# STUDY ON PARAMETER ESTIMATION VIA MULTI-STAGE SAMPLING WITH APPLICATIONS

by

Francis Bilson Darku

APPROVED BY SUPERVISORY COMMITTEE:

Bhargab Chattopadhyay, Co-Chair

Frank Konietschke, Co-Chair

Swati Biswas

Pankaj K. Choudhary

*This dissertation is dedicated to my family, fiancée and all who believed in me and have supported me one way or the other.*

# STUDY ON PARAMETER ESTIMATION VIA MULTI-STAGE SAMPLING WITH APPLICATIONS

by

FRANCIS BILSON DARKU, BSc, MSc

DISSERTATION

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY IN

STATISTICS

THE UNIVERSITY OF TEXAS AT DALLAS

August 2018

# ACKNOWLEDGMENTS

I take this opportunity to express my heartfelt gratitude to all who have influenced and guided me in producing this wonderful work. First and foremost, I would like to express my appreciation to Dr. Bhargab Chattopadhyay and Dr. Frank Konietschke for their immense guidance, support, advice and time throughout my journey as a PhD student. Their patience and tolerance made them very approachable at all times (literally). My thanks also goes to Dr. Swati Biswas and Dr. Pankaj Kumar Choudhary for serving on my advising committee, providing me with valuable feedback, and granting me the opportunities to serve as their teaching assistant. Furthermore, I am thankful to Dr. Ken Kelley of University of Notre Dame for his collaboration and advice on my future career. I would also like to acknowledge all the Statistics professors in the Department of Mathematical Sciences for the knowledge they imparted in me throughout my studies. To all the secretaries in Mathematical Sciences Department (both present and past), I say a big thank you for patiently helping me with all my paperwork and making my studies easier at UTD. My special thanks also goes to Dr. Robert Serfling for his assistance during my admission process and the Department of Mathematical Sciences for the funding opportunity to pursue my PhD. To all the TAs and RAs, it has been a great time working with you all, as well as serving some of you as your TA. I am grateful to Joanna Agyapomaa Aduhene, Dorcas Ofori-Boateng, Josephine Sarpong Akosa and Dr. Marinelle Rodrigues for proofreading my thesis amidst their busy schedule and different area of expertise. I also acknowledge the support of the following families who made my stay in Dallas wonderful: the Ogbolu family, the Ayisah-Eyeson family and the Rozario family. I also appreciate Newman Catholic Ministry members, the African and Ghanaian communities, and my roommates for their company and support. Lastly, but most importantly, I would like to thank the Almighty God for His grace and mercy that has brought me thus far.

April 2018

# STUDY ON PARAMETER ESTIMATION VIA MULTI-STAGE SAMPLING WITH APPLICATIONS

Francis Bilson Darku, PhD
The University of Texas at Dallas, 2018

Supervising Professors: Bhargab Chattopadhyay, Co-Chair
Frank Konietschke, Co-Chair

Over the past few decades, researchers have been (and are still being) encouraged to report confidence intervals along with their parameter estimates instead of just the binary outcome of a hypothesis testing based on an arbitrary cut of value for $p - value$ (mostly $\alpha = 5\%$). However, researchers traditionally define their sample sizes in advance before sampling is done. This naturally may lead to wide confidence intervals. Ceteris paribus, wider confidence intervals indicate higher uncertainty and this discourages researchers from reporting the confidence intervals. As a remedy to this problem, sample size planning methods, such as accuracy in parameter estimation and power analysis, were developed. These methods seek to determine the appropriate sample sizes needed to obtain sufficiently narrow confidence intervals in the case of accuracy in parameter estimation, or high power in the case of power analysis. One drawback of these methods is that they require researchers to provide the values of some population parameters which are generally unknown in advance. Thus, the use of suppose population values, which are different from their true population values, in these methods will result in wrong sample size calculations. Incorrect sample sizes then also lead to incorrect inferences or decisions. Another drawback of these traditional methods is the assumption of the distribution from which the data are sampled. There is no reason to assume that data will always follow a particular distribution, say normal, in every situation. To overcome these challenging assumptions, multi-stage procedures which have been around for more than half a century can be used. We therefore develop multi-stage sampling procedures for constructing sufficiently narrow confidence intervals for parameters with a pre-specified confidence level and pre-specified upper bound on the width of the confidence interval. We do this for a general class of effect sizes, different types of correlation measures, and the Gini index. Our methods do not require the knowledge of population parameters or the distribution from which the data are sampled. In other words, our methods work in a distribution-free environment with no requirement for knowledge of population values. In our procedure, the sample size needed to obtain a sufficiently narrow confidence is not specified a priori. Rather, a stopping rule, which will be defined, determines whether after a pilot sample is obtained, additional samples will be needed or not. We provide theorems with their proofs to support our procedures and demonstrate their characteristics with Monte Carlo simulations. In the case of the Gini index, we also provide an application to the 64th National Sample Survey in India.

## TABLE OF CONTENTS

ACKNOWLEDGMENTS . . . . . . . . . . v

ABSTRACT . . . . . . . . . . vi

LIST OF TABLES . . . . . . . . . . xii

CHAPTER 1 INTRODUCTION . . . . . . . . . . 1

1.1 Overview . . . . . . . . . . 1

1.2 Bounded-Width Confidence Interval Problem . . . . . . . . . . 1

1.3 Multi-Stage Sampling Procedures . . . . . . . . . . 3

1.3.1 Two-Stage Procedure . . . . . . . . . . 4

1.3.2 Purely Sequential Procedure . . . . . . . . . . 4

1.3.3 Other Procedures . . . . . . . . . . 5

1.3.4 Properties of Sequential Approach . . . . . . . . . . 7

1.4 Random Central Limit Theorem . . . . . . . . . . 10

1.5 U-statistics, Hoeffding's Decomposition, and Asymptotic Variance . . . . . . 11

1.6 Chapter Organization . . . . . . . . . . 12

CHAPTER 2 A PURELY SEQUENTIAL APPROACH TO ACCURACY IN PARAMETER ESTIMATION FOR A GENERAL CLASS OF EFFECT SIZES . . . . . . 14

2.1 Introduction . . . . . . . . . . 14

2.2 Effect Sizes Based on Ratio of Two Parameters: A General Framework . . . 18

2.2.1 Example 1: Standardized Mean Difference . . . . . . . . . . 19

2.2.2 Example 2: Coefficient of Variation . . . . . . . . . . 21

2.2.3 Example 3: The Standardized Mean . . . . . . . . . . 21

2.2.4 Example 4: Regression Coefficient in Simple Linear Model . . . . . . 22

2.2.5 Example 5: Effect Size for Ordinal Data . . . . . . . . . . 23

2.2.6 Example 6: Contrasts . . . . . . . . . . 23

2.2.7 Example 7: Univariate Parameters and Their Functions . . . . . . . . 24

2.3 Central Limit Theorem for $T_n$ . . . . . . . . . . 24

2.4 Accuracy in Parameter Estimation of General Effect Size Parameter: A Fixed-Sample Size Approach . . . . . . . . . . 27

2.5 Accuracy in Parameter Estimation via a Sequential Optimization Procedure 29

2.5.1 Characteristics of Our Sequential Procedure . . . . . . . . . . . . . . 31

2.6 Random Central Limit Theorem for $T_n$ . . . . . . . . . . . . . . . . . . . . . 32

2.7 Application to Some Widely Used Effect Sizes . . . . . . . . . . . . . . . . 36

2.7.1 Standardized Mean Difference . . . . . . . . . . . . . . . . . . . . . 36

2.7.2 Coefficient of Variation . . . . . . . . . . . . . . . . . . . . . . . . . . 38

2.7.3 Regression Coefficient: Simple Linear Model . . . . . . . . . . . . . . 41

2.8 An Extension: Unbalanced Design . . . . . . . . . . . . . . . . . . . . . . . 43

2.9 Concluding Remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 45

CHAPTER 3 A PURELY SEQUENTIAL APPROACH TO ACCURACY IN PARAMETER ESTIMATION FOR POPULATION CORRELATION COEFFICIENTS . 51

3.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 51

3.2 Accuracy in Parameter Estimation of Pearson's Product Moment Correlation Coefficient . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55

3.3 Accuracy in Parameter Estimation via a Sequential Optimization Procedure 57

3.4 Lemmas and Theorems . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

3.5 Characteristics of the Final Sample Size for Pearson's Product Moment Correlation: A Simulation Study . . . . . . . . . . . . . . . . . . . . . . . . 62

3.6 Alternative Confidence Intervals for Pearson's Correlation Coefficient . . . . 65

3.6.1 Confidence Interval by Corty and Corty (2011) . . . . . . . . . . . . . 66

3.6.2 Confidence Interval in Method 4 by Moinester and Gottfried (2014) . 68

3.6.3 Simulation Study . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 69

3.7 Sequential AIPE for Kendall's Tau and Spearman's Rho . . . . . . . . . . . 70

3.7.1 AIPE for Kendall's $\tau$ . . . . . . . . . . . . . . . . . . . . . . . . . . . 71

3.7.2 AIPE for Spearman's $\rho$ . . . . . . . . . . . . . . . . . . . . . . . . . . 73

3.8 An Extension: Squared Multiple Correlation Coefficient . . . . . . . . . . . . 77

3.8.1 Derivation of Optimal and Pilot Sample Sizes for Confidence Intervals for Squared Multiple Correlation Coefficient . . . . . . . . . . . . . 81

3.8.2 Simulation Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83

3.9 Discussion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83

CHAPTER 4 GINI INDEX ESTIMATION WITHIN PRE-SPECIFIED ERROR BOUND APPLIED TO INDIAN HOUSEHOLD SURVEY DATA . . . . . . . . . . . . . . . 90
4.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 90
4.1.1 Contributions of This Paper . . . . . . . . . . . . . . . . . . . . 93
4.2 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94
4.3 Sequential Methodology . . . . . . . . . . . . . . . . . . . . . . . . . 96
4.3.1 Estimation for $\xi^2$ . . . . . . . . . . . . . . . . . . . . . . . . . 96
4.3.2 Purely Sequential Procedure . . . . . . . . . . . . . . . . . . . . 98
4.3.3 Two-stage Procedure . . . . . . . . . . . . . . . . . . . . . . . . 99
4.3.4 Pilot Sample Size . . . . . . . . . . . . . . . . . . . . . . . . . . 100
4.4 Application . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 100
4.4.1 Application of Purely Sequential Procedure . . . . . . . . . . . . . . 102
4.4.2 Application of Two-Stage Procedure . . . . . . . . . . . . . . . . . . 103
4.5 Characteristics of the Procedures . . . . . . . . . . . . . . . . . . . . 107
4.6 Replication Using Empirical Data . . . . . . . . . . . . . . . . . . . . 109
4.6.1 Replication Using Empirical Data: Purely Sequential Procedure (PSP) 109
4.6.2 Replication Using Empirical Data: Two-Stage procedure . . . . . . . 110
4.7 Replication Using a Pseudo Population . . . . . . . . . . . . . . . . . 113
4.8 Concluding Thoughts . . . . . . . . . . . . . . . . . . . . . . . . . . . 117
4.9 Acknowledgement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 118
CHAPTER 5 SUMMARY AND CONCLUSION . . . . . . . . . . . . . . . . . 119
5.1 Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 119
5.2 Concluding Remarks . . . . . . . . . . . . . . . . . . . . . . . . . . . 120
5.3 Future Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 120
APPENDIX A CENTRAL LIMIT THEOREM FOR SOME EFFECT SIZES . . . . 122
A.1 Standardized Mean Difference . . . . . . . . . . . . . . . . . . . . . . 122
A.2 Coefficient of Variation . . . . . . . . . . . . . . . . . . . . . . . . . . 123
A.3 Regression Coefficient: Simple Linear Model . . . . . . . . . . . . . . . 124

APPENDIX B CONSISTENT ESTIMATOR FOR ASYMPTOTIC VARIANCE OF PEARSON'S CORRELATION COEFFICIENT . . . . . . . . . . . . . . . . . . . . . . 126

APPENDIX C SIMULATION RESULTS FOR CONFIDENCE INTERVAL FOR THE GINI INDEX . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 128

C.1 Pure Sequential Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 128

C.2 Two-Stage Procedure . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 132

REFERENCES . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 134

BIOGRAPHICAL SKETCH . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 144

CURRICULUM VITAE

## LIST OF TABLES

2.1 Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\delta$ with $\omega = 0.2$ 39

2.2 Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\kappa$ with $\omega = 0.04$ 42

3.1 Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\rho$ . . . . . 63

3.2 Simulation results for $100(1-\alpha)\%$ confidence interval for $\rho$ for fixed $n$ . . . . . 65

3.3 Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\rho$ using Corty and Corty (2011) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 70

3.4 Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\rho$ using Moinester and Gottfried (2014) . . . . . . . . . . . . . . . . . . . . . . . . . . 71

3.5 Summary of final sample size for $100(1-\alpha)\%$ confidence interval for Kendall's $\tau$ using asymptotic distribution . . . . . . . . . . . . . . . . . . . . . . . . . . . 74

3.6 Summary of final sample size for $100(1-\alpha)\%$ confidence interval for Spearman's rho, $\rho_s$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 78

3.7 Summary of final sample size for 90% confidence interval for $P^2$ using Bonett and Wright (2011) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 86

3.8 Summary of final sample size for 95% confidence interval for $P^2$ using Bonett and Wright (2011) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 87

3.9 Summary of final sample size for 90% confidence interval for $P^2$ using Olkin and Finn (1995) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 88

3.10 Summary of final sample size for 95% confidence interval for $P^2$ using Olkin and Finn (1995) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89

4.1 Application results for PSP on NSS 64th Round Data for $\omega = 0.02$ . . . . . . . . 103

4.2 Application results for PSP on NSS 64th Round Data for $\omega = 0.025$ . . . . . . . 104

4.3 Application results for the two-stage procedure on NSS 64th Round Data for $\omega = 0.02$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106

4.4 Application results for the two-stage procedure on NSS 64th Round Data for $\omega = 0.025$ . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 107

4.5 Replication results for PSP on 64th NSS Data ($\omega = 0.02$) . . . . . . . . . . . . . 111

4.6 Replication results for PSP on 64th NSS Data ($\omega = 0.025$) . . . . . . . . . . . . 112

4.7 Replication results for two-stage procedure on 64th NSS Data ($\omega = 0.02$) . . . . 114

4.8 Replication results for two-stage procedure on 64th NSS Data ($\omega = 0.025$) . . . . 115

C.1 Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.02$, $m' = 1$) . 128

C.2 Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.02$, $m' = 1$) . 128
C.3 Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.025$, $m' = 1$) 129
C.4 Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.025$, $m' = 1$) . 129
C.5 Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.02$, $m' = 10$) 129
C.6 Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.02$, $m' = 10$) . 130
C.7 Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.025$, $m' = 10$) 130
C.8 Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.025$, $m' = 10$) 130
C.9 Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.02$, $m' = 20$) 131
C.10 Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.02$, $m' = 20$) . 131
C.11 Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.025$, $m' = 20$) 131
C.12 Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.025$, $m' = 20$) 132
C.13 Simulation results for two-stage procedure ($\alpha = 10\%$, $\omega = 0.02$) . . . . . . . . . 132
C.14 Simulation results for two-stage procedure ($\alpha = 5\%$, $\omega = 0.02$) . . . . . . . . . . 132
C.15 Simulation results for two-stage procedure ($\alpha = 10\%$, $\omega = 0.025$) . . . . . . . . . 133
C.16 Simulation results for two-stage procedure ($\alpha = 5\%$, $\omega = 0.025$) . . . . . . . . . 133

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

With many researchers and regulatory authorities advocating for the reporting of effect sizes and their corresponding confidence intervals, this dissertation is concerned with improving the accuracy and precision of such confidence intervals. This is achieved by developing sequential procedures for constructing confidence intervals that have sufficiently narrow width at given confidence level. The methods do not rely on data distribution, that is, they do not make any distribution assumption about data. This chapter may contain some overlapping information with the remaining chapters because they are heavily drawn from fully developed manuscripts that either have been published, are under review, or yet to be submitted to a journal for publication.

## 1.2 Bounded-Width Confidence Interval Problem

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables from a proper distribution function $F(x)$ with an unknown parameter $\theta$. Let $\hat{\theta}_n$ be a consistent estimator of $\theta$ based on a sample of size $n$. Without loss of generality, let

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta)}{\xi} \to \mathscr{F}(x) \quad \text{as} \quad n \to \infty \tag{1.1}$$

then a $100(1-\alpha)\%$ confidence interval for $\theta$ is given as

$$J_n = \left(\hat{\theta}_n - q_{\alpha/2}\frac{\xi}{\sqrt{n}}, \hat{\theta}_n + q_{\alpha/2}\frac{\xi}{\sqrt{n}}\right) \tag{1.2}$$

where, $q_\alpha$ is the $100\alpha\%$ upper quantile of the distribution $\mathscr{F}$. The width of $J_n$ is

$$w_n = \frac{2q_{\alpha/2}\xi}{\sqrt{n}}. \tag{1.3}$$

All things being equal, a narrow confidence interval is preferred as it shows higher precision and accuracy (Anscombe, 1949). Thus, it is appropriate to bound the width of the confidence interval by a pre-specified value, say $\omega$. This means that the maximum width of the confidence interval that is acceptable is $\omega$ and any width $w_n$ less than $\omega$ is also preferred. If $\xi^2$ is known, then the smallest sample size needed to achieve the desired width $\omega$ and the confidence level $100(1-\alpha)\%$ is given by

$$n \geq \left\lceil \frac{4q^2_{\alpha/2}\xi^2}{\omega^2} \right\rceil, \tag{1.4}$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to $x$. On the other hand, if $\xi^2$ is unknown but it is known that $\xi^2$ is bounded above by $\xi_0^2$, i.e. $0 < \xi^2 \leq \xi_0^2$ then

$$n \geq \left\lceil \frac{4q^2_{\alpha/2}\xi_0^2}{\omega^2} \right\rceil \tag{1.5}$$

will guarantee that we meet our pre-specified goal (Ghosh and Mukhopadhyay, 1976). Nonetheless, $\xi^2$ (or its upper bound $\xi_0^2$) is unknown in most, if not all, situations. The use of supposed values of $\xi^2$ may lead to inaccurate sample sizes that may not satisfy Equation (1.4) or (1.5).

Unfortunately, there is no fixed sample size procedure that can guarantee that the population parameter $\theta$ will lie within the confidence interval with pre-assigned confidence level and width (Anscombe, 1953; Dantzig, 1940; Ghosh and Mukhopadhyay, 1976; Mukhopadhyay and De Silva, 2009). To solve this bounded-width problem, two steps have to be followed according to Anscombe (1953): (i) obtain a suitable sampling distribution of $\hat{\theta}_n$ using large sample size theory, as exact small-sample solutions may be very difficult to formulate and at the same time achieve a narrow width (Anscombe, 1949), (ii) develop an appropriate sequential procedure to obtain the sample size $n$. Thus, by step (ii), the solution to this bounded-width confidence interval problem lies within the domain of sequential analysis. The next section discusses several sequential procedures and this will be followed by the large sample size theory for sampling distributions of statistics under sequential procedures.

## 1.3 Multi-Stage Sampling Procedures

The first person known to have used the term *sequential* in the statistics context is Abraham Wald in his work on hypothesis testing in which the sample size was not fixed in advance (Anscombe, 1953). He proposed a sequential method of testing a hypothesis based on three decisions - (1) accept the hypothesis, (2) reject the hypothesis, or (3) sample additional observations. Thus, given a set of observations, one of the three decisions is made. If either decision (1) or (2) is made, sampling is terminated. If decision (3) is made then an additional observation is collected and the hypothesis is tested again for any of the three mentioned outcome. The process is continued sequentially until either decision (1) or (2) is made. In addition, Wald (1947) suggested in his *Sequential Analysis* book that sequential procedures can be used for constructing confidence intervals for unknown parameters.

In 1949, Anscombe showed that with the use of sequential procedures, one can estimate unknown parameters with some level of accuracy while using fixed-sample size formulae. However, his work was challenged by Cox (1952) who argued that Anscombe's work was more heuristic and lacked theoretical backing. In light of this, Anscombe developed the large-sample theory for sequential estimation in 1952. His works along with the results of Stein (1945), Ray (1957), Chow and Robbins (1965), among others, pioneered the field of sequential analysis. (Readers are referred to Mukhopadhyay and Chattopadhyay (2012) for more contributions of Anscombe.)

With the introduction of sequential analysis, the field of statistics can be pushed beyond the limits and assumptions of fixed sample size procedures. Researchers do not need to know the population values of a parameter(s) a priori or use supposed values. Sequential procedures have extended the boundaries of sample size calculations, power analysis, and accuracy in parameter estimation, in that researchers can now proceed with estimating and inferring parameters with little or no assumptions about the population distributions and required sample sizes. In some situations, sequential analyses can reduce the average sample size needed to minimize the desired level of error or maximize the level of accuracy of an estimator (Robbins, 1952).

A variety of sequential procedures, developed over the course of time, are discussed below in the next three subsections.

### 1.3.1 Two-Stage Procedure

The two-stage sequential procedure is known to have been introduced and used by Mahalanobis in 1940 to estimate the acreage and yield of Jute crop in Bengal, India (Mahalanobis, 1967). He had the idea of using pilot samples in large scale sample surveys to gain information about the sampling error. However, it was Stein's seminal work in 1945 and 1949 that laid down the statistical foundations for the two-stage procedure. In his work, he solved the problem of constructing a confidence interval with a pre-specified width and confidence level, for the mean $\mu$ of a normal distribution with unknown population variance $\sigma^2$. He proposed that observations should be taken in two stages: In the first stage, also called the pilot stage, a sample $X_1, \ldots, X_m$ of size $m(\geq 2)$ (pilot sample size) should be taken, based on which the sample variance $S_m^2$ is obtained. In the second stage, the final sample size $N$ needed to obtain a confidence interval whose width and confidence level meet the set criteria is computed. If $N = m$, no additional observations are needed. On the contrary, if $N > m$, $N - m$ additional observations are sampled in the second stage. This proposed procedure by Stein has become the bedrock for applying the two-stage procedure in several situations, with several modifications.

### 1.3.2 Purely Sequential Procedure

The two-stage procedure by Stein (1945) is known to overestimate the optimal sample size (Ghosh et al., 1997). As an alternative to the two-stage procedure, Anscombe (1952) introduced the idea of a purely sequential procedure which was later expanded by Ray (1957) and Chow and Robbins (1965). The goal of this procedure is to improve the estimate of the nuisance parameter(s) (which in this case is the variance $\sigma^2$ of the normal distribution) as each observation is collected until a stopping criterion is met. The improvement in the estimation of $\sigma^2$ eventually lead to an improvement in the estimate of the sample size needed to achieve the pre-specified width and confidence coefficient.

The purely sequential procedure is described as follows: Obtain the pilot sample $X_1, \ldots, X_m$ of size $m(\geq 2)$ at the initial stage and estimate $\sigma^2$ (using $S_m^2$) and estimate the final sample size $C$. If $m \geq C$ stop sampling. If $m < C$, add one more observation $X_{m+1}$ and update the estimation of $\sigma^2$ to $S_{m+1}^2$. Check if $m+1$ greater or equal to new estimate of the sample size $C$. If $m+1 \geq C$, sampling is terminated, otherwise an additional observation is collected. This process goes on until a sample of size $n$ that it is greater than or equal to the estimated sample size $C$ is attained. At this point sampling is terminated and the confidence interval is constructed as stated in Equation (1.2).

### 1.3.3 Other Procedures

Aside from the two main principal procedures already described many researchers have tried to develop hybrid and improved versions to harness the benefits of both methods. Some of these versions are discussed below.

#### Modified Two-Stage Procedure

In implementing Stein's (1945) two-stage procedure, one chooses the pilot sample size $m$ arbitrarily. However, there is still the question of how large or small should $m$ be. As the confidence level is increased, while the confidence interval width and others remain constant, it is expected that the optimal sample size and the estimated sample size increase. Also, as the confidence interval width gets narrower, while the confidence level and other parameters remain constant, the optimal and estimated sample sizes are expected to increase. In both of these cases, the pilot sample size should increase accordingly but not too much to exceed the optimal sample size. To solve this problem, Mukhopadhyay (1980) proposed a *modified* two-stage procedure. In his work, he proposed a formula for $m$ that depends on the pre-specified width and confidence coefficient of the confidence interval. As a result, the pilot sample size increases as the pre-specified width decreases or the confidence coefficient increases.

### Three-Stage Procedure

Since Stein's (1945) two-stage procedure oversamples most of the time and Anscombe's (1949) purely sequential procedure is operationally infeasible at times, Mukhopadhyay (1976) introduced the three-stage procedure. This procedure has an intermediated sampling stage between the pilot samples and the final samples. The aim of the intermediate stage is to improve the accuracy of the final sample size estimation. The result from this procedure is typically closer to the optimal sample size than what is obtained from the two-stage procedure. At the same time, it also involves fewer steps than the purely sequential procedure in the sense that sampling is done in only three stages as compared to the one-by-one sequential sampling used in the purely sequential procedure. Further research on the three-stage procedure can be found in Hall (1981), Ghosh et al. (1997), and Mukhopadhyay and De Silva (2009).

### Accelerated Sequential Procedure

As mentioned above, the purely sequential procedure may be inconvenient sometimes in some situations and one may require a procedure that can accurately estimate the final sample size but also reduce the sampling stages. Another procedure that is a good candidate for this situation, apart from the three-stage procedure, is the accelerated sequential procedure. This method is designed to first collect observations in a purely sequential manner up to a point where the number of observations $n$ is greater or equal to a fraction $\gamma$ of the estimated final sample size $C$, that is $n \geq \gamma C = t$ for $0 < \gamma < 1$. Then based on the collected observations, the final sample size can be estimated as $C = t/\gamma$ and used to determine how many more observations need to be collected in a single batch. More details about this procedure can be found in Hall (1983), Mukhopadhyay and Solanky (1991), Mukhopadhyay (1996), and Kumar et al. (2012).

**Parallel Piecewise Sequential Procedure**

In 1993, Mukhopadhyay and Sen introduced the parallel piecewise sequential procedure to take advantage of processes that naturally employ several operations that occur concurrently or in parallel forms. Examples of such processes include conducting a survey in different regions of a state or country, several bank tellers attending to customers, computer servers processing data simultaneously, customers checking out at supermarkets, etc. In the parallel piecewise sequential procedure, $k$ independent sequences of random variables $X_{ij}$, $j = 1, \ldots, n_i, \ldots$ and $k = 1, \ldots, k$ are observed concurrently from a distribution $F$. At any point, the observations are pooled together to estimate the unknown parameter(s) of interest and the appropriate inferences are made when the stopping rule(s) is(are) met. Readers can find more work on this procedure by Bose and Mukhopadhyay (1994), Mukhopadhyay and Datta (1994), and Mukhopadhyay and de Silva (1998).

### 1.3.4 Properties of Sequential Approach

There are some desirable properties that sequential procedures may exhibit. A sequential procedure that possess more of these properties may be preferred to one that has fewer. Below, we mention and discuss some of these properties.

### Consistency or Exact Consistency

A sequential procedure for constructing a $100(1-\alpha)\%$ confidence interval $J_N$ for an unknown parameter $\theta$ is said to be *(exactly) consistent* if

$$P_\theta \left\{\theta \in J_N\right\} \geq 1-\alpha, \quad \forall \theta \in \Theta \tag{1.6}$$

(Chow and Robbins, 1965). Here, $\Theta$ defines the parameter space of $F$. This means that after sampling is terminated for a sequential procedure the confidence interval constructed will achieve at least $100(1-\alpha)\%$ coverage probability as desired.

### Asymptotic Consistency

Now, to talk about asymptotic theory in sequential procedures, i.e. the final sample size $N$ is assumed to be large ($N \to \infty$). As the maximum acceptable width $\omega$ of the confidence interval tends to 0 ($\omega \to 0$), the sample size $N$ gets larger. This is intuitive because as a large sample size is needed to make a narrow confidence interval that still maintains $100(1-\alpha)\%$ coverage probability. Hence, the asymptotic properties of the sequential procedures introduced and/or discussed in this manuscript will be stated with limits as $\omega \to 0$.

An important large sample property that researchers would like sequential procedures to have is the *asymptotic consistency.* For a given stopping rule for constructing a $100(1-\alpha)\%$ confidence interval for an unknown parameter $\theta$, the asymptotic consistency property is defined as

$$\lim_{\omega \downarrow 0} P_\theta \left\{\theta \in J_N\right\} = 1-\alpha, \quad \forall \theta \in \Theta. \tag{1.7}$$

This property ensures that the confidence interval constructed from the given sequential procedure will have the anticipated $1-\alpha$ coverage probability for large sample sizes. It is worth noting that a procedure that posses the asymptotic consistency property does not necessarily possess the exact consistency property. An example of such a case is the purely sequential procedure for constructing $1-\alpha$ confidence interval for mean $\mu$ from a normal population with an unknown variance $\sigma^2$ (see Mukhopadhyay and De Silva, 2009, pp 111).

### Asymptotic (First Order) Efficiency

Another important property is the *asymptotic first order efficiency* or *asymptotic efficiency*. A sequential procedure is asymptotically efficient if

$$\lim_{\omega \downarrow 0} \frac{\mathrm{E}(N)}{n_\omega} = 1. \tag{1.8}$$

This ensures that as the desired width $\omega$ of the confidence interval tends to 0, the ratio of the estimated sample size, obtained through the given procedure, to the theoretical optimal sample size will be 1 on the average. In other words, a desired procedure should not overestimate or underestimate the optimal sample size on the average. Stein and Wald's ((1947), (1949)) two-stage procedure does not possess this property (Mukhopadhyay, 1980). It actually overestimates the optimal sample size (Ghosh et al., 1997).

### Other Properties

Aside the above main properties, there are other properties that most sequential procedures seek to achieve.

(i) **Finite Sample Size:** This is a very important property as it ensures that sampling will be terminated. This is usually stated as

$$P(N < \infty) = 1. \tag{1.9}$$

(ii) **Almost Sure Convergence:** A sequential procedure converges almost surely if

$$P\left(\lim_{\omega \downarrow 0} \frac{N}{n_\omega} = 1\right) = 1. \tag{1.10}$$

This property ensures that for large sample size (or very narrow width) the estimated final sample size will be close to the optimal sample size almost surely.

(iii) **Asymptotically Second Order Efficiency:** Ghosh and Mukhopadhyay (1981) introduced the notion of asymptotically second order efficiency to show that the purely sequential procedure for constructing a fixed-width confidence interval for mean with unknown variance performs better than the modified two-stage procedure. If a stopping rule for $N$ is asymptotically second order efficient then

$$\lim_{\omega\downarrow 0} \mathrm{E}(N - n_\omega) < \infty \tag{1.11}$$

## 1.4 Random Central Limit Theorem

For large fixed sample sizes, the central limit theorem (CLT) can be invoked to obtain the sampling distribution for most statistics. However, for a random sample size $N$, a justification for the use of the central limit theorem is needed. Anscombe (1952) introduced an analogous version of CLT for random sample sizes known as the *random central limit theorem.* In order to apply Anscombe's random central limit theorem for the computation of the asymptotic distribution of a statistic, Anscombe's *uniform continuity in probability* (u.c.i.p.) condition as stated below must be fulfilled:

Let $\{Y_n\}, n = 1, 2, \ldots$ be a sequence of i.i.d. random variable. Suppose that there exists a real number $\theta$, a sequence of positive numbers $w_n$, and a distribution function $\mathscr{F}(x)$, such that the following conditions are satisfied:

i. *Convergence of* $\{Y_n\}$*:* For any $x$ such that $\mathscr{F}(x)$ is continuous,

$$P(Y_n - \theta \leq xw_n) \rightarrow \mathscr{F}(x) \quad \text{as} \quad n \rightarrow \infty \tag{1.12}$$

ii. *Uniform continuity in probability of* $\{Y_n\}$*:* Given any small positive $\varepsilon$ and $\eta$, there is a large $\nu$ and small positive $c$ such that, for any $n > \nu$,

$$P\left\{|Y_{n'} - Y_n| < \varepsilon w_n \text{ simultaneously for all integers } n' \text{ such that } |n' - n| < cn\right\} > 1 - \eta \tag{1.13}$$

**Theorem 1.1** (Anscombe's (1952) Random Central Limit Theorem). *Let $\{n_r\}$ be an increasing sequence of positive integers tending to infinity, and let $\{N_r\}$ be a sequence of proper random variables taking positive integer values such that $N_r/n_r \to 1$ in probability as $r \to \infty$. Then if the sequence of random variables $\{Y_n\}$ satisfies conditions* (1.12) *and* (1.13),

$$P(Y_{N_r} - \theta \leq x w_{n_r}) \to \mathscr{F}(x) \quad \textit{as} \quad r \to \infty \tag{1.14}$$

*at all continuity points $x$ of $\mathscr{F}$.*

With the above theorem and the procedures discussed in subsection 1.3, the remaining chapters of this dissertation solve different bounded-width problems in the field of statistics and other disciplines.

## 1.5 U-statistics, Hoeffding's Decomposition, and Asymptotic Variance

U-statistics are a class of statistics introduced by Hoeffding (1948), which can be used to construct an unbiased estimator of some parameters associated with any unknown distribution function. The U-statistic associated with some parameter $\phi^{(r)}$ can be defined as

$$H_n^{(r)} = \binom{n}{r}^{-1} \sum_{(n,r)} h^{(r)}(X_{i_1}, ..., X_{i_r}), \tag{1.15}$$

where the summation is over all possible combinations of indices $(i_1, \ldots, i_r)$ such that $1 \leq i_1 < i_2 < \cdots < i_r \leq n$, $r < n$ and $h^{(r)}(\cdot)$ is a symmetric kernel of degree $r$ such that $\text{E}[h^{(r)}(X_{i_1}, ..., X_{i_r})] = \theta$. The degree, $r$, is the smallest number of random variables required to estimate the parameter, $\phi^{(r)}$, unbiasedly. Examples of U-statistics can be found in Chattopadhyay and Kelley (2016, 2017).

Sproule (1969, 1985) developed the framework for constructing a fixed-width confidence interval for the mean of a U-statistic for a given coverage probability. In his work, he only required the existence of the second moment of the kernel. With the use of U-statistics, it becomes easier to derive the asymptotic distribution and the asymptotic variance of the estimator using Hoeffding's (1948) decomposition. Sproule also proved that a fixed-width sequential procedure for a U-statistic is asymptotically efficient.

Next, proceeding along the lines of Lee (1990), for $c = 1, \ldots, r$, define $h_c(x_1, \ldots, x_c) = \mathrm{E}[h(X_1, \ldots, X_r)|(x_1, \ldots, x_c)] - \phi^{(r)}$. Next, define

$$\psi_1(x_1) = h_1(x_1); \quad \psi_2(x_1, x_2) = h_2(x_1, x_2) - \psi_1(x_1) - \psi_1(x_2)$$
$$\psi_r(x_1, \ldots, x_r) = h_r(x_1, \ldots, x_r) - \sum_{c=1}^{r} \psi_c(x_c) - \sum_{1 \le i_1 < i_2 \le r} \psi_2(x_{i_1}, x_{i_2}) - \cdots -$$
$$\sum_{1 \le i_1 < \ldots < i_{r-1} \le r} \psi_{r-1}(x_1, \ldots, x_{i_{r-1}}) \tag{1.16}$$

Then, the U-statistic, using Hoeffding's decomposition, can be defined as

$$H_n^{(r)} - \phi^{(r)} = \frac{r}{n} \sum_{c=1}^{r_1} \psi_1(X_c) + M_n, \tag{1.17}$$

where $M_n$ is the remainder term composed of $\psi_2, \ldots, \psi_{r-1}$, such that $M_n = O_p(n^{-1})$ if $\mathrm{E}\left[h^{(r)}(X_{i_1}, ..., X_{i_r})\right]^2 < \infty$. Using Lee (1990), the variance of $H_n^{(r)}$ is $\frac{r^2}{n}\psi^2 + O(n^{-2})$, where $\psi^2 = \mathrm{E}\left[\psi_1^2(X_1)\right]$. Thus, the asymptotic variance of $H_n^{(r)}$ is

$$\mathrm{Var}\left(H_n^{(r)}\right) = \frac{r^2}{n}\psi^2. \tag{1.18}$$

## 1.6 Chapter Organization

The remaining chapters of this dissertation are organised as follows: Chapter 2 develops a sequential approach to solving the bounded-width confidence interval problem for a general class of effect sizes. This work has already been published by Kelley, Darku, and Chattopadhyay (2018) in Psychological Methods - a journal by American Psychological Association. Chapter 3 is an extension of Chapter 2 to include correlation coefficients which is a special class of effect sizes. This chapter develops bounded-width confidence interval for the different measures of correlation coefficient, including multiple $R^2$, using sequential approaches. This work has been submitted for publication and is currently under review (Kelley et al., 2017). In Chapter 4, we develop the sequential methodology for constructing a sufficiently narrow confidence interval for Gini index under a survey whose sampling involves stratification and clustering, ie. a complex survey. In this same chapter, the sequential methodology under complex survey is applied to data from National Sample Survey (NSS) Organization in India. This work has also been submitted for publication (Bilson Darku et al., 2018). Chapter 5 presents a summary of the works in the manuscripts, gives directions for future works and provides concluding remarks.

Simulations and computations found in this dissertation were all done with codes and software packages from R software (R Core Team, 2017).

# CHAPTER 2

# A PURELY SEQUENTIAL APPROACH TO ACCURACY IN PARAMETER ESTIMATION FOR A GENERAL CLASS OF EFFECT SIZES[1]

## 2.1 Introduction

The concept of effect size as a primary outcome of interest has gained much traction over the last decade and is widely recognized as an important part of research studies. This concept is different from, though it can be complementary to, the binary outcome of a null hypothesis significance test that either rejects or fails-to-reject one or more null hypotheses. Effect size has been defined as “a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest” (Kelley and Preacher, 2012). Effect sizes such as the standardized mean difference, coefficient of determination, regression coefficient, path coefficient, correlation, among others, are widely used in many disciplines. The emphasis on effect sizes in modern research seems to have stemmed from methodologists heavily emphasizing their importance for many years (e.g., Morrison and Henkel, 1970; Meehl, 1997; Cohen, 1994; Thompson, 2002), professional organizations (American Psychological Association, 2010; Task Force on Reporting of Research Methods in AERA Publications, 2006; Association for Psychological Science, 2014) requiring them in scholarly work, journal editors pushing for more emphasis on effect size as a way to quantify practical meaning from a study, and journal reviewers, many of whom have themselves embraced the call for more effect sizes. It is clear that effect size now plays an important role in the research landscape of many disciplines.

The need to focus on effect sizes, the importance of confidence intervals for population effect sizes, and the limitations of null hypothesis significance tests based on a $p$-value that

[1] Copyright ©2017 by the American Psychological Association. Adapted with permission. Kelley, K., F. Bilson Darku, and B. Chattopadhyay (2017). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods.*

is less than or greater than a specified Type I error rate has recently been put at the front of methodological consideration. In particular, warnings and recommendations long made by methodologists within statistics, psychology, and others, about an over-reliance on null hypothesis significance tests and the corresponding $p$-value, has now been echoed by the American Statistical Association (ASA) in what is "the first time the ASA has spoken so publicly about a fundamental part of statistical theory and practice" (American Statistical Association, 2016). In an editorial by the ASA's Executive Director, on behalf of the ASA Board of Directors (Wasserstein, 2016), six principles are addressed that could "improve the conduct or interpretation of quantitative science" (2016, p. X). The ASA's conclusions comes 50 years after Bakan stated "the test of statistical significance in psychological research may be taken as an instance of a kind of essential mindlessness in the conduct of research" (1966, p. 436) but that he also acknowledged that his ideas were not original but what "everybody knows" (p. 423).

The ASA editorial goes on to say that "in view of the prevalent misuses of and misconceptions concerning $p$-values, some statisticians prefer to supplement or even replace $p$-values with other approaches". The suggestions for supplementing or replacing $p$-values are "methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address *the size of an effect* (*and its associated uncertainty*) or whether the hypothesis is correct" (emphasis added). This work addresses the size of the effect and its uncertainty explicitly and, importantly, without the prior specification of likely unknown population values as is typical in research design texts. This work is believed to be both timely and important for helping to advance psychology and related disciplines by focusing explicitly on estimation of effect sizes of interest.

We begin with the premise that point estimates almost certainly differ from their population analogs. Correspondingly, as many others have stated, it is important for effect sizes to be accompanied by confidence intervals in order to convey the uncertainty of the estimate at some specified level of confidence. Depending on sample size, among other factors, the confidence interval width (i.e., the length of the confidence interval), even with samples from the same population, will vary. Holding everything else constant, narrower confidence intervals at a specific confidence level (e.g., 95%) provide more precise information about the parameter of interest than a wider confidence interval does at the same confidence level. In an effort to construct sufficiently narrow confidence intervals, the accuracy in parameter estimation (AIPE) approach to sample size planning (e.g. Kelley and Maxwell, 2003; Kelley and Lai, 2011; Kelley and Rausch, 2006; Kelley, 2008, 2007b; Terry and Kelley, 2012; Lai and Kelley, 2011b; Pornprasertmanit and Schneider, 2014), also known as "the confidence interval approach", has been developed for a variety of important effect sizes. Such approaches are similar to the "fixed-width confidence interval problem," where instead of having an upper bound on the length of the confidence interval, the length of the confidence interval is exactly the desired width (e.g., Sproule, 1985; Mukhopadhyay and De Silva, 2009; Mukhopadhyay and Chattopadhyay, 2012).

The AIPE approach to sample size planning, as it has been developed thus far in the literature, is a fixed-sample size approach based on supposed values of one or more parameters in an effort to obtain a sufficiently narrow confidence interval at the specified level (e.g., 95%, 99%). However, a potential problem is that if the supposed value of the population parameter(s) is/are incorrect, then the (fixed) sample size from the AIPE perspective may be very different than what the (fixed) sample size would have been if the true population parameter(s) was (were) used. This problem also happens in power analysis when the sample size is based on the supposed population value (e.g., Cohen, 1988; cf. basing sample size on the minimum value of the parameter that would be of theoretical interest, Lipsey, 1990;

Murphy and Myors, 2004). A remedy to needing the generally unknown population values in traditional applications of the AIPE approach is a sequential analysis approach. Here, the population parameters are not pre-specified, and as a result, the sample size cannot be fixed in advance. That is, the procedure is not of the "fixed-$n$" research design framework. Rather, the sample size deemed appropriate in sequential estimation procedure depends on collecting observations until an a priori specified criterion or *stopping rule* is satisfied.

Sequential methods have been developed in various areas of statistics beginning 75 years ago (e.g., Wald, 1945, 1943). In sequential medical trials, Armitage (1960, pp. 9–10) advocated the use of estimates of difference in effects of two treatments with some desired standard error rather than basing a decision on null hypothesis significance tests. In the context of clinical trials, Lai (1998) discussed a sequential procedure for constructing fixed-width confidence intervals for some population characteristics of interest. Recently, for allocation of two treatments in clinical trials, Bandyopadhyay and Biswas (2015) developed fixed-width confidence intervals for response-adaptive allocation design. However, sequential methods for inference have not had much impact in psychology and related disciplines as of yet, with the notable exception of item response theory (e.g., in the context of computer adaptive tests; Chang and Ying, 2009 and the references therein). The focus of this work is in the research design context when inferences are made about population parameters, such as mean differences, correlations, a variety of standardized effect sizes, et cetera, where it is proposing an alternative to fixed-$n$ procedures such as how power analysis and accuracy in parameter estimation are commonly employed.

In this work, a sequential procedure is used to estimate a general effect size in order for the confidence interval for the population effect size of interest to be sufficiently narrow. This idea, which will be called *sequential AIPE*, is a generalization of much of the literature that currently exists in this framework that requires the specification of population values. The work then discusses the AIPE problem for this general class of effect size, special cases of which are many commonly used effect sizes, and propose a sequential estimation procedure. This is followed with the desired result of the sequential optimization procedure for constructing the sufficiently narrow confidence interval with a pre-specified level of confidence. Additionally, and importantly, these developments are made in a distribution-free environment. The distribution-free environment is important, though challenging, because there is often no reason to assume that the underlying distribution of the data for which an effect size will be calculated would be known (e.g., gamma, lognormal, normal). For example, in applied research, normal distributions may be rare (e.g., Micceri, 1989), and thus basing important decisions on assumptions that are not realized may be problematic. Such issues are avoided and focus is placed on the most robust context of a distribution-free environment (e.g., Wilcox, 2012). Thus, these developments offer a great deal of generality and flexibility. In the following section, we introduce a class of effect sizes that will be considered in this work and the framework for their estimation.

## 2.2 Effect Sizes Based on Ratio of Two Parameters: A General Framework

In this work, we consider a general family of effect sizes, which we define as the ratio of two parameters, with each parameter being a function of one or more other parameters. In particular, the sequential procedure developed is not for any one particular effect size, but rather for a general effect size which has many special cases.

Consider the general effect size parameter $\theta$, which can be expressed as a ratio of functions of two parameters, $\theta_1$ and $\theta_2$, such that $\theta$ is defined as

$$\theta = \frac{g_1\left(\theta_1\right)}{g_2\left(\theta_2\right)}, \tag{2.1}$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are two continuous functions, $\theta_1$ and $\theta_2$ are parameters each involving linear combinations of parameters corresponding to $k$ groups or $k$ different parameters from the same group, provided $g_2\left(\theta_2\right) \neq 0$. For $i = 1, \ldots, k$, suppose that $\theta_{1i}$ and $\theta_{2i}$ are the parameters for the $i^{\text{th}}$ group or is the $i^{\text{th}}$ parameter from the same group, such that

$$\theta_1 = \sum_{i=1}^{k} l_{1i}\theta_{1i} \tag{2.2}$$

and

$$\theta_2 = \sum_{i=1}^{k} l_{2i}\theta_{2i}, \tag{2.3}$$

where $l_{1i}$s and $l_{2i}$s are known constants. Suppose the goal is to estimate the effect size parameter $\theta$, the population ratio of functions of $\theta_1$ and $\theta_2$, on the basis of $n$ observations from each of the $k$ groups. Let the observations from the $i^{\text{th}}$ group be $X_{i1}, \ldots, X_{in}$. Further, let $T_{1n}$ and $T_{2n}$ be the two estimators for $\theta_1$ and $\theta_2$ respectively, where $T_{1n} = \sum_{i=1}^{k} l_{1i}U_{in}$ is a linear combination of $k$ independent U-statistics and $T_{2n} = \sum_{i=1}^{k} l_{2i}V_{in}$ is another linear combination of $k$ independent U-statistics. Now assume that for $i = 1, 2, ..., k$ the U-statistic $U_{in}$ is an unbiased and consistent estimator of $\theta_{1i}$ and the U-statistic $V_{in}$ is an unbiased and consistent estimator of $\theta_{2i}$. The U-statistics, that is, unbiased estimators of the parameters of interest, are discussed in detail in Section 1.5. The estimator of the effect size parameter $\theta$, based on estimators of $\theta_1$ and $\theta_2$, is given by,

$$T_n = \frac{g_1(T_{1n})}{g_2(T_{2n})}. \tag{2.4}$$

This effect size, $\theta$ in the population and $T_n$ in a sample of size $n$, is very general. Using several examples it will be shown that some widely used effect size parameters and their corresponding estimators are special cases of the forms given in Equation (2.1) and Equation (2.4), respectively.

### 2.2.1 Example 1: Standardized Mean Difference

Consider the standardized mean difference, which is a standardized measure of separation between two group means. The population standardized mean difference is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \tag{2.5}$$

where $\theta_1 = \mu_1 - \mu_2$ and $\theta_2 = \sigma^2$, with $\mu_1$ and $\mu_2$ being the population means from groups 1 and 2, respectively, and $\sigma$ being the population standard deviation of scores within the two groups under the homogeneity of variance assumption ($\sigma_1^2 = \sigma_2^2 = \sigma^2$). In practice, $\delta$ itself is unknown as the population values of the means for groups 1 and 2, $\mu_1$ and $\mu_2$ respectively, and common standard deviation, $\sigma$, are unknown. Let $\bar{X}_{1n}$ and $\bar{X}_{2n}$ denote the sample mean of scores on an outcome of interest from groups 1 and 2 respectively. Let $s_{1n}^2$ and $s_{2n}^2$ represent the usual unbiased estimators of population variances from groups 1 and 2, respectively. $\bar{X}_{1n}$ and $\bar{X}_{2n}$ are each U-statistics of degree 1 and $s_{pn}^2$ is a function of two U-statistics, $s_{1n}^2$ and $s_{2n}^2$, both of degree 2 such that

$$s_{pn} = \sqrt{\frac{s_{1n}^2 + s_{2n}^2}{2}}, \tag{2.6}$$

is the square root of the pooled sample variance here. (Note that the sample sizes for both groups are the same.) In this case, from Equation (2.4), $T_{1n} = \bar{X}_{1n} - \bar{X}_{2n}$ is the difference of means from two groups and $T_{2n} = s_{pn}^2$ is the pooled sample variance. Thus, $U_{in} = \bar{X}_{in}$ and $V_{in} = s_{in}^2$. The known coefficients are $l_{11} = 1, l_{12} = -1, l_{21} = 1/2$ and $l_{22} = 1/2$ and $k = 2$. The numerator and denominator are, the difference in means and square root of the pooled variance, respectively. More formally, $g_1(T_{1n}) = \bar{X}_{1n} - \bar{X}_{2n}$ and $g_2(T_{2n}) = s_p$. Thus the effect size estimator for the population standardized mean difference is

$$d_n = \frac{\bar{X}_{1n} - \bar{X}_{2n}}{s_{pn}}. \tag{2.7}$$

Consider now a variant of Equation (2.5) in which the control group standard deviation is used as the divisor. Let subscript 1 be treatment group ($T$) and subscript 2 denote the control group ($C$). Then, the parameter of interest would be Glass' $\Delta$ which is given by

$$\Delta = \frac{\mu_T - \mu_C}{\sigma_C} \tag{2.8}$$

(Glass and Smith, 1979). For $\Delta$ the homogeneity of variance need not be assumed, as only one standard deviation is used. Here, the known coefficients are $l_{11} = l_{1T} = 1, l_{12} = l_{1C} =$

$-1, l_{21} = l_{2T} = 0$, $l_{22} = l_{2C} = 1$ and $k = 2$. Thus effect size estimator for the population standardized mean difference is

$$d_{Cn} = \frac{(\bar{X}_{Tn} - \bar{X}_{Cn})}{s_{Cn}}. \tag{2.9}$$

Here, the functions are $g_1(T_{1n}) = T_{1n}$ and $g_2(T_{2n}) = \sqrt{T_{2n}}$, where $T_{1n} = \bar{X}_{Tn} - \bar{X}_{Cn}$ and $T_{2n} = s^2_{Cn}$. The notation $d_{Cn}$ is used to show that groups can be used in the numerator but not the denominator, or vice versa.

### 2.2.2 Example 2: Coefficient of Variation

Consider the coefficient of variation, where the population value is defined as

$$\kappa = \frac{\sigma}{\mu}. \tag{2.10}$$

From Equation (2.1), $\theta_1 = \sigma^2$ and $\theta_2 = \mu$, where $\mu$ is the population mean and $\sigma$ is the population standard deviation (with $k = 1$). For estimating the unknown population coefficient of variation, $\kappa$, the corresponding estimator is

$$k_n = \frac{s_n}{\bar{X}_n}. \tag{2.11}$$

From Equation (2.4), $T_{1n} = s^2_n$, which is the sample variance and $T_{2n} = \bar{X}_n$ is the sample mean of $n$ observations. Because $k = 1$, $U_{1n} = s^2_n$ and $V_{1n} = \bar{X}_n$. The known coefficients are $l_{11} = 1$ and $l_{21} = 1$. The function $g_1(T_{1n}) = s_n$ and $g_2(T_{2n}) = \bar{X}_n$. Thus we see that $k_n$ is a ratio of two functions of two U-statistics: $s^2_n$ (a U-statistic of degree 2; of which, with $g_1(\cdot)$ we take the square root) and $\bar{X}_n$ (a U-statistic of degree 1).

### 2.2.3 Example 3: The Standardized Mean

The standardized mean, which is the reciprocal of the coefficient of variation, that is, $\mu/\sigma$ is also an effect size of interest in some situations (e.g., Cohen, 1988; Kelley, 2007a). From

Equation (2.1), $\theta_1 = \mu$ and $\theta_2 = \sigma^2$. For estimating $\mu/\sigma$, the estimator is $\bar{X}_n/s_n$. According to Equation (2.4), $T_{1n} = \bar{X}_n$ is the sample mean and $T_{2n} = s_n^2$ is the sample variance score from a sample of $n$ observations. Here, $k = 1$, so $U_{1n} = \bar{X}_n$ and $V_{1n} = s_n^2$. The known coefficients are $l_{11} = 1$ and $l_{21} = 1$. The functions $g_1(T_{1n}) = \bar{X}_n$ and $g_2(T_{1n}) = s_n$.

### 2.2.4 Example 4: Regression Coefficient in Simple Linear Model

Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are pairs of observations from a simple linear regression model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{2.12}$$

where $Y_i$ is the dependent variable, $X_i$ is the independent variable, $\varepsilon_i$'s are the independent and identically distributed errors, and $\beta_0$ is the population intercept parameter and $\beta_1$ is the population slope parameter. Now consider the effect of $X$ on $Y$, which is the population slope defined as

$$\beta_1 = \frac{\sigma_{XY}}{\sigma_X^2}, \tag{2.13}$$

where $\sigma_{XY}$ is the population covariance between $X$ and $Y$, and $\sigma_X^2$ is the population variance of $X$. Because the value of $\beta_1$ is unknown in practice, it must be estimated, which is generally done using least squares criterion, by

$$b_{1n} = \frac{\sum(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum(x_i - \bar{x}_n)^2} = \frac{s_{XYn}}{s_{Xn}^2}, \tag{2.14}$$

where $s_{XYn}$ is the unbiased estimator for covariance between $X$ and $Y$, and $s_{Xn}^2$ is the unbiased estimator for the variance of $X$, based on a sample of size $n$. These estimators are both U-statistics of degree 2. From Equation (2.4), $g_1(T_{1n}) = s_{XYn}$ and $g_2(T_{2n}) = s_X^2$ and $l_{11} = l_{21} = 1$. Hence, $U_{1n} = s_{XY,n}$ and $V_{1n} = s_X^2$. Thus, the estimator for the regression parameter $\beta_1$ is a ratio of two functions of U-statistics with degree 2.

### 2.2.5 Example 5: Effect Size for Ordinal Data

In the case of ordinal data, Cliff's delta can be used, which we illustrate here (see, for example, Cliff, 1993). Cliff's delta is a measure of how often randomly sampled values in one distribution are larger than the randomly sampled values in a second distribution. Suppose there are two sets of ordinal data of sizes $n_1$ and $n_2$, potentially from two groups or distributions. Then, the sample estimator of Cliff's delta for the two groups or distributions is given by

$$\frac{\#(x_i > y_j) - \#(x_i < y_j)}{n_1 n_2} = \frac{2U}{n_1 n_2} - 1, \tag{2.15}$$

where # is defined as the number of times and $U$ is the Mann-Whitney U-statistic (the test statistic used in nonparametric two-sample location test). For details on Mann-Whitney U-statistic, readers are referred to Kumar and Chattopadhyay (2013).

### 2.2.6 Example 6: Contrasts

Let us now consider contrasts, which are often used in analysis of variance. For the $i^{\text{th}}$ group, suppose $X_{i1}, \ldots, X_{in}$ are independent and identically distributed random variables with means $\mu_i$ and variances $\sigma_i^2$, $i = 1, \ldots, K$. Thus, in total, there are $Kn$ observations from $K$ groups. Then, the population contrast related to the corresponding scenario is given by

$$\psi = \sum_{i=1}^{K} c_i \mu_i, \tag{2.16}$$

where $c_1, \ldots, c_K$ are known coefficients and $\sum_{i=1}^{K} c_i$=0.

An estimator of the contrast $\psi$ is $\widehat{\psi}_n = \sum_{i=1}^{K} c_i \bar{X}_{in}$, where $\bar{X}_{1n}, \ldots, \bar{X}_{Kn}$ are the group means. In this case, from Equation (2.4), we use, $T_{1n} = \sum_{i=1}^{K} c_i \bar{X}_{in}$, so $U_{in} = \bar{X}_{in}$. The known coefficients are $l_{1i} = c_i$ for $i = 1, 2..., k$ and $k = K$. Here, $g_1(T_{1n}) = T_{1n}$ and $g_2(\cdot)$=1.

### 2.2.7 Example 7: Univariate Parameters and Their Functions

The parameters such as population mean, difference of population means, population variance, population Gini's mean difference can be shown to satisfy Equation (2.1) with $g_1(\theta_1)$ as the parameter of interest and $g_2(\theta_2) = 1$. In fact, the sum or difference of the above parameters themselves satisfy Equation (2.1) (i.e., the difference in means, the difference in variances, etc.).

In all of the above mentioned examples, the effect sizes satisfy Equation (2.1) and the corresponding estimators satisfy Equation (2.4). Correspondingly, $\theta$ is described as the *general effect size.* Note that the subscript $n$ is used on the effect size estimator to denote the sample size on which it is based. This is very important as the properties of the effect size estimator $T_n$ based on different sample sizes are considered. At this point, a general effect size has been developed and illustrated with several examples. In what follows, we develop the central limit theorem for the general effect size.

## 2.3 Central Limit Theorem for $T_n$

Our procedure depends on the Central Limit Theorem for the effect size parameter $\theta$, defined in Equation (2.1), due to the distribution-free scenario we have used. As noted earlier, $T_{1n}$ and $T_{2n}$ are linear combinations of U-statistics. We note that, $T_{1n} = \left(\sum_{i=1}^{k} l_{1i} U_{in}\right)$ and $T_{2n} = \left(\sum_{i=1}^{k} l_{2i} V_{in}\right)$, where $U_{in}$ values are U-statistics with kernel $h_{1i}$ of degree $r_{1i}$ for estimating the $\theta_{1i}$ and $V_{in}$ values are U-statistics with kernel $h_{2i}$ of degree $r_{2i}$ for estimating $\theta_{2i}$.

**Theorem 2.1.** *Suppose the parent distribution(s) is(are) such that $E[U_{in}^2] < \infty$ and $E[V_{in}^2]) < \infty$ for $i = 1, \ldots, k$. Then, the central limit theorem corresponding to $T_n$ is*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} N(0, \xi^2), \tag{2.17}$$

*where* $\xi^2$ *is the asymptotic variance given by* $\xi^2 = \mathbf{D}'\mathbf{\Sigma}\mathbf{D}$, *and* $\mathbf{D}' = \left[\frac{g_1'(\theta_1)}{g_2(\theta_2)}, \frac{-g_1(\theta_1)g_2'(\theta_2)}{g_2^2(\theta_2)}\right]$ *is a vector and,*

$$\mathbf{\Sigma} = \begin{bmatrix} \xi_1^2 & \xi_{12} \\ \xi_{12} & \xi_2^2 \end{bmatrix}.$$

*Here,* $\xi_1^2$ *and* $\xi_2^2$ *are, respectively, the asymptotic variances of* $\sqrt{n}(T_{1n} - \theta_1)$ *and* $\sqrt{n}(T_{2n} - \theta_2)$ *and the asymptotic covariance of* $\sqrt{n}(T_{1n} - \theta_1)$ *and* $\sqrt{n}(T_{2n} - \theta_2)$ *is* $\xi_{12}$.

Before proving the main theorem, the following lemma will be stated and proved.

**Lemma 2.1.** *Asymptotic variances of* $\sqrt{n}(T_{1n} - \theta_1)$ *and* $\sqrt{n}(T_{2n} - \theta_2)$ *are* $\xi_1^2$ *and* $\xi_2^2$ *and the asymptotic covariance of* $\sqrt{n}(T_{1n} - \theta_1)$ *and* $\sqrt{n}(T_{2n} - \theta_2)$ *is* $\xi_{12}$.

*Proof.* Using Hoeffding's (1948) decomposition as in Equation (1.17), $U_{in}$ and $V_{in}$ can be written as

$$\begin{aligned} U_{in} &= \frac{r_{1i}}{n} \sum_{c=1}^{r_{1i}} \psi_{1i}(X_{ic}) + O_p(n^{-1}) \\ V_{in} &= \frac{r_{2i}}{n} \sum_{c=1}^{r_{2i}} \psi_{2i}(X_{ic}) + O_p(n^{-1}), \end{aligned} \tag{2.18}$$

where $\psi_{1i}(X_{ic})$ and $\psi_{2i}(X_{ic})$ are, respectively,

$$\begin{aligned} \psi_{1i}(X_{ic}) &= \mathrm{E}_F[h_{1i}(X_{i1}, \ldots, X_{ir_{1i}}) | X_{ic} = x_{ic}] - \theta_{1i} \\ \psi_{2i}(X_{ic}) &= \mathrm{E}_F[h_{2i}(X_{i1}, \ldots, X_{ir_{2i}}) | X_{ic} = x_{ic}] - \theta_{2i}. \end{aligned} \tag{2.19}$$

Suppose that $\xi_{1i}^2 = \mathrm{E}\left[\psi_{11}^2(X_{ic})\right]$ and $\xi_{2i}^2 = \mathrm{E}\left[\psi_{21}^2(X_{ic})\right]$ and $\xi_{1i2i} = \mathrm{cov}\left[\psi_{1i}^2(X_{ic}), \psi_{2i}^2(X_{ic})\right]$. Using Equation (1.18), the asymptotic variance of $U_{in}$ is $r_{1i}^2 \xi_{1i}^2 / n$ for $i = 1, 2, ..., k$ and the asymptotic variance of $V_{in}$ is $r_{2i}^2 \xi_{2i}^2 / n$ for $i = 1, 2, ..., k$. Then, we have

$$\mathrm{E}\left[\sum_{i=1}^{k} l_{1i} U_{in}\right] = \sum_{i=1}^{k} l_{1i}\theta_{1i}, \tag{2.20}$$

$$\mathrm{E}\left[\sum_{i=1}^{k} l_{2j} V_{in}\right] = \sum_{j=1}^{k} l_{2i}\theta_{2i}, \tag{2.21}$$

and

$$\mathrm{Var}\left[\sum_{i=1}^{k} l_{1i}U_{in}\right] = \sum_{i=1}^{k} l_{1i}^2 r_{1i}^2 \xi_{1i}^2/n = \xi_1^2/n, \tag{2.22}$$

$$\mathrm{Var}\left[\sum_{i=1}^{k} l_{2i}V_{in}\right] = \sum_{i=1}^{k} l_{2i}^2 r_{2i}^2 \xi_{2i}^2/n = \xi_2^2/n. \tag{2.23}$$

The asymptotic covariance of $T_{1n}$ and $T_{2n}$ is,

$$\begin{aligned}\mathrm{cov}\left[\sum_{i=1}^{k} l_{1i}U_{in}, \sum_{j=1}^{k_2} l_{2j}V_{jn}\right] &= \sum_{i=1}^{k}\sum_{i=1}^{k} l_{1i}l_{2i}\mathrm{cov}\left[U_{in}, V_{in}\right] \\ &= \sum_{i=1}^{k}\sum_{i=1}^{k} l_{1i}l_{2i}r_{1i}r_{2i}\xi_{1i2i}/n = \xi_{12}/n \end{aligned} \tag{2.24}$$

■

Now, let us prove Theorem 2.1.

*Proof.* Using Lee (1990), $\mathbf{Y}_n = [\sqrt{n}(T_{1n} - \theta_1), \sqrt{n}(T_{2n} - \theta_2)]' \xrightarrow{\mathcal{L}} N_2(\mathbf{0}, \mathbf{\Sigma})$, where,

$$\mathbf{\Sigma} = \begin{bmatrix} \xi_1^2 & \xi_{12} \\ \xi_{12} & \xi_2^2 \end{bmatrix}.$$

Now, define the ratio $R(u, v) = \frac{g_1(u)}{g_2(v)}$, if $g_2(v) \neq 0$. Using Taylor's expansion, we can write

$$\sqrt{n}(T_n - \theta) = \sqrt{n}(R(T_{1n}, T_{2n}) - R(\theta_1, \theta_2)) = \mathbf{D}'\mathbf{Y}_n + \epsilon_n||\mathbf{Y}_n||_2, \tag{2.25}$$

where $\mathbf{D}' = \left[\frac{g_1'(\theta_1)}{g_2(\theta_2)}, \frac{-g_1(\theta_1)g_2'(\theta_2)}{g_2^2(\theta_2)}\right]$, and $\epsilon_n \to 0$ if $||(T_{1n}, T_{2n})' - (\theta_1, \theta_2)'||_2 \to 0$. Hence, $\epsilon_n||\mathbf{Y}_n||_2 \xrightarrow{P} 0$ as $n \to \infty$. Thus, the central limit theorem for the effect size of the type defined in Equation (2.4) shows that as $n \to \infty$,

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{L}} N(0, \xi^2), \tag{2.26}$$

where $\xi^2$ is the asymptotic variance given by $\xi^2 = \mathbf{D}'\mathbf{\Sigma}\mathbf{D}$ and $\xrightarrow{\mathcal{L}}$ indicates convergence in distribution. ■

Using the central limit theorem developed, the application of accuracy in parameter estimation approach for the general effect size parameter under a fixed-sample scenario will be discussed.

## 2.4 Accuracy in Parameter Estimation of General Effect Size Parameter: A Fixed-Sample Size Approach

The goal of this work is to obtain a sufficiently narrow $100(1-\alpha)\%$ confidence interval for the effect size parameter $\theta$ under the distribution-free scenario. As this work is being done in a distribution-free scenario, no assumption is made about the distribution of the scores from which the sufficiently narrow confidence interval for $\theta$ will be calculated. In practice, the distribution of the scores is generally unknown. In other words, because of the untenability of knowing the distribution from which scores are sampled, we do not assume any specific distribution of the scores. Correspondingly, what follows are developments in a distribution-free scenario. Thus, the exact distribution of $T_n$ cannot be obtained. To be clear, this is not a limitation of the method per se, but rather with the distribution-free scenario more generally. Sproule (1985) developed a method to construct a fixed-width confidence interval under distribution-free scenario using large sample theory, but that method cannot be applied directly in this problem as the general effect size may involve the ratio of functions of one or more parameters. In this work, large sample theory will be used to find the asymptotic distribution of $T_n$. With this, a sufficiently narrow $100(1-\alpha)\%$ confidence interval for $\theta$ will be constructed, and for practical purposes, will be shown to yield intervals that tend to work well.

Using Theorem 2.1, the approximate $100(1-\alpha)\%$ confidence interval for $\theta$ is given by

$$J_n = \left(T_n - z_{\alpha/2}\frac{\xi}{\sqrt{n}}, T_n + z_{\alpha/2}\frac{\xi}{\sqrt{n}}\right), \tag{2.27}$$

where $\xi^2/n$ is the asymptotic variance of $T_n$. The width of the confidence interval $J_n$ is given by

$$w_n = 2z_{\alpha/2}\frac{\xi}{\sqrt{n}}. \tag{2.28}$$

In AIPE problems, the sample size required to achieve sufficient accuracy is solved so that the width of the confidence interval is no larger than $\omega$. Thus, for a given $\omega$, we have

$$2z_{\alpha/2}\frac{\xi}{\sqrt{n}} \leq \omega, \tag{2.29}$$

which implies that the necessary sample size to construct $100(1-\alpha)\%$ confidence interval for $\theta$ will be

$$n \geq \left\lceil \frac{4z_{\alpha/2}^2\xi^2}{\omega^2} \right\rceil \equiv n_\omega, \tag{2.30}$$

where $\lceil x \rceil$ is the ceiling function which is the least integer greater than or equal to $x$ (e.g., $\lceil 95.2 \rceil = 96$). The expression in the Equation (2.30) can be found by solving for $n$ in Equation (2.29). Thus, $n_\omega$ is the theoretically optimal sample size required to make the $100(1-\alpha)\%$ confidence interval for $\theta$ provided $\xi^2$ (recall that $\xi^2/n$ is the asymptotic variance of $T_n$) is known. Because in reality $\xi^2$ is generally unknown, the optimal sample size $n_\omega$ is also unknown as $n_\omega$ depends on $\xi^2$. In order to estimate the optimal sample size $n_\omega$, a consistent estimator, $\widehat{\xi}_n^2$ is used for estimating $\xi^2$. We note that any value of $\widehat{\xi}_n^2$ does not guarantee that the condition in Equation (2.30) is satisfied and thereby estimates the optimal sample size $n_\omega$. Also often, in several sample size planning methods, a researcher will use a supposed value (say $\widetilde{\xi}^2$) of the population parameter, $\xi^2$, to compute $n_\omega$. However, $\widetilde{\xi}^2$ may differ considerably from $\xi^2$, which can have a large impact on the appropriate sample size. Further, and more troubling, even if $\widetilde{\xi}^2$ differs from $\xi^2$ by a relatively small degree, there can (still) be a large impact on the appropriate sample size. Thus, using unknown population values to estimate $\xi^2$ can lead to potentially poor choices for the appropriate sample sizes. So, a sequential procedure, which does not require supposed population parameter value, will be used to find out the sample size but will satisfy the condition given in Equation (2.30). This approach w

## 2.5 Accuracy in Parameter Estimation via a Sequential Optimization Procedure

As opposed to fixed-sample procedures, in sequential procedures, the sample size is not fixed in advance. As previously discussed, no fixed sample-size procedure can provide a solution to the accuracy in parameter estimation problems without making assumptions about the distribution of the data. Here we propose a purely sequential procedure to construct a $100(1-\alpha)\%$ confidence interval for the general effect size parameter $\theta$. Recalling that the effect size $\theta$ subsumes many special cases, and that we work within a distribution-free environment, our work is thus a general and novel treatment and one that subsumes many potential special cases that could have independently been developed.

In a sequential procedure, the estimation of parameter(s) occurs in stages until a stopping rule is met. In the first stage, a small sample called a pilot sample is observed and then the parameters are estimated to check a pre-defined condition in a *stopping rule*. A stopping rule is a rule that indicates, after every stage, whether further sampling of one (or more) observation(s) is necessary or should be stopped. Thus, further sampling of observations is carried out if the pre-defined condition in the stopping rule is not met and further sampling is stopped once the pre-defined condition in the stopping rule is satisfied. At a particular stage, if the pre-defined condition is not met, the researcher collects one (or more) observation(s) and then estimates the parameter of interest based on the collected observation(s). This process is repeated until the pre-defined condition is met. For details about the general theory of sequential estimation procedures, interested readers are referred to Sen (1981), Ghosh and Sen (1991), Mukhopadhyay and Chattopadhyay (2012), or Chattopadhyay and Mukhopadhyay (2013).

Recall that the optimal sample size $n_\omega$ is unknown due to $\xi^2$ being unknown. We use the consistent estimator of $\xi^2$, namely $\widehat{\xi}_n^2$, which is based on $n$ observations drawn from the

$k$ groups. We now develop an algorithm to find an estimate of the optimal sample size via the purely sequential estimation procedure.

**Stage 1:** Scores of $m$ randomly selected individuals are collected from each of the $k$ groups. Following Mukhopadhyay (1980) we recommend using the pilot sample size $m$ given as

$$m = \max\left\{m_0, \left\lceil \frac{2z_{\alpha/2}}{\omega} \right\rceil \right\}, \tag{2.31}$$

where $m_0(> 0)$ is the least possible sample size required to estimate $\xi^2$ and $\lceil \cdot \rceil$ is the ceiling function of the term—the ceiling being the smallest integer not less than $(2z_{\alpha/2}/\omega)$. Based on this pilot sample of size $m$, an estimate of $\xi^2$ is obtained by computing $\widehat{\xi}^2_m$. If $m < \left\lceil \frac{4z^2_{\alpha/2}}{\omega^2}\left(\widehat{\xi}^2_m + \frac{1}{m}\right)\right\rceil$, then proceed to the next step. Otherwise, if $m \geq \left\lceil \frac{4z^2_{\alpha/2}}{\omega^2}\left(\widehat{\xi}^2_m + \frac{1}{m}\right)\right\rceil$, stop sampling and set the final sample size equal to $m$ from each group.

**Stage 2:** Obtain an additional $m'(\geq 1)$ observations. At this stage there are $(m + m')$ observations from each of the $k$ groups. Update the estimate of $\xi^2$ by computing $\widehat{\xi}^2_{m+m'}$. Now check whether $m+m' \geq \left\lceil \frac{4z^2_{\alpha/2}}{\omega^2}\left(\widehat{\xi}^2_{m+m'} + \frac{1}{m+m'}\right)\right\rceil$. If $m+m' < \left\lceil \frac{4z^2_{\alpha/2}}{\omega^2}\left(\widehat{\xi}^2_{m+m'} + \frac{1}{m+m'}\right)\right\rceil$ then go to the next step. Otherwise, if $m + m' \geq \left\lceil \frac{4z^2_{\alpha/2}}{\omega^2}\left(\widehat{\xi}^2_{m+m'} + \frac{1}{m+m'}\right)\right\rceil$ then stop further sampling and report that the final sample size is $(m + m')$ from each group.

This process of collecting one (or more) observation(s) in each stage after stage 1 continues until there are $N$ observations such that $N \geq \left\lceil \frac{4z^2_{\alpha/2}}{\omega^2}\left(\widehat{\xi}^2_N + \frac{1}{N}\right)\right\rceil$. At this stage, we stop further sampling because the stopping rule has been satisfied and report that the final sample size is $N$ for single group designs or $N$ within each group for multiple group designs (and thus the total sample size is $KN$ in multiple group designs, as we have assumed equal sample size per group).

Based on the algorithm just outlined, a stopping rule for the sampling can be defined as follows:

$$N \text{ is the smallest integer } n(\geq m) \text{ such that } n \geq \frac{4z^2_{\alpha/2}}{\omega^2}\left(\widehat{\xi}^2_n + \frac{1}{n}\right), \tag{2.32}$$

where the term $n^{-1}$ is a correction term which ensures that the sampling process does not stop too early for the optimal sample size because of the use of the approximate expression. After each and every stage, the stopping rule indicates whether the collected sample size is more than the estimated optimal sample size, then additional $m'$ observations are collected in the next stage. At some stage, when the collected sample size becomes equal to or more than the estimated optimal sample size, sampling is terminated. Thus $N$ in Equation (2.32) is regarded as the estimator of the theoretically optimal sample size $n_\omega$ required to make $100(1-\alpha)\%$ confidence interval for $\theta$ provided $\xi^2$ is known.

For details about the correction term, refer to Chattopadhyay and De (2016), Sen and Ghosh (1981), Chattopadhyay and Kelley (2016, 2017). Note that for large sample sizes, $\widehat{\xi}_n^2 + n^{-1}$ converges to $\xi^2$.

### 2.5.1 Characteristics of Our Sequential Procedure

Based on the algorithm just outlined, it is important to ensure that the sampling of infinite number of observations is not possible. If observations are collected using Equation (2.32), sampling will stop at some stage with probability one. This is proved in Lemma 2.2, which says that under appropriate conditions, $P(N < \infty) = 1$. This result is very important as it mathematically ensures that sampling will be terminated eventually.

Note that, from Equation (2.32), $N$ is a random variable because $N$ depends on the estimator of $\xi^2$, which itself is a random variable. Theorem 2.2 implies that the $100(1-\alpha)\%$ confidence interval for $\theta$

$$\left[T_N - \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}, T_N + \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}\right] \tag{2.33}$$

formed using $N$ observations achieves the specified coverage probability $1-\alpha$ asymptotically. This property is called *asymptotic consistency.* Thus, in Theorem 2.2 we have proven that our purely sequential procedure has the asymptotic consistency property. Additionally, Theorem 2.2 proves that the confidence interval for $\theta$ given in Equation (2.33) always achieves a sufficiently narrow width (less than $\omega$).

We now state the lemmas and another theorem of this work along with their proofs.

## 2.6 Random Central Limit Theorem for $T_n$

Before the main theorem is stated, the following two lemmas are needed.

**Lemma 2.2.** *Under the assumption that $E[\widehat{\xi}_n^2]$ exist, for any $\omega > 0$, the stopping time $N$ is finite, that is, $P(N < \infty) = 1$.*

*Proof.* Note that $\widehat{\xi}_n^2$ is a consistent estimator of $\xi^2$. Therefore, for any fixed $\omega > 0$,

$$P(N > \infty) = \lim_{n\to\infty} P(N > n) \leq \lim_{n\to\infty} P\left\{ n < \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \left(\widehat{\xi}_n^2 + \frac{1}{n}\right)\right\} = 0. \tag{2.34}$$

The last equality is obtained since $\widehat{\xi}_n^2 \to \xi$ almost surely as $n \to \infty$. Thus, $P(N < \infty) = 1$. ■

**Lemma 2.3.** *If the parent distribution(s) is(are) such that $E[\widehat{\xi}_n^2]$ exists, then the stopping rule in Equation* (2.32) *yields*

$$\frac{N}{n_\omega} \xrightarrow{P} 1 \text{ as } \omega \downarrow 0 \tag{2.35}$$

*where, $\xrightarrow{P}$ indicates convergence in probability.*

*Proof.* The definition of stopping rule $N$ in Equation (2.32) yields

$$\left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \widehat{\xi}_N^2 \leq N \leq mI(N = m) + \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \left(\widehat{\xi}_{N-1}^2 + (N-1)^{-1}\right). \tag{2.36}$$

Since $N \to \infty$ asymptotically as $\omega \downarrow 0$ and $\widehat{\xi}_n \to \xi$ in probability as $n \to \infty$, by Theorem 2.1 of Gut (2009), $\widehat{\xi}_N^2 \to \xi^2$ in probability. Hence, dividing all sides of Equation (3.18) by $n_\omega$ and letting $\omega \downarrow 0$, we prove $N/n_\omega \to 1$ asymptotically as $\omega \downarrow 0$. ■

Now with all the necessary lemmas and theorem laid down, the main theorem of this work will be stated and proved.

**Theorem 2.2** (Main Theorem)**.** *If the parent distribution(s) is(are) such that* $E[\widehat{\xi}_n^2]$ *exist, then the stopping rule in Equation* (2.32) *yields:*

$$\begin{aligned}
&\textit{Part 1: } P\left(T_N - \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} < \theta < T_N + \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}\right) \to 1-\alpha \textit{ as } N \to \infty. \\
&\textit{Part 2: } \frac{2z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} \le \omega
\end{aligned} \tag{2.37}$$

*Proof.* Here we proceed along the lines of (Chattopadhyay and De, 2016).

Part 1: We define $n_1 = (1-\rho)n_\omega$ and $n_2 = (1+\rho)n_\omega$ for $0 < \rho < 1$. From Lee (1990),

$$\mathbf{Y}_n = \left[\sqrt{n}(T_{1n} - \theta_1), \sqrt{n}(T_{2n} - \theta_2)\right]' \xrightarrow{\mathcal{L}} N_2(\mathbf{0}, \mathbf{\Sigma}),$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \xi_1^2 & \xi_{12} \\ \xi_{12} & \xi_2^2 \end{bmatrix}.$$

So we need to show that $\mathbf{Y}_N \xrightarrow{\mathcal{L}} N_2(\mathbf{0}, \mathbf{\Sigma})$. Let $\mathbf{D}' = [a_0, a_1]$. Then $\mathbf{D}'\mathbf{Y}_N = \mathbf{D}'\mathbf{Y}_{n_\omega} + (\mathbf{D}'\mathbf{Y}_N - \mathbf{D}'\mathbf{Y}_{n_\omega})$. It is therefore sufficient to show that $(\mathbf{D}'\mathbf{Y}_N - \mathbf{D}'\mathbf{Y}_{n_\omega}) \xrightarrow{P} 0$ as $N \to \infty$. Note that

$$(\mathbf{D}'\mathbf{Y}_N - \mathbf{D}'\mathbf{Y}_{n_\omega}) = a_0\sqrt{N}(T_{1N} - T_{1n_\omega}) + a_1\sqrt{N}(T_{2N} - T_{2n_\omega}) + \left(\sqrt{\frac{N}{n_\omega}} - 1\right)\mathbf{D}'\mathbf{Y}_{n_\omega} \tag{2.38}$$

For a fixed $\varepsilon > 0$,

$$
\begin{aligned}
&P\left\{\left|a_0\sqrt{N}(T_{1N} - T_{1n_\omega}) + a_1\sqrt{N}(T_{2N} - T_{2n_\omega})\right| > \varepsilon\right\} \\
&\quad \le P\left\{\left|a_0\sqrt{N}(T_{1N} - T_{1n_\omega}) + a_1\sqrt{N}(T_{2N} - T_{2n_\omega})\right| > \varepsilon, |N - n_\omega| < \rho n_\omega\right\} \\
&\qquad + P\left\{|N - n_\omega| > \rho n_\omega\right\} \\
&\quad \le P\left\{\max_{n_1<n<n_2} \sqrt{n}\,|T_{1n} - T_{1n_\omega}| > \frac{\varepsilon}{2|a_0|}\right\} + P\left\{\max_{n_1<n<n_2} \sqrt{n}\,|T_{2n} - T_{2n_\omega}| > \frac{\varepsilon}{2|a_1|}\right\} \\
&\qquad + P\left\{|N - n_\omega| > \rho n_\omega\right\} \\
&\quad \le P\left\{\max_{n_1<n<n_2} \sqrt{n}\left|\sum_{i=1}^{k} l_{1i}U_{in} - \sum_{i=1}^{k} l_{1i}U_{in_\omega}\right| > \frac{\varepsilon}{2|a_0|}\right\} \\
&\qquad + P\left\{\max_{n_1<n<n_2} \sqrt{n}\left|\sum_{i=1}^{k} l_{2i}V_{in} - \sum_{i=1}^{k} l_{2i}V_{in_\omega}\right| > \frac{\varepsilon}{2|a_1|}\right\} \\
&\qquad + P\left\{|N - n_\omega| > \rho n_\omega\right\} \\
&\quad \le \sum_{i=1}^{k} P\left\{\max_{n_1<n<n_2} \sqrt{n}\,|U_{in} - U_{in_\omega}| > \frac{\varepsilon}{2|a_0|kl_{1i}}\right\} \\
&\qquad + \sum_{i=1}^{k} P\left\{\max_{n_1<n<n_2} \sqrt{n}\,|V_{in} - V_{in_\omega}| > \frac{\varepsilon}{2|a_1|kl_{2i}}\right\} \\
&\qquad + P\left\{|N - n_\omega| > \rho n_\omega\right\}
\end{aligned}
$$

Using Lemma 2.3, we have $N/n_\omega \xrightarrow{P} 1$, and $U_{in}$ and $V_{in}$, $i = 1, \ldots, k$, are U-statistics which satisfy Anscombe's uniform continuous in probability condition, we conclude that for all $\varepsilon > 0$, $\exists \eta > 0, N_0 > 0$ such that

$$
P\left\{\left|a_0\sqrt{N}(T_{1N} - T_{1n_\omega}) + a_1\sqrt{N}(T_{2N} - T_{2n_\omega})\right| > \varepsilon\right\} < \eta,\ \forall N > N_0.
$$

This implies that $a_0\sqrt{N}(T_{1N} - T_{1n_\omega}) + a_1\sqrt{N}(T_{2N} - T_{2n_\omega}) \xrightarrow{P} 0$ as $N \to \infty$. Now, $\mathbf{D}'\mathbf{Y}_{n_\omega} \xrightarrow{\mathcal{L}} N_2(\mathbf{0}, \mathbf{\Sigma})$ and using Lemma 2.3, we have $N/n_\omega \xrightarrow{P} 1$ and, then $(\sqrt{N/n_\omega} - 1)\mathbf{D}'\mathbf{Y}_{n_\omega} \xrightarrow{P} 0$ as $N \to \infty$. Therefore, from Equation (2.38), we know that $(\mathbf{D}'\mathbf{Y}_N - \mathbf{D}'\mathbf{Y}_{n_\omega}) \xrightarrow{P} 0$, that is, $\mathbf{Y}_N \xrightarrow{\mathcal{L}} N_2(\mathbf{0}, \mathbf{\Sigma})$. We define $R(u, v) = \frac{g_1(u)}{g_2(v)}$, if $g_2(v) \neq 0$. By Taylor series expansion, we can expand $R(T_{1N}, T_{2N})$ around $(\theta_1, \theta_2)$ as

$$R(T_{1N}, T_{2N}) = R(\theta_1, \theta_2) + \frac{g_1'(\theta_1)}{g_2(\theta_2)}(T_{1N} - \theta_1) - \frac{g_1(\theta_1)g_2'(\theta_2)}{g_2^2(\theta_2)}(T_{2N} - \theta_2) + h_N,$$

where

$$\begin{aligned} h_N = \frac{1}{2}\Bigg\{ & \frac{g_1''(a)}{g_2(b)}(T_{1N} - \theta_1) - \frac{2g_1'(a)g_2'(b)}{g_2^2(b)}(T_{1N} - \theta_1)(T_{2N} - \theta_2) \\ & + g_1(a)\left(\frac{g_2''(b)g_2^2(b) - 2g_2'(b)g_2(b)}{g_2^4(b)}\right)(T_{2N} - \theta_2)^2\Bigg\}, \end{aligned} \tag{2.39}$$

$$a = \theta_1 + p(T_{1N} - \theta_1), \quad b = \theta_2 + p(T_{2N} - \theta_2), \quad \text{and} \quad p \in (0, 1).$$

Thus,

$$\sqrt{N}\left(R(T_{1N}, T_{2N}) - R(\theta_1, \theta_2)\right) = \mathbf{D}'\mathbf{Y}_N + \sqrt{N}h_N \tag{2.40}$$

where $\mathbf{D}' = \left[\frac{g_1'(\theta_1)}{g_2(\theta_2)}, -\frac{g_1(\theta_1)g_2'(\theta_2)}{g_2^2(\theta_2)}\right]$.

From Lee (1990) and Anscombe's (1952) CLT, $\sqrt{N}(U_{iN} - \theta_{1i})$ and $\sqrt{N}(V_{iN} - \theta_{2i})$ converge in distribution to normal distributions. This implies that $\sqrt{N}(T_{1N} - \theta_1)$ and $\sqrt{N}(T_{2N} - \theta_2)$ also converge in distribution to normal. Also, both $(T_{1N} - \theta_1)$ and $(T_{2N} - \theta_2)$ converge to 0 almost surely. Hence, $\sqrt{N}h_N \xrightarrow{P} 0$.

Therefore,

$$\sqrt{N}\left(T_N - \theta\right) = \sqrt{N}\left(R(T_{1N}, T_{2N}) - R(\theta_1, \theta_2)\right) \xrightarrow{\mathcal{L}} N(0, \mathbf{D}'\mathbf{\Sigma}\mathbf{D}) \qquad \text{as} \qquad N \to \infty.$$

Part 2: Using stopping rule $N$ in Equation (2.32) we have, for all $N$,

$$\begin{aligned} \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \widehat{\xi}_N^2 \le N &\implies 4z_{\alpha/2}^2\frac{\widehat{\xi}_N^2}{N} \le \omega^2 \\ &\implies 2z_{\alpha/2}\frac{\widehat{\xi}_N}{\sqrt{N}} \le \omega \end{aligned}$$

■

## 2.7 Application to Some Widely Used Effect Sizes

As an illustration of our sequential procedure, we will show its application in detail for the standardized mean difference, coefficient of variation, and the regression coefficient (slope) from a simple linear model. Other effect sizes, as we previously explained, as well as linear functions of those effect sizes for multiple groups, can be implemented in a similar way. We focus on these three effect sizes because of their wide usage in psychology and related fields.

### 2.7.1 Standardized Mean Difference

Suppose $X_{11}, X_{12}, \ldots, X_{1n}$ are independent random samples from a distribution $F_1$ with mean $\mu_1$ and variance $\sigma^2$, and $X_{21}, X_{22}, \ldots, X_{2n}$ are independent random samples from another distribution $F_2$ with mean $\mu_2$ and variance $\sigma^2$. The population standardized mean difference, from Equation (2.5) is estimated by the sample standardized mean difference as

$$d_n = \frac{(\bar{X}_{1n} - \bar{X}_{2n})}{s_{pn}}, \tag{2.41}$$

where $s_{pn} = \sqrt{\frac{1}{2}(s_{1n}^2 + s_{2n}^2)}$ is the square root of the pooled sample variance. Using Theorem A.1 the asymptotic distribution of the sample standardized mean difference, $d_n$ is given by

$$\sqrt{n}\,(d_n - \delta) \xrightarrow{\mathcal{L}} N(0, \xi^2), \tag{2.42}$$

where the asymptotic variance of $d_n$ is given by

$$\xi^2 = 2 - \frac{(\mu_1 - \mu_2)(\mu_{13} - \mu_{23})}{\sigma^4} + \frac{(\mu_1 - \mu_2)^2}{4\sigma^6}\left(\frac{\mu_{14} + \mu_{24}}{4} - \frac{\sigma^4}{2}\right) \tag{2.43}$$

and $\mu_{ij}$ is the $j$th central moment of distribution $F_i$, for $i = 1, 2$. Thus, we have a consistent estimator of $\xi^2$, which is given as

$$\widehat{\xi}_n^2 = \max\{V_n^2, n^{-3}\} \tag{2.44}$$

with $V_n^2$ given by

$$V_n^2 = 2 - \frac{(\bar{X}_{1n} - \bar{X}_{2n})(\hat{\mu}_{13n} - \hat{\mu}_{23n})}{s_{pn}^4} + \frac{(\bar{X}_{1n} - \bar{X}_{2n})^2}{4s_{pn}^6}\left(\frac{\hat{\mu}_{14n} + \hat{\mu}_{24n}}{4} - \frac{s_{pn}^4}{2}\right). \tag{2.45}$$

where for $i = 1, 2$, $\widehat{\mu}_{i3n}$ and $\widehat{\mu}_{i4n}$ are U-statistics for $\mu_{i3}$ and $\mu_{i4}$, respectively, which are defined in Equations (A.1) and (A.2). Theorem A.2 in the Appendix shows that the (approximate) $100(1-\alpha)\%$ confidence interval for the population standardized mean difference, $\delta$, is given by

$$\left(d_N - \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}, d_N + \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}\right), \tag{2.46}$$

which is formed using $N$ observations and achieves the specified coverage probability of $1-\alpha$, asymptotically. Additionally, Theorem A.2 proves that the confidence interval for $\delta$ given in Equation (2.46) always achieves a sufficiently narrow width (less than $\omega$).

The sequential procedure we developed can be used in constructing an approximate $100(1-\alpha)\%$ confidence interval for the parameter $\delta$, so that the width of the confidence interval is less than $\omega$ under a distribution-free framework. Additionally, using Theorem A.2, it can be shown that for large sample sizes, the confidence interval will also achieve the specified coverage probability $1-\alpha$ asymptotically. Nevertheless, for different distributions the sampling distribution of the final sample size will vary and this distribution has no known way to be analytically derived. We illustrate the properties of the final sample size empirically with a Monte Carlo simulation. Note that our method is mathematically justified and we provide Monte Carlo simulation results for descriptive purposes as well as to illustrate the properties of our method for a variety of finite sample sizes. We acknowledge that our large sample theory framework may not work well in all finite sample size situations for arbitrary distributions.

**Characteristics of the Final Sample Size: An Empirical Demonstration**

We now demonstrate the properties of our method using a Monte Carlo simulation for constructing $100(1-\alpha)\%$ confidence interval for population standardized mean difference, $\delta$, such that the width of the confidence interval is less than $\omega$ and the confidence interval achieves, asymptotically, the specified coverage probability $1-\alpha$. This is done by implementing the sequential procedure via Monte Carlo simulations by drawing random samples from three distributions; gamma, lognormal, and normal under several parameter combinations.

We implement the proposed sequential procedure and, for the sample size ($N$), we estimate the mean sample size ($\overline{N}$), the standard error of $\overline{N}$ (s($\overline{N}$)), coverage probability ($p$), the standard error of estimated coverage probability($s_p$), and average length of confidence intervals $\bar{w}_N$. We use 5,000 replications for each condition of the simulation study. We chose parameters of the distributions that, in our experience, are reasonable scenarios in applied research. In each replication, we first draw $m$ observations from the populations and then follow the algorithm of the purely sequential procedure outlined in Section 2.5 by drawing $m' = 1$ observations at each stage after the pilot stage. We summarize our findings in Table 2.1. In all cases in Table 2.1, the seventh column suggests that the coverage probability is close to the target coverage probability of either 90% or 95%, respectively. Also, in all cases, the average width is less than $\omega$. The fifth column indicates the ratio of the average final sample size ($N$) to the optimal sample size ($n_\omega$) is close to 1. Furthermore, notice that the mean confidence interval width is just below the desired width. Thus, our procedure is shown to work well in a variety of situations, demonstrating empirically (for finite samples) what has been shown mathematically (under large sample theory).

### 2.7.2 Coefficient of Variation

Suppose $X_1, X_2, \ldots, X_n$ are independent random samples from a distribution $F$ with mean $\mu$ and variance $\sigma^2$, then using Theorem 2.1, the asymptotic distribution of the sample coefficient

Table 2.1. Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\delta$ with $\omega = 0.2$

| Distributions | $\delta$ | $\bar{N}$ | $s_{\bar{N}}$ | $n_\omega$ | $\bar{N}/n_\omega$ | $p$ | $s_p$ | $\bar{w}_N$ |
|---|---|---|---|---|---|---|---|---|
| | | | $\alpha = 0.10$ | | | | | |
| N(10, 1), N(9.7, 1) | 0.3 | 548.5 | 0.0013 | 548 | 1.0010 | 0.8986 | 0.0043 | 0.1998 |
| N(10, 1), N(9.6, 1) | 0.4 | 553.3 | 0.0017 | 552 | 1.0020 | 0.9056 | 0.0041 | 0.1998 |
| N(10, 1), N(9.5, 1) | 0.5 | 559.5 | 0.0022 | 559 | 1.0010 | 0.8990 | 0.0043 | 0.1998 |
| LN(2.991, 0.09975), LN(2.96, 0.1028) | 0.3 | 548.7 | 0.0051 | 549 | 0.9994 | 0.9012 | 0.0042 | 0.1995 |
| LN(2.991, 0.09975), LN(2.95, 0.1039) | 0.4 | 554.2 | 0.0058 | 555 | 0.9986 | 0.9016 | 0.0042 | 0.1995 |
| LN(2.991, 0.09975), LN(2.939, 0.105) | 0.5 | 560.0 | 0.0082 | 562 | 0.9964 | 0.8962 | 0.0043 | 0.1990 |
| Ga(100, 0.1), Ga(94.09, 0.1031) | 0.3 | 548.6 | 0.0037 | 548 | 1.0010 | 0.8968 | 0.0043 | 0.1997 |
| Ga(100, 0.1), Ga(92.16, 0.1042) | 0.4 | 553.7 | 0.0046 | 554 | 0.9995 | 0.8970 | 0.0043 | 0.1996 |
| Ga(100, 0.1), Ga(90.25, 0.1053) | 0.5 | 559.8 | 0.0056 | 560 | 0.9997 | 0.8948 | 0.0043 | 0.1996 |
| | | | $\alpha = 0.05$ | | | | | |
| N(10, 1), N(9.7, 1) | 0.3 | 778.3 | 0.0016 | 777 | 1.0020 | 0.9538 | 0.0003 | 0.1999 |
| N(10, 1), N(9.6, 1) | 0.4 | 785.0 | 0.0021 | 784 | 1.0010 | 0.9494 | 0.0031 | 0.1999 |
| N(10, 1), N(9.5, 1) | 0.5 | 793.6 | 0.0026 | 793 | 1.0010 | 0.9456 | 0.0032 | 0.1999 |
| LN(2.991, 0.09975), LN(2.96, 0.1028) | 0.3 | 779.4 | 0.0046 | 779 | 1.0000 | 0.9468 | 0.0032 | 0.1998 |
| LN(2.991, 0.09975), LN(2.95, 0.1039) | 0.4 | 787.3 | 0.0054 | 787 | 1.0000 | 0.9464 | 0.0032 | 0.1998 |
| LN(2.991, 0.09975), LN(2.939, 0.105) | 0.5 | 796.6 | 0.0086 | 798 | 0.9983 | 0.9458 | 0.0032 | 0.1995 |
| Ga(225, 0.0667), Ga(216.1, 0.0680) | 0.3 | 778.5 | 0.0037 | 778 | 1.0010 | 0.9428 | 0.0033 | 0.1998 |
| Ga(225, 0.0667), Ga(213.2, 0.0685) | 0.4 | 785.2 | 0.0046 | 785 | 1.0000 | 0.9448 | 0.0032 | 0.1998 |
| Ga(225, 0.0667), Ga(210.2, 0.0690) | 0.5 | 793.9 | 0.0063 | 794 | 0.9999 | 0.9448 | 0.0032 | 0.1997 |

Note: $\delta$ is the population standardized mean difference; $\bar{N}$ is the mean final sample size; $p$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $\delta$; $s(\bar{N})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_\omega$ is the theoretical sample size if the procedure is used with the population parameters; $s(p)$ is the standard error of $p$; $\bar{w}_N$ average length of confidence intervals for $\delta$ based on $N$ observations; tabled values are based on 5,000 replications of a Monte Carlo simulation study from distributions Normal (N) with parameters mean and variance, lognormal (LN) with parameters log-mean and log-sd and Gamma (Ga) with parameters shape and scale.

of variation $k_n = s_n/\bar{X}_n$ is

$$\sqrt{n}\,(k_n - \kappa) \xrightarrow{\mathcal{L}} N(0, \xi^2), \tag{2.47}$$

where

$$\xi^2 = \frac{\mu_4}{4\mu^2\sigma^2} - \frac{\sigma^2}{4\mu^2} - \frac{\mu_3}{\mu^3} + \frac{\sigma^4}{\mu^4} \tag{2.48}$$

and $\mu_\nu = \mathrm{E}\left[(X-\mu)^\nu\right]$, (for $\nu = 3, 4$, provided the fourth moment exists). This approach yields the same asymptotic variance as found by Albrecher et al. (2010) (although they used a different method to derive the expression). Thus, $k_n$, which is a consistent estimator of the population coefficient of variation $\kappa = \sigma/\mu$, is asymptotically distributed as normal with mean $\kappa$ and variance, $\xi^2/n$.

Using Heffernan (1997) and Abbasi et al. (2010), we have estimators based on U-statistics for the population third central moment, namely, $\mu_3 = \mathrm{E}[X-\mu]^3$, and the population fourth central moment, namely $\mu_4 = \mathrm{E}[X-\mu]^4$. Let the estimators be denoted as $\hat{\mu}_{3n}$ and $\hat{\mu}_{4n}$ respectively. The expressions of $\hat{\mu}_{3n}$ and $\hat{\mu}_{4n}$ are given in Equations (A.4) and (A.5) in the Appendix. Thus we have a consistent estimator of $\xi^2$ which is given by

$$\widehat{\xi_n^2} = \max\left\{V_n^2, n^{-3}\right\} \tag{2.49}$$

where $V_n^2$ is given by

$$V_n^2 = \frac{\hat{\mu}_{4n}}{4\bar{X}_n^2 s_n^2} - \frac{s_n^2}{4\bar{X}_n^2} - \frac{\hat{\mu}_{3n}}{\bar{X}_n^3} + \frac{s_n^4}{\bar{X}_n^4}. \tag{2.50}$$

The small positive term $n^{-3}$, for large sample size $n$, is used to ensure that we do not get a negative estimate of $\xi^2$ as there is a nonzero chance, though it may be small, that the sample estimate of $V_n^2$ may be negative. Theorem A.3 shows that the $100(1-\alpha)\%$ confidence interval for coefficient of variation is given by

$$\left(k_N - \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}, k_N + \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}\right) \tag{2.51}$$

achieves the specified coverage probability of $100(1-\alpha)\%$, asymptotically. Additionally, Theorem A.3 proves that the confidence interval for $\kappa$ given in Equation (2.51) always achieves a sufficiently narrow width (less than $\omega$).

### Characteristics of the Final Sample Size: An Empirical Demonstration

We now demonstrate the properties of our method using a Monte Carlo simulation for constructing $100(1-\alpha)\%$ confidence interval for population coefficient of variation, $\kappa$, such that the width of the confidence interval is less than $\omega$ and the confidence interval achieves, asymptotically, the specified coverage probability $100(1-\alpha)\%$. This is done by implementing the sequential procedure via Monte Carlo simulations by drawing random samples from a pair of distributions.

We implement the purely sequential procedure and, for the sample size ($N$), we estimate the mean sample size ($\overline{N}$), the standard error ($s(\overline{N})$) of $\overline{N}$, coverage probability ($p$), the standard error of estimated coverage probability ($s_p$), and average length of confidence intervals $\bar{w}_N$, based on 5,000 replications by drawing random samples from several distributions: gamma, lognormal, normal. The parameters of the distribution are chosen to represent possible scenarios in research. In all cases, the number of replications used is 5,000. In each replication, we first draw $m$ observations from the populations and then follow the algorithm of the purely sequential procedure by drawing $m' = 1$ observations at each stage after the pilot stage. We summarize our findings in Table 2.2. In all cases in Table 2.2, the seventh column suggests that the coverage probability is close to the target coverage probability of 90% and 95%, respectively. Also in all cases, the average width is less than $\omega$. The fifth column indicates the ratio of the average final sample size ($N$) to the optimal sample size ($n_\omega$) is close to 1.

### 2.7.3 Regression Coefficient: Simple Linear Model

Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are pairs of observation from a simple linear regression model of the form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2.52}$$

where $Y_i$ is the dependent variable, $X_i$ is the independent variable, $\varepsilon_i$s are independent and identically distributed errors, and $\beta_0$ and $\beta_1$ are the unknown regression parameters. Now, we consider the effect of $X$ on $Y$, which is the regression coefficient $\beta_1 = \sigma_{XY}/\sigma_X^2$, where $\sigma_{XY}$ is the population covariance between $X$ and $Y$, and $\sigma_X^2$ is the population variance of

Table 2.2. Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\kappa$ with $\omega = 0.04$

| Distribution | $\kappa$ | $\bar{N}$ | $s_{\bar{N}}$ | $n_\omega$ | $\bar{N}/n_\omega$ | $p$ | $s_p$ | $\bar{w}_N$ |
|---|---|---|---|---|---|---|---|---|
| | | | | $\alpha = 0.10$ | | | | |
| N(10, 4) | 0.2 | 178.8 | 0.0049 | 147 | 1.2170 | 0.8844 | 0.0049 | 0.0351 |
| N(10, 9) | 0.3 | 368.6 | 0.0104 | 360 | 1.0240 | 0.8890 | 0.0044 | 0.0389 |
| N(10, 16) | 0.4 | 713.5 | 0.0164 | 715 | 0.9979 | 0.8916 | 0.0044 | 0.0397 |
| LN(1, 0.1980) | 0.2 | 182.3 | 0.0080 | 159 | 1.1470 | 0.8650 | 0.0048 | 0.0350 |
| LN(1, 0.2936) | 0.3 | 397.6 | 0.0245 | 429 | 0.9268 | 0.8680 | 0.0048 | 0.0388 |
| LN(1, 0.3853) | 0.4 | 852.5 | 0.0592 | 971 | 0.8780 | 0.8598 | 0.0049 | 0.0397 |
| Ga(25, 0.6) | 0.2 | 173.0 | 0.0054 | 141 | 1.2270 | 0.8794 | 0.0046 | 0.0347 |
| Ga(11.11, 0.6) | 0.3 | 335.2 | 0.0129 | 332 | 1.0100 | 0.8780 | 0.0046 | 0.0386 |
| Ga(6.25, 0.6) | 0.4 | 610.0 | 0.0239 | 628 | 0.9713 | 0.8810 | 0.0046 | 0.0396 |
| | | | | $\alpha = 0.10$ | | | | |
| N(10, 4) | 0.2 | 241.2 | 0.0063 | 208 | 1.1600 | 0.9422 | 0.0033 | 0.0363 |
| N(10, 9) | 0.3 | 519.1 | 0.0126 | 510 | 1.0180 | 0.9408 | 0.0033 | 0.0392 |
| N(10, 16) | 0.4 | 1014.0 | 0.0195 | 1015 | 0.9992 | 0.9458 | 0.0032 | 0.0398 |
| LN(1, 0.1980) | 0.2 | 247.4 | 0.0105 | 225 | 1.1000 | 0.9234 | 0.0038 | 0.0362 |
| LN(1, 0.2936) | 0.3 | 570.3 | 0.0322 | 608 | 0.9381 | 0.9244 | 0.0037 | 0.0392 |
| LN(1, 0.3853) | 0.4 | 1243.0 | 0.0770 | 1378 | 0.9022 | 0.9210 | 0.0038 | 0.0398 |
| Ga(25, 0.6) | 0.2 | 233.0 | 0.0071 | 200 | 1.1650 | 0.9342 | 0.0035 | 0.0359 |
| Ga(11.11, 0.6) | 0.3 | 472.5 | 0.0163 | 472 | 1.0010 | 0.9356 | 0.0035 | 0.0390 |
| Ga(6.25, 0.6) | 0.4 | 871.3 | 0.0301 | 892 | 0.9768 | 0.9402 | 0.0034 | 0.0397 |

Note: $\kappa$ is the population coefficient of variation; $\bar{N}$ is the mean final sample size; $p$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $\delta$; $s(\bar{N})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_\omega$ is the theoretical sample size if the procedure is used with the population parameters; $s(p)$ is the standard error of $p$; $\bar{w}_N$ average length of confidence intervals for $\delta$ based on $N$ observations; tabled values are based on 5,000 replications of a Monte Carlo simulation study from distributions Normal (N) with parameters mean and variance, lognormal (LN) with parameters

$X$. Since the value of $\beta_1$ is unknown in practice, we estimate it by

$$b_{1n} = \frac{\sum(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum(x_i - \bar{x}_n)^2} = \frac{s_{XYn}}{s^2_{Xn}} \tag{2.53}$$

where $s_{XYn}$ is the unbiased estimator for covariance between $X$ and $Y$ for a sample of size $n$ and $s^2_{Xn}$ is the unbiased estimator for variance of $X$ for a sample of size $n$. Using

Theorem A.4, the central limit theorem for $b_{1n} = s_{XYn}/s_{Xn}^2$ is

$$\sqrt{n}\,(b_{1n} - \beta_1) \xrightarrow{\mathcal{L}} N(0, \xi^2), \tag{2.54}$$

where the asymptotic variance is given by

$$\xi^2 = \frac{\mu_{22}}{\sigma_X^4} - \frac{2\sigma_{XY}\mu_{31}}{\sigma_X^6} + \frac{\sigma_{XY}^2\mu_{40}}{\sigma_X^8}. \tag{2.55}$$

A consistent estimator for $\xi^2$, similar to that given in Equation (A.11) in the Appendix, can be used to construct $100(1-\alpha)\%$ confidence interval for the regression parameter $\beta_1$ without considering any normality assumption of the errors. Theorem A.5 shows that the $100(1-\alpha)\%$ confidence interval for regression parameter, $\beta_1$ is given by

$$\left( b_{1N} - \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}, b_{1N} + \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} \right) \tag{2.56}$$

achieves the specified coverage probability of $1-\alpha$, asymptotically. Additionally, Theorem A.5 proves that the confidence interval for $\beta_1$ given in Equation (2.56) achieves a sufficiently narrow width (less than $\omega$) using the sequentially estimated sample size, which is an estimate of the theoretically optimal sample size.

## 2.8 An Extension: Unbalanced Design

There are situations in which the sample size per group may be different. Under such designs, we can also use the sequential procedure. As an example, we consider a single-factor between-subjects unbalanced design related to Example 6. For the $k$th group, suppose $X_{k1}, \ldots, X_{kn_k}$ are independent and identically distributed random variables with unknown means $\mu_k$ and unknown variances $\sigma_k^2$. Thus, in total, there are $n = \sum_{k=1}^K n_k$ observations from $K$ groups. Then, the population contrast related to the corresponding scenario is given by

$$\psi = \sum_{k=1}^{K} c_k \mu_k \tag{2.57}$$

where $c_1, \ldots, c_K$ are known coefficients and $\sum_{k=1}^K c_k = 1$. An estimator of the contrast $\psi$ is $\hat{\psi}_n = \sum_{k=1}^K c_k \bar{X}_{kn_k}$, where $\bar{X}_{1n_1}, \ldots, \bar{X}_{Kn_K}$ are the sample group means. Now,

$$\mathrm{Var}(\hat{\psi}_n) = \sum_{k=1}^K \frac{c_k^2 \sigma_k^2}{n_k}. \tag{2.58}$$

Thus, the $100(1-\alpha)\%$ confidence interval for $\psi$ is given by

$$\left( \hat{\psi}_n - z_{\alpha/2} \sqrt{\sum_{k=1}^K \frac{c_k^2 \sigma_k^2}{n_k}}, \hat{\psi}_n + z_{\alpha/2} \sqrt{\sum_{k=1}^K \frac{c_k^2 \sigma_k^2}{n_k}} \right). \tag{2.59}$$

The length of the confidence interval given by Equation (2.59) is

$$w_n = 2z_{\alpha/2} \sqrt{\sum_{k=1}^K \frac{c_k^2 \sigma_k^2}{n_k}}. \tag{2.60}$$

Here, we need to find the minimum total sample size with the restriction

$$2z_{\alpha/2} \sqrt{\sum_{k=1}^K \frac{c_k^2 \sigma_k^2}{n_k}} \le \omega \implies \sum_{k=1}^K \frac{c_k^2 \sigma_k^2}{n_k} \le \frac{\omega^2}{4z_{\alpha/2}^2}. \tag{2.61}$$

Using the Lagrange multiplier method, we define a function

$$g_{n_k \lambda} = \sum_{k=1}^K n_k + \lambda \left( \sum_{k=1}^K \frac{c_k^2 \sigma_k^2}{n_k} - \frac{\omega^2}{4z_{\alpha/2}^2} \right). \tag{2.62}$$

We note that the Lagrange multiplier method is a method which can be used to find local minimum or maximum of a function under equality constraints (for e.g. Vapnyarskii, 2001). By partial differentiation of $g_{n_k \lambda}$ with respect to $n_k$ and $\lambda$, we have for $i = 1, 2, \ldots, K$

$$\frac{\partial}{\partial n_k} g_{n_k \lambda} = 0 \implies 1 - \lambda \frac{c_k^2 \sigma_k^2}{n_k} = 0 \implies n_k = \sqrt{\lambda} c_k \sigma_k \tag{2.63}$$

and

$$\frac{\partial}{\partial \lambda} g_{n_k \lambda} = 0 \implies \sum_{k=1}^K \frac{c_k^2 \sigma_k^2}{n_k} = \frac{\omega^2}{4z_{\alpha/2}^2} \implies \sum_{k=1}^K \frac{c_k^2 \sigma_k^2}{\sqrt{\lambda} c_k \sigma_k} = \frac{\omega^2}{4z_{\alpha/2}^2}$$

$$\implies \sqrt{\lambda} = \frac{4z_{\alpha/2}^2}{\omega^2} \sum_{k=1}^K c_k \sigma_k. \tag{2.64}$$

Using Equations (2.63) and (2.64), the optimum sample size for the $k$th group is given by

$$n_{k\omega} = \frac{4c_k\sigma_c z^2_{\alpha/2}}{\omega^2}\sum_{k=1}^{K} c_k\sigma_k. \tag{2.65}$$

Thus $n_{k\omega}$ $(k = 1, 2, \ldots, K)$ is the optimum sample size that is required from the $k$th group so as to have a confidence interval of width less than $\omega$. But for $k = 1, 2, \ldots, K$, $n_{k\omega}$ are unknown. So, as before, in order to estimate the optimum sample size for all $K$ groups, we use a sequential method.

## 2.9 Concluding Remarks

In psychology and related disciplines, estimating effect sizes is important and so is quantifying their uncertainties. Correspondingly, wide confidence intervals are undesirable and illustrate the uncertainty with which the population value has been estimated, at some specified level of confidence. Intervals that illustrate a wide range for the population value of the parameter of interest have been termed "embarrassingly large" (see, Cohen, 1994, p. 1102), with Cohen speculating that the reason researchers seldomly, at the time, reported confidence intervals was due to their (embarrassingly large) widths. The AIPE approach to sample size sought to solve embarrassingly large widths by explicitly planning sample size for sufficiently narrow intervals. Although these methods are useful, they have their own shortcoming, namely traditional applications of AIPE tend to require knowledge or speculation of parameters and distribution in order to plan the necessary sample size. Further, traditional applications of AIPE require assumptions about the type of distribution from which the scores were sampled (e.g., normal). This work solves this problem of requiring population parameters and known distributional forms in order to implement the AIPE approach to sample size planning and it does so for a general class of effect size. Importantly, we have worked in a distribution-free environment, where we have not made untenable assumptions about the distribution of the scores from which the observations are sampled. We believe that our approach advances the fields of effect size and sample size planning to improve the state of research design and analysis in psychology and related disciplines. The accuracy in parameter estimation for effect sizes of interest is an important issue (e.g., Maxwell et al., 2008, for a review) and now the approach can be implemented easily without assumptions of known population values and known distributional forms.

For any given population, the variance of the effect size estimator decreases as sample size increases, holding everything else constant. This, in turn, decreases the width of the corresponding confidence interval for the effect size parameter as well as the coverage probability, the probability that the confidence interval will contain the true effect size parameter value. An optimal sample size is desired which can be used to construct a $100(1-\alpha)\%$ confidence interval for the effect size parameter, such that the confidence interval will be as narrow as specified (i.e., the width of the confidence interval will be less than $\omega$ and the coverage probability of the interval is approximately $100(1-\alpha)\%$). A fixed sample size procedure cannot achieve both the coverage probability and the width less than $\omega$ simultaneously. We have shown in this work how to solve this problem for the general effect size parameter under a distribution-free environment.

Our method, unlike fixed sample size procedures, is not based on the assumption on the distribution of the data and the population parameters required to estimate the theoretically optimal sample size (i.e., the sample size if all parameters were known). In this work, we have developed a sequential procedure which provides an estimate of the theoretically true optimal sample size required to construct a $100(1-\alpha)\%$ confidence interval for the effect size parameter such that the confidence interval will be narrow – that is the width of the confidence interval will be less than $\omega$ and the coverage probability of the interval will be approximately $100(1-\alpha)\%$ without assuming any specific distribution for the data. The lack of any assumption on the distribution of the data is a key part of the contribution, as in many situations there is no reason to believe that the distribution of the scores is gamma, lognormal, normal, or some other distribution.

The sequential procedure we developed in this work ensures that the width of the confidence interval for the general effect size will be less than the pre-specified upper bound, $\omega$, and also the coverage probability is approximately $100(1-\alpha)\%$, assuming throughout that the observations are independent and identically distributed but with no assumption of the distribution of the data. Additionally, the ratio of the average final sample size and the theoretically optimal sample size is approximately 1, as we showed with theorems as well as demonstrating empirically via the Monte Carlo simulations.

The traditional AIPE procedure, unlike a sequential AIPE procedure, requires the knowledge or speculation of parameters in order to plan the necessary sample size. After getting the complete data, with sample size as given by the traditional AIPE procedure, the required confidence interval for the effect size is computed. In the sequential AIPE, the analysis of the data is carried out in stages, as it comes, and then finally the confidence interval for the effect size is computed. Unlike traditional AIPE, in the sequential AIPE the data collection always stops after the width of the confidence interval is smaller than $\omega$. The traditional AIPE procedure can be used when the population parameters necessary to compute the required sample size are fully known, however this is not practically possible. In fact, Sen and Ghosh (1981) argued that sequential procedures are economical in terms of sample size.

There are several limitations of our method because the method does not directly consider (a) the problem of continuous availability of participants or observations after each stage; (b) potential difficulty in specification of $m'$ (c) difficulty in specification of $\omega$ and confidence coefficient $100(1-\alpha)\%$; (d) no knowledge of the final sample size at the beginning of the study; and (e) the problem of unbounded confidence intervals (e.g., single-sided confidence intervals which have a limit of positive or negative infinity).

The first limitation is the problem of assuming that the participants or observations can be readily available as and when required. In some situations, after applying stopping rules for the observations collected up to a certain stage, we may need to wait until another opportunity to obtain another $m'$ observations. However, a similar kind of situation may also arise when using a traditional AIPE method as well.

The second limitation is that we need to pre-specify the values of the choice of $m'$, which represents the number of observations that will be added in each stage after the pilot sampling stage or first stage. In some situations it is as easy to collect more than one observation as it is collecting a single observation at every stage. So, as per convenience, the value of $m'$ should be accordingly decided based on economic considerations. For example, Chattopadhyay and Kelley (2017) discussed the choice of $m'$ using an application of a sequential procedure that considered both cost and accuracy for estimating standardized mean difference of the reading scores while studying the impact of same language subtitling (SLS) on reading ability. Suppose the data collection on the reading ability of the students is performed during in-school visits by a surveyor. On any working day at the school, suppose the surveyor is allowed only 2 hours for interviewing students and every day, a certain amount of money is provided to the surveyor for travel cost and an hourly wage, say $60 for 2 hours of work and travel. Because there will be two groups, the choice of $m'$ could be 1, or any other value such as 5 or 10. As the surveyor may just as easily collect $m' = 10$ as $m' = 1$, we generally recommend a larger value of $m'$, all other things being equal. Nevertheless, there is no uniform method which can help take a decision on $m'$ that will fit all scenarios.

In conceptually similar situations, but yet in a different context, consider the way in which a computer adaptive test (CAT), in which the final number of items is usually unknown, additional items are presented to an examinee until the desired accuracy in the estimation of examinee's ability is achieved. Obtaining one more sample (in CAT, presenting an additional item), that is taking $m' = 1$ is better than giving 10 more additional items ($m' = 10$) at a time after pilot stage may ultimately result in oversampling. Suppose after a certain stage, only two more additional items are actually required, but due to pre-specification of $m' = 10$ we have to present eight more items. This will require participants taking more items than is actually required.

The third limitation is that of specifying the value of $\omega$, as there is no uniformly appropriate value. This limitation, however, also exists in traditional AIPE method. This is similar to some extent to the question of "what is the appropriate value of statistical power?" The answer has rules of thumb (e.g., 80% power, power $= 1 - \alpha$, power $= 1 - 2\alpha$, etc.), but with no universal agreement on what should be used. We see this limitation as a type of paradox of choice (e.g., Schwarz, 2004), in that $\omega$ can be specified as any (positive) value, in which the smaller $\omega$ the more accurate an estimate. Nevertheless, by requiring a specific value of $\omega$, researchers may decide that because there is no obvious value, they do not implement the procedure.

The fourth limitation of the sequential AIPE procedure is that of not knowing the final sample size at the start of the study. Since, our sequential procedure is a data-driven procedure, it may lead to a sample size that is so large that it is unreasonable to obtain with the available resources. Nevertheless, the problem of not knowing the true final sample size at the beginning of the study can be palliated by using a sensitivity analysis with parameters and supposed distributions in the sequential framework. This will provide a lot of information about the sensitivity of the final sample size in a variety of scenarios.

The fifth limitation of the sequential AIPE procedure is due to the ratio form. If the numerator of an estimate is non-zero but the denominator is near zero, the ratio can be extremely large in an absolute sense. However, this is not just a problem in our case but is true for effect sizes (estimates) that are (a) functions of ratios and (b) not bounded. Bounded ratios, such as the correlation coefficient, will not suffer from this potential issue. A similar situation in the context of mediation is discussed in Preacher and Kelley (2011) (see also Fieller, 1954).

The procedure we developed for the sequential accuracy in parameter estimation problem of general effect size is applied to several effect sizes such as coefficient of variation, standardized mean difference, and regression coefficient among others. The basic theory of sequential methods is based on the idea of “learn-as-you-go” with the stopping rule instructing a research to continue or stop sampling. Based on the limitations of fixed sample size planning procedures with regard to assumed data distribution and assumed knowledge of population parameters, use of sequential procedures in psychology and related fields can be beneficial. Recent methodological advances for sequential methods, for example, consider the standard error and study cost for the coefficient of variation (Chattopadhyay and Kelley, 2016) and the standardized mean difference (Chattopadhyay and Kelley, 2017). This is the first work, however, to make developments for AIPE in the context of sequential methods and to do so for a general class of effect size measures.

# CHAPTER 3

# A PURELY SEQUENTIAL APPROACH TO ACCURACY IN PARAMETER ESTIMATION FOR POPULATION CORRELATION COEFFICIENTS[1]

## 3.1 Introduction

The correlation coefficient provides a scale-free measure of the magnitude, direction, and strength of the linear relationship between two variables and lies in the interval $[-1, 1]$. The Pearson product moment correlation coefficient is often used when variables are quantitative. However, other correlations exist when variables are ordinal, such as the Kendall's tau or Spearman's rho rank correlation coefficients. In this work, we develop sequential methods for obtaining accurate estimates of population correlation coefficients. We begin with the Pearson product moment correlation due to its popularity in psychology and related fields before generalizing to other correlations and, ultimately, to the squared multiple correlation coefficient in the multiple regression framework. Our work builds on Kelley et al. (2018), but is distinct in important ways as we will discuss.

Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a bivariate distribution of arbitrary form, $F$, with covariance $\sigma_{XY}$ and with the marginal distributions of $X$ and $Y$ having population variances $\sigma_X^2$ and $\sigma_Y^2$, respectively. Throughout the article, it is assumed that observations are drawn from a homogeneous population. The population Pearson product moment correlation coefficient of $X$ and $Y$, is given by

$$\rho = \frac{\sigma_{XY}}{\sqrt{\sigma_X^2 \sigma_Y^2}}. \tag{3.1}$$

In most disciplines, the correlation coefficient is often a primary outcome variable of interest. For this reason, many authors have heavily invested in methodological work for estimation

[1] This chapter is based on Kelley, Bilson Darku, and Chattopadhyay (2017) which is under review for publication.

of and inference for the population correlation coefficient in an effort to better describe quantitative relationships, plan studies that will estimate the correlation coefficient, and perform inferential procedures for the population correlation coefficient. For example, Wolf and Cornell (1986) and Bonett and Wright (2000) emphasized the importance of estimating the population correlation coefficient with a narrow confidence interval, specifically under the assumption of a bivariate normal distribution. Under the same distribution assumptions, Moinester and Gottfried (2014) provided a review of several methods for constructing a narrow confidence interval for the population correlation coefficient. Holding constant the population of interest, the effect size of interest, any bias of the estimator, and the confidence interval coverage, a narrower confidence interval for the parameter is preferred to a wider confidence interval because it illustrates more precision about the parameter of interest. Holding constant or decreasing any bias, one way of increasing precision, and thereby improving accuracy, is to increase the sample size (e.g., Kelley and Maxwell, 2003; Maxwell et al., 2008).

The existing approaches to plan sample size for obtaining a narrow confidence interval for the population correlation coefficient are based on supposed values of one or more population parameters in the context of bivariate normal distributions (e.g., Bonett and Wright, 2000; Corty and Corty, 2011; Moinester and Gottfried, 2014). A framework of sample size planning known as accuracy in parameter estimation (AIPE), which has been developed for constructing sufficiently narrow confidence intervals for several population effect sizes, has traditionally been based on supposed population parameter values (e.g., Kelley and Maxwell, 2003; Kelley and Rausch, 2006; Kelley, 2007b, 2008; Kelley and Lai, 2011; Terry and Kelley, 2012; Lai and Kelley, 2011a,b; Pornprasertmanit and Schneider, 2014). However, a potential problem is the requirement of one or more supposed values of the population parameters, which will generally be unknown.[2] Existing approaches aimed at sample size planning for

[2] In the context of statistical power, an alternative approach that does not require the specification of the population parameter(s) is to specify the minimally important effect size of interest, which basis statistical

obtaining a narrow confidence interval for the population correlation coefficient rely on supposed population values. Further, applications of AIPE and power analysis also rely on supposed population values. When using supposed population values, that is, treating a supposed value as if it were the true population value, the obtained sample size estimates can differ dramatically from what the theoretically optimal sample size value would be if the population parameters were known. In such situations, even small differences in the supposed and actual value of a parameter can lead to large differences in the planned versus actually required sample size, due to power curves being nonlinear.

Fisher's (1915) $z$-transform method can be used to find the confidence interval for the population correlation coefficient, but it is based on the assumption of bivariate normality. However, because we are working under a distribution-free scenario, our confidence interval procedure is built upon the asymptotic distribution of sample correlation coefficient proposed by Lee (1990). Unlike Bonett and Wright (2000), Corty and Corty (2011), and Moinester and Gottfried (2014), our approach is more flexible because (a) it does not require the assumption of the bivariate normal distribution of the two variables and (b) supposed values of the population parameters are not needed to plan the sample size. These two points are critical.

We use a *sequential approach* to find a narrow confidence interval for the population correlation coefficient, which we call *sequential AIPE*. This approach is similar to the "fixed-width confidence interval" method, in which the width of the confidence interval is pre-specified. Sequential AIPE differs from the fixed-width confidence interval approach because sequential AIPE aims to find the minimum value of the sample size such that the confidence interval is sufficiently narrow by pre-specifying the upper bound on the confidence interval width

power based on the minimum parameter value of interest that would be practically of interest or theoretically interesting. O'Brien and Castelloe (2007) discuss that this approach has potential problems, because for important outcomes in which any non-zero effect is important, the planned sample sizes can be extremely large.

(e.g., Sproule, 1985; Mukhopadhyay and De Silva, 2009; Mukhopadhyay and Chattopadhyay, 2012). In particular, the fixed-width confidence interval procedure deals with construction of a confidence interval for a population parameter that has a width which is exactly equal to the pre-specified value of the confidence interval width. By contrast, in sequential AIPE, the aim is to obtain a sufficiently narrow confidence interval for a population parameter such that the confidence interval is not wider than the pre-specified width.

Under the distribution-free scenario, the exact sampling distribution of the sample correlation coefficient cannot be obtained. This is because in the distribution-free environment no underlying distribution is assumed, such as a bivariate normal. Unlike Fisher's method, we are working under the distribution-free scenario, we use the asymptotic distribution of the sample correlation coefficient developed by Lee (1990) to obtain a sufficiently narrow confidence interval for the population correlation coefficient, $\rho$, using the smallest possible sample size.

We first discuss the (traditional) AIPE for the Pearson product moment correlation coefficient and propose a sequential estimation procedure, which extends the ideas of Kelley et al. (2018). The methods of Kelley et al. (2018) were for a generalized effect size consisting of the ratio of linear functions. We then extend the methods used for the Pearson product moment correlation coefficient to Kendall's tau rank correlation coefficient and Spearman's rank correlation coefficient. Thus, the methods that we develop here are for three types of correlation coefficients, which represents a fundamentally different type of effect size than that given in Kelley et al. (2018). Nevertheless, in both cases, our methods use a sequential procedure for constructing a sufficiently narrow confidence interval (i.e., no larger than specified width) with a specified level of confidence without requiring supposed population values. Importantly, we make all of these developments in a distribution-free environment. The distribution-free environment is important because there is often no reason to assume that the underlying distribution of the data for which the correlation coefficient will be calculated would be known (e.g., bivariate normal, bivariate gamma). Finally, we provide an extension of the sequential procedure in order to obtain a sufficiently narrow confidence interval for the squared multiple correlation coefficient, yet here we assume multivariate normality rather than working in a distribution-free environment (due to the current limitations in the distribution-free literature for confidence intervals for the population squared multiple correlation coefficient). For all of the different correlation coefficients discussed, we provide Monte Carlo simulation results that illustrate the characteristics of the procedures in a variety of scenarios.

## 3.2 Accuracy in Parameter Estimation of Pearson's Product Moment Correlation Coefficient

Pearson's product moment correlation coefficient continues to serve as an important role in many disciplines. The sample correlation coefficient based on $n$ observations is given by

$$r_n = \frac{S_{XYn}}{\sqrt{S^2_{Xn}S^2_{Yn}}}. \tag{3.2}$$

where

$$S_{XYn} \equiv U_{1n} = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \tfrac{1}{2}(X_i - X_j)(Y_i - Y_j) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n), \tag{3.3}$$

$$S^2_{Xn} = U_{2n} = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \tfrac{1}{2}(X_i - X_j)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2, \text{ and} \tag{3.4}$$

$$S^2_{Yn} = U_{2n} = \binom{n}{2}^{-1} \sum_{1\le i<j\le n} \tfrac{1}{2}(Y_i - Y_j)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 \tag{3.5}$$

$$\tag{3.6}$$

are the sample covariance of $X$ and $Y$, sample variance of $X$, and sample variance of $Y$ respectively. We use a subscript $n$ to denote the current sample size used in the estimation. For technical details regarding the equivalence, we refer the reader to Lee (1990) and

Mukhopadhyay and Chattopadhyay (2013, 2014). The U-statistics form of the expressions of sample covariance and sample variances are required for proving Theorem 3.1.

Using Lee (1990), the asymptotic variance of $r_n$ is $\xi_\rho^2/n$, where

$$\xi_\rho^2 = \frac{\rho^2}{4}\left(\frac{\mu_{40}}{\sigma_X^4} + \frac{\mu_{04}}{\sigma_Y^4} + \frac{2\mu_{22}}{\sigma_X^2\sigma_Y^2} + \frac{4\mu_{22}}{\sigma_{XY}^2} - \frac{4\mu_{31}}{\sigma_{XY}\sigma_X^2} - \frac{4\mu_{13}}{\sigma_{XY}\sigma_Y^2}\right) \tag{3.7}$$

and $\mu_{ij} = \mathrm{E}\left[(X-\mu_X)^i(Y-\mu_Y)^j\right]$. Then the approximate $100(1-\alpha)\%$ confidence interval for $\rho$ is given by

$$\left(r_n - z_{\alpha/2}\frac{\xi_\rho}{\sqrt{n}}, r_n + z_{\alpha/2}\frac{\xi_\rho}{\sqrt{n}}\right), \tag{3.8}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. The width of the confidence interval defined in Equation (3.8) is given by

$$w_n = 2z_{\alpha/2}\frac{\xi_\rho}{\sqrt{n}}. \tag{3.9}$$

In AIPE problems the sample size required to achieve the sufficient accuracy is solved by specifying the upper bound on the width of the confidence interval, $\omega$. So for a given $\omega$, we have

$$2z_{\alpha/2}\frac{\xi_\rho}{\sqrt{n}} \leq \omega, \tag{3.10}$$

which implies that the necessary sample size to construct $100(1-\alpha)\%$ confidence interval for $\rho$ will be

$$n \geq \left\lceil \frac{4z_{\alpha/2}^2\xi_\rho^2}{\omega^2} \right\rceil \equiv n_\omega, \tag{3.11}$$

where $\lceil x \rceil$ is the ceiling function which is the smallest integer greater than or equal to $x$ (e.g., $\lceil 49.2 \rceil = 50$). Here, $n_\omega$ is the theoretical optimal sample size required to make the $100(1-\alpha)\%$ confidence interval for $\rho$ provided that the asymptotic variance, $\xi_\rho^2$, is known. The optimal sample size $n_\omega$, is unknown as in reality $\xi_\rho^2$ is unknown. We note that the supposed values of $\xi_\rho^2$ cannot be used to estimate $n_\omega$ as this may not guarantee that the condition in Equation (3.11) is satisfied. So, we will use a sequential procedure, which does not need a supposed population parameter value to find out the sample size but will satisfy the condition given in Equation (3.11). Here, we use a consistent estimator of $\xi_\rho^2$ in our sequential procedure to estimate the optimal sample size. The consistent estimator of $\xi_\rho^2$ is given by

$$\hat{\xi}_{\rho n}^2 = \max\left\{V_{\rho n}^2, n^{-\gamma}\right\}, \gamma > 0, \tag{3.12}$$

where

$$V_{\rho n}^2 = \frac{r_n^2}{4}\left(\frac{\hat{\mu}_{40n}}{S_{Xn}^4} + \frac{\hat{\mu}_{04n}}{S_{Yn}^4} + \frac{2\hat{\mu}_{22n}}{S_{Xn}^2 S_{Yn}^2} + \frac{4\hat{\mu}_{22n}}{S_{XYn}} - \frac{4\hat{\mu}_{31n}}{S_{XYn}S_{Xn}^2} - \frac{4\hat{\mu}_{13n}}{S_{XYn}^2 S_{Yn}^2}\right). \tag{3.13}$$

and $\hat{\mu}_{ijn}$'s, the estimators of $\mu_{ij}$'s, are defined in Equations (B.2) – (B.6) in the Appendix. We note that the estimator $V_{\rho n}^2$ is a moment-based estimator of $\xi_\rho^2$ and there is a chance, even though negligible, that it may come out to be negative in some situations. In order to avoid such scenario, if it arises, we use the term $n^{-\gamma}$. Any choice of $\gamma$ will not affect the consistency property of $\hat{\xi}_{\rho n}^2$; hereafter we use $\gamma = 3$.

## 3.3 Accuracy in Parameter Estimation via a Sequential Optimization Procedure

The sample size is not fixed in advance in sequential methodologies as opposed to fixed sample-size procedures. Here, we propose a sequential procedure to construct a $100(1-\alpha)\%$ confidence interval for the correlation coefficient $\rho$ within a distribution-free environment. For details about the general theory of sequential estimation procedures, we refer interested readers to Sen (1981), Ghosh and Sen (1991), Chattopadhyay and Mukhopadhyay (2013), Chattopadhyay and Kelley (2016, 2017), and De and Chattopadhyay (2017). Recall that the optimal sample size $n_\omega$ is unknown due to $\xi^2$ being unknown. We use the consistent estimator of $\xi^2$, namely $\widehat{\xi}_n^2$, which is based on $n$ observations on both $X$ and $Y$. We now develop an algorithm to find an estimate of the optimal sample size via the purely sequential estimation procedure.

**Stage 1:** Observations are collected on variables $X$ and $Y$ for a randomly selected pilot sample of size $m$. We recommend using the pilot sample size, $m$, following Mukhopadhyay (1980), as

$$m = \max\left\{4, \left\lceil \frac{2z_{\alpha/2}}{\omega} \right\rceil\right\}. \tag{3.14}$$

Based on this pilot sample of size $m$, we estimate $\xi^2$ by computing $\widehat{\xi}_m^2$. If $m < \left\lceil \frac{4z_{\alpha/2}^2}{\omega^2}\left(\widehat{\xi}_m^2 + \frac{1}{m}\right)\right\rceil$, then proceed to the next step. Otherwise, if $m \geq \left\lceil \frac{4z_{\alpha/2}^2}{\omega^2}\left(\widehat{\xi}_m^2 + \frac{1}{m}\right)\right\rceil$, stop sampling and set the final sample size equal to $m$.

**Stage 2:** Obtain an additional $m'(\geq 1)$ observations, where $m'(\geq 1)$ is the number of paired observations that are added to the sample in every stage after the pilot stage. Thus, for adding a single pair to the collected data, $m' = 1$. However, if five additional pairs are taken at each stage, then $m' = 5$. Thus, after collecting the pilot sample and the sampling at the next stage, there are $(m + m')$ observations on both $X$ and $Y$. After updating the estimate of $\xi^2$ by computing $\widehat{\xi}_{m+m'}^2$, a check is performed to determine whether $m + m' \geq \left\lceil \frac{4z_{\alpha/2}^2}{\omega^2}\left(\widehat{\xi}_{m+m'}^2 + \frac{1}{m+m'}\right)\right\rceil$. If $m + m' < \left\lceil \frac{4z_{\alpha/2}^2}{\omega^2}\left(\widehat{\xi}_{m+m'}^2 + \frac{1}{m+m'}\right)\right\rceil$ then go to the next step. Otherwise, if $m + m' \geq \left\lceil \frac{4z_{\alpha/2}^2}{\omega^2}\left(\widehat{\xi}_{m+m'}^2 + \frac{1}{m+m'}\right)\right\rceil$ then stop further sampling and report the final sample size as $(m + m')$.

This process of collecting one (or more) observation(s) in each stage after the first stage continues until there are $N$ observations such that $N \geq \left\lceil \frac{4z_{\alpha/2}^2}{\omega^2}\left(\widehat{\xi}_N^2 + \frac{1}{N}\right)\right\rceil$. At this stage, we stop further sampling and report the final sample size as $N$.

Based on the algorithm just outlined, a sampling stopping rule can be defined as follows:

$$N \text{ is the smallest integer } n(\geq m) \text{ such that } n \geq \frac{4z_{\alpha/2}^2}{\omega^2}\left(\widehat{\xi}_n^2 + \frac{1}{n}\right), \tag{3.15}$$

where the term $n^{-1}$ is a correction term ensuring that the sampling process does not stop too early for the optimal sample size because of the use of the approximate expression.

The inclusion of the correction term retains the convergence property of $\widehat{\xi}_n^2 + n^{-1}$, thus $\widehat{\xi}_n^2 + n^{-1}$ converges to $\xi^2$ for large sample sizes. For details of the correction term, refer to Chattopadhyay and De (2016), Sen (1981), and Chattopadhyay and Kelley (2016, 2017).

Following the sequential procedure, the $100(1-\alpha)\%$ confidence interval for the population correlation coefficient, $\rho$, is given by

$$\left(r_N - \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}, r_N + \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}\right). \tag{3.16}$$

The width of the confidence interval in Equation (3.16) will be less than $\omega$, in accord with our method's specifications. Lemma 3.1 proves that the estimated sample size from sequential procedure, $N$, is finite. Also, Theorem 3.1 in the Appendix proves that the confidence interval achieves the specified coverage probability $1-\alpha$ asymptotically using $N$ which is the estimate of the smallest possible sample size ($n_\omega$). Here the smallest possible sample size indicates that the sample size required to obtain a sufficiently narrow $100(1-\alpha)\%$ confidence interval is $n_\omega$. Since $n_\omega$ is unknown, using sequential procedure we can find a consistent estimator, $N$, of $n_\omega$ (proved in Lemma 2). Additionally, Theorem 3.1 proves that the confidence interval for $\rho$ given in Equation (3.16) always achieves a sufficiently narrow width (less than $\omega$).

In the next section, we state and prove some lemmas and theorem associated with the sequential procedure described for constructing a bounded-width confidence interval for Pearson's correlation coefficient.

### 3.4 Lemmas and Theorems

**Lemma 3.1.** *Under the assumption that $E[\widehat{\xi}_{\rho n}^2]$ exist, for any $\omega > 0$, the stopping rule $N$ is finite, that is, $P(N < \infty) = 1$.*

*Proof.* Using Lemma A1 of Chattopadhyay and De (2016), we can prove this lemma. ■

**Lemma 3.2.** *If the parent distribution(s) is(are) such that* $E[\widehat{\xi}^2_{\rho n}]$ *exists, then the stopping rule in Equation* (3.15) *yields*

$$\frac{N}{n_\omega} \xrightarrow{P} 1 \text{ as } \omega \to 0, \tag{3.17}$$

*where* $\xrightarrow{P}$ *indicates convergence in probability.*

*Proof.* To prove this lemma, we proceed along the lines of Chattopadhyay and Kelley (2017) (see also De and Chattopadhyay, 2017). The definition of stopping rule $N$ in Equation (3.15) yields

$$\left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \widehat{\xi}^2_{\rho N} \leq N \leq mI(N=m) + \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \left(\widehat{\xi}^2_{\rho,N-1} + \frac{1}{N-1}\right). \tag{3.18}$$

Since $N \to \infty$ asymptotically as $\omega \downarrow 0$ and $\widehat{\xi}^2_{\rho n} \to \xi^2_\rho$ in probability as $n \to \infty$, by Theorem 2.1 of Gut (2009), $\widehat{\xi}^2_{\rho N} \to \xi^2_\rho$ in probability. Hence, dividing all sides of Equation (3.18) by $n_\omega$ and letting $\omega \downarrow 0$, we prove $N/n_\omega \to 1$ asymptotically as $\omega \downarrow 0$. ■

**Theorem 3.1.** *Suppose the parent distribution $F$ is such that* $E[U^2_{in}] < \infty$ *for* $i = 1, 2, 3$, *then the stopping rule in Equation* (3.15) *yields:*

$$\textit{Part 1: } P\left(r_N - \frac{z_{\alpha/2}\hat{\xi}_{\rho N}}{\sqrt{N}} < \rho < r_N + \frac{z_{\alpha/2}\hat{\xi}_{\rho N}}{\sqrt{N}}\right) \to 1-\alpha \quad as \quad \omega \to 0,$$

$$\textit{Part 2: } \frac{2z_{\alpha/2}\widehat{\xi}_{\rho N}}{\sqrt{N}} \leq \omega. \tag{3.19}$$

*Proof.* Part 1: We now proceed along the lines of De and Chattopadhyay (2017). Let $\mathbf{U}_n = [U_{1n}, U_{2n}, U_{3n}]'$ and $\theta = [\sigma_{XY}, \sigma^2_X, \sigma^2_Y]'$, then from Lee (1990), we know that

$$\mathbf{Y}_n = \sqrt{n}[\mathbf{U}_n - \theta] \xrightarrow{\mathcal{L}} N_3(\mathbf{0}, \mathbf{\Sigma}), \tag{3.20}$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} \mu_{22} - \sigma^2_{XY} & \mu_{31} - \sigma_{XY}\sigma^2_X & \mu_{13} - \sigma_{XY}\sigma^2_Y \\ \mu_{31} - \sigma_{XY}\sigma^2_X & \mu_{40} - \sigma^4_X & \mu_{22} - \sigma^2_X\sigma^2_Y \\ \mu_{13} - \sigma_{XY}\sigma^2_Y & \mu_{22} - \sigma^2_X\sigma^2_Y & \mu_{04} - \sigma^4_Y \end{bmatrix}.$$

We define $\mathbf{D}' = [a_1, a_2, a_3]$ and note that $\mathbf{D}'\mathbf{Y}_N = \mathbf{D}'\mathbf{Y}_{n_\omega} + (\mathbf{D}'\mathbf{Y}_N - \mathbf{D}'\mathbf{Y}_{n_\omega})$. To prove that $\mathbf{Y}_N \xrightarrow{\mathcal{L}} N_3(\mathbf{0}, \mathbf{\Sigma})$, we have to show that $\mathbf{D}'(\mathbf{Y}_N - \mathbf{Y}_{n_\omega}) \xrightarrow{P} 0$ as $\omega \to 0$. We write

$$\mathbf{D}'(\mathbf{Y}_N - \mathbf{Y}_{n_\omega}) = \sum_{i=1}^{3} a_i \sqrt{N} \left(U_{iN} - U_{in_\omega}\right) + \left(\sqrt{N/n_\omega} - 1\right) \mathbf{D}'\mathbf{Y}_{n_\omega}. \tag{3.21}$$

Let $n_1 = (1-\gamma)n_\omega$ and $n_2 = (1+\gamma)n_\omega$ for $\gamma \in (0,1)$. For a fixed $\varepsilon > 0$,

$$\begin{aligned} & P\left\{\left|\sum_{i=1}^{3} a_i \sqrt{N} \left(U_{iN} - U_{in_\omega}\right)\right| > \varepsilon\right\} \\ & \le P\left\{\left|\sum_{i=1}^{3} a_i \sqrt{N} \left(U_{iN} - U_{in_\omega}\right)\right| > \varepsilon, |N - n_\omega| < \gamma n_\omega\right\} + P\left\{|N - n_\omega| > \gamma n_\omega\right\} \qquad (3.22) \\ & \le \sum_{i=1}^{3} P\left\{\max_{n_1 < n < n_2} \sqrt{n} \left|U_{in} - U_{in_\omega}\right| > \frac{\varepsilon}{3|a_i|}\right\} + P\left\{|N - n_\omega| > \gamma n_\omega\right\}. \end{aligned}$$

Because $N/n_\omega \xrightarrow{P} 1$ and $U_{in}, i = 1, 2, 3$ are U-statistics which satisfy Anscombe's uniformly continuous in probability condition (see Anscombe, 1952, or Theorem 1.4), we conclude that $\sum_{i=1}^{3} a_i \sqrt{N} \left(U_{iN} - U_{in_\omega}\right) \xrightarrow{P} 0$. Also, $(\sqrt{N/n_\omega} - 1)\mathbf{D}'\mathbf{Y}_{n_\omega} \xrightarrow{P} 0$ as $\omega \to 0$ and $\mathbf{D}'\mathbf{Y}_{n_\omega} \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{D}'\mathbf{\Sigma}\mathbf{D})$. Hence, we conclude from Equation (3.21) that $\mathbf{D}'(\mathbf{Y}_N - \mathbf{Y}_{n_\omega}) \xrightarrow{P} 0$, that is, $\mathbf{Y}_N \xrightarrow{\mathcal{L}} N_3(\mathbf{0}, \mathbf{\Sigma})$. Now, we define $g(u_1, u_2, u_3) = \frac{u_1}{\sqrt{u_2 u_3}}$ for $u_2, u_3 > 0$ and rewrite $r_n = g(\mathbf{U}_n)$ using Taylor's expansion about $\theta$:

$$g(\mathbf{U}_n) = g(\theta) + \frac{U_{1N} - \sigma_{XY}}{\sigma_X \sigma_Y} - \frac{\sigma_{XY}}{2\sigma_X^3 \sigma_Y}(U_{2N} - \sigma_X^2) - \frac{\sigma_{XY}}{2\sigma_X \sigma_Y^3}(U_{3N} - \sigma_Y^2) + R_N, \tag{3.23}$$

where

$$R_N = \frac{1}{2}(\mathbf{U}_N - \theta)'\{D^2 g(\mathbf{a})\}(\mathbf{U}_N - \theta) \tag{3.24}$$

and $D^2 g(\mathbf{a})$ is the Hessian matrix of $g(\mathbf{U}_n)$ evaluated at $\mathbf{a} = (1-\gamma)\theta + \gamma\mathbf{U}_n$ for $\gamma \in (0,1)$. Thus,

$$\begin{aligned} \sqrt{N}(r_N - \rho) &= \sqrt{N}\frac{\rho}{2}\left(\frac{2}{\sigma_{XY}}(U_{1N} - \sigma_{XY}) - \frac{1}{\sigma_X^2}(U_{2N} - \sigma_X^2) - \frac{1}{\sigma_Y^2}(U_{3,N} - \sigma_3^2)\right) + \sqrt{N}R_N \\ &= \mathbf{D}'\mathbf{Y}_N + \sqrt{N}R_N, \qquad (3.25) \end{aligned}$$

where $\rho = g(\theta)$ and $\mathbf{D}' = \frac{\rho}{2}\left[\frac{2}{\sigma_{XY}}, -\frac{1}{\sigma_X^2}, -\frac{1}{\sigma_Y^2}\right]$. According to Lee (1990) and Anscombe's CLT (see Anscombe, 1952, or Theorem 1.4), $\sqrt{N}(U_{1N} - \sigma_{XY})$, $\sqrt{N}(U_{2N} - \sigma_X^2)$, and $\sqrt{N}(U_{3N} - \sigma_Y^2)$ converge to normal distributions and $(U_{1N} - \sigma_{XY})$, $(U_{2N} - \sigma_X^2)$, and $(U_{3N} - \sigma_Y^2)$ converge to 0 almost surely. This implies $\sqrt{N}R_N \xrightarrow{P} 0$ as $N \to \infty$. Hence, $\sqrt{N}(r_N - \rho) \xrightarrow{\mathcal{L}} N(0, \xi^2)$ as $\omega \to 0$, where

$$\xi^2 = \mathbf{D}'\mathbf{\Sigma}\mathbf{D} = \frac{\rho^2}{4}\left(\frac{\mu_{40}}{\sigma_X^4} + \frac{\mu_{04}}{\sigma_Y^4} + \frac{2\mu_{22}}{\sigma_X^2\sigma_Y^2} + \frac{4\mu_{22}}{\sigma_{XY}^2} - \frac{4\mu_{31}}{\sigma_{XY}\sigma_X^2} - \frac{4\mu_{13}}{\sigma_{XY}\sigma_Y^2}\right) \tag{3.26}$$

and $\mu_{ij} = \mathrm{E}\left[(X_1 - \mu_X)^i(Y_1 - \mu_Y)^j\right]$.

Part 2: We can prove by using Kelley et al. (2018) directly. ■

## 3.5 Characteristics of the Final Sample Size for Pearson's Product Moment Correlation: A Simulation Study

Recall that our procedure is asymptotically correct but its effectiveness in smaller sample size situations is not fully known, given the methods of confidence interval construction are themselves asymptotically correct. Correspondingly, we now demonstrate the properties of our method using a Monte Carlo simulation for constructing $100(1-\alpha)\%$ confidence intervals for population correlation coefficients from a variety of different bivariate distributions. To implement the sequential AIPE procedure, we specify a maximum confidence interval width of $\omega = 0.1$ (say) and a confidence coefficient of 90%. We compute the pilot sample size by using the formula given in the algorithm $m = 33$ $\left(= \max\left\{4, \lceil 2z_{0.1/2}/0.1\rceil\right\}\right)$. The estimate of the asymptotic variance of correlation coefficient is calculated using the pilot sample, and we check if the stopping rule in Equation (3.15) is met. If the stopping rule is met, the sampling stops. Otherwise, an additional sample is generated and asymptotic variance recalculated. This continues until the stopping rule in Equation (3.15) is met. The simulation results are based on two different distributions: bivariate normal and bivariate gamma distributions. For bivariate normal and the bivariate gamma distribution from Theorem 2 of Nadarajah and Gupta (2006), the simulation study was done for correlation coefficient $\rho$ of 0.1, 0.3, and 0.5, $\omega = 0.1, 0.2$ and $\alpha = 0.1, 0.05$. In all cases, 5,000 replications were used.

Table 3.1. Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\rho$

| $\omega$ | Distribution | $\rho$ | $\bar{N}$ | $se(\bar{N})$ | $n_\omega$ | $\bar{N}/n_\omega$ | $p$ | $s_p$ | $\bar{w}_N$ | $se(\bar{w}_N)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\alpha = 0.10$ | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 1056.0 | 0.9581 | 1061 | 0.9957 | 0.8944 | 0.0043 | 0.0999 | $9.33 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 891.2 | 0.9900 | 897 | 0.9935 | 0.8902 | 0.0044 | 0.0999 | $1.21 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 601.0 | 0.9920 | 609 | 0.9869 | 0.8868 | 0.0045 | 0.0997 | $2.04 \times 10^{-6}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 1124.0 | 1.5480 | 1138 | 0.9876 | 0.8984 | 0.0043 | 0.0999 | $1.66 \times 10^{-6}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 1049.0 | 1.6510 | 1066 | 0.9840 | 0.8982 | 0.0043 | 0.0999 | $1.97 \times 10^{-6}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 774.4 | 1.6730 | 797 | 0.9717 | 0.8862 | 0.0045 | 0.0995 | $7.24 \times 10^{-5}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 260.3 | 0.5085 | 266 | 0.9784 | 0.8762 | 0.0047 | 0.1985 | $1.48 \times 10^{-4}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 216.4 | 0.6056 | 225 | 0.9617 | 0.8666 | 0.0048 | 0.1963 | $2.86 \times 10^{-4}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 138.0 | 0.6628 | 153 | 0.9022 | 0.8124 | 0.0055 | 0.1873 | $5.60 \times 10^{-4}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 272.0 | 0.7694 | 285 | 0.9545 | 0.8784 | 0.0046 | 0.1978 | $2.07 \times 10^{-4}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 245.9 | 0.9045 | 267 | 0.9209 | 0.8546 | 0.0050 | 0.1947 | $3.69 \times 10^{-4}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 164.0 | 1.0100 | 200 | 0.8198 | 0.7686 | 0.0060 | 0.1806 | $7.25 \times 10^{-4}$ |
| | | | | | $\alpha = 0.05$ | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 1502.0 | 1.1200 | 1507 | 0.9969 | 0.9442 | 0.0032 | 0.0999 | $6.44 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 1267.2 | 1.1890 | 1273 | 0.9956 | 0.9460 | 0.0032 | 0.0999 | $8.13 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 857.0 | 1.1920 | 865 | 0.9908 | 0.9464 | 0.0032 | 0.0998 | $7.91 \times 10^{-6}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 1600.0 | 1.8780 | 1615 | 0.9910 | 0.9498 | 0.0031 | 0.0999 | $1.23 \times 10^{-7}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 1497.0 | 1.9960 | 1513 | 0.9840 | 0.9490 | 0.0031 | 0.0999 | $1.24 \times 10^{-6}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 1109.0 | 1.9540 | 1132 | 0.9794 | 0.9430 | 0.0033 | 0.0998 | $3.90 \times 10^{-5}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 372.2 | 0.5746 | 377 | 0.9872 | 0.9396 | 0.0034 | 0.1992 | $8.16 \times 10^{-5}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 312.5 | 0.6112 | 319 | 0.9796 | 0.9332 | 0.0035 | 0.1989 | $9.71 \times 10^{-5}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 204.0 | 0.7435 | 217 | 0.9400 | 0.9012 | 0.0042 | 0.1938 | $3.83 \times 10^{-4}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 391.1 | 0.8966 | 404 | 0.9681 | 0.9412 | 0.0033 | 0.1989 | $1.15 \times 10^{-4}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 360.1 | 1.0160 | 379 | 0.9501 | 0.9252 | 0.0037 | 0.1977 | $2.30 \times 10^{-4}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 251.6 | 1.1640 | 283 | 0.8889 | 0.8766 | 0.0047 | 0.1901 | $5.30 \times 10^{-4}$ |

Note: $\rho$ is the population correlation coefficient; $\bar{N}$ is the mean final sample size; $p$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $\rho$; $se(\bar{N})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_\omega$ is the theoretical sample size if the procedure is used with the population parameters; $se(p)$ is the standard error of $p$; $\bar{w}_N$ average length of confidence intervals for $\rho$ based on $N$ observations; $se(\bar{w}_N)$ is the standard error of $\bar{w}$; tabled values are based on 5,000 replications of a Monte Carlo simulation study from distributions: Bivariate Normal ($N_2$) distribution with parameters $\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$ and $\rho$, respectively, and Bivariate Gamma ($Ga_2$) with parameters $a_1$, $a_2$, $c$, and $\mu$, respectively, based on Theorem 2 of Nadarajah and Gupta (2006).

Table 3.1 shows the mean final sample size $\bar{N}$ (estimates E$[N]$), coverage probability $p$, and average confidence interval width $\bar{w}_N$ (estimates E$[w_n]$). The values $se(\bar{N}), se(p)$, and $se(\bar{w}_N)$ represent the standard errors of $\bar{N}$, $p$, and $\bar{w}_N$ respectively. None of the confidence interval widths, $w_N$, obtained from the final sample sizes, $N$, exceeded the maximum specified width, $\omega$. The table also shows that, in most cases, the ratio of the mean final sample size to the theoretical sample sizes is satisfactory, if not highly so. However, in some situations the the ratio of the mean final sample size to the theoretical sample sizes is not on target (e.g., $< 85\%$ empirical coverage in the situation of a 90% confidence interval). These situations, however, occur only when the empirical confidence interval coverage differs markedly from the nominal coverage. We will discuss this limitation below, which is due to the confidence interval procedure and not the sequential AIPE procedure.

Table 3.1 shows that, in most situations, our sequential procedure works well. However, there are some situations where (a) the ratio of the mean final sample size to the theoretical sample size (i.e., $\bar{N}/n_\omega$) is considerably less than 1.0, such as in Table 3.1 in the final column. There, the ratio is 0.82. In particular, the mean final sample size was 164 but the theoretical sample size is 200. However, also note that the confidence interval coverage, nominally set to 90%, was shown to be only 76.86%. Consideration of this issue led to a separate simulation study to evaluate if the problem was with (a) the sequential procedure or (b) the confidence interval method itself. We conducted another Monte Carlo simulation study using a fixed-$n$ approach at the theoretical sample size. Because the only thing manipulated was the method, fixed-$n$ or sequential, any differences would be due to that. However, if the confidence interval coverage performance is the same, then it is a failure of the confidence interval procedure itself, not the sequential approach.

Our results (see Table 3.2), based on 10,000 replications, show that the empirical confidence interval coverage is too low for some of the situations, particularly for the smaller sample sizes (e.g., 90% coverage and larger values of $\rho$). We believe that the shortcoming in

Table 3.2. Simulation results for $100(1-\alpha)\%$ confidence interval for $\rho$ for fixed $n$

| | | | $\alpha = 0.10$ | | $\alpha = 0.05$ | |
|---|---|---|---|---|---|---|
| $\omega$ | Distribution | $\rho$ | $n_\omega$ | $p$ | $n_\omega$ | $p$ |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 1061 | 0.8998 | 1507 | 0.9477 |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 897 | 0.9013 | 1273 | 0.9528 |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 609 | 0.8941 | 865 | 0.9466 |
| | $Ga_2(5,5,50,10)$ | 0.1 | 1138 | 0.8991 | 1615 | 0.9505 |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 1066 | 0.8942 | 1513 | 0.9447 |
| | $Ga_2(5,5,10,10)$ | 0.5 | 797 | 0.8951 | 1132 | 0.9456 |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 266 | 0.8924 | 377 | 0.9454 |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 225 | 0.8985 | 319 | 0.9446 |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 153 | 0.8894 | 217 | 0.9392 |
| | $Ga_2(5,5,50,10)$ | 0.1 | 285 | 0.8899 | 404 | 0.9470 |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 267 | 0.8881 | 379 | 0.9424 |
| | $Ga_2(5,5,10,10)$ | 0.5 | 200 | 0.8840 | 283 | 0.9381 |

Note: $\omega$ is the upper bound of the desired confidence interval length; $\rho$ is the population correlation coefficient; $n_\omega$ is the optimal sample size given $\omega$, $\rho$ and the specified distribution; $p$ is the simulated coverage probability for a fixed sample size of $n_\omega$ given $\rho$ and the distribution.

the confidence interval coverage is due to the bias of the estimator of the kurtosis parameter (namely, $\mu_{40}/\sigma_X^4$) in $\xi^2$ given in Equation (3.7). This biased estimator of the kurtosis that is used in this article, along with five other estimators, was studied by An and Ahmed (2008). In order to get better results, a robust consistent estimator for kurtosis parameters may be used, but this is an active area of research and we are limited by what already exists in the literature.

## 3.6 Alternative Confidence Intervals for Pearson's Correlation Coefficient

Our proposed sequential procedure can be extended to other forms of confidence intervals for the population correlation coefficient, such as those proposed by Corty and Corty (2011) and Moinester and Gottfried (2014). We discuss how our methods apply to the methods recommended by these authors. We are agnostic to which method should be used, but rather want to show how our methods work for both situations.

### 3.6.1 Confidence Interval by Corty and Corty (2011)

Corty and Corty (2011) used Fisher's $z$-transform and thereby proposed a way to estimate the sample size for a given choice of sample correlation coefficient, confidence level, and $\omega$.

Moinester and Gottfried (2014) noted that the optimal sample size required to achieve a $100(1-\alpha)\%$ confidence interval for correlation coefficient ($\rho$), proposed by Corty and Corty (2011), with width no larger than $\omega$ is

$$n_{CC} = \left(\frac{4z_{\alpha/2}}{\ln(B)}\right)^2 + 3, \tag{3.27}$$

where

$$B = \frac{(1+|\rho|+\omega/2)(1-|\rho|+\omega/2)}{(1+|\rho|-\omega/2)(1-|\rho|-\omega/2)}, \tag{3.28}$$

and $|\rho|$ being the absolute value of the population correlation coefficient. Now the supposed value of the population correlation coefficient, whose confidence interval we would like to construct, can differ markedly from the true population value. As discussed, our sequential procedure does not require inserting supposed population values. Our sequential stopping rule which helps find the estimate of the optimal sample size is as follows:

$$\begin{aligned} &N_{CC} \text{ is the smallest integer } n(\geq m_{CC}) \text{ such that} \\ &n \geq 16z_{\alpha/2}^2\left[\left(\ln\left(\frac{(1+|r_n|+\omega/2)(1-|r_n|+\omega/2)}{(1+|r_n|-\omega/2)(1-|r_n|-\omega/2)}\right)\right)^{-2} + \frac{1}{n}\right] + 3. \end{aligned} \tag{3.29}$$

Following the sample of size $N_{CC}$, collected using the sequential stopping rule stated in Equation (3.29), the $100(1-\alpha)\%$ confidence interval for the population correlation coefficient, $\rho$, can be constructed by applying the confidence interval formula as in Corty and Corty (2011). We suggest the pilot sample size, $m_{CC}$, as

$$m_{CC} = \max\left\{4, \left\lceil \frac{1}{2}\left[3 + \frac{16z_{\alpha/2}^2}{(\ln b)^2} + \sqrt{\left(3 + \frac{16z_{\alpha/2}^2}{(\ln b)^2}\right)^2 + \left(8z_{\alpha/2}\right)^2}\right]\right\rceil\right\} \tag{3.30}$$

where

$$b = \frac{\left(2+\frac{\omega}{2}\right)\left(1+\frac{\omega}{2}\right)}{\left(1-\frac{\omega}{2}\right)\left(\frac{1-2\omega}{4}\right)} \tag{3.31}$$

for $r_n < 1 - \frac{\omega}{2}$ and $\omega < 0.5$.

The optimal sample size, $n_{CC}$, can be estimated by adopting the sequential stopping rule defined in Equation (3.29). Note that, in practice, $n_{CC}$ is usually unknown because $\rho$ will usually be unknown. Thus, when one uses supposed values of parameters the final sample size is known but is based on a value that is almost certainly not true. The derivation of pilot sample size formula in Equation (3.30) is shown in the next subsection.

### Derivation of Pilot Sample Size Confidence Interval by Corty and Corty (2011)

For $r_n < 1 - \frac{\omega}{2}$ and $\omega < 0.5$,

$$B = \frac{\left(1+|r_n|+\frac{\omega}{2}\right)\left(1-|r_n|+\frac{\omega}{2}\right)}{\left(1+|r_n|-\frac{\omega}{2}\right)\left(1-|r_n|-\frac{\omega}{2}\right)} < \frac{\left(2+\frac{\omega}{2}\right)\left(1+\frac{\omega}{2}\right)}{\left(1-\frac{\omega}{2}\right)\left(\frac{1-2\omega}{4}\right)} = b$$

$$\implies \ln B < \ln b \quad \implies \quad \frac{1}{\ln B} > \frac{1}{\ln b}. \tag{3.32}$$

From the stopping rule in Equation (3.29) and using Equation (3.32), we have

$$\begin{aligned} n &\geq 16z_{\alpha/2}^2\left(\frac{1}{(\ln B)^2}+\frac{1}{n}\right)+3 \geq 16z_{\alpha/2}^2\left(\frac{1}{(\ln b)^2}+\frac{1}{n}\right)+3 \\ n^2 &\geq \left(\frac{16z_{\alpha/2}^2}{(\ln b)^2}+3\right)n + 16z_{\alpha/2}^2 \\ n^2 &- \left(\frac{16z_{\alpha/2}^2}{(\ln b)^2}+3\right)n - 16z_{\alpha/2}^2 \geq 0. \end{aligned} \tag{3.33}$$

Using quadratic formula and $n > 0$, we have

$$n \geq \frac{1}{2}\left[3+\frac{16z_{\alpha/2}^2}{(\ln b)^2}+\sqrt{\left(3+\frac{16z_{\alpha/2}^2}{(\ln b)^2}\right)^2+\left(8z_{\alpha/2}\right)^2}\right] \tag{3.34}$$

Thus, the pilot sample size $m_{CC}$ is

$$m_{CC} = \max\left\{4, \left\lceil\frac{1}{2}\left[3+\frac{16z_{\alpha/2}^2}{(\ln b)^2}+\sqrt{\left(3+\frac{16z_{\alpha/2}^2}{(\ln b)^2}\right)^2+\left(8z_{\alpha/2}\right)^2}\right]\right\rceil\right\}. \tag{3.35}$$

### 3.6.2 Confidence Interval in Method 4 by Moinester and Gottfried (2014)

In method 4 of Moinester and Gottfried (2014), the $100(1-\alpha)\%$ confidence interval for the population correlation coefficient, when observations are assumed to be from a bivariate-normal distribution, is

$$\left(r_n - z_{\alpha/2}\frac{1-r_n^2}{\sqrt{n-1}}, r_n + z_{\alpha/2}\frac{1-r_n^2}{\sqrt{n-1}}\right). \tag{3.36}$$

The optimal sample size required to achieve a $100(1-\alpha)\%$ confidence interval for correlation coefficient ($\rho$) with width no larger than $\omega$ is

$$n_{MG} = \left\lceil \frac{2z_{\alpha/2}(1-\rho^2)}{\omega} \right\rceil^2 + 1. \tag{3.37}$$

Our sequential stopping rule, which does not take into account the supposed value of the population correlation coefficient, is as follows:

$$N_{MG} \text{ is the smallest integer } n(\geq m_{MG}) \text{ such that } n \geq \frac{4z_{\alpha/2}}{\omega^2}\left[\left(1-r_n^2\right)^2 + \frac{1}{n}\right] + 1. \tag{3.38}$$

Following the sample of size $N_{MG}$ collected using the sequential stopping rule of Equation (3.38), the $100(1-\alpha)\%$ confidence interval for $\rho$ is

$$\left(r_{N_{MG}} - z_{\alpha/2}\frac{1-r_{N_{MG}}^2}{\sqrt{N_{MG}-1}}, r_{N_{MG}} + z_{\alpha/2}\frac{1-r_{N_{MG}}^2}{\sqrt{N_{MG}-1}}\right). \tag{3.39}$$

We suggest the pilot sample size, $m_{CC}$, of

$$m_{MG} = \max\left\{4, \left\lceil \frac{1+\sqrt{1+\left(4z_{\alpha/2}/\omega\right)^2}}{2} \right\rceil\right\}. \tag{3.40}$$

The optimal sample size, $m_{MG}$, can be estimated by following the sequential stopping rule defined in Equation (3.38). The next subsection shows derivation of the pilot sample size given in Equation (3.30).

**Derivation of Pilot Sample Size for Confidence Interval in Method 4 by Moinester and Gottfried (2014)**

From the stopping rule defined in Equation (3.38), we have

$$\begin{aligned} n &\geq \frac{4z_{\alpha/2}^2}{\omega^2}\left[\left(1-r_n^2\right)^2+\frac{1}{n}\right]+1 \geq \frac{4z_{\alpha/2}^2}{n\omega^2}+1 \\ n^2-n-\frac{4z_{\alpha/2}^2}{\omega^2} &\geq 0. \end{aligned} \tag{3.41}$$

Solving for $n > 0$, we have

$$n \geq \frac{1+\sqrt{1+\left(4z_{\alpha/2}/\omega\right)^2}}{2} \tag{3.42}$$

The pilot sample size $m_{MG}$ is therefore defined as

$$m_{MG} = \max\left\{4, \left\lceil \frac{1+\sqrt{1+\left(4z_{\alpha/2}/\omega\right)^2}}{2} \right\rceil\right\}. \tag{3.43}$$

### 3.6.3 Simulation Study

We now compare the characteristics of the stopping rule defined in Equation (3.15) with the stopping rules defined in Equations (3.29) and (3.38) using a Monte Carlo simulation study for constructing $100(1-\alpha)\%$ confidence intervals for population correlation coefficients from bivariate distributions. For bivariate normal and the bivariate gamma distribution from Theorem 2 of Nadarajah and Gupta (2006), the simulation study was done for correlation coefficient $\rho$ of 0.1, 0.3, and 0.5 and $\omega = 0.1, 0.2$. In both cases, 5,000 replications were used. Tables 3.3 and 3.4 show the estimates of mean final sample size, coverage probability, and average confidence interval width and also the corresponding standard errors.

Comparing the characteristics of the stopping rule defined in Equation (3.15) with the stopping rules defined in Equations (3.29) and (3.38), we observe that the behavior of the coverage probability as well as ratio of average sample size estimate and the optimal sample size are almost similar in all three procedures.

Table 3.3. Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\rho$ using Corty and Corty (2011)

| $\omega$ | Distribution | $\rho$ | $\bar{N}_{CC}$ | $se(\bar{N}_{CC})$ | $n_{CC}$ | $\bar{N}_{CC}/n_{CC}$ | $p_{CC}$ | $se(p_{CC})$ | $\bar{w}_{N_{CC}}$ | $se(\bar{w}_{N_{CC}})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\alpha = 0.10$ | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 1060.0 | 0.1942 | 1062 | 0.9979 | 0.8980 | 0.0043 | 0.1000 | $2.75 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 893.6 | 0.5208 | 897 | 0.9962 | 0.8986 | 0.0043 | 0.1000 | $7.71 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 604.4 | 0.7567 | 609 | 0.9924 | 0.8868 | 0.0045 | 0.1001 | $1.32 \times 10^{-5}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 1060.0 | 0.1966 | 1062 | 0.9982 | 0.8906 | 0.0044 | 0.1000 | $3.03 \times 10^{-7}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 894.5 | 0.5549 | 897 | 0.9972 | 0.8728 | 0.0047 | 0.1000 | $1.19 \times 10^{-6}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 604.5 | 0.8638 | 609 | 0.9925 | 0.8464 | 0.0051 | 0.1001 | $1.52 \times 10^{-5}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 264.6 | 0.1173 | 267 | 0.9909 | 0.8892 | 0.0044 | 0.1999 | $1.16 \times 10^{-5}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 222.1 | 0.2994 | 225 | 0.9872 | 0.8916 | 0.0044 | 0.2002 | $3.70 \times 10^{-5}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 146.1 | 0.4565 | 153 | 0.9551 | 0.8692 | 0.0048 | 0.2021 | $1.14 \times 10^{-4}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 264.1 | 0.1665 | 267 | 0.9892 | 0.8842 | 0.0045 | 0.1999 | $3.10 \times 10^{-5}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 221.8 | 0.3441 | 225 | 0.9859 | 0.8606 | 0.0049 | 0.2002 | $5.22 \times 10^{-5}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 143.1 | 0.5639 | 153 | 0.9351 | 0.7942 | 0.0057 | 0.2027 | $1.54 \times 10^{-4}$ |
| | | | | | $\alpha = 0.05$ | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 1504.0 | 0.2268 | 1507 | 0.9983 | 0.9474 | 0.0032 | 0.1000 | $1.92 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 1269.0 | 0.6185 | 1273 | 0.9971 | 0.9462 | 0.0032 | 0.1000 | $5.64 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 857.6 | 0.8464 | 863 | 0.9938 | 0.9452 | 0.0032 | 0.1001 | $1.43 \times 10^{-6}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 1505.0 | 0.2318 | 1507 | 0.9985 | 0.9400 | 0.0034 | 0.1000 | $2.10 \times 10^{-7}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 1270.0 | 0.6581 | 1273 | 0.9975 | 0.9354 | 0.0035 | 0.1000 | $8.25 \times 10^{-7}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 859.1 | 0.9800 | 863 | 0.9955 | 0.9090 | 0.0041 | 0.1001 | $9.37 \times 10^{-6}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 375.2 | 0.1206 | 377 | 0.9952 | 0.9478 | 0.0031 | 0.1999 | $1.69 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 315.6 | 0.3343 | 318 | 0.9925 | 0.9444 | 0.0032 | 0.2002 | $2.81 \times 10^{-5}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 210.1 | 0.4924 | 215 | 0.9774 | 0.9400 | 0.0034 | 0.2017 | $7.54 \times 10^{-5}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 375.1 | 0.1250 | 377 | 0.9950 | 0.9418 | 0.0033 | 0.1999 | $1.93 \times 10^{-6}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 315.7 | 0.3694 | 318 | 0.9926 | 0.9248 | 0.0037 | 0.2002 | $2.60 \times 10^{-5}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 208.2 | 0.5970 | 215 | 0.9686 | 0.8866 | 0.0045 | 0.2019 | $9.94 \times 10^{-5}$ |

Note: $\rho$ is the population correlation coefficient; $\bar{N}_{CC}$ is the mean final sample size; $p_{CC}$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $\rho$; $se(\bar{N}_{CC})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_{CC}$ is the theoretical sample size if the procedure is used with the population parameters; $se(p_{CC})$ is the standard error of $p_{CC}$; $\bar{w}_{N_{CC}}$ average length of confidence intervals for $\rho$ based on $N_{CC}$ observations; $se(\bar{w}_{N_{CC}})$ is the standard error of $\bar{w}_{N_{CC}}$; tabled values are based on 5,000 replications of a Monte Carlo simulation study from distributions: Bivariate Normal ($N_2$) distribution with parameters $\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$ and $\rho$, respectively, and Bivariate Gamma ($Ga_2$) with parameters $a_1$, $a_2$, $c$, and $\mu$, respectively, based on Theorem 2 of Nadarajah and Gupta (2006).

## 3.7 Sequential AIPE for Kendall's Tau and Spearman's Rho

We now discuss the sequential approach related to the accuracy in parameter estimation problem for estimating Kendall's rank correlation coefficient, popularly known as Kendall's

Table 3.4. Summary of final sample size for $100(1-\alpha)\%$ confidence interval for $\rho$ using Moinester and Gottfried (2014)

| $\omega$ | Distribution | $\rho$ | $\bar{N}_{MG}$ | $se(\bar{N}_{MG})$ | $n_{MG}$ | $\bar{N}_{MG}/n_{MG}$ | $p_{MG}$ | $se(p_{MG})$ | $\bar{w}_{N_{MG}}$ | $se(\bar{w}_{N_{MG}})$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\alpha = 0.10$ | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 1060.0 | 0.1936 | 1062 | 0.9978 | 0.8944 | 0.0043 | 0.1000 | $2.74 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 893.9 | 0.5182 | 898 | 0.9955 | 0.8990 | 0.0043 | 0.0999 | $7.62 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 606.1 | 0.7103 | 610 | 0.9935 | 0.8932 | 0.0044 | 0.0999 | $1.90 \times 10^{-6}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 264.5 | 0.1032 | 267 | 0.9907 | 0.8868 | 0.0045 | 0.1998 | $2.37 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 222.7 | 0.2779 | 226 | 0.9853 | 0.8848 | 0.0045 | 0.1995 | $8.62 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 149.1 | 0.4043 | 154 | 0.9681 | 0.8642 | 0.0048 | 0.1989 | $2.21 \times 10^{-5}$ |
| | | | | | $\alpha = 0.05$ | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 1505.0 | 0.2233 | 1508 | 0.9982 | 0.9498 | 0.0032 | 0.1000 | $1.94 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 1270.0 | 0.6142 | 1274 | 0.9970 | 0.9452 | 0.0032 | 0.1000 | $5.13 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 860.5 | 0.8382 | 866 | 0.9937 | 0.9454 | 0.0032 | 0.0999 | $1.33 \times 10^{-6}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 375.8 | 0.1204 | 378 | 0.9942 | 0.9440 | 0.0033 | 0.1998 | $1.65 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 317.1 | 0.3343 | 320 | 0.9908 | 0.9452 | 0.0032 | 0.1997 | $4.26 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 213.9 | 0.4407 | 218 | 0.9812 | 0.9376 | 0.0034 | 0.1992 | $2.66 \times 10^{-5}$ |

Note: $\rho$ is the population correlation coefficient; $\bar{N}_{MG}$ is the mean final sample size; $p_{MG}$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $\rho$; $se(\bar{N}_{MG})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_{MG}$ is the theoretical sample size if the procedure is used with the population parameters; $se(p_{MG})$ is the standard error of $p_{MG}$; $\bar{w}_{N_{MG}}$ average length of confidence intervals for $\rho$ based on $N_{MG}$ observations; $se(\bar{w}_{N_{MG}})$ is the standard error of $\bar{w}_{N_{MG}}$; tabled values are based on 5,000 replications of a Monte Carlo simulation study from Bivariate Normal distribution ($N_2$) with parameters: means, variances and correlation.

tau and denoted here by $\tau$, and Spearman's rank correlation coefficient, popularly known as Spearman's $\rho$ and denoted here by $\rho_s$.

### 3.7.1 AIPE for Kendall's $\tau$

Kendall's $\tau$ is a statistic which can be used to measure the ordinal association between two variables. Suppose $(X, Y)$ denote a pair of random observations with a joint distribution function $F$. If $(X_1, Y_1)$ and $(X_2, Y_2)$ are random observations from $F$, then Kendall's tau which measures the association between variables $X$ and $Y$ can be defined as

$$\tau = \mathrm{E}\left[\mathrm{sgn}(X_1 - X_2)\mathrm{sgn}(Y_1 - Y_2)\right] \tag{3.44}$$

where

$$\operatorname{sgn}(x) = \begin{cases} -1, & \text{when } x < 0, \\ \ \ 0 & \text{when } x = 0, \\ +1 & \text{when } x > 0, \end{cases}$$

An estimator of Kendall's $\tau$ is given by

$$r_{\tau,n} = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \operatorname{sgn}(X_i - X_j)\operatorname{sgn}(Y_i - Y_j) \tag{3.45}$$

which is a U-statistics (see Lee, 1990). Hoeffding (1948) as well as Daniels and Kendall (1947) have shown that the asymptotic distribution of $\tau$, defined in Equation (3.44), is given by

$$\sqrt{n}\left(r_{\tau,n} - \tau\right) \xrightarrow{D} N\left(0, \xi_\tau^2\right), \tag{3.46}$$

where the expression of the asymptotic variance, $\xi_\tau^2$, is given by

$$\begin{aligned} \xi_\tau^2 &= 4\text{Var}\left\{\text{E}\left[\operatorname{sgn}\left(X_1 - X_2\right)\operatorname{sgn}\left(Y_1 - Y_2\right) \middle| X_1, Y_1\right]\right\} \\ &= 4\text{Var}\left\{1 - 2F_1(X_1) - 2F_2(Y_1) + 4F(X_1, Y_1)\right\}, \end{aligned} \tag{3.47}$$

provided $F_1$ and $F_2$ are the marginal distributions of $X$ and $Y$ respectively. Proceeding along the same lines as in Equations (3.8)–(3.11), we can find that the sample size required to achieve the sufficient accuracy with pre-specified upper bound ($\omega$) on the width of the confidence interval for $\tau$ will be

$$n \ge \left\lceil \frac{4z_{\alpha/2}^2 \xi_\tau^2}{\omega^2} \right\rceil \equiv n_{KT}, \tag{3.48}$$

where $\xi_\tau^2$ is defined as in Equation (3.47). In reality, $\xi_\tau^2$ is unknown, so we use a consistent estimator, which is given by

$$\hat{\xi}_{n,KT}^2 = \frac{16}{n-1} \sum_{i=1}^{n} (W_i - \bar{W})^2, \tag{3.49}$$

where

$$W_i = \frac{2}{n}\sum_{k=1}^{n} 1\left\{R_{x,k} \leq R_{x,i}, R_{y,k} \leq R_{y,i}\right\} - \frac{R_{x,i}}{n+1} - \frac{R_{y,i}}{n+1} \tag{3.50}$$

with $1\{A\}$ denoting the indicator function of set $A$ (see Kojadinovic and Yan, 2010; Genest and Favre, 2007). Because $\xi_\tau^2$ is unknown in reality, in order to compute the required sample size, $n_{KT}$, we use the sequential procedure outlined. Our sequential stopping rule which helps find the estimate of the optimal sample size is as follows:

$$N_{KT} \text{ is the smallest integer } n(\geq m_{KT}) \text{ such that } \frac{4z_{\alpha/2}^2}{\omega^2}\left(\widehat{\xi}_{n,KT}^2 + n^{-1}\right) \tag{3.51}$$

where $m_{KT}$ is the pilot sample, the same as that given in Equation (3.14).

We now find the characteristics of the stopping rule defined in Equation 3.51 using Monte Carlo simulation for constructing $100(1-\alpha)\%$ confidence intervals for population correlation coefficients from bivariate distributions – bivariate normal and the bivariate gamma distribution from Theorem 2 of Nadarajah and Gupta (2006). The simulation study was done for correlation coefficient $\tau$ corresponding of 0.1, 0.3, and 0.5 and $\omega = 0.1, 0.2$. In both cases, 5,000 replications were used. Table 3.5 shows the estimates of mean final sample size, coverage probability, and average confidence interval width and also the corresponding standard errors for 90% and 95% confidence interval coverage, respectively.

The width of the confidence interval given by the sequential procedure with stopping rule defined in Equation (3.51) did not exceed the maximum specified width $\omega$. Further, the coverage probability estimates are close to the corresponding confidence level. Also, the ratio of average sample size estimate and the optimal sample size is close to 1.

### 3.7.2 AIPE for Spearman's $\rho$

Let $(X, Y)$ be a random observation with common distribution function $F$ with marginals $F_1(x)$ and $F_2(y)$ respectively for $X$ and $Y$. The popular nonparametric correlation measure

Table 3.5. Summary of final sample size for $100(1-\alpha)\%$ confidence interval for Kendall's $\tau$ using asymptotic distribution

| $\omega$ | Distribution | $\rho$ | $\tau$ | $\bar{N}_{KT}$ | $se(\bar{N}_{KT})$ | $n_{KT}$ | $\bar{N}_{KT}/n_{KT}$ | $p_{KT}$ | $se(p_{KT})$ | $\bar{w}_{N_{KT}}$ | $se(\bar{w}_{N_{KT}})$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $\alpha = 0.10$ | | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 0.0638 | 478.4 | 0.0597 | 477 | 1.003 | 0.8946 | 0.0043 | 0.0997 | $4.815 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 0.1940 | 443.0 | 0.1669 | 442 | 1.002 | 0.8888 | 0.0044 | 0.0997 | $8.224 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 0.3333 | 371.1 | 0.2575 | 369 | 1.006 | 0.8876 | 0.0045 | 0.0995 | $1.562 \times 10^{-6}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 0.0638 | 478.4 | 0.0626 | 477 | 1.003 | 0.8984 | 0.0043 | 0.0997 | $4.901 \times 10^{-7}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 0.1940 | 443.1 | 0.1783 | 442 | 1.002 | 0.8952 | 0.0043 | 0.0996 | $1.022 \times 10^{-6}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 0.3330 | 370.9 | 0.2881 | 369 | 1.005 | 0.8930 | 0.0044 | 0.0995 | $2.200 \times 10^{-6}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 0.0638 | 120.9 | 0.0354 | 120 | 1.008 | 0.8824 | 0.0046 | 0.1977 | $3.834 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 0.1904 | 112.2 | 0.0846 | 111 | 1.011 | 0.8840 | 0.0045 | 0.1972 | $7.805 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 0.3333 | 94.5 | 0.1250 | 93 | 1.016 | 0.8862 | 0.0045 | 0.1960 | $1.615 \times 10^{-5}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 0.0638 | 120.9 | 0.0364 | 120 | 1.008 | 0.8858 | 0.0045 | 0.1977 | $4.113 \times 10^{-6}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 0.1940 | 112.0 | 0.0915 | 111 | 1.009 | 0.8866 | 0.0045 | 0.1972 | $1.037 \times 10^{-5}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 0.3333 | 94.2 | 0.1457 | 93 | 1.012 | 0.8756 | 0.0047 | 0.1958 | $2.273 \times 10^{-5}$ |
| | | | | | $\alpha = 0.05$ | | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 0.0638 | 678.4 | 0.0698 | 677 | 1.002 | 0.9506 | 0.0031 | 0.0998 | $3.429 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 0.1940 | 628.2 | 0.1984 | 627 | 1.002 | 0.9408 | 0.0033 | 0.0998 | $5.691 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 0.3333 | 526.0 | 0.3058 | 524 | 1.004 | 0.9444 | 0.0032 | 0.0996 | $1.084 \times 10^{-6}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 0.6380 | 678.4 | 0.0741 | 677 | 1.002 | 0.9490 | 0.0031 | 0.0998 | $3.410 \times 10^{-7}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 0.1940 | 628.2 | 0.2084 | 627 | 1.002 | 0.9512 | 0.0030 | 0.0998 | $7.487 \times 10^{-7}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 0.3333 | 526.1 | 0.3452 | 524 | 1.004 | 0.9394 | 0.0034 | 0.0996 | $1.539 \times 10^{-6}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 0.0638 | 170.9 | 0.0371 | 170 | 1.005 | 0.9440 | 0.0033 | 0.1984 | $3.028 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 0.1940 | 158.2 | 0.1004 | 157 | 1.010 | 0.9378 | 0.0034 | 0.1980 | $5.218 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 0.3333 | 133.2 | 0.1514 | 131 | 1.017 | 0.9318 | 0.0036 | 0.1972 | $1.045 \times 10^{-5}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 0.6380 | 170.8 | 0.0425 | 170 | 1.005 | 0.9428 | 0.0033 | 0.1984 | $3.335 \times 10^{-6}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 0.1940 | 158.3 | 0.1121 | 157 | 1.009 | 0.9364 | 0.0035 | 0.1980 | $6.536 \times 10^{-6}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 0.3330 | 133.1 | 0.1725 | 131 | 1.016 | 0.9332 | 0.0035 | 0.1971 | $1.470 \times 10^{-5}$ |

Note: $\rho$ is the population Pearson's correlation coefficient; $\tau$ is the population Kendall's $\tau$ (computed using bootstrap method for bivariate Gamma distribution); $\bar{N}_{KT}$ is the mean final sample size; $p_{KT}$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $\tau$; $se(\bar{N}_{KT})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_{KT}$ is the theoretical sample size if the procedure is used with the population parameters; $se(p_{KT})$ is the standard error of $p_{KT}$; $\bar{w}_{N_{KT}}$ average length of confidence intervals for $\rho$ based on $N$ observations; tabled values are based on 5,000 replications of a Monte Carlo simulation study from distributions: Bivariate Normal ($N_2$) with parameters$\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$ and $\rho$ respectively, and Bivariate Gamma ($Ga_2$) with parameters $a_1$, $a_2$, $c$, and $\mu$, respectively, based on Theorem 2 of Nadarajah and Gupta (2006).

proposed by Spearman (1904), which is equivalent to the Pearson correlation for the ranks of observations, is defined as

$$\rho_s = \text{Corr}\left(F_1(X)F_2(Y)\right) = 12\text{E}\left[F_1(X)F_2(Y)\right] - 3. \tag{3.52}$$

For more details, we refer readers to Croux and Dehon (2010), Kojadinovic and Yan (2010), Genest and Favre (2007), and Borkowf (1999). A consistent estimator for Spearman's $\rho$, $\rho_s$ based on observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, is given by

$$r_{s,n} = \frac{\sum_{i=1}^{n}(R_{x,i} - \bar{R}_x)(R_{y,i} - \bar{R}_y)}{\sqrt{\sum_{i=1}^{n}(R_{x,i} - \bar{R}_x)^2 \sum_{i=1}^{n}(R_{y,i} - \bar{R}_y)^2}} \tag{3.53}$$

$$= \frac{12}{n(n+1)(n-1)} \sum_{i=1}^{n} R_{x,i}R_{y,i} - 3\frac{n+1}{n-1} \tag{3.54}$$

$$= 1 - \frac{6\sum_{i=1}^{n}(R_{x,i} - R_{y,i})^2}{n^3 - n}, \tag{3.55}$$

where $R_{x,i}$ and $R_{y,i}$ are respectively ranks of $X_i$ among all $X$'s and $Y_i$ among all $Y$'s. Using Borkowf (2002) and Hoeffding (1948), the asymptotic distribution of $r_{s,n}$ is

$$\sqrt{n}\left(r_{s,n} - \rho_s\right) \xrightarrow{D} N\left(0, \xi_{\rho_s}^2\right), \tag{3.56}$$

where

$$\xi_{\rho_s}^2 = 144(-9\theta_1^2 + \theta_3 + 2\theta_4 + 2\theta_5 + 2\theta_6), \tag{3.57}$$

and

$$\theta_1 = \mathrm{E}\left[F_1(X_1)F_2(Y_1)\right] \tag{3.58}$$

$$\theta_3 = \mathrm{E}\left[S_1(X_1)^2 S_2(Y_1)^2\right] \tag{3.59}$$

$$\theta_4 = \mathrm{E}\left[S(X_1, Y_2)S_1(X_2)S_2(Y_1)\right] \tag{3.60}$$

$$\theta_5 = \mathrm{E}\left[S(\max\{X_1, X_2\})S_2(Y_1)S_2(Y_2)\right] \tag{3.61}$$

$$\theta_6 = \mathrm{E}\left[S_1(X_1)S_1(X_2)S(\max\{Y_1, Y_2\})\right] \tag{3.62}$$

$$S_i(x) = 1 - F_i(x), \quad i \in \{1, 2\} \tag{3.63}$$

$$S(x, y) = 1 - F_1(x) - F_2(y) + F(x, y). \tag{3.64}$$

Proceeding along the same lines as given in Equations (3.8)–(3.11), we can find that the sample size required to achieve the sufficient accuracy with pre-specified upper bound ($\omega$)

on the width of the confidence interval for $\rho_s$ will be

$$n \geq \left\lceil \frac{4z_{\alpha/2}^2 \xi_{\rho_s}^2}{\omega^2} \right\rceil \equiv n_{SR}, \tag{3.65}$$

where $\xi_{\rho_s}^2$ is defined as in Equation (3.57). In practice, $\xi_{\rho_s}^2$ is usually unknown and we use a consistent estimator. According to Genest and Favre (2007), an estimator of $\xi_{\rho_s}^2$ is given by

$$V_{GF,n}^2 = 144(-9\mathcal{A}_n + \mathcal{B}_n + 2\mathcal{C}_n + 2\mathcal{D}_n + 2\mathcal{E}_n) \tag{3.66}$$

where

$$\mathcal{A}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{R_{x,i}}{n+1} \frac{R_{y,i}}{n+1} \tag{3.67}$$

$$\mathcal{B}_n = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{R_{x,i}}{n+1} \right)^2 \left( \frac{R_{y,i}}{n+1} \right)^2 \tag{3.68}$$

$$\mathcal{C}_n = \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{R_{x,i}}{n+1} \frac{R_{y,i}}{n+1} 1 \left\{ R_{x,k} \leq R_{x,i}, R_{x,k} \leq R_{x,j} \right\} + \frac{1}{4} - \mathcal{A}_n \tag{3.69}$$

$$\mathcal{D}_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{R_{y,i}}{n+1} \frac{R_{y,j}}{n+1} \max \left\{ \frac{R_{x,i}}{n+1} \frac{R_{x,j}}{n+1} \right\} \tag{3.70}$$

$$\mathcal{E}_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{R_{x,i}}{n+1} \frac{R_{x,j}}{n+1} \max \left\{ \frac{R_{y,i}}{n+1} \frac{R_{y,j}}{n+1} \right\}. \tag{3.71}$$

consistent estimator of $\xi_{\rho_s}^2$ as

$$V_{KY,n}^2 = \frac{144}{n-1} \sum_{i=1}^{n} (Z_i - \bar{Z})^2 \tag{3.72}$$

where

$$Z_i = \frac{R_{x,i}}{n+1} \frac{R_{y,i}}{n+1} + \frac{1}{n} \sum_{k=1}^{n} 1 \left\{ R_{x,i} \leq R_{x,k} \right\} \frac{R_{y,k}}{n+1} + \frac{1}{n} \sum_{k=1}^{n} 1 \left\{ R_{y,i} \leq R_{y,k} \right\} \frac{R_{x,k}}{n+1}. \tag{3.73}$$

Since $\xi_{\rho_s}^2$ is unknown in reality, in order to compute the required sample size, $n_{SR}$, we use sequential procedure. Our sequential stopping rule which helps find the estimate of the optimal sample size is as follows:

$$N_{SR} \text{ is the smallest integer } n (\geq m_{SR}) \text{ such that } \frac{4z_{\alpha/2}^2}{\omega^2} \left( \widehat{\xi}_{\rho_s n}^2 + n^{-1} \right), \tag{3.74}$$

where $m_{SR}$ is the pilot sample which is same as the pilot sample size as defined in Equation (3.14) and $\widehat{\xi}^2_{\rho_s n} = V^2_{GF,n} = V^2_{KY,n}$.

We now find the characteristics of the stopping rule defined in Equation 3.74 using Monte Carlo simulation for constructing $100(1-\alpha)\%$ confidence intervals for population correlation coefficients from bivariate distributions – bivariate normal and the bivariate gamma distribution from Theorem 2 of Nadarajah and Gupta (2006). The simulation study was done for correlation coefficient $\tau$ corresponding to $\rho$ of 0.1, 0.3, and 0.5 and $\omega = 0.1, 0.2$. In both cases, 5,000 replications were used. Table 3.6 shows the estimates of mean final sample size, coverage probability, and average confidence interval width and also the corresponding standard errors at the 90% and 95%.

The width of the confidence interval given by the sequential procedure with stopping rule defined in Equation (3.74) did not exceed the maximum specified width $\omega$. The coverage probability estimates are close to the corresponding confidence level. Also, the ratio of average sample size estimate and the optimal sample size is close to 1.

One can proceed along the same lines as in this manuscript (in the Appendix) and in Kelley et al. (2018) to prove that the respective coverage probabilities for the confidence interval for Kendall's tau ($\tau$) and Spearman's rho ($\rho_s$) are approximately close to the confidence level, and also the width of the confidence interval is less than $\omega$.

## 3.8 An Extension: Squared Multiple Correlation Coefficient

The sequential procedure that is proposed for correlation coefficient in the previous sections may be extended for finding a sufficiently narrow confidence interval for multiple correlation coefficient. In this section, we focus on the sequential AIPE procedure for multiple correlation coefficient under multivariate normal assumption only. We first formulate the corresponding AIPE problem. Although the previous parts of the article was distribution-free, here we assume multivariate normality because there is not, to our knowledge, a sufficient analytic

Table 3.6. Summary of final sample size for $100(1-\alpha)\%$ confidence interval for Spearman's rho, $\rho_s$

| $\omega$ | Distribution | $\rho$ | $\rho_s$ | $\bar{N}_{SR}$ | $se(\bar{N}_{SR})$ | $n_{SR}$ | $\bar{N}_{SR}/n_{SR}$ | $p_{SR}$ | $se(p_{SR})$ | $\bar{w}_{N_{SR}}$ | $se(\bar{w}_{N_{SR}})$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha = 0.10$ | | | | | | | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 0.0955 | 1065.0 | 0.4098 | 1066 | 0.9988 | 0.8998 | 0.0042 | 0.0999 | $4.070 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 0.2876 | 931.3 | 0.6019 | 933 | 0.9982 | 0.8964 | 0.0043 | 0.0999 | $5.599 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 0.4826 | 679.3 | 0.8218 | 683 | 0.9946 | 0.8876 | 0.0045 | 0.0998 | $1.063 \times 10^{-6}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 0.0915 | 1069.0 | 0.4031 | 1116 | 0.9579 | 0.9002 | 0.0042 | 0.0999 | $3.943 \times 10^{-7}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 0.2821 | 948.9 | 0.6067 | 957 | 0.9916 | 0.9004 | 0.0042 | 0.0999 | $5.432 \times 10^{-7}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 0.4765 | 714.2 | 0.8437 | 695 | 1.0280 | 0.8998 | 0.0042 | 0.0998 | $9.701 \times 10^{-7}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 0.0955 | 265.4 | 0.2140 | 267 | 0.9940 | 0.8898 | 0.0044 | 0.1993 | $3.409 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 0.2876 | 231.5 | 0.3236 | 234 | 0.9895 | 0.8890 | 0.0044 | 0.1990 | $1.780 \times 10^{-5}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 0.4826 | 166.5 | 0.4690 | 171 | 0.9735 | 0.8598 | 0.0049 | 0.1977 | $5.669 \times 10^{-5}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 0.0915 | 266.2 | 0.2102 | 279 | 0.9542 | 0.8924 | 0.0044 | 0.1993 | $3.444 \times 10^{-6}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 0.2821 | 235.5 | 0.3279 | 240 | 0.9913 | 0.8894 | 0.0044 | 0.1990 | $1.256 \times 10^{-5}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 0.4765 | 174.9 | 0.4832 | 174 | 1.0050 | 0.8618 | 0.0049 | 0.1980 | $4.947 \times 10^{-5}$ |
| $\alpha = 0.05$ | | | | | | | | | | | |
| 0.1 | $N_2(0,0,1,1,0.1)$ | 0.1 | 0.0955 | 1512.0 | 0.4830 | 1513 | 0.9994 | 0.9524 | 0.0030 | 0.0999 | $2.780 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 0.2876 | 1323.0 | 0.7309 | 1325 | 0.9986 | 0.9526 | 0.0030 | 0.0999 | $3.870 \times 10^{-7}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 0.4826 | 966.2 | 0.9759 | 970 | 0.9961 | 0.9410 | 0.0033 | 0.0999 | $7.030 \times 10^{-7}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 0.0915 | 1518.0 | 0.4826 | 1584 | 0.9583 | 0.9530 | 0.0030 | 0.0999 | $2.770 \times 10^{-7}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 0.2821 | 1348.0 | 0.7112 | 1358 | 0.9927 | 0.9526 | 0.0030 | 0.0999 | $3.740 \times 10^{-7}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 0.4765 | 1015.0 | 0.9828 | 986 | 1.0290 | 0.9508 | 0.0031 | 0.0999 | $6.472 \times 10^{-7}$ |
| 0.2 | $N_2(0,0,1,1,0.1)$ | 0.1 | 0.0955 | 377.4 | 0.2461 | 379 | 0.9958 | 0.9362 | 0.0035 | 0.1995 | $2.371 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.3)$ | 0.3 | 0.2876 | 329.7 | 0.3701 | 332 | 0.9931 | 0.9394 | 0.0034 | 0.1993 | $3.428 \times 10^{-6}$ |
| | $N_2(0,0,1,1,0.5)$ | 0.5 | 0.4826 | 238.8 | 0.5289 | 243 | 0.9829 | 0.9300 | 0.0036 | 0.1986 | $3.026 \times 10^{-5}$ |
| | $Ga_2(5,5,50,10)$ | 0.1 | 0.0915 | 378.7 | 0.2450 | 396 | 0.9562 | 0.9478 | 0.0031 | 0.1995 | $2.352 \times 10^{-6}$ |
| | $Ga_2(5,5,16.67,10)$ | 0.3 | 0.2821 | 336.1 | 0.3653 | 340 | 0.9887 | 0.9410 | 0.0033 | 0.1994 | $3.349 \times 10^{-6}$ |
| | $Ga_2(5,5,10,10)$ | 0.5 | 0.4765 | 250.9 | 0.5369 | 247 | 1.0160 | 0.9340 | 0.0035 | 0.1987 | $3.231 \times 10^{-5}$ |

Note: $\rho$ is the population Pearson's correlation coefficient; $\rho_s$ is the population Spearman's rho (computed using bootstrap method for bivariate Gamma distribution); $\bar{N}_{SR}$ is the mean final sample size; $p_{SR}$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $\rho_s$; $se(\bar{N}_{SR})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_{SR}$ is the theoretical sample size if the procedure is used with the population parameters (computed using bootstrap method for bivariate Ga distribution); $se(p_{SR})$ is the standard error of $p_{SR}$; $\bar{w}_{N_{SR}}$ average length of confidence intervals for $\rho_s$ based on $N$ observations; tabled values are based on 5,000 replications of a Monte Carlo simulation study from distributions: Bivariate Normal ($N_2$) distribution with parameters $\mu_1$, $\mu_2$, $\sigma_1^2$, $\sigma_2^2$ and $\rho$, respectively, and Bivariate Gamma ($Ga_2$) with parameters $a_1$, $a_2$, $c$, and $\mu$, respectively, based on Theorem 2 of Nadarajah and Gupta (2006).

method for forming a confidence interval for the population squared multiple correlation coefficient that is distribution-free.

Suppose, for the $i^{\text{th}}(i = 1, 2, \ldots, n)$ individual out of $n$ individuals, $Y_i$ is the score corresponding to the response variable and $X_{ij}$ is the observed score corresponding to the $j^{\text{th}}$

$(j = 1, 2, \ldots, k)$ predictor variable. Let $\mathbf{Y}$ denote the random vector of responses and $\mathbf{X}$ denote the corresponding random design matrix. The univariate linear regression model in matrix form is

$$\mathbf{Y} = \beta_0 \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.75}$$

where $\mathbf{1}$ is the vector where all elements are 1, $\boldsymbol{\beta}$ is the vector of fixed regression parameters and $\boldsymbol{\epsilon}$ is the random error vector. The population squared multiple correlation coefficient is given by

$$P^2 = \frac{\boldsymbol{\sigma}_{YX} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\sigma}_{XY}}{\sigma_Y^2} \tag{3.76}$$

where $\boldsymbol{\Sigma}_{XX}^{-1}$ is the inverse of the $k \times k$ population covariance matrix of the $k$ predictors, $\boldsymbol{\sigma}_{XY}$ is the $k$ dimensional column vector of covariances of the $k$ predictors with the response $Y$, $\boldsymbol{\sigma}_{YX}$ is the $k$ dimensional row vector of covariances of the $k$ predictors with the response $Y$ $(\boldsymbol{\sigma}'_{XY} = \boldsymbol{\sigma}_{YX})$, and $\sigma_Y^2$ is the population variance of the response $Y$.

A well known consistent estimator of the population squared multiple correlation coefficient, also known as R-squared ($R^2$) or multiple R squared, is given by

$$R^2 = \frac{\boldsymbol{s}_{YX} \mathbf{S}_{XX}^{-1} \boldsymbol{s}_{XY}}{s_Y^2} \tag{3.77}$$

where $\mathbf{S}_{XX}^{-1}$ is the inverse of the $k \times k$ sample covariance matrix of the $k$ predictors, $\boldsymbol{s}_{XY}$ is the $k$ dimensional column vector of sample covariances of the $k$ predictors with the response $Y$, $\boldsymbol{s}_{YX}$ is the $k$ dimensional row vector of sample covariances of the $k$ predictors with the response $Y$ $(\boldsymbol{s}'_{XY} = \boldsymbol{s}_{YX})$, and $s_Y^2$ is the sample variance of the response $Y$.

The approximate $100(1-\alpha)\%$ confidence interval for the squared population multiple correlation coefficient as given in Bonett and Wright (2011) is

$$1 - \exp\left(\ln\left(1 - R^2\right) \pm z_{\alpha/2} \frac{2P}{\sqrt{n-k-2}}\right), \tag{3.78}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)^{\text{th}}$ percentile of the standard normal distribution. The approximate $100(1-\alpha)\%$ Wald-type confidence interval for the population squared multiple correlation coefficient using Olkin and Finn (1995) is given by

$$R^2 \pm z_{\alpha/2} \frac{2P(1-P^2)}{\sqrt{n-k-2}}. \tag{3.79}$$

In AIPE of squared population multiple correlation coefficient, in order to have a sufficient narrow confidence interval, an upper bound ($\omega$) of the confidence interval width is prespecified. Using the width constraint, we can find the optimal sample size for the confidence interval given in Equation (3.78) which is given by

$$n \geq 2 + k + 4P^2 z_{\alpha/2}^2 \left\{ \ln \left[ \frac{1}{2} \left( \frac{\omega}{(1-P^2)} + \sqrt{\frac{\omega^2}{(1-P^2)^2} + 4} \right) \right] \right\}^{-2} \equiv n_{BW}, \tag{3.80}$$

and the optimal sample size for the confidence interval given in Equation (3.79) is

$$n \geq \frac{16 z_{\alpha/2}^2}{\omega^2} P^2 \left(1 - P^2\right)^2 + k + 2 \equiv n_{OF}. \tag{3.81}$$

The derivation of the optimal sample sizes is given in the next subsection. Thus, $n_{BW}$ is the optimal sample size which is required to get a sufficiently narrow confidence interval, of the form given in Equation (3.78), of the multiple correlation coefficient $P^2$. $n_{OF}$ is the optimal sample size which is required to get a sufficiently narrow confidence interval, of the form given in Equation (3.79), of the multiple correlation coefficient $P^2$.

Because $P^2$ is unknown, both $n_{BW}$ and $n_{OF}$ are also unknown. Thus, in order to obtain a sufficiently narrow confidence interval for $P^2$, we need to estimate $n_{BW}$ and $n_{OF}$. This can be done using sequential procedure similar to what was described earlier.

The stopping rule related to the sequential procedure for estimating $n_{BW}$ is given by: $N_{BW}$ is the smallest integer $n (\geq m_{BW})$ such that

$$n \geq 2 + k + 4R^2 z_{\alpha/2}^2 \left\{ \ln \left[ \frac{1}{2} \left( \frac{\omega}{(1-R^2)} + \sqrt{\frac{\omega^2}{(1-R^2)^2} + 4} \right) \right] \right\}^{-2}. \tag{3.82}$$

We propose the corresponding pilot sample size to be $m_{BW} = k + 2$. Now, the stopping rule related to the sequential procedure for estimating $n_{OF}$ is given by:

$N_{OF}$ is the smallest integer $n(\geq m_{OF})$ such that

$$n \geq \frac{16z_{\alpha/2}^2}{\omega^2}\left[\left(R^2+\frac{1}{n}\right)\left(1-\left(R^2+\frac{1}{n}\right)\right)^2\right] \tag{3.83}$$

where $m_{OF} = \max\{k, 4z_{\alpha/2}/\omega\}$ is the corresponding pilot sample size. We note that, in Equation 3.83, we use $R^2 + 1/n$ which is a consistent estimator of $P^2$. Next, the expression for the stopping rules in Equations (3.78) & (3.79) and their respective pilot sample sizes $m_{BW}$ and $m_{OF}$ will be derived.

### 3.8.1 Derivation of Optimal and Pilot Sample Sizes for Confidence Intervals for Squared Multiple Correlation Coefficient

For the situation in which the population squared multiple correlation coefficient is known, the optimal sample size for confidence interval provided by Bonett and Wright (2011) in Equation (3.78) can be derived as follows:

$$\begin{aligned}
\exp\left(\ln\left(1-R^2\right)+\frac{2Pz_{\alpha/2}}{\sqrt{n-k-2}}\right)-\exp\left(\ln\left(1-R^2\right)-\frac{2Pz_{\alpha/2}}{\sqrt{n-k-2}}\right) \leq&\omega\\
\left(1-R^2\right)\left[\exp\left(\frac{2Pz_{\alpha/2}}{\sqrt{n-k-2}}\right)-\exp\left(\frac{-2Pz_{\alpha/2}}{\sqrt{n-k-2}}\right)\right] \leq&\omega\\
\exp\left(\frac{2Pz_{\alpha/2}}{\sqrt{n-k-2}}\right)-\exp\left(\frac{-2Pz_{\alpha/2}}{\sqrt{n-k-2}}\right) \leq&\frac{\omega}{(1-R^2)}\\
\exp\left(\frac{4P^2z_{\alpha/2}^2}{n-k-2}\right)-\frac{\omega}{(1-R^2)}\exp\left(\frac{2Pz_{\alpha/2}}{\sqrt{n-k-2}}\right)-1 \leq&0. \qquad (3.84)
\end{aligned}$$

One may note that $\exp\left(\frac{2Pz_{\alpha/2}}{\sqrt{n-k-2}}\right) > 0$. Thus, using Equation (3.84), we have

$$\begin{aligned}\exp\left(\frac{2Pz_{\alpha/2}}{\sqrt{n-k-2}}\right) &\leq \frac{1}{2}\left(\frac{\omega}{(1-R^2)} + \sqrt{\frac{\omega^2}{(1-R^2)^2}+4}\right)\\ \frac{2Pz_{\alpha/2}}{\sqrt{n-k-2}} &\leq \ln\left[\frac{1}{2}\left(\frac{\omega}{(1-R^2)} + \sqrt{\frac{\omega^2}{(1-R^2)^2}+4}\right)\right]\\ \sqrt{n-k-2} &\geq 2Pz_{\alpha/2}\left\{\ln\left[\frac{1}{2}\left(\frac{\omega}{(1-R^2)} + \sqrt{\frac{\omega^2}{(1-R^2)^2}+4}\right)\right]\right\}^{-1}\\ n &\geq 2+k+4P^2z_{\alpha/2}^2\left\{\ln\left[\frac{1}{2}\left(\frac{\omega}{(1-R^2)} + \sqrt{\frac{\omega^2}{(1-R^2)^2}+4}\right)\right]\right\}^{-2}.\end{aligned}$$

Next, using the stopping rule in Equation (3.82), the corresponding pilot sample can be derived by

$$n \geq 2+k+4R^2z_{\alpha/2}^2\left\{\ln\left[\frac{1}{2}\left(\frac{\omega}{(1-R^2)} + \sqrt{\frac{\omega^2}{(1-R^2)^2}+4}\right)\right]\right\}^{-2} \implies n \geq 2+k. \quad (3.85)$$

Thus, the pilot sample size $m_{BW}$ is defined as

$$m_{BW} = k+2. \quad (3.86)$$

Furthermore, given that the population squared multiple correlation coefficient is known, the optimal sample size for the Wald-type confidence interval provided by Olkin and Finn (1995) in Equation (3.79) can be derived as follows:

$$\begin{aligned}4z_{\alpha/2}\frac{P(1-P^2)}{\sqrt{n-k-2}} \leq \omega &\implies n-k-2 \geq 16z_{\alpha/2}^2\frac{P^2(1-P^2)^2}{\omega^2}\\ &\implies n \geq 16z_{\alpha/2}^2\frac{P^2(1-P^2)^2}{\omega^2}+k+2\end{aligned}$$

The corresponding pilot sample size can be derive as

$$n \geq \frac{16z_{\alpha/2}^2}{\omega^2}\left(R^2\left(1-R^2\right)^2+\frac{1}{n}\right) \geq \frac{16z_{\alpha/2}^2}{n\omega^2} \implies n^2 \geq \frac{16z_{\alpha/2}^2}{\omega^2} \implies n \geq \frac{4z_{\alpha/2}}{\omega}. \quad (3.87)$$

The pilot sample size $m_{OF}$ is therefore defined as

$$m_{OF} = \max\left\{k+2, \frac{4z_{\alpha/2}}{\omega}\right\}. \quad (3.88)$$

### 3.8.2 Simulation Results

We now find the characteristics of the stopping rules defined in Equations (3.82) and (3.83) using Monte Carlo simulation for constructing $100(1-\alpha)\%$ confidence intervals for squared population multiple correlation coefficient from multivariate distribution with mean parameter vector and dispersion matrix respectively given by

$$\mu = (0, \ldots, 0)' = \mathbf{0}_{(k+1)\times 1} \qquad \text{and} \qquad \mathbf{\Sigma}_{(k+1)\times(k+1)} = \begin{bmatrix} 1 & \boldsymbol{\gamma}' \\ \boldsymbol{\gamma} & \mathbf{I} \end{bmatrix}$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_k)' = \boldsymbol{\sigma_{YX}}$ with $\gamma_i = \sqrt{P^2/k}$ for $i = 1, \ldots, k$, and $\mathbf{I}$ is a $k \times k$ identity matrix (i.e. $\text{cov}(X) = \mathbf{I}$). The simulation study was done for squared population multiple correlation under several scenarios with replication size 5,000.

Tables 3.7–3.10 show the estimates of mean final sample size, coverage probability, and also the corresponding standard errors and the average confidence interval width. The simulation results show that the width of the confidence interval given by the sequential procedure with stopping rules defined in Equations (3.82)–(3.83) did not exceed the maximum specified width $\omega$. Except for smaller optimal sample sizes, the coverage probability estimates are close to the corresponding confidence level and also, the ratio of average sample size estimate and the optimal sample size is close to 1. Overall, the results for the sequential procedure corresponding to the confidence interval given by Olkin and Finn (1995) performed better than the procedure given by Bonett and Wright (2011) in terms of required optimal sample size.

## 3.9 Discussion

The correlation coefficient is a widely used effect size in psychology and related fields and estimating the population value is of great importance. The necessity of using a $100(1-\alpha)\%$ confidence interval that brackets a wide range of values in order to include the true value, with the specified level of confidence, represents an important problem. Correspondingly, a method to obtain a sufficiently narrow $100(1-\alpha)\%$ confidence interval for the population correlation coefficient with a confidence interval width no larger than desired is very advantageous in many research contexts. However, until now, all such approaches required the specification of unknown population values and bivariate normality. Our approach overcomes both of these limitations. We discuss a distribution-free confidence interval approach for the population correlation coefficients viz. Pearson's product moment correlation coefficient, Kendall's $\tau$ rank correlation coefficient, and Spearman's $\rho$ rank correlation coefficient. We then use the distribution-free framework to develop a sequential approach to accuracy in parameter estimation of the correlation coefficients.

It is known that, holding everything else constant, a narrower confidence interval provides more information about the parameter than a wider confidence interval. Given a value of the upper bound of the confidence interval ($\omega$), an approximate $100(1-\alpha)\%$ confidence interval for the population correlation coefficient can be constructed by using an a priori sample size planning approach, which requires a hypothesized value of the population parameters. Using a supposed population value based on theory, an estimate from one or more other studies, or a conjecture based on a rule of thumb can lead to sample size estimates that grossly differ from what the theoretically optimal sample size would be if the population parameters were known and assumptions specified. We overcome such a limitation by proposing a sequential procedure which can be used to construct an approximate $100(1-\alpha)\%$ confidence interval for the population correlation coefficients within a pre-specified width ($\omega$) without assuming any distribution of the data. Unlike a priori sample size planning approaches, our sequential procedure does not require knowledge of population parameters in order to obtain a sufficiently narrow confidence interval.

We discuss a sequential approach to construct a sufficiently narrow confidence interval for the population correlation coefficient assuming homogeneity of the data distribution. Some studies (e.g., Stanley et al., 2017; van Erp et al., 2017) have shown that heterogeneity of the data distribution is possible due to which the population effect size may change (e.g., parameter drift). However, the incorporation of heterogeneity or parameter drift is beyond the scope of the article. To the extent that heterogeneity or parameter drift exists over the time frame in which data are collected, it would be a limitation. Additionally, the methods we use for confidence interval construction do not work well in all situations, particularly for small sample sizes combined with non-normal data. Nevertheless, without assuming any particular distribution, the distribution-free methods we use for Pearson's, Spearman's, and Kendall's correlations, provided sample size is not too small for the particular situation, will work well as sample size gets larger. Another limitation in this regard is that our sequential methods for the squared multiple correlation coefficient requires multivariate normality, as a well developed distribution-free confidence interval method for the squared multiple correlation coefficient is not yet available.

As a general overview of our procedure, we first obtain a pilot sample size. After collecting the pilot data, we then use a sequential sampling procedure where, at each stage, we check whether a stopping rule has been satisfied. If not, additional observation(s) from one or more individuals, depending on the selected sample size at each stage, on both variables are collected and the check is performed again. This process continues until the stopping rule is satisfied. Our method ensures that the length of the confidence interval for correlation coefficient is less than the desired width and also attains the coverage probability asymptotically while using the smallest possible sample size. Based on the limitation of existing sample size procedures with regard to distribution assumption and assumed knowledge of population parameters, our sequential procedure has the potential to be widely used in psychology and related disciplines. To help researchers, we have provided freely available R functions via the MBESS package.

Table 3.7. Summary of final sample size for 90% confidence interval for $P^2$ using Bonett and Wright (2011)

| $k$ | $\omega$ | $P^2$ | $\bar{N}_{BW}$ | $se(\bar{N}_{BW})$ | $n_{BW}$ | $\bar{N}_{BW}/n_{BW}$ | $p_{BW}$ | $se(p_{BW})$ | $\bar{w}_{N_{BW}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.05 | 0.10 | 1308.44 | 5.4457 | 1407 | 0.9299 | 0.8344 | 0.0053 | 0.0489 |
| | | 0.30 | 2529.40 | 3.0025 | 2551 | 0.9915 | 0.8866 | 0.0045 | 0.0499 |
| | | 0.50 | 2149.81 | 2.6762 | 2171 | 0.9902 | 0.8916 | 0.0044 | 0.0499 |
| | | 0.70 | 1071.65 | 2.3025 | 1098 | 0.9760 | 0.8886 | 0.0044 | 0.0496 |
| | | 0.90 | 140.63 | 0.8780 | 164 | 0.8575 | 0.7894 | 0.0058 | 0.0481 |
| | 0.10 | 0.10 | 297.69 | 2.0105 | 355 | 0.8386 | 0.7404 | 0.0062 | 0.0959 |
| | | 0.30 | 623.99 | 1.3225 | 642 | 0.9719 | 0.8768 | 0.0046 | 0.0992 |
| | | 0.50 | 533.22 | 1.1170 | 547 | 0.9748 | 0.8858 | 0.0045 | 0.0992 |
| | | 0.70 | 262.79 | 0.9621 | 280 | 0.9385 | 0.8554 | 0.0050 | 0.0982 |
| | | 0.90 | 34.84 | 0.3222 | 47 | 0.7412 | 0.7688 | 0.0060 | 0.0908 |
| 5 | 0.05 | 0.10 | 1414.03 | 2.8253 | 1410 | 1.0029 | 0.8948 | 0.0043 | 0.0498 |
| | | 0.30 | 2525.96 | 3.5884 | 2554 | 0.9890 | 0.8916 | 0.0044 | 0.0498 |
| | | 0.50 | 2130.43 | 3.9169 | 2174 | 0.9800 | 0.8846 | 0.0045 | 0.0497 |
| | | 0.70 | 1056.26 | 2.8159 | 1101 | 0.9594 | 0.8722 | 0.0047 | 0.0494 |
| | | 0.90 | 129.40 | 0.9586 | 167 | 0.7748 | 0.7196 | 0.0064 | 0.0473 |
| | 0.10 | 0.10 | 364.23 | 1.3204 | 358 | 1.0174 | 0.8802 | 0.0046 | 0.0989 |
| | | 0.30 | 630.03 | 1.2387 | 645 | 0.9768 | 0.8816 | 0.0046 | 0.0992 |
| | | 0.50 | 524.81 | 1.4068 | 550 | 0.9542 | 0.8716 | 0.0047 | 0.0987 |
| | | 0.70 | 252.39 | 1.1333 | 283 | 0.8918 | 0.8216 | 0.0054 | 0.0970 |
| | | 0.90 | 30.90 | 0.3098 | 50 | 0.6179 | 0.6668 | 0.0067 | 0.0871 |
| 10 | 0.05 | 0.10 | 1444.09 | 3.0292 | 1415 | 1.0206 | 0.8938 | 0.0044 | 0.0497 |
| | | 0.30 | 2503.95 | 5.1902 | 2559 | 0.9785 | 0.8908 | 0.0044 | 0.0496 |
| | | 0.50 | 2101.63 | 5.2572 | 2179 | 0.9645 | 0.8720 | 0.0047 | 0.0493 |
| | | 0.70 | 1033.60 | 3.5467 | 1106 | 0.9345 | 0.8510 | 0.0050 | 0.0489 |
| | | 0.90 | 113.23 | 1.0222 | 172 | 0.6583 | 0.6032 | 0.0069 | 0.0455 |
| | 0.10 | 0.10 | 391.00 | 1.3881 | 363 | 1.0771 | 0.8580 | 0.0049 | 0.0981 |
| | | 0.30 | 621.40 | 1.8054 | 650 | 0.9560 | 0.8506 | 0.0050 | 0.0982 |
| | | 0.50 | 509.91 | 1.8875 | 555 | 0.9188 | 0.8298 | 0.0053 | 0.0973 |
| | | 0.70 | 234.17 | 1.3801 | 288 | 0.8131 | 0.7476 | 0.0061 | 0.0950 |
| | | 0.90 | 27.13 | 0.2533 | 55 | 0.4932 | 0.5132 | 0.0071 | 0.0817 |

Note: $P^2$ is the population multiple correlation; $\bar{N}_{BW}$ is the mean final sample size; $p_{BW}$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $P^2$; $se(\bar{N}_{BW})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_{BW}$ is the theoretical sample size if the procedure is used with known population value of $P^2$; $se(p_{BW})$ is the standard error of $p_{BW}$; $\bar{w}_{BW}$ average length of confidence intervals for $P^2$ based on $N_{BW}$ observations; tabled values are based on 5,000 replications of a Monte Carlo simulation study from Multivariate Normal distribution ($N_k$) with parameters: mean vector $\mu$ and variance covariance matrix $\mathbf{\Sigma}$.

Table 3.8. Summary of final sample size for 95% confidence interval for $P^2$ using Bonett and Wright (2011)

| $k$ | $\omega$ | $P^2$ | $\bar{N}_{BW}$ | $se(\bar{N}_{BW})$ | $n_{BW}$ | $\bar{N}_{BW}/n_{BW}$ | $p_{BW}$ | $se(p_{BW})$ | $\bar{w}_{N_{BW}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.05 | 0.10 | 1893.25 | 6.6567 | 1996 | 0.9485 | 0.8966 | 0.0043 | 0.0492 |
| | | 0.30 | 3596.31 | 3.8162 | 3620 | 0.9935 | 0.9388 | 0.0034 | 0.0499 |
| | | 0.50 | 3063.19 | 2.8966 | 3080 | 0.9945 | 0.9448 | 0.0032 | 0.0499 |
| | | 0.70 | 1537.45 | 2.3780 | 1557 | 0.9874 | 0.9410 | 0.0033 | 0.0498 |
| | | 0.90 | 212.23 | 0.9910 | 230 | 0.9228 | 0.8940 | 0.0044 | 0.0490 |
| | 0.10 | 0.10 | 435.52 | 2.5998 | 503 | 0.8659 | 0.8132 | 0.0055 | 0.0968 |
| | | 0.30 | 892.21 | 1.5674 | 910 | 0.9805 | 0.9374 | 0.0034 | 0.0995 |
| | | 0.50 | 760.73 | 1.3081 | 775 | 0.9816 | 0.9386 | 0.0034 | 0.0995 |
| | | 0.70 | 380.06 | 1.0911 | 395 | 0.9622 | 0.9254 | 0.0037 | 0.0989 |
| | | 0.90 | 50.55 | 0.4177 | 64 | 0.7899 | 0.8432 | 0.0051 | 0.0936 |
| 5 | 0.05 | 0.10 | 2005.04 | 3.3728 | 1999 | 1.0030 | 0.9432 | 0.0033 | 0.0498 |
| | | 0.30 | 3600.12 | 3.8252 | 3623 | 0.9937 | 0.9486 | 0.0031 | 0.0499 |
| | | 0.50 | 3055.20 | 3.5829 | 3083 | 0.9910 | 0.9432 | 0.0033 | 0.0499 |
| | | 0.70 | 1520.37 | 3.2019 | 1560 | 0.9746 | 0.9322 | 0.0036 | 0.0496 |
| | | 0.90 | 198.56 | 1.1441 | 233 | 0.8522 | 0.8384 | 0.0052 | 0.0482 |
| | 0.10 | 0.10 | 513.15 | 1.5693 | 506 | 1.0141 | 0.9334 | 0.0035 | 0.0993 |
| | | 0.30 | 897.13 | 1.5073 | 913 | 0.9826 | 0.9342 | 0.0035 | 0.0993 |
| | | 0.50 | 754.62 | 1.6021 | 778 | 0.9699 | 0.9284 | 0.0036 | 0.0991 |
| | | 0.70 | 369.88 | 1.2873 | 398 | 0.9294 | 0.9074 | 0.0041 | 0.0982 |
| | | 0.90 | 45.27 | 0.4144 | 67 | 0.6757 | 0.7616 | 0.0060 | 0.0910 |
| 10 | 0.05 | 0.10 | 2034.40 | 3.6812 | 2004 | 1.0152 | 0.9446 | 0.0032 | 0.0498 |
| | | 0.30 | 3588.62 | 5.2427 | 3628 | 0.9891 | 0.9446 | 0.0032 | 0.0498 |
| | | 0.50 | 3028.29 | 5.4085 | 3088 | 0.9807 | 0.9402 | 0.0034 | 0.0497 |
| | | 0.70 | 1499.93 | 4.0235 | 1565 | 0.9584 | 0.9228 | 0.0038 | 0.0493 |
| | | 0.90 | 182.09 | 1.2902 | 238 | 0.7651 | 0.7526 | 0.0061 | 0.0473 |
| | 0.10 | 0.10 | 541.94 | 1.5947 | 511 | 1.0606 | 0.9308 | 0.0036 | 0.0990 |
| | | 0.30 | 893.74 | 1.9830 | 918 | 0.9736 | 0.9236 | 0.0038 | 0.0990 |
| | | 0.50 | 744.97 | 2.0460 | 783 | 0.9514 | 0.9100 | 0.0040 | 0.0985 |
| | | 0.70 | 354.90 | 1.6083 | 403 | 0.8806 | 0.8572 | 0.0049 | 0.0969 |
| | | 0.90 | 39.85 | 0.3800 | 72 | 0.5534 | 0.6460 | 0.0068 | 0.0879 |

Note: $P^2$ is the population multiple correlation; $\bar{N}_{BW}$ is the mean final sample size; $p_{BW}$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $P^2$; $se(\bar{N}_{BW})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_{BW}$ is the theoretical sample size if the procedure is used with known population value of $P^2$; $se(p_{BW})$ is the standard error of $p_{BW}$; $\bar{w}_{BW}$ average length of confidence intervals for $P^2$ based on $N_{BW}$ observations; tabled values are based on 5,000 replications of a Monte Carlo simulation study from Multivariate Normal distribution with parameters: mean vector $\mu$ and variance covariance matrix $\mathbf{\Sigma}$.

Table 3.9. Summary of final sample size for 90% confidence interval for $P^2$ using Olkin and Finn (1995)

| $k$ | $\omega$ | $P^2$ | $\bar{N}_{OF}$ | $se(\bar{N}_{OF})$ | $n_{OF}$ | $\bar{N}_{OF}/n_{OF}$ | $p_{OF}$ | $se(p_{OF})$ | $\bar{w}_{N_{OF}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.05 | 0.10 | 1406.47 | 2.4778 | 1403 | 1.0025 | 0.8902 | 0.0044 | 0.0497 |
| | | 0.30 | 2548.32 | 0.2678 | 2546 | 1.0009 | 0.8986 | 0.0043 | 0.0499 |
| | | 0.50 | 2165.16 | 0.9286 | 2165 | 1.0001 | 0.9024 | 0.0042 | 0.0499 |
| | | 0.70 | 1091.74 | 1.2002 | 1091 | 1.0007 | 0.9026 | 0.0042 | 0.0498 |
| | | 0.90 | 174.44 | 0.4987 | 156 | 1.1182 | 0.8898 | 0.0044 | 0.0461 |
| | 0.10 | 0.10 | 347.98 | 1.3031 | 351 | 0.9914 | 0.8460 | 0.0051 | 0.0962 |
| | | 0.30 | 638.92 | 0.1593 | 637 | 1.0030 | 0.8930 | 0.0044 | 0.0995 |
| | | 0.50 | 543.16 | 0.4694 | 542 | 1.0021 | 0.8962 | 0.0043 | 0.0993 |
| | | 0.70 | 275.69 | 0.6141 | 273 | 1.0099 | 0.8830 | 0.0045 | 0.0981 |
| | | 0.90 | 68.66 | 0.0963 | 39 | 1.7605 | 0.8678 | 0.0048 | 0.0733 |
| 5 | 0.05 | 0.10 | 1427.36 | 2.3615 | 1403 | 1.0174 | 0.8900 | 0.0044 | 0.0497 |
| | | 0.30 | 2549.42 | 0.2636 | 2546 | 1.0013 | 0.8938 | 0.0044 | 0.0499 |
| | | 0.50 | 2161.49 | 0.9365 | 2165 | 0.9984 | 0.8998 | 0.0042 | 0.0499 |
| | | 0.70 | 1087.11 | 1.2317 | 1091 | 0.9964 | 0.8908 | 0.0044 | 0.0498 |
| | | 0.90 | 170.13 | 0.4898 | 156 | 1.0906 | 0.8684 | 0.0048 | 0.0458 |
| | 0.10 | 0.10 | 375.77 | 1.0788 | 351 | 1.0706 | 0.8888 | 0.0044 | 0.0976 |
| | | 0.30 | 640.31 | 0.1399 | 637 | 1.0052 | 0.8932 | 0.0044 | 0.0995 |
| | | 0.50 | 539.22 | 0.4790 | 542 | 0.9949 | 0.8854 | 0.0045 | 0.0993 |
| | | 0.70 | 268.82 | 0.6312 | 273 | 0.9847 | 0.8604 | 0.0049 | 0.0980 |
| | | 0.90 | 67.94 | 0.0812 | 39 | 1.7420 | 0.8406 | 0.0052 | 0.0710 |
| 10 | 0.05 | 0.10 | 1458.75 | 2.2223 | 1403 | 1.0397 | 0.8964 | 0.0043 | 0.0497 |
| | | 0.30 | 2550.97 | 0.2510 | 2546 | 1.0020 | 0.9016 | 0.0042 | 0.0499 |
| | | 0.50 | 2157.60 | 0.9383 | 2165 | 0.9966 | 0.8962 | 0.0043 | 0.0499 |
| | | 0.70 | 1080.75 | 1.2345 | 1091 | 0.9906 | 0.8898 | 0.0044 | 0.0498 |
| | | 0.90 | 163.12 | 0.4639 | 156 | 1.0456 | 0.8106 | 0.0055 | 0.0452 |
| | 0.10 | 0.10 | 403.12 | 0.9458 | 351 | 1.1485 | 0.8736 | 0.0047 | 0.0981 |
| | | 0.30 | 641.16 | 0.1292 | 637 | 1.0065 | 0.8820 | 0.0046 | 0.0995 |
| | | 0.50 | 534.97 | 0.4924 | 542 | 0.9870 | 0.8736 | 0.0047 | 0.0993 |
| | | 0.70 | 261.45 | 0.6568 | 273 | 0.9577 | 0.8236 | 0.0054 | 0.0979 |
| | | 0.90 | 67.00 | 0.0584 | 39 | 1.7179 | 0.7294 | 0.0063 | 0.0661 |

Note: $P^2$ is the population multiple correlation coefficient; $\bar{N}_{OF}$ is the mean final sample size; $p_{OF}$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $P^2$; $se(\bar{N}_{OF})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_{OF}$ is the theoretical sample size if the procedure is used with known population value of $P^2$; $se(p_{OF})$ is the standard error of $p_{OF}$; $\bar{w}_{N_{OF}}$ average length of confidence intervals for $P^2$ based on $N_{OF}$ observations; tabled values are based on 5,000 replications of a Monte Carlo simulation study from Multivariate Normal distribution with parameters: mean vector $\mu$ and variance covariance matrix $\mathbf{\Sigma}$.

Table 3.10. Summary of final sample size for 95% confidence interval for $P^2$ using Olkin and Finn (1995)

| $k$ | $\omega$ | $P^2$ | $\bar{N}_{OF}$ | $se(\bar{N}_{OF})$ | $n_{OF}$ | $\bar{N}_{OF}/n_{OF}$ | $p_{OF}$ | $se(p_{OF})$ | $\bar{w}_{N_{OF}}$ |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0.05 | 0.10 | 1997.80 | 2.9059 | 1992 | 1.0029 | 0.9404 | 0.0033 | 0.0498 |
| | | 0.30 | 3616.98 | 0.3207 | 3615 | 1.0005 | 0.9432 | 0.0033 | 0.0500 |
| | | 0.50 | 3074.28 | 1.1125 | 3074 | 1.0001 | 0.9498 | 0.0031 | 0.0499 |
| | | 0.70 | 1550.08 | 1.4774 | 1549 | 1.0007 | 0.9412 | 0.0033 | 0.0498 |
| | | 0.90 | 237.06 | 0.6750 | 222 | 1.0679 | 0.9270 | 0.0037 | 0.0473 |
| | 0.10 | 0.10 | 495.01 | 1.5851 | 498 | 0.9940 | 0.9152 | 0.0039 | 0.0976 |
| | | 0.30 | 906.47 | 0.1733 | 904 | 1.0027 | 0.9474 | 0.0032 | 0.0996 |
| | | 0.50 | 769.43 | 0.5548 | 769 | 1.0006 | 0.9522 | 0.0030 | 0.0995 |
| | | 0.70 | 389.45 | 0.7446 | 388 | 1.0037 | 0.9328 | 0.0035 | 0.0987 |
| | | 0.90 | 84.17 | 0.1493 | 56 | 1.5030 | 0.9222 | 0.0038 | 0.0790 |
| 5 | 0.05 | 0.10 | 2016.71 | 2.8057 | 1992 | 1.0124 | 0.9492 | 0.0031 | 0.0498 |
| | | 0.30 | 3618.07 | 0.3129 | 3615 | 1.0008 | 0.9504 | 0.0031 | 0.0500 |
| | | 0.50 | 3070.70 | 1.1156 | 3074 | 0.9989 | 0.9502 | 0.0031 | 0.0499 |
| | | 0.70 | 1544.58 | 1.4603 | 1549 | 0.9971 | 0.9492 | 0.0031 | 0.0498 |
| | | 0.90 | 230.38 | 0.6648 | 222 | 1.0377 | 0.9108 | 0.0040 | 0.0471 |
| | 0.10 | 0.10 | 522.35 | 1.3550 | 498 | 1.0489 | 0.9442 | 0.0032 | 0.0983 |
| | | 0.30 | 907.40 | 0.1615 | 904 | 1.0038 | 0.9442 | 0.0032 | 0.0996 |
| | | 0.50 | 766.17 | 0.5611 | 769 | 0.9963 | 0.9450 | 0.0032 | 0.0995 |
| | | 0.70 | 383.90 | 0.7328 | 388 | 0.9894 | 0.9320 | 0.0036 | 0.0986 |
| | | 0.90 | 82.94 | 0.1290 | 56 | 1.4810 | 0.8988 | 0.0043 | 0.0769 |
| 10 | 0.05 | 0.10 | 2047.95 | 2.6772 | 1992 | 1.0281 | 0.9508 | 0.0031 | 0.0498 |
| | | 0.30 | 3619.31 | 0.3047 | 3615 | 1.0012 | 0.9522 | 0.0030 | 0.0500 |
| | | 0.50 | 3066.39 | 1.0971 | 3074 | 0.9975 | 0.9538 | 0.0030 | 0.0499 |
| | | 0.70 | 1540.85 | 1.4559 | 1549 | 0.9947 | 0.9466 | 0.0032 | 0.0498 |
| | | 0.90 | 223.07 | 0.6671 | 222 | 1.0048 | 0.8720 | 0.0047 | 0.0469 |
| | 0.10 | 0.10 | 552.43 | 1.1856 | 498 | 1.1093 | 0.9406 | 0.0033 | 0.0986 |
| | | 0.30 | 908.41 | 0.1527 | 904 | 1.0049 | 0.9384 | 0.0034 | 0.0996 |
| | | 0.50 | 762.34 | 0.5808 | 769 | 0.9913 | 0.9360 | 0.0035 | 0.0995 |
| | | 0.70 | 376.14 | 0.7833 | 388 | 0.9694 | 0.9038 | 0.0042 | 0.0986 |
| | | 0.90 | 81.28 | 0.0972 | 56 | 1.4513 | 0.8304 | 0.0053 | 0.0728 |

Note: $P^2$ is the population multiple correlation coefficient; $\bar{N}_{OF}$ is the mean final sample size; $p_{OF}$ is the estimated coverage probability; $\omega$ is the upper bound of the length of the confidence interval for $P^2$; $se(\bar{N}_{OF})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); $n_{OF}$ is the theoretical sample size if the procedure is used with known population value of $P^2$; $se(p_{OF})$ is the standard error of $p_{OF}$; $\bar{w}_{N_{OF}}$ average length of confidence intervals for $P^2$ based on $N_{OF}$ observations; tabled values are based on 5,000 replications of a Monte Carlo simulation study from Multivariate Normal distribution with parameters: mean vector $\mu$ and variance covariance matrix $\mathbf{\Sigma}$.

# CHAPTER 4

# GINI INDEX ESTIMATION WITHIN PRE-SPECIFIED ERROR BOUND APPLIED TO INDIAN HOUSEHOLD SURVEY DATA[1]

## 4.1 Introduction

The Gini index is a widely used measure of economic inequality which arises due to the disparity in income or wealth that exists in all countries, states or societies. The most celebrated population Gini index is given by

$$G_F \equiv G_F(X) = \frac{2}{\mu}\int_0^\infty xF(x)\, dF(x) - 1, \quad \mu = \mathrm{E}(X). \tag{4.1}$$

This index lies between 0 and 1, with 0 being perfect equality and 1 being perfect inequality. The Gini index of a country or a region can be computed using household surveys, generally known as complex household surveys, which involve stratification and clustering sampling methods. For example, the National Sample Survey (NSS) conducts such household surveys in India. Similar complex survey designs are used by survey agencies in United States, European Union and others. For details about the surveys adopted by various agencies in different countries, please refer to Bhattacharya (2005, 2007). The complex household survey design, as described by Bhattacharya (2007) is as follows:

Assume that the population is divided into $s = 1, 2, \ldots, S$ strata. The $s^{\text{th}}$ stratum is divided into $H_s$ clusters. The clusters of the $s^{\text{th}}$ stratum are labelled by $c_s = 1, \ldots, H_s$. Under the $c_s^{\text{th}}$ cluster in stratum $s$, there is a group of $M_{sc_s}$ households with $\nu_{sc_sh}$ individuals or members. Therefore, the total number of clusters in the population is $H = \sum_{s=1}^{S} H_s$. The number of households in a stratum will be $M_s = \sum_{c_s=1}^{H_s} M_{sc_s}$ and the total number of households in the population will be $M = \sum_{s=1}^{S} M_s = \sum_{s=1}^{S}\sum_{c_s=1}^{H_s} M_{sc_s}$.

[1] This chapter is based on Bilson Darku, Konietschke, and Chattopadhyay (2018) which has been submitted for publication

From the population, a sample of $n_s$ clusters is selected from the $s^{\text{th}}$ stratum by simple random sampling with replacement. A sample of $k$ households is then taken from each of the selected clusters and indexed by $h = 1, 2, \ldots, k$. Let the total number of clusters selected from the population be

$$n = \sum_{s=1}^{S} n_s \quad \text{with} \quad n_s = a_s n \quad \text{and} \quad a_s = \frac{H_s}{H}. \tag{4.2}$$

Thus, the total number of households in the sample will be $nk = k\sum_{s=1}^{S} n_s$. For the $h^{\text{th}}$ household in the $c_s^{\text{th}}$ cluster from the $s^{\text{th}}$ stratum, the observed data (that is, household monthly income, monthly expenditure, per capita income or others) is denoted as $x_{sc_sh}$ and assigned with a weight $W_{sc_sh}$. With the presence of stratification and clustering, the households are assigned with different weights because the probability of inclusion in the sample will be different. The assigned weight is computed as the inverse of the probability of inclusion in the sample (see Binder and Kovacevic, 1995; Horvitz and Thompson, 1952; Lee and Forthofer, 2006). For the given survey framework, weights are assigned to the data $(x_{sc_sh})$ with respect to the number of observations in the population. The attached weight is given by

$$W_{sc_sh} = \frac{M_{sc_sh} H_s}{k n_s} \nu_{sc_sh}, \tag{4.3}$$

and, for computational purposes, standardized as

$$w_{sc_sh} = \frac{W_{sc_sh}}{\sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} W_{sc_sh}}. \tag{4.4}$$

Under the above framework, a consistent estimator of the Gini index for such a population with $n$ number of clusters, is given by

$$\hat{G}_n = 1 - \frac{2}{\hat{\mu}} \sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} w_{sc_sh} x_{sc_sh} \left(1 - \hat{F}\left(x_{sc_sh}\right)\right), \tag{4.5}$$

where the weighted average income ($\hat{\mu}$) and the empirical distribution ($\hat{F}(\cdot)$) of income are given by

$$\hat{\mu} = \sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} w_{sc_sh} x_{sc_sh} \quad \text{and} \tag{4.6}$$

$$\hat{F}\left(x_{sc_sh}\right) = \sum_{i=1}^{S} \sum_{j=1}^{n_s} \sum_{l=1}^{k} w_{ijl} I_{\left[x_{ijl} \leq x_{sc_sh}\right]}. \tag{4.7}$$

respectively (see Bhattacharya, 2007).

In this work, we seek to find a confidence interval for the Gini index in which both the confidence level $(1-\alpha)$ and the upper bound on the length of the confidence interval are pre-specified. It is known that the precision of a confidence interval is given by its width, and the accuracy is given by the confidence level. For a fixed sample size, if the width of a confidence interval with a pre-specified confidence level increases, the precision of the estimate decreases. So, in an effort to get a more precise confidence interval estimate of the Gini index, a narrow confidence interval is preferred. This work deals with constructing a sufficiently narrow $100(1-\alpha)\%$ confidence interval for the Gini index when the data come from stratified and clustered household surveys with large number of clusters per stratum.

Using Binder and Kovacevic (1995) and Bhattacharya (2007), one may construct confidence intervals for the Gini index under complex survey designs, but it cannot be used to find a sufficiently narrow $100(1-\alpha)\%$ confidence interval for the population Gini index. This can only be done by using at least a two stage sequential procedure. These procedures fall in the domain of sequential analysis. In contrast to other procedures, sequential procedures do not require fixing the sample size (cluster size, here) in advance. However, sampling is done in stages and the analysis is performed at each stage of the procedure until a pre-defined stopping rule is met. We refer readers to Ghosh and Sen (1991), Ghosh et al. (1997), Mukhopadhyay and De Silva (2009), Chattopadhyay and Kelley (2017), Kelley et al. (2018) and others for more on the sequential analysis literature.

We are not the first to advocate the use of sequential analysis in economics. Several economics and econometrics journal articles pursued the idea of sequential analysis earlier. Stein (1945, 1949) first proposed a two-stage procedure which aimed at estimating the normal mean when the population variance was unknown. This was the seminal work which gave the mathematical formulations of sequential analysis after Mahalanobis (1940) described the design and implementation of multi-stage sampling methodologies in large-scale surveys in India. Kanninen (1993), Greene (1998), Arcidiacono and Jones (2003), Aguirregabiria and Mira (2007), and many others are some of the authors that have contributed to the development of sequential analysis. Recently, Chattopadhyay and De (2016) and De and Chattopadhyay (2017) developed a sequential procedure for inference problems related to the Gini index under independent and identically distributed (i.i.d.) conditions, but the proposed methodology cannot be used for finding a sufficiently narrow $100(1-\alpha)\%$ confidence interval for the population Gini index under complex household survey designs.

The next subsection discusses the contribution that this work adds to the existing literature in statistical inference and economics.

### 4.1.1 Contributions of This Paper

Several authors developed procedures for inference problems related to the Gini index under the framework of i.i.d. random variables and complex survey designs. Examples include the manuscripts of Beach and Davidson (1983), Davidson (2009), Davidson and Duclos (2000), Gastwirth (1972), Bhattacharya (2007), Xu (2007), Chattopadhyay and De (2016), (De and Chattopadhyay, 2017) and others. However, none of these methods can be used to find a sufficiently narrow $100(1-\alpha)\%$ confidence interval for the population Gini index under complex household survey designs. We propose a two stage procedure and a purely sequential procedure to find an estimate of the optimal number of clusters which is required to find the sufficiently narrow confidence interval under a distribution-free scenario. Both the two-stage and purely sequential procedures are applied to the 64th round of household survey data collected in India. Further, a simulation study is carried out on observations collected in the Indian household survey data and from known income distributions to explore the properties of the procedures.

The remainder of this paper is organized as follows. In Section 4.2, the problem of finding a sufficiently narrow confidence interval for the Gini index and the reason for non-applicability of a procedure with fixed cluster size are formulated. In Section 4.3, the purely sequential, as well as the two-stage, procedure is developed followed by an application of both procedures on synthetic and real datasets in Section 4.4. The characteristics of the procedures are discussed in Section 4.5 with concluding comments provided in Section 4.8.

## 4.2 Problem Statement

Consider the complex household survey design described in Section 4.1, as originally proposed in Bhattacharya (2007). A consistent estimator of the population Gini index is $\hat{G}_n$, given in Equation (4.5). Using Bhattacharya (2007), under uniform consistency and weak convergence, if for every stratum $s$, $\mathrm{E}\left(|X|\right) < \infty$, then as $n_s \to \infty$ for each $s$ at the same rate,

$$\sqrt{n}\left(\hat{G} - G_F\right) \xrightarrow{D} N\left(0, \xi^2\right) \tag{4.8}$$

where $\xi^2$ is the asymptotic variance which can be found in Bhattacharya (2007). Thus, the $100(1-\alpha)\%$ confidence interval for $G_F$ is given by

$$\left(\hat{G}_n - z_{\alpha/2}\frac{\xi}{\sqrt{n}}, \hat{G}_n + z_{\alpha/2}\frac{\xi}{\sqrt{n}}\right) \tag{4.9}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)^{\text{th}}$ percentile of the standard normal distribution $N(0,1)$. The goal of this work is to construct a confidence interval for the population Gini index such that

$$P\left(\hat{G}_n - z_{\alpha/2}\frac{\xi}{\sqrt{n}} < G_F < \hat{G}_n + z_{\alpha/2}\frac{\xi}{\sqrt{n}}\right) \geq 1 - \alpha \tag{4.10}$$

and that the width of the confidence interval is no more than $\omega$ (a pre-specified value), that is,

$$2z_{\alpha/2}\frac{\xi}{\sqrt{n}} \leq \omega. \tag{4.11}$$

Based on the asymptotic normality distribution of $\hat{G}_n$,

$$P\left\{\left|\hat{G}_n - G_F\right| \leq \frac{\omega}{2}\right\} \approx 2\Phi\left(\frac{\omega\sqrt{n}}{2\xi}\right) - 1. \tag{4.12}$$

So, the coverage probability of such a confidence interval is approximately $(1-\alpha)$ if

$$\frac{\omega\sqrt{n}}{2\xi} \geq z_{\alpha/2} \implies n \geq \frac{4z_{\alpha/2}^2\xi^2}{\omega^2} \equiv C. \tag{4.13}$$

Therefore to achieve a confidence interval with width being at most $\omega$ and coverage probability approximately $100(1-\alpha)\%$, the required optimal number of clusters to be sampled from the $s^{\text{th}}$ stratum $(s = 1, 2, \ldots, S)$ will be $C_s = Ca_s$. Here, $C$ is the optimal total number of clusters from all strata, and $a_s$ is known and given in Equation (4.2). Thus if $C$ is known, that is $\xi^2$ is known, one can find the sufficiently narrow confidence interval by computing

$$\left(\hat{G}_C - z_{\alpha/z}\frac{\xi}{\sqrt{C}}, \hat{G}_C + z_{\alpha/z}\frac{\xi}{\sqrt{C}}\right) \tag{4.14}$$

which will satisfy Equation (4.13). However without knowing the underlying distribution of income (or assets or expenditure) within the population, the value of $\xi^2$ is unknown in practical scenarios. Thus, the optimal cluster size from all the $S$ strata, $C$, is also unknown. It is worth noting that the supposed value (or previous survey estimate) of $\xi^2$ may be used to obtain a value of $C$. However, a potential problem they may arise is that the supposed value of $\xi^2$ may be different from the actual value. Moreover, using previous survey estimates in many situations is not advised as they may not be applicable to the current population. This is because of a possible change in socio-economic conditions that may arise due to the change in distribution of income or expenditure as a result of change in economic policies or situations. Due to all these factors, the value of $C$ may widely differ from what it would have been if $\xi^2$ is known and will not guarantee that Equation (4.13) is satisfied. Since the total cluster size is unknown, one must use at least a two-stage procedure, which does not need a supposed value (or prior survey estimate) of $\xi^2$, to find out $C$ that will satisfy Equation (4.13). In this work, both two-stage and purely sequential procedures have been proposed to estimate the optimal cluster size and thereby ensure a sufficiently narrow $100(1-\alpha)\%$ confidence interval for the Gini index.

## 4.3 Sequential Methodology

This section describes the two-stage and purely sequential procedures which can be used to collect data so that we can find the sufficiently narrow confidence interval. Since, $\xi^2$ is unknown, an estimator of $\xi^2$ will first be discussed.

### 4.3.1 Estimation for $\xi^2$

Several articles published in statistics and economics journals have proposed different estimators of the asymptotic variance parameter of the Gini index under different sampling schemes. Readers are referred to Langel and Tillé (2013) for a discussion on several techniques used in estimating the asymptotic variance of the Gini index for various sampling designs. Under the current framework, Binder and Kovacevic (1995) proposed an estimator of $\xi^2$ for $n$ clusters as

$$V_{n,1}^2 = \sum_{s=1}^{S} \frac{n_s}{n_s - 1} \sum_{c_s=1}^{n_s} (u_{sc_s} - \bar{u}_s)^2 \tag{4.15}$$

where

$$u_{sc_s} = \frac{2}{\hat{\mu}} \sum_{h=1}^{k} w_{sc_sh} \left[ A(x_{sc_sh}) x_{sc_sh} + B(x_{sc_sh}) - \frac{\hat{\mu}}{2}(\hat{G}_n) + 1 \right], \tag{4.16}$$

$$\bar{u}_s = \frac{1}{n_s} \sum_{c_s=1}^{n_s} u_{sc_s}, \tag{4.17}$$

$$A(x_{sc_sh}) = \hat{F}(x_{sc_sh}) - \frac{\hat{G}_n + 1}{2}, \quad \text{and} \tag{4.18}$$

$$B(x_{sc_sh}) = \sum_{a=1}^{S} \sum_{b=1}^{n_s} \sum_{c=1}^{k} w_{abc} x_{abc} I(x_{abc} \geq x_{sc_sh}), \tag{4.19}$$

and another estimator of $\xi^2$ proposed by Bhattacharya (2007) is given by

$$V_{n,2}^2 = \sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} w_{sc_sh}^2 \hat{\psi}_{sc_sh}^2 + \sum_{s=1}^{S} \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} \sum_{h' \neq h} w_{sc_sh} \hat{\psi}_{sc_sh} w_{sc_sh'} \hat{\psi}_{sc_sh'}$$
$$- \sum_{s=1}^{S} \frac{1}{n_s} \left( \sum_{c_s=1}^{n_s} \sum_{h=1}^{k} w_{sc_sh} \hat{\psi}_{sc_sh} \right)^2 \tag{4.20}$$

where

$$\hat{\psi}_{sc_sh} = -\frac{2}{\mu} \sum_{g=1}^{kn} w_g \left[ x_{sc_sh} I(x_{sc_sh} < x_{(g)}) + x_{(g)}(\hat{F}(x_{(g)})) - I(x_{sc_sh} < x_{(g)}) \right]$$
$$+ \frac{2}{\hat{\mu}^2} \sum_{g=1}^{kn} \left[ \left\{ \sum_{a=1}^{S} \sum_{b=1}^{n_s} \sum_{c=1}^{k} w_{abc} x_{abc} I(x_{abc} < x_{(g)}) \right\} x_{sc_sh} \right], \tag{4.21}$$

$$kn = k \sum_{s=1}^{S} n_s \quad \text{i.e total number of observations,} \tag{4.22}$$

$$x_{(g)} \quad \text{is the } g^{\text{th}} \text{ ordered observation (among all } x_{sc_sh}). \tag{4.23}$$

Recently, Hoque and Clarke (2015) showed that the estimators of $\xi^2$ by Binder and Kovacevic (1995) and Bhattacharya (2007) are actually the same, that is $V_{n,1}^2 = V_{n,2}^2$ $(= V_n^2,$ say). However, the estimator proposed by (Binder and Kovacevic, 1995) is computationally more efficient. So, in this work, we use the plug-in estimator $V_n^2$ given in Equation (4.15) (and originally proposed by Binder and Kovacevic (1995)) as an estimator of $\xi^2$. Next, we describe a purely sequential procedure for constructing a sufficiently narrow confidence interval for the Gini index.

### 4.3.2 Purely Sequential Procedure

Recall that at least $C_s$ clusters from the $s^{\text{th}}$ stratum $(s = 1, 2, ..., S)$ are needed to achieve the desired confidence interval. First, a pilot cluster size of $t_s$ from each stratum $s$ is chosen. Momentarily, the value of the pilot cluster size will be discussed. Therefore, the total number of clusters in the pilot stage will be $t = \sum_{s=1}^{S} t_s$. Within each selected cluster, there are $k$ randomly selected households. Now, pilot observations $x_{s11}, \ldots, x_{s1k}, \ldots, x_{st_s1}, \ldots, x_{st_sk}$ with $s = 1, \ldots, S$ which represent per capita monthly expenditure (or any other data) from $k$ households from $t$ clusters belonging to all $S$ strata are collected.

Based on the pilot cluster size $t$, the estimate of $\xi^2$ is computed to examine a pre-defined condition in a *stopping rule*. A stopping rule indicates, after every stage, whether further sampling of cluster(s) is(are) required or to be stopped. So, at a particular stage, if the condition in the stopping rule is not satisfied, the surveyor collects data from additional $m'(\geq 1)$ clusters, with $k$ randomly chosen households, from each stratum that has $n_s \leq \hat{C}a_s$, where $\hat{C} = \frac{4z^2_{\alpha/2}}{\omega^2}\left(V_n^2 + \frac{1}{n}\right)$. Then $\xi^2$ is estimated based on all the observations collected up to that stage, and the stopping condition is checked. This process is repeated until the condition in the stopping rule is satisfied. It should be noted that $m'(\geq 1)$ can be any integer that is appropriate, suitable or feasible for the survey. Based on the sequential process, the stopping rule can be defined as:

$$\begin{aligned} &N \equiv N_\omega(\leq H) \text{ is the smallest integer } n(\geq t) \text{ such that} \\ &n \geq \frac{4z^2_{\alpha/2}}{\omega^2}\left(V_n^2 + \frac{1}{n}\right) = \hat{C} \quad \text{and} \quad n_s \geq \hat{C}_s = C\hat{a}_s, \forall s. \end{aligned} \tag{4.24}$$

The term $1/n$ is a correction term incorporated to avoid early stopping of the sequential procedure as $V_n^2$ (the estimator of $\xi^2$) may be very small. However, as the cluster size $n$ becomes large, $V_n^2 + 1/n$ converges to $\xi^2$ since $V_n^2$ is a consistent estimator of $\xi^2$. The final sample size $N$ constitutes $N_s$ clusters from each stratum $s$ where

$$N_s = Na_s, \text{for } s = 1, 2, \ldots, S. \tag{4.25}$$

Based on the sampled data $x_{sc_sh}$ and their corresponding standardized weight $w_{sc_sh}$ with $s = 1, \ldots, S$, $c_s = 1, \ldots, N_s$, and $h = 1, \ldots, k$, the $100(1-\alpha)\%$ confidence interval for the Gini index $G$, with width not larger than $\omega$, is given by

$$\left(\hat{G}_N - z_{\alpha/2}\frac{V_N}{\sqrt{N}}, \hat{G}_N + z_{\alpha/2}\frac{V_N}{\sqrt{N}}\right). \tag{4.26}$$

### 4.3.3 Two-stage Procedure

Unlike the purely sequential procedure, the two-stage procedure comprises of two stages. The first stage is called the pilot stage, wherein a sample is drawn from the population. That is, first a pilot sample of clusters, $t_s$, is selected from each stratum $s$. Based on the sample from the pilot stage, $\xi^2$ is computed using the estimator given in Equation (4.15). Then the total final cluster size from all strata can be estimated by

$$Q = \min\left\{H, \max\left\{t, \left\lceil \frac{4z_{\alpha/2}^2}{\omega^2} V_t^2 \right\rceil \right\}\right\} = \min\left\{H, Q^*\right\} \tag{4.27}$$

where $Q^*$ is the (unbounded) optimal sample size and $\lceil\cdot\rceil$ is the ceiling function, that is, $\lceil x\rceil$ is the smallest integer that is greater than or equal to $x$. Thus, the estimated number of clusters to be sampled from the $s^{\text{th}}$ stratum is given by

$$Q_s = \min\{H_s, [Qa_s]\}, \tag{4.28}$$

with $a_s$ as defined in Equation (4.2) and $[\cdot]$ being the nearest integer function. So, in the second stage, observations from $k$ households will be collected from $Q_s - t_s$ clusters from each stratum $s$. Using the combined data from the two stages, the estimate of $\xi^2$ is updated and the approximate $100(1-\alpha)\%$ confidence interval for the Gini index is given by

$$\left(\hat{G}_Q - z_{\alpha/2}\frac{V_Q}{\sqrt{Q}}, \hat{G}_Q + z_{\alpha/2}\frac{V_Q}{\sqrt{Q}}\right). \tag{4.29}$$

It can be noted that the final cluster size using either the two-stage procedure or the purely sequential procedure can be shown to be always finite. The proof is straight forward and the details are left out for brevity. In the next subsection, the pilot cluster size formula will be derived.

### 4.3.4 Pilot Sample Size

Using Equation (4.24),

$$n \geq \frac{4z_{\alpha/2}^2}{\omega^2}\left(V_n^2 + \frac{1}{n}\right) \geq \frac{4z_{\alpha/2}^2}{\omega^2}\frac{1}{n} \implies n \geq \frac{2z_{\alpha/2}}{\omega}. \tag{4.30}$$

Thus the total number of sampled clusters is at least $2z_{\alpha/2}/\omega$. The maximum number of clusters from the $s^{\text{th}}$ stratum is $H_s$ and also the minimum number of clusters to be sampled is 2. Considering all the constraints in Equation (4.24), the number of clusters to be sampled from the $s^{\text{th}}$ stratum at the pilot stage is

$$t_s = \max\left\{2, \min\left\{H_s, \left\lceil \frac{2a_s z_{\alpha/2}}{\omega} \right\rceil\right\}\right\}. \tag{4.31}$$

This ensures that the minimum cluster size is met as well as the total possible cluster size is not exceeded.

## 4.4 Application

We now apply the sequential procedures to construct confidence intervals for the Gini index using the per capita monthly expenditures data collected in the 64th Round National Sample Survey (NSS). The 64th NSS was a stratified multi-staged survey design running from between July 2007 till June 2008 and covered almost the whole of India.[2] As of 2008, the country was divided into 28 states and 7 union territories, and each is subdivided into districts. Within each district, two basic sectors were formed; all rural areas constituted the rural sector while all urban areas constituted the urban sector. Nonetheless, for the urban areas in a district, separate basic strata were formed for each town that had at least a population of 10 lakhs[3]

[2] "The survey excluded (i) Leh (Ladakh) and Kargil districts of Jammu & Kashmir (for central sample), (ii) interior villages of Nagaland situated beyond 5km of the bus route and (ii) villages of Andaman and Nicobar Islands which remain inaccessible throughout the year."(National Sample Survey Organization, 2007)

[3] 1 lakh is 100,000

(1 million) as at 2001 population census and remaining areas were grouped as another basic stratum (National Sample Survey Organization, 2007). For the rural sector, the sampling frame was made up of villages while for the urban sector, it was towns/blocks.

Census villages and the Urban Frame Survey (UFS) blocks are the first stage units (FSU) in the rural and urban sectors respectively. From each strata, FSUs are selected with replacement from the rural sector with probability proportional to the size and without replacement from the urban sector by using simple random sampling. Within the FSU, the households in each sector were considered as the smallest unit of grouping, which is also referred to as the Ultimate Stage Units (USU). Households were selected by simple random sampling without replacement and various information about the households were recorded during the survey. Some of the information include the demographics, household size, expenditure on education, food, clothing, etc. In order to make inferences from the data, the information from the households were weighted. A detailed description of the NSS Data can be found at `http://164.100.34.62/index.php/catalog/15`.

This work considers only the per capita monthly expenditure for the households. The "Stratum" variable in the 64$^{\text{th}}$ NSS dataset will be used to stratify the states/sectors while "FSUno" (First Stage Unit Number) variable will be used to cluster the households under each stratum. We discuss the results obtained from applying the proposed sequential methodologies which were applied to the data collected from two of the most populous states in India, namely Uttar Pradesh and West Bengal. Additionally, the report includes the results for the whole state as well as rural and urban sectors of the state. Here, all the households in each cluster were considered since we are sampling from a survey that already has few number of households per cluster. However, the weight per household was adjusted at each sampling stage. The per capita monthly expenditure (with their adjusted weights) are used to estimate $\xi$.

In applying the sequential methodologies, the pilot sample sizes $t_s$ for each stratum $s$ are computed using Equation (4.31). At the outset, $t_s$ number of clusters are selected from stratum $s$ for $s = 1, \ldots, S$. Where $t_s$ is same for both the purely sequential procedure and the two-stage methodology. We apply each of the procedures considering the survey data as our population.

### 4.4.1 Application of Purely Sequential Procedure

The proposed purely sequential procedure, with observations from one cluster collected at each stage after the pilot stage, is applied to the NSS $64^{\text{th}}$ round data. The results for different combinations of pre-specified width ($\omega \in \{0.020, 0.025\}$) and confidence level ($1-\alpha, \alpha \in \{0.05, 0.10\}$) can be found in Tables 4.1 – 4.2. The third column of the tables shows the collected cluster size $N$ using the stopping rule in Equation (4.24) and the pilot sample size $t$. The fourth column shows the value of $\hat{C}$ when the procedure ended, $\hat{C}$ is the estimated optimal sample size as in Equation (4.24). $\hat{G}_N$ and $se(\hat{G}_N)$ are the estimated Gini index and its standard error respectively based on $N$ clusters. The lower and upper limits of the confidence intervals obtained with the stopping rule in Equation (4.24) are also reported. Furthermore, column $w_N$ is the estimated width of the confidence interval, and $p(N_s < \hat{C}_s)$ shows the proportion of strata that had their collected cluster size $N_s$ from the purely sequential procedure being less than their estimated optimal cluster size $\hat{C}_s$ ($N_s$ is the final number of clusters selected from stratum $s$ while $\hat{C}_s$ is the estimated optimal number of clusters to be sampled from stratum $s$).

Note that the stopping rule was not met under the urban sectors in both states. This is because all strata do not have enough clusters (that is, $p(N_s < \hat{C}_s) = 1$). However, in the other cases, even though $N > \hat{C}$, some strata had $N_s < \hat{C}_s$. This is because some strata had more than enough clusters while others did not. For example, it can be seen from Table 4.1 that in the rural sector of Uttar Pradesh, 40% of the strata did not have enough clusters even though at the end the confidence interval was 0.0186 wide which was less than the desired width of 0.02.

Table 4.1. Application results for PSP on NSS $64^{\text{th}}$ Round Data for $\omega = 0.02$

| Region | $\hat{G}_H$ | $H$ | $\hat{C}$ | $N$ | $\hat{G}_N$ | Lower CI | Upper CI | $w_N$ | $p(N_s < \hat{C}_s)$ |
|---|---|---|---|---|---|---|---|---|---|
| | $se(\hat{G}_H)$ | | | $(t)$ | $se(\hat{G}_N)$ | | | | |
| | | | | $\alpha = 0.10$ | | | | | |
| *Uttar Pradesh* | | | | | | | | | |
| All | 0.2163 | 1262 | 622 | 672 | 0.2116 | 0.2023 | 0.2209 | 0.0186 | 0.2138 |
| | (0.0042) | | | (321) | (0.0057) | | | | |
| Rural | 0.1997 | 903 | 505 | 523 | 0.2024 | 0.1931 | 0.2117 | 0.0186 | 0.4000 |
| | (0.0041) | | | (198) | (0.0057) | | | | |
| Urban | 0.2229 | 359 | 903 | 359 | 0.2229 | 0.2077 | 0.2381 | 0.0304 | 1.0000 |
| | (0.0092) | | | (180) | (0.0092) | | | | |
| *West Bengal* | | | | | | | | | |
| All | 0.2320 | 878 | 587 | 593 | 0.2334 | 0.2239 | 0.2430 | 0.0191 | 0.1282 |
| | (0.0051) | | | (190) | (0.0058) | | | | |
| Rural | 0.1812 | 551 | 450 | 450 | 0.1816 | 0.1723 | 0.1909 | 0.0186 | 0.2353 |
| | (0.0048) | | | (172) | (0.0057) | | | | |
| Urban | 0.2609 | 327 | 612 | 327 | 0.2609 | 0.2482 | 0.2736 | 0.0254 | 1.0000 |
| | (0.0077) | | | (185) | (0.0077) | | | | |
| | | | | $\alpha = 0.05$ | | | | | |
| *Uttar Pradesh* | | | | | | | | | |
| All | 0.2163 | 1262 | 834 | 878 | 0.2117 | 0.2022 | 0.2212 | 0.0190 | 0.2138 |
| | (0.0042) | | | (333) | (0.0048) | | | | |
| Rural | 0.1997 | 903 | 643 | 667 | 0.2024 | 0.1930 | 0.2117 | 0.0187 | 0.4000 |
| | (0.0041) | | | (226) | (0.0048) | | | | |
| Urban | 0.2229 | 359 | 1282 | 359 | 0.2229 | 0.2048 | 0.2410 | 0.0362 | 1.0000 |
| | (0.0092) | | | (254) | (0.0092) | | | | |
| *West Bengal* | | | | | | | | | |
| All | 0.2320 | 878 | 906 | 878 | 0.2320 | 0.2221 | 0.2419 | 0.0198 | 1.0000 |
| | (0.0051) | | | (223) | (0.0051) | | | | |
| Rural | 0.1812 | 551 | 552 | 551 | 0.1812 | 0.1719 | 0.1906 | 0.0187 | 1.0000 |
| | (0.0048) | | | (203) | (0.0048) | | | | |
| Urban | 0.2609 | 327 | 869 | 327 | 0.2609 | 0.2458 | 0.2761 | 0.0303 | 1.0000 |
| | (0.0077) | | | (207) | (0.0077) | | | | |

### 4.4.2 Application of Two-Stage Procedure

Using the estimate for $\xi^2$ in Equation (**??**) obtained from the pilot stage, the final sample size $Q^*$ is computed using Equation (4.27). $Q^*$ is then adjusted to account for the limited availability of clusters per stratum in the NSS data to obtain the possible number of clusters $Q$ that can be sampled (Equation (4.27)). Here, $Q$ is distributed over $S$ strata as $Q_s$ for

Table 4.2. Application results for PSP on NSS 64th Round Data for $\omega = 0.025$

| Region | $\hat{G}_H$ | $H$ | $\hat{C}$ | $N$ | $\hat{G}_N$ | Lower CI | Upper CI | $w_N$ | $p(N_s < \hat{C}_s)$ |
|---|---|---|---|---|---|---|---|---|---|
| | $se(\hat{G}_H)$ | | | $(t)$ | $se(\hat{G}_N)$ | | | | |
| | | | | $\alpha = 0.10$ | | | | | |
| *Uttar Pradesh* | | | | | | | | | |
| All | 0.2163 | 1262 | 401 | 540 | 0.2138 | 0.2035 | 0.2242 | 0.0207 | 0.0000 |
| | (0.0042) | | | (302) | (0.0063) | | | | |
| Rural | 0.1997 | 903 | 386 | 400 | 0.2014 | 0.1899 | 0.2130 | 0.0231 | 0.1714 |
| | (0.0041) | | | (168) | (0.0070) | | | | |
| Urban | 0.2229 | 359 | 578 | 359 | 0.2229 | 0.2077 | 0.2381 | 0.0304 | 1.0000 |
| | (0.0092) | | | (168) | (0.0092) | | | | |
| *West Bengal* | | | | | | | | | |
| All | 0.2320 | 878 | 324 | 319 | 0.2288 | 0.2175 | 0.2401 | 0.0226 | 0.1795 |
| | (0.0051) | | | (158) | (0.0069) | | | | |
| Rural | 0.1812 | 551 | 276 | 289 | 0.1829 | 0.1721 | 0.1937 | 0.0216 | 0.2353 |
| | (0.0048) | | | (138) | (0.0066) | | | | |
| Urban | 0.2609 | 327 | 392 | 327 | 0.2609 | 0.2482 | 0.2736 | 0.0254 | 1.0000 |
| | (0.0077) | | | (142) | (0.0077) | | | | |
| | | | | $\alpha = 0.05$ | | | | | |
| *Uttar Pradesh* | | | | | | | | | |
| All | 0.2163 | 1262 | 572 | 653 | 0.2123 | 0.2010 | 0.2236 | 0.0226 | 0.2138 |
| | (0.0042) | | | (728) | (0.0058) | | | | |
| Rural | 0.1997 | 903 | 496 | 510 | 0.2010 | 0.1893 | 0.2128 | 0.0234 | 0.1714 |
| | (0.0041) | | | (197) | (0.0060) | | | | |
| Urban | 0.2229 | 359 | 821 | 359 | 0.2229 | 0.2048 | 0.2410 | 0.0362 | 1.0000 |
| | (0.0092) | | | (717) | (0.0092) | | | | |
| *West Bengal* | | | | | | | | | |
| All | 0.2320 | 878 | 517 | 519 | 0.2318 | 0.2199 | 0.2437 | 0.0238 | 0.1538 |
| | (0.0051) | | | (186) | (0.0061) | | | | |
| Rural | 0.1812 | 551 | 351 | 352 | 0.1815 | 0.1703 | 0.1927 | 0.0223 | 0.2353 |
| | (0.0048) | | | (163) | (0.0057) | | | | |
| Urban | 0.2609 | 327 | 556 | 327 | 0.2609 | 0.2458 | 0.2761 | 0.0303 | 1.0000 |
| | (0.0077) | | | (162) | (0.0077) | | | | |

stratum $s$; rounding off if $Q_s$ is not an integer. The sum of $Q_s$ gives the actual number of clusters, $\tilde{Q} = \sum_{s=1}^{S} Q_s$, that were sampled from all strata. Using $\tilde{Q}$ clusters, the Gini index and $\xi^2$ are re-estimated (or updated) and a $100(1-\alpha)\%$ confidence interval is constructed according to Equation (4.29).

Similar to the application of the purely sequential procedure, the two-stage procedure is applied to the NSS 64th round data for different combinations of pre-specified precision ($\omega$)

and accuracy $(1-\alpha)$ with the results shown in Tables **??** – 4.4.The second column of the tables indicates the total number of clusters $H$ in the unit (i.e. the whole state, rural or urban sector) of the NSS data. The third column displays estimated optimal number of cluster $(Q^*)$ that are required in order to achieve the desired precision and accuracy. Below $Q^*$ is the pilot number of clusters $t$. The next column shows the estimated optimal sample sizes $Q$ taking into account the total number of clusters available in the data, that is, the number of clusters are finite and limited. $\tilde{Q}$ is the actual number of clusters that can be sampled from all strata considering the fact that we can only sample integer number of clusters from each strata (i.e. rounding off where there are decimals in the number of clusters to be sampled from a stratum). Using Equation (4.5) and (4.15), the Gini index estimate, $\hat{G}_H$, for the unit is computed using all $H$ clusters with its standard error as $se(\hat{G}_H)$. The selected clusters are used to estimate the Gini index and this is denoted as $\hat{G}_{\tilde{Q}}$, with it standard error as $se(\hat{G}_{\tilde{Q}})$. Lower CI and Upper CI are the lower and upper limits of the $100(1-\alpha)\%$ confidence interval of the Gini index using a sample of size $\tilde{Q}$, respectively. The last column shows the length of the confidence interval, $w_{\tilde{Q}}$. It must be noted that $Q^*$ is unbounded while on the other hand, $Q$ and $\tilde{Q}$ cannot exceed $H$. $\tilde{Q}$ can be less than, equal to, or greater than $Q$ depending on the rounding off. $Q^*$ will be equal to $Q$ if and only if $Q^*$ is less than or equal to $H$.

In Tables 4.3 and 4.4, it can be observed that in all cases, except for the urban sectors for both states, the confidence interval widths were less than $\omega$. These results were achieved because the optimal number of clusters required $(Q^*)$, according to the two-stage procedure, were less than the number available $(H)$. On the other hand, in both Uttar Pradesh and West Bengal, the estimated optimal cluster sizes $\hat{C}$ for the urban sector exceeded the available number of clusters in the data $H$. As a consequence of this, the confidence interval widths for the Gini index in the urban sectors were larger than the pre-specified bound, that is $w_{\tilde{Q}} > \omega$.

Table 4.3. Application results for the two-stage procedure on NSS 64th Round Data for $\omega = 0.02$

| Region | $H$ | $Q^*$ | $\tilde{Q}$ | $\hat{G}_H$ | $\hat{G}_{\tilde{Q}}$ | Lower CI | Upper CI | $w_{\tilde{Q}}$ |
|---|---|---|---|---|---|---|---|---|
| | | $(t)$ | $(Q)$ | $(se(\hat{G}_H))$ | $(se(\hat{G}_{\tilde{Q}}))$ | | | |
| | | | | $\alpha = 0.10$ | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 1262 | 1146 | 1171 | 0.2163 | 0.2137 | 0.2072 | 0.2203 | 0.0131 |
| | | (321) | (1146) | (0.0042) | (0.0040) | | | |
| Rural | 903 | 398 | 406 | 0.1997 | 0.2027 | 0.1940 | 0.2114 | 0.0174 |
| | | (198) | (398) | (0.0041) | (0.0053) | | | |
| Urban | 359 | 1177 | 359 | 0.2229 | 0.2229 | 0.2077 | 0.2381 | 0.0304 |
| | | (180) | (359) | (0.0092) | (0.0092) | | | |
| *West Bengal* | | | | | | | | |
| All | 878 | 624 | 626 | 0.2320 | 0.2307 | 0.2216 | 0.2398 | 0.0182 |
| | | (190) | (624) | (0.0051) | (0.0055) | | | |
| Rural | 551 | 422 | 420 | 0.1812 | 0.1785 | 0.1707 | 0.1862 | 0.0155 |
| | | (173) | (422) | (0.0048) | (0.0047) | | | |
| Urban | 327 | 857 | 327 | 0.2609 | 0.2609 | 0.2482 | 0.2736 | 0.0254 |
| | | (185) | (327) | (0.0077) | (0.0077) | | | |
| | | | | $\alpha = 0.05$ | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 1262 | 1665 | 1262 | 0.2163 | 0.2163 | 0.2081 | 0.2245 | 0.0164 |
| | | (333) | (1262) | (0.0042) | (0.0042) | | | |
| Rural | 903 | 593 | 595 | 0.1997 | 0.2000 | 0.1914 | 0.2085 | 0.0171 |
| | | (226) | (593) | (0.0041) | (0.0044) | | | |
| Urban | 359 | 1712 | 359 | 0.2229 | 0.2229 | 0.2048 | 0.2410 | 0.0362 |
| | | (254) | (359) | (0.0092) | (0.0092) | | | |
| *West Bengal* | | | | | | | | |
| All | 878 | 874 | 878 | 0.2320 | 0.2320 | 0.2221 | 0.2419 | 0.0198 |
| | | (223) | (874) | (0.0051) | (0.0051) | | | |
| Rural | 551 | 535 | 534 | 0.1812 | 0.1814 | 0.1719 | 0.1910 | 0.0191 |
| | | (203) | (535) | (0.0048) | (0.0049) | | | |
| Urban | 327 | 1110 | 327 | 0.2609 | 0.2609 | 0.2458 | 0.2761 | 0.0303 |
| | | (207) | (327) | (0.0077) | (0.0077) | | | |

Table 4.4. Application results for the two-stage procedure on NSS $64^{\text{th}}$ Round Data for $\omega = 0.025$

| Region | $H$ | $Q^*$ | $\tilde{Q}$ | $\hat{G}_H$ | $\hat{G}_{\tilde{Q}}$ | Lower CI | Upper CI | $w_{\tilde{Q}}$ |
|---|---|---|---|---|---|---|---|---|
| | | $(t)$ | $(Q)$ | $(se(\hat{G}_H))$ | $(se(\hat{G}_{\tilde{Q}}))$ | | | |
| $\alpha = 0.10$ | | | | | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 1262 | 688 | 680 | 0.2163 | 0.2104 | 0.2023 | 0.2185 | 0.0162 |
| | | (302) | (688) | (0.0042) | (0.0049) | | | |
| Rural | 903 | 299 | 308 | 0.1997 | 0.2026 | 0.1927 | 0.2126 | 0.0199 |
| | | (168) | (299) | (0.0041) | (0.0061) | | | |
| Urban | 359 | 1087 | 359 | 0.2229 | 0.2229 | 0.2077 | 0.2381 | 0.0304 |
| | | (168) | (359) | (0.0092) | (0.0092) | | | |
| *West Bengal* | | | | | | | | |
| All | 878 | 396 | 396 | 0.2320 | 0.2293 | 0.2171 | 0.2414 | 0.0243 |
| | | (158) | (396) | (0.0051) | (0.0074) | | | |
| Rural | 551 | 275 | 275 | 0.1812 | 0.1750 | 0.1660 | 0.1840 | 0.0180 |
| | | (138) | (275) | (0.0048) | (0.0055) | | | |
| Urban | 327 | 582 | 327 | 0.2609 | 0.2609 | 0.2482 | 0.2736 | 0.0254 |
| | | (142) | (327) | (0.0077) | (0.0077) | | | |
| $\alpha = 0.05$ | | | | | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 1262 | 976 | 947 | 0.2163 | 0.2124 | 0.2041 | 0.2207 | 0.0166 |
| | | (302) | (946) | (0.0042) | (0.0042) | | | |
| Rural | 903 | 364 | 353 | 0.1997 | 0.2032 | 0.1922 | 0.2142 | 0.0220 |
| | | (197) | (364) | (0.0041) | (0.0056) | | | |
| Urban | 359 | 1081 | 359 | 0.2229 | 0.2229 | 0.2048 | 0.2410 | 0.0362 |
| | | (177) | (359) | (0.0092) | (0.0092) | | | |
| *West Bengal* | | | | | | | | |
| All | 878 | 607 | 608 | 0.2320 | 0.2315 | 0.2204 | 0.2427 | 0.0224 |
| | | (186) | (607) | (0.0051) | (0.0057) | | | |
| Rural | 551 | 391 | 392 | 0.1812 | 0.1759 | 0.1670 | 0.1849 | 0.0178 |
| | | (163) | (391) | (0.0048) | (0.0045) | | | |
| Urban | 327 | 754 | 327 | 0.2609 | 0.2609 | 0.2458 | 0.2761 | 0.0303 |
| | | (162) | (327) | (0.0077) | (0.0077) | | | |

## 4.5 Characteristics of the Procedures

The purely sequential procedure and the two-stage procedure for constructing a sufficiently narrow confidence interval for the Gini index - unlike fixed sample size procedures - require sample size which are obtained from data. So, the respective sample sizes $N$ and $Q$ are random in nature. The following theorem provides some asymptotic properties (as $\omega \to 0$) of the final sample sizes of the above procedures with sufficiently large $H$.

**Theorem 4.1.** *If the parent distribution(s) is(are) such that $E[V_n^2]$ exists and $H$ (fixed) is sufficiently large, then as $\omega \to 0$*

*(i)* $\dfrac{N}{C} \to 1$ *in probability,*

*(ii)* $\dfrac{Q}{C} \to 1$ *in probability, and*

*(iii)* $\dfrac{2z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} \leq \omega.$

*Proof.* (i) The definition of stopping rule $N$ associated with the purely sequential procedure in Equation (4.24) yields

$$\left(\frac{2z_{\alpha/2}}{\omega}\right)^2 V_N^2 \leq N \leq tI(N=t) + \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 \left(V_{N-1}^2 + (N-1)^{-1}\right). \tag{4.32}$$

Since $N \to \infty$ as $\omega \downarrow 0$ and $V_n^2 \to \xi^2$ in probability as $n \to \infty$, by applying theorem 2.1 of Gut (2009), $V_N^2 \to \xi^2$ in probability.

Also, $tP(N=t)/C \leq t/C \to 0$ as $\omega \downarrow 0$. Hence, dividing all sides of Equation (4.32) by $C$ and letting $\omega \downarrow 0$, we prove $N/C \to 1$ in probability as $\omega \downarrow 0$.

(ii) The definition of final sample size $Q$ related to the two-stage procedure in Equation (4.27) yields

$$\left(\frac{2z_{\alpha/2}}{\omega}\right)^2 V_t^2 \leq Q \leq tI(Q=t) + \left(\frac{2z_{\alpha/2}}{\omega}\right)^2 (V_t^2 + 1/t). \tag{4.33}$$

Also, $tP(Q=t)/C \leq t/C \to 0$ as $\omega \downarrow 0$. Now, $V_t^2 \to \xi^2$ in probability as $\omega \downarrow 0$. Hence, dividing all sides of Equation (4.33) by $C$ and letting $\omega \downarrow 0$, we prove $Q/C \to 1$ in probability as $\omega \downarrow 0$.

(iii) Using stopping rule $N$ in Equation (4.24) we have, for all $N$,

$$\begin{aligned}\left(\frac{2z_{\alpha/2}}{\omega}\right)^2 V_N^2 \leq N &\implies \frac{4z_{\alpha/2}^2}{N} V_N^2 \leq \omega^2 \\ &\implies 2z_{\alpha/2}\frac{V_N}{\sqrt{N}} \leq \omega\end{aligned}$$

■

Parts (i) and (ii) of the theorem shows that the final sample size as obtained from the purely sequential and the two-stage procedure is a consistent estimator of the sample size provided $\xi^2$ being known. Part (iii) of the theorem shows that the sufficiently smaller confidence interval (that is length less than $\omega$) will be obtained by the purely sequential procedure. The same result can never be proven for the two-stage procedure.

## 4.6 Replication Using Empirical Data

Next, we use simulation study to illustrate and compare the properties of our purely sequential and the two stage procedures in constructing a $100(1-\alpha)\%$ confidence interval for the Gini index whose width is less than $\omega$ under a complex survey. We presented two different simulation studies with 5000 as the simulation size - (a) simulation runs using the NSS survey data as the population and (b) a Monte Carlo simulation in which the observations are drawn from three populations, each of which has been drawn using three different distributions, namely; Pareto, Gamma and Lognormal distributions.

### 4.6.1 Replication Using Empirical Data: Purely Sequential Procedure (PSP)

To begin with, we describe the simulation procedure for the purely sequential methodology. From any of the given population mentioned previously, $t_s$ clusters are randomly sampled from the $s^{\text{th}}$ stratum without replacement. From there, four households are selected from each cluster using simple random sampling without replacement and these households from all $t$ clusters will constitute the pilot sample. From the collected pilot sample, the asymptotic variance of the Gini index $\xi^2$ is estimated using Equation (4.15), and from Equation (4.24), the optimal number of clusters $C$ is estimated. The stopping rule is checked and if it is satisfied, sampling is terminated. On the other hand, if the stopping rule is not satisfied, the strata whose number of clusters selected are less than the expected, that is $\{s : t_s < \hat{C}_s\}$, are identified and additional $m'$ number of clusters are randomly selected from them without replacement. Here, $m'$ is chosen to be either 1 or 10 or 20. In each of the selected $m'$ clusters, four households are randomly selected without replacement. At this stage, with the total number of sampled clusters being $n$, the value of $V_n^2$ is updated and the stopping rule is checked. If the rule is met, sampling is stopped, otherwise the strata without enough clusters are identified again and additional $m'$ clusters are collected from each of them. This process will continue until and unless the stopping rule is met. At that point, based on say $N$ number of clusters sampled from all strata, the $100(1-\alpha)\%$ confidence interval for the Gini index is constructed as given in Equation (4.26).

The purely sequential procedure was replicated 5000 times on the NSS data. The results of the simulation study of our purely sequential procedure are found in Tables 4.5 and 4.6. In the tables, the average optimal sample sizes $\bar{N}$ and their respective standard errors $se(N)$ are indicated in the fourth column. $p(w_N > \omega)$ indicates the proportion of confidence interval widths that exceeded the desired bound $\omega$, while $p(N < \hat{C})$ is the proportion of estimated optimal cluster sizes that could not meet the stopping rule.

Apart from the urban sectors, the average confidence interval width $\bar{w}_N$ are less than $\omega = 0.02$ and they have small standard errors. Also, none of the width exceeded the specified value of $\omega$.

### 4.6.2 Replication Using Empirical Data: Two-Stage procedure

Unlike the purely sequential procedure described above, the two-stage procedure has only two stages. The simulation algorithm for the two-stage is as follows. From a given popula-

Table 4.5. Replication results for PSP on 64th NSS Data ($\omega = 0.02$)

| Region | $\hat{G}_H$ | $H$ | $\overline{\hat{C}}$ | $\bar{N}$ | $\overline{\hat{G}}_N$ | $\bar{w}_N$ | $p(w_N > \omega)$ | $p(N < \hat{C})$ |
|---|---|---|---|---|---|---|---|---|
| | $se(\hat{G}_H)$ | $t$ | $se(\hat{C})$ | $se(N)$ | $se(\hat{G}_N)$ | $se(\bar{w}_N)$ | $se(p(w_N > \omega))$ | |
| | | | | $\alpha = 0.10$ | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 0.2163 | 1262 | 597.3592 | 637.7388 | 0.2160 | 0.0186 | 0.0000 | 0.0000 |
| | 0.0042 | 321 | 112.5162 | 117.8815 | 0.0042 | 0.0006 | 0.0000 | |
| Rural | 0.1997 | 903 | 431.5098 | 450.9542 | 0.1993 | 0.0180 | 0.0000 | 0.0000 |
| | 0.0041 | 198 | 77.5388 | 78.3839 | 0.0041 | 0.0007 | 0.0000 | |
| Urban | 0.2229 | 359 | 903.0000 | 359.0000 | 0.2229 | 0.0304 | 1.0000 | 1.0000 |
| | 0.0092 | 180 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |
| *West Bengal* | | | | | | | | |
| All | 0.2320 | 878 | 682.2386 | 691.2456 | 0.2315 | 0.0193 | 0.0000 | 0.0000 |
| | 0.0051 | 190 | 85.2230 | 85.8925 | 0.0036 | 0.0002 | 0.0000 | |
| Rural | 0.1812 | 551 | 391.9578 | 397.1726 | 0.1808 | 0.0179 | 0.0000 | 0.0000 |
| | 0.0048 | 172 | 51.0689 | 52.2322 | 0.0032 | 0.0006 | 0.0000 | |
| Urban | 0.2609 | 327 | 612.0000 | 327.0000 | 0.2609 | 0.0254 | 1.0000 | 1.0000 |
| | 0.0077 | 185 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |
| | | | | $\alpha = 0.05$ | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 0.2163 | 1262 | 836.5332 | 871.9902 | 0.2159 | 0.0190 | 0.0000 | 0.0000 |
| | 0.0042 | 321 | 111.8031 | 113.5042 | 0.0030 | 0.0004 | 0.0000 | |
| Rural | 0.1997 | 903 | 592.8814 | 612.3790 | 0.1992 | 0.0185 | 0.0000 | 0.0000 |
| | 0.0041 | 198 | 88.4435 | 89.3062 | 0.0029 | 0.0005 | 0.0000 | |
| Urban | 0.2229 | 359 | 1282.0000 | 359.0000 | 0.2229 | 0.0362 | 1.0000 | 1.0000 |
| | 0.0092 | 180 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |
| *West Bengal* | | | | | | | | |
| All | 0.2320 | 878 | 888.4526 | 866.3402 | 0.2316 | 0.0197 | 0.0000 | 0.8218 |
| | 0.0051 | 190 | 42.1969 | 31.2673 | 0.0013 | 0.0002 | 0.0000 | |
| Rural | 0.1812 | 551 | 531.0336 | 530.8254 | 0.1807 | 0.0186 | 0.0000 | 0.7352 |
| | 0.0048 | 172 | 42.5598 | 41.2625 | 0.0014 | 0.0003 | 0.0000 | |
| Urban | 0.2609 | 327 | 869.0000 | 327.0000 | 0.2609 | 0.0303 | 1.0000 | 1.0000 |
| | 0.0077 | 185 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |

tion, $t_s$ number of clusters are randomly selected without replacement from the $s^{\text{th}}$ stratum and four households are randomly sampled from each of selected clusters without replacement. The per monthly capita expenditure $x_{sc_sh}$ from the selected households, with their respective weight $W_{sc_sh}$, are used to estimate the asymptotic variance of the Gini index (from Equation (4.15)). This is followed by using Equation (4.27) to obtain the optimal number

Table 4.6. Replication results for PSP on 64th NSS Data ($\omega = 0.025$)

| Region | $\hat{G}_H$ | $H$ | $\overline{\hat{C}}$ | $\bar{N}$ | $\overline{\hat{G}}_N$ | $\bar{w}_N$ | $p(w_N > \omega)$ | $p(N < \hat{C})$ |
|---|---|---|---|---|---|---|---|---|
| | $se(\hat{G}_H)$ | $t$ | $se(\hat{C})$ | $se(N)$ | $se(\hat{G}_N)$ | $se(\bar{w}_N)$ | $se(p(w_N > \omega))$ | |
| | | | | $\alpha = 0.10$ | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 0.2163 | 1262 | 397.4636 | 446.5638 | 0.2162 | 0.0222 | 0.0000 | 0.0000 |
| | 0.0042 | 321 | 87.6632 | 86.6652 | 0.0058 | 0.0012 | 0.0000 | |
| Rural | 0.1997 | 903 | 290.1868 | 323.1846 | 0.1995 | 0.0211 | 0.0000 | 0.0000 |
| | 0.0041 | 168 | 61.1267 | 66.2810 | 0.0054 | 0.0012 | 0.0000 | |
| Urban | 0.2229 | 359 | 578.0000 | 359.0000 | 0.2229 | 0.0304 | 1.0000 | 1.0000 |
| | 0.0092 | 180 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |
| *West Bengal* | | | | | | | | |
| All | 0.2320 | 878 | 458.0644 | 468.1314 | 0.2315 | 0.0235 | 0.0000 | 0.0000 |
| | 0.0051 | 190 | 92.1715 | 93.6177 | 0.0057 | 0.0006 | 0.0000 | |
| Rural | 0.1812 | 551 | 268.8992 | 274.2660 | 0.1812 | 0.0213 | 0.0000 | 0.0000 |
| | 0.0048 | 172 | 44.5474 | 45.7673 | 0.0049 | 0.0013 | 0.0000 | |
| Urban | 0.2609 | 327 | 389.7278 | 326.3296 | 0.2607 | 0.0254 | 0.9760 | 0.9760 |
| | 0.0077 | 185 | 14.7339 | 5.0234 | 0.0005 | 0.1531 | 0.0022 | |
| | | | | $\alpha = 0.05$ | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 0.2163 | 1262 | 549.4374 | 588.3550 | 0.2161 | 0.0231 | 0.0000 | 0.0000 |
| | 0.0042 | 321 | 107.3979 | 110.8163 | 0.0046 | 0.0008 | 0.0000 | |
| Rural | 0.1997 | 903 | 399.6852 | 417.5508 | 0.1994 | 0.0223 | 0.0000 | 0.0000 |
| | 0.0041 | 198 | 78.7000 | 75.5232 | 0.0043 | 0.0010 | 0.0000 | |
| Urban | 0.2229 | 359 | 821.0000 | 359.0000 | 0.2229 | 0.0362 | 1.0000 | 1.0000 |
| | 0.0092 | 180 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |
| *West Bengal* | | | | | | | | |
| All | 0.2320 | 878 | 632.3722 | 641.8234 | 0.2315 | 0.0240 | 0.0000 | 0.0000 |
| | 0.0051 | 190 | 92.1456 | 93.1628 | 0.0040 | 0.0004 | 0.0000 | |
| Rural | 0.1812 | 551 | 361.3896 | 366.2682 | 0.1809 | 0.0222 | 0.0000 | 0.0000 |
| | 0.0048 | 172 | 50.9696 | 52.3207 | 0.0036 | 0.0009 | 0.0000 | |
| Urban | 0.2609 | 327 | 556.0000 | 327.0000 | 0.2609 | 0.0303 | 1.0000 | 1.0000 |
| | 0.0077 | 185 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | |

of clusters $Q$ needed to achieved the desired confidence level and width. If $Q > t$, additional $Q_s - t_s$ number of clusters are randomly selected without replacement from each stratum $s$. In each of the additional clusters, four households are also randomly selected without replacement. Finally, per capita monthly expenditure of all households from the $Q$ number of clusters are used to construct the $100(1-\alpha)\%$ confidence interval for the Gini index as stated in Equation (4.29).

Tables 4.7 to 4.8 show the results of the simulation study of our two-stage procedure. The third column shows the average estimated Gini Index and its standard error. The fourth through to the seventh columns show the average values of $Q^*$, $Q$, $\tilde{Q}$, and $w_{\tilde{Q}}$ respectively with their standard errors. The last column indicates the proportion of times that the length of the confidence intervals obtained ($w_{\tilde{Q}}$) exceeded the pre-specified upper bound $\omega$.

As the application of the two-stage procedure was repeated 5000 times, results show that the optimal cluster size is less than the number of clusters in the NSS Data on the average irrespective of the households that were in the pilot sample. However, the estimated optimal cluster sizes ($Q^*$) in the urban clusters always exceeded the available number of clusters irrespective of which clusters and households were initially sampled. This is clearly seen in the last column of Tables 4.7 and 4.8. The urban sectors always had confidence interval width more than desired due to inadequate number of clusters. Also, from the penultimate column, it could be seen that there are chances that this procedure results in a larger width.

The above described process for the two procedures are replicated 5000 times on the three synthetic populations. In each replicate, different clusters and households are used as a pilot sample.

### 4.7 Replication Using a Pseudo Population

First, a pseudo population is created considering the structure of the survey described in this paper. This population is subdivided into three strata ($S = 3$, i.e. $s = 1, 2, 3$) with each stratum having 600 clusters ($H_s = 600$ for $s = 1, 2, 3$). The clusters are further divided into households. In all, the total number of households $M$ is 900000, and each $s^{th}$ stratum having $M_s = 300000$ number of households. Each household is randomly assigned a household size $\nu_{sc_sh}$. To create three different synthetic populations, we allow the per capita

Table 4.7. Replication results for two-stage procedure on 64th NSS Data ($\omega = 0.02$)

| Region | $\hat{G}_H$ | $H$ | $\overline{\hat{G}}_{\tilde{Q}}$ | $\overline{Q^*}$ | $\overline{Q}$ | $\overline{\tilde{Q}}$ | $\overline{w}_{\tilde{Q}}$ | $p(w_{\tilde{Q}} > \omega)$ |
|---|---|---|---|---|---|---|---|---|
| | $se(\hat{G}_H)$ | $t$ | $se(\hat{G}_{\tilde{Q}})$ | $se(Q^*)$ | $se(Q)$ | $se(\tilde{Q})$ | $se(w_{\tilde{Q}})$ | $se(p(w_{\tilde{Q}} > \omega))$ |
| | | | | $\alpha = 0.10$ | | | | |
| *Utter Pradesh* | | | | | | | | |
| All | 0.2163 | 1262 | 0.2158 | 746.0006 | 736.7442 | 747.7408 | 0.0175 | 0.1940 |
| | 0.0042 | 321 | 0.0038 | 292.0440 | 270.7270 | 260.7141 | 0.0028 | 0.0059 |
| Rural | 0.1997 | 903 | 0.1993 | 521.3184 | 521.0958 | 521.2806 | 0.0173 | 0.2060 |
| | 0.0041 | 198 | 0.0038 | 173.9302 | 173.4148 | 173.5880 | 0.0029 | 0.0057 |
| Urban | 0.2229 | 359 | 0.2229 | 921.8320 | 358.9994 | 359.0000 | 0.0304 | 1.0000 |
| | 0.0092 | 180 | 0.0000 | 234.4798 | 0.0424 | 0.0000 | 0.0000 | 0.0000 |
| *West Bengal* | | | | | | | | |
| All | 0.2320 | 878 | 0.2317 | 783.1276 | 719.0094 | 719.0098 | 0.0184 | 0.2248 |
| | 0.0051 | 190 | 0.0029 | 240.1110 | 151.8134 | 152.2045 | 0.0021 | 0.0059 |
| Rural | 0.1812 | 551 | 0.1809 | 491.3992 | 459.2752 | 459.2734 | 0.0170 | 0.0706 |
| | 0.0048 | 173 | 0.0024 | 123.0809 | 78.5944 | 78.6132 | 0.0017 | 0.0036 |
| Urban | 0.2609 | 327 | 0.2609 | 696.5434 | 327.0000 | 327.0000 | 0.0254 | 1.0000 |
| | 0.0077 | 185 | 0.0000 | 113.4849 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | | | | $\alpha = 0.05$ | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 0.2163 | 1262 | 0.2160 | 1050.3342 | 938.7144 | 937.7854 | 0.0189 | 0.3372 |
| | 0.0042 | 333 | 0.0028 | 427.1341 | 253.3348 | 255.7325 | 0.0026 | 0.0067 |
| Rural | 0.1997 | 903 | 0.1995 | 737.4964 | 660.4210 | 660.4842 | 0.0185 | 0.3296 |
| | 0.0041 | 226 | 0.0027 | 280.1239 | 147.8075 | 147.9787 | 0.0026 | 0.0066 |
| Urban | 0.2229 | 359 | 0.2229 | 1438.5742 | 359.0000 | 359.0000 | 0.0362 | 1.0000 |
| | 0.0092 | 254 | 0.0000 | 324.5592 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| *West Bengal* | | | | | | | | |
| All | 0.2320 | 878 | 0.2319 | 1067.1052 | 833.7042 | 833.7296 | 0.0203 | 0.2620 |
| | 0.0051 | 223 | 0.0014 | 292.5594 | 78.2840 | 78.5031 | 0.0011 | 0.0062 |
| Rural | 0.1812 | 551 | 0.1811 | 666.8132 | 534.2016 | 534.2286 | 0.0189 | 0.1012 |
| | 0.0048 | 203 | 0.0010 | 160.5609 | 31.3909 | 31.4040 | 0.0007 | 0.0043 |
| Urban | 0.2609 | 327 | 0.2609 | 950.6524 | 327.0000 | 327.0000 | 0.0303 | 1.0000 |
| | 0.0077 | 207 | 0.0000 | 132.2993 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

monthly expenditure $x_{sc_sh}$ for the households to follow a chosen theoretical distribution. The distributions chosen were Pareto(scale=20000, shape=5), Lognormal(mean = 2.185, sd = 0.562) and Gamma(shape = 2.649, rate = 0.84). The choice of the parameter values of the distributions are the same as in Ransom and Cramer (1983), Chattopadhyay and De (2016), and De and Chattopadhyay (2017).

Table 4.8. Replication results for two-stage procedure on 64th NSS Data ($\omega = 0.025$)

| Region | $\hat{G}_H$ | $H$ | $\overline{\hat{G}_{\tilde{Q}}}$ | $\overline{Q^*}$ | $\overline{Q}$ | $\overline{\tilde{Q}}$ | $\overline{w}_{\tilde{Q}}$ | $p(w_{\tilde{Q}} > \omega)$ |
|---|---|---|---|---|---|---|---|---|
| | $se(\hat{G}_H)$ | $t$ | $se(\hat{G}_{\tilde{Q}})$ | $se(Q^*)$ | $se(Q)$ | $se(\tilde{Q})$ | $se(w_{\tilde{Q}})$ | $se(p(w_{\tilde{Q}} > \omega))$ |
| | | | | $\alpha = 0.10$ | | | | |
| *Uttar Pradesh* | | | | | | | | |
| All | 0.2163 | 1262 | 0.2157 | 511.2518 | 511.2518 | 540.5866 | 0.0199 | 0.0374 |
| | 0.0042 | 302 | 0.0052 | 174.7202 | 174.7202 | 151.6189 | 0.0026 | 0.0027 |
| Rural | 0.1997 | 903 | 0.1994 | 374.6068 | 374.5782 | 374.7222 | 0.0204 | 0.1462 |
| | 0.0041 | 168 | 0.0053 | 176.9589 | 176.8717 | 177.1875 | 0.0043 | 0.0050 |
| Urban | 0.2229 | 359 | 0.2228 | 658.1510 | 357.3200 | 357.6046 | 0.0304 | 0.9980 |
| | 0.0092 | 168 | 0.0009 | 227.5936 | 8.9762 | 9.9289 | 0.0005 | 0.0000 |
| *West Bengal* | | | | | | | | |
| All | 0.2320 | 878 | 0.2316 | 529.8012 | 527.2098 | 528.3812 | 0.0215 | 0.1524 |
| | 0.0051 | 158 | 0.0047 | 170.4682 | 164.0081 | 162.7237 | 0.0033 | 0.0051 |
| Rural | 0.1812 | 551 | 0.1811 | 337.3440 | 337.1524 | 337.1660 | 0.0197 | 0.0546 |
| | 0.0048 | 138 | 0.0042 | 92.0419 | 91.5658 | 91.6090 | 0.0028 | 0.0032 |
| Urban | 0.2609 | 327 | 0.2609 | 460.0384 | 325.2216 | 325.3178 | 0.0255 | 0.9930 |
| | 0.0077 | 142 | 0.0007 | 85.7729 | 8.1272 | 0.0007 | 0.0004 | 0.0012 |
| | | | | $\alpha = 0.05$ | | | | |
| *Utter Pradesh* | | | | | | | | |
| All | 0.2163 | 1262 | 0.2159 | 698.7892 | 695.6476 | 709.7158 | 0.0213 | 0.1336 |
| | 0.0042 | 302 | 0.0040 | 254.9395 | 246.8035 | 234.0947 | 0.0032 | 0.0048 |
| Rural | 0.1997 | 903 | 0.1993 | 476.1066 | 476.1066 | 475.9842 | 0.0214 | 0.1758 |
| | 0.0041 | 197 | 0.0041 | 157.4488 | 157.4488 | 158.1605 | 0.0035 | 0.0054 |
| Urban | 0.2229 | 359 | 0.2229 | 850.5696 | 358.9922 | 358.9976 | 0.0362 | 1.0000 |
| | 0.0092 | 177 | 0.0002 | 212.7411 | 0.5515 | 0.1697 | 0.0001 | 0.0000 |
| *West Bengal* | | | | | | | | |
| All | 0.2320 | 878 | 0.2316 | 730.4786 | 686.8248 | 686.7118 | 0.0225 | 0.2094 |
| | 0.0051 | 186 | 0.0033 | 232.2337 | 167.1106 | 167.6879 | 0.0029 | 0.0058 |
| Rural | 0.1812 | 551 | 0.1809 | 452.0266 | 432.7092 | 432.7040 | 0.0208 | 0.0688 |
| | 0.0048 | 163 | 0.0028 | 116.9002 | 87.0609 | 87.1087 | 0.0024 | 0.0036 |
| Urban | 0.2609 | 327 | 0.2609 | 628.6992 | 327.0000 | 327.0000 | 0.0303 | 1.0000 |
| | 0.0077 | 162 | 0.0000 | 104.9581 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

For each of the datasets, the upper bound of the confidence interval width $\omega$ and the confidence level $1 - \alpha$ are pre-specified. The pilot sample size for each $s^{th}$ stratum is then computed by using the pilot sample size formula. For example, for $\omega = 0.02$ and $\alpha = 5\%$, we start with a pilot sample size of $t_s = \max\left\{2, \min\left\{600, \left\lceil \frac{2a_s z_{0.05/2}}{0.02} \right\rceil\right\}\right\} = 66$ for $s = 1, 2, 3$. Thus, the initial number of clusters to be taken from the population is $t = \sum_{s=1}^{3} t_s = 198$. Now, a simple random sample of $t_s$ clusters is drawn from stratum $s$. Then for each selected cluster, a simple random sample of $k = 4$ households is drawn. The per capita monthly expenditure $x_{sc_sh}$ of the $h^{th}$ household belonging to the $c_s^{th}$ cluster from the $s^{th}$ stratum is recorded and weighted with $w_{sc_sh} = \frac{M_{sc_s}H_s}{kt_s}\nu_{sc_sh}$. The estimator of $\xi^2$, $V_t^2$ is computed.

Now we apply the purely sequential procedure and the two-stage procedure for $\alpha = 0.05, 0.10$, $\omega = 0.020, 0.025$ and $m^{'} = 1, 10, 20$ with replication size being 5000.

Tables C.1 through C.12 in Appendix C show the results of the Monte Carlo simulations on the three pseudo populations using the purely sequential procedure. The second column of the tables gives the theoretical Gini index ($G$) given the distribution in the first column. $\bar{N}$ and $s_N$ indicate the estimate of the average optimal sample size and its standard error respectively. The coverage probability ($p$) of the 5000 confidence intervals is displayed in column 4 with its corresponding standard error being $s_p$. The average length of the confidence intervals ($\bar{w}_N$) is in column 5 and $s_{w_N}$ is its standard error. The last column denotes the proportion of confidence intervals that were wider than the specified upper bound $\omega$ ($p(w_N > \omega)$).

Results in Tables C.1 to C.12 show that the average widths were all less than the pre-specified value of $\omega$. Since $p(w_N > \omega) = 0$ for all the results, the width of the confidence intervals will be less than $\omega$. This was also indicated while replicating the procedure using NSS data. The coverage probabilities obtained were approximately equal to $1 - \alpha$.

The results in Tables C.13 – C.16 (in Appendix C) show the properties of the Monte Carlo simulations on the three pseudo populations using the two-stage methodology. The second column of the tables gives the theoretical Gini index ($G$) given the distribution in the first column. The average final sample size after 5000 replications is denoted as $\bar{Q}$ along with the corresponding standard error $s_Q$. The coverage probability ($p$) of the 5000 confidence intervals is in column 4 with its standard error being $s_p$. The average length of the confidence intervals ($\bar{w}_Q$) is in column 5 and $s_{w_Q}$ is its standard error. The last column denotes the proportion of confidence intervals that were wider than the specified upper bound $\omega$.

The results in Tables C.13 through C.16 show that the width of the confidence intervals may not be less than $\omega$. This was also indicated while replicating the procedure using NSS data. However, the coverage probabilities obtained were approximately equal to $1 - \alpha$.

From the simulations, we find that the coverage probability for the confidence intervals for both purely sequential procedure and the two-stage procedure are approximately close to the desired confidence level provided that the cluster size (in all strata) is large, which is also a basic criterion while proving the asymptotic normality in Equation (4.8). However, the width of the confidence intervals for the two stage procedure, unlike the purely sequential procedure, may result in confidence intervals of width larger than the pre-specified value of $\omega$. This outcome is not surprising since the two-stage procedure is based on only the pilot sample which is usually taken to be small. So, there is a higher variability in the estimate of $\xi^2$. The optimal cluster sizes for the purely sequential procedure is less than that of the two-stage procedure.

## 4.8 Concluding Thoughts

Working within the asymptotic purview for complex survey data, developed by Bhattacharya (2005, 2007), we have developed purely sequential and two-stage procedures for constructing sufficiently narrow confidence intervals for the Gini index which is one of the most popular measure of economic inequality. Our procedure may be applied for surveys when the stratified clustered sample data are drawn from a large number of clusters per stratum, which is a reasonable assumption to make.

It turns out that the two-stage procedure is practically more feasible under this survey design than the purely sequential procedure. The confidence intervals of both procedures yielded a coverage probability closer to the desired confidence level, however, the purely sequential procedure produces confidence intervals whose width is always less than the desired bound $\omega$. The two-stage procedure is also known to over-estimate the optimal sample size than the purely sequential procedure (see Mukhopadhyay and De Silva, 2009) and this property can be seen in results from the simulation and the application to the NSS data. Furthermore, the estimated optimal sample sizes have smaller standard error under purely sequential procedure as compared to two-stage procedure.

After the first stage, the purely sequential procedure requires observations from additional $m'$ clusters every time the condition in the stopping rule is not met. Thus, there is a need to fix the value of $m'$. In some situations, it is as easy to collect observations from more than one cluster as it is to collect observations from a single cluster at every stage. So, as per convenience, the value of $m'$ should be accordingly decided based on economic considerations. In fact, the purely sequential procedure is not affected by the choice of $m'$, the larger the value of $m'$, the lesser will be the number of stages.

We believe, this is the first article to make developments on having sufficiently narrow confidence interval of economic inequality index based on complex household survey. Developing a survey design that takes economic factors into account was a very important issue that was raised by Bhattacharya (2005). We feel that our work is a first step to address that important issue in the sense of achieving a sufficiently narrow confidence interval. This work can be extended to include simultaneous confidence intervals or confidence ellipsoids for several economic indices that are can be obtained from sample surveys.

### 4.9 Acknowledgement

The research work of Dr. Bhargab Chattopadhyay was part of the project sanctioned by Science and Engineering Research Board, Government of India (ECR/2017/001213). The author is also grateful to the Ministry of Statistics and Program Implementation, Government Of India for sharing the household data related to the consumer expenditure for the year 2007–2008 (Round 64, Schedule 1.0).

# CHAPTER 5
# SUMMARY AND CONCLUSION

## 5.1 Summary

The reporting of effect sizes and their respective confidence intervals in research have gain more traction in recent years within statistics and related research fields. However, researchers do not report their confidence intervals due to intervals being embarrassingly wide. A wider confidence interval indicates more uncertainty as compared to a narrower one with all things being equal. Thus, researchers prefer narrower confidence intervals. A narrower confidence interval can be achieved by increasing the sample size. Thus the interval becomes more accurate given the same confidence level. To achieve sufficiently narrow confidence intervals, sample size calculation approaches could be used to obtain the necessary sample size. However, these methods also required the knowledge of population parameters or the distribution of the data which are unknown in advance.

To resolve the drawbacks of these approaches, we developed a purely sequential procedure to construct sufficiently narrow confidence intervals at a pre-specified level in a distribution-free environment. We do this for effect sizes that are ratios of linear combinations of parameters. Some of these effect sizes include standardized mean difference, coefficient of variation and simple linear regression slope. We also developed purely sequential procedures to construct sufficiently narrow confidence intervals for different measures of correlation including multiple $R^2$. We proposed both purely sequential and two-stage procedures for obtaining a narrow confidence interval for Gini index under a complex survey. Unlike fixed sample size procedure, in all our procedure, we defined stopping rules that indicate whether further sampling is required or should be terminated after sampling at a pilot stage. Moreover, apart from multiple $R^2$ where data were assumed to be normally distributed due to our limitation to available literature, we did not make distributional assumptions for the effect sizes or the Gini index. Instead, we relied on their asymptotic distributions.

Using theorems and simulation studies, we showed that our multi-stage sampling procedures achieved the desired narrow width and at that same time are asymptotically consistent and efficient. That is, they produced confidence intervals whose width did not exceed the targeted upper bound $\omega$ and had $100(1-\alpha)\%$ coverage probability. More so, the ratio of their average final sample sizes to their respective optimal sample sizes were close to 1 as the sample size increased.

## 5.2 Concluding Remarks

This work adds to and expands the existing literature on sequential methodology. It is the first work to provide a general framework for a general class of effect size and different types of correlation measures under purely sequential procedure. It is also the first work to develop both the purely sequential and the two-stage procedures for estimating and constructing bounded-width confidence interval width for the Gini index under complex household survey design. Even though other types of multi-stage sampling were not discussed, the procedures developed in this work can easily be extended to include their respective two-stage, three-stage, and other procedures that were mentioned in Section 1.3.

## 5.3 Future Work

An area of this work that needs future consideration is the choice of $m'$, that is, the size of the additional sample added at each stage after the pilot stage. Currently in this work, $m'$ is chosen arbitrarily without considering the rate of convergence of the procedure. A smaller choice of $m'$ may lead to longer purely sequential procedure while a larger choice may lead to oversampling. Thus, formulating $m'$ to be a function that depends on data sampled could be one way to solve this problem. Also, another way is to design $m'$ as a function of the sampling cost.

Furthermore, as discussed in this work, the choice of estimators for the unknown population parameters could affect the estimation of the final sample size. In this regard, bootstrap could be employed in the multi-stage sampling, especially the second stage of the two-stage procedure, to improve the estimation of the final sample size. This in effect can guarantee that oversampling is reduced, confidence intervals are robust, and final sample sizes are asymptotically efficient.

Sampling and estimation cost which are vital part of any research work were not considered in procedures developed in this work. This is another extension of this work that could be considered for future research, especially in the case of Gini index estimation where national budgets are mostly assigned to the surveys.

Last but not least, it is our hope to provide, in the near future, an R package for implementing all the procedures that have been developed and discussed in this manuscript, as well as other sequential procedures not mentioned here, to aid in their implementation and application.

# APPENDIX A

# CENTRAL LIMIT THEOREM FOR SOME EFFECT SIZES

## A.1 Standardized Mean Difference

**Theorem A.1.** *If the parent distribution for both groups is such that the corresponding fourth moments exists, then the stopping rule (2.32) adapted for the standardized mean difference yields the asymptotic consistency property, that is:*

$$\sqrt{n}\,(d_n - \delta) \xrightarrow{\mathcal{L}} N(0, \xi^2),$$

*where*

$$\xi^2 = 2 - \frac{(\mu_1 - \mu_2)(\mu_{13} - \mu_{23})}{\sigma^4} + \frac{(\mu_1 - \mu_2)^2}{4\sigma^6}\left(\frac{\mu_{14} + \mu_{24}}{4} - \frac{\sigma^4}{2}\right)$$

*Proof.* The asymptotic joint distribution of the sample mean difference $\bar{X}_{1n} - \bar{X}_{2n}$ and the pooled sample variance $s_{pn}^2 = \sqrt{\frac{1}{2}(s_{1n}^2 + s_{2n}^2)}$ is given as

$$\sqrt{n}\begin{bmatrix} (\bar{X}_{1n} - \bar{X}_{2n}) - (\mu_1 - \mu_2) \\ s_{pn}^2 - \sigma^2 \end{bmatrix} \xrightarrow{\mathcal{L}} N_2\left(\mathbf{0}, \mathbf{\Sigma}\right)$$

where

$$\mathbf{\Sigma} = \begin{bmatrix} 2\sigma^2 & \frac{1}{2}(\mu_{13} - \mu_{23}) \\ \frac{1}{2}(\mu_{13} - \mu_{23}) & \frac{1}{4}(\mu_{14} + \mu_{24} - 2\sigma^2) \end{bmatrix}.$$

Applying the delta method, we have the asymptotic distribution of the sample standardized mean difference $d_n = (\bar{X}_{1n} - \bar{X}_{2n})/s_{pn}$, an estimator of the population standardized mean difference $\delta = (\mu_1 - \mu_2)/\sigma$, to be

$$\sqrt{N}\,(d_N - \delta) \xrightarrow{\mathcal{L}} N(0, \xi^2),$$

where

$$\xi^2 = 2 - \frac{(\mu_1 - \mu_2)(\mu_{13} - \mu_{23})}{\sigma^4} + \frac{(\mu_1 - \mu_2)^2}{4\sigma^6}\left(\frac{\mu_{14} + \mu_{24}}{4} - \frac{\sigma^4}{2}\right)$$

and $\mu_{ij}$ is the $j$th central moment of distribution $F_i$, for $i = 1, 2$. ■

An estimator based on U-statistics for the population third central moment, that is, $\mu_{i3} = \mathrm{E}[X_i - \mu_i]^3$ and population fourth central moment $\mu_{i4} = \mathrm{E}[X_i - \mu_i]^4$ are

$$\widehat{\mu}_{i3n} = \frac{n}{(n-1)(n-2)} \sum_{k=1}^{n} (X_{ik} - \bar{X}_{in})^3, \tag{A.1}$$

and

$$\begin{aligned}\widehat{\mu}_{i4n} = &\frac{n^2}{(n-1)(n-2)(n-3)} \sum_{k=1}^{n} (X_{ik} - \bar{X}_{in})^4 - \frac{2n-3}{(n-1)(n-2)(n-3)} \sum_{k=1}^{n} X_{ik}^4 + \\ &\frac{8n-12}{(n-1)(n-2)(n-3)} \bar{X}_{in} \sum_{k=1}^{n} X_{ik}^3 - \frac{6n-9}{n(n-1)(n-2)(n-3)} \left( \sum_{k=1}^{n} X_{ik}^2 \right)^2, \end{aligned} \tag{A.2}$$

respectively.

**Theorem A.2.** *If the parent distribution for both groups is such that the corresponding fourth moments exist, then the stopping rule* (2.32) *adapted for the standardized mean difference yields:*

$$\textit{Part 1: } P\left( d_N - \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} < \delta < d_N + \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} \right) \to 1 - \alpha \textit{ as } N \to \infty.$$

$$\textit{Part 2: } \frac{2z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} \leq \omega \tag{A.3}$$

*Proof.* This can be proved by using the proof of Theorem 2.2. ■

## A.2 Coefficient of Variation

Using Heffernan (1997) or Abbasi et al. (2010), an estimator based on U-statistics for the population third central moment, that is, $\mu_3 = \mathrm{E}[X - \mu]^3$ and population fourth central moment $\mu_4 = \mathrm{E}[X - \mu]^4$ are respectively given by:

$$\widehat{\mu}_{3n} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (X_i - \bar{X}_n)^3, \tag{A.4}$$

and

$$\widehat{\mu}_{4n} = \frac{n^2}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^4 - \frac{2n-3}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n} X_i^4 + \frac{8n-12}{(n-1)(n-2)(n-3)}\bar{X}_n\sum_{i=1}^{n} X_i^3 - \frac{6n-9}{n(n-1)(n-2)(n-3)}\left(\sum_{i=1}^{n} X_i^2\right)^2. \quad \text{(A.5)}$$

The quantity $\widehat{\mu}_{3n}$ is a U-statistic of degree 3 and is an unbiased and consistent estimate of $\mu_3$, whereas $\widehat{\mu}_{4n}$ is a U-statistic of degree 4 and is an unbiased and consistent estimators of $\mu_4$.

**Theorem A.3.** *If the parent distribution $F$ is such that the fourth moment exists, then the stopping rule* (2.32) *adapted for the coefficient of variation yields:*

$$\textit{Part 1: } P\left(k_N - \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} < \kappa < k_N + \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}\right) \to 1-\alpha \textit{ as } N \to \infty.$$

$$\textit{Part 2: } \frac{2z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} \leq \omega \quad \text{(A.6)}$$

*Proof.* This can be proved by using the proof of Theorem 2.2. ■

## A.3 Regression Coefficient: Simple Linear Model

**Theorem A.4.** *Suppose for $i, j = 0, 1, 2, 3, 4$, $\mu_{ij} = E[(X-\mu_x)^i(Y-\mu_Y)^j]$, $\sigma_X, \sigma_{XY}$ and $\sigma_Y$ all exist. Then, the central limit theorem corresponding to the regression coefficient $\beta_1$ of the simple linear model is*

$$\sqrt{n}\,(b_{1n} - \beta_1) \xrightarrow{\mathcal{L}} N(0, \xi^2), \quad \text{(A.7)}$$

*where*

$$\xi^2 = \frac{\mu_{22}}{\sigma_X^4} - \frac{2\sigma_{XY}\mu_{31}}{\sigma_X^6} + \frac{\sigma_{XY}^2\mu_{40}}{\sigma_X^8}.$$

*Proof.* For proving a central limit theorem for the regression coefficient in simple linear model, defined in Equation (2.53), we first find the asymptotic joint distribution of the sample covariance $s_{XYn}$ and the sample variance of $X$, $s^2_{Xn}$. This is given by

$$\sqrt{n}\left(s_{XYn} - \sigma_{XY}, s^2_{Xn} - \sigma^2_X\right)' \xrightarrow{\mathcal{L}} N_2\left(\mathbf{0}, \mathbf{\Sigma}\right) \tag{A.8}$$

where the asymptotic variance of sample correlation coefficient is given by

$$\mathbf{\Sigma} = \begin{bmatrix} \mu_{22} - \sigma^2_{XY} & \mu_{31} - \sigma_{XY}\sigma^2_X \\ \mu_{31} - \sigma_{XY}\sigma^2_X & \mu_{40} - \sigma^4_X \end{bmatrix}. \tag{A.9}$$

An application of the delta method will give the central limit theorem for $\beta$ as in Equation (A.7). ■

A consistent estimator for $\xi^2$ is given by

$$\widehat{\xi}^2_n = \max\left\{V^2_n, n^{-3}\right\}, \tag{A.10}$$

where,

$$V^2_n = \frac{\hat{\mu}_{22n}}{s^4_{Xn}} - \frac{2s_{XYn}\hat{\mu}_{31n}}{s^6_{Xn}} + \frac{s^2_{XYn}\hat{\mu}_{40n}}{s^8_{Xn}}. \tag{A.11}$$

**Theorem A.5.** *If the error distribution is such that $E[\widehat{\xi}^2_n]$ exist, then the stopping rule* (2.32) *adapted for the regression coefficient $\beta_1$ yields:*

*Part 1:* $P\left(b_{1N} - \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} < \beta_1 < b_{1N} + \frac{z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}}\right) \to 1 - \alpha$ *as* $N \to \infty$.

*Part 2:* $$\frac{2z_{\alpha/2}\widehat{\xi}_N}{\sqrt{N}} \leq \omega \tag{A.12}$$

*Proof.* This can be proved by using the proof of Theorem 2.2. ■

# APPENDIX B

# CONSISTENT ESTIMATOR FOR ASYMPTOTIC VARIANCE OF PEARSON'S CORRELATION COEFFICIENT

Using Lee (1990), a consistent estimator of $\xi_\rho^2$ is $\hat{\xi}_{\rho n}^2 = \max\left\{V_{\rho n}^2, n^{-3}\right\}$, where $V_{\rho n}^2$ is

$$V_{\rho n}^2 = \frac{r_n^2}{4}\left(\frac{\hat{\mu}_{40n}}{S_{Xn}^4} + \frac{\hat{\mu}_{04n}}{S_{Yn}^4} + \frac{2\hat{\mu}_{22n}}{S_{Xn}^2 S_{Yn}^2} + \frac{4\hat{\mu}_{22n}}{S_{XYn}} - \frac{4\hat{\mu}_{31n}}{S_{XYn} S_{Xn}^2} - \frac{4\hat{\mu}_{13n}}{S_{XYn}^2 S_{Yn}^2}\right). \tag{B.1}$$

$\hat{\mu}_{40n}$ and $\hat{\mu}_{04n}$ are the respective unbiased estimators of the fourth central moment of $X$ $(\mu_{40})$ and $Y$ $(\mu_{04})$ which are given respectively as:

$$\begin{aligned}\widehat{\mu}_{40n} = {} & \frac{n^2}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(X_i - \bar{X}_n)^4 - \frac{2n-3}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n} X_i^4 + \\ & \frac{8n-12}{(n-1)(n-2)(n-3)}\bar{X}_n\sum_{i=1}^{n} X_i^3 - \frac{6n-9}{n(n-1)(n-2)(n-3)}\left(\sum_{i=1}^{n} X_i^2\right)^2\end{aligned} \tag{B.2}$$

and

$$\begin{aligned}\widehat{\mu}_{04n} = {} & \frac{n^2}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^4 - \frac{2n-3}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n} Y_i^4 + \\ & \frac{8n-12}{(n-1)(n-2)(n-3)}\bar{Y}_n\sum_{i=1}^{n} Y_i^3 - \frac{6n-9}{n(n-1)(n-2)(n-3)}\left(\sum_{i=1}^{n} Y_i^2\right)^2.\end{aligned} \tag{B.3}$$

According to Cook (1951) and Fisher (1930), the remaining estimators can be defined as

$$\hat{\mu}_{13n} = k_{13} + 3k_{02}k_{11} \tag{B.4}$$

$$\hat{\mu}_{22n} = k_{22} + k_{20}k_{02} + 2k_{11}^2 \tag{B.5}$$

$$\hat{\mu}_{31n} = k_{31} + 3k_{20}k_{11} \tag{B.6}$$

where

$$
\begin{aligned}
W_{pq} &= \sum_i X_i^p Y_i^q \\
k_{02} &= \frac{1}{n-1}\left(W_{02} - \frac{1}{n}W_{01}^2\right) = S_{Yn}^2 \\
k_{11} &= \frac{1}{n-1}\left(W_{11} - \frac{1}{n}W_{10}W_{01}\right) = S_{XYn} \\
k_{13} &= \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)W_{13} - \frac{n+1}{n}W_{03}W_{10} - \frac{3(n-1)}{n}W_{11}W_{02}\right. \\
&\quad \left. - \frac{3(n+1)}{n}W_{12}W_{01} + \frac{6}{n}W_{11}W_{01}^2 + \frac{6}{n}W_{02}W_{01}W_{10} - \frac{6}{n^2}W_{10}W_{01}^3\right\} \\
k_{20} &= \frac{1}{n-1}\left(W_{20} - \frac{1}{n}W_{10}^2\right) = S_{Xn}^2 \\
k_{22} &= \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)W_{22} - \frac{2(n+1)}{n}W_{21}W_{01} - \frac{n+1}{n}W_{12}W_{10}\right. \\
&\quad \left. - \frac{n-1}{n}W_{20}W_{02} - \frac{2(n-1)}{n}W_{11}^2 + \frac{8}{n}W_{11}W_{01}W_{10} + \frac{2}{n}W_{02}W_{10}^2 + \frac{2}{n}W_{20}W_{01}^2 - \frac{6}{n^2}W_{10}^2W_{01}^2\right\} \\
k_{31} &= \frac{n}{(n-1)(n-2)(n-3)}\left\{(n+1)W_{31} - \frac{n+1}{n}W_{30}W_{01} - \frac{3(n-1)}{n}W_{11}W_{20}\right. \\
&\quad \left. - \frac{3(n+1)}{n}W_{21}W_{10} + \frac{6}{n}W_{11}W_{10}^2 + \frac{6}{n}W_{20}W_{10}W_{01} - \frac{6}{n^2}W_{01}W_{10}^3\right\} \qquad \text{(B.7)}
\end{aligned}
$$

# APPENDIX C

# SIMULATION RESULTS FOR CONFIDENCE INTERVAL FOR THE GINI INDEX

## C.1 Pure Sequential Procedure

Table C.1. Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.02$, $m' = 1$)

| Distribution | $G$ | $\bar{N}$<br>$(s_N)$ | $p$<br>$(s_p)$ | $\bar{w}_N$<br>$(s_{w_N})$ | $p(w_N > \omega)$<br>$se(p(w_N > \omega))$ |
|---|---|---|---|---|---|
| Pareto<br>(scale=20000, shape=5) | 0.1111 | 273.3894<br>(41.2829) | 0.9020<br>(0.0042) | 0.0155<br>(0.0012) | 0.0000<br>(0.0000) |
| Lognormal<br>(mean = 2.185, sd = 0.562) | 0.3089 | 441.5382<br>(49.2216) | 0.8944<br>(0.0043) | 0.0184<br>(0.0003) | 0.0000<br>(0.0000) |
| Gamma<br>(shape = 2.649, rate = 0.84) | 0.3308 | 403.3110<br>(25.5081) | 0.8976<br>(0.0043) | 0.0182<br>(0.0002) | 0.0000<br>(0.0000) |

Table C.2. Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.02$, $m' = 1$)

| Distribution | $G$ | $\bar{N}$<br>$(s_N)$ | $p$<br>$(s_p)$ | $\bar{w}_N$<br>$(s_{w_N})$ | $p(w_N > \omega)$<br>$se(p(w_N > \omega))$ |
|---|---|---|---|---|---|
| Pareto<br>(scale=20000, shape=5) | 0.1111 | 354.2580<br>(55.5185) | 0.9484<br>(0.0031) | 0.0163<br>(0.0010) | 0.0000<br>(0.0000) |
| Lognormal<br>(mean = 2.185, sd = 0.562) | 0.3089 | 602.5182<br>(61.3798) | 0.9448<br>(0.0032) | 0.0188<br>(0.0002) | 0.0000<br>(0.0000) |
| Gamma<br>(shape = 2.649, rate = 0.84) | 0.3308 | 546.0798<br>(32.2857) | 0.9514<br>(0.0030) | 0.0186<br>(0.0002) | 0.0000<br>(0.0000) |

Table C.3. Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.025$, $m' = 1$)

| Distribution | $G$ | $\bar{N}$ | $p$ | $\bar{w}_N$ | $p(w_N > \omega)$ |
|---|---|---|---|---|---|
| | | $(s_N)$ | $(s_p)$ | $(s_{w_N})$ | $se(p(w_N > \omega))$ |
| Pareto | 0.1111 | 199.3122 | 0.8952 | 0.0181 | 0.0000 |
| (scale=20000, shape=5) | | (29.6017) | (0.0043) | (0.0019) | (0.0000) |
| Lognormal | 0.3089 | 300.9054 | 0.8942 | 0.0223 | 0.0000 |
| (mean = 2.185, sd = 0.562) | | (36.0870) | (0.0043) | (0.0006) | (0.0000) |
| Gamma | 0.3308 | 278.0940 | 0.8940 | 0.0219 | 0.0000 |
| (shape = 2.649, rate = 0.84) | | (18.6691) | (0.0044) | (0.0004) | (0.0000) |

Table C.4. Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.025$, $m' = 1$)

| Distribution | $G$ | $\bar{N}$ | $p$ | $\bar{w}_N$ | $p(w_N > \omega)$ |
|---|---|---|---|---|---|
| | | $(s_N)$ | $(s_p)$ | $(s_{w_N})$ | $se(p(w_N > \omega))$ |
| Pareto | 0.1111 | 253.7610 | 0.9486 | 0.0191 | 0.0000 |
| (scale=20000, shape=5) | | (38.1767) | (0.0031) | (0.0016) | (0.0000) |
| Lognormal | 0.3089 | 404.7870 | 0.9408 | 0.0229 | 0.0000 |
| (mean = 2.185, sd = 0.562) | | (46.0777) | (0.0033) | (0.0005) | (0.0000) |
| Gamma | 0.3308 | 370.9584 | 0.9484 | 0.0226 | 0.0000 |
| (shape = 2.649, rate = 0.84) | | (23.9455) | (0.0031) | (0.0003) | (0.0000) |

Table C.5. Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.02$, $m' = 10$)

| Distribution | $G$ | $\bar{N}$ | $p$ | $\bar{w}_N$ | $p(w_N > \omega)$ |
|---|---|---|---|---|---|
| | | $(s_N)$ | $(s_p)$ | $(s_{w_N})$ | $se(p(w_N > \omega))$ |
| Pareto | 0.1111 | 288.3120 | 0.9020 | 0.0152 | 0.0000 |
| (scale = 20000, shape = 5) | | (43.1982) | (0.0042) | (0.0012) | (0.0000) |
| Lognormal | 0.3089 | 456.3240 | 0.8976 | 0.0182 | 0.0000 |
| (mean = 2.185, sd = 0.562) | | (49.8155) | (0.0043) | (0.0004) | (0.0000) |
| Gamma | 0.3308 | 417.1740 | 0.8950 | 0.0179 | 0.0000 |
| (shape = 2.649, rate = 0.84) | | (26.9280) | (0.0043) | (0.0003) | (0.0000) |

Table C.6. Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.02$, $m' = 10$)

| Distribution | $G$ | $\bar{N}$<br>$(s_N)$ | $p$<br>$(s_p)$ | $\bar{w}_N$<br>$(s_{w_N})$ | $p(w_N > \omega)$<br>$se(p(w_N > \omega))$ |
|---|---|---|---|---|---|
| Pareto<br>(scale = 20000, shape = 5) | 0.1111 | 368.7300<br>(55.9612) | 0.9466<br>(0.0032) | 0.0160<br>(0.0010) | 0.0000<br>(0.0000) |
| Lognormal<br>(mean = 2.185, sd = 0.562) | 0.3089 | 615.7560<br>(61.8045) | 0.9450<br>(0.0032) | 0.0186<br>(0.0003) | 0.0000<br>(0.0000) |
| Gamma<br>(shape = 2.649, rate = 0.84) | 0.3308 | 558.6360<br>(33.2710) | 0.9516<br>(0.0030) | 0.0184<br>(0.0002) | 0.0000<br>(0.0000) |

Table C.7. Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.025$, $m' = 10$)

| Distribution | $G$ | $\bar{N}$<br>$(s_N)$ | $p$<br>$(s_p)$ | $\bar{w}_N$<br>$(s_{w_N})$ | $p(w_N > \omega)$<br>$se(p(w_N > \omega))$ |
|---|---|---|---|---|---|
| Pareto<br>(scale = 20000, shape = 5) | 0.1111 | 213.9060<br>(30.8347) | 0.8996<br>(0.0043) | 0.0175<br>(0.0018) | 0.0000<br>(0.0000) |
| Lognormal<br>(mean = 2.185, sd = 0.562) | 0.3089 | 315.6540<br>(37.4517) | 0.8984<br>(0.0043) | 0.0218<br>(0.0007) | 0.0000<br>(0.0000) |
| Gamma<br>(shape = 2.649, rate = 0.84) | 0.3308 | 291.9000<br>(20.6367) | 0.8924<br>(0.0044) | 0.0213<br>(0.0006) | 0.0000<br>(0.0000) |

Table C.8. Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.025$, $m' = 10$)

| Distribution | $G$ | $\bar{N}$<br>$(s_N)$ | $p$<br>$(s_p)$ | $\bar{w}_N$<br>$(s_{w_N})$ | $p(w_N > \omega)$<br>$se(p(w_N > \omega))$ |
|---|---|---|---|---|---|
| Pareto<br>(scale = 20000, shape = 5) | 0.1111 | 268.6380<br>(39.6310) | 0.9488<br>(0.0031) | 0.0187<br>(0.0016) | 0.0000<br>(0.0000) |
| Lognormal<br>(mean = 2.185, sd = 0.562) | 0.3089 | 418.0260<br>(47.0880) | 0.9400<br>(0.0034) | 0.0226<br>(0.0005) | 0.0000<br>(0.0000) |
| Gamma<br>(shape = 2.649, rate = 0.84) | 0.3308 | 383.4900<br>(25.1437) | 0.9478<br>(0.0031) | 0.0222<br>(0.0004) | 0.0000<br>(0.0000) |

Table C.9. Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.02$, $m' = 20$)

| Distribution | $G$ | $\bar{N}$ | $p$ | $\bar{w}_N$ | $p(w_N > \omega)$ |
|---|---|---|---|---|---|
| | | $(s_N)$ | $(s_p)$ | $(s_{w_N})$ | $se(p(w_N > \omega))$ |
| Pareto | 0.1111 | 305.1360 | 0.8984 | 0.0148 | 0.0000 |
| (scale=20000, shape=5) | | (45.3104) | (0.0043) | (0.0013) | (0.0000) |
| Lognormal | 0.3089 | 471.8640 | 0.8980 | 0.0179 | 0.0000 |
| (mean = 2.185, sd = 0.562) | | (52.4671) | (0.0043) | (0.0005) | (0.0000) |
| Gamma | 0.3308 | 432.0480 | 0.8954 | 0.0176 | 0.0000 |
| (shape = 2.649, rate = 0.84) | | (31.5691) | (0.0043) | 0.0004 | (0.0000) |

Table C.10. Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.02$, $m' = 20$)

| Distribution | $G$ | $\bar{N}$ | $p$ | $\bar{w}_N$ | $p(w_N > \omega)$ |
|---|---|---|---|---|---|
| | | $(s_N)$ | $(s_p)$ | $(s_{w_N})$ | $se(p(w_N > \omega))$ |
| Pareto | 0.1111 | 384.3720 | 0.9484 | 0.0157 | 0.0000 |
| (scale=20000, shape=5) | | (59.4090) | (0.0031) | (0.0011) | (0.0000) |
| Lognormal | 0.3089 | 631.2240 | 0.9442 | 0.0184 | 0.0000 |
| (mean = 2.185, sd = 0.562) | | (63.9250) | (0.0032) | (0.0004) | (0.0000) |
| Gamma | 0.3308 | 573.5520 | 0.9516 | 0.0182 | 0.0000 |
| (shape = 2.649, rate = 0.84) | | (36.2823) | (0.0030) | (0.0003) | (0.0000) |

Table C.11. Simulation results for purely sequential procedure ($\alpha = 10\%$, $\omega = 0.025$, $m' = 20$)

| Distribution | $G$ | $\bar{N}$ | $p$ | $\bar{w}_N$ | $p(w_N > \omega)$ |
|---|---|---|---|---|---|
| | | $(s_N)$ | $(s_p)$ | $(s_{w_N})$ | $se(p(w_N > \omega))$ |
| Pareto | 0.1111 | 226.2000 | 0.8984 | 0.0171 | 0.0000 |
| (scale=20000, shape=5) | | (37.8894) | (0.0043) | (0.0017) | (0.0000) |
| Lognormal | 0.3089 | 331.5240 | 0.9010 | 0.0213 | 0.0000 |
| (mean = 2.185, sd = 0.562) | | (39.7815) | (0.0042) | (0.0009) | (0.0000) |
| Gamma | 0.3308 | 308.9280 | 0.8940 | 0.0208 | 0.0000 |
| (shape = 2.649, rate = 0.84) | | (20.8713) | (0.0044) | (0.0008) | (0.0000) |

Table C.12. Simulation results for purely sequential procedure ($\alpha = 5\%$, $\omega = 0.025$, $m' = 20$)

| Distribution | $G$ | $\bar{N}$<br>$(s_N)$ | $p$<br>$(s_p)$ | $\bar{w}_N$<br>$(s_{w_N})$ | $p(w_N > \omega)$<br>$se(p(w_N > \omega))$ |
|---|---|---|---|---|---|
| Pareto<br>(scale=20000, shape=5) | 0.1111 | 285.5040<br>(42.7734) | 0.9480<br>(0.0031) | 0.0182<br>(0.0017) | 0.0000<br>(0.0000) |
| Lognormal<br>(mean = 2.185, sd = 0.562) | 0.3089 | 433.5600<br>(49.7480) | 0.9412<br>(0.0033) | 0.0222<br>(0.0007) | 0.0000<br>(0.0000) |
| Gamma<br>(shape = 2.649, rate = 0.84) | 0.3308 | 398.8440<br>(27.4951) | 0.9530<br>(0.0030) | 0.0218<br>(0.0006) | 0.0000<br>(0.0000) |

## C.2 Two-Stage Procedure

Table C.13. Simulation results for two-stage procedure ($\alpha = 10\%$, $\omega = 0.02$)

| Distribution | $G$ | $\bar{Q}$<br>$(s_Q)$ | $p$<br>$(s_p)$ | $\bar{w}_Q$<br>$(s_{w_Q})$ | $p(w_Q > \omega)$<br>$se(p(w_Q > \omega))$ |
|---|---|---|---|---|---|
| Pareto<br>(scale = 20000, shape = 5) | 0.1111 | 342.0360<br>(86.7413) | 0.8996<br>(0.0043) | 0.0141<br>(0.0017) | 0.0068<br>(0.0012) |
| Lognormal<br>(mean = 2.185, sd = 0.562) | 0.3089 | 546.3234<br>(97.0866) | 0.8968<br>(0.0043) | 0.0167<br>(0.0013) | 0.0140<br>(0.0017) |
| Gamma<br>(shape = 2.649, rate = 0.84) | 0.3308 | 499.1904<br>(47.0225) | 0.9034<br>(0.0042) | 0.0164<br>(0.0007) | 0.0000<br>(0.0000) |

Table C.14. Simulation results for two-stage procedure ($\alpha = 5\%$, $\omega = 0.02$)

| Distribution | $G$ | $\bar{Q}$<br>$(s_Q)$ | $p$<br>$(s_p)$ | $\bar{w}_Q$<br>$(s_{w_Q})$ | $p(w_Q > \omega)$<br>$se(p(w_Q > \omega))$ |
|---|---|---|---|---|---|
| Pareto<br>(scale = 20000, shape = 5) | 0.1111 | 447.1206<br>(111.9348) | 0.9474<br>(0.0032) | 0.0148<br>(0.0017) | 0.0112<br>(0.0056) |
| Lognormal<br>(mean = 2.185, sd = 0.562) | 0.3089 | 737.7720<br>(126.1189) | 0.9432<br>(0.0033) | 0.0172<br>(0.0012) | 0.0190<br>(0.0019) |
| Gamma<br>(shape = 2.649, rate = 0.84) | 0.3308 | 670.0152<br>(60.5521) | 0.9512<br>(0.0030) | 0.0169<br>(0.0007) | 0.0002<br>(0.0020) |

Table C.15. Simulation results for two-stage procedure ($\alpha = 10\%$, $\omega = 0.025$)

| Distribution | $G$ | $\bar{Q}$<br>$(s_Q)$ | $p$<br>$(s_p)$ | $\bar{w}_Q$<br>$(s_{w_Q})$ | $p(w_Q > \omega)$<br>$se(p(w_Q > \omega))$ |
|---|---|---|---|---|---|
| Pareto | 0.1111 | 246.1068 | 0.8980 | 0.0165 | 0.0052 |
| (scale = 20000, shape = 5) | | (63.1499) | (0.0043) | (0.0021) | (0.0010) |
| Lognormal | 0.3089 | 375.8364 | 0.8968 | 0.0201 | 0.0098 |
| (mean = 2.185, sd = 0.562) | | (69.7613) | (0.0043) | (0.0016) | (0.0014) |
| Gamma | 0.3308 | 345.9612 | 0.8990 | 0.0197 | 0.0000 |
| (shape = 2.649, rate = 0.84) | | (33.8431) | (0.0043) | (0.0008) | (0.0000) |

Table C.16. Simulation results for two-stage procedure ($\alpha = 5\%$, $\omega = 0.025$)

| Distribution | $G$ | $\bar{Q}$<br>$(s_Q)$ | $p$<br>$(s_p)$ | $\bar{w}_Q$<br>$(s_{w_Q})$ | $p(w_Q > \omega)$<br>$se(p(w_Q > \omega))$ |
|---|---|---|---|---|---|
| Pareto | 0.1111 | 316.3188 | 0.9476 | 0.0174 | 0.0068 |
| (scale = 20000, shape = 5) | | (79.7332) | (0.0032) | (0.0021) | (0.0012) |
| Lognormal | 0.3089 | 502.0482 | 0.9454 | 0.0208 | 0.0120 |
| (mean = 2.185, sd = 0.562) | | (90.2360) | (0.0032) | (0.0016) | (0.0015) |
| Gamma | 0.3308 | 459.2406 | 0.9512 | 0.0204 | 0.0000 |
| (shape = 2.649, rate = 0.84) | | (43.7022) | (0.0030) | (0.0009) | (0.0000) |

# REFERENCES

Abbasi, N., S. Hemati, and A. Jafarei (2010). Simple proof of the theorem: Tending u-statistics to central moments of sample. *Int. J. Contemp. Math. Sciences 5*(37), 1807–1811.

Aguirregabiria, V. and P. Mira (2007). Sequential estimation of dynamic discrete games. *Econometrica 75*(1), 1–53.

Albrecher, H., S. A. Ladoucette, and J. L. Teugels (2010). Asymptotics of the sample coefficient of variation and the sample dispersion. *Journal of Statistical Planning and Inference 140*(2), 358–368.

American Psychological Association (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

American Statistical Association (2016, March, 7). American statistical association releases statement on statistical significance and P-values: Provides principles to improve the conduct and interpretation of quantitative science. `http://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf`.

An, L. and S. E. Ahmed (2008). Improving the performance of kurtosis estimator. *Computational Statistics & Data Analysis 52*(5), 2669 – 2681.

Anscombe, F. J. (1949). Large-sample theory of sequential estimation. *Biometrika 36*(3/4), 455–458.

Anscombe, F. J. (1952). Large-sample theory of sequential estimation. *Math. Proc. Camb. Phil. Soc. 48*(04), 600–607.

Anscombe, F. J. (1953). Sequential estimation. *Journal of the Royal Statistical Society. Series B (Methodological) 15*(1), 1–29.

Arcidiacono, P. and J. B. Jones (2003). Finite mixture distributions, sequential likelihood and the em algorithm. *Econometrica 71*(3), 933–946.

Armitage, P. (1960). *Sequential Medical Trials.* Blackwell Scientific Publications.

Association for Psychological Science (2014). Submission guidelines. Website. last checked: 09.25.2015.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin 66*(6), 423–437.

Bandyopadhyay, U. and A. Biswas (2015). Sequential and two-stage fixed-width confidence interval estimation in response-adaptive designs. *Sequential Analysis 34*(3), 350–363.

Beach, C. M. and R. Davidson (1983). Distribution-free statistical inference with lorenz curves and income shares. *The Review of Economic Studies 50*(4), 723–735.

Bhattacharya, D. (2005). Asymptotic inference from multi-stage samples. *Journal of Econometrics 126*(1), 145–171.

Bhattacharya, D. (2007). Inference on inequality from household survey data. *Journal of Econometrics 137*(2), 674–707.

Bilson Darku, F., F. Konietschke, and B. Chattopadhyay (2018). Gini index estimation within pre-specified error bound applied to indian household survey data. *Manuscript submitted for publication.*

Binder, D. A. and M. S. Kovacevic (1995). Estimating some measures of income inequality from survey data: an application of the estimating equations approach. *Survey Methodology 21*, 137–146.

Bonett, D. G. and T. A. Wright (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika 65*(1), 23–28.

Bonett, D. G. and T. A. Wright (2011). Sample size requirements for multiple regression interval estimation. *Journal of Organizational Behavior 32*(6), 822–830.

Borkowf, C. B. (1999). A new method for approximating the asymptotic variance of spearman's rank correlation. *Statistica Sinica 9*(2), 535–558.

Borkowf, C. B. (2002). Computing the nonnull asymptotic variance and the asymptotic relative efficiency of spearman's rank correlation. *Computational Statistics & Data Analysis 39*(3), 271 – 286.

Bose, A. and N. Mukhopadhyay (1994). Sequential estimation via replicated piecewise stopping number in a towparameter exponential family of distributions. *Sequential Analysis 13*(1), 1–10.

Chang, H.-H. and Z. Ying (2009). Nonlinear sequential designs for logistic item response theory models with applications to computerized adaptive tests. *The Annals of Statistics 37*(3), 1466–1488.

Chattopadhyay, B. and S. K. De (2016). Estimation of Gini index within pre-specified error bound. *Econometrics 4*(3), 30.

Chattopadhyay, B. and K. Kelley (2016). Estimation of the coefficient of variation with minimum risk: A sequential method for minimizing sampling error and study cost. *Multivariate Behavioral Research 51*(5), 627–648.

Chattopadhyay, B. and K. Kelley (2017). Estimating the standardized mean difference with minimum risk: Maximizing accuracy and minimizing cost with sequential estimation. *Psychological Methods 22*(1), 94–113.

Chattopadhyay, B. and N. Mukhopadhyay (2013). Two-stage fixed-width confidence intervals for a normal mean in the presence of suspect outliers. *Sequential Analysis 32*(2), 134–157.

Chow, Y. S. and H. Robbins (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics 36*(2), 457–462.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin 114*(3), 494–509.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist 49*(12), 997–1003.

Cook, M. B. (1951). Bi-variate k-statistics and cumulants of their joint sampling distribution. *Biometrika 38*(1-2), 179–195.

Corty, E. W. and R. W. Corty (2011). Setting sample size to ensure narrow confidence intervals for precise estimation of population values. *Nursing Research 60*(2), 148–153.

Cox, D. R. (1952). A note on the sequential estimation of means. *Mathematical Proceedings of the Cambridge Philosophical Society 48*(3), 447–450.

Croux, C. and C. Dehon (2010). Influence functions of the spearman and kendall correlation measures. *Statistical Methods & Applications 19*(4), 497–515.

Daniels, H. E. and M. G. Kendall (1947). The significance of rank correlations where parental correlation exists. *Biometrika 34*(3–4), 197–208.

Dantzig, G. B. (1940). On the non-existence of tests of student's hypothesis having power functions independent of $\sigma$. *Ann. Math. Statist. 11*(2), 186–192.

Davidson, R. (2009). Reliable inference for the gini index. *Journal of Econometrics 150*(1), 30–40.

Davidson, R. and J.-Y. Duclos (2000). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. *Econometrica 68*(6), 1435–1464.

De, S. K. and B. Chattopadhyay (2017). Minimum risk point estimation of gini index. *Sankhya B*.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 175–185.

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika 10*, 507–521.

Fisher, R. A. (1930). Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society s2-30*(1), 199–238.

Gastwirth, J. L. (1972). The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, 306–316.

Genest, C. and A.-C. Favre (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering 12*(4), 347–368.

Ghosh, B. K. and P. K. Sen (1991). *Handbook of sequential analysis*, Volume 118. New York, NY: CRC Press.

Ghosh, M. and N. Mukhopadhyay (1976). On two fundamental problems of sequential estimation. *Sankhya: The Indian Journal of Statistics, Series B (1960-2002) 38*(3), 203–218.

Ghosh, M. and N. Mukhopadhyay (1981). Consistency and asymptotic efficiency of two stage and sequential estimation procedures. *Sankhya: The Indian Journal of Statistics, Series A (1961-2002) 43*(2), 220–227.

Ghosh, M., N. Mukhopadhyay, and P. K. Sen (1997). *Sequential Estimation*. New York, NY: John Wiley & Sons, Inc.

Glass, G. V. and M. L. Smith (1979). Meta-analysis of research on class size and achievement. *Educational Evaluation and Policy Analysis 1*(1), 2–16.

Greene, W. H. (1998). Gender economics courses in liberal arts colleges: Further results. *The Journal of Economic Education 29*(4), 291–300.

Gut, A. (2009). *Stopped Random Walks*. Springer New York.

Hall, P. (1981). Asymptotic theory of triple sampling for sequential estimation of a mean. *The Annals of Statistics 9*(6), 1229–1238.

Hall, P. (1983). Sequential estimation saving sampling operations. *Journal of the Royal Statistical Society. Series B (Methodological) 45*(2), 219–223.

Heffernan, P. M. (1997). Unbiased estimation of central moments by using u-statistics. *Journal of the Royal Statistical Society. Series B (Methodological) 59*(4), pp. 861–863.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics 19*(3), 293–325.

Hoque, A. A. and J. A. Clarke (2015, jan). On variance estimation for a gini coefficient estimator obtained from complex survey data. *Communications in Statistics: Case Studies, Data Analysis and Applications 1*(1), 39–58.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association 47*(260), 663–685.

Kanninen, B. J. (1993). Design of sequential experiments for contingent valuation studies. *Journal of Environmental Economics and Management 25*(1), S1–S11.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software 20*(8), 1–24.

Kelley, K. (2007b). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods 39*(4), 755–766.

Kelley, K. (2008). Sample size planning for the squared multiple correlation coefficient: Accuracy in parameter estimation via narrow confidence intervals. *Multivariate Behavioral Research 43*(4), 524–555.

Kelley, K., F. Bilson Darku, and B. Chattopadhyay (2017). Sequential accuracy in parameter estimation for population correlation coefficients. *Under Review*.

Kelley, K., F. B. Darku, and B. Chattopadhyay (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods 23*(2), 226–243.

Kelley, K. and K. Lai (2011). Accuracy in parameter estimation for the root mean square error of approximation: Sample size planning for narrow confidence intervals. *Multivariate Behavioral Research 46*, 1–32.

Kelley, K. and S. E. Maxwell (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods 8*(3), 305–321.

Kelley, K. and K. J. Preacher (2012). On effect size. *Psychological Methods 17*(2), 137–152.

Kelley, K. and J. R. Rausch (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods 11*(4), 363–385.

Kojadinovic, I. and J. Yan (2010). Comparison of three semiparametric methods for estimating dependence parameters in copula models. *Insurance: Mathematics and Economics 47*(1), 52–63.

Kumar, N. and B. Chattopadhyay (2013). On the class of U-statistics based two-sample test for location. *Journal of Combinatorics, Information & System Sciences 38*(1–4), 191–201.

Kumar, S., R. Gupta, and Y. Raj (2012). An accelerated sequential class to minimize combined risk for simultaneous estimation of parameters of several. *Pakistan Journal of Statistics and Operation Research 8*(4), 737–748.

Lai, K. and K. Kelley (2011a). Accuracy in parameter estimation for ANCOVA and ANOVA contrasts: Sample size planning via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology 65*(2), 350–370.

Lai, K. and K. Kelley (2011b). Accuracy in parameter estimation for targeted effects in structural equation modeling: Sample size planning for narrow confidence intervals. *Psychological Methods 16*(2), 127–148.

Lai, T. L. (1998). Sequential analysis. In *Encyclopedia of Biostatistics*, Volume 5, pp. 4074–4079. Wiley, New York.

Langel, M. and Y. Tillé (2013). Variance estimation of the gini index: revisiting a result several times published. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 176*(2), 521–540.

Lee, A. J. (1990). *U-statistics: Theory and Practice*. New York, NY: CRC Press.

Lee, E. and R. Forthofer (2006). *Analyzing Complex Survey Data*. SAGE Publications, Inc.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*, Volume 19. Newbury Park, CA: Sage.

Mahalanobis, P. C. (1940). A sample survey of the acreage under jute in bengal. *Sankhyā: The Indian Journal of Statistics*, 511–530.

Mahalanobis, P. C. (1967). The sample census of the area under jute in bengal in 1940. *Sankhya: The Indian Journal of Statistics, Series B (1960-2002) 29*(1/2), 81–182.

Maxwell, S. E., K. Kelley, and J. R. Rausch (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology 59*, 537–563.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Eds.), *What if there where no significance tests?*, pp. 393–426. Mahwah, NJ: Lawrence Erlbaum Associates.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin 105*, 156–166.

Moinester, M. and R. Gottfried (2014). Sample size estimation for correlations with prespecified confidence interval. *The Quantitative Methods for Psychology 10*(2), 124–130.

Morrison, D. E. and R. E. Henkel (1970). *The Significance Test Controversy*. New Brunswick, CT: Aldine Transaction.

Mukhopadhyay, N. (1976). Fixed-width confidence intervals for the mean using a three-stage procedure. *Unpublished report*.

Mukhopadhyay, N. (1980). A consistent and asymptotically efficient two-stage procedure to construct fixed width confidence intervals for the mean. *Metrika 27*(1), 281–284.

Mukhopadhyay, N. (1996, jan). An alternative formulation of accelerated sequential procedures with applications to parametric and nonparametric estimation. *Sequential Analysis 15*(4), 253–269.

Mukhopadhyay, N. and B. Chattopadhyay (2012). A tribute to Frank Anscombe and random central limit theorem from 1952. *Sequential Analysis 31*(3), 265–277.

Mukhopadhyay, N. and B. Chattopadhyay (2013). On a new interpretation of the sample variance. *Statistical Papers*, 827–837.

Mukhopadhyay, N. and B. Chattopadhyay (2014). A note on the construction of a sample variance. *Sri Lankan Journal of Applied Statistics 15*(1), 71–80.

Mukhopadhyay, N. and S. Datta (1994). Replicated piecewise multistage sampling with applications. *Sequential Analysis 13*(3), 253–276.

Mukhopadhyay, N. and B. M. de Silva (1998). Multistage partial piecewise sampling and its applications. *Sequential Analysis 17*(1), 63–90.

Mukhopadhyay, N. and B. M. De Silva (2009). *Sequential methods and their applications*. Boca Raton, FL: CRC Press.

Mukhopadhyay, N. and P. K. Sen (1993). Replicated piecewise stopping numbers and sequential analysis. *Sequential Analysis 12*(2), 179–197.

Mukhopadhyay, N. and T. K. S. Solanky (1991, jan). Second order properties of accelerated stopping times with applications in sequential estimation. *Sequential Analysis 10*(1-2), 99–123.

Murphy, K. R. and B. Myors (2004). *Statistical Power Analysis: A Simple and General Model for Traditional and Modern Hypothesis Tests* (2nd ed.). Mahwah, NJ: Erlbaum.

Nadarajah, S. and A. K. Gupta (2006). Some bivariate gamma distributions. *Applied Mathematics Letters 19*(8), 767 – 774.

National Sample Survey Organization (2007). Note on estimation procedure of NSS 64th round. http://catalog.ihsn.org/index.php/catalog/1906/download/35538.

O'Brien, R. G. and J. Castelloe (2007). Sample-size analysis for traditional hypothesis testing: Concepts and issues. In A. Dmitrienko, C. Chuang-Stein, and R. D'Agostino (Eds.), *Pharmaceutical statistics using SAS: A practical guide*, pp. 237–272. SAS Institute.

Olkin, I. and J. D. Finn (1995). Correlations redux. *Psychological Bulletin 118*(1), 155 – 164.

Pornprasertmanit, S. and W. J. Schneider (2014). Accuracy in parameter estimation in cluster randomized designs. *Psychological Methods 19*(3), 356–379.

Preacher, K. J. and K. Kelley (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods 16*(2), 93–115.

R Core Team (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ransom, M. R. and J. S. Cramer (1983). Income distribution functions with disturbances. *European Economic Review 22*(3), 363–372.

Ray, W. D. (1957). Sequential confidence intervals for the mean of a normal population with unknown variance. *Journal of the Royal Statistical Society. Series B (Methodological) 19*(1), 133–143.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc. 58*(5), 527–535.

Schwarz, B. (2004). *The paradox of choice: Why more is less*, Volume 6. Harper Collins, New York.

Sen, P. K. (1981). *Sequential nonparametrics: Invariance principles and statistical inference.* Wiley New York.

Sen, P. K. and M. Ghosh (1981). Sequential point estimation of estimable parameters based on U-statistics. *Sankhyā: The Indian Journal of Statistics, Series A*, 331–344.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology 15*(2), 201–292.

Sproule, R. (1969). *A sequential fixed-width confidence interval for the mean of a U-statistic.* Ph. D. thesis, Ph. D. dissertation, Univ. of North Carolina.

Sproule, R. N. (1985). Sequential nonparametric fixed-width confidence intervals for U-statistics. *The Annals of Statistics 13*(1), 228–235.

Stanley, S. K., M. S. Wilson, and T. L. Milfont (2017). Exploring short-term longitudinal effects of right-wing authoritarianism and social dominance orientation on environmentalism. *Personality and Individual Differences 108*, 174–177.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics 16*(3), 243–258.

Stein, C. (1949). Some problems in sequential estimation. *Econometrica 17*(1), 77–78.

Stein, C. and A. Wald (1947). Sequential confidence intervals for the mean of a normal distribution with known variance. *Ann. Math. Statist. 18*(3), 427–433.

Task Force on Reporting of Research Methods in AERA Publications (2006). *Standards for Reporting on Empirical Social Science Research in AERA Publications, American Educational.* Washington, DC: American Educational Research Association.

Terry, L. J. and K. Kelley (2012). Sample size planning for composite reliability coefficients: Accuracy in parameter estimation via narrow confidence intervals. *British Journal of Mathematical and Statistical Psychology 65*(3), 371–401.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher 31*(3), 25–32.

van Erp, S., A. Verhagen, R. Grasman, and E. Wagenmakers (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in. *Psychological Bulletin*, 1990–2013.

Vapnyarskii, I. B. (2001). Lagrange multipliers. *Hazewinkel, Michiel, Encyclopedia of Mathematics, Springer, ISBN 978*, 1–55.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society 54*(3), 426–482.

Wald, A. (1945). Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics 16*(2), 117–186.

Wald, A. (1947). *Squential Analysis.* John Wiley & Sons, Inc.

Wasserstein, R. L. (2016). ASA statement on statistical significance and P-values. *The American Statistician X*, X–X.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). New York, NY: Elsevier.

Wolf, F. M. and R. G. Cornell (1986). Interpreting behavioral, biomedical, and psychological relationships in chronic disease from 2×2 tables using correlation. *Journal of Chronic Diseases 39*(8), 605–608.

Xu, K. (2007). U-statistics and their asymptotic results for some inequality and poverty measures. *Econometric Reviews 26*(5), 567–577.

# BIOGRAPHICAL SKETCH

Francis Bilson Darku is a Ghanaian, born and raised in Sekondi, located in the Western Region of Ghana. After completing St. Augustine's College - Cape Coast, in 2006, Francis attended Kwame Nkrumah University of Science and Technology, Kumasi from 2007 to 2011 and attained his Bachelor's degree in Actuarial Science. He worked as a Teaching and Research Assistant in the same university from 2011 to 2013. Thereafter, he entered The University of Texas at Dallas to study for his PhD in Statistics.

## CURRICULUM VITAE

# Francis Bilson Darku

May 31, 2018

Department of Mathematical Sciences
University of Texas at Dallas
800 W. Campbell Rd, FO35
Richardson, TX 75080

Email: Francis.BilsonDarku@utdallas.com
BilsonDarku@gmail.com
Website: sites.google.com/site/FBDarku
LinkedIn: www.linkedin.com/in/FBDarku

## EDUCATION

- **PhD Statistics** Aug 2018
  University of Texas at Dallas, Richardson USA
  Thesis: Study on Parameter Estimation via Multi-stage Sampling with Applications
  Advisors: Bhargab Chattopadhyay, Frank Konietschke
- **MSc Statistics** May 2017
  University of Texas at Dallas, Richardson USA
- **BSc Actuarial Science** May 2011
  Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
  Thesis: Regression Analysis on Road Accident Fatalities: Case Study - Central Region of Ghana
  Supervisor: Nana Kena Frempong
- **Actuarial Professional Exams**
  - Exam FM/2 (Financial Mathematics)
  - Exam P/1 (Probability)
  - VEE: Corporate Finance, Economics, Applied Statistical Methods

## TEACHING & WORK

- **Teaching Assistant** Aug 2013 - May 2018
  University of Texas at Dallas, Richardson Texas, USA
- **Teaching/Research Assistant** Sep 2011 - Jun 2013
  Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
- **Tutor** May 2012 - Apr 2013
  Winners Professional Institute, Accra, Ghana

- **Intern – Customer Mobilization Clerk** Jul 2010 - Aug 2010
  Ahantaman Rural Bank Limited, Sekondi, Ghana

- **Intern: Regional Account Office** Jun 2009 - Jul 2009
  Vodafone Telecommunication, Takoradi, Ghana

- **Pupils' Teacher** Oct 2006 - Jul 2007
  Baptist Preparatory and Junior High School, Sekondi, Ghana

## RESEARCH INTEREST

- Sequential Analysis
- Nonparametric Methods
- Econometrics

## PUBLICATIONS

1. Kelley, K., Darku, F. B., and Chattopadhyay, B. (2018). Accuracy in Parameter Estimation for Effect Sizes: A Sequential Approach. Psychological Methods, 23(2), 226–243. `http://psycnet.apa.org/doi/10.1037/met0000127`

2. Dontwi, I. K., Obeng-Denteh, W., Darku, F. B., Owusu-Mensah, I., & Amoah-Mensah, J. (2014). On the study of memorization trends. Physical Sciences Research International, 2(2), 44–48.

3. Dontwi, I. K., Obeng-Denteh, W., Bilson Darku, F., Tackie Otoo, J., Batsa Tetteh, F., Aboraa Sarpong, J., Ewoenam, K., Owusu-Mensah, I., and Amoah-Mensah, J (2013). Modeling memorization and forgetfulness using differential equations. Progress in Applied Mathematics, 6(1), 1-11. `DOI:10.3968/j.pam.1925252820130601.3421`

## CONFERENCES & PRESENTATIONS

- Contributed Talk: Joint Statistical Meeting, Baltimore, MD Aug 2017
- 3MT Competition: University of Texas at Dallas, Richardson, TX Apr 2017
- Poster Presentation: Conference of Texas Statisticians, Dallas, TX Mar 2017
- Speed Talk & Poster Presentation: Joint Statistical Meeting, Chicago, IL Aug 2016
- Workshop: Industrial Mathematical and Statistical Modeling Workshop, Raleigh, NC Jul 2016
- Poster Presentation: Conference of Texas Statisticians, San Antonio, TX Apr 2016
- Contributed Talk: UNCG Regional Mathematics and Statistics Conference, Greensboro, NC Nov 2015

## AFFILIATIONS

- Founding Member, Knights of Columbus, UTD, Richardson Mar 2017 - Present
- Student Member, Royal Statistical Society (RSS) Feb 2013 - Present
- Student Member, American Statistical Association (ASA) Dec 2014 - Present
- Student Member, Society of Industrial and Applied Mathematics Jan 2014 - Present
- Student Member, American Mathematical Society (AMS) Oct 2013 - Present
- Student Member, International Association of Black Actuaries Jul 2013 - Present
- Member, African Students Union, UTD, Richardson Aug 2013 - Present
- Member, Newman Catholic Ministry, UTD, Richardson Aug 2013 - Present

## HONORS & AWARDS

- Outstanding Teaching Assistant of the Year, UTD Maths Dept Apr 2018
- Quality & Productivity Section JSM Student Travel Award May 2017
- Betty & Gifford Johnson Scholarship Award Apr 2017
- Finalist for Three Minute Thesis (3MT) Competition, UTD Apr 2017
- Best Poster in Applied Statistics/Interdisciplinary category for Doctoral Students, Conference of Texas Statisticians Mar 2017
- ASA Travel Award to JSM Diversity Workshop and Mentoring Program Aug 2016
- SAMSI Travel Award to IMSM Workshop Jul 2016
- NSM Scholarship, University of Texas at Dallas Aug 2013 - Present
- IABA Foundation Scholarship, USA Jun 2013
- Overall Best Student, College of Science, 45th Congregation, KNUST Jun 2011
- Overall Best Student, Department of Actuarial Science, KNUST, Ghana Apr 2011
- Overall Best Student, Vice Chancellor's Awards, KNUST, Ghana Feb 2011
- Winner of Inter-Tertiary Institution Mathematics Competition, Ghana Apr 2010
- Dean's Honour List, KNUST, Ghana Apr 2009

## COMPUTER SKILLS

- **Programming**: R, SAS, Stata, VBA, MATLAB
- **Microsoft Office**: Word, Excel, PowerPoint, Access, Publisher, OneNote, Outlook
- **Other Packages**: SPSS, Minitab, LISREL, HLM, Mplus, LaTeX

## LEADERSHIP

- **Financial Secretary** Mar 2017 - Present
  Knights of Columbus, UTD, Richardson TX

- **Organizing Committee Head** Mar 2012
  3rd Inter-Tertiary Institution Conference for Tertiary Students' Marshallan Association of Ghana (TESMAG)
- **Co-Founder** Oct 2011
  Tertiary Students' Marshallan Association of Ghana, KNUST Chapter, Kumasi
- **Outdoor Function Committee Head** Aug 2010 - May 2011
  Legion of Mary, IMCS Pax Romana, KNUST Local, Kumasi, Ghana
- **Preasidium Vice President, Curia Vice President** Aug 2010 - May 2011
  Legion of Mary, IMCS Pax Romana, KNUST Local, Kumasi, Ghana
- **Committee Head** Aug 2010 - May 2011
  Trade and Technology Fair Committee for Actuarial Science Students' Association (ASSA), KNUST
- **Hostel Coordinator** Aug 2009 - May 2011
  International Movement of Catholic Students (IMCS Pax Romana), KNUST Local
- **Program Committee Head** Aug 2010
  2010 Western and Central Zonal Congress for the Junior Knights and Ladies of Marshall
- **House Secretary** Sep 2005 - May 2006
  St. Luke's House, St. Augustine's College, Cape Coast

## VOLUNTEER & SERVICE EVENTS

- Volunteered with UTD Newman Catholic Ministry at Vogel Alcove, a Dallas shelter geared towards homeless children. We set up, served and cleaned after lunch for the children and their families. Apr 2014

- Participated in outreach programme to rural areas such as the Rural Development and Education Project (RUDEP) organised by IMCS Pax Romana to the Sawla-Tuna-Kalba District (Northern Region of Ghana). May 2011