My Final College Paper

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Naomi A. Boss

December 2020

Approved for the Division
(Mathematics)

Andrew P. Bray

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Chapter 1

thesisdown::thesis_gitbook: default

Placeholder

Chapter 2

Literature Review

2.1 Survey Methodology

```
  (2)When you are trying to make asertain something about a population, the most a
  - Sampling Procedures
(1-9)

Probability sampling is used in surveying.
     - Methods (section likely mixed rather than divided like so)
```

Simple Random Sampling (1)Simple reandom sampling is a procedure in which you take n independent samples from a population of size N. There are two ways of executing a simple random sample, with replacement and without replacement. If you survey with replacement, every sample has probabiliy 1/N of being drawn. The advantage of surveying without replacement is that you do not have repeated observations in the sample. In finite population sampling a duplicate observation does not provide new information, so it is more usefule to sample without replacement. The samples in the population are still considered indepdent samples; however, since the population decreases by one every draw, the probability becomes $\frac{n!\text{N-n!}}{N!}$. (1)

(2)When you draw n distinct units of observation(unit) from a population of size N without replacement, it is called *Simple Random Sampling*. Each unit is indepdent of one another with equal probabilty of being drawn. The probability of drawing a specific unit from the population is 1/N. You can use a random number generator on a computer to determine which units to include in your sample. Since each unit has equal probability of being drawn, the population mean is the average of the observations:

$$\mu = \frac{1}{N}\left(x_1 + x_2 + \cdots + x_n\right) = \frac{1}{N}\sum_{i=1}^{N} x_i$$

with a population variance of

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{n}\left(x_i - \mu\right)$$

Similarly, sample mean and variance are:

$$\overline{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) = \sum_{i=1}^{n} x_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_1 - \overline{x}\right)^2$$

You can also calculate the variance of the estimators themselves. If you use the population variance, you derive the variance of the estimator:

$$\operatorname{var}\left(\overline{x}\right) = \left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}$$

If you use the sample variance, then you get an unbiased estimator of the variance:

$$\widehat{\operatorname{var}}\left(\overline{x}\right) = \left(\frac{N-n}{N}\right)\frac{s^2}{n}$$

A variation of simple random sampling takes samples withoutreplacement. The draws are still independent of one another, but repeated observations is now allowed. This is used in situations where you do not want repeated observations. Now, instead of each draw being equal proability, we have that every sequence of samples is of equal probability. This slightly alters the sample mean

$$\overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

with a variance of

$$\operatorname{var}\left(\overline{x}_n\right) = \frac{1}{nN}\sum_{i=1}^{N}\left(x_i = \mu\right)^2 = \frac{N-1}{nN}\sigma^2$$

The unbiased estimator of the variance of the sample mean is

$$\widehat{\operatorname{var}}\left(\overline{x}_n\right) = \frac{s^2}{n}$$

(2) **Stratified:** (1)Stratified random sampling is a variation of sampling in which you partition the population into disjoint subpopulations that you sample from. The overall outcome of the survey is the combination of independent probability samples taken from the disjoint groups. There are several reasons why stratified random sampling can be advantageous. Stratified sampling can provide more precise estimates, especially if your aim involves comparing subgroups. Moreover, it allows flexibility in the method of conducting the surveye by customizing for each strata.(1)

(2)Another method of sampling is called stratified sampling, where you separate the population into disjoint subpopulations algorithmically. Then you can sample from each strata independenty. You can combine the mean and variations of each strata to caluclate the total variance and mean of the sample. One goal when designing the partition, is to minimize the variance within each strata. If the sampling procedure of each strata is simple random sampling, then the method is referred to as stratified random sampling. Both the sample mean and the ith observation are random variables Here we can define elements and probabilities for stratified sampling. Let h refer to a specific strata, with the population partitioned into L stratas. Let us define the terms and variables associated with stratified sampling. Let h be a designator of the strata. The total population size is the sum of the sizes of each strata, N_h. The sample size is the sum of the samples taken from each strata, denoted n_h. We define τ as the sum of the unit values in the population, and τ_h to be the sum of the unit values in strata h. The population mean is μ/N and the mean of each strata is $\mu_h = \tau_h/N_h$. Let $\hat{\tau}_h$ be the unbiased estimator of τ_h. The unbiased estimator of the population is

$$\hat{\tau}_{st} = \sum_{h=1}^{L} \hat{\tau}_h$$

The variance of this estimator is the sum of the variances of the strata estimators:

$$\operatorname{var}\left(\hat{\tau}_{st}\right) = \sum_{h=1}^{L} \operatorname{var}\left(\hat{\tau}_h\right)$$

with an unbiased variance is similarly equal to the sum of the unbiased variances of each strata. Regardless of sampling method, we can define the stratified estimator of the mean to be

$$\hat{\mu}_{st} = \hat{\tau}_{st}/N$$

The unbiased variance of the estimator, given that the sample selection from each strata are indepdent, is

$$\widehat{\operatorname{var}}\left(\hat{\mu}_{st}\right) = \frac{1}{N^2} \sum_{h=1}^{L} \hat{\operatorname{var}}\left(\hat{\tau}_{st}\right)$$

(2) **Cluster:** M: total number of seconary units N: the number of primary units M_i total number of secondary units in the ith primary unit x_i: the total of the values in the ith primary unit x_{ij}: the jth secondary unit in the ith primary unit τ: population total of secondary units μ: population mean per secondary unit μ_1: population mean per primary unit $\hat{\tau}$: unbiased estimator of population total var($\hat{\tau}$): variance of the unbiased population total estimator σ_u^2: variance of the primary unit totals $\widehat{\operatorname{var}}(\hat{\tau})$: unbiased estimate of the variance s_u^2: sample variance of the primary unit totals $\hat{\tau}_r$: ratio estimator of population total r: sample ratio p_i: probability of draw proportional to ith primary unit size.

(1)Cluster sampling is similar to stratified sampling because both methods involve segmenting the primary population; however, the choice for how the population is split is different. A unit is included in the sample if it falls within the partition of the cluster. When calculating probabilities, you need to consider how many clusters there are, as well as how many units are within each cluster. A cluster sample can be advantageous because it is cheaper to conduct than other methods. Moreover, populations are often distributed in natural clusters such as comparing universities. (1)

(2)The structure of cluster sampling and systemic sampling is the same. First the population is partitioned into units called *primary units*. Within each unit was have *secondary units*, or the units of observations. The primary unit of a cluster sample is determined by spacing proximity. On the other hand, the primary unit of a systemic sample is determined by equally spaced secondary units. The key aspect of cluster sampling and systematic sampling is that you choose the sample by selecting primary units, therefore, all secondary units in a primary unit will be included in the sample. There is variability in characterisitics of the primary units in cluster sampling, such as size and shape, which can provide supplementary information but can affect the efficiency of the sample.

The selection of primary units is conducted by other sampling procedures.You can use simple random sampling to select a primary unit. Calculating the unbiased estimators using simple random sampling without replacement: Estimator of population total

$$\hat{\tau} = \frac{N}{n} \sum_{i=1}^{n} = N\overline{y}$$

where $\overline{y} = (1/n)\sum_{i=1}^{n} y_i$ is the sample mean of the primary unit totals. Let σ_u^2 be the finite population variance of the primary unit totals

$$\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_i)^2$$

Then the variance of the population total estimator is

$$\text{var}\left(\hat{\tau}\right) = N\left(N-n\right)\frac{\sigma_u^2}{n}$$

and an unbiased variance is

$$\widehat{\text{var}}\left(\tau^2\right) = N\left(N-n\right)\frac{s_u^2}{n}$$

s_u^2 is the sample variance of the primary unit totals, calculated by

$$s_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

The ratio estimator based on size can be more efficient than simple random sampling. The primary unit totals for cluster sampling do not have to be the same size, so the unit total is highly correlated to unit size. The sample ratio r is defined as

$$r = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} M_i}$$

Then the ratio estimator of population total is

$$\widehat{\tau}_r = rM$$

You can approximate the variance or mean squared error of the ratio estimator

$$\operatorname{var}(\hat{\tau}_r) \approx \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{N} (x_i - M_i\mu)^2$$

which has an estimator

$$\widehat{\operatorname{var}}(\hat{\tau}_r) \approx \frac{N(N-n)}{n(N-1)} \sum_{i=1}^{n} (x_i - rM_i)^2$$

The population mean per primary unit is calculated by dividing the ratio estimator by the number of primary units. (2)

(1) Systemactic sampling is a specific kind of cluster sampling. First decide on the number of samples n. Choose a random number between 1 and the number of clusters, this will (1)

(2) The Hansen-Hurwits estimator is a cluster sampling method which proportionalizes the probability of drawing the primary unit based on the size. The unbiased Hansen-Hurwitz estimator is

$$\hat{\tau}_p = \frac{M}{n} \sum_{i=1}^{n} \frac{x_i}{M_i}$$

```
- Advantages/Disadvantages
pg 255
Unequal sampling - chp 6
pg 49
113-125
```

- Data Collection
  - Designing survey (sample unit, sample size, hypothesis/motivation role) pg 247 pg 258 31-33
  - Survey method (brief list e.g. interview, questionnaire)
- Data Analysis
  - Hypothesis testing
  - Inference

2.2 Non-Probability Sampling

```
- Sampling Procedures
     - Methods
          - Convenience sampling
          - Consecutive sampling
          - Judgmental Sampling
          -Hansen-Hurwitz
          46-49
     ##Advantages/Disadvantages
- Data Collection
     - Designing Survey (sample unit, sample size, hypothesis/motivation role)
     - Survey method (brief list e.g. interview, questionnaire)
- Data Analysis
67
94-99
     - Hypothesis testing
     - Inference
```

2.3 Election Polling

```
- Purpose
- Design
     - Population
          - Sample selection
     - Methods
     - Sample size
     - Question design
          - Wording impacting response
- Data collection
     - Implementation
     - Errors
          - Sampling Error
          - Non-Sampling Error
               - Questionaire design
               - Interviewer effects
               - Responder effects
- Data analysis
```

Unclear Questions: Simple random sample implies that an unit has a 1/N probability of being drawn. What is the distinction from the probability than an ith unit of the popultion is included, which is n/N

Soruces: 1) “Sampling: Design and Analysis” Sharon L. Lohr
2)“Sampling” Steven K. Thompson

3) “Design an Inference in Finite Population Sampling” A. S. Hedayat, Bikas K. Sinha

Chapter 3

tbd

Chapter 4

Tables, Graphics, References, and Labels

Placeholder

4.1 Tables

4.2 Figures

4.3 Footnotes and Endnotes

4.4 Bibliographies

4.5 Anything else?

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

In Chapter 4:

Appendix B

The Second Appendix, for Fun

References

Placeholder