratio}) = \sqrt{Var(\tilde{\mu}_{ratio})}$$

3.3 Regression Estimation

We define the regression estimate of μ_Y

$$\hat{\mu}_{reg} = \hat{\alpha} + \hat{\beta}\mu_X = \hat{\mu}_y + \hat{\beta}(\mu_x - \hat{\mu}_x)$$

where

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} \quad \hat{\beta} = \frac{\sum_{i\in s}(x_i - \overline{x})y_i}{\sum_{i\in s}(x_i - \overline{x})^2}$$

3.3.1 Expectation and Variance

- $E(\tilde{\mu}_{reg}) \approx \mu_Y$
- $Var(\tilde{\mu}_{reg}) = (1 - \frac{n}{N})\frac{\hat{\sigma}_{reg}^2}{n}$ where $\sigma_{reg}^2 = \frac{1}{n-1}\sum_{i\in S}[y_i - \hat{\alpha} - \hat{\beta}x_i]^2$

3.3.2 Confidence Interval

$$\hat{\mu}_{reg} \pm c \times s.e.(\hat{\mu}_{reg})$$

where

$$s.e.(\hat{\mu}_{reg}) = \sqrt{\frac{1}{n}(1 - \frac{n}{N})\frac{\sum_{i\in S}(y_i - \overline{y} - \hat{\beta}(x_i - \overline{x}))^2}{n-1}}$$

4 Stratified Sample

Divide population into H mutualy exclusive and non-empty strata, each statum has size N_h
$N = N_1 + N_2 + \cdots + N_h$ is the number of units in the population
Variate of interest will be y, and write

- y_{hj}: the observed value for jth element in the hth stratum
- $\mu_h = \frac{1}{N_h}\sum_{j=1}^{N_h} y_{hj}$: the mean for stratum h (population level)
- $\mu = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj}$: the overall population mean
- $\sigma_h^2 = \frac{1}{N_h - 1}\sum_{j=1}^{N_h}(y_{hj} - \mu_h)^2$: variance for statrum h
- $\sigma^2 = \frac{1}{N-1}\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj} - \mu)^2$: population variance

4.0.1 Population Mean and Stratum Weight

We can also write the population mean as

$$\mu = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj} = \frac{1}{N}\sum_{h=1}^{H} N_h \frac{\sum_{j=1}^{N_h} y_{hj}}{N_h} = \sum_{h=1}^{H}\frac{N_h}{N}\mu_h = \sum_{h=1}^{H} W_h\mu_h$$

where $W_h = \frac{N_h}{N}$ is the stratum weight

4.1 Estimation

For a particular stratum h, $h = 1, \cdots, H$

- $\hat{\mu}_h = \overline{y}_h = \frac{1}{n_h}\sum_{j\in s_h} y_{hj}$: sample stratum mean
- $\hat{\sigma}_h^2 = \frac{1}{n_h - 1}\sum_{j\in s_h}(y_{hj} - \hat{\mu}_h)^2$: sample stratum variance

Suppose SRSWOR sampling used within each stratum, then

- $E(\tilde{\mu}_h) = \mu_h$
- $Var(\tilde{\mu}_h) = (1 - \frac{n_h}{N_h})\frac{\sigma_h^2}{n_h}$

The estimate of population average is

$$\hat{\mu}_s = \sum_{h=1}^{H} W_h \hat{\mu}_h = \sum_{h=1}^{H} \frac{N_H}{N} \hat{\mu}_h$$

with corresponding estimator $\tilde{\mu}_s$

- $E[\tilde{\mu}_s(y)] = \mu$
- $Var(\tilde{\mu}_s) = \sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{\sigma_h^2}{n_h}$ where $\sigma_h^2 = \frac{1}{n_h - 1} \sum_{j \in s_h} (y_{hj} - \mu_h)^2$

4.2 Confidence Interval for Mean μ

A $100(1-\alpha)\%$ confidence interval for μ based on a random stratified sample is

$$\hat{\mu}_s \pm c \times SE(\hat{\mu}_s)$$

- $\hat{\mu}_s = \sum_{h=1}^{H} W_h \hat{\mu}_h$
- $SE(\hat{\mu}_s) = \sqrt{\sum_{h=1}^{H} W_h^2 (1 - \frac{n_h}{N_h}) \frac{\hat{\sigma}_h^2}{n_h}}$
- $P(Z \leq c) = 1 - \frac{\alpha}{2}$, use Z over T when σ is known

4.3 Proportional Allocation

We select the same proportion of units from each stratum, $n_h = \frac{n}{N} N_h$
Then we have $W_h = \frac{N_h}{N} = \frac{n_h}{n} = w_h$, the variance of the stratified estimator is

$$Var(\tilde{\mu}_s) = \frac{1}{n} (1 - \frac{n}{N}) \sum_{h=1}^{H} W_h \sigma_h^2$$

4.3.1 Sample Size Determination

We have

$$L = c \sqrt{\frac{1}{n} (1 - \frac{n}{N}) \sum_{h=1}^{H} W_h \hat{\sigma}_h^2}$$

where $W_h = \frac{N_h}{N} = \frac{n_h}{n}$, use $\frac{N_h}{N}$ here to keep only one n

4.4 Post Stratification

The post-stratified population average estimate is

$$\hat{\mu}_{post} = W_1\hat{\mu}_1 + \cdots + W_H\hat{\mu}_H$$

where $\hat{\mu}_h = \frac{1}{n_h}\sum_{j\in s_h} y_{hj}$ is the stratum h sample average

4.4.1 Conditional Expectation and Variance

Condition on the fact that $\tilde{n}_1 = n_1, \cdots, \tilde{n}_H = n_H$
Then we have

$$E(\tilde{\mu}_{post}|\tilde{n}_1 = n_1, \cdots, \tilde{n}_H = n_H) = \mu$$

$$Var(\tilde{\mu}_{post}|\tilde{n}_1 = n_1, \cdots, \tilde{n}_H = n_H) = \sum_{h=1}^{H} W_h^2(1 - \frac{n_h}{N_h})\frac{\sigma_h^2}{n_h}$$

4.4.2 Unconditional Expectation and Variance

$$E[\tilde{\mu}_{post}] = E[E(\tilde{\mu}_{post}|\tilde{n}_1 = n_1, \cdots, \tilde{n}_H = n_H)] \approx \mu$$

$$Var(\tilde{\mu}_{post}) = \sum_{h=1}^{H} W_h^2(\frac{1}{n_h} - \frac{1}{N_h})\sigma_h^2$$

4.5 Cluster Sampling

4.5.1 Equal Size

- N is number of clusters
- number of units within each cluster is M
- population size is $NM = T$
- y_{ij} is the response from unit j in cluster i
- cluster average is $\mu_i = \overline{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$
- population average is

$$\mu = \frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} = \frac{1}{N}\sum_{i=1}^{N}\overline{y}_i$$

The point estimate for population average is

$$\hat{\mu} = \frac{1}{nM}\sum_{i\in s}\sum_{j=1}^{M} y_{ij} = \frac{1}{n}\sum_{i\in s}\overline{y}_i$$

$$E[\tilde{\mu}] = \mu_z = \frac{1}{N}\sum_{i=1}^{N}\overline{y}_i = \mu$$

$$Var(\tilde{\mu}) = (1 - \frac{n}{N})\frac{\sigma_z^2}{n}$$

where

$$\sigma_z^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\overline{y}_i - \mu)^2$$

4.5.2 Unequal Size

- N is number of clusters
- number of units within each cluster is M_i
- population size is $T = \sum_{i=1}^{N} M_i$
- y_{ij} is response from unit j in cluster i
- cluster average is $\mu_i = \overline{y}_i = \frac{1}{M_i}\sum_{j=1}^{M_i} y_{ij}$
- cluster total is $t_i = \sum_{j=1}^{M_i} y_{ij}$
- population average is

$$\mu = \frac{1}{T}\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij} = \frac{\mu_T}{\mu_M}$$

$$Var(\hat{\mu}) = \frac{1}{\hat{\mu}_M^2}(1 - \frac{n}{N})\frac{\hat{\sigma}_u^2}{n}$$

where

$$\hat{\sigma}_u^2 = \frac{1}{n-1}\sum_{i\in s}(t_i - \hat{\mu}M_i)^2$$

4.6 Non-Response

Response rate is

$$RR = \frac{\text{number of complete interviews with reporting units}}{\text{number of eligible reporting units in the sample}}$$

We have

Component	Notation	Number of
Eligible responders	I	Complete interviews
	P	Partial interviews
Eligible non-responders	R	Refusals or break-offs
	NC	Non-contacts
	O	Other eligible non-respondents
Non-responders with unknown eligibility	UH	Unknown household occupancy
	UO	Others with unknown eligibility

We care most about about RR1 and RR6

- RR1: $\dfrac{I}{I+P+R+NC+O+UH+UO}$
- RR6: $\dfrac{I+P}{I+P+R+NC+O}$

4.6.1 Estimation

- N_R: numbere of respondents
- N_M: number of non-respondents
- μ_R: mean of respondents
- μ_M: mean of non-respondents

$$\mu = W_R\mu_R + W_M\mu_M = \frac{N_R}{N}\mu_R = \frac{N_M}{N}\mu_M$$

5 Fundamental of Experimental Plans

5.1 The Fundamental Principles

5.1.1 Observational Study

An observational study is one in which data are collected about a population or process without any attempt to change the value of one or more variates for the sampled units

5.1.2 Experimental Study

In experimental study we deliberately change one or more of the process input to investigate the effect of the change on the process output

Difference

Observational plans provide insight about the effect of changes in varying inputs, but con't conlude causality
In experimental plans, people who conduct the experiment change one or more input variates on the sample units before the response variate is measured

5.1.3 Terminology

- Factor: a single explanatory variate (input) that will be changed or set on each unit
  - prescribed treatment
  - additional therapy

  Can be
  - Quantitative: occlusion dose in hours
  - Qualitative: type of therapy
- Factor levels: the set of values assigned to any factor in the plan
  - low or high does (2 levels)
  - glasses, drops, none (3 levels)
- Treatment: combination of levels of factors that can be applied to a unit
- Experimental unit: object or individual that we apply treatment to
- Response variate: outcome or observation in the study. Can be continuous

5.1.4 Fundamental Principles

- Replication: applying each treatment to more than one experimental unit
- Random assignment (random allocation): process of assigning treatments to experimental unit using a probabilistic mechanism
- Blocking: groups in which one or more explanatory variates are held fixed while different treatments are applied to the units within the group

5.1.5 Single Treatment

If we want to use the model

$$Y_i = \mu + R_i \quad i = 1, 2, \cdots, 20$$

with $R_i \sim N(0, \sigma^2)$, and all R_i are mutually independent We know the estimate for the residual variance σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \hat{\mu})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

The corresponding estimators have the following properties

$$\tilde{\mu} \sim N(\mu, \frac{\sigma^2}{n}) \quad \frac{\tilde{\sigma}^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

The pivotal quantity is

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

The $100(1-\alpha)\%$ confidence interval is computed by

$$\hat{\mu} \pm c \times s.e.(\hat{\mu})$$

where c is chosen such that $P(|T_{n-1}| \leq c) = 1-\alpha$, and $s.e.(\hat{\mu}) = \hat{\sigma}/\sqrt{n}$

5.2 Comparative Experimental Plans without Blocking

5.2.1 A Model for Comparing Two Treatments

$$Y_{ij} = \mu + \tau_i + R_{ij} \quad i = 1, 2 \quad j = 1, 2, \cdots, n$$

where

- Y_{ij} is response for the jth unit that received treatment i
- $R_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$ is the experimental variability (error), with equal variance across groups

Also apply constraint $n_1\tau_1 + n_2\tau_2 = 0$ so that

- μ is the mean response across our two treatment groups
- τ_1, τ_2 represent treatment effect, increase or decrease from the mean response due to treatment

5.2.2 Notation

- $Y_{i+} = \sum_{j=1}^{n_i} Y_{ij}$
- $\overline{Y}_{i+} = \frac{1}{n_i} Y_{i+}$
- $Y_{++} = \sum_{i=1}^{2} \sum_{j=1}^{n_i} Y_{ij}$
- $\overline{Y}_{++} = \frac{1}{n_1+n_2} Y_{++} = (\frac{n_1}{n_1+n_2})\overline{Y}_{1+} + (\frac{n_2}{n_1+n_2})\overline{Y}_{2+}$

where

$$E\left[\overline{Y}_{++}\right] = \mu$$

- $E[\overline{Y}_{i+}] = \mu + \tau_i$

- $E[\overline{Y}_{1+} - \overline{Y}_{2+}] = \tau_1 - \tau_2 = \theta$

$\tilde{\theta} = \overline{Y}_{1+} - \overline{Y}_{2+}$ is the unbiased estimator for θ
We want to minimize μ, τ_1, τ_2 from

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \mu - \tau_i)^2$$

such that $\sum_{i=1}^{2} n_i\tau_i = 0$
We get

- $\hat{\mu} = \overline{y}_{++}$
- $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}$
- $\hat{\tau}_1 - \hat{\tau}_2 = \overline{y}_{1+} - \overline{y}_{2+}$

Variance Estimation

The sample variance within treatment i is

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

The pooled overall sample variance is

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

5.2.3 Confidence Interval and Hypothesis Test

Given the model for comparative studies without blocking

$$Y_{ij} = \mu + \tau_i + R_{ij} \quad R_{ij} \overset{iid}{\sim} N(0, \sigma^2) \quad i = 1, 2 \quad j = 1, \cdots, n_i$$

subject to $\sum_{i=1}^{2} n_i\tau_i = 0$, we have $\overline{Y}_{i+} \sim N(\mu + \tau_i, \sigma^2/n_i)$

Estimators

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

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

- Variance: $\frac{(n_i - 1)\tilde{\sigma}_i^2}{\sigma^2} \sim \chi^2_{n_i - 1}$, and so

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

Confidence Interval for $\theta = \tau_1 - \tau_2$

Putting all together to get

$$\frac{\tilde{\theta} - \theta}{\tilde{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\tilde{\tau}_1 - \tilde{\tau}_2) - (\tau_1 - \tau_2)}{\tilde{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Thus the $100(1-\alpha)\%$ confidence interval for θ is

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

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

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

Hypothesis Test for $\theta = \tau_1 - \tau_2$

To test $H_0 : \theta = \theta_0$, calculate observed test statistic (or discrepancy measure)

$$t_{obs} = \frac{(\hat{\tau}_1 - \hat{\tau}_2) - \theta_0}{\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

There is evidence against $H_0 : \theta = \theta_0$ if p-value is small

5.3 Comparative Experimental Plans with Blocking

5.3.1 Model

$$Y_{ij} = \mu + \tau_i + \beta_j + R_{ij} \quad i = 1, 2 \quad j = 1, 2, \cdots, b \quad R_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

subject to $\sum_{i=1}^{2} \tau_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$

- μ: overall average improvement in visual acuity
- τ_i: treatment effect for treatment i
- β_j: block effect for block j
- $n = 2b$: sample size

5.3.2 Estimation

Want to minimize $\mu, \tau_i, \beta_j, \lambda_1, \lambda_2$ for

$$\sum_{i=1}^{2}\sum_{j=1}^{b}(y_{ij} - \mu - \tau_i - \beta_j)^2 + \lambda_1 \sum_{i=1}^{2}\tau_i + \lambda_2 \sum_{j=1}^{b}\beta_j$$

This gives

- $\hat{\mu} = \overline{y}_{++}$ for overall mean
- $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}$ for treatment effect
- $\hat{\beta}_j = \overline{y}_{+j} - \overline{y}_{++}$ for block effect

We are interested in estimating the difference between treatment effects

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

The estimated variance is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{b-1}\sum_{i=1}^{2}\sum_{j=1}^{b}(y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j)^2 \\ &= \frac{1}{b-1}\sum_{i=1}^{2}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{i+} - \overline{y}_{+j} + \overline{y}_{++})^2\end{aligned}$$

5.3.3 Degrees of Freedom for Two Treatments with Blocking

The degree of freedom is

$$sample\ size - number\ of\ parameters\ estimated + number\ of\ constraints$$

5.3.4 Confidence Interval and Hypothesis Test

Estimators for τ_i is

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

Then

$$\tilde{\theta} = \tilde{\theta}_1 - \tilde{\theta}_2 = \overline{D} \sim N(\tau_1 - \tau_2, \frac{2\sigma^2}{b})$$

Since $\tilde{\sigma}^2 \sim \chi^2_{b-1}$, we have

$$\frac{\tilde{\theta} - \theta}{\tilde{\sigma}\sqrt{2/b}} \sim t_{b-1}$$

The $(1-\alpha)\times 100\%$ confidence interval for $\theta = \tau_1 - \tau_2$ is

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

where c is chosen such that $P(|T_{b-1}| \le c) = 1 - \alpha$, and $s.e.(\hat{\tau}_1 - \hat{\tau}_2) = \hat{\sigma}\sqrt{2/b}$

6 Experiemntal Plans for More Than Two Treatments

6.1 Completely Randomized Designs

6.1.1 Balanced Design

The design is balanced if the same number of units receive each treatment
Assuming a balanced plan, we observe a response y_{ij} for the j^{th} unit receiving the i^{th} treatment, and the model is

$$Y_{ij} = \mu + \tau_i + R_{ij} \quad i = 1, 2, \cdots, t \quad j = 1, 2, \cdots, n \quad R_{ij} \overset{iid}{sim} N(0, \sigma^2)$$

where

- r: number of replicates for each tretment
- t: number of treatments
- μ: mean response across all treatments
- τ_i: treatment effect for treatment i
- The balanced design constraint is $\sum_{i=1}^{t} n_i \tau_i = 0$

We have

$$E(\overline{Y}_{++}) = E(\tilde{\mu}) = \mu$$
$$E(\overline{Y}_{i+}) = \mu + \tau_i \Rightarrow \tau_i = E(\overline{Y}_{i+}) = E(\overline{Y}_{i+} - \overline{Y}_{++})$$

6.1.2 Parameter Estimation

Using the Lagrange multiplier method and solve the system of $t + 2$ equations, we get

- $\hat{\mu} = \overline{y}_{++}$: overall average
- $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}$: treatment effect
- $\hat{\tau}_i - \hat{\tau}_h = \overline{y}_{i+} - \overline{y}_{h+}$: difference in treatment effect

The estimate of the error variance σ^2 is

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

The degree of freedom in the denominator is $rt - t$

$$total\ sample\ size - number\ of\ parameters\ estimated + number\ of\ constraints$$

6.1.3 Dsitributions of the Estimators

- $Y_{ij} = \mu + \tau_i + R_{ij}$, $R_{ij} \overset{iid}{\sim} N(0, \sigma^2)$
- $\tilde{\tau}_i = \overline{Y}_{i+} - \overline{Y}_{++}$, $\tilde{\tau}_i - \tilde{\tau}_h = \overline{Y}_{i+} - \overline{Y}_{h+}$
- $\overline{Y}_{i+} \sim N(\mu + \tau_i, \sigma^2/n)$
- Then, $\overline{Y}_{i+} - \overline{Y}_{h+} \sim N(\tau_i - \tau_h, \sigma^2(1/r + 1/r))$
- $Z = \dfrac{\tilde{\tau}_i - \tilde{\tau}_h - (\tau_i - \tau_h)}{\sqrt{\sigma^2(1/r + 1/r)}} \sim N(0, 1)$
- $U = \dfrac{t(r-1)\tilde{\sigma}^2}{\sigma^2} \sim \chi^2_{t(r-1)}$, independent of Z
- Therefore,

$$\frac{Z}{\sqrt{U/(t(r-1))}} = \frac{\tilde{\tau}_i - \tilde{\tau}_h - (\tau_i - \tau_h)}{\sqrt{\sigma^2(1/r + 1/r)}} \sim t_{t(r-1)}$$

6.1.4 Contrast

For this course, we need the following condition to be satisfied

$$\theta = \sum_{i=1}^{t} a_i \tau_i = 0 \quad \sum_{i=1}^{t} a_i = 0$$

Estimate and Estimator

- Parameter: $\theta = \sum_{i=1}^{t} a_i \tau_i$
- Estimate: $\hat{\theta} = \sum_{i=1}^{t} a_i \hat{\tau}_i = \sum_{i=1}^{t} a_i \overline{y}_{i+}$
- Estimator: $\tilde{\theta} = \sum_{i=1}^{t} a_i \tilde{\tau}_i = \sum_{i=1}^{t} a_i \overline{Y}_{i+}$

We have

$$\tilde{\theta} \sim N(\theta, \sigma^2 \sum_{i=1}^{t} a_i^2 (1/r))$$

6.1.5 Confidence Interval and Hypothesis Test

$$\tilde{\theta} \sim N(\theta, \sigma^2 \sum_{i=1}^{t} a_i^2 (1/r)) \Rightarrow Z = \frac{\tilde{\theta} - \theta}{\sqrt{\sigma^2 \sum_{i=1}^{t} a_i^2 (1/r)}} \sim N(0, 1)$$

It turns that that

$$\frac{Z}{\sqrt{U/(t(r-1))}} = \frac{\tilde{\theta} - \theta}{\sqrt{\sigma^2 \sum_{i=1}^{t} a_i^2 (1/r)}} \sim t_{t(r-1)}$$

A $100(1-\alpha)\%$ confidence interval for θ is

$$\hat{\theta} \pm c \times s.e.(\hat{\theta})$$

where c is chosen such that $P(T_{t(r-1)} \leq c) = 1 - \alpha/2$, and $s.e.(\hat{\theta}) = \hat{\sigma}\sqrt{\sum_{i=1}^{t} a_i^2(1/r)}$
Similarly, the observed test statistic for $H_0 : \theta = \theta_0$ is

$$t_{obs} = \frac{\hat{\theta} - \theta_0}{\hat{\sigma}\sqrt{\sum_{i=1}^{t} a_i^2(1/r)}}$$

6.2 Unbalanced Design

6.2.1 Model

$$Y_{ij} = \mu + \tau_i + R_{ij} \quad i = 1, 2, \cdots, t \quad j = 1, 2, \cdots, n_i \quad R_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

where

- constraint: $\sum_{i=1}^{t} n_i \tau_i = 0$
- $\mu = E[\overline{Y}_{++}] = \dfrac{\sum_{i=1}^{t} r_i E[\overline{Y}_{i+}]}{\sum_{i=1}^{t} r_i}$, a weighted average of treatment means
- $\mu + \tau_i =$ average response for treatment i

6.2.2 Parameter Estimation

- Estimates: $\hat{\mu} = \overline{y}_{++}$, $\hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++}$
- Estimators: $\tilde{\mu} = \overline{Y}_{++}$, $\tilde{\tau}_i = \overline{Y}_{i+} - \overline{Y}_{++}$
- Variance:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \overline{y}_{i+})^2}{\sum_i (r_i - 1)}$$

6.2.3 Anova Table

Source	SS	df	MS	F
Treatment	$\sum_{i=1}^{t} r_i(\overline{y}_{i+} - \overline{y}_{++})^2$	$t-1$	$\dfrac{SS_{treatment}}{t-1}$	$\dfrac{MS_{treatment}}{MS_{error}}$
Residuals	$\sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \overline{y}_{i+})^2$	$\sum_{i=1}^{t}(r_i - 1)$	$\dfrac{SS_{error}}{\sum_{i=1}^{t}(r_i-1)}$	
Total	$\sum_{i=1}^{t} \sum_{j=1}^{r_i} (y_{ij} - \overline{y}_{++})^2$	$\sum_{i=1}^{t} r_i - 1$		

As from the ANOVA table, we have

$$F = \frac{MS_T}{MS_R} = \frac{\sum_i r_i(\overline{Y}_{i+} - \overline{Y}_{++})^2/(t-1)}{\sum_i \sum_j (y_{ij} - \overline{Y}_{i+})^2 / \sum_i (r_i - 1)} \sim F_{t-1, \sum_i (r_i - 1)}$$

6.2.4 Contrasts

Each treatment average has the distribution

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

So for a contrast in unequal sample size we have

$$tilde\sigma \sim N(\theta, \sigma^2 \sum_{i=1}^{t} a_i^2/r_i)$$

Hence a $100(1-\alpha)\%$ confidence interval for θ is

$$\hat{\theta} \pm t * \hat{\sigma} \sqrt{\sum_{i=1}^{t} a_i^2/r_i}$$

where $t*$ is chosen such that $P(T_{t-1} \leq t*) = 1 - \alpha/2$

6.3 Randomized Block Designs

6.3.1 Model

$$Y_{ij} = \mu + \tau_i + \beta_j + R_{ij} \quad i = 1, 2, \cdots, t \quad j = 1, 2, \cdots, b \quad R_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

where

- $\sum_{i=1}^{t} \tau_i = \sum_{j=1}^{b} \beta_j = 0$
- μ: overall mean response
- τ_i: treatment effect for treatment i
- β_j: block effect for block j

6.3.2 Estimation

Least squares from the model gives us

$$\hat{\mu} = \overline{y}_{++} \quad \hat{\tau}_i = \overline{y}_{i+} - \overline{y}_{++} \quad \hat{\beta}_j = \overline{y}_{+j} - \overline{y}_{++}$$

where

$$\overline{y}_{++} = \frac{1}{tb} \sum_{i=1}^{t} \sum_{j=1}^{b} y_{ij} \quad \overline{y}_{i+} = \frac{1}{b} \sum_{j=1}^{b} y_{ij} \quad \overline{y}_{+j} = \frac{1}{t} \sum_{i=1}^{t} y_{ij}$$

The variance estimate is

$$\hat{\sigma}^2 = \frac{1}{(t-1)(b-1)} \sum_{i=1}^{t} \sum_{j=1}^{b} (y_{ij} - \hat{\mu} - \hat{\tau}_i - \hat{\beta}_j)^2$$

6.3.3 ANOVA Table

Source	SS	df	MS	F
Treatment	$\sum_{i=1}^{t} b(\overline{y}_{i+} - \overline{y}_{++})^2$	$t-1$	$\frac{SS_{treatment}}{t-1}$	$\frac{MS_T}{MS_R}$
Blocks	$\sum_{j=1}^{b} t(\overline{y}_{+j} - \overline{y}_{++})^2$	$b-1$	$\frac{SS_{blocks}}{b-1}$	
Residuals	$\sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{i+} - \overline{y}_{+j} + \overline{y}_{++})^2$	$(t-1)(b-1)$	$\frac{SS_{error}}{(t-1)(b-1)}$	
Total	$\sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{++})^2$	$tb-1$		

6.3.4 Contrasts

$$\tilde{\theta} = \sum_i a_i \overline{Y}_{i+} \sim N(\theta, \sigma^2 \sum_i a_i^2/b)$$

Then

$$(\tilde{\theta} - \theta)/\sqrt{\hat{\sigma}_r^2 \sum_i a_i^2/b} \sim t_{(t-1)(b-1)}$$

- Confidence interval for θ: $\hat{\theta} \pm t * \hat{\sigma}\sqrt{\sum_i a_i^2/b}$
- Obversed test statistic for $H_0 : \theta = \theta_0$: $t_{obs} = (\hat{\theta} - \theta_0)/\hat{\sigma}\sqrt{\sum_i a_i^2/b}$

6.3.5 Notes

Use blocking design makes it easier to detect a treatment effect if there truly is one to detect in the first place

7 Factorial Treatment Structure and Interaction

Interaction occurs if the effect of one factor on the response depends on the level of a second factor

7.1 Two Factors at Two Levels

7.1.1 Model

To express interaction explicitly we write the model as

$$Y_{ijk} = \mu + \alpha_i + \lambda_j + \gamma_{ij} + R_{ijk} \quad i = 1, 2 \quad j = 1, 2 \quad k = 1, 2, \cdots, r$$

where

- μ: mean response across all treatments
- α_i: effect of factor A at level i
- λ_j: effect of factor B at level j
- γ_{ij}: interaction effect of factor A at level i and factor B at level j

Constraints

- For the main effects we have $\alpha_1 + \alpha_2 = 0$ and $\lambda_1 + \lambda_2 = 0$
- Then for the interaction effects we have

$$\sum_i \gamma_{ij} = 0 \quad \sum_j \gamma_{ij} = 0$$

7.2 Two Factors with More Than Two Levels

7.2.1 General Interaction Model

$$Y_{ijkl} = \mu + \alpha_i + \lambda_j + \gamma_{ij} + R_{ijk}$$

where

- μ: overall mean response
- α_i: effect of factor A at level i
- λ_j: effect of factor B at level j
- γ_{ij}: interaction effect of factor A at level i and factor B at level j

7.2.2 Interaction and Blocking Model

$$Y_{ijkl} = \mu + \alpha_i + \lambda_j + \gamma_{ij} + \beta_k + R_{ijk}$$

where

- μ: overall mean response
- α_i: effect of factor A at level i
- λ_j: effect of factor B at level j
- γ_{ij}: interaction effect of factor A at level i and factor B at level j
- β_k: effect of block k

The estimates we got are

- $\hat{\mu} = \overline{y}_{+++}$
- $\hat{\alpha}_i = \overline{y}_{i++} - \overline{y}_{+++}$
- $\hat{\lambda}_j = \overline{y}_{+j+} - \overline{y}_{+++}$
- $\hat{\gamma}_{ij} = \overline{y}_{ij+} - \overline{y}_{i++} - \overline{y}_{+j+} + \overline{y}_{+++}$
- $\hat{\beta}_k = \overline{y}_{++k} - \overline{y}_{+++}$

7.2.3 ANOVA Table

Source	Sum of squares	DF	Mean square
Treatments	$b\sum_{i,j}(\overline{y}_{ij+} - \overline{y}_{+++})^2$	$t_a t_b - 1$	$\frac{SS_T}{(t_a t_b - 1)}$
- Factor A	$t_b b\sum_{i}(\overline{y}_{i++} - \overline{y}_{+++})^2$	$t_a - 1$	$\frac{SS_a}{(t_a - 1)}$
- Factor B	$t_a b\sum_{j}(\overline{y}_{+j+} - \overline{y}_{+++})^2$	$t_b - 1$	$\frac{SS_b}{(t_b - 1)}$
- Interactions	$\mathrm{b}\sum_{i,j}(\overline{y}_{ij+} - \overline{y}_{i++} - \overline{y}_{+j+} + \overline{y}_{+++})^2$	$(t_a - 1)(t_b - 1)$	$\frac{SS_I}{(t_a - 1)(t_b - 1)}$
Blocks	$t_a t_b\sum_{k}(\overline{y}_{++k} - \overline{y}_{+++})^2$	$b - 1$	$\frac{SS_B}{(b - 1)}$
Residual	$\sum_{i,j,k}(y_{ijk} - \overline{y}_{ij+} - \overline{y}_{++k} + \overline{y}_{+++})^2$	$(t_a t_b - 1)(b - 1)$	$\frac{SS_R}{(t_a t_b - 1)(b - 1)}$
Total	$\sum_{i,j,k}(y_{ijk} - \overline{y}_{+++})^2$	$t_a t_b b - 1$	

The test statistic is the mean square ration between the test and error