CONTRIBUTIONS TO NUMERICAL FORMAL CONCEPT ANALYSIS, BAYESIAN PREDICTIVE INFERENCE AND SAMPLE SIZE DETERMINATION

by

JUNHENG MA

Submitted in partial fulfillment of the requirements
For the degree of Doctor of Philosophy

Dissertation Advisor: Dr. Jiayang Sun and Dr. Joe Sedransk

Department of Statistics
CASE WESTERN RESERVE UNIVERSITY

January 2011
CASE WESTERN RESERVE UNIVERSITY
SCHOOL OF GRADUATE STUDIES

We hereby approve the dissertation of

JUNHENG MA

candidate for the Doctor of Philosophy degree

Committee Chair: ________________________________
Dr. Jiayang Sun
Dissertation Advisor
Professor
Department of Statistics

Committee: ________________________________
Dr. Joe Sedransk
Dissertation Advisor
Professor
Department of Statistics

Committee: ________________________________
Dr. Mark Schluchter
Professor
Department of Epidemiology & Biostatistics

Committee: ________________________________
Dr. Guo-Qiang Zhang
Professor
Department of Electrical Engineering & Computer Science

Committee: ________________________________
Dr. Tomas Radivoyevitch
Assistant Professor
Department of Epidemiology & Biostatistics

September 20, 2010
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table of Contents</td>
<td>iii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>viii</td>
</tr>
<tr>
<td>Acknowledgement</td>
<td>x</td>
</tr>
<tr>
<td>Abstract</td>
<td>xiii</td>
</tr>
</tbody>
</table>

## I Numerical Formal Concept Analysis

### 1 Formal Concept Analysis

1.1 Introduction to FCA | 2
1.2 Mathematics of FCA | 3
1.3 Properties of FCA | 5
1.4 Algorithms for FCA | 7
1.5 An Example of FCA | 8
1.6 Interpretation of Hasse Diagram | 10
1.7 Challenges to FCA | 11
References | 12

### 2 Numerical Formal Concept Analysis (nFCA)

2.1 Motivation of nFCA | 13
2.2 Introduction to nFCA | 15
2.3 Choice of Threshold
   2.3.1 Histogram Method | 17
   2.3.2 Confidence Interval Method | 18
2.4 Analysis of nFCA | 19
2.5 Algorithm for nFCA | 21
References | 23

### 3 Example, Application, Evaluation and Simulation Study

3.1 An Illustrative Example for nFCA | 26
3.2 Application of nFCA to Cardiovascular (CV) Traits | 37
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Gibbs sampler</td>
<td>92</td>
</tr>
<tr>
<td>12</td>
<td>Model Inference, Discussion and Further Work</td>
<td>96</td>
</tr>
<tr>
<td>12.1</td>
<td>Model Inference</td>
<td>96</td>
</tr>
<tr>
<td>12.2</td>
<td>Discussion</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>97</td>
</tr>
<tr>
<td>13</td>
<td>Introduction</td>
<td>101</td>
</tr>
<tr>
<td>13.1</td>
<td>Introduction and Literature Review</td>
<td>101</td>
</tr>
<tr>
<td>13.2</td>
<td>General Approaches for Sample Size Determination</td>
<td>103</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>104</td>
</tr>
<tr>
<td>14</td>
<td>Frequentist Approach</td>
<td>105</td>
</tr>
<tr>
<td>14.1</td>
<td>Original Frequentist Approach</td>
<td>105</td>
</tr>
<tr>
<td>14.2</td>
<td>Improvements for Frequentist Approach</td>
<td>110</td>
</tr>
<tr>
<td>14.2.1</td>
<td>Bootstrap Method</td>
<td>111</td>
</tr>
<tr>
<td>14.2.2</td>
<td>Minimum Difference Method</td>
<td>111</td>
</tr>
<tr>
<td>14.2.3</td>
<td>Correction Method</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>113</td>
</tr>
<tr>
<td>15</td>
<td>Bayesian approach</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>References</td>
<td>118</td>
</tr>
<tr>
<td>16</td>
<td>Simulation study</td>
<td>119</td>
</tr>
<tr>
<td>16.1</td>
<td>Experiment 1 (Small Differences)</td>
<td>119</td>
</tr>
<tr>
<td>16.2</td>
<td>Experiment 2 (Medium Differences)</td>
<td>122</td>
</tr>
<tr>
<td>16.3</td>
<td>Experiment 3 (Large Differences)</td>
<td>123</td>
</tr>
<tr>
<td>16.4</td>
<td>Experiment 4</td>
<td>124</td>
</tr>
<tr>
<td>17</td>
<td>Application, Discussion, and Conclusion</td>
<td>128</td>
</tr>
<tr>
<td>17.1</td>
<td>Application to Leukoplakia Lesions</td>
<td>128</td>
</tr>
<tr>
<td>17.2</td>
<td>Discussion and Conclusion</td>
<td>130</td>
</tr>
<tr>
<td>18</td>
<td>Appendix</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>137</td>
</tr>
</tbody>
</table>
# List of Tables

1.1 Cross table for the example binary data .................................. 4  
1.2 Formal concepts for the example binary data ................................. 9  
2.1 Height values for 3 person .......................................................... 13  
2.2 Scale context for example data ..................................................... 14  
2.3 Discretized/Reduced data ............................................................. 14  
2.4 Correlation coefficient matrix for CV traits .................................... 18  
3.1 Made-up correlation coefficient matrix for the Family example ............ 27  
3.2 Binary matrix after 1st thresholding for the Family example ............... 28  
3.3 Concept lattice constructed using faster concept analysis for the Family example .......................................................... 29  
3.4 1st updated correlation coefficient matrix for the Family example ........ 30  
3.5 Binary matrix after 2nd thresholding for the Family example ............. 31  
3.6 Concept lattice constructed by faster concept analysis after 2nd thresholding for the Family example .................................................. 31  
3.7 2nd updated correlation coefficient matrix for the Family example ........ 32  
3.8 Binary matrix after third thresholding for the Family example ........... 33  
3.9 Concept lattice constructed by faster concept analysis after 3rd thresholding for the Family example .................................................. 34  
3.10 Descriptions of 13 CV traits ......................................................... 38  
3.11 Correlation coefficient matrix of 13 CV traits ................................ 38  
3.12 Estimated correlation coefficient matrix of 13 CV traits .................... 46  
3.13 Performance comparison for population correlation coefficient matrix \( (n = 30) \) ................................................................. 48  
3.14 Performance comparison for sample correlation coefficient matrix \( (n = 30) \) ........................................................................ 49  
3.15 Performance comparison for population correlation coefficient matrix \( (n = 60) \) ................................................................. 49  
3.16 Performance comparison for sample correlation coefficient matrix \( (n = 60) \) ................................................................. 49
14.1 Contingency table for the count data $n_{ij}$ ............... 105
16.1 Averaged MSE of sample sizes ....................... 122
16.2 Averaged MSE of sample sizes ....................... 123
16.3 Averaged MSE of sample sizes ....................... 123
16.4 Winner table .................................................. 124
16.5 Performance comparison ............................... 127
17.1 Leukoplakia lesion locations regarding to bidi and non-bidi smoking habits ................................. 129
17.2 Sample sizes for Leukoplakia lesion ................... 130
List of Figures

1.1 Hasse diagram for the power set of \( \{x, y, z\} \) ........................................ 7
1.2 Hasse diagram for the example binary data ........................................ 10

2.1 Histogram of the CV traits correlation coefficient matrix ................. 19

3.1 Histogram of the correlation coefficients from the Family example ........................................ 28
3.2 Hasse diagram of the concept lattice for the Family example .......... 29
3.3 Histogram of the correlation coefficients after 1st thresholding for the Family example ........................................ 31
3.4 Hasse diagram of the concept lattice after 2nd thresholding for the Family example ........................................ 32
3.5 Histogram of the correlation coefficients after 2nd thresholding for the Family example ........................................ 33
3.6 Hasse diagram of the concept lattice ........................................ 35
3.7 Hierarchical cluster for the Family data ........................................ 36
3.8 Inherent structure for the Family data ........................................ 37
3.9 Hierarchical cluster of CV traits by nFCA: nFCA H-graph ............ 40
3.10 Inherent structure of CV traits by nFCA: nFCA I-graph ............ 41
3.12 Hierarchical cluster of CV traits for negative correlations by nFCA ........................................ 42
3.13 Inherent structure of CV traits for negative correlations by nFCA ........................................ 43
3.14 nFCA H-graph of CV traits for positive part ........................................ 44
3.15 nFCA I-graph of CV traits for positive part ........................................ 45
3.16 nFCA H-graph of CV traits for negative part ........................................ 46
3.17 nFCA I-graph of CV traits for negative part ........................................ 47

4.1 Visualization hierarchical cluster in Graphviz ........................................ 57
4.2 Visualization inherent structure in Graphviz ........................................ 58

16.1 Side-by-side modified Boxplot for Experiment 1 ........................................ 120
16.2 Side-by-side modified Boxplot for Experiment 1 after adding limit 900 ................................. 121
16.3 Side-by-side modified Boxplot for Experiment 2 after adding limit 900 ................................. 122
16.4 Side-by-side modified Boxplot for Experiment 3 after adding limit 900 ................................. 124
16.5 Plot of estimated sample sizes based on pilot sample size 30 for Experiment 1 .......................... 125
16.6 Plot of estimated sample sizes based on pilot sample size 80 for Experiment 1 .......................... 125
16.7 Plot of estimated sample sizes based on pilot sample size 20 for Experiment 2 .......................... 126
16.8 Plot of estimated sample sizes based on pilot sample size 30 for Experiment 2 .......................... 126
ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my co-advisors, Professor Jiayang Sun and Professor Joe Sedransk. They offered me tremendous help both in research and life. They guided me through this journey and encouraged me to think and create. They inspired me to become a better individual and statistician. I could not have asked for better advisors. I would like to thank Professor Jiayang Sun for the financial assistantship support during my graduate studies and also for meeting me at her lovely house. I also would like to thank Professor Joe Sedransk for letting me stay in his beautiful home. I have enjoyed my time there so much.

My committee members, Professor Mark Schluchter, Professor Guo-Qiang Zhang, and Professor Tomas Radivojevitch deserve a big thank you for the insightful discussions and valuable advice regarding my projects that has improved my thesis. I would like to thank Professor Jill Korbin, Professor Wojbor Woyczynski, Professor Patricia Williamson, Professor Stanislaw Szarek, Professor David Gurarie, and Professor Pingfu Fu for their advice and help through my years in graduate school. I would like to thank Professor James Alexander and Professor Jill Korbin for the graduate assistantship during my graduate studies. I also would like to thank Adam Troy and Professor Guo-Qiang Zhang for the permission of using their faster concept analysis program.

I would like to thank Ms. Sharon Dingess, Mr. Vinay Augustine, and Mr. Joe Teagno for their assistance and help through all these years. I also would like to thank my classmates and friends in the Department of Statistics, Mathematics, Electrical Engineering & Computer Science, and Epidemiology & Biostatistics for their friendship and all the fun at Case. In addition, I would to thank Ms. Van Anh Tran for helping me with the faster concept analysis program.

The NSF is sincerely appreciated for providing funding for part of my research.

Last, but not the least, I would thank my parents, Erhei Ma and Gaicui Fan, and
my sisters, Junlan Ma and Ruilan Ma, for their unconditional love and support. A special thank you to my wife, Junying Liu, for her support accompanied me through every success and failure. Thank you to my daughter, Sydney – her smile always cheers me up.
to Junying and Sydney
Contributions to Numerical Formal Concept Analysis, Bayesian Predictive
Inference and Sample Size Determination

Abstract

by

JUNHENG MA

This dissertation contributes to three areas in Statistics: Numerical Formal
Concept Analysis (nFCA), Bayesian predictive inference and sample size deter-
mination, and has applications beyond statistics.

Formal concept analysis (FCA) is a powerful data analysis tool, popular in
Computer Science (CS), to visualize binary data and its inherent structure. In the
first part of this dissertation, Numerical Formal Concept Analysis (nFCA) is de-
veloped. It overcomes FCA’s limitation to provide a new methodology for analyzing
more general numerical data. It combines the Statistics and Computer Science
graphical visualization to provide a pair of nFCA graphs, H-graph and I-graph, to
reveal the hierarchical clustering and inherent structure among the data. Compar-
ing with conventional statistical hierarchical clustering methods, nFCA provides
more intuitive and complete relational network among the data. nFCA performs
better than the conventional hierarchical clustering methods in terms of the Cophe-
netic correlation coefficient which measures the consistency of a dendrogram to
the original distance matrix. We have also applied nFCA to cardiovascular (CV)
traits data. nFCA produces consistent results to the earlier discovery and provides
a complete relational network among the CV traits.
In the second part of this dissertation, Bayesian predictive inference is investigated for finite population quantities under informative sampling, i.e., unequal selection probabilities. Only limited information about the sample design is available, i.e., only the first-order selection probabilities corresponding to the sampled units are known. We have developed a full Bayesian approach to make inference for the parameters of the finite population and also predictive inference for the non-sampled units. Thus we can make inference for any characteristic of the finite population quantities. In addition, our methodology, using Markov chain Monte Carlo, avoids the necessity of using asymptotic approximations.

Sample size determination is one of the most important practical tasks for statisticians. There has been extensive research to develop appropriate methodology for sample size determination, say, for continuous, or ordered categorical outcome data. However, sample size determination for comparative studies with unordered categorical data remains largely untouched. In terms of statistical terminology, one is interested in finding the sample size needed to detect a specified difference between the parameters of two multinomial distributions. For this purpose, in the third part of this dissertation, we have developed a frequentist approach based on a chi-squared test to calculate the required sampled size. Three improvement for the original frequentist approach (using bootstrap, minimum difference and asymptotic correction) have been proposed and investigated. In addition, using an extension of a posterior predictive p-value, we further develop a simulation-based Bayesian approach to determine the required sample size. The performance of these methods is evaluated via both a simulation study and a real application to Leukoplakia lesion data. Some asymptotic are also provided.
Part I

Numerical Formal Concept Analysis
Chapter 1

Formal Concept Analysis

1.1 Introduction to FCA

Formal Concept Analysis (FCA) was first introduced by Rudolf Wille in 1982 and originally developed as a subfield of applied mathematics based on the “concept” and “concept hierarchy”. Formal Concept Analysis now refers to an unsupervised machine learning technique which provides conceptual tools to identify conceptual structure among the elements of a data set. It is much more popular in computer science than in statistics and other fields but has wide applications in data mining, graphical knowledge representation, and information management. Since FCA’s inception in 1982, hundreds of articles have been published. Ganter and Wille (1999)[5] provided the mathematical foundation of FCA. Major books include Ganter and Wille (1999)[5], Davey and Priestley (2002)[4], and Carpineto and Romano (2004)[2]. The Darmstadt group in Germany was the driving force in the early days of FCA. Nowadays, FCA has developed into a very active research field nationally and internationally. Every year three major international conferences are dedicated to FCA and related methods: International Conference on Formal Concept Analysis (ICFCA), International Conference on Conceptual Structures (ICCS), and Concept Lattices and their Applications (CLA). FCA now has applications in many different disciplines such as computer science, statistics,
applied mathematics, medicine, psychology, social science, artificial intelligence (AI), and information retrieval.

Specifically, FCA is a data analysis method to visualize binary data and its inherent structures, implications and dependencies. The input of FCA is a binary data matrix which represents objects and their properties (called attributes); the output of FCA is all the natural clusters of attributes and all the natural clusters of objects in the original binary data matrix. The results of FCA then can be graphically visualized and represented as a Hasse diagram (also called a line diagram) as explained in Section 1.6. Therefore, one nice feature of FCA is its capability of producing graphical visualization of the inherent structures among the binary data. The binary data are visualized in the form of a concept lattice (also called Galois lattice), which makes the hidden structure in the data more transparent and informative. In other words, one can extract relevant ontology (from Wikipedia: ontology is the philosophical study of the nature of being, existence or reality in general, as well as the basic categories of being and their relations) from a set of objects and their attributes.

1.2 Mathematics of FCA

Given a set of objects and corresponding attributes, the relationship between the objects and the attributes can be represented in a cross table. In the cross table, a Boolean value, e.g., 1 or 0, or “×” or empty, is given to indicate if an object possesses an attribute. See Table 1.1, for example. The mathematical structure that is used to describe the table crosses is called a formal context or context. Formally, a triple \((G,M,I)\) is called a formal context if \(G\) and \(M\) are object and attribute sets \(^1\) and \(I \subseteq G \times M\) denotes a relation between object set \(G\) and attribute set \(M\). The elements in \(G\) are called objects and the elements in \(M\) are called attributes. \(I\) is called the incidence of the formal context \((G,M,I)\). For \(A \subseteq G\), we

\(^1\) \(G\) stands for “Gegenstand” and \(M\) stands for “Merkmal”, which are object and attribute in German
Table 1.1: Cross table for the example binary data

<table>
<thead>
<tr>
<th></th>
<th>composite</th>
<th>odd</th>
<th>even</th>
<th>prime</th>
<th>square</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
<td>×</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>×</td>
<td></td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

define

\[ A' = \{ m \in M | (g, m) \in I \text{ for all } g \in A \}, \] (1.1)

the set of of all attributes of the objects in A, i.e., \( A' \) is a dual set of \( A \). Similarly, given \( B \subseteq M \) we define

\[ B' = \{ g \in G | (g, m) \in I \text{ for all } m \in B \}, \] (1.2)

the set of all objects of the attributes in \( G \), i.e., \( B' \) is a dual set of \( B \).

A *formal concept* consists of a set of objects \( G_i \), a set of attributes \( M_i \), and the indications of their binary relation \( G_i \times M_i \), i.e., which object has which attributes, such that

1. \( G_i \subseteq G \),
2. \( M_i \subseteq M \),
3. Every object in \( G_i \) has all the attributes in \( M_i \),
4. For any object which is not in \( G_i \), there is at least one attribute in \( M_i \) that the object doesn’t have,
5. For any attribute which is not in $M_i$, there is at least one object in $G_i$ which doesn’t have this attribute.

Further, $G_i$ is called the extent of the formal concept and $M_i$ is called the intent of the formal concept. The extent covers all the objects belonging to the concept and the intent comprises all the attributes. A formal concept usually is represented as the triple $(G_i, M_i, I_i)$. The elements of $G_i$ are called subjects and the elements of $M_i$ are called attributes.

A natural cluster of objects is the set of all objects that share a common subset of attributes. A natural cluster of attributes is the set of all attributes shared by one of the natural object clusters. The natural clusters of objects have a one-to-one correspondence with the natural clusters of attributes. A concept lattice is the collection of all the natural clusters of objects and all the natural clusters of attributes combined together. The pair of a natural attribute cluster and a corresponding natural object cluster constitutes a formal concept. Thus, the family of all ordered formal concepts forms a concept lattice. Formally, the set of all ordered formal concepts in $(G, M, I)$ is called the concept lattice of $(G, M, I)$.

### 1.3 Properties of FCA

Regarding to the objects and attributes, for $A, A_1, A_2 \subseteq G$ and $B, B_1, B_2 \subseteq M$, we have following simple relationships,

1. $A_1 \subseteq A_2 \implies A'_1 \subseteq A'_2,$

2. $B_1 \subseteq B_2 \implies B'_1 \subseteq B'_2,$

3. $A \subseteq A'''$ and $A' = A''',$

4. $B \subseteq B'''$ and $B' = B''',$

5. $A \subseteq B' \iff A' \subseteq B,$
where $A'' = (A')'$, $A''' = (A'')'$, $B'' = (B')'$, and $B''' = (B'')'$.

An object and attribute set, $(A,B)$, is a formal concept of $(G,M,I)$ if and only if

$$A \subseteq G, \quad B \subseteq M, \quad A' = B, \quad \text{and} \quad B' = A. \quad (1.3)$$

For a formal concept, $(A,B)$, $A$ is called the formal objects of $B$ and $B$ is called the formal attributes of $A$.

The formal concepts of a concept lattice are naturally ordered. Suppose $(A_1,B_1)$ and $(A_2,B_2)$ are two formal concepts, the order between formal concepts is determined by subconcept-superconcept relation $(\leq$ and $\geq)$ which is defined as

$$(A_1,B_1) \leq (A_2,B_2) \iff A_1 \subseteq A_2 \iff B_1 \subseteq B_2, \quad (1.4)$$

$$(A_1,B_1) \geq (A_2,B_2) \iff A_1 \supseteq A_2 \iff B_1 \supseteq B_2, \quad (1.5)$$

where $A_1, A_2 \subseteq G$ and $B_1, B_2 \subseteq M$. The subconcept-superconcept relation is transitive.

A concept lattice can be graphically visualized via a Hasse or line diagram. A Hasse diagram is a simple picture of a finite partially ordered set, a drawing of the transitive reduction of the partial order. For example, the power set of $\{x,y,z\}$ can be partially ordered by inclusion and represented by the Hasse Diagram shown in Figure 1.1.
The Hasse diagram constructed from the concept lattice contains enough information which can be used to fully recover the original binary data matrix. In other words, there is no information lost during the construction of the diagram from the concept lattice.

There are many other properties of FCA which are not presented here. Regarding the properties of FCA, Ganter and Wille (1999)[5] is one of the excellent references.

1.4 Algorithms for FCA

A lot of algorithms have been developed to construct the concept lattice from a given binary data matrix. See Kuznetsov and Obiedkov (2001)[6] for a comprehensive survey and evaluation of those FCA algorithms. Implementation of the FCA algorithm may vary in many different aspects, but the main idea is based on the fact that each edge of the Hasse diagram connects two formal concepts and thus one can build the concept lattice by starting with the empty set and finding new formal concepts from the neighbor of the known formal concepts. The foundation of the main idea comes from the general construction of a concept lattice
from a given binary data matrix, described in Birkhoff’s *lattice theory* (Birkhoff (1940)[1]).

When the numbers of objects and attributes become large, a FCA algorithm could take a long time to run. Thus, developing efficient algorithms has become an important goal. Among these algorithms an efficient FCA algorithm, “faster concept analysis”, developed by Troy, Zhang, and Tian (2007)[7], exhibits significant improvement for a variety of input profiles.

### 1.5 An Example of FCA

A simple example of FCA will be given to illustrate the basic concepts. We want to show how FCA works and how to use FCA to graphically visualize the concept lattice and the inherent structures among the elements of a data set.

The example binary data set under consideration has ten objects

\[ G = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \]

and five attributes

\[ M = \{\text{composite, odd, even, prime, square}\}, \]

representing the properties of these numbers (objects). The objects and attributes in this example along with their binary relations have been represented in Table 1.1.

Each cross in the cross table indicates that the corresponding object possesses the corresponding attribute. For example, 4 is a composite, even, and square number. And an empty cell indicates that the corresponding object does not have the corresponding attribute. All the objects, attributes, and their binary relations in this cross table is called a *formal context*. In addition, \( G = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\} \) is called the *object set* and \( M = \{\text{composite, odd, even, prime, square}\} \) is called the *attribute set*. The binary relation \( I \) between \( G \) and \( M \), i.e., the crosses in Table 1.1, is called the *incidence* of the formal context.
An example of a formal concept is the pair of object set \( \{2, 4, 6, 8, 10\} \) and attribute set \( \{\text{even}\} \). This can be easily verified by using the definition of formal concept. The whole set of formal concepts is listed below in Table 1.2.

<table>
<thead>
<tr>
<th>Objects</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>{1,2,3,4,5,6,7,8,9,10}</td>
<td>\emptyset</td>
</tr>
<tr>
<td>{4,6,8,9,10}</td>
<td>{\text{composite}}</td>
</tr>
<tr>
<td>{1,4,9}</td>
<td>{\text{square}}</td>
</tr>
<tr>
<td>{2,4,6,8,10}</td>
<td>{\text{even}}</td>
</tr>
<tr>
<td>{1,3,5,7,9}</td>
<td>{\text{odd}}</td>
</tr>
<tr>
<td>{2,3,5,7}</td>
<td>{\text{prime}}</td>
</tr>
<tr>
<td>{4,6,8,10}</td>
<td>{\text{composite, even}}</td>
</tr>
<tr>
<td>{4,9}</td>
<td>{\text{composite, square}}</td>
</tr>
<tr>
<td>{1,9}</td>
<td>{\text{odd, square}}</td>
</tr>
<tr>
<td>{4}</td>
<td>{\text{composite, even, square}}</td>
</tr>
<tr>
<td>{9}</td>
<td>{\text{composite, odd, square}}</td>
</tr>
<tr>
<td>{2}</td>
<td>{\text{even, prime}}</td>
</tr>
<tr>
<td>{3,5,7}</td>
<td>{\text{odd, prime}}</td>
</tr>
<tr>
<td>\emptyset</td>
<td>{\text{composite, odd, even, prime, square}}</td>
</tr>
</tbody>
</table>

Table 1.2: Formal concepts for the example binary data

The collection of all the ordered formal concepts is called the concept lattice. And the concept lattice can be represented and graphically visualized by a Hasse diagram. For this example, the corresponding Hasse diagram is Figure 1.2.
The interpretation of Figure 1.2 is given in Section 1.6.

1.6 Interpretation of Hasse Diagram

The Hasse diagram constructed from the concept lattice consists of all the formal concepts of the concept lattice and the subconcept-superconcept relations between the formal concepts. Each node of the Hasse diagram represents a formal concept and the edge connecting two nodes corresponds to the order relation between those two formal concepts. For each node, the first part denotes the formal object of the
corresponding formal concept and the second part is the formal attribute of the formal concept. The order between two formal concepts can be determined based on the subconcept-superconcept relation which is described in Section 1.3.

The top and bottom nodes in the Hasse diagram are special. The top node takes all the attributes as its intension and its extension usually is empty but is not necessarily empty. The bottom node takes all the objects as its extension. Similar to the top node, the intension of the bottom node usually is empty but is not necessarily empty either. The top node can be thought as the null formal concept and the bottom node can be thought as the universal formal concept.

Each node in the Hasse diagram indicates a natural cluster of objects and corresponding attributes. For example, the object set \{1, 4, 9\} forms a cluster since they share the same attribute \{square\}. Within this small cluster, object 9 is more dominant than object 1 and object 4 in the sense that it possesses more attributes than each of the other two objects. Generally speaking, the nodes in the Hasse diagram exhibit the cluster information for the objects and their corresponding attributes, and the subconcept-superconcept relation determines the relationship within the cluster. Therefore, FCA is a clustering method and can provide useful and informative clustering results for many different applications.

1.7 Challenges to FCA

As we have already mentioned FCA was built for binary relational data, and this limits its application. Various attempts, e.g., Choi, Huang, Lam, Potter, Laubenbacher, and Duca (2006)[3], have been made to extend FCA to numerical data. However, most of them discretize or reduce the numerical data to binary data and then use FCA, which would lose much of the information in the original numerical data matrix.

Another obvious disadvantage of FCA is that it will produce a huge number of formal concepts for a large-scale data set. This will make the concept lattice
and the Hasse diagram less informative and overwhelm the analyst.

References


Chapter 2

Numerical Formal Concept Analysis (nFCA)

2.1 Motivation of nFCA

Our objective is to extend FCA to accommodate numerical data. We start with an example. Suppose that objects are individual persons and the attribute is \textit{height}. Clearly, \textit{height} is not a binary-valued attribute since it may take multiple values. Table 2.1 below shows the values of the attribute \textit{height} for 3 individuals. In order to convert this numerical data to binary data and apply FCA, one may discretize \textit{height} into 3 categories as shown in Table 2.2. This kind of discretization is called \textit{scale context}. Based on the above defined scale context, we can represent the data in Table 2.1 as binary data under FCA framework as given in Table 2.3.

<table>
<thead>
<tr>
<th>Object</th>
<th>Attribute (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.85</td>
</tr>
<tr>
<td>2</td>
<td>1.75</td>
</tr>
<tr>
<td>3</td>
<td>1.65</td>
</tr>
</tbody>
</table>

Table 2.1: Height values for 3 person

After reducing the numerical data to the binary data, to which FCA can be applied to conduct the data analysis. This is the procedure often used by researchers...
and analysts to extend FCA to numerical data and is formally called conceptual scaling. The key of conceptual scaling is to define an appropriate scale context to discretize or reduce the multi-valued data to binary data. If the attributes of a data set include more than one multi-valued attribute, conceptual scaling can be applied to each of them to discretize multi-valued attributes to single-valued attributes, i.e., binary data. For example, Choi, Huang, Lam, Potter, Laubenbacher, and Duca (2006)[1] applied conceptual scaling to gene expression data to reduce it to binary data to apply FCA. Priss (2006)[3] discussed the applications of conceptual scaling in information science. Yang, Kim, Hwang, and Choi (2008)[5] developed new data mining tools for numerical data using conceptual scaling. A special case of conceptual scaling is thresholding the numerical data. Note that thresholding a numerical data is equivalent to discretizing the continuous domain of the numerical data into two separate domains. However, with only one threshold, much numerical information is lost; with multiple thresholds, how to merge information from multiple FCAs consistently becomes a challenge.

Inspired by FCA and conceptual scaling, our proposed new statistical procedure uses multiple thresholds to threshold a numerical data matrix and then combines the clustering information from each thresholding consistently. And finally, the combined clustering information can show the hierarchical clustering structure

<table>
<thead>
<tr>
<th>Height (H)</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H \geq 1.80$</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1.70 \leq H &lt; 1.80$</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H &lt; 1.70$</td>
<td>×</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.2: Scale context for example data

<table>
<thead>
<tr>
<th>Object</th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>×</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3: Discretized/Reduced data
of the numerical data and also describe the inherent structure among the objects under consideration. Note that conceptual scaling only applies for a single multivalued attribute. In order to apply conceptual scaling to the whole numerical data we will focus on correlation coefficient matrix or more broadly distance matrix of a set of objects. Later on, we will discuss the way to generalize nFCA to general numerical objects and attributes data. Using multiple thresholds determined statistically to threshold the numerical data matrix will lose much less information than single conceptual scaling. In other words, the proposed procedure extends FCA to numerical data by applying multiple thresholdings along with the combination of the clustering results from each individual thresholding to describe the hierarchical clustering as well as the inherent structure among the original numerical data.

2.2 Introduction to nFCA

Given numerical data, their correlation coefficient matrix or distance matrix can be computed among data points. If the interest is at the relationship between the variables, we should work with the correlation coefficient matrix \( R = (r_{ij}) \) and apply nFCA to it. If the interest is at clustering data points, we should apply nFCA to the distance matrix \( D = (d_{ij}) \) computed based on the objective of a study. No matter it is a correlation matrix or a distance matrix, we have a numerical matrix. See Xu (2007)[4] and Johnson and Wichern (2007)[2] for different uses of distances. In this section, we develop nFCA to show the hierarchical clustering among the variables based on their correlation coefficient matrix. We are focusing on correlation coefficient matrix now. It will be clear that nFCA can work for any numerical matrix.

Given a correlation coefficient matrix of a set of variables, a threshold can be chosen and then used to threshold the numerical data matrix. The cells in the numerical data matrix are set to 1 if they are greater than or equal to the specified
threshold and 0 otherwise. After the thresholding, we will have a binary data and FCA can be applied to construct the concept lattices. Based on the concept lattice, a Hasse diagram can be built to illustrate the cluster information among the objects. We are particularly interested in the small clusters in the Hasse diagram. For each qualified small cluster in the Hasse diagram, we group the members of the cluster together and treat them as one single new object. After all the small clusters in the diagram are grouped, we compute the correlation coefficient matrix for newly defined objects and other non-grouped objects based on a choice of distance measure, e.g., single linkage method, complete linkage method, and average linkage method. The choice of distance measure highly depends on the application. For correlation coefficient matrix, single linkage method is used to compute the correlation coefficient matrix. After the correlation coefficient or distance matrix is computed, we repeat the above procedure to find the cluster information among the newly defined objects and other non-grouped objects. This procedure is applied iteratively until all the objects are grouped together.

The hierarchical cluster among the set of objects can be described based on the resulting cluster information from each thresholding step. For example, during a thresholding step objects A, B and C are grouped together and from the Hasse diagram we also know object C dominates objects A and B, i.e., object C is the kernel or core object of the cluster formed by objects A, B, and C. Then objects A, B and C are grouped together and treated as a new object D. In the following thresholding step, if object D is further dominated by another object E, among the cluster \( \{A, B, C, E\} \) we can see object E dominants the small cluster \( \{A, B, C\} \) and among this small cluster objects A and B are dominated by object C. Following this basic principal, the hierarchical structure can be generated by combining the clustering information from each thresholding step. The resulting hierarchical structure can be graphically visualized.

The inherent structure among the numerical data can be built based on the resulting hierarchical cluster. Starting from one of the kernel clusters or core clus-
ters which usually come from the first thresholding step, the objects which directly connect to the chosen kernel cluster are connected to the objects in the kernel cluster which have the highest correlation coefficients or smallest distances with the objects. The length of the connection denotes the correlation coefficient or distance between the two objects. Short lengths indicate strong correlations or small distances between the two objects, while long lengths indicate weak correlations or large distances between the two objects. If there is a small cluster immediately outside of the kernel cluster, each object in the small cluster is connected to the kernel cluster in the same way. After all the directly connected objects or clusters are connected, we group them together into the kernel cluster and treat it as a new kernel cluster. This procedure is repeated until all the objects are grouped into the kernel cluster. The resulting inherent structure can be graphically visualized.

For correlation coefficient matrix, the hierarchical cluster is constructed based on the correlation coefficients between the variables. The inherent structure among the data shows the most significant connections between the variables. nFCA is applicable to a general numerical matrix representing more general distances between variables or objects.

### 2.3 Choice of Threshold

Choice of threshold is one of the key steps in nFCA. For the sorted numerical values in a given numerical data matrix, the threshold is chosen from the upper part in order to capture the strong or major connections among all the objects. Thus, without considering the random errors in the original numerical data matrix, the optimal choice of the threshold should be the one close to the largest value in a given numerical data matrix. In practice, we suggest to choose threshold in two different ways.
2.3.1 Histogram Method

Histogram is used to graphically show the distribution of frequencies in the data. For the choice of threshold, the basic guideline is to choose the threshold to capture the strong relationship among the data matrix. So in order to extract the major relationship between the objects, the threshold should be chosen such that a reasonable set of objects are involved in the thresholding, i.e., the total number of objects involved in the thresholding should not be either too small or too large. If there are only few objects involved, not too much useful information will be given after applying the FCA algorithm. On the other hand, if too many objects are involved in the thresholding, the major relationship between objects is overwhelmed. Thus, a reasonable threshold can be chosen by looking at the Histogram of the numerical values in the original numerical data matrix. Using Histogram to choose threshold is relatively simple and easy to implement.

Here is one example of choosing threshold based on the Histogram. Suppose the CV traits correlation coefficient matrix from Nadeau, Burrage, Restivo, Pao, Churchill, and Hoit (2003)[23] under consideration is shown in Table 2.4. The

<table>
<thead>
<tr>
<th></th>
<th>0.92</th>
<th>0.28</th>
<th>0.10</th>
<th>-0.02</th>
<th>0.67</th>
<th>0.42</th>
<th>0.41</th>
<th>0.14</th>
<th>0.71</th>
<th>-0.04</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.92</td>
<td>0.29</td>
<td>0.11</td>
<td>-0.05</td>
<td>0.68</td>
<td>0.57</td>
<td>0.29</td>
<td>0.17</td>
<td>0.68</td>
<td>-0.05</td>
</tr>
<tr>
<td>0.92</td>
<td>1</td>
<td>0.29</td>
<td>0.11</td>
<td>-0.05</td>
<td>0.68</td>
<td>0.57</td>
<td>0.29</td>
<td>0.17</td>
<td>0.68</td>
<td>-0.05</td>
</tr>
<tr>
<td>0.28</td>
<td>0.29</td>
<td>1</td>
<td>0.90</td>
<td>-0.76</td>
<td>0.84</td>
<td>0.32</td>
<td>0.61</td>
<td>0.92</td>
<td>-0.42</td>
<td>-0.64</td>
</tr>
<tr>
<td>0.10</td>
<td>0.11</td>
<td>0.90</td>
<td>1</td>
<td>-0.95</td>
<td>0.70</td>
<td>0.24</td>
<td>0.53</td>
<td>0.78</td>
<td>-0.52</td>
<td>-0.75</td>
</tr>
<tr>
<td>-0.02</td>
<td>-0.05</td>
<td>-0.76</td>
<td>-0.95</td>
<td>1</td>
<td>-0.55</td>
<td>-0.22</td>
<td>-0.36</td>
<td>-0.58</td>
<td>0.52</td>
<td>0.72</td>
</tr>
<tr>
<td>0.67</td>
<td>0.68</td>
<td>0.84</td>
<td>0.70</td>
<td>-0.55</td>
<td>1</td>
<td>0.47</td>
<td>0.67</td>
<td>0.80</td>
<td>0.02</td>
<td>-0.51</td>
</tr>
<tr>
<td>0.42</td>
<td>0.57</td>
<td>0.32</td>
<td>0.24</td>
<td>-0.22</td>
<td>0.47</td>
<td>1</td>
<td>-0.33</td>
<td>0.27</td>
<td>0.26</td>
<td>-0.35</td>
</tr>
<tr>
<td>0.41</td>
<td>0.29</td>
<td>0.61</td>
<td>0.53</td>
<td>-0.36</td>
<td>0.67</td>
<td>-0.33</td>
<td>1</td>
<td>-0.59</td>
<td>-0.13</td>
<td>-0.22</td>
</tr>
<tr>
<td>0.14</td>
<td>0.17</td>
<td>0.92</td>
<td>0.78</td>
<td>-0.58</td>
<td>0.80</td>
<td>0.27</td>
<td>-0.59</td>
<td>1</td>
<td>-0.52</td>
<td>-0.58</td>
</tr>
<tr>
<td>0.71</td>
<td>0.68</td>
<td>-0.42</td>
<td>-0.52</td>
<td>0.52</td>
<td>0.02</td>
<td>0.26</td>
<td>-0.13</td>
<td>-0.52</td>
<td>1</td>
<td>0.41</td>
</tr>
<tr>
<td>-0.04</td>
<td>-0.05</td>
<td>-0.64</td>
<td>-0.75</td>
<td>0.72</td>
<td>-0.51</td>
<td>-0.35</td>
<td>-0.22</td>
<td>-0.58</td>
<td>0.41</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2.4: Correlation coefficient matrix for CV traits

Histogram of the numerical values in this correlation coefficient matrix is shown in Figure 2.1. A reasonable choice of the threshold based on the Histogram for this numerical data is about 0.8 as the cutoff value for the largest Histogram bar.
is 0.8. If 0.8 is used as the threshold, we see from Table 2.4 there will be in fact six objects involved in the thresholding. Specifically, there are five values greater than 0.8. The first value 0.92 relates objects 1, 2, and 9; the second value 0.9 relates object 3 and object 4; the third value 0.84 relates object 3 and object 6; the forth value 0.8 relates objects 6 and 9. Hence, there are total 6 distinct objects \{1,2,3,4,6,9\} involved in this thresholding.

### 2.3.2 Confidence Interval Method

If the given numerical data matrix is correlation coefficient matrix and also the number of samples used to calculate the sample correlation coefficients is known, the threshold can be chosen by confidence interval method. For the sample correlation coefficient, after applying the Fisher transformation, it approximately has a
normal distribution. To be specific, suppose $r$ is the sample correlation coefficient between 2 random variables, $X$ and $Y$, based on $n$ independent sample points of $(X, Y)$, and $Z$ is the Fisher transformation of $r$, i.e.,

$$Z = \frac{1}{2} \ln \frac{1+r}{1-r}.$$ 

Then $Z$ approximately has a normal distribution,

$$Z \sim N \left( \frac{1}{2} \ln \frac{1+\rho}{1-\rho}, \frac{1}{n-3} \right),$$

where $\rho$ is the population correlation coefficient between $X$ and $Y$. As shown in the formula, the mean of the normal distribution is related to the true population correlation coefficient $\rho$ and the variance only involves the sample size $n$. The confidence interval of the population correlation coefficient can be obtained by transforming back the confidence interval related to this normal distribution. For above normal distribution, the $100(1-\alpha)\%$ confidence interval for the mean $\mu = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ is

$$\left( Z - z_{(1-\alpha/2)} \sqrt{\frac{1}{n-3}}, \quad Z + z_{(1-\alpha/2)} \sqrt{\frac{1}{n-3}} \right),$$

where $Z$ is the Fisher transformation of $r$, $n$ is the sample size, and $z_{(1-\alpha/2)}$ is the upper $\alpha/2$ quantile of the standard normal distribution. After transforming back the above confidence interval, the $100(1-\alpha)\%$ confidence interval for population correlation coefficient $\rho$ is

$$\left( \frac{\exp(2l_z) - 1}{\exp(2l_z) + 1}, \quad \frac{\exp(2u_z) - 1}{\exp(2u_z) + 1} \right),$$

where

$$l_z = Z - z_{(1-\alpha/2)} \sqrt{\frac{1}{n-3}}, \quad u_z = Z + z_{(1-\alpha/2)} \sqrt{\frac{1}{n-3}}.$$
In order to choose the threshold using the confidence interval, we start with the largest value in the numerical data matrix. After getting the confidence interval of the largest value, we check how many numerical values in the original data matrix fall into this confidence interval. If there are only a few numerical values fall into the confidence interval, then we move to the second largest value. Otherwise, we take the lower bound of the confidence interval as the threshold. This method only applies to the correlation coefficient matrix for which the sample size \( n \) used in computing these correlation coefficients is known.

For illustration purpose, suppose the sample size used to calculate the previous correlation coefficient matrix, i.e., Table 2.4, is 30. We start with the largest sample correlation coefficient value, 0.92. Using the above formulas, the 95% confidence interval for the population correlation coefficient is \((0.837, 0.962)\). There are four numerical values in the correlation coefficient matrix fall into this confidence interval. Therefore, based on the confidence interval method 0.837 will be a good choice as a threshold for the example correlation coefficient data.

No matter which method is used to choose the threshold the number of involved subjects should not be too small or too large. Typically, a number between 3 and 10 is about right.

### 2.4 Analysis of nFCA

Suppose \( R = (r_{ij}) \) is the correlation coefficient matrix under consideration. Denote the distribution of \( r_{ij} \) to be \( f(x) \) by ignoring their membership information. If there exists a simple random sample (SRS) from \( f(x) \), with an appropriate choice of the window width, \( \hat{f}_{Hist}(x) \) is consistent to \( f(x) \), where \( \hat{f}_{Hist}(x) \) is the Histogram density estimator of \( f(x) \). In real data situations, \( r_{ij} \) may not be available, but were estimated using the sampled correlation coefficient \( \hat{r}_{ij} \) based on \( n \) random sample points. So they are correlated. In this case, we show that \( \hat{f}_{Hist}(x) \) is still consistent to \( f(x) \) under some regularity conditions. This justifies our Histogram method for
threshold choice, and asymptotically, the nFCA result should be consistent to the true structure of the underline population.

Specifically, let $R = (\hat{r}_{ij})_{k \times k}$ be a sample correlation coefficient matrix and $f^*(x)$ be the distribution of $\hat{r}_{ij}$. Suppose that $\hat{f}_{Hist}(x)$ is the Histogram density estimate based on elements in the upper triangular part of $R$ without the diagonal elements. Denote this set of $r_{ij}$ in the upper triangular $R$ by $\tilde{X} = \{X_i, i = 1, 2, \ldots, t\}$, where $t = k(k - 1)/2$. For $x \in B_j$, where $B_j = [b_j, b_{j+1})$ is the $j$th bin of the Histogram, the Histogram density estimate can then be written as

$$\hat{f}_{Hist}(x) = \frac{1}{th} \sum_{i=1}^{t} I_{x_i \in B_j}$$

where $h = b_{j+1} - b_j$ is the bin width of the Histogram, and $I$ is an indicator function. When $\hat{r}_{ij}$ is the sample correlation coefficient based on $n$ iid multivariate sample, $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n \in \mathbb{R}^k$, with $\text{cor}(\tilde{X}) = (r_{ij})$, $\hat{r}_{ij} = r_{ij} + (\hat{r}_{ij} - r_{ij}) \sim f^*(x) = f(x) * g_n(x)$, which is the convolution of $f(x)$ and $g_n(x)$, where $g_n(x)$ is a normal distribution with variance in the order of $1/n$. In terms of the mean and variance of $\hat{f}_{Hist}(x)$ we have the following proposition.

**Proposition 2.4.1.** The mean of $\hat{f}_{Hist}(x)$ is

$$E(\hat{f}_{Hist}(x)) = f^*(x) + \frac{1}{2} f^{*'}(x)(h + 2b_j - 2x) + O(h^2)$$

$$= f(x) * g_n(x) + \frac{1}{2} f^{*'}(x)(h + 2b_j - 2x) + O(h^2)$$

as $t \to \infty, h \to 0$. If we further assume that $\sum_{m=l+1}^{t} \left( f_{x_l,x_m}^*(x) - f^*_2(x) \right) < M$, where $f_{x_l,x_m}^*(x) \times x$ is the joint density function of $(X_l, X_m)$ evaluated at $x_l = x_m = x$ and $M$ is not related to $t$, the variance of $\hat{f}_{Hist}(x)$ is

$$\text{Var}(\hat{f}_{Hist}(x)) = \frac{1}{th} f^*(x) + O \left( \frac{C(x)}{t} \right) + O(B(x)h).$$
The detailed proof of Proposition 2.4.1 is provided in Appendix. The rational for the consistency in the mean integrated square error (MISE) of \( \hat{f}_{Hist}(x) \) is that under some regularity conditions

\[
MISE(\hat{f}_{Hist}(x)) = E \int (\hat{f}_{Hist}(x) - f(x))^2 \, dx \\
= \int \text{Var}(\hat{f}_{Hist}(x)) \, dx + \int (E(\hat{f}_{Hist}(x)) - f(x))^2 \, dx \\
= \int \left( \frac{1}{lh^2} f^*(x) + O \left( \frac{C(x)}{l} \right) + O(B(x)h) \right) \, dx \\
+ \int \left( f(x) - f(x) * g_n(x) + \frac{1}{2} f^*(x)(h + 2b_j - 2x) + O(h^2) \right)^2 \, dx \\
= \frac{1}{lh} + O\left( \frac{1}{\sqrt{n}} \right) + O\left( \frac{1}{l} \right) + O(h).
\]

Hence, as \( n, k \to \infty, h \to 0, \text{ and } th \to +\infty \), we have \( MISE(\hat{f}_{Hist}(x)) \to 0 \). Since \( g_n(x) \) is a normal distribution with variance in the order of \( 1/n \), as \( n \to +\infty \) we have \( f^*(x) \to f(x) \).

Next, we take a close look at the assumption we made in Proposition 2.4.1. \( \sum_{m=l+1}^{l'} (f^*_{(x_l,x_m)}(x,x) - f^{*2}(x)) < M \), may be interpreted as that in a society of many or infinite people, only at most a small population of them would know many people while each of the majority knows just a finite number of people well. This is consistent to have an interesting clustered community rather than a community of intertwined relationship with everyone.

## 2.5 Algorithm for nFCA

Next, we give step by step nFCA algorithm, Algorithm 1, based on a correlation coefficient matrix \( R = (r_{ij}) \). For distance matrix, the algorithm is similar except the threshold can not be chosen by confidence interval method.
Data: $m \times m$ numerical data matrix $X$ involving $m$ objects

Result: Graphical file to show the hierarchical clustering structure

Calculate correlation coefficient matrix $R = (r_{ij}), i = 1, \cdots, m; j = 1, \cdots, m$ if necessary;
Create an empty output file $\text{graph.txt}$;

while not all the objects $O$ are clustered do
  // Choose a threshold $t$ using either Histogram or confidence interval method, use Histogram method if $n$ is missing
  if Use Histogram method to choose the threshold $t$ then
    Put all the $r_{ij}$'s in $R$ which are not equal to 1 in a vector $V$;
    Draw the Histogram of the vector $V$ and denoted by $\text{Hist}$;
    $t < -\text{Hist\$breaks[length(\text{Hist\$breaks}) - 1]}$;
  end

  if Use confidence interval method to choose the threshold $t$ then
    Sort all the $r_{ij}$ values in $R$ which are not equal to 1 in increasing order;
    Calculate the 100$(1 - \alpha)$% confidence interval for the largest one using formula 2.1;
    while too many ($> 10$) $r_{ij}$'s in $R$ fall into the confidence interval do
      Decrease $\alpha$ by $\frac{1}{2}$ or $\frac{1}{3}$ and recalculate the confidence interval;
    end
    while only a few ($< 3$) $r_{ij}$'s in $R$ fall into the confidence interval do
      Calculate the confidence interval for the next largest $r_{ij}$;
    end
    Choose the lower bound of the confidence interval as the threshold $t$;
  end

Create an empty binary matrix $B$;
for each value $r_{ij}$ in $R$ do
  if $R_{ij} \geq t$ then
    $B[i, j] < -1$;
  else
    $B[i, j] < -0$;
  end
end

Use faster concept analysis algorithm to construct the concept lattice $C$ for $B$;
Sort the concepts in the concept lattice $C$ in increasing order;
for each concept $C_1$ in the ordered concept lattice do
  for each concept $C_2$ in the ordered concept lattice do
    if $C_1 \neq C_2$ then
      if $C_1$ is a subset of $C_2$ then
        if $(C_1, C_2)$ is not in $\text{graph.txt}$ then
          Append $(C_1, C_2)$ to $\text{graph.txt}$ ;
          $C_1 = C_2$ ;
        end
      end
    end
  end
Group all the newly appended objects as one cluster;
end

Group separately the objects within each cluster together as a new object;
Update the correlation coefficient matrix $R$;
end

Algorithm 1: nFCA algorithm
References


Chapter 3
Example, Application, Evaluation and Simulation Study

3.1 An Illustrative Example for nFCA

In order to show how nFCA actually works and also check whether it will produce a reasonable result, in this illustrative example we apply nFCA to an made-up family correlation coefficient matrix. The family members involved in this correlation coefficient matrix are Husband(H), Wife(W), Son(S), Daughter(D), Husband’s Father(HF), Husband’s Mother(HM), Wife’s Father(WF), Wife’s Mother(WM), Husband’s Friend(H1), Wife’s Friend(W1), and Others(O). The correlation coefficient matrix is constructed based on the following principles,

1. Husband and Wife are highly correlated, such as between H and W, between HF and HW, and between WF and WM,

2. both H and W are highly correlated with S and D, but not as high as the correlation between them $^2$,

3. H and W along with S and D forms a nuclear family,

4. this nuclear family further has high correlation with grandparents, such as HF and HM, WF and WM .

$^2$Although this may be hotly debated if H and W are in a cold war, here we are talking about a normal happy family.
5. H1 and W1 have strong friendship with the nuclear family, but not as strong as the relationship with grandparents,

6. O has even less correlation with the nuclear family than H1 and W1 do.

Following table, Table 3.1, is this made-up correlation coefficient data set for the Family example. Next, we show step by step how nFCA works on this correlation coefficient matrix to finally lead to a social network.

<table>
<thead>
<tr>
<th>H</th>
<th>W</th>
<th>S</th>
<th>D</th>
<th>HF</th>
<th>HM</th>
<th>WF</th>
<th>WM</th>
<th>H1</th>
<th>W1</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.95</td>
<td>0.90</td>
<td>0.90</td>
<td>0.85</td>
<td>0.85</td>
<td>0.80</td>
<td>0.80</td>
<td>0.40</td>
<td>0.30</td>
<td>0.10</td>
</tr>
<tr>
<td>0.95</td>
<td>1</td>
<td>0.88</td>
<td>0.88</td>
<td>0.70</td>
<td>0.70</td>
<td>0.87</td>
<td>0.87</td>
<td>0.20</td>
<td>0.45</td>
<td>0.15</td>
</tr>
<tr>
<td>0.90</td>
<td>0.88</td>
<td>1</td>
<td>0.82</td>
<td>0.80</td>
<td>0.80</td>
<td>0.76</td>
<td>0.76</td>
<td>0.10</td>
<td>0.10</td>
<td>0.02</td>
</tr>
<tr>
<td>0.90</td>
<td>0.88</td>
<td>0.82</td>
<td>1</td>
<td>0.82</td>
<td>0.82</td>
<td>0.81</td>
<td>0.81</td>
<td>0.10</td>
<td>0.10</td>
<td>0.05</td>
</tr>
<tr>
<td>0.85</td>
<td>0.70</td>
<td>0.80</td>
<td>0.82</td>
<td>1</td>
<td>0.95</td>
<td>0.65</td>
<td>0.65</td>
<td>0.08</td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>0.85</td>
<td>0.70</td>
<td>0.80</td>
<td>0.82</td>
<td>0.95</td>
<td>1</td>
<td>0.60</td>
<td>0.60</td>
<td>0.06</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>0.80</td>
<td>0.87</td>
<td>0.76</td>
<td>0.81</td>
<td>0.65</td>
<td>0.60</td>
<td>1</td>
<td>0.95</td>
<td>0.03</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>0.80</td>
<td>0.87</td>
<td>0.76</td>
<td>0.81</td>
<td>0.65</td>
<td>0.60</td>
<td>0.95</td>
<td>1</td>
<td>0.01</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>0.40</td>
<td>0.20</td>
<td>0.10</td>
<td>0.10</td>
<td>0.08</td>
<td>0.08</td>
<td>0.06</td>
<td>0.06</td>
<td>0.03</td>
<td>0.01</td>
<td>1</td>
</tr>
<tr>
<td>0.30</td>
<td>0.45</td>
<td>0.10</td>
<td>0.10</td>
<td>0.06</td>
<td>0.07</td>
<td>0.05</td>
<td>0.05</td>
<td>0.02</td>
<td>0.02</td>
<td>1</td>
</tr>
<tr>
<td>0.10</td>
<td>0.15</td>
<td>0.02</td>
<td>0.05</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 3.1: Made-up correlation coefficient matrix for the Family example

**Step 1:** First thresholding and first FCA

1.1 First we need to choose a threshold for this correlation coefficient. If Histogram method is used to choose the threshold, we use the Histogram of the correlation coefficients in Figure 3.1. From the Histogram we choose 0.9 as the threshold.

1.2 After the threshold is chosen we use it to threshold the correlation coefficient matrix. Table 3.2 is the resulting binary matrix after the thresholding.

1.3 Faster concept analysis algorithm is used to construct the concept lattice of above binary matrix shown in Table 3.2. The concept lattice is shown in the Table 3.3.

1.4 After we have the concept lattice, we can graphically visualize it via a Hasse diagram as shown in Figure 3.2. There are three small clusters
Figure 3.1: Histogram of the correlation coefficients from the Family example

Table 3.2: Binary matrix after 1st thresholding for the Family example

<table>
<thead>
<tr>
<th>H</th>
<th>W</th>
<th>S</th>
<th>D</th>
<th>HF</th>
<th>HM</th>
<th>WF</th>
<th>WM</th>
<th>HI</th>
<th>W1</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

in the Hasse diagram: \{H, W, S, D\}, \{HF, HM\}, and \{WF, WM\}. Then we group the objects in each of these clusters together and form new objects, one for each cluster. For example, H, W, S, and D are grouped together and will be treated as a single new object, say Cluster1. Similarly, after grouping HF and HM together the new object is
<table>
<thead>
<tr>
<th>Objects</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>{H}</td>
<td>{H, W, S, D}</td>
</tr>
<tr>
<td>{H, W}</td>
<td>{H, W}</td>
</tr>
<tr>
<td>{WF, WM}</td>
<td>{WF, WM}</td>
</tr>
<tr>
<td>{H, S}</td>
<td>{H, S}</td>
</tr>
<tr>
<td>{H1}</td>
<td>{H1}</td>
</tr>
<tr>
<td>{H, D}</td>
<td>{H, D}</td>
</tr>
<tr>
<td>{W1}</td>
<td>{W1}</td>
</tr>
<tr>
<td>{HF, HM}</td>
<td>{HF, HM}</td>
</tr>
<tr>
<td>{O}</td>
<td>{O}</td>
</tr>
<tr>
<td>{H, W, S, D}</td>
<td>{H, W, S, D}</td>
</tr>
<tr>
<td>φ</td>
<td>φ</td>
</tr>
<tr>
<td>{HM, H, WF, W, WM, S, H1, D, W1, HF, O}</td>
<td>{H, W, S, D, HF, HM, WF, WM, H1, W1, O}</td>
</tr>
</tbody>
</table>

Table 3.3: Concept lattice constructed using faster concept analysis for the Family example

called Cluster2. Cluster3 is used to denote the object by grouping WF and WM together. Now the newly formed objects along with the other non-grouped objects, i.e., \{Cluster1, Cluster2, Cluster3, H1, W1, O\}, are our new object set.

Figure 3.2: Hasse diagram of the concept lattice for the Family example
Next we calculate the correlation coefficient matrix for the new object set. For example, the correlation coefficient between Cluster1 and Cluster2 is taken as the largest value among the correlation coefficients for all the pairs of objects, one from each of Cluster1 and Cluster2. The correlation coefficient between Cluster1 and Cluster2 is 0.85, i.e.,

$$\max(0.85, 0.70, 0.80, 0.82, 0.85, 0.70, 0.80, 0.82).$$

Table 3.4 is the calculated correlation coefficient matrix for the new object set.

<table>
<thead>
<tr>
<th>Cluster1</th>
<th>Cluster2</th>
<th>Cluster3</th>
<th>H1</th>
<th>W1</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>0.85</td>
<td>0.87</td>
<td>0.40</td>
<td>0.45</td>
<td>0.15</td>
</tr>
<tr>
<td>0.85</td>
<td>1.00</td>
<td>0.65</td>
<td>0.08</td>
<td>0.07</td>
<td>0.01</td>
</tr>
<tr>
<td>0.87</td>
<td>0.65</td>
<td>1.00</td>
<td>0.03</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>0.40</td>
<td>0.08</td>
<td>0.03</td>
<td>1.00</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>0.45</td>
<td>0.07</td>
<td>0.05</td>
<td>0.02</td>
<td>1.00</td>
<td>0.02</td>
</tr>
<tr>
<td>0.15</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 3.4: 1st updated correlation coefficient matrix for the Family example

In summary, the tables and figures to illustrate this step are: Table 3.2, Table 3.3, Table 3.4, Figure 3.1, and Figure 3.2.

**Step 2:** Second thresholding and second FCA

2.1 A second threshold needs to be chosen for the new correlation coefficient matrix. Suppose we use Histogram method, the corresponding Histogram is shown in Figure 3.3. A reasonable choice for the threshold is 0.8. We use this threshold to threshold the new correlation coefficient matrix and Table 3.5 is the resulting binary matrix.

2.2 The concept lattice constructed by faster concept analysis for the binary matrix after second thresholding is shown in Table 3.6.

2.3 The concept lattice can be graphically visualized via the Hasse diagram in Figure 3.4. From this diagram we can see that Cluster1, Cluster2, and Cluster3 form a new small cluster. After grouping them together, a new object set is defined similarly as before and the calculated correlation coefficient matrix is shown in Table 3.7. Cluster4 is the new object by grouping Cluster1, Cluster2, and Cluster3 together.
Figure 3.3: Histogram of the correlation coefficients after 1st thresholding for the Family example

Table 3.5: Binary matrix after 2nd thresholding for the Family example

<table>
<thead>
<tr>
<th>Objects</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>{ O }</td>
<td>{ Cluster1, Cluster2, Cluster3 } Cluster4</td>
</tr>
<tr>
<td>{ Cluster1 }</td>
<td>{ Cluster1, Cluster3 }</td>
</tr>
<tr>
<td>{ Cluster1, Cluster3 }</td>
<td>{ Cluster1, Cluster3 }</td>
</tr>
<tr>
<td>{ Cluster1, Cluster2 }</td>
<td>{ Cluster1, Cluster2 }</td>
</tr>
<tr>
<td>{ H1 }</td>
<td>{ H1 }</td>
</tr>
<tr>
<td>{ W1 }</td>
<td>{ W1 }</td>
</tr>
<tr>
<td></td>
<td>{ Cluster1, Cluster2, Cluster3, H1, W1, O }</td>
</tr>
<tr>
<td>( \phi )</td>
<td>( \phi )</td>
</tr>
</tbody>
</table>

Table 3.6: Concept lattice constructed by faster concept analysis after 2nd thresholding for the Family example

In summary, the tables and figures in this step are: Table 3.5, Table 3.6,
Figure 3.4: Hasse diagram of the concept lattice after 2nd thresholding for the Family example

<table>
<thead>
<tr>
<th></th>
<th>Cluster4</th>
<th>H1</th>
<th>W1</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster4</td>
<td>1.00</td>
<td>0.40</td>
<td>0.45</td>
<td>0.15</td>
</tr>
<tr>
<td>0.40</td>
<td>1.00</td>
<td>0.02</td>
<td>0.01</td>
<td></td>
</tr>
<tr>
<td>0.45</td>
<td>0.02</td>
<td>1.00</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.01</td>
<td>0.02</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.7: 2nd updated correlation coefficient matrix for the Family example

Table 3.7, Figure 3.3, and Figure 3.4.

**Step 3:** Third thresholding and third FCA

3.1 The corresponding Histogram for newly updated correlation coefficient matrix is shown in Figure 3.5. Suppose we choose the threshold
as 0.3, Table 3.8 shows the binary matrix after third thresholding.

![Histogram of correlation coefficients](image)

**Figure 3.5:** Histogram of the correlation coefficients after 2nd thresholding for the Family example

<table>
<thead>
<tr>
<th>Cluster4</th>
<th>H1</th>
<th>W1</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 3.8:** Binary matrix after third thresholding for the Family example

3.2 The concept lattice for the above binary matrix is constructed and shown in Table 3.9.

3.3 The Hasse diagram used to visualize the concept lattice after third thresholding is shown in Figure 3.6. It is clear that after third thresholding step, Cluster4, H1, and W1 can be grouped together and treated as a single new object, Cluster5. After grouping them together we only have two objects, Cluster5 and O. There is no need to repeat the procedure since for two objects they have to be grouped together. Finally we group Cluster5 and O together and denoted by Cluster6.
Table 3.9: Concept lattice constructed by faster concept analysis after 3rd thresholding for the Family example

<table>
<thead>
<tr>
<th>Objects</th>
<th>Attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>{ O }</td>
<td>{ O }</td>
</tr>
<tr>
<td>{ Cluster4 }</td>
<td>{ Cluster4, H1, W1 }</td>
</tr>
<tr>
<td>{ Cluster4, H1 }</td>
<td>{ Cluster4, H1 }</td>
</tr>
<tr>
<td>{ Cluster4, W1 }</td>
<td>{ Cluster4, W1 }</td>
</tr>
<tr>
<td>{ Cluster4, H1, W1 }</td>
<td>{ Cluster4 }</td>
</tr>
<tr>
<td>{ Cluster4, H1, W1, O }</td>
<td>{ Cluster4, H1, W1, O }</td>
</tr>
</tbody>
</table>

In summary, the tables and figures in this step are: Table 3.8, Table 3.9, Figure 3.5, and Figure 3.6.

**Step 4:** Final graph combined step

Finally three thresholding steps have been conducted to group all the objects together. The clustering information from each thresholding step is combined together in inverse order. For example, in this illustrative example, we start from Cluster6 which contains Cluster 5 and object O. And Cluster5 further contains Cluster4 and objects W1 and H1. We can continue this process until all the clusters and objects are visited. Figure 3.7, *nFCA H-graph*, provides the combined clustering information which shows the hierarchical structure among the numerical data. In this hierarchical structure, the connections among the objects within a cluster are shown in black lines with or without arrows. If the number of objects in a cluster is 2, the objects are connected using a black line without arrows. If there are more than 2 objects in a cluster, we use black lines with arrows to show their connections and dominance. In this case, the common start of the arrows is the kernel or core object of the cluster. The label of the connection indicates how strong the connection is. The value of the label comes from the thresholding procedure. Large value of the label means strong connection and small value of the label means weak connection. The connections between the objects which connect to a same kernel or core object are shown in red lines without arrows. The value of the label for a red line is calculated using average linkage method. To be specific, for example the connection between the cluster formed by HM and HF and the cluster formed by WF
and WM is determined by the average of all the pairwise correlation coefficients between these two clusters. The calculated value of the connection is shown in red label.

For the hierarchical structure from nFCA, it clearly shows that H together with W, S, and D forms a nuclear family. This nuclear family then extends to

Figure 3.6: Hasse diagram of the concept lattice
two small clusters contains HF and HM, and WF and WM separately. HF and HW are grouped together since they are highly correlated. WF and WM have the similar situation. This extended family includes H, W, S, D, HF, HM, WF, and WM, which further extends to include W1 and H1, the friends of the wife and the husband respectively and finally O since it has the weakest connection with any of the extended family. This hierarchical structure is consistent to the principles we used to create the original correlation coefficient matrix.

This nFCA H-graph is intuitive and will be shown to compare with the standard hierarchical clustering method. However, it dose not show, say, how W1 is related to which particular member of the nuclear family. For this, we provide a companion to the nFCA H-graph, called nFCA I-graph, which depicts all the point-to-cluster relationship and also the strength of each relationship in a nice tree fashion. See the nFCA I-graph for the Family data in Figure 3.8. The construction of the inherent structure is described in Section 2.2. This structure reveals specific links and the most significant connections among the data. For example, HF and HM have strong connections with H while WF and WM have
strong connections with W. Similarly, H1 has relatively strong connection with H while W1 has relatively strong connection with W. Comparing with the hierarchical cluster in Figure 3.7, this inherent structure provides additional information and the result reveals the underline principles we used the created the data.

Figure 3.8: Inherent structure for the Family data

3.2 Application of nFCA to Cardiovascular (CV) Traits

One fundamental problem in biological research is to understand how genes within an organism correlate to each other and work together to provide the structures and functions of the organism. It is also equally important to study the relationships among the attributes or properties of an organism at a higher functional level. Nadeau, Burrage, Restivo, Pao, Churchill, and Hoit (2003) conducted a proof-of-concept study for the functional network among cardiovascular (CV) traits. They used A/J and C57BL/6J strains of mice to perturb the CV functions of mice in a non-pathologic way and investigated the interactions between the CV traits. Their resulting functional network of CV traits is built directly on the correlation coefficient matrix and shows some useful information on the relationship among the CV traits. However, other useful information and a complete hierarchical structure for the CV traits are still unknown.

Next, by applying nFCA to the same CV traits data, we study the way that CV traits interact to each other and show a more complete hierarchical clustering
information among the CV traits than the earlier work.

Their study involves 13 CV traits. Table 3.10 gives a brief description of these 13 traits. The correlation coefficient matrix of the CV traits they used is shown in

<table>
<thead>
<tr>
<th>EDD</th>
<th>end diastolic dimension</th>
<th>SWTH</th>
<th>end systolic dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWTW</td>
<td>posterior wall thickness</td>
<td>LVmass</td>
<td>septal wall thickness</td>
</tr>
<tr>
<td>FS</td>
<td>fractional shortening</td>
<td>SV</td>
<td>left ventricular mass</td>
</tr>
<tr>
<td>Thr</td>
<td>relative wall thickness</td>
<td>HR</td>
<td>stroke volume</td>
</tr>
<tr>
<td>CO</td>
<td>cardiac output</td>
<td>EST</td>
<td>heart rate</td>
</tr>
<tr>
<td>BW</td>
<td>body weight</td>
<td>LVBW</td>
<td>exercise</td>
</tr>
<tr>
<td>LVBW</td>
<td>LVmass/BW</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.10: Descriptions of 13 CV traits

Table 3.11: Correlation coefficient matrix of 13 CV traits

In this application, we examine the relationship in 2 ways. First, we rescale the numerical values \( x \) in the correlation coefficient matrix to \([0, 1]\) by using transformation \((1 + x)/2\). Second, we only look at values that are positive and negative, respectively. Now we start with the transformed values. In this case, the nFCA contains four thresholding steps at thresholds 0.90, 0.85, 0.80, and 0.70. After combining the cluster information from each thresholding step, we present nFCA H-graph, the hierarchical structure for CV traits in Figure 3.9, and nFCA I-graph in Figure 3.10 which shows the inherent structure for CV traits. The functional network of CV traits presented in the original paper is shown in Figure 3.11 where solid lines indicate positive relationships and dotted lines indicate negative relationships. The way to interpret Figure 3.9 and Figure 3.10 is similar as we did in the illustrative example. First of all, it is obvious that EDD, ESD, SV, and LVmass forms the nuclear “family” and EDD is the kernel or core of this nuclear
family. This is equivalent to say that end diastolic dimension is highly correlated with end systolic dimension, stroke volume, and left ventricular mass. EDD seems to be the driving force. Then cardiac output has strong connection with this nuclear family than the others. After extending the nuclear family to include the cardiac output, this extended nuclear family has strong connection with the “ratio between left ventricular mass and body weight”. It also highly correlated with posterior wall thickness, septal wall thickness, and relative wall thickness. It further shows that posterior wall thickness has strong connections with septal wall thickness and relative wall thickness. Now we extend the nuclear family to encompass LVBW, PWTH, SWTH, and Thr. From the hierarchical structure we can see that the extended nuclear family has some connections with body weight, fractional shortening, and heart rate. In addition, heart rate and fractional shortening have some connections also. Last from the hierarchical structure we can see exercise does not have as strong connection as the others do among themselves. This maybe relates to the mice in the experiment and will be different for human. The inherent structure in Figure 3.10 provides additional information to the hierarchical cluster. It illustrates the most significant connections among CV traits. In addition, the inherent structure also shows how the hierarchical cluster extends from the kernel to the outside in terms of the major connections for each CV trait.

Comparing Figures 3.9 and Figure 3.10 with Figure 3.11, it is clear that the hierarchical cluster and inherent structure from nFCA confirm the functional relationship among the CV traits as shown in Figure 3.11. The resulting figures from nFCA provide more complete information via the hierarchical cluster and inherent structure among the CV traits in the nFCA H-graph and I-graph. It clearly shows how the CV traits are interacting with each other and also the dominant traits within each cluster.

Notice that the functional network by Nadeau, Burrage, Restivo, Pao, Churchill, and Hoit (2003)[1] as shown in Figure 3.11 also includes the negative relationships among the CV traits. In order to conduct a fair comparison, we also apply nFCA to the CV traits data to reveal the negative relationships. For the convenience purpose, we apply \((1 - x)/2\) transformation to the non-diagonal correlation coefficients in the original CV traits data. After applying nFCA, the hierarchical cluster and the inherent structure are shown in Figure 3.12 and 3.13 separately.
From the results of nFCA, we see ESD and FS have strong negative relationships and form the kernel or core of the hierarchical cluster. And EDD has strong negative relationship with FS and HR has strong negative relationship with ESD. In addition, EDD and HR have relatively strong negative relationships. These results are consistent with the relationship shown in Figure 3.11. In fact, this pair of nFCA graphs provides a complete relational network among CV traits in terms of negative relationships. For example, ESD and FS have very strong negative relationship and HR has strong negative relationship with the nuclear cluster formed by ESD and FS since it has strong negative relationship with ESD. However, the negative relationship between FS and ESD is not as strong as the negative relationship between ESD and FS.

Next, we apply nFCA to the positive part and negative part respectively, i.e., the negative values in the original correlation coefficient matrix are set to 0 for the positive case and the non-diagonal positive values in the original correlation
Figure 3.10: Inherent structure of CV traits by nFCA: nFCA I-graph

coefficient matrix are set to 0 and times -1 for the negative case. Then, we apply nFCA to these two resulting numerical matrices to illustrate the positive and negative relationship respectively. Figure 3.14 and Figure 3.15 are the nFCA H-graph and I-graph for the positive case and Figure 3.16 and Figure 3.17 are the nFCA H-graph and I-graph for the negative case. For both cases, the nFCA H-graph and I-graph are consistent with the nFCA resulting graphs, H-graph and I-graph, using the transformations. This indicates that our nFCA procedure is robust against noise at lower levels.
3.3 Evaluating the Performance of nFCA

The hierarchical clustering using nFCA is different from conventional hierarchical clustering. For hierarchical clustering of nFCA, the elements in a sub-cluster
Figure 3.13: Inherent structure of CV traits for negative correlations by nFCA

are not just listed there, instead arrows have been used to show their directional relationships. The label of the connection also indicates how strong the relationship is. Large value of the label means stronger relationship. From the directions of the arrows we can tell which one is the kernel of this sub-cluster. This feature provides additional information to what conventional hierarchical clustering provides. For example, in the cardiovascular application the hierarchical clustering from nFCA shows that EDD, ESD, SV, and LVmass belong to the same sub-cluster and EDD is the kernel of this sub-cluster which may be the most important measure to monitor in terms of CV health. This also explains why EDD, ESD, SV, and LVmass are grouped together.

Next, we compare nFCA graphs and the conventional hierarchical clustering systematically using the Cophenetic correlation coefficient. In statistics, Cophenetic correlation coefficient is used to measure the consistency of a dendrogram to the original distance matrix. Formally, the Cophenetic correlation coefficient is
Figure 3.14: nFCA H-graph of CV traits for positive part

defined as

\[
\frac{\sum_{i<j} (x(i, j) - \bar{x}) (c(i, j) - \bar{c})}{\sqrt{\sum_{i<j} (x(i, j) - \bar{x})^2 \sum_{i<j} (c(i, j) - \bar{c})^2}},
\]

(3.1)

where \(x(i, j)\) is the correlation or distance between \(i\)th and \(j\)th objects, \(c(i, j)\) is the correlation or distance between \(i\)th and \(j\)th objects calculated from the dendrogram or hierarchical clustering, and \(\bar{x}\) and \(\bar{c}\) are the mean values of \(x(i, j)\)'s and \(c(i, j)\)'s separately. Cophenetic correlation coefficient can be easily calculated if we have the estimated correlation coefficient matrix or distance matrix from the hierarchical cluster result.

For nFCA, estimating the correlation coefficient matrix is quite straightforward. Since in the hierarchical cluster from nFCA the label of each connection indicates how strong the connections is. So we can easily compute an estimated correlation coefficient or distance matrix from hierarchical cluster. Next, we use the additional connection information provided by the inherent structure of nFCA
to improve the estimation of the correlation coefficient. Once we get the estimated correlation coefficient or distant matrix from nFCA, Cophenetic correlation coefficient is ready to compute.

Next, as an example we show how to compute the Cophenetic correlation coefficient for the hierarchical clustering, i.e., Figure 3.9, of CV traits using nFCA. Based on the hierarchical cluster and inherent structure of CV traits, the estimated correlation coefficient matrix is shown in Table 3.12. Based on this estimated correlation coefficient matrix, one can easily compute the Cophenetic correlation coefficient as 0.844. If we apply hierarchical clustering on the same data using average linkage, single linkage, and complete linkage separately. The corresponding computed Cophenetic correlation coefficients are 0.748, 0.490, and 0.523. This
Figure 3.16: nFCA H-graph of CV traits for negative part

<table>
<thead>
<tr>
<th></th>
<th>PWTH</th>
<th>SWTH</th>
<th>EDD</th>
<th>ESD</th>
<th>FS</th>
<th>LVmass</th>
<th>BW</th>
<th>LV/BW</th>
<th>SV</th>
<th>Thr</th>
<th>HR</th>
<th>EST</th>
<th>CO</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWTH</td>
<td>1.000</td>
<td>0.960</td>
<td>0.532</td>
<td>0.532</td>
<td>0.357</td>
<td>0.835</td>
<td>0.627</td>
<td>0.595</td>
<td>0.532</td>
<td>0.855</td>
<td>0.357</td>
<td>0.563</td>
<td>0.532</td>
</tr>
<tr>
<td>SWTH</td>
<td>0.960</td>
<td>1.000</td>
<td>0.532</td>
<td>0.532</td>
<td>0.357</td>
<td>0.840</td>
<td>0.785</td>
<td>0.595</td>
<td>0.532</td>
<td>0.848</td>
<td>0.357</td>
<td>0.563</td>
<td>0.532</td>
</tr>
<tr>
<td>EDD</td>
<td>0.532</td>
<td>0.532</td>
<td>1.000</td>
<td>0.950</td>
<td>0.357</td>
<td>0.920</td>
<td>0.627</td>
<td>0.673</td>
<td>0.960</td>
<td>0.532</td>
<td>0.357</td>
<td>0.563</td>
<td>0.770</td>
</tr>
<tr>
<td>ESD</td>
<td>0.532</td>
<td>0.532</td>
<td>0.950</td>
<td>1.000</td>
<td>0.357</td>
<td>0.850</td>
<td>0.627</td>
<td>0.673</td>
<td>0.890</td>
<td>0.532</td>
<td>0.357</td>
<td>0.563</td>
<td>0.770</td>
</tr>
<tr>
<td>FS</td>
<td>0.357</td>
<td>0.357</td>
<td>0.357</td>
<td>0.357</td>
<td>1.000</td>
<td>0.357</td>
<td>0.357</td>
<td>0.760</td>
<td>0.860</td>
<td>0.563</td>
<td>0.357</td>
<td>0.563</td>
<td>0.357</td>
</tr>
<tr>
<td>LVmass</td>
<td>0.835</td>
<td>0.840</td>
<td>0.920</td>
<td>0.850</td>
<td>0.357</td>
<td>1.000</td>
<td>0.627</td>
<td>0.835</td>
<td>0.900</td>
<td>0.510</td>
<td>0.357</td>
<td>0.563</td>
<td>0.770</td>
</tr>
<tr>
<td>BW</td>
<td>0.627</td>
<td>0.785</td>
<td>0.627</td>
<td>0.627</td>
<td>0.358</td>
<td>1.000</td>
<td>0.627</td>
<td>0.627</td>
<td>0.358</td>
<td>0.563</td>
<td>0.627</td>
<td>0.563</td>
<td>0.627</td>
</tr>
<tr>
<td>LV/BW</td>
<td>0.595</td>
<td>0.595</td>
<td>0.673</td>
<td>0.673</td>
<td>0.357</td>
<td>0.835</td>
<td>0.627</td>
<td>1.000</td>
<td>0.673</td>
<td>0.595</td>
<td>0.357</td>
<td>0.685</td>
<td>0.673</td>
</tr>
<tr>
<td>SV</td>
<td>0.532</td>
<td>0.532</td>
<td>0.960</td>
<td>0.890</td>
<td>0.357</td>
<td>0.900</td>
<td>0.627</td>
<td>0.673</td>
<td>1.000</td>
<td>0.532</td>
<td>0.357</td>
<td>0.563</td>
<td>0.865</td>
</tr>
<tr>
<td>Thr</td>
<td>0.855</td>
<td>0.840</td>
<td>0.532</td>
<td>0.532</td>
<td>0.760</td>
<td>0.510</td>
<td>0.627</td>
<td>0.595</td>
<td>0.532</td>
<td>1.000</td>
<td>0.705</td>
<td>0.563</td>
<td>0.532</td>
</tr>
<tr>
<td>HR</td>
<td>0.357</td>
<td>0.357</td>
<td>0.357</td>
<td>0.357</td>
<td>0.860</td>
<td>0.357</td>
<td>0.357</td>
<td>0.357</td>
<td>0.705</td>
<td>1.000</td>
<td>0.563</td>
<td>0.357</td>
<td>0.357</td>
</tr>
<tr>
<td>EST</td>
<td>0.563</td>
<td>0.563</td>
<td>0.563</td>
<td>0.563</td>
<td>0.563</td>
<td>0.563</td>
<td>0.563</td>
<td>0.685</td>
<td>0.563</td>
<td>0.563</td>
<td>1.000</td>
<td>0.563</td>
<td>0.563</td>
</tr>
<tr>
<td>CO</td>
<td>0.532</td>
<td>0.532</td>
<td>0.770</td>
<td>0.770</td>
<td>0.357</td>
<td>0.770</td>
<td>0.627</td>
<td>0.673</td>
<td>0.865</td>
<td>0.357</td>
<td>0.563</td>
<td>1.000</td>
<td>0.563</td>
</tr>
</tbody>
</table>

Table 3.12: Estimated correlation coefficient matrix of 13 CV traits

shows that hierarchical cluster result from nFCA for CV traits data has a better performance than the conventional hierarchical cluster results. Next, in our simulation study we will compare the performance of nFCA with conventional hierarchical cluster, based on random samples sampled from the family model.
3.4 Simulation Study

The purpose of this simulation study is to evaluate and compare the performance of nFCA and convention hierarchical cluster method. We use the same family model in Section 3.1. The Family correlation coefficient data serves as the true population correlation coefficient matrix. Next, we construct the sample correlation coefficients from the population correlation coefficients. For each sample correlation coefficient between $i$th and $j$th object $r_{ij}$ the Fisher transformation is denoted by $z_{ij}$. $z_{ij}$ approximately has a normal distribution with mean $\frac{1}{2} \ln \frac{\rho_{ij} + 1}{\rho_{ij} - 1}$ and variance $\frac{1}{n - 3}$. Here $n$ is the sample size used to calculate the sample correla-
tion coefficient and \( \rho_{ij} \) is the population correlation coefficient between \( i \)th object and \( j \)th object. A random sample of \( z_{ij} \) is drawn from this normal distribution and transformed back to get the sample correlation coefficient, i.e.,

\[
    r_{ij} = \frac{\exp(2z_{ij} - 1)}{\exp(2z_{ij} - 1)}.
\]

In this simulation the sample size \( n \) is taken to be 30. We repeat this procedure for all the \( i \)'s and \( j \)'s and finally we have a sample correlation coefficient matrix. We apply conventional hierarchical cluster method and nFCA to this sample correlation coefficient matrix and evaluate their performance by comparing their Cophenetic correlation coefficients. In our simulation, we generate 300 sample correlation coefficient matrices and compare the Cophenetic correlation coefficients from nFCA and conventional hierarchical cluster method.

First, for each generated sample correlation coefficient the Cophenetic correlation coefficients for nFCA and conventional hierarchical cluster method are computed based on the population correlation coefficient matrix and the hierarchical cluster results from nFCA and conventional hierarchical cluster methods. Following table, Table 3.13, shows the percentage of the samples for which nFCA has a better performance, i.e., higher Cophenetic correlation coefficient, than conventional hierarchical cluster method. Three different linkages, i.e., average linkage, single linkage, and complete linkage, have been used in conventional hierarchical cluster method.

<table>
<thead>
<tr>
<th>Percentage of samples for which nFCA has better performance</th>
<th>Conventional hierarchical cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average linkage</td>
</tr>
<tr>
<td></td>
<td>70%</td>
</tr>
</tbody>
</table>

Table 3.13: Performance comparison for population correlation coefficient matrix \((n = 30)\)

Second, rather than using the population correlation coefficients as \( x(i, j) \) in Table 3.1, we use sampled correlation coefficient as \( x(i, j) \). We then compute the Cophenetic correlation coefficients of the cluster results from nFCA and conventional hierarchical cluster methods with three different linkages. Table 3.14 shows the percentage of the samples for which nFCA has a better performance, i.e., higher Cophenetic correlation coefficient than conventional hierarchical cluster method.
If we change the sample size $n$ from 30 to 60 and redo the simulation, Table 3.15 and Table 3.16 show the percentages of the samples for which nFCA has a better performance than conventional cluster method regarding to population correlation coefficient matrix and sample correlation efficient matrix separately.

Table 3.14: Performance comparison for sample correlation coefficient matrix $(n = 30)$

<table>
<thead>
<tr>
<th>Percentage of samples for which nFCA has better performance</th>
<th>Average linkage</th>
<th>Single linkage</th>
<th>Complete linkage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99%</td>
<td>99%</td>
<td>99%</td>
</tr>
</tbody>
</table>

Table 3.15: Performance comparison for population correlation coefficient matrix $(n = 60)$

<table>
<thead>
<tr>
<th>Percentage of samples for which nFCA has better performance</th>
<th>Average linkage</th>
<th>Single linkage</th>
<th>Complete linkage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>96.7%</td>
<td>98.3%</td>
<td>99.3%</td>
</tr>
</tbody>
</table>

Table 3.16: Performance comparison for sample correlation coefficient matrix $(n = 60)$

<table>
<thead>
<tr>
<th>Percentage of samples for which nFCA has better performance</th>
<th>Average linkage</th>
<th>Single linkage</th>
<th>Complete linkage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>99.3%</td>
<td>99.3%</td>
<td>99.3%</td>
</tr>
</tbody>
</table>

From above simulation results we see that nFCA has a much better performance than conventional hierarchical cluster regarding to the population correlation coefficient matrix. And the performance of nFCA becomes even better as the sample size increasing. In terms of sample population correlation coefficient matrix, nFCA performs almost uniformly better than conventional hierarchical cluster.

References

Chapter 4

Implementation and Generalization of nFCA

4.1 Implementation of nFCA

The nFCA algorithm has been programmed and implemented in R, with an interface to a fast FCA Ruby code coauthored by Adam Troy and Professor Guo-Qiang Zhang. As a final step of showing both nFCA H-graph and I-graph, we use a graphical visualization software Graphviz. Hence, in order to use our nFCA, one must have

1. **R**: available from [http://www.r-project.org](http://www.r-project.org)
4. **nFCA R code**: available from [http://sun.case.edu/~jiayang/nFCA.tgz](http://sun.case.edu/~jiayang/nFCA.tgz)

installed on either PC, Mac or Linux. We have tested our code extensively on PC with Ruby and Graphviz interfaces, although it should also work for Mac and Linux. Specifically, the links for the PC versions of R, Ruby and Graphviz (that can be found from the url above) are as of September 20, 2010:


Next, we show the detailed steps needed to configure and run nFCA successfully.

1. Install R
   - Download R from above link
   - Locate the folder where R is saved in your machine and then install R

2. Install Ruby
   - Download Ruby from above link
   - Locate the folder where Ruby is saved in your machine and then install Ruby

3. Install Graphviz
   - Download Graphviz from above link
   - Locate the folder where Graphviz is saved and install Graphviz

4. Install nFCA program
   - Download nFCA package from above link
   - Locate the folder where nFCA is saved and extract to a directory

5. Using nFCA
   - Start R
   - Change R working directory to the folder where nFCA code is extracted to
   - Source the nFCA code in an R command window using:
     ```r
     source('nFCA.r')
     ```
   In the case when R is not interacting with Ruby correctly, check if the Ruby path is added to the system path setting file in the two R code: nFCA.r and nFCAfunctions.r. See the README file that comes with the nFCA cod for more details.
Run nFCA by typing in the command window
nFCA('datafile',man=1)
where the “datafile” is the name of the input, or the numerical data
you are working with, and manual is the indicator of how you want to
choose the threshold. It only takes value 0 and 1. If manual is set to
be 1, one may choose a threshold manually for each thresholding step
based on a Histogram shown automatically on the left of a panel. If it is
0, nFCA program will choose threshold automatically. nFCA program
takes the second largest value of the Histogram breaks as the automati-
cally chosen threshold. The automatically chosen procedure will fail
if only a few objects or too many objects grouped together after us-
ing the chosen value to threshold the matrix. In this situation, nFCA
will generate an error message and notify user to choose a threshold
manually. After an appropriate threshold is chosen manually, nFCA
program will continue and next time it will choose threshold auto-
matically again. A standard description for using nFCA program is
provided in Appendix. See the README file for more details.
The output includes a sequence of files which contain the clus-
tering information from each thresholding step and two files which contain the
resulting combined clustering information (called “Hgraph.txt” by de-
fault) and generated inherent structure information (called “Igraph.txt”).

6. Visualization of H-graph and I-Graph

- Open Graphviz and choose “open” menu to open “Hgraph.txt”, i.e.,
nFCA H-graph, then choose “run” or press F5 to bring up a dialog box
and make sure that the layout engine for H-graph is “fdp”, and finally
click okay to see the H-graph.
- Open Graphviz and choose “open” menu to open “Igraph.txt”, i.e.,
nFCA I-graph, then choose “run” or press F5 to bring up the same
dialog box as the above but use the layout engine for I-graph, “neato”
and then click okay to let it run.
The other return value of nFCA include: status, nFCACop, hcCopAve, hcCopSin, hcCopCom, and eMatrix. The return “status” value will be 0 if the nFCA program runs successfully. See the help file in Section 4.2 for the description of nFCACop, hcCopAve, hcCopSin, hcCopCom, and eMatrix.

Next, we take the Family data as an example to show how to run nFCA program and visualize the resulting hierarchical structure and inherent structure. In this example the parameter manual is set to 0 and the Family data file is data1.txt. In R console, following is the process to run nFCA program,

```
> source( "nFCA.r" )
> nFCA( "data1.txt", manual = 0, method = "Histogram" )
[[1]]
[1] 0

[[2]]
[1] 0.9924285

[[3]]
[1] 0.9719541 0.9418811 0.9639018

[[4]]
[1,] 1.000 0.950 0.900 0.900 0.850 0.850 0.810 0.810 0.810 0.400 0.143 0.039
[2,] 0.950 1.000 0.880 0.880 0.792 0.792 0.870 0.870 0.122 0.450 0.150
[3,] 0.900 0.880 1.000 0.820 0.792 0.792 0.810 0.810 0.122 0.143 0.039
[4,] 0.900 0.880 0.820 1.000 0.792 0.792 0.810 0.810 0.122 0.143 0.039
[5,] 0.850 0.792 0.792 0.792 1.000 0.950 0.625 0.625 0.122 0.143 0.039
[6,] 0.850 0.792 0.792 0.792 0.950 1.000 0.625 0.625 0.122 0.143 0.039
[7,] 0.810 0.870 0.810 0.810 0.625 0.625 0.625 1.000 0.950 0.122 0.143 0.039
[8,] 0.810 0.870 0.810 0.810 0.625 0.625 0.625 0.950 1.000 0.122 0.143 0.039
[9,] 0.400 0.122 0.122 0.122 0.122 0.122 0.122 0.122 0.122 0.122 0.122 0.122
[10,] 0.143 0.450 0.143 0.143 0.143 0.143 0.143 0.143 0.020 1.000 0.039
[11,] 0.039 0.150 0.039 0.039 0.039 0.039 0.039 0.039 0.039 0.039 0.039
```

Once nFCA program is finished, we use Graphviz to visualize the resulting hierarchical cluster result and inherent structure from nFCA as shown in Figure 4.1 and Figure 4.2.
4.2 Help File of nFCA

Description
*nFCA* is used to generate the hierarchical cluster and inherent structure for correlation coefficient or distance matrix.

Usage
```
nFCA( data, manual=1, method="Histogram", cutoff1=3, cutoff2=10, n=30, alpha=0.05 )
```

Arguments
- **data** a numerical data matrix. It can a correlation coefficient matrix or a distance matrix.
- **manual** numerical. If manual=1 the threshold is selected manually. If manual=0 the threshold is selected automatically.
- **method** the method to be used. If method="Histogram" the threshold is chosen using Histogram method. If method="Confidence" the threshold is chosen using confidence interval method.
- **cutoff1** the minimum number of objects involved in the thresholding.
- **cutoff2** the maximum number of objects involved in the thresholding.
- **n** the sample size for confidence interval method.
- **alpha** the confidence level for confidence interval method.

Details
For Histogram method, the automated choice of threshold is the second breaks of the Histogram. Confidence interval method only applies when the data is a correlation coefficient matrix and the sample size used to calculate the sample correlation coefficients is known. The automated choice of threshold for confidence interval method is the lower bound of the calculated confidence interval.

Value
nFCA returns a list which contains the following components:
\textbf{status} \[[1]\], the indicator whether the nFCA program is successful or not. 0 indicates success and 1 indicates fail.

\textbf{nFCACop} \[[2]\], the Cophenetic correlation coefficient for nFCA.

\textbf{hcCopAve} \[[3] [[1]]\], the Cophenetic correlation coefficient for conventional hierarchical cluster using average linkage.

\textbf{hcCopSin} \[[3] [[2]]\], the Cophenetic correlation coefficient for conventional hierarchical cluster using single linkage.

\textbf{hcCopCom} \[[3] [[3]]\], the Cophenetic correlation coefficient for conventional hierarchical cluster using complete linkage.

\textbf{eMatrix} \[[4]\], the estimated correlation coefficient matrix using nFCA.

\section*{4.3 Generalization of nFCA}

nFCA was initially developed as a data analysis method for a particular type of numerical data such as correlation or distance matrix of a set of objects. It can be further generalized to accommodate more general numerical data. For conceptual scaling, the generalization from one multi-valued attribute to multiple multi-valued attributes is quite simple and straightforward. One may apply conceptual scaling to each of the multi-valued attributes to discretize or reduce the numerical data to binary data. Suppose the numerical objects and attributes data under consideration includes multiple multi-valued and single-valued attributes. The following procedure can be used to generalize nFCA to this general numerical data.

1. Normalizing each of the multi-valued attributes,

2. Rescaling each of the multi-valued attributes to the interval \([0, 1]\),

3. Applying nFCA to the normalized and rescaled numerical data along with the rest single-valued attributes in the original data.
Note that the original single-valued attributes in the numerical data will not affect the choice of threshold and the thresholding result. They will also not affect the correlation or distance matrix updating step in nFCA since they are only involved in the step of applying FCA. By using above procedure, we extend nFCA to general numerical objects and attributes data and make it more applicable.
Figure 4.1: Visualization hierarchical cluster in Graphviz
Figure 4.2: Visualization inherent structure in Graphviz
5.1 Discussion and Conclusion

The main value of nFCA is to reveal the underline hierarchical structure among the data and illustrate the relationship network among variables. As a hierarchical cluster technique nFCA provides the relationship among multiple members within a sub-cluster and uses arrows along with their directions and lengths to give a detailed information about the relationships.

One issue of nFCA is that the hierarchical clustering result for a given data is not unique. Different choices of threshold may produce different hierarchical clustering results. Theoretically speaking, different hierarchical clustering results illustrate the underline structure from different angles and different levels. How to choose the most accurate and efficient one becomes an important issue.

5.2 Future Work

Since the choice of threshold is important in nFCA, we may further develop a more sophisticated threshold choice algorithm to achieve better performance. Ideally, the threshold can be chosen using an optimization algorithm or genetic algorithm such that the final sequence of thresholds will lead to the maximum Cophenetic correlation value. Hence the nFCA based on this sequence of thresholds
will have the best performance among all the possible sequences of thresholds.
Chapter 6

Appendix

6.1 Proof of Proposition 2.4.1

First we calculate the mean of \( \hat{f}_{\text{Hist}}(x) \) as following

\[
E \left( \hat{f}_{\text{Hist}}(x) \right) = E \left( \frac{1}{th} \sum_{i=1}^{t} I_{X_i \in B_j} \right) \\
= \frac{1}{h} P \left( b_j < X < b_{j+1} \right), \quad \text{where} \quad X \sim f^*(x) \\
= \frac{1}{h} \int_{b_j}^{b_{j+1}} f^*(y)dy \\
= \frac{1}{h} \int_{b_j}^{b_{j+1}} \left( f^*(x) + f^\prime(x)(y-x) + f^\prime\prime(\xi)(y-x)^2 \right) dy \\
= f^*(x) + \frac{1}{h} f^\prime(x) \frac{1}{2} (y-x)^2 \big|_{b_j}^{b_{j+1}} + O(h^2) \\
= f^*(x) + \frac{1}{2} f^\prime(x)(h+2b_j - 2x) + O(h^2).
\]
For the variance of \( \hat{f}_{\text{Hist}}(x) \), we have

\[
\text{Var} \left( \hat{f}_{\text{Hist}}(x) \right) = \frac{1}{t^2 h^2} \text{Var} \left( \sum_{i=1}^{t} I(X_i \in B_j) \right)
\]

\[
= \frac{1}{t^2 h^2} \left( \sum_{i=1}^{t} \text{Var} \left( I(X_i \in B_j) \right) + 2 \sum_{l < m} \text{Cov} \left( I(X_l \in B_j), I(X_m \in B_j) \right) \right)
\]

\[
= A + B,
\]

where

\[
A = \frac{1}{t^2 h^2} \left( \sum_{i=1}^{t} \text{Var} \left( I(X_i \in B_j) \right) \right)
\]

and

\[
B = \frac{1}{t^2 h^2} \left( 2 \sum_{l < m} \text{Cov} \left( I(X_l \in B_j), I(X_m \in B_j) \right) \right).
\]

Now we calculate \( A \) first as following

\[
A = \frac{1}{th^2} \left( E \left( I^2(X_1 \in B_j) \right) - \left( E \left( I(X_1 \in B_j) \right) \right)^2 \right)
\]

\[
= \frac{1}{th^2} \left( E \left( I(X_1 \in B_j) \right) - \left( E \left( I(X_1 \in B_j) \right) \right)^2 \right)
\]

\[
= \frac{1}{th} \left( f^*(x) + \frac{1}{2} f''(x)(h + 2b_j - 2x) + O(h^2) \right.
\]

\[
- h \left( f'^2(x) + O(h^2) \right) \}
\]

\[
= \frac{1}{th} f^*(x) + O\left( \frac{C_1(x)}{n} \right).
\]

where \( C_1(x) \) is a function of \( x \). For \( B \) we have

\[
B = \frac{2}{t^2 h^2} \sum_{l < m} \text{Cov} \left( I(X_l \in B_j), I(X_m \in B_j) \right)
\]

\[
= \frac{2}{t^2 h^2} \sum_{l < m} \left( E \left( I(X_l \in B_j) I(X_m \in B_j) \right) - \left( E \left( I(X_l \in B_j) \right) \right)^2 \right).
\]
Notice that $E \left( I_{(X_l \in B_j)} I_{(X_m \in B_j)} \right) = P \left\{ (X_l \in B_j) \cap (X_m \in B_j) \right\}$. Suppose $f^*(x_l, x_m)$ is the joint density function of $(x_l, x_m)$, then this probability can be written as

$$
P \left\{ (X_l \in B_j) \cap (X_m \in B_j) \right\} = \int_{b_j}^{b_{j+1}} \int_{b_j}^{b_{j+1}} f^*(u, v) dudv
$$

$$
= \int_{b_j}^{b_{j+1}} \int_{b_j}^{b_{j+1}} \left( f^*(x, x) + \frac{\partial}{\partial u} f^*(x, x)(u - x) + \frac{\partial}{\partial v} f^*(x, x)(v - x) + O(h^2) \right) dudv
$$

$$
= f^*(x, x) h^2 + O(h^3).
$$

We also have

$$
\left( E \left( I_{(X_l \in B_j)} \right) \right)^2 = f^{*2}(x) h^2 + O(h^3).
$$

Based on the assumption, $\sum_{l < m} (f^*(x, x) - f^{*2}(x)) < M$, $B$ can be calculated as

$$
B = \frac{2}{t^2 h^2} \sum_{l < m} (f^*(x, x) - f^{*2}(x)) h^2 + O(h^3)
$$

$$
= O(B(x)h) + O\left( \frac{C_2(x)}{t} \right)
$$

where $C_2(x)$ is a function of $x$. Finally, the variance of $\hat{f}_{Hist}(x)$ can be calculated as

$$
Var(\hat{f}_{Hist}(x)) = A + B
$$

$$
= \frac{1}{th} f^*(x) + O\left( \frac{C(x)}{t} \right) + O(B(x)h)
$$

where $C(x)$ and $B(x)$ are functions of $x$. 

Part II

Bayesian Predictive Inference for Finite Population Quantities under Informative Sampling
Chapter 7

Introduction

7.1 Inference for Survey Data

Survey data can be considered as the outcome of two random processes: The first random process is to generate the finite population and the second process is to select a sample from the finite population. Here, define the population model as the model used to generate the finite population. Survey data are used to make inference for the parameters of the population model and predictive inference for the unobserved (i.e., non-sampled) units in the finite population. Thus one can make inference for any characteristic of the finite population values, e.g., mean, median, and quantiles.

7.2 Informative and Ignorable Sampling

Survey data are usually collected from samples drawn with unequal selection probabilities from a finite population. Most often, a set of weights for the observed (i.e., sampled) units is presented to reflect the unequal selection probabilities. These are called the sampling weights. Under this circumstance, the resulting survey data are a biased sample from the target population. If the selection probabilities of a sample are related to the response values of the model (even after conditioning on the model covariates), we say the sampling is informative and the selection process is called informative sampling. Under informative sampling
the distribution holding for the sample is different from the distribution holding for the finite population. These two distributions are called the sample distribution and (finite) population distribution respectively. Making inferences based on the sample without accounting for the informative sampling may produce large biases.

Suppose \( Y \) is the response variable, \( \tilde{\theta} \) is the vector of parameters for the population model setting. Letting \( U \) denote the finite population indices and \( S \) the sampled unit indices, the sample distribution for a single element \( y_i \), \( f_s(y_i|\tilde{\theta}) \equiv f(y_i|\tilde{\theta}, i \in S) \), and the population distribution for \( y_i \), \( f_p(y_i|\tilde{\theta}) \equiv f(y_i|\tilde{\theta}, i \in U) \), are different under informative sampling. Specifically, we have

\[
f_s(y_i|\tilde{\theta}) = \frac{Pr(i \in S|y_i) f_p(y_i|\tilde{\theta})}{Pr(i \in S)}.
\]

From (7.1) we can see that \( f_s(y_i|\tilde{\theta}) \) and \( f_p(y_i|\tilde{\theta}) \) are different for informative sampling unless \( Pr(i \in S|y_i) = Pr(i \in S) \).

If the sample distribution and population distribution remain the same under a sampling scheme, we say this sampling scheme is ignorable. For example, suppose \( \tilde{Z} \) denotes the population values of all the design variables. Then the sample distribution can be written as

\[
f_s(y_i|\tilde{\theta}, \tilde{Z}) \equiv f(y_i|\tilde{\theta}, \tilde{Z}, i \in S) = Pr(i \in S|y_i, \tilde{Z}) f_p(y_i|\tilde{\theta}, \tilde{Z}) / Pr(i \in S|\tilde{Z}).
\]

It is clear that the sample distribution, \( f_s(y_i|\tilde{\theta}, \tilde{Z}) \), is not the same as the population distribution, \( f_p(y_i|\tilde{\theta}, \tilde{Z}) \), unless the selection probability only depends on \( \tilde{Z} \). If we assume that the selection probability only depends on \( \tilde{Z} \), i.e.,

\[
Pr(i \in S|y_i, \tilde{Z}) = Pr(i \in S|\tilde{Z}),
\]

this sampling scheme is ignorable and standard inference procedures can be applied to make inferences.

### 7.3 Probability Proportional to Size Sampling

Probability proportional to size (PPS) sampling is the most common type of sampling technique with unequal selection probabilities. For PPS sampling, the probability of selecting a sampling unit is proportional to its “size”, i.e., an auxiliary
variable. This auxiliary variable, i.e., size measure, is assumed to be correlated with the response variable. Suppose $v_i$ is the size measure for unit $i$ and $I_i$ is the indicator variable, where $I_i = 1$ if unit $i$ is selected into the sample and $I_i = 0$ otherwise. Then the probability for unit $i$ to be selected in a single draw is

$$Pr(I_i = 1) = \frac{v_i}{\sum_{j=1}^{N} v_j}$$

where $N$ is the total number of units in the finite population.

For a discussion of PPS sampling with and without replacement, see Cochran (1977)[1].

References

Chapter 8

Literature Review

Recently, researchers have tackled several difficult problems when the sample is selected with unequal selection probabilities and only a limited amount of information is available. Chambers, Dorfman, and Wang (1998)[1], henceforth CDW, consider a situation where one wishes to model the population process used to generate the finite population. They assume that the sample is selected with unequal selection probabilities from the finite population and only the first-order selection probabilities for the sampled units are known. They establish a theoretical framework for their maximum likelihood approach. However, they also state that “it is almost impossible to proceed without fixing ideas on an example.” Let $\Theta$ denote the parameters of the population distribution and $Y_s, D_s, \pi_s$ and $I$ denote the values of the variate of interest for the sampled units, the design information for the sampled units, the vector of inclusion probabilities for the sampled units and the sample inclusion indicator respectively. The score function for $\Theta$ given $Y_s, D_s, \pi_s$, and $I$, is

$$sc_\Theta(\Theta;Y_s, D_s, \pi_s, I) = E(\partial \Theta \left[ \log \{ f(Y, D, I) \} \right] | Y_s, D_s, \pi_s, I)$$

(8.1)

where $\partial \Theta$ symbolizes differentiation with respect to $\Theta$ and $(Y, D)$ are the population counterparts of $(Y_s, D_s)$. The maximum likelihood estimator (MLE) for $\Theta$ is obtained by setting (8.1) equal to 0. Their methodology is quite important since they provide a solution for such a complicated situation. However, their methodology is limited in that it only makes inference for the population process and inference depends on asymptotic results.
Nandram and Sedransk (2004)[2] consider a problem similar to the one investigated by CDW (1998)[1]. They extend the probabilistic specification of CDW (1998)[1] so that it will accommodate many sample surveys. Unlike that of CDW (1998)[1], their full-Bayesian approach using Markov chain Monte Carlo avoids the necessity of using asymptotic approximations. Their methodology not only can make exact inference for the population parameters but also can make predictive inference for the non-sampled units. However, they did not handle the non-identifiability problem in their model in a fully satisfactory way, and did not account for all the restrictions needed to ensure that all of the unknown selection probabilities are in $[0, 1]$.

Pfeffermann, Krieger, and Rinott (1998)[3] proposed a general method to make inference for the population distribution under informative sampling by approximating the parametric distribution of the sample measurements. They start with the conditional marginal sample distribution $f_s(y_i|x_i)$,

$$f_s(y_i|x_i) = f_p(y_i|x_i, I_i = 1) = \frac{Pr(I_i = 1|x_i, y_i)f_p(y_i|x_i)}{Pr(I_i = 1|x_i)}$$

where $I_i$ is the indicator variable, $y_i$ is the model response, and $x_i$ is the vector of covariates. $Pr(I_i = 1|x_i, y_i)$ can be further written as

$$Pr(I_i = 1|x_i, y_i) = \int Pr(I_i = 1|\pi_i, x_i, y_i)f_p(\pi_i|x_i, y_i)d\pi_i = \int \pi_i f_p(\pi_i|x_i, y_i)d\pi_i = E_p(\pi_i|x_i, y_i)$$

where $\pi_i$ is the inclusion probability of unit $i$ which may depend on the population values of design variables $\tilde{Z}$. Then the conditional marginal sample distribution $f_s(y_i|x_i)$ can be rewritten as

$$f_s(y_i|x_i) = f_p(y_i|x_i, I_i = 1) = \frac{E_p(\pi_i|x_i, y_i)f_p(y_i|x_i)}{E_p(\pi_i|x_i)}.$$ 

Assuming the asymptotic independence of the sample measurements, proved in Pfeffermann, Krieger, and Rinott (1998)[3], the full distribution of $(Y_s, I)$ can be
shown to be,

\[
f(y_s, I|X) = \prod_{i \in S} \left[ \frac{\pi(y_i, x_i) f_p(y_i|x_i)}{\pi_0 i} \right] \prod_{i \in S} \pi_0 i \prod_{i \notin S} (1 - \pi_0 i) \tag{8.2}
\]

where \( \pi(y_i, x_i) = E(\pi|y_i, x_i) \), \( \pi_0 i = \int \pi(y_i, x_i) f_p(y_i|x_i) dy_i = Pr(i \in S|x_i) \) and \( X = \{x_i, i \in U\} \). Since \( \prod_{i \notin S} (1 - \pi_0 i) \) depends on all the individual vectors \( x_i, i \notin S \), and these vectors are usually not part of the data, they use the following approximation to make inference for the parameters of the population distribution, i.e.,

\[
f(y_s, I|X) \approx \prod_{i \in S} \left[ \frac{\pi(y_i, x_i) f_p(y_i|x_i)}{\pi_0 i} \right]. \tag{8.3}
\]

This approach provides a general solution to make inference for the parameters of the population distribution. However, it is based on the asymptotic independence assumption for the sample measurements in (8.2) and also the approximation of the likelihood in (8.3) which is hard to defend.

Pfeffermann and Sverchkov (2009)[5] give a excellent review of inference under informative sampling. In the review, they investigate the relationship between the sample distribution and the population distribution and show how to use the sample distribution for inference about the population distribution parameters. Pfeffermann and Sverchkov (1999)[4] show that for a general pair of vector random variables \((\nu_1, \nu_2)\) measured for unit \( i \in U \):

\[
E_p(\nu_1|\nu_2) = \frac{E_s(w_i \nu_1|\nu_2)}{E_s(w_i|\nu_2)}
\]

and

\[
E_p(\pi_i|\nu_2) = \frac{1}{E_s(w_i|\nu_2)}
\]

where \( w_i \) is the sampling weight for unit \( i \) and \( w_i = \frac{1}{\pi_i} \).

After adding parameters and defining \( A_s \) as the selected sample, the sample distribution \( f_s(y_i|x_i; \theta, \gamma) \) is

\[
f_s(y_i|x_i; \theta, \gamma) = \frac{E_p(\pi_i|y_i, x_i; \gamma) f_p(y_i|x_i; \theta)}{E_p(\pi_i|x_i; \theta, \gamma)} \]

where

\[
E_p(\nu_1|\nu_2) = \frac{E_s(w_i \nu_1|\nu_2)}{E_s(w_i|\nu_2)}
\]

and

\[
E_p(\pi_i|\nu_2) = \frac{1}{E_s(w_i|\nu_2)}
\]

where \( w_i \) is the sampling weight for unit \( i \) and \( w_i = \frac{1}{\pi_i} \).
where $\theta$ is the population parameter and $\gamma$ is the parameter governing $Pr(A_s = 1|Y_p = y_p; \gamma)$. Then the population distribution $f_p(y_i|x_i; \theta)$ can be written as

$$f_p(y_i|x_i; \theta) = \frac{E_s(w_i|y_i,x_i; \gamma)f_s(y_i|x_i; \theta, \gamma)}{E_s(w_i|x_i; \theta, \gamma)} = \frac{E_s(w_i|y_i,x_i; \gamma)}{E_s(w_i|x_i; \theta, \gamma)} f_s(y_i|x_i; \theta).$$

The inference for the parameters of the population distribution is based on the framework in Pfeffermann, Krieger, and Rinott (1998)[3]. Now, the $w_i$’s are known for the sampled units. Then $E_s(w_i|y_i,x_i; \gamma)$ can be estimated by regressing $w_i$ against $(y_i,x_i)$ using classical model fitting procedures, $f_s(y_i|x_i; \theta)$ can be estimated using classical procedures applied to the observed data, and $E_s(w_i|x_i; \theta, \gamma)$ can be estimated in the similar way. Thus $f_p(y_i|x_i; \theta)$ can be estimated based on only the sample outcomes and the weights without knowing the design variables.

The approach proposed by Pfeffermann and Sverchkov (2009)[5] shows how to estimate the population distribution based only on the sample outcomes and the weights without knowing the values of design variables. However, as with Pfeffermann, Krieger, and Rinott (1998)[3] their approach is based on the asymptotic independence assumption for the sample measurements and also the likelihood approximation (see (8.3)). Their approach provides both point estimates and variance estimates for the population distribution parameters.

References


Chapter 9

Notation and Background

We assume that the finite population under consideration consists of \( N \) units \( \{Y_1, Y_2, \ldots, Y_N\} \) denoted by \( \tilde{Y} \) which are drawn from a distribution with parameters \( \mu \) and \( \sigma^2 \). A sample of size \( n \), \( S \), is drawn from this finite population with unequal selection probabilities. The selection indicator variables are denoted by \( I_1, I_2, \ldots, I_N \), where \( I_i \) takes value 1 if unit \( i \) is selected into the sample and 0 otherwise. After a sample is selected, the indicator vector \( \tilde{I}, \tilde{I} = (I_1, I_2, \ldots, I_N) \), is observed and \( \sum_{i=1}^{N} I_i = n \). For the theoretical development, we assume, as CDW do, that the selection mechanism can be approximated by a version of Poisson sampling. Thus, the \( I_i, i = 1, 2, \ldots, N \), are assumed to be independent Bernoulli variables such that \( \pi_i \equiv Pr(I_i = 1) = 1 - Pr(I_i = 0) \). Thus, conditioning on \( \tilde{\pi} = (\pi_1, \pi_2, \ldots, \pi_N) \), we have

\[
Pr(\tilde{I} | \tilde{\pi}) = \prod_{i \in S} Pr(I_i = 1) \prod_{i \notin S} Pr(I_i = 0) = \prod_{i \in S} \pi_i \prod_{i \notin S} (1 - \pi_i).
\]

The size measure used for Poisson sampling is denoted by \( v \). Following CDW, we assume the selection probability for unit \( i \) is

\[
\pi_i \equiv Pr(I_i = 1) = \frac{n \nu_i}{\sum_{j=1}^{N} \nu_j} \tag{9.1}
\]

where \( \sum_{j=1}^{N} \nu_j \) is denoted by \( t \) and given \( \beta_0, \beta_1, \sigma^2_e \) and the \( Y_i \),

\[
\nu_i = \beta_0 + \beta_1 Y_i + e_i, \quad i = 1, 2, \ldots, N
\]
where \( e_i | \sigma_e^2 \sim \text{iid } N(0, \sigma_e^2) \). It is clear that the selection probability \( \pi_i \) for unit \( i \) is related to its response value \( Y_i \) in a nonlinear way. This is a realistic specification for many establishment surveys where the measure of size \( \nu_i \) for unit \( i \) is linearly related to the response value \( Y_i \).

For convenience, we assume the units indexed by \( \{1, 2, \cdots, n\} \) are the sampled units and the units indexed by \( \{n+1, n+2, \cdots, N\} \) are non-sampled. Thus, we have \( I_i = 1 \) for \( i = 1, 2, \cdots, n \) and \( I_i = 0 \) for \( i = n+1, n+2, \cdots, N \). The selection probabilities, \( \{\pi_i : i \in S\} = \{\pi_1, \pi_2, \cdots, \pi_n\} \), and the response values, \( \{y_i : i \in S\} = \{y_1, y_2, \cdots, y_n\} \), for the sampled units are assumed known and denoted by \( \tilde{\pi}_s \) and \( \tilde{Y}_s \). We use \( \tilde{\pi}_{ns} \) and \( \tilde{Y}_{ns} \) to denote the selection probabilities and response values for the non-sampled units. The latent variable vector \( \{\nu_1, \nu_2, \cdots, \nu_N\} \) is denoted by \( \tilde{\nu} \). We also use \( \tilde{\psi} \) and \( \tilde{\eta} \) to denote the parameters \( (\mu, \sigma^2) \) and \( (\beta_0, \beta_1, \sigma_e^2) \), respectively.

For convenience, we transform \( \tilde{\nu} = \{\nu_1, \nu_2, \cdots, \nu_N\} \) to \( \tilde{Z} = \{Z_1, Z_2, \cdots, Z_N\} \), as follows so that the \( Z_i \)’s are centered at 0,

\[
Z_1 = v_1 - \tilde{\nu}, \quad Z_2 = v_2 - \tilde{\nu}, \quad \cdots, \quad Z_{N-1} = v_{N-1} - \tilde{\nu}, \quad Z_N = \tilde{\nu}
\]

where \( \tilde{\nu} = \frac{1}{N} \sum_{j=1}^{N} v_j = \frac{1}{N} tr(v) \).

Next, we develop full-Bayesian approaches under two different model assumptions, i.e.,

1. Finite population follows normal distribution and the total of measures of size, \( t \), is known.
2. Finite population follows log-normal distribution and the total of measures of size, \( t \), is known.
Chapter 10

Normal Model with Known $t$

We assume the population distribution is normal with mean $\mu$ and variance $\sigma^2$, i.e.,

$$Y_i | \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2)$$

where $i = 1, 2, \cdots, N$ and $N(\mu, \sigma^2)$ denotes the normal distribution with mean $\mu$ and variance $\sigma^2$. From (9.1) we see that $(v_1, v_2, \cdots, v_N)$ and $(cv_1, cv_2, \cdots, cv_N)$ will produce the same selection probability $\pi_i$ for $c \neq 0$. This will cause the non-identifiability problem. To avoid the non-identifiability problem, we assume that the total of measures of size for the finite population, $\sum_{j=1}^{N} v_j = t$, is known. Since $\{\pi_i : i = 1, 2, \cdots, n\}$ and $t$ are known, it’s clear from (9.1) that $\tilde{v}_n = (v_1, v_2, \cdots, v_n)$ is known. The size measures for the non-sampled units, $\{v_{n+1}, v_{n+2}, \cdots, v_N\}$, denoted by $\tilde{v}_{ns}$, are not known. From the transformation from $\tilde{v}$ to $\tilde{Z}$, $(z_1, z_2, \cdots, z_n, z_N)$ are known and denoted by $\tilde{z}_s$ while $(z_{n+1}, z_{n+2}, \cdots, z_{N-1})$ are unknown and denoted by $\tilde{z}_{ns}$.

10.1 Distribution of $\tilde{Z}$ Under Restrictions

Based on the transformation from $\tilde{v}$ to $\tilde{Z}$, the reverse transformation is

$$v_1 = Z_1 + Z_N, \quad v_2 = Z_2 + Z_N, \quad \cdots, \quad v_{N-1} = Z_{N-1} + Z_N, \quad v_N = Z_N - \sum_{i=1}^{N-1} Z_i.$$
The Jacobian of this transformation is

\[
J = \begin{pmatrix}
1 & 0 & 0 & \cdots & 0 & 1 \\
0 & 1 & 0 & \cdots & 0 & 1 \\
0 & 0 & 1 & \cdots & 0 & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 1 \\
-1 & -1 & -1 & \cdots & -1 & -1
\end{pmatrix}_{N \times N}
\]

and \(|J| = N\). For \(i = 1, 2, \cdots, N\), the distribution of \(v_i\) given \(\beta_0, \beta_1, \sigma_e^2\), and \(Y_i\) is

\[
v_i | \beta_0, \beta_1, \sigma_e^2, Y_i \sim N(\beta_0 + \beta_1 Y_i, \sigma_e^2)
\]

i.e.,

\[
f_{v_i}(v_i) = \frac{1}{\sqrt{2\pi}\sigma_e} \exp\left\{ -\frac{(v_i - \beta_0 - \beta_1 Y_i)^2}{2\sigma_e^2} \right\}.
\]

Since the \(v_i\) are independent the joint distribution of \(\bar{v} = (v_1, v_2, \cdots, v_N)\) given \(\beta_0, \beta_1, \sigma_e^2\), and \(\bar{Y}\) is

\[
f_{\bar{v}}(v_1, v_2, \cdots, v_N) = \prod_{i=1}^{N} \left[ \frac{1}{\sqrt{2\pi}\sigma_e} \exp\left\{ -\frac{(v_i - \beta_0 - \beta_1 Y_i)^2}{2\sigma_e^2} \right\} \right].
\]

Then the distribution of \(\bar{Z}\) given \(\beta_0, \beta_1, \sigma_e^2\), and \(\bar{Y}\) can be written as

\[
f_{\bar{Z}}(z_1, z_2, \cdots, z_N) = N \left( \sqrt{2\pi}\sigma_e \right)^{-N} \exp\left\{ -\frac{1}{2\sigma_e^2} \left[ Nz_N^2 - 2z_N \sum_{i=1}^{N} \theta_i \right. \right.
\]

\[
+ \left( \sum_{i=1}^{N} z_i + \theta_N \right)^2 + \sum_{i=1}^{N} (z_i - \theta) \left. \right] \right\}
\]

where \(\theta_j = \beta_0 + \beta_1 y_j\) for \(j = 1, 2, \cdots, N\).

In order to find the restrictions for \(\bar{Z}\), we start from the restrictions for \(\bar{v}\). The restrictions for \(\bar{v}\) can be summarized as following

1. \(0 \leq \frac{n_v}{\sum_{j=1}^{N} v_j} \leq 1\) for \(i = 1, 2, \cdots, N\);
2. $\sum_{j=1}^{N} \nu_j = t$.

Then from the transformation for $\tilde{\nu}$ to $\tilde{Z}$, the restrictions for $\tilde{Z}$ are

1. $z_N = \frac{t}{N}$;

2. $-\frac{t}{N} \leq z_i \leq \frac{t}{n} - \frac{t}{N}$ for $i = 1, 2, \ldots, N - 1$;

3. $\frac{t}{N} - \frac{t}{n} \leq \sum_{j=1}^{N-1} z_j \leq \frac{t}{N}$.

Given $Z_N = \frac{t}{N}$, the conditional distribution of $(Z_1, Z_2, \ldots, Z_{N-1})$ is

$$f \left( z_1, z_2, \ldots, z_{N-1} | Z_N = \frac{t}{N} \right) = \frac{1}{f(Z_N = \frac{t}{N})} f \left( z_1, z_2, \ldots, z_{N-1}, Z_N = \frac{t}{N} \right).$$

Then under the second and third restrictions for $\tilde{Z}$ the conditional distribution of $(Z_1, Z_2, \ldots, Z_{N-1})$ is

$$f_{R_0} \left( z_1, z_2, \ldots, z_{N-1} | Z_N = \frac{t}{N} \right) = \frac{f \left( z_1, z_2, \ldots, z_{N-1} | Z_N = \frac{t}{N} \right)}{\int_{R_0} f \left( z_1, z_2, \ldots, z_{N-1} | Z_N = \frac{t}{N} \right) dz_1 dz_2 \cdots dz_{N-1}}$$

$$= \frac{f \left( z_1, z_2, \ldots, z_{N-1}, Z_N = \frac{t}{N} \right)}{\int_{R_0} f \left( z_1, z_2, \ldots, z_{N-1}, Z_N = \frac{t}{N} \right) dz_1 dz_2 \cdots dz_{N-1}}$$

where

$$R_0 = \left\{ (z_1, z_2, \ldots, z_{N-1}) \mid -\frac{t}{N} \leq z_i \leq \frac{t}{n} - \frac{t}{N} \text{ for } i = 1, 2, \ldots, N - 1; \right\}$$

$$\frac{t}{N} - \frac{t}{n} \leq \sum_{j=1}^{N-1} z_j \leq \frac{t}{N} \right\}.$$ 

Finally, the distribution of $(Z_1, Z_2, \ldots, Z_{N-1}, Z_N)$ under the restrictions is

$$f_{R_0} (z_1, z_2, \ldots, z_{N-1}, z_N) = \frac{1}{C(\bar{y}_N, \bar{\psi}, \bar{\eta})} N(\sqrt{2\pi}\sigma_e)^{-N} \exp \left\{ -\frac{1}{2\sigma_e^2} \left[ \frac{t^2}{N} - 2 \frac{t}{N} \sum_{i=1}^{N} \theta_i \right] + \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right\}.$$
where \( Z_N = \frac{t}{N}, (Z_1, Z_2, \ldots, Z_{N-1}) \in R_0 \) and

\[
C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta}) = \int_{R_0} N(\sqrt{2\pi}\sigma_e)^{-N} \exp \left\{ -\frac{1}{2\sigma_e^2} \frac{t^2}{N} - \frac{2t}{N} \sum_{i=1}^{N} \theta_i + \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 \right. \\
\left. + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right\} dz_1 dz_2 \cdots dz_{N-1}.
\]

### 10.2 Posterior Distribution

The posterior distribution for \( \tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi} \) and \( \tilde{\eta} \) can be written as

\[
P(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \tilde{y}_s, \tilde{z}_s, I) \propto P(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta}, I) \\
= P(\tilde{I} | \tilde{y}, \tilde{\psi}, \tilde{\eta}) P(\tilde{z} | \tilde{I}, \tilde{\psi}, \tilde{\eta}) P(\tilde{\psi}) P(\tilde{\eta}),
\]

where \( \tilde{I} \subseteq R \) and

\[
R = \left\{ (z_{n+1}, z_{n+2}, \ldots, z_{N-1}) \mid \frac{t}{N} \leq z_i \leq \frac{t}{n} - \frac{t}{N} \text{ for } i = n + 1, n + 2, \ldots, N - 1; \right. \\
\left. \frac{t}{N} - \frac{t}{n} - \sum_{j=1}^{n} z_j \leq \sum_{j=n+1}^{N-1} z_j \leq \frac{t}{N} - \sum_{j=1}^{n} z_j \right\}.
\]

For each component of the posterior distribution we have

\[
P(\tilde{I} | \tilde{y}, \tilde{\psi}, \tilde{\eta}) = \prod_{i \in S} \pi_i \prod_{i \notin S} (1 - \pi_i) \\
= \left( \frac{n}{Nz_N} \right)^n \prod_{i=1}^{n} (z_i + z_N) \prod_{i=n+1}^{N-1} \left[ 1 - \frac{n}{Nz_N} (z_i + z_N) \right] \\
\times \left[ 1 - \left( \frac{n}{Nz_N} \right) \left( z_N - \sum_{i=1}^{N-1} z_i \right) \right],
\]

\[
P(\tilde{z} | \tilde{y}, \tilde{\psi}, \tilde{\eta}) = \frac{1}{C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta})} N(\sqrt{2\pi}\sigma_e)^{-N} \exp \left\{ -\frac{1}{2\sigma_e^2} \frac{t^2}{N} - \frac{2t}{N} \sum_{i=1}^{N} \theta_i \\
+ \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right\},
\]
\[
P(\tilde{y}, \tilde{\eta}) = (\sqrt{2\pi}\sigma)^{-N} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2 \right\}
\]

and \(P(\tilde{\psi}, \tilde{\eta})\) is the prior distribution. For \((\tilde{\psi}, \tilde{\eta})\), we use non-informative prior distribution, i.e., \(P(\tilde{\psi}, \tilde{\eta}) \propto \sigma^{-2}\sigma_v^{-2}\).

The resulting posterior distribution is
\[
\begin{align*}
P(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \tilde{y}_s, \tilde{z}_s, \tilde{I}) & \propto (\sigma_e\sigma)^{-N-2} \prod_{i=n+1}^{N-1} \left[ 1 - \frac{n}{Nz_N} (z_i + z_N) \right] \\
& \times \left[ 1 - \left( \frac{n}{Nz_N} \right) \left( z_N - \sum_{i=1}^{N-1} z_i \right) \right] \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2 \right\} \\
& \exp \left\{ -\frac{1}{2\sigma^2} \left[ \frac{1}{N} \sum_{i=1}^{N} \theta_i + \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right] \right\}
\end{align*}
\]

where \(\tilde{z}_{ns} \in \mathbb{R}\).

### 10.3 Sampling Importance Resampling (SIR)

We use the Gibbs sampler (Gelfand and Smith 1990[1]) to draw samples from the posterior distribution in (10.1). For convenience, the posterior distribution can be further rewritten as
\[
P(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \tilde{y}_s, \tilde{z}_s, \tilde{I}) \propto \frac{1}{L} \pi_a(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \tilde{y}_s, \tilde{z}_s, \tilde{I})
\]

where \(\tilde{z}_{ns} \in \mathbb{R}\) and
\[
L = \prod_{i=n+1}^{N-1} \left[ 1 - \frac{n}{Nz_N} (z_i + z_N) \right] \left[ 1 - \left( \frac{n}{Nz_N} \right) \left( z_N - \sum_{i=1}^{N-1} z_i \right) \right].
\]
The Gibbs sampler is used to draw samples from \( \pi_u(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \bar{y}_s, \bar{z}_s, \bar{I}) \) and the SIR algorithm is used to subsample the Gibbs sample to get a sample from the posterior distribution, \( P(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \bar{y}_s, \bar{z}_s, \bar{I}) \).

Letting \( \Omega^{(k)} = (\tilde{y}^{(k)}_{ns}, \tilde{z}^{(k)}_{ns}, \tilde{\psi}^{(k)}, \tilde{\eta}^{(k)}) \), \( k = 1, 2, \cdots, M \) where \( M \) is the number of iterates obtained from the Gibbs sampler for draws made from \( \pi_u(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \bar{y}_s, \bar{z}_s, \bar{I}) \), the weights in the SIR algorithm are \( w_k = \tilde{w}_k / \sum_{k'}^M \tilde{w}_{k'} \), \( k = 1, 2, \cdots, M \) where

\[
\tilde{w}_k = \frac{P(\tilde{y}^{(k)}_{ns}, \tilde{z}^{(k)}_{ns}, \tilde{\psi}^{(k)}, \tilde{\eta}^{(k)} | \bar{y}_s, \bar{z}_s, \bar{I})}{\pi_u(\tilde{y}^{(k)}_{ns}, \tilde{z}^{(k)}_{ns}, \tilde{\psi}^{(k)}, \tilde{\eta}^{(k)} | \bar{y}_s, \bar{z}_s, \bar{I})}.
\]

By (10.2), \( \tilde{w} \propto L(\hat{z}^{(k)}_{ns}) / C(\hat{y}^{(k)}_{ns}, \hat{\psi}^{(k)}, \hat{\eta}^{(k)}) \), \( k = 1, 2, \cdots, M \). Thus, the SIR algorithm is performed by drawing \( M_0 \) iterates from \( (\Omega^{(k)}, w_k), k = 1, 2, \cdots, M \) without replacement.

### 10.4 Gibbs Sampler

We use the Gibbs sampler to draw samples from \( \pi_u(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \bar{y}_s, \bar{z}_s, \bar{I}) \). In order to perform the Gibbs sampler we need to find the conditional distributions of \( \tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \) and \( \tilde{\eta} \) respectively given everything else.

The conditional distribution of \( \tilde{y}_{ns} \) given \( (\bar{y}_s, \bar{z}_N, \bar{\psi}, \bar{\eta}, \bar{I}) \) is

\[
P(\tilde{y}_{ns} | \bar{y}_s, \bar{z}_N, \bar{\psi}, \bar{\eta}, \bar{I}) \propto \exp \left\{ T_1 \sum_{i=n+1}^{N} \left( y_i + \frac{T_i}{2 T_1} \right)^2 + T_1 \left( y_N + \frac{T_2}{2 T_1} \right)^2 \right\}
\]
where

\[
T_1 = -\frac{1}{2\sigma^2} - \frac{\beta_1^2}{2\sigma_e^2}, \quad T_2 = \frac{\mu}{\sigma^2} - \frac{\beta_1 (\beta_0 + \sum_{i=1}^{N-1} z_i - z_N)}{\sigma_e^2},
\]

\[
T_i = \frac{\mu}{\sigma^2} - \frac{\beta_1 (\beta_0 - z_i - z_N)}{\sigma_e^2}, \quad i = n + 1, n + 2, \ldots, N - 1.
\]

From (10.3), we see that given \((y_{n+1}, \ldots, y_{N-1}, \tilde{y}_s, \tilde{z}_N, \tilde{\psi}, \tilde{\eta}, \tilde{I})\), \(y_N\) has a normal distribution with mean \(-\frac{T_2}{2T_1}\) and variance \(-\frac{1}{2T_1}\), i.e.,

\[
y_N | y_{n+1}, \ldots, y_{N-1}, \tilde{y}_s, \tilde{z}_N, \tilde{\psi}, \tilde{\eta}, \tilde{I} \sim N \left( -\frac{T_2}{2T_1}, -\frac{1}{2T_1} \right).
\]

Also, given \((y_{n+1}, \ldots, y_{i-1}, y_i, \tilde{y}_s, \tilde{z}_N, \tilde{\psi}, \tilde{\eta}, \tilde{I})\), \(y_i\) has a normal distribution with mean \(-\frac{T_i}{2T_1}\) and variance \(-\frac{1}{2T_1}\), i.e.,

\[
y_i | y_{n+1}, \ldots, y_{i-1}, y_i, y_{N-1}, \tilde{y}_s, \tilde{z}_N, \tilde{\psi}, \tilde{\eta}, \tilde{I} \sim N \left( -\frac{T_i}{2T_1}, -\frac{1}{2T_1} \right)
\]

for \(i = n + 1, n + 2, \ldots, N - 1\).

The conditional distribution of \(z_{ns}\) given \((\tilde{y}_N, \tilde{z}_s, \tilde{\psi}, \tilde{\eta}, \tilde{I})\) is

\[
P(z_{ns} | \tilde{y}_N, \tilde{z}_s, \tilde{\psi}, \tilde{\eta}, \tilde{I}) \propto \exp \left\{ -\frac{1}{2\sigma_e^2} \left[ \left( \sum_{n+1}^{N-1} z_i + T_3 \right)^2 + \sum_{i=n+1}^{N-1} (z_i - \theta_i)^2 \right] \right\}
\]

(10.4)

where \(z_{ns} \in R\) and \(T_3 = \sum_{i=1}^{n} z_i + \theta_N\). Notice that the right side of (10.4) is the kernel of a multivariate normal distribution, \(MVN(\bar{\mu}_0, \Sigma)\), where

\[
\bar{\mu}_0 = \begin{pmatrix}
\mu_{01} \\
\mu_{02} \\
\vdots \\
\mu_{0(N-n-1)}
\end{pmatrix} = \begin{pmatrix}
\frac{1}{N-n}([N-n-1]\theta_{n+1} - \theta_{n+2} - \cdots - \theta_{N-1} - T_3) \\
\frac{1}{N-n}([N-n-1]\theta_{n+2} - \theta_{n+3} - \cdots - \theta_{N-1} - T_3) \\
\vdots \\
\frac{1}{N-n}([N-n-1]\theta_{N-1} - \theta_{N+1} - \cdots - \theta_{N-2} - T_3)
\end{pmatrix} \in (N-n-1) \times 1
\]

(10.5)
and
\[
\Sigma^{-1} = \frac{1}{\sigma_e^2} \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{pmatrix}_{(N-n-1) \times (N-n-1)}
\] (10.6)

From (10.4), (10.5), and (10.6) we see that \( \tilde{z}_{ns} \) given \((\tilde{y}_N, \tilde{z}_s, \tilde{\psi}, \tilde{\eta}, \tilde{I}) \) has a multivariate normal distribution with mean \( \bar{\mu} \) and variance \( \tilde{\Sigma} \) restricted to region \( R \), i.e.,

\[
\tilde{z}_{ns} | \tilde{y}_N, \tilde{z}_s, \tilde{\psi}, \tilde{\eta}, \tilde{I} \sim MVN(\bar{\mu}, \tilde{\Sigma})
\]

where \( \tilde{z}_{ns} \in R \). One can draw samples from \( z_{ns} \) given everything else by generating random samples from the multivariate normal density, \( MVN(\bar{\mu}, \tilde{\Sigma}) \), and only keeping the samples which fall in \( R \). This procedure performs poorly because there are too many rejections. Thus, we proceed by sampling the members of \( \tilde{z}_{ns} \) one at a time, i.e., for \( z_i, i = n+1, n+2, \ldots, N-1 \) given \((z_{n+1}, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{N-1}, \tilde{z}_s, \tilde{y}_N, \tilde{\psi}, \tilde{\eta}, \tilde{I}) \), we have

\[
P(z_i | z_{n+1}, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{N-1}, \tilde{z}_s, \tilde{y}_N, \tilde{\psi}, \tilde{\eta}, \tilde{I}) \propto \\
\exp \left\{ -\frac{1}{2} \left( z_i + \frac{1}{2} \left( T_3 + \sum_{j=n+1}^{N-1} z_j - z_i - \theta_i \right) \right)^2 \right\}
\]

where \( T_{\text{maxi}} \leq z_i \leq T_{\text{mini}} \) and

\[
T_{\text{maxi}} = \max \left( -z_N, \frac{t}{n} \left( n - 1 - \sum_{i=1}^{n} \pi_i - \frac{n(N-n-1)}{N} \right) - \sum_{j=n+1}^{N-1} z_j + z_i \right),
\]

\[
T_{\text{mini}} = \min \left( \frac{N}{n} - 1, z_N, \frac{t}{n} \left( n - \sum_{i=1}^{n} \pi_i - \frac{n(N-n-1)}{N} \right) - \sum_{j=n+1}^{N-1} z_j + z_i \right).
\]

Then given \((z_{n+1}, \ldots, z_{i-1}, z_{i+1}, \ldots, z_{N-1}, \tilde{z}_s, \tilde{y}_N, \tilde{\psi}, \tilde{\eta}, \tilde{I}) \), \( z_i, i = n+1, n+2, \ldots, N-1 \), has a truncated normal distribution, i.e.,

\[
\tilde{z}_i | \tilde{z}_{n+1}, \ldots, \tilde{z}_{i-1}, \tilde{z}_{i+1}, \ldots, \tilde{z}_{N-1}, \tilde{z}_s, \tilde{y}_N, \tilde{\psi}, \tilde{\eta}, \tilde{I} \sim N \left( -\frac{1}{2} \left( T_3 + \sum_{j=n+1}^{N-1} z_j - z_i - \theta_i \right), \frac{1}{2} \sigma_e^2 \right)
\]
where \( T_{\text{maxi}} \leq z_i \leq T_{\text{mini}} \).

The conditional distribution of \( \tilde{\psi} \) given \((\tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})\) is

\[
P(\tilde{\psi}|\tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I}) \propto \sigma^{-N-2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2 \right\}.
\]

Then the conditional distribution of \( \mu \) given \((\sigma^2, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})\) is

\[
P(\mu|\sigma^2, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I}) \propto \sigma^{-N-2} \exp \left\{ -\frac{N}{2\sigma^2} \left[ \mu - \frac{1}{N} \sum_{i=1}^{N} y_i \right]^2 \right\}.
\]

Thus, given \((\sigma^2, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})\), \( \mu \) has a normal distribution with mean \( \frac{1}{N} \sum_{i=1}^{N} y_i \) and variance \( \frac{1}{N} \sigma^2 \), i.e.,

\[
\mu|\sigma^2, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I} \sim N \left( \frac{1}{N} \sum_{i=1}^{N} y_i, \frac{1}{N} \sigma^2 \right).
\]

Similarly, we have the conditional distribution of \( \sigma^2 \) given \((\mu, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})\)

\[
P(\sigma^2|\mu, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I}) \propto \sigma^{-N-2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \mu)^2 \right\}.
\]

Then, given \((\mu, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})\), \( \sigma^2 \) has an Inverse-Gamma distribution with shape parameter \( N/2 \) and scale parameter \( \frac{1}{2} \sum_{i=1}^{N} (y_i - \mu)^2 \), i.e.,

\[
\sigma^2|\mu, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I} \sim \text{Inverse-Gamma} \left( \frac{N}{2}, \frac{1}{2} \sum_{i=1}^{N} (y_i - \mu)^2 \right).
\]

The conditional distribution of \( \tilde{\eta} \) given \((\tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I})\) is

\[
P(\tilde{\eta}|\tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I}) \propto \sigma^{-N-2} \exp \left\{ -\frac{1}{2\sigma^2} \left[ N\tilde{z}_N^2 - 2\sum_{i=1}^{N} (\beta_0 + \beta_1 y_i) \right. \right. \\
\left. \left. + \left( \sum_{i=1}^{N} z_i + \beta_0 + \beta_1 y_N \right)^2 + \sum_{i=1}^{N} (z_i - \beta_0 - \beta_1 y_i)^2 \right] \right\}.
\]
Then the conditional distribution of $\sigma^2_e$ given $(\beta_0, \beta_1, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I})$ is

$$P(\sigma^2_e | \beta_0, \beta_1, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I}) \propto \sigma^{-N-2}_e \exp \left\{-\frac{1}{2\sigma^2_e} T_4 \right\}$$

where

$$T_4 = N\tilde{z}_N^2 - 2\tilde{z}_N \sum_{i=1}^{N} (\beta_0 + \beta_1 y_i) + \left( \sum_{i=1}^{N-1} z_i + \beta_0 + \beta_1 y_N \right)^2$$

$$+ \sum_{i=1}^{N-1} (z_i - \beta_0 - \beta_1 y_i)^2.$$

So, given $(\beta_0, \beta_1, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I})$, $\sigma^2_e$ has an Inverse-Gamma distribution with shape parameter $N/2$ and scale parameter $T_4/2$, i.e.,

$$\sigma^2_e | \beta_0, \beta_1, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I} \sim \text{Inverse - Gamme} \left( \frac{N}{2}, \frac{T_4}{2} \right).$$

The conditional distribution of $\beta_0$ given $(\sigma^2_e, \beta_1, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I})$ is

$$P(\beta_0 | \sigma^2_e, \beta_1, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I}) \propto \exp \left\{-\frac{1}{2\sigma^2_e} \left[ N\beta_0^2 + 2 \left( \beta_1 \sum_{i=1}^{N} y_i - t \right) \beta_0 \right] \right\}.$$

This implies that, given $(\sigma^2_e, \beta_1, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I})$, $\beta_0$ has the following normal distribution

$$\beta_0 | \sigma^2_e, \beta_1, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I} \sim N \left( -\frac{1}{N} \left( \beta_1 \sum_{i=1}^{N} y_i - t \right), \frac{1}{\sigma^2_e} \right).$$

Similarly, the conditional distribution of $\beta_1$ given $(\sigma^2_e, \beta_0, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I})$ is

$$P(\beta_1 | \sigma^2_e, \beta_0, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I}) \propto \exp \left\{-\frac{1}{2\sigma^2_e} \left[ \beta_1^2 \sum_{i=1}^{N} y_i^2 + 2T_5 \beta_1 \right] \right\}$$

where

$$T_5 = \sum_{i=1}^{N-1} y_i (\beta_0 - z_i) + y_N \left( \beta_0 + \sum_{i=1}^{N-1} z_i \right) - \frac{t}{N} \sum_{i=1}^{N} y_i.$$
Then, given \((\sigma_e^2, \beta_0, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I})\), \(\beta_1\) has the following normal distribution

\[
\beta_1 | \sigma_e^2, \beta_0, \tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I} \sim N \left( -\frac{T_5}{\sum_{i=1}^{N} y_i^2}, \frac{1}{\sum_{i=1}^{N} y_i^2} \sigma_e^2 \right).
\]

After having all the conditional distributions we use the Gibbs sampler to draw samples from \(\pi_\alpha(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \tilde{y}_s, \tilde{z}_s, \tilde{I})\).

### 10.5 Evaluating \(C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta})\)

The weights in the SIR algorithm are related to the normalization constant \(C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta})\) as shown in Section 10.1. Given \((\tilde{y}_N, \tilde{\psi}, \tilde{\eta})\), we need to compute \(C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta})\).

First, notice that \(C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta})\) can be further rewritten as

\[
C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta}) = 
\int_{R^0} N(\sqrt{2\pi}\sigma_e)^{-N} \exp \left\{ -\frac{1}{2\sigma_e^2} \left[ \frac{t^2}{N} - \frac{2t}{N} \sum_{i=1}^{N} \theta_i + \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right] \right\} dz_1 dz_2 \cdots dz_{N-1}
\]

\[
= N(\sqrt{2\pi}\sigma_e)^{-N} \exp \left\{ -\frac{1}{2\sigma_e^2} \left[ \frac{t^2}{N} - \frac{2t}{N} \sum_{i=1}^{N} \theta_i \right] \right\}
\times \int_{R^0} \exp \left\{ -\frac{1}{2\sigma_e^2} \left[ \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right] \right\} dz_1 dz_2 \cdots dz_{N-1}.
\]

Let \(g(\tilde{z}_N, \tilde{y}_N, \tilde{\psi}, \tilde{\eta})\) denote

\[
\exp \left\{ -\frac{1}{2\sigma_e^2} \left[ \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right] \right\}.
\]
We also define $G(\tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta})$ as

$$G(\tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta}) = \int_{R_0} \exp \left\{ -\frac{1}{2\sigma^2} \left[ (\sum_{i=1}^{N-1} z_i + \theta_N)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right] \right\} dz_1 dz_2 \cdots dz_{N-1}$$

$$= \int_{R_0} g(\tilde{z}_N, \tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta}) dz_1 dz_2 \cdots dz_{N-1}.$$ 

Second, we separate region $R_0$ into two parts, $R_1$ and $R_2$:

$$R_1 = \left\{ (z_1, z_2, \cdots, z_{N-1}) \mid -\frac{t}{N} \leq z_i \leq \frac{t}{n} - \frac{t}{N} \right\},$$

$$R_2 = \left\{ (z_1, z_2, \cdots, z_{N-1}) \mid \frac{t}{N} - \frac{t}{n} \leq \sum_{i=1}^{N-1} z_i \leq \frac{t}{N} \right\}.$$

Then we have

$$G(\tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta}) = \int_{R_0} g(\tilde{z}_N, \tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta}) dz_1 dz_2 \cdots dz_{N-1}$$

$$= \int_{R_1} I_{R_2} g(\tilde{z}_N, \tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta}) dz_1 dz_2 \cdots dz_{N-1}$$

$$= \int_{R_1} (I_{R_2} C_0) \frac{1}{C_0} g(\tilde{z}_N, \tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta}) dz_1 dz_2 \cdots dz_{N-1},$$

where $C_0$ is defined as

$$C_0 = \int_{R_1} g(\tilde{z}_N, \tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta}) dz_1 dz_2 \cdots dz_{N-1}$$

$$= \int_{R_1} \exp \left\{ -\frac{1}{2\sigma^2} \left[ (\sum_{i=1}^{N-1} z_i + \theta_N)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right] \right\} dz_1 dz_2 \cdots dz_{N-1}.$$ 

From the above definition we see that $\frac{1}{C_0} g(\tilde{z}_N, \tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta})$ is a multivariate normal density restricted to $R_1$ and $C_0$ is the corresponding normalization constant. Now consider the following integral,

$$\int_{R_1} I_{R_2} \frac{1}{C_0} g(\tilde{z}_N, \tilde{\gamma}_N, \tilde{\psi}, \tilde{\eta}) dz_1 dz_2 \cdots dz_{N-1}.$$
One can compute this integral by drawing samples from the multivariate normal density $\frac{1}{(2\pi)^{N/2}} g(\tilde{z}_N, \tilde{y}_N, \tilde{\psi}, \tilde{\eta})$ restricted to $R_1$ and counting how many samples fall in $R_2$. It is our experience that this proportion usually is close to 1. Notice that $g(\tilde{z}_N, \tilde{y}_N, \tilde{\psi}, \tilde{\eta})$ is the kernel of the multivariate normal distribution with mean $\tilde{\mu}'$ and variance $\tilde{\Sigma}'$ where

$$\tilde{\mu}' = \begin{pmatrix} \mu'_1 \\ \mu'_2 \\ \vdots \\ \mu'_{N-1} \end{pmatrix} = \begin{pmatrix} \frac{1}{N} [(N-1) \theta_1 - \theta_2 - \cdots - \theta_{N-1} - \theta_N] \\ \frac{1}{N} [(N-1) \theta_2 - \theta_1 - \cdots - \theta_{N-1} - \theta_N] \\ \vdots \\ \frac{1}{N} [(N-1) \theta_{N-1} - \theta_1 - \cdots - \theta_{N-2} - \theta_N] \end{pmatrix}_{(N-1) \times 1}$$

and

$$\tilde{\Sigma}'^{-1} = \frac{1}{\sigma^2_e} \begin{pmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{pmatrix}_{(N-1) \times (N-1)}$$

We still need to evaluate $C_0$, the normalization constant for $g(\tilde{z}_N, \tilde{y}_N, \tilde{\psi}, \tilde{\eta})$. $C_0$ can be rewritten as

$$C_0 = \int_{R_1} \exp \left\{ -\frac{1}{2\sigma^2_e} \left[ \frac{N-1}{2} \left( \sum_{i=1}^{N-1} \theta_i \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right] \right\} dz_1 dz_2 \cdots dz_{N-1}$$

$$= \int_{R_1} \exp \left\{ -\frac{1}{2\sigma^2_e} \left[ \theta_N^2 + \sum_{i=1}^{N-1} \theta_i^2 - \left( \sum_{i=1}^{N-1} \mu'_i \right)^2 - \sum_{i=1}^{N-1} (\mu'_i)^2 \right] \right\} (2\pi)^{N-1} \sqrt{|\Sigma'|}$$

$$\times \frac{1}{(2\pi)^{N/2} \sqrt{|\Sigma'|}} \exp \left\{ -\frac{1}{2} \left[ (\tilde{Z}_{N-1} - \tilde{\mu}') \Sigma'^{-1} (\tilde{Z}_{N-1} - \tilde{\mu}') \right] \right\} dz_1 dz_2 \cdots dz_{N-1}$$

$$= \exp \left\{ -\frac{1}{2\sigma^2_e} \left[ \theta_N^2 + \sum_{i=1}^{N-1} \theta_i^2 - \left( \sum_{i=1}^{N-1} \mu'_i \right)^2 - \sum_{i=1}^{N-1} (\mu'_i)^2 \right] \right\} (2\pi)^{N-1} \sqrt{|\Sigma'|}$$

$$\times \int_{R_1} \frac{1}{(2\pi)^{N/2} \sqrt{|\Sigma'|}} \exp \left\{ -\frac{1}{2} \left[ (\tilde{Z}_{N-1} - \tilde{\mu}') \Sigma'^{-1} (\tilde{Z}_{N-1} - \tilde{\mu}') \right] \right\} dz_1 dz_2 \cdots dz_{N-1}.$$
Notice that
\[
\int_{R_1} \frac{1}{\sqrt{(2\pi)^{-1} \sqrt{|\Sigma|}}} \exp \left\{ -\frac{1}{2} \left[ (\tilde{Z}_{N-1} - \tilde{\mu}')\Sigma'^{-1}(\tilde{Z}_{N-1} - \tilde{\mu}') \right] \right\} dz_1dz_2\cdots dz_{N-1}
\]

is the probability of a multivariate normal distribution over region \( R_1 \). This is a standard problem and one can use the algorithm proposed by Genz (1992)[2] to compute this probability. By putting everything together, we see how \( C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta}) \) is evaluated, i.e.,
\[
C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta}) = N(\sqrt{2\pi} \sigma) \exp \left\{ -\frac{1}{2} \frac{1}{\sigma^2} \left[ \sum_{i=1}^{N-1} z_i + \theta_N \right]^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right\} d\tilde{z}_1d\tilde{z}_2\cdots d\tilde{z}_{N-1}
\]
\[
\times \int_{R_0} \exp \left\{ -\frac{1}{2\sigma^2} \left[ \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right] \right\} d\tilde{z}_1d\tilde{z}_2\cdots d\tilde{z}_{N-1}
\]
\[
= N(\sqrt{2\pi} \sigma)^{-N} \exp \left\{ -\frac{1}{2\sigma^2} \left[ \frac{t^2}{N} - \frac{2t}{N} \sum_{i=1}^{N} \theta_i \right] \right\}
\]
\[
\times C_0 \int_{R_1} \frac{1}{C_0} g(\tilde{z}_N, \tilde{\psi}, \tilde{\eta})d\tilde{z}_1d\tilde{z}_2\cdots d\tilde{z}_{N-1}
\]
\[
= N(\sqrt{2\pi} \sigma)^{-N} \exp \left\{ -\frac{1}{2\sigma^2} \left[ \frac{t^2}{N} - \frac{2t}{N} \sum_{i=1}^{N} \theta_i \right] \right\}
\]
\[
\times \left[ \exp \left\{ -\frac{1}{2\sigma^2} \left[ \theta_N^2 + \sum_{i=1}^{N-1} \theta_i^2 - \left( \sum_{i=1}^{N-1} \mu_i \right)^2 - \sum_{i=1}^{N-1} (\mu_i)^2 \right] \right\} (2\pi)^{-\frac{N+1}{2}} \sqrt{|\Sigma|} \right]
\]
\[
\times \int_{R_1} \frac{1}{(2\pi)^{-\frac{N+1}{2}} \sqrt{|\Sigma|}} \exp \left\{ -\frac{1}{2} \left[ (\tilde{Z}_{N-1} - \tilde{\mu}')\Sigma'^{-1}(\tilde{Z}_{N-1} - \tilde{\mu}') \right] \right\} d\tilde{z}_1d\tilde{z}_2\cdots d\tilde{z}_{N-1}
\]
\[
\times \int_{R_1} \frac{1}{C_0} g(\tilde{z}_N, \tilde{\psi}, \tilde{\eta})d\tilde{z}_1d\tilde{z}_2\cdots d\tilde{z}_{N-1}.
\]

In the SIR algorithm the computational burden is to compute this normalization constant, \( C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta}) \). This has to be done at each Gibbs sampler step. Specifically, at each Gibbs sampler step, we need to draw samples from \( \frac{1}{C_0} g(\tilde{z}_N, \tilde{\psi}, \tilde{\eta}) \)
using another Gibbs sampler to compute
\[ \int_{R_1} I_{R_2} \frac{1}{C_0} g(\tilde{z}_N, \tilde{y}_N, \tilde{\psi}, \tilde{\eta}) dz_1 dz_2 \cdots dz_{N-1}. \]

We also need to calculate the \((N-1)\) dimensional multivariate normal probability
\[ \int_{R_1} \frac{1}{(2\pi)^{N-1} \sqrt{|\Sigma|}} \exp \left\{ -\frac{1}{2} \left[ (\tilde{Z}_{N-1} - \tilde{\mu}') \Sigma^{-1} (\tilde{Z}_{N-1} - \tilde{\mu}') \right] \right\} dz_1 dz_2 \cdots dz_{N-1} \]
at each Gibbs sampler step. It is our experience that \(C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta})\) usually is a very small number. For this purpose, we have developed a parallel computing algorithm for SIR in a high-performance cluster environment.

References


Chapter 11

Log-Normal Model with Known \( t \)

Now we assume that the population distribution is log-normal with mean \( \mu \) and variance \( \sigma^2 \) in the log scale, i.e.,

\[
Y_i | \mu, \sigma^2 \overset{iid}{\sim} \text{Log-normal}(\mu, \sigma^2)
\]

where \( i = 1, 2, \cdots N \). As in Chapter 10 we assume that the total of measures of size for the finite population, \( \sum_{j=1}^{N} v_j = t \), is known in order to avoid the non-identifiability problem. Also, we use the same notation as in Chapter 10.

11.1 Posterior distribution

The posterior distribution for the log-normal model can be written as

\[
P(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta} | \tilde{y}_{s}, \tilde{z}_{s}, \tilde{I}) \propto P(\tilde{y}, \tilde{z}, \tilde{\psi}, \tilde{\eta}, \tilde{I})
\]

\[
= P(I | \tilde{y}, \tilde{z}, \tilde{\psi}, \tilde{\eta})P(\tilde{z} | \tilde{y}, \tilde{\psi}, \tilde{\eta})P(\tilde{y} | \tilde{\psi}, \tilde{\eta})P(\tilde{\psi}, \tilde{\eta})
\]
where \( \tilde{z}_{ns} \in R, \tilde{y}_{ns} > 0 \) and

\[
P(\tilde{I}|\tilde{y}, \tilde{z}, \tilde{\psi}, \tilde{\eta}) = \prod_{i \in S} \pi_i \prod_{i \notin S} (1 - \pi_i)
\]

\[
= \left( \frac{n}{Nz_N} \right)^n \prod_{i=1}^n (z_i + z_N) \prod_{i=n+1}^{N-1} \left[ 1 - \frac{n}{Nz_N} (z_i + z_N) \right]
\]

\[
\left[ 1 - \left( \frac{n}{Nz_N} \right) \left( z_N - \sum_{i=1}^{N-1} z_i \right) \right]
\]

\[
P(\tilde{z}|\tilde{y}, \tilde{\psi}, \tilde{\eta}) = \frac{1}{C(\tilde{y}, \tilde{\psi}, \tilde{\eta})} N(\sqrt{2\pi} \sigma_e)^{-N} \exp \left\{ -\frac{1}{2\sigma_e^2} \left[ \frac{t^2}{N} - 2 \frac{t}{N} \sum_{i=1}^{N} \theta_i \right. \right.
\]

\[
+ \left. \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right\}
\]

\[
P(\tilde{y}|\tilde{\psi}, \tilde{\eta}) = (\sqrt{2\pi} \sigma)^{-N} \prod_{i=1}^{N} \frac{1}{y_i} \exp \left\{ -\frac{1}{2\sigma^2} (\log y_i - \mu)^2 \right\}
\]

and \( P(\tilde{\psi}, \tilde{\eta}) \) is the prior distribution. For \((\tilde{\psi}, \tilde{\eta})\), we use non-informative prior distribution, i.e., \( P(\tilde{\psi}, \tilde{\eta}) \propto \sigma^{-2} \sigma_e^{-2} \).

The resulting posterior distribution is

\[
P(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta}|\tilde{y}_{s}, \tilde{z}_{s}, \tilde{I}) \propto (\sigma_e \sigma)^{-N-2} \frac{1}{C(\tilde{y}, \tilde{\psi}, \tilde{\eta})} \prod_{i=n+1}^{N-1} \left[ 1 - \frac{n}{Nz_N} (z_i + z_N) \right]
\]

\[
\left[ 1 - \left( \frac{n}{Nz_N} \right) \left( z_N - \sum_{i=1}^{N-1} z_i \right) \right] \prod_{i=1}^{N-1} \left[ \frac{1}{y_i} \exp \left\{ -\frac{1}{2\sigma^2} (\log y_i - \mu)^2 \right\} \right]
\]

\[
\exp \left\{ -\frac{1}{2\sigma_e^2} \left[ \frac{t^2}{N} - 2 \frac{t}{N} \sum_{i=1}^{N} \theta_i + \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right\} \right\}
\]

where \( \tilde{z}_{ns} \in R \) and \( \tilde{y}_{ns} > 0 \).
11.2 Sampling importance resampling (SIR)

We proceed in the same way as in Section 10.3. However, \( \pi_a(\tilde{y}_{ns}, \tilde{z}_{ns}, \psi, \tilde{\eta}|\tilde{y}_s, \tilde{z}_s, \tilde{I}) \) is different from its counterpart in the normal model. For the log-normal model, \( \pi_a(\tilde{y}_{ns}, \tilde{z}_{ns}, \psi, \tilde{\eta}|\tilde{y}_s, \tilde{z}_s, \tilde{I}) \) is

\[
\pi_a(\tilde{y}_{ns}, \tilde{z}_{ns}, \psi, \tilde{\eta}|\tilde{y}_s, \tilde{z}_s, \tilde{I}) = (\sigma_c \sigma)^{-N-2} \prod_{i=1}^{N} \left\{ \exp \left( -\frac{1}{2\sigma^2} (\log y_i - \mu)^2 \right) \right\} \\
\exp \left\{ -\frac{1}{2\sigma^2} \left[ \frac{t^2}{N} - 2 \frac{t}{N} \sum_{i=1}^{N} \theta_i + \left( \sum_{i=1}^{N-1} z_i + \theta_N \right)^2 + \sum_{i=1}^{N-1} (z_i - \theta_i)^2 \right] \right\}
\]

where \( \tilde{z}_{ns} \in R \) and \( \tilde{y}_{ns} > 0 \). Notice that the normalization constant \( C(\tilde{y}_N, \psi, \tilde{\eta}) \) is the same for both the normal and log-normal models.

11.3 Gibbs sampler

In order to draw samples from \( \pi_a(\tilde{y}_{ns}, \tilde{z}_{ns}, \psi, \tilde{\eta}|\tilde{y}_s, \tilde{z}_s, \tilde{I}) \) we need to find the conditional distributions of \( \tilde{y}_{ns}, \tilde{z}_{ns}, \psi, \) and \( \tilde{\eta} \) respectively given everything else.

The conditional distribution of \( \tilde{y}_{ns} \) given \( (\tilde{y}_s, \tilde{z}_s, \psi, \tilde{\eta}, \tilde{I}) \) is

\[
P(\tilde{y}_{ns} | \tilde{y}_s, \tilde{z}_N, \psi, \tilde{\eta}, \tilde{I}) \propto \exp \left\{ -\frac{1}{2\sigma^2} \left[ \frac{2t}{N} \beta_1 \sum_{i=n+1}^{N} y_i + \sum_{i=n+1}^{N-1} (z_i - \beta_0 - \beta_1 y_i)^2 \right] \right\} \prod_{i=n+1}^{N} y_i^{-1} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=n+1}^{N} (\log y_i - \mu)^2 \right\}
\]

where \( \tilde{y}_{ns} > 0 \). From the above conditional distribution we see that, given everything else, the conditional distribution of \( y_i, i = n + 1, n + 2, \cdots, N - 1 \) is

\[
P(y_i|\text{others}) \propto \exp \left\{ -\frac{1}{2\sigma^2} \left[ \frac{2t}{N} \beta_1 y_i + (z_i - \beta_0 - \beta_1 y_i)^2 \right] \right\} y_i^{-1}
\times \exp \left\{ -\frac{1}{2\sigma^2} (\log y_i - \mu)^2 \right\}
\]
where $y_i > 0$. We use rejection sampling to draw samples from the conditional distribution of $y_i$ given everything else. Define a new distribution of $y_i$ given everything else, as

$$P_1(y_i|\text{others}) \propto \exp \left\{ -\frac{1}{2\sigma^2} \left( \frac{\beta_1 y_i}{N} + (z_i - \beta_0 - \beta_1 y_i)^2 \right) \right\}$$

where $y_i > 0, i = n + 1, n + 2, \cdots, N - 1$. Then we have

$$\frac{P(y_i|\text{others})}{P_1(y_i|\text{others})} = \frac{C_1}{C_2} y_i^{-1} \exp \left\{ -\frac{1}{2\sigma^2} (\log y_i - \mu)^2 \right\}$$

$$\leq \frac{C_1}{C_2} \exp \left\{ \frac{1}{2} \sigma^2 - \mu \right\} = M$$

where $C_1$ and $C_2$ are the normalization constants for the two conditional distributions of $y_i$, $P(y_i|\text{others})$ and $P_1(y_i|\text{others})$ respectively and $M$ is a constant given $(y_{n+1}, \cdots, y_i-1, y_i+1, \cdots, y_{N-1}, y_N, \bar{y}, \bar{z}, \bar{\psi}, \bar{\eta}, \bar{I})$. For rejection sampling, first we draw a sample from $P_1(y_i|\text{others})$. Second, we calculate the acceptance probability

$$\frac{P(y_i|\text{others})}{M \times P_1(y_i|\text{others})} = \exp \left\{ \mu - \frac{1}{2} \sigma^2 \right\} \exp \left\{ -\frac{1}{2\sigma^2} (\log y_i - \mu)^2 \right\} y_i^{-1}. \quad (11.1)$$

Then the sample from $P_1(y_i|\text{others})$ is selected as a sample from $P(y_i|\text{others})$ with the calculated acceptance probability. In terms of $P_1(y_i|\text{others})$, we have

$$y_i|\text{others} \sim N \left( \frac{1}{\beta_1} \left( z_i + \frac{t}{N} - \beta_0 \right), \left( \frac{\sigma^2}{\beta_1} \right)^2 \right) \quad (11.2)$$

where $y_i > 0$ and $i = n + 1, n + 2, \cdots, N - 1$. For $y_N$, the conditional distribution of $y_N$ given everything else is

$$P(y_N|\text{others}) \propto y_N^{-1} \exp \left\{ -\frac{1}{2\sigma^2} (\log y_N - \mu)^2 \right\}$$

$$\times \exp \left\{ -\frac{1}{2\sigma^2} \left[ -\frac{2t}{N} \beta_1 y_N + \left( \sum_{i=1}^{N-1} z_i + \beta_0 + \beta_1 y_N \right)^2 \right] \right\}$$
where \( y_N > 0 \). Accordingly, \( P_1(y_N|\text{others}) \) is defined as

\[
P_1(y_N|\text{others}) \propto \exp \left\{ -\frac{1}{2\sigma^2} \left[ \frac{2t}{N} \beta_1 y_N + \left( \sum_{i=1}^{N-1} z_i + \beta_0 + \beta_1 y_N \right)^2 \right] \right\}
\]

(11.3)

where \( y_N > 0 \). Then we have

\[
\frac{P(y_N|\text{others})}{P_1(y_N|\text{others})} = \frac{C_3}{C_4} y_N^{-1} \exp \left\{ -\frac{1}{2\sigma^2} (\log y_N - \mu)^2 \right\}
\]

(11.4)

\[
\leq \frac{C_3}{C_4} \exp \left\{ \frac{1}{2} \sigma^2 - \mu \right\}
\]

(11.5)

\[
= M_0
\]

(11.6)

where \( C_3 \) and \( C_4 \) are the normalization constants for the two conditional distributions of \( y_N, P(y_N|\text{others}) \) and \( P_1(y_N|\text{others}) \) respectively and \( M_0 \) is a constant given \((y_{n+1}, \ldots, y_i, \ldots, y_{N-1}, \tilde{y}_s, \tilde{z}_N, \tilde{\psi}, \tilde{\eta}, \tilde{I})\). The acceptance probability for the rejection sampling is

\[
\frac{P(y_N|\text{others})}{M \times P(y_N|\text{others})} = \exp \left\{ \mu - \frac{1}{2} \sigma^2 \right\} \exp \left\{ -\frac{1}{2\sigma^2} (\log y_N - \mu)^2 \right\} y_N^{-1}.
\]

Thus a sample from \( P_1(y_N|\text{others}) \) is chosen as a sample from \( P(y_N|\text{others}) \) with the acceptance probability in (101). For \( P_1(y_N|\text{others}) \), we have

\[
y_N|\text{others} \sim N \left( \frac{1}{\beta_1} \left( \frac{t}{N} - \beta_0 - \sum_{i=1}^{N-1} z_i \right) , \left( \frac{\sigma_0}{\beta_1} \right)^2 \right)
\]

where \( y_N > 0 \).

The conditional distribution of \( \tilde{z}_{n} \) given \((\tilde{y}_N, \tilde{z}_s, \tilde{\psi}, \tilde{\eta}, \tilde{I})\) is exactly the same as for the normal model with known \( t \). See Section 10.4 and the related discussion.

The conditional distribution of \( \tilde{\psi} \) given \((\tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})\) is

\[
P(\tilde{\psi}|\tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I}) \propto \sigma^{-N-2} \prod_{i=1}^{N} \left[ \frac{1}{\gamma_i} \exp \left\{ -\frac{1}{2\sigma^2} (\log y_i - \mu)^2 \right\} \right].
\]

(11.7)
Then the conditional distribution of $\mu$ given $(\sigma^2, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})$ is

$$P(\mu|\sigma^2, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I}) \propto \exp \left\{ -\frac{N}{2\sigma^2} \left[ \mu - \frac{1}{N} \sum_{i=1}^{N} \log y_i \right]^2 \right\}. \quad (11.8)$$

Thus, given $(\sigma^2, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})$, $\mu$ has a normal distribution with mean $\frac{1}{N} \sum_{i=1}^{N} \log y_i$ and variance $\frac{1}{N} \sigma^2$, i.e.,

$$\mu|\sigma^2, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I} \sim N \left( \frac{1}{N} \sum_{i=1}^{N} \log y_i, \frac{1}{N} \sigma^2 \right). \quad (11.9)$$

Similarly, we have the conditional distribution of $\sigma^2$ given $(\mu, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})$, i.e.,

$$P(\sigma^2|\mu, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I}) \propto \sigma^{N-2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (\log y_i - \mu)^2 \right\}. \quad (11.10)$$

Then, given $(\mu, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I})$, $\sigma^2$ has an Inverse-Gamma distribution with shape parameter $N/2$ and scale parameter $\frac{1}{2} \sum_{i=1}^{N} (\log y_i - \mu)^2$, i.e.,

$$\sigma^2|\mu, \tilde{y}_N, \tilde{z}_N, \tilde{\eta}, \tilde{I} \sim \text{Inverse-Gamma} \left( \frac{N}{2}, \frac{1}{2} \sum_{i=1}^{N} (\log y_i - \mu)^2 \right). \quad (11.11)$$

The conditional distribution of $\tilde{\eta}$ given $(\tilde{y}_N, \tilde{z}_N, \tilde{\psi}, \tilde{I})$ is exactly the same as for the normal model with known $t$. See Section 10.4 and the related discussion.

After having all the conditional distributions we use the Gibbs sampler to draw samples from $\pi_u(\tilde{y}_{ns}, \tilde{z}_{ns}, \tilde{\psi}, \tilde{\eta}|\tilde{y}_s, \tilde{z}_s, \tilde{I})$. As with the normal model, the computational burden is to evaluate the normalization constant $C(\tilde{y}_N, \tilde{\psi}, \tilde{\eta})$. This has to be done at each Gibbs sampler step.
Chapter 12

Model Inference, Discussion and Further Work

12.1 Model Inference

By using the methodology in Section 3.2 and Section 3.3, a set of samples from the posterior distribution is generated. Based on this set of samples one can make inference for the parameters of the population distribution and make predictive inference for the non-sampled units. Thus, we can make inference for any characteristic of the finite population.

Supposing that we have $M$ samples from the posterior distribution, the point estimate of a population parameter $\eta$ is the average of the $M$ values of $\eta$. The Bayesian $100(1 - \alpha)\%$ credible interval for $\eta$ is obtained by using the $(\alpha/2)M$ and $(1 - \alpha/2)M$ values of the $M$ order statistics for $\eta$. For the non-sampled units, the point estimate of $\tilde{y}_{ns}$ is taken as the average of the $M$ values of $\tilde{y}_{ns}$. For each individual non-sampled unit, the Bayesian $100(1 - \alpha)\%$ credible interval can be obtained by using the $(\alpha/2)M$ and $(1 - \alpha/2)M$ values of the $M$ order statistics for this individual non-sampled unit.

12.2 Discussion

There are two general approaches to make inference for the parameters of the population distribution when there is informative sampling; i.e., use either an ig-
norable or a non-ignorable model. For a non-ignorable model, the selection probability is related to the response variable while for an ignorable model, they are not related; see Sugden and Smith (1984) and Krieger and Pfeffermann (1992). Our proposed methodology is based on using a non-ignorable model. Referring to the non-ignorable model Pfeffermann, Krieger and Rinott (1998) state that “A different approach in wide use to deal with the effects of informative sampling is to replace the ordinary sample estimates or estimating equations by weighted analogues obtained by weighting the sample observations inversely proportional to the sample selection probabilities. The use of this approach is restricted in general to point estimation and does not permit the use of standard inference tools such as likelihood based inference or residual analysis. Probabilistic statements require large sample normality assumptions.” From this statement, it’s clear that the non-ignorable model uses weighted sample quantities or estimating equations to estimate the parameters of the population distribution.

For the estimation of a finite population quantity when there is informative sampling, the problem is more complex than inference for the parameters of the population distribution. The reason is that the informative sampling tends to make the sampled units “large” and the non-sampled units “small”. Such an adjustment could be carried out using our methodology, but it is very difficult to do using the other non-ignorable approaches that have been proposed.

Our methodology is appropriate for many establishment surveys when the $\pi_i$ are known only for the sampled units - a common situation because, typically, only these $\pi_i$ are retained. We have shown how to conduct a complete analysis of a complicated problem using survey data; i.e., where there is only limited information about the survey design and there is a selection bias. Our approach using Markov chain and Monte Carlo avoids the necessity of using asymptotic approximations which are necessary for the other approaches that have been proposed.

References


Part III

Sample Size Determination for Unordered Categorical Data
Chapter 13

Introduction

13.1 Introduction and Literature Review

Sample size determination is one of the most important practical tasks for statisticians. There has been extensive research to develop appropriate methodology, e.g., methods for sample size determination for continuous, or ordered categorical outcome data. However, sample size determination for comparative studies with unordered categorical data remains largely untouched. For example researchers may be interested in comparing the distributions of the locations of leukoplakia lesions for different smoking or chewing habits (see, e.g., Pindborg, Kiaer, Gupta, and Chawla (1967)[29]). In clinical trials, the five-year survival rate is often used to evaluate the efficacy of different treatments (see, e.g., Krstevska and Crvenkova (2006)[19]). However, the difference could be seen at other important landmark years, such as 1, 2, 3, 10, 15 (see Figures 2 and 3 in Krstevska and Crvenkova (2006)[19]). We can compare their survival distributions, which requires finding the simultaneous confidence bands for the survival functions. Alternatively, instead of comparing 5-year survival rates or finding the simultaneous confidence bands for the survival functions, we could compare their survival rates for different time periods. For example, we can compare their 1, 2, 3, 5, 10, 15-year survival rates under different treatments. Another example will be comparing the response patterns for two or more different doses in clinical trials (see, e.g., Dillman, Davis, Green, Weiss, Gottlieb, Caplan, Kopel, Preisler, McIntyre, and Schiffer (1991)[11]). In other words we would like to find the required sample
sizes needed to detect the differences between the response patterns for different doses.

The applications just described involve sample size determination for unordered categorical data, which come from the following 2-multinomial model. Specifically, if $n_{i1}, n_{i2}, \ldots, n_{ik}$ were the observed cell frequencies or data in a sample of size $n_i$ from a multinomial distribution with parameter vector $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{ik})$, where $\sum_{j=1}^{k} n_{ij} = n_i, \sum_{j=1}^{k} p_{ij} = 1$, for $i = 1, 2$ respectively, one would be able to test if there is a difference between $\mathbf{p}_1$ and $\mathbf{p}_2$. Here, we are interested in, instead, determining the minimum sample sizes $n_1$ (for population 1) and $n_2$ (for population 2) needed to detect a difference between $\mathbf{p}_1$ and $\mathbf{p}_2$ that is at least $\Delta_0$. Here $\Delta_0$ is called the precision and needs to be prespecified.

There has been a substantial literature on sample size determination, for example, the books by Desu and Raghavarao (1990)[10] and Chow, Shao, and Wang (2007)[7], an excellent review article by Adcock (1997)[2], a representative Bayesian paper by Wang and Gelfand (2002)[35], and references therein. Desu and Raghavarao (1990)[10] presented commonly-used methods for determining the sample sizes in experiments and sample surveys. Chow et al. (2007)[7] provided a comprehensive and unified presentation of statistical procedures (Bayesian and non-Bayesian) for sample size calculation needed at various phases of clinical research. Wang and Gelfand (2002)[35] proposed a simulation–based approach to sample size determination (SSD) in two situations. The first is to find the sample size needed to achieve specified performance with regard to one or more features of a model. The second one is about selecting a sample size to achieve a specified separation of two models.

The following are some references on sample size determination for one multinomial population. Tortora (1978)[33] and Thompson (1987)[32] considered sample size determination for simultaneously estimating the parameters of a multinomial distribution with a specified confidence interval width while Adcock (1993)[1] developed a Bayesian approach to select the sample size such that the parameter of interest is contained in a tolerance region with specified probability (which holds on average over all possible samples). Sison and Glaz (1995)[31] developed two procedures to construct alternative simultaneous confidence intervals for the multinomial proportions and proposed two corresponding sample size determina-
tion methods to achieve a specified coverage probability.

In this part we find the sample sizes needed to detect a difference (with specified precision $\Delta_0$) between two vectors of multinomial proportions using both frequentist and Bayesian approaches. In the frequentist approach, our starting point is the distribution under the alternative hypothesis of a $\chi^2$ test for testing the equality of two vectors of multinomial proportions, given by Meng and Chapman (1966) [1].

This part is organized as follows. In Chapter 13 we set up and review general approaches for sample size determination. In Chapter 14 we study chi-squared tests for contingency tables, derive several frequentist procedures for sample size determination and investigate their asymptotic properties. In Chapter 14, we also develop bootstrap and other improvements to the basic frequentist approach. In Chapter 15 we use the concepts of posterior predictive p-values (Meng (1994) [22], Gelman, Meng, and Stern (1996) [15], and Gelman, Carlin, Stern, and Rubin (2004) [14]) and calibrated posterior predictive p-values (first proposed by Hjort, Dahl, and Steinbakk (2006) [17]) to develop a Bayesian approach to sample size determination. The simulation studies and real data application and their results are described and summarized in Chapter 16 and 17. Our discussion and conclusions are in Chapter 17.

### 13.2 General Approaches for Sample Size Determination

All sample size determination methods require that the sample sizes satisfy a specified target level of accuracy or maximize a specified objective function. In general, the approaches for sample size determination fall into two broad groups, i.e., Bayesian and frequentist.

In the frequentist approach, we first specify a null and an alternative hypothesis for the parameter of interest, $\mu$, say:

$$H_0 : \mu \in \omega, \quad H_1 : \mu \in \Omega - \omega.$$  

Second, we specify a desired test size $\alpha$ and power $(1 - \beta)$. Let $c$ be the critical value of a test such that the probability of making a type I error is close to but not
larger than \( \alpha \), i.e.,

\[
Pr \{ \text{reject } H_0 | \mu \} \leq \alpha, \quad \mu \in \omega.
\]

Third, based on this critical value, we choose the smallest sample size \( n \) to satisfy

\[
Pr \{ \text{reject } H_0 | \mu \} \geq 1 - \beta, \quad \text{for some } \mu \in \Omega - \omega.
\]

This method usually requires the existence of a pivotal quantity that is related to the test statistic and the computation is specific to the problem under consideration. Frequentist methods have been used widely in practice and can also serve as a starting point for a Bayesian approach.

In the Bayesian approach, we first specify a prior distribution for the model parameter(s). Second, we calculate the posterior distribution of the parameters based on the prior distribution and data. Third, we find a test quantity (also called a discrepancy measure), which is a function of the data and parameters. Fourth, we compute the posterior predictive \( p \)-value based on the test quantity. Then we look for the minimum sample size such that

\[
P_{H_1} \left\{ p_q(n_{\text{obs}}) \leq \alpha \right\} \geq 1 - \beta
\]

where \( p_q(n_{\text{obs}}) \) is the posterior predictive \( p \)-value based on the observed value \( n_{\text{obs}} \), “\( H_1 \)” denotes a specified set of values of \( \mu, \mu \in \Omega - \omega \), and \((\alpha, \beta)\) are choices comparable to those for the frequentist method. In Chapter 13 we replace the posterior predictive \( p \)-value, \( p_q(n_{\text{obs}}) \), with the calibrated posterior predictive \( p \)-value (Hjort et al. (2006)[17]) to overcome the possible non-uniformity of the distribution of the posterior predictive \( p \)-values under the null (hypothesis) model.

Next we provide the frequentist and Bayesian approaches for the multinomial model.

**References**

Chapter 14

Frequentist Approach

14.1 Original Frequentist Approach

Consider the hypotheses:

\[ H_0 : p_1 = p_2 = p, \quad \text{vs} \quad H_1 : p_1 \neq p_2, \]

where \( p = (p_1, \ldots, p_k) \) is some unknown parameter vector s.t. \( \sum_{i=1}^{k} p_i = 1 \). A reasonable test statistic for testing this set of hypotheses is

\[ X^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - n_i p_j)^2}{n_i p_j} \]

where \( n_{ij} \) is the cell frequency which is often arranged into a contingency table, Table 1, and

\[ p_j = \frac{t_j}{N} \]

<table>
<thead>
<tr>
<th></th>
<th>Cell 1</th>
<th>Cell 2</th>
<th>...</th>
<th>Cell k</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multinomial 1</td>
<td>( n_{11} )</td>
<td>( n_{12} )</td>
<td>...</td>
<td>( n_{1k} )</td>
<td>( n_1 )</td>
</tr>
<tr>
<td>Multinomial 2</td>
<td>( n_{21} )</td>
<td>( n_{22} )</td>
<td>...</td>
<td>( n_{2k} )</td>
<td>( n_2 )</td>
</tr>
<tr>
<td>Total</td>
<td>( t_1 )</td>
<td>( t_2 )</td>
<td>...</td>
<td>( t_k )</td>
<td>( N )</td>
</tr>
</tbody>
</table>

Table 14.1: Contingency table for the count data \( n_{ij} \)
Under the null hypothesis $H_0$, as $n \to \infty$, under reasonable regularity conditions,

$$X^2 \overset{appr}{\sim} \chi^2_{(k-1)},$$  

(14.1)

by using properties of the generalized likelihood ratio test. Therefore we may reject $H_0$ at level $\alpha$ if the observed value of $X^2 > \chi^2_{(k-1)}(\alpha)$, where $\chi^2_{(k-1)}(\alpha)$ is the upper $\alpha$ quantile of the chi-squared distribution with $(k-1)$ degrees of freedom.

When the alternative hypothesis $H_1$ is true, i.e., $p_1 \neq p_2$, the test statistic $X^2$ has approximately a noncentral chi-squared distribution with $(k-1)$ degrees of freedom and noncentrality parameter $\lambda$ as shown by Meng and Chapman (1966)[21].

We advocate rewriting $\lambda$ in a way that provides a symmetric expression. This will lead to a more accurate approximation to the sample size than the one without symmetrization. Specifically, under $H_1$, we denote $p_1 = p + \delta_1$ and $p_2 = p + \delta_2$, where $p = (p_1 + p_2)/2$ and $\delta_i = p_i - p$ for $i = 1, 2$ respectively. A general formula to calculate the noncentrality parameter $\lambda$ for a $(2 \times k)$ contingency table is:

$$\lambda = N \times q_1 \times q_2 \times \sum_{j=1}^{k} \left\{ \frac{(p_{1j} - p_{2j})^2}{p_j} \right\},$$

where $q_1 = n_1/N, q_2 = n_2/N$. In a balanced design, we have $n_1 = n_2 = n$ and then

$$\lambda = \frac{n}{2} \sum_{j=1}^{k} \left\{ \frac{(p_{1j} - p_{2j})^2}{p_j} \right\},$$  

(14.2)

where $p_j = (p_{1j} + p_{2j})/2$. To have large power, we require that

$$Pr_{H_1} \left\{ X^2 > \chi^2_{(k-1)}(\alpha) \right\} \geq 1 - \beta.$$  

Since $X^2$ has a noncentral chi-squared distribution with $(k-1)$ degrees of freedom and noncentrality parameter $\lambda$, this implies

$$\chi^2_{(k-1), \lambda}(1 - \beta) \geq \chi^2_{(k-1)}(\alpha),$$  

(14.3)
where \( \chi^2_{(k-1),\lambda}(1-\beta) \) is the upper \((1-\beta)\) quantile of the noncentral chi-squared distribution with \((k-1)\) degrees of freedom and noncentrality parameter \(\lambda\).

Given \(\alpha, \beta\) and \(k\), we can find the value of \(\lambda_0\) from a noncentral chi-squared table that makes the equality part in (14.3) hold. For example, for \(k = 5\) and \(\alpha = 0.05, \beta = 0.2\), \(\lambda_0 = 11.94\). Also, note that \(\chi^2_{(k-1),\lambda}(1-\beta)\) increases in \(\lambda\), so the minimal \(n\) required to have power \(1-\beta\) to detect the difference when \(\lambda \geq \lambda_0\) is

\[
n = \left\lceil 2\lambda_0 \left( \sum_{j=1}^{k} \left(\frac{p_{1j} - p_{2j}}{p_j}\right)^2 \right)^{-1} \right\rceil, \tag{14.4}\]

where \(\lceil x \rceil\) denotes the smallest integer that is greater than or equal to \(x\).

We notice that \(n\) in SSD formula (14.4) depends on unknown parameters \(\delta_1, \delta_2\) and \(p\). In practice, we recommend using the following two methods to deal with these unknown parameters.

1. Specify a minimum average difference between \(p_1\) and \(p_2\), \(d\), and a minimum relative difference, \(r\), that we want to detect. In other words, if the average difference, \(D\), and relative difference, \(R_j\), are defined to be

\[
D = \frac{1}{k} \sum_j |p_{1j} - p_{2j}|, \quad R_j = \left| \frac{p_{1j} - p_{2j}}{p_j} \right|, \quad j = 1, 2, \cdots, k,
\]

then the sample size needed to detect a difference with \(D \geq d, R_j \geq r\) for \(j = 1, 2, \cdots, k\) with power at least \(1-\beta\) is

\[
\tilde{n} = \left\lceil 2\lambda_0 (r \cdot kd)^{-1} \right\rceil. \tag{14.5}\]

2. Replace the \(\{\delta_{ij} : i = 1, 2; j = 1, 2, \cdots, k\}\) and \(p\) by estimates computed from historical, pilot, or approximate data. Then

\[
\hat{n} = \left\lceil 2\lambda_0 \left( \sum_{j=1}^{k} \frac{\hat{\lambda}_j^2}{\hat{p}_j} \right)^{-1} \right\rceil, \tag{14.6}\]

where \(\hat{\lambda}_j = \hat{\delta}_{1j} - \hat{\delta}_{2j} = \hat{p}_{1j} - \hat{p}_{2j}, \hat{p}_j = \frac{1}{2}(\hat{p}_{1j} + \hat{p}_{2j})\).
In Method 2, since the $\hat{\Delta}_j$ and $\hat{p}$ in (14.6) are estimates, we may develop a confidence interval for $n$ in (14.4). One situation where (14.6) is appropriate is when an investigator observes $\hat{\Delta}_j$ from a pilot sample or prior study and thinks that if $\Delta_j = \delta_{1j} - \delta_{2j}$ was, in fact, no less than $\hat{\Delta}_j$, $j = 1, \cdots, J$, he/she would like to be able to reject $H_0 : p_1 = p_2 = p$ with probability $(1 - \beta)$ when $H_1$ is true. Then, the total sample size needed to achieve this objective is given by (14.6).

In a balanced design, suppose the sample size from a pilot sample is $m$ for each multinomial distribution. Then

$$E(\hat{p}_j) = p_j, \quad \text{Var}(\hat{p}_j) = \frac{1}{4m} \left( p_{1j}(1 - p_{1j}) + p_{2j}(1 - p_{2j}) \right).$$

For the $\hat{\Delta}_j$, we also have

$$E(\hat{\Delta}_j) = \Delta_j = p_{1j} - p_{2j},$$

$$\text{Var}(\hat{\Delta}_j) = \frac{1}{m} \left( p_{1j}(1 - p_{1j}) + p_{2j}(1 - p_{2j}) \right).$$

Let $t_j = (p_{1j}(1 - p_{1j}) + p_{2j}(1 - p_{2j})).$ For large $m$, by the Central Limit Theorem, it’s easy to see that

$$\hat{p}_j \xrightarrow{d} p_j + Z_{j1} \sqrt{\frac{1}{m} t_j} + o_p \left( \frac{1}{m} \right)$$

and

$$\hat{\Delta}_j \xrightarrow{d} \Delta_j + Z_{j2} \sqrt{\frac{1}{m} t_j} + o_p \left( \frac{1}{m} \right),$$

where “$d$” represents equality in distribution and $Z_{j1}$ and $Z_{j2}$ are standard normal random variables. Using this asymptotic expansion, we have the following lemma and proposition. The proofs are in the Appendix.

**Lemma 14.1.1.** Given a pilot sample of size $m$ from the two multinomial populations, if $m$ is large, then

$$\hat{\alpha}_j = 1 - \frac{A}{n} \left( \sum_{j=1}^{k} \frac{2\Delta_j \sqrt{T_j}}{p_j} Z_{j2} - \sum_{j=1}^{k} \frac{\Delta_j^2 \sqrt{T_j}}{2p_j} Z_{j1} \right) + o \left( \frac{1}{m} \right), \quad (14.7)$$
where
\[
A = 2\lambda \cdot \frac{1}{\sqrt{m}} \cdot \left( \sum_{j=1}^{k} \frac{\Delta_j^2}{p_j} \right)^{-2} = O \left( \frac{1}{\sqrt{m}} \right),
\]  
(14.8)

and \( Z_{j1} \) and \( Z_{j2} \) are standard normal random variables.

Further, as \( m \to \infty \),
\[
\text{Var}(\hat{n}) = A^2 \cdot B + o \left( \frac{1}{m^2} \right),
\]  
(14.9)

where
\[
B = \sum_{j=1}^{k} (a_j^2 + b_j^2) + \sum_{j=1}^{k} \sum_{i=1}^{k} \left\{ (b_ib_j - a_ia_j)(t_it_j)^{-1/2} (p_{1i}p_{1j} + p_{2i}p_{2j}) \right\} +
\sum_{j=1}^{k} \sum_{i=1}^{k} \left\{ a_jb_i(t_it_j)^{-1/2} (p_{1i}p_{1j} - p_{2i}p_{2j}) \right\}.
\]

**Remark 1**: From (14.7), for the estimator \( \hat{n} \),
\[
\frac{\hat{n}}{n} = 1 + O_P \left( \frac{1}{\sqrt{m}} \right).
\]

Remark 1 gives a sense of the magnitude of \( \epsilon \), where \( \epsilon = \hat{n} - n \).

**Remark 2**: From (14.9) we can see that \( \text{Var}(\hat{n}) = O \left( \frac{1}{m} \right) \). Hence (14.7) and (14.9) give us some measure of accuracy of \( \hat{n} \) in estimating \( n \) when \( m \) is not too small. We shall examine by simulation how good the estimate \( \hat{n} \) is in estimating \( n \) for finite \( n \).

**Remark 3**: All these big or little o’s should be interpreted as the approximation when the pilot sample size is not too small. Although we can do standard asymptotics to say that \( \text{var}(\hat{n}) = O \left( \frac{1}{m} \right) \to 0 \) and “the right side of \( \hat{n} \) in (14.7)” \( \sim N(n, \text{var}(\hat{n})) \) as \( m \to \infty \), they would only be “true” if the “pilot” sample of size \( m \) is a proxy or historical sample used to estimate parameters and not included in the total sample size \( n \) of the current study. In practice, it’d be foolish to not
reuse the pilot sample if it’s from the same experiment under study. This shows the limitation of an asymptotics study in some cases and confirms the need for our simulation study in Chapter 14.

**Proposition 14.1.2.** Define Power($\hat{n}$) as the power of $\hat{n}$ under $H_1$, i.e., $\text{Power}(\hat{n}) = P_{H_1}(X^2 \geq \chi^2_{k-1}(\alpha))$. Then if $|\hat{n} - n| < \varepsilon$, it can be shown that $\text{Power}(\hat{n}) = 1 - \beta + O(\varepsilon)$.

This means the actual power of $\hat{n}$ tends to $(1 - \beta)$ as $\hat{n} \to n$ if the pilot sample gives a good estimate of the unknown parameters or as $m \to \infty$. See Appendix for the proof and a bound for the $O(\varepsilon)$ term.

For a pilot study with sample size $m$, we use the $\hat{\Delta}_j$’s and $\hat{p}_j$’s to estimate the $\Delta_j$’s and $p_j$’s, and let

$$
\sigma_0^2 = \hat{\Lambda}^2 \left\{ \sum_{j=1}^{k} (\hat{a}_j^2 + \hat{b}_j^2) + \sum_{j=1}^{k} \sum_{i=1}^{k} \left\{ (\hat{b}_i \hat{b}_j - \hat{a}_i \hat{a}_j)(\hat{i}_i \hat{i}_j)^{-\frac{1}{2}}(\hat{p}_{1i} \hat{p}_{1j} + \hat{p}_{2i} \hat{p}_{2j}) \right\} + \sum_{j=1}^{k} \sum_{i=1}^{k} \left\{ \hat{a}_i \hat{b}_i (\hat{i}_i \hat{i}_j)^{-\frac{1}{2}}(\hat{p}_{1i} \hat{p}_{1j} - \hat{p}_{2i} \hat{p}_{2j}) \right\} \right\}.
$$

Then, a 100$(1 - \alpha)$% approximate confidence interval for $n$ is:

$$(\hat{n} - z_{\alpha/2} \sigma_0, \hat{n} + z_{\alpha/2} \sigma_0),$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

### 14.2 Improvements for Frequentist Approach

Notice that formula (14.6), which can be used to calculate the required sample size given pilot or proxy data, is based on the estimated parameters from the pilot data. In order to reduce the sampling error from the pilot data and also stabilize the estimated required sample size, based on the above frequentist methodology, we propose the following improvements which will also be examined in Chapter 14.
14.2.1 Bootstrap Method

The basic idea of the bootstrap method is to draw bootstrap samples from the pilot data and take the mean or median of the calculated sample sizes from the bootstrap samples as the required sample size. Since there are two possible ways to draw a bootstrap sample, parametric and non-parametric, we have four bootstrap methods: parametric bootstrap-mean method, non-parametric bootstrap-mean method, parametric bootstrap-median method, and non-parametric bootstrap-median method.

**Parametric Bootstrap-mean and Bootstrap-median Methods**

With these two methods, bootstrap samples are drawn parametrically from multinomial distributions where the parameters of the multinomial distributions are estimated from the pilot data. Then we calculate the sample sizes using (14.6) for each of the bootstrap samples, for $i = 1, 2, \ldots, B$,

$$n^{(s_i)} = 2\lambda_0 \left( \sum_{j=1}^{k} \frac{\hat{\Delta}^{(s_i)}_{ij}^2}{\hat{p}^{(s_i)}_{ij}} \right)^{-1},$$

where $\hat{\Delta}^{(s_i)}_{ij} = \hat{p}^{(s_i)}_{1j} - \hat{p}^{(s_i)}_{2j}$, $\hat{p}_{ij} = \frac{1}{2} \left( \hat{p}^{(s_i)}_{1j} + \hat{p}^{(s_i)}_{2j} \right)$ and the $\hat{p}^{(s_i)}_{lj}$, $l = 1, 2$, are the usual estimates but based on the $i$th bootstrap sample and $B$ is the bootstrap sample size. Then take the mean and median of $\{n^{s_i}, i = 1, 2, \ldots, B\}$ as the required sample sizes for the bootstrap-mean method and for the bootstrap-median method, respectively.

**Non-parametric Bootstrap-mean and Median Methods**

For the non-parametric bootstrap method, we proceed as for the parametric case except that the bootstrap samples are drawn directly nonparametrically, i.e., from the pilot data using random sampling with replacement.

14.2.2 Minimum Difference Method

Another way to reduce the effect of sampling error from the pilot data is to specify the minimum difference between the two estimated parameters from the pilot data.
in (14.6). Let \(c\) denote the minimum difference between the two parameter vectors that we want to detect. Then the required sample size can be calculated as

\[
\hat{n} = 2\lambda_0 \left( \sum_{j=1}^{k} \frac{C_j^2}{\hat{p}_j} \right)^{-1},
\]

where \(C_j = \max(c, |\Delta_j|)\). This formula is same as that in (14.6) except that \(\hat{\Delta}_j\) is replaced by \(\hat{C}_j\). The common \(\hat{p}_j\) does not change because when \(\Delta_j\) is changed to \(\hat{C}_j\), \(\hat{p}_{j1}\) is changed to \(\hat{p}_{j1}^* = \frac{1}{2}(\hat{p}_{j1} + \hat{p}_{j2})\) and \(\hat{p}_{j2}\) to \(\hat{p}_{j2}^* = \hat{p}_{j1} - C_j\), which imply that \(\hat{p}_{j}^* = \frac{1}{2}(\hat{p}_{j1}^* + \hat{p}_{j2}^*) = \frac{1}{2}(\hat{p}_{j1} + \hat{p}_{j2}) = \hat{p}_j\). In practice, \(c\) is determined by scientist as the smallest difference that would be useful and meaningful to detect.

### 14.2.3 Correction Method

From (14.7), we get

\[
n = \hat{n} + A \cdot \left( \sum_{j=1}^{k} \frac{2\Delta_j \sqrt{T_j}}{p_j} Z_{j2} - \sum_{j=1}^{k} \frac{\Delta_j^2 \sqrt{T_j}}{2p_j} Z_{j1} \right) + o \left( \frac{1}{m} \right),
\]

where \(o\) is in probability and

\[
A = 2\lambda \cdot \frac{1}{\sqrt{m}} \cdot \left( \sum_{j=1}^{k} \frac{\Delta_j^2}{p_j} \right)^{-2},
\]

and \(Z_{j1}\) and \(Z_{j2}\) are standard normal random variables that are independent to each other. Now

\[
A \cdot \left( \sum_{j=1}^{k} \frac{2\Delta_j \sqrt{T_j}}{p_j} Z_{j2} - \sum_{j=1}^{k} \frac{\Delta_j^2 \sqrt{T_j}}{2p_j} Z_{j1} \right)
\]

can be treated as a bias correction term. In order to correct the estimated sample size, \(N\) pairs of random samples for \((Z_{j1}, Z_{j2})\) are drawn from standard normal distributions and the new corrected sample size, \(\hat{n}_0\), is

\[
\hat{n}_0 = \hat{n} + \frac{1}{N} \sum_{i=1}^{N} \left[ A \cdot \left( \sum_{j=1}^{k} \frac{2\Delta_j \sqrt{T_j}}{p_j} Z_{j2}^{(i)} - \sum_{j=1}^{k} \frac{\Delta_j^2 \sqrt{T_j}}{2p_j} Z_{j1}^{(i)} \right) \right].
\]

(14.10)
All the parameters in (14.10), including those in $A$, are replaced by their estimates.

References

Chapter 15

Bayesian approach

Under the null model, i.e., $H_0$, the two multinomial distributions have the same probability parameter $p^* = (p^*_1, p^*_2, \cdots, p^*_k)$, s.t.

$$n_i = (n_{i1}, n_{i2}, \cdots, n_{ik}) \sim \text{multinomial}(p^*, n_i), \quad i = 1, 2,$$

where $n_i = \sum_{j=1}^{k} n_{ij}$.

For a multinomial distribution, the conjugate prior distribution is Dirichlet with density function,

$$f(p^*|\alpha) \propto \prod_{j=1}^{k} p_j^{\alpha_j-1},$$

where the distribution is restricted to nonnegative $p_j$'s with $\sum_{j=1}^{k} p_j = 1$ and $\alpha_j > 0$.

The resulting posterior distribution of the $p_j$'s, given $n = (n_{ij}, i = 1, 2; j = 1, 2, \cdots, k)$ under $H_0$, is

$$f(p^*|n) \propto \prod_{j=1}^{k} p_j^{(\alpha_j+n_{j}-1)}, \quad (15.1)$$

where $n_{.j} = \sum_{i=1}^{2} n_{ij}$.

Let $t(n, p)$ be a discrepancy measure and let $n^{obs}$ be an observed value of $n$. 

114
Then the posterior predictive p-value is defined as

$$p_{ppp}\{n^{obs}\} = Pr_{H_0}\left\{t(n^{pred}, p^*) > t(n^{obs}, p^*)|n^{obs}\right\}, \quad (15.2)$$

where $p^*$ is the common vector of proportions for the two multinomial distributions. The probability in (15.2) is taken over the joint distribution of $(n^{pred}, p^*)$ given $n^{obs}$, in which $p^* \sim f(p^*|n^{obs})$ in (11) and $n^{pred}$ is a predictive value of $n$. Here, given $p^*$, $n^{pred}$ is assumed to be independent of $n^{obs}$ and $n_i^{pred}$ has the multinomial distribution, $\text{mult}(p^*, n_i)$.

One may reject the null model at level $\alpha$ if $p_{ppp}\{n^{obs}\} < \alpha$.

A natural choice of $t(n, p)$, analogous to

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - m_j)^2}{m_j},$$

for our test is

$$t(n, p) = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(n_{ij} - n_i p_j)^2}{n_i p_j},$$

where $n = (n_{ij} : i = 1, 2; j = 1, 2, \cdots, k)$, $p = (p_1, p_2, \cdots, p_k)$ and $n_i = \sum_{j=1}^{k} n_{ij}$.

Hjort et al. (2006)[17] explain problems with the posterior predictive p-value, i.e.,

“the ppp calculation uses the data twice, first updating the prior to fit the data better and then estimating how surprising the data are relative to the posterior parameter distribution. Thus it is not surprising that its distribution across likely values of $y_{obs}$ is not uniform; we can, in fact, demonstrate various extreme aspects of nonuniformity in several situations. This makes the interpretation and comparison of ppp values a difficult and risky matter.”

To correct the possible non-uniformity of the distribution of the posterior predictive p-value, Hjort et al. (2006)[17] proposed a calibrated posterior predictive p-value, i.e.,

$$p_{cppp}\{n^{obs}\} = Pr_{H_0}\left\{p_{ppp}\{n\} \leq p_{ppp}\{n^{obs}\}\right\}, \quad (15.3)$$
where $n$ comes from the marginal distribution derived from the joint distribution of $(p^*, n)$ in which $n$ has the multinomial distribution with parameter $p^*$, which comes from the prior distribution $\pi(p^*)$.

Since the calibrated posterior predictive p-value has a uniform distribution under the null model, it can play the role of a classical p-value.

Double-simulation can be used to estimate the value of $p_{cppp}(n^{obs})$. In our case, given specified sample sizes $n_1 = n_2 = m$, the value of $p_{cppp}(n^{obs})$ can be evaluated by simulation,

$$p_{cppp}(n^{obs}) \approx \frac{1}{A} \sum_{i=1}^{A} I\left(t(n_{i}^{pred}, p_i^*) \geq t(n^{obs}, p_i^*)\right), \quad (15.4)$$

for a large simulation size $A$, where each $p_i^*$ is simulated from the posterior distribution $f(p^*|n^{obs})$ and $n^{pred}$ is simulated from the multinomial distribution with parameters $p_i^*$ and $m$. The simulation is repeated independently $A$ times.

Then for a large number $B$, the calibrated posterior predictive p-value can be estimated by simulation,

$$p_{cppp}(n^{obs}) \approx \frac{1}{B} \sum_{j=1}^{B} I\left(p_{cppp}(n^{(j)}) \leq p_{cppp}(n^{obs})\right), \quad (15.5)$$

where $p^*$ is simulated from the prior distribution $\pi(p^*)$ and each $n^{(j)}$ is simulated from a multinomial distribution with parameter $p^*$ and $m$. For each $n^{(j)}$, $p_{cppp}(n^{(j)})$ can be approximated using (15.4). The simulation is repeated independently $B$ times.

Given a significance level $\alpha$ and power $1 - \beta$, we want to find the minimum sample size for each multinomial distribution such that

$$Pr_{H_1}\left\{p_{cppp}(n^{obs}) \leq \alpha \right\} \geq 1 - \beta,$$

where "$H_1$" denotes a specified value in $\Omega - \omega$.

In a balanced design, suppose that the sample size is $m$ for each multinomial. We generate samples from the two multinomials $N$ times,

$$n_i^j = (n_{i1}^j, n_{i2}^j, \ldots, n_{ik}^j),$$
where $j = 1, 2, \cdots, N$ and $i = 1, 2$.

Let $n^{\text{obs}}_j = (n^1_j, n^2_j)$ for $j = 1, 2, \cdots, N$. We need to evaluate the calibrated posterior predictive p-value, $p_{\text{cppp}}(n^{\text{obs}}_j)$, for each $n^{\text{obs}}_j$. Then we compare each calibrated posterior predictive p-value with the specified significance level $\alpha$ and calculate the proportion for which $p_{\text{cppp}}(n^{\text{obs}}_j) \leq \alpha$.

If the calculated proportion is less than $1 - \beta$, we increase the sample size $m$ and repeat the whole process. Otherwise, we decrease the sample size $m$ and repeat the process. Finally, we can find the minimum sample size $m$ such that the calculated proportion is at least $1 - \beta$.

Since this process is very computationally intensive, a good starting point of $m$ is very important. In practice, we may use the hypothesis test method to obtain a reasonable starting point.

The following is our proposed calibrated posterior predictive p-value based procedure to find the desired $m$:

- **Choose a starting value of** $m$.
- **Select** $p_1$ and $p_2$ from $\Omega - \omega$. Here, $p_1$ and $p_2$ are the parameters representing the minimal differences one would like to detect, or, possibly, estimates from pilot data.
- **Generate samples** $n_1$ and $n_2$ $N$ times,

$$
n^i_j \sim \text{multi}(m, p_i),
$$

where $j = 1, 2, \cdots, N$ and $i = 1, 2$.
- **Let** $n^i = (n^1_i, n^2_i)$ for $j = 1, 2, \cdots, N$. Evaluate the calibrated posterior predictive p-value $p_{\text{cppp}}(n^i)$ for each $n^i$.
- **Compare each calibrated posterior predictive p-value with the specified significance level** $\alpha$ and calculate the proportion for which $p_{\text{cppp}}(n^i) \leq \alpha$.

For a specified $\beta$, choose alternative values of $m$ until the proportion for which $p_{\text{cppp}}(n^i) \leq \alpha$ is about $1 - \beta$. 
References


Chapter 16

Simulation study

In this comprehensive simulation study, we evaluate the performance of both the frequentist and Bayesian methods. There are 3 different parameter settings for the frequentist methods, and for each parameter setting we chose several pilot sample sizes. For a given parameter setting and fixed pilot sample size, we generate pilot data 3,000 times and for each pilot data set, five frequentist methods, the original method using (14.6), and the improvements in Section 14.2, are evaluated simultaneously. For the minimum difference method (Section 14.2), the specified minimum difference between the proportions of two multinomial distributions is set to be 0.02, i.e., \( c = 0.02 \). For the bootstrap based methods, we only present the results for the parametric and nonparametric bootstrap-mean improvement methods since the performance of the bootstrap-mean methods is better than the performance of the bootstrap-median methods in our experiments.

16.1 Experiment 1 (Small Differences)

We first set the proportions for the two multinomial distributions to be

\[
p_1 = (0.10, 0.25, 0.30, 0.20, 0.15)
\]

and

\[
p_2 = (0.15, 0.20, 0.25, 0.30, 0.10).
\]
Clearly, the true differences between $p_{1j}$ and $p_{2j}$ are small, mostly 0.05. Given $\alpha = 0.05$, $\beta = 0.2$ and if these values of $p_1$ and $p_2$ are known, the true required sample size from (14.4) can be calculated to be 239. In practice, two values of $p_1$ and $p_2$ are unknown. Figure 16.1 is our side-by-side modified Boxplot comparison of the estimated sample sizes for the 5 frequentist methods when the values of $p_1$ and $p_2$ are unknown but some pilot data are available. These modified Boxplots have all the information provided by the standard Boxplots except for the points which are outside of the whisker; the counts of these outside points are presented at the top of each modified Boxplot. The horizontal line in the middle of the plot indicates the true sample size, 239. The circle and triangle indicate the mean and standard deviation (over the 5000 replicates) of the estimated sample size using each method. This type of modified Boxplot gives a compact, informative view.

![Figure 16.1: Side-by-side modified Boxplot for Experiment 1](image)

From the simulation results we see that all five methods produce means that approach the required sample size 239 as the pilot sample size increases. However, for the smaller sample sizes there is a significant underestimation. The non-parametric bootstrap method has somewhat smaller means and smaller standard deviations than the original and correction methods. The minimum difference method yields smaller standard deviations. This is probably due to the additional constraint ($c = 0.02$). Setting a minimum detectable difference is not uncommon in practice.

Depending on the pilot data the calculated sample sizes can be extremely large.
In a practical situation one would not conduct the main study if the required sample size was too large. So we repeated the simulation study (with 3000 replicates), but omitted all cases where the required sample size exceeded 900. The results, presented in Figure 16.2, show that for all five methods the means approach the required sample size, 239, as the pilot sample increases. For the smaller pilot sample sizes, there is a significant underestimation. The five methods have similar means, slightly smaller for both bootstrap methods. However the standard deviation is smaller for the minimum difference method, and the smallest for the non-parametric bootstrap method.

A summary assessment of these methods is given in Table 16.1 where the methods are compared using the average mean square error (MSE). Here, it is clear that the original and correction methods are essentially equivalent as are the two bootstrap methods. However, the average MSE is much smaller for the bootstrap methods, e.g., a 52% reduction for pilot size 150 and a 63% reduction for pilot size 200. Finally, the mini-diff method has MSE’s less than the original method but consistently larger than the bootstrap methods.
### Table 16.1: Averaged MSE of sample sizes

<table>
<thead>
<tr>
<th>Pilot Sample Size</th>
<th>30</th>
<th>80</th>
<th>120</th>
<th>150</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Method</td>
<td>9.37</td>
<td>7.28</td>
<td>6.56</td>
<td>6.05</td>
<td>5.63</td>
</tr>
<tr>
<td>Correction Method</td>
<td>9.42</td>
<td>7.29</td>
<td>6.57</td>
<td>6.09</td>
<td>5.63</td>
</tr>
<tr>
<td>NpB-Mean Method</td>
<td>8.56</td>
<td>5.39</td>
<td>4.40</td>
<td>3.99</td>
<td>3.46</td>
</tr>
<tr>
<td>PB-Mean Method</td>
<td>8.57</td>
<td>5.40</td>
<td>4.40</td>
<td>4.00</td>
<td>3.46</td>
</tr>
<tr>
<td>Min-Diff Method</td>
<td>9.00</td>
<td>6.64</td>
<td>5.93</td>
<td>5.47</td>
<td>5.03</td>
</tr>
</tbody>
</table>

#### 16.2 Experiment 2 (Medium Differences)

In this experiment, the proportions for the two multinomial distributions are set to be

\[
p_1 = (0.10, 0.25, 0.30, 0.20, 0.15)
\]

and

\[
p_2 = (0.17, 0.32, 0.36, 0.10, 0.05),
\]

and the true sample size calculated from (14.4) is 104. The side-by-side Boxplots analogous to those in Figure 16.2 (with sample sizes truncated at 900) are shown in Figure 16.3. The results in Figure 16.3 lead to the same conclusion as those in Section 16.1 for experiment 1.

![Side-by-side modified Boxplot for Experiment 2 after adding limit 900](image)

Figure 16.3: Side-by-side modified Boxplot for Experiment 2 after adding limit 900

The summarized averaged MSE is in Table 16.2.
16.3 Experiment 3 (Large Differences)

In experiment 3, we consider the case in which the differences between the parameters from the two populations are moderately large, i.e.,

\[ p_1 = (0.10, 0.25, 0.30, 0.20, 0.15) \]

and

\[ p_2 = (0.30, 0.10, 0.20, 0.10, 0.30). \]

The true required sample size calculated from (14.4) is 45. The side-by-side Box-plots analogous to those in Figure 16.2 (with sample sizes truncated at 900) are shown in Figure 16.4. The averaged MSEs are in Table 16.3. The recommendation is to use the original method for both small and large samples. If additional computation is easy to do, nonparametric bootstrap methods may be used. See Table 16.4.

<table>
<thead>
<tr>
<th>Pilot Sample Size</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>50</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Method</td>
<td>0.43</td>
<td>0.40</td>
<td>0.36</td>
<td>0.22</td>
<td>0.12</td>
</tr>
<tr>
<td>Correction Method</td>
<td>0.44</td>
<td>0.41</td>
<td>0.37</td>
<td>0.22</td>
<td>0.12</td>
</tr>
<tr>
<td>NpB-Mean Method</td>
<td>0.41</td>
<td>0.38</td>
<td>0.34</td>
<td>0.21</td>
<td>0.12</td>
</tr>
<tr>
<td>PB-Mean Method</td>
<td>0.41</td>
<td>0.38</td>
<td>0.34</td>
<td>0.21</td>
<td>0.11</td>
</tr>
<tr>
<td>Min-Diff Method</td>
<td>0.39</td>
<td>0.37</td>
<td>0.35</td>
<td>0.21</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Table 16.3: Averaged MSE of sample sizes

Table 16.4 provides the summary of our recommendations based on all simulation results. If some knowledge about the size of differences is known, one could follow the winner table in choosing a method with the help of program and code. Otherwise, if minimum difference is known and no code, one can use minimum difference method.
In this experiment, we use the same parameter settings as those in Experiment 1 and Experiment 2. First we generate 50 pilot data sets. For each of these pilot data sets, we use the original frequentist method to calculate the estimated sample size using (14.6), \( n_0 \). Using 3 different starting values, \( n_0, n_0 - 15, \) and \( n_0 + 15, \) we apply our Bayesian method to each of the pilot data sets. Then we compare the performance of the Bayesian method with these three different starting values. Figure 16.5 and Figure 16.6 present the simulation results for pilot sample sizes 30 and 80 from Experiment 1, respectively. Figure 16.7 and Figure 16.8 present the simulation results for pilot sample sizes 20 and 30 from Experiment 2. Each figure is a scatter plot of indices 1 to 50 versus the 50 sorted \( n_0 \), superimposed with the three matching Bayesian estimates at each index.

Overall, these four estimates are very close to each other, indicating some ro-
bustness in the Bayesian method. Taking a microscopic look, the original method and Bayesian method with starting value $n_0$ produce the closest estimates. On the other hand, the calculated sample sizes for the Bayesian method using starting value $n_0 - 15$ ($n_0 + 15$) are usually less (greater) than the calculated sample sizes from the original method. Therefore, it’s possible that the Bayesian method starting from $n_0 + 15$ produce more estimates that are close to the target than the original $n_0$ would, even though the differences are small. (See the summary in
Table 16.5)

<table>
<thead>
<tr>
<th>Sample Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Method</td>
</tr>
<tr>
<td>Bayesian Method (n0)</td>
</tr>
<tr>
<td>Bayesian Method (n0-15)</td>
</tr>
<tr>
<td>Bayesian Method (n0+15)</td>
</tr>
</tbody>
</table>

Figure 16.7: Plot of estimated sample sizes based on pilot sample size 20 for Experiment 2

Figure 16.8: Plot of estimated sample sizes based on pilot sample size 30 for Experiment 2

In Table 16.5, we count the number of times that $\|\hat{n}_B - n_T\| > \|\hat{n}_O - n_T\|$ for each data set, where $\hat{n}_B$ is the estimated sample size from Bayesian method, $\hat{n}_O$ is the estimated sample size from original method, and $n_T$ is the target sample size. The results are summarized in Table 16.5.

Based on the results in Table 16.6 and Figures 16.5-16.8, if one has the computation power, it is recommended to use the sample size from the pppp with
<table>
<thead>
<tr>
<th>Starting Value</th>
<th>Figure 16.5</th>
<th>Figure 16.6</th>
<th>Figure 16.7</th>
<th>Figure 16.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_0 - 15$</td>
<td>73%</td>
<td>42%</td>
<td>70%</td>
<td>30%</td>
</tr>
<tr>
<td>$n_0$</td>
<td>32%</td>
<td>12%</td>
<td>40%</td>
<td>20%</td>
</tr>
<tr>
<td>$n_0 + 15$</td>
<td>20%</td>
<td>20%</td>
<td>56%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Table 16.5: Performance comparison

starting value $n_0 + 15$ in the small differences and small sample size cases and $n_0$ otherwise.
Chapter 17

Application, Discussion, and Conclusion

17.1 Application to Leukoplakia Lesions

Leukoplakia is associated with several factors such as poor diet, poor oral hygiene, local irritants, alcohol and tobacco (Pindborg et al. (1967)[29]). The locations of leukoplakia lesion are closely related to different smoking habits. Here, we want to compare the distribution of lesion locations for two types of different smoking, i.e., Bidi smoking and non-Bidi smoking. According to Neville and Day (2002)[25], the locations of oral leukoplakia are significantly correlated with the frequency of finding dysplastic or malignant changes at biopsy. Pindborg et al. (1967)[29] present a data set giving the locations of lesions for 363 Bidi smoking individuals and for 142 non-Bidi smoking individuals. These data are presented in Table 17.1 and we regard them as the pilot data for the main study that we wish to design. Unlike the previous simulation study, the pilot sample sizes in this application are not equal. In order to determine the required sample sizes based on the unbalanced pilot data, we set the ratio of two required sample sizes equal to the ratio of two pilot sample sizes. In this application, after removing the Tongue category (since the observations are equal for both chewing habits), the ratio of two pilot sample sizes, $r$, is $(363 - 7)/(142 - 7) = 2.64$. Suppose the required sample size for bidi smoking is $n_1$, the required sample size for non-bidi smoking is $n_2 = n_1/2.64$. 

128
After fixing the ratio between two required sample sizes, the required sample size for bidi smoking, \( n_1 \), can be easily written as

\[
\frac{(r + 1)\lambda_0 \left( \sum_{j=1}^{k} \frac{(p_{1j} - p_{2j})^2}{p_j} \right)^{-1}}{1}, \tag{17.1}
\]

where the notation is the same as in Section 1 (with “1” indicating the bidi group), \( k = 9 \) and \( \lambda_0 \) is the minimum value of \( \lambda \) such that (14.3) holds given \( \alpha \) and \( (1 - \beta) \).

Using the estimates from the pilot data, the estimated sample size for bidi smoking, \( \hat{n}_1 \), can be calculated from the above formula. The corresponding frequentist improvements for unbalanced case are similar to those for balanced case except for correction method. The correction method is developed based on balanced design and can not be applied to unbalanced case directly. The calculated sample sizes for bidi smoking and non-bidi smoking are provided in Table 17.2.

For mini-diff method, the minimum difference we want to detect between two proportions is set to be 0.01, i.e., \( c = 0.01 \). Then the calculated sample sizes using minimum difference methods are 196 and 75 for bidi and non-bidi smoking respectively.

<table>
<thead>
<tr>
<th>Location of lesion</th>
<th>Bidi smoking</th>
<th>non-bidi smoking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labial commissure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>101</td>
<td>24</td>
</tr>
<tr>
<td>Left</td>
<td>88</td>
<td>25</td>
</tr>
<tr>
<td>Buccal mucosa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>70</td>
<td>31</td>
</tr>
<tr>
<td>Left</td>
<td>70</td>
<td>35</td>
</tr>
<tr>
<td>Lip</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upper</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Lower</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>Alveolar ridge</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Tongue</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Gingiva</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Palate</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>363</td>
<td>142</td>
</tr>
</tbody>
</table>

Table 17.1: Leukoplakia lesion locations regarding to bidi and non-bidi smoking habits
If each individual patient has only one Leukoplakia lesion or his/her primary lesion is of concern, the multinomial model would be a reasonable one for the data and the sample sizes in Table 17.2 would be the required sample sizes for the desired power. If there is more than one lesion for some patients (which is often true, but the number of overlaps is unknown a priori) and primary lesion can not be identified the sample sizes calculated from our procedures will serve as an upper bound for the true required sample size as there will be more occurrences of each lesion, even though the occurrences are likely to be positively correlated. And this upper bound will provide useful information for choosing the sample size.

### 17.2 Discussion and Conclusion

In this part we have developed both frequentist and Bayesian approaches to the calculation of the sample size needed to contrast two multinomial populations. The original frequentist approach is based on asymptotic theory while the Bayesian approach is based on computationally intensive simulation. We also have developed several methods to improve the original frequentist approach, one using the bootstrap and another specifying the minimum difference between the parameters that the investigator wishes to detect. We have found that both the original method and Bayesian method provide similar choices for the sample sizes while the bootstrap methods provide better performance than the others. The practical recommendations are provided. Our approaches are developed and studied based on unordered categorical data. For ordered categorical data, the related reference will be Whitehead (1993)[36]. Alternatively, instead of using non-parametric approaches like ours, parametric solutions to this problem are also possible for unordered categorical data. Please note that although there are sample size calculations based on (similar) Chi-squared tests, they are for the hypotheses different from ours. For example, see Fleiss (1981), Haber (1983, eenland (1983), Rochon

<table>
<thead>
<tr>
<th>Methods</th>
<th>Original</th>
<th>Non-para Bootstrap</th>
<th>Para Bootstrap</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1$</td>
<td>291</td>
<td>275</td>
<td>273</td>
</tr>
<tr>
<td>$n_2$</td>
<td>111</td>
<td>105</td>
<td>104</td>
</tr>
</tbody>
</table>

Table 17.2: Sample sizes for Leukoplakia lesion
Chapter 18

Appendix

18.1 Proof of Proposition 14.1.2:

Proof. Let $\hat{e}_0 = \hat{n} - n$ and

$$c = \frac{1}{2} \sum_{j=1}^{k} \left\{ \frac{(\delta_{1j} - \delta_{2j})^2}{p_j} \right\}.$$ 

Conditioning on the pilot sample, for each fixed numerical value of $\hat{n}$ (i.e. not random value), the power can be evaluated as following:

$$\hat{P}_{H_1} \left( X^2 \geq \chi^2_{k-1}(\alpha) \right)$$

$$= 1 - \sum_{j=0}^{\infty} e^{-\frac{1}{2}} \cdot \left( \frac{\lambda}{2} \right)^j \cdot \frac{j!}{\Gamma \left( j + \frac{k-1}{2} \right)} \cdot \gamma \left( j + \frac{k-1}{2}, \frac{\chi^2_{k-1}(\alpha)}{2} \right).$$

(18.1)

(18.2)

where $\Gamma(x)$ is the gamma function, $\gamma(a, x) = \int_0^x t^{a-1} \cdot e^{-t} dt$, which is lower incomplete gamma function and

$$\hat{\lambda} = \frac{\hat{n}}{2} \sum_{j=1}^{k} \left\{ \frac{(\delta_{1j} - \delta_{2j})^2}{p_j} \right\}.$$ 

If $\hat{n}$ equals to the true sample size in (14.4), then the power from (14.6) is exactly $(1 - \beta)$ by the construction in (14.3). Next, we will see the effect of
the difference between $\hat{n}$ and $n$, and hence the difference between $\hat{\lambda}$ and $\lambda$

to (14.6).

Under $H_1$, the power is

$$
\hat{P}_{H_1} \left( X^2 \geq \chi^2_{k-1}(\alpha) \right)
\quad = \quad 1 - \sum_{j=0}^{\infty} e^{-\frac{1}{2}} \cdot \left( \frac{1}{2} \right)^j \frac{\gamma \left( j + \frac{k-1}{2}, \frac{\chi^2_{k-1}(\alpha)}{2} \right)}{\Gamma \left( j + \frac{k-1}{2} \right)}
\quad = \quad 1 - \sum_{j=0}^{\infty} e^{-\frac{1}{2}} \cdot \left( \frac{1}{2} \right)^j \frac{\gamma \left( j + \frac{k-1}{2}, \frac{\chi^2_{k-1}(\alpha)}{2} \right)}{\Gamma \left( j + \frac{k-1}{2} \right)} + C
\quad = \quad (1 - \beta) + C,
$$

where

$$
C = \sum_{j=0}^{\infty} \left[ e^{-\frac{1}{2}} \cdot \left( \frac{1}{2} \right)^j / j! - e^{-\frac{1}{2}} \cdot \left( \frac{1}{2} \right)^j / j! \right] \cdot \frac{\gamma \left( j + \frac{k-1}{2}, \frac{\chi^2_{k-1}(\alpha)}{2} \right)}{\Gamma \left( j + \frac{k-1}{2} \right)}.
$$

Let

$$
g(j) = \frac{\gamma \left( j + \frac{k-1}{2}, \frac{\chi^2_{k-1}(\alpha)}{2} \right)}{\Gamma \left( j + \frac{k-1}{2} \right)}.
$$

Notice $g(j) \leq 1$.

Then

$$
C = \sum_{j=0}^{M} \left[ e^{-\frac{1}{2}} \cdot \left( \frac{1}{2} \right)^j / j! - e^{-\frac{1}{2}} \cdot \left( \frac{1}{2} \right)^j / j! \right] \cdot g(i) + 
\sum_{j=M+1}^{\infty} \left[ e^{-\frac{1}{2}} \cdot \left( \frac{1}{2} \right)^j / j! - e^{-\frac{1}{2}} \cdot \left( \frac{1}{2} \right)^j / j! \right] \cdot g(i) \quad (18.3)
$$
For any given $\delta > 0$, we want to find $M$ such that the second part of (16.3) satisfies

$$\left| \sum_{j=M+1}^{\infty} \left[ e^{-\frac{1}{2}} \cdot \left( \frac{j}{2} \right)^j \right] \cdot g(i) \right| \leq \frac{\delta}{4}. \quad (18.4)$$

Since

$$\left| \sum_{j=M+1}^{\infty} \left[ e^{-\frac{1}{2}} \cdot \left( \frac{j}{2} \right)^j \right] \cdot g(i) \right| \leq \sum_{j=M+1}^{\infty} \left[ e^{-\frac{1}{2}} \cdot \left( \frac{j}{2} \right)^j \right] + \sum_{j=M+1}^{\infty} \left[ e^{-\frac{1}{2}} \cdot \left( \frac{j}{2} \right)^j \right]$$

$$= \sum_{j=M+1}^{\infty} \left[ \frac{\lambda}{j!} \cdot e^{-\frac{1}{2}} \cdot \frac{\lambda}{2} \cdot \frac{j}{2} \right] + \sum_{j=M+1}^{\infty} \left[ \frac{\hat{\lambda}}{j!} \cdot e^{-\frac{1}{2}} \cdot \frac{\hat{\lambda}}{2} \cdot \frac{j}{2} \right]$$

$$= \left[ 1 - \frac{\Gamma(M+1, \frac{\lambda}{2})}{M!} \right] + \left[ 1 - \frac{\Gamma(M+1, \frac{\hat{\lambda}}{2})}{M!} \right]$$

$$= \int_0^{\frac{\lambda}{2}} \frac{x^{M-1} e^{-x}}{(M-1)!} \cdot \frac{x}{M} dx + \int_0^{\frac{\hat{\lambda}}{2}} \frac{x^{M-1} e^{-x}}{(M-1)!} \cdot \frac{x}{M} dx$$

$$\leq \frac{\lambda}{2M} \int_0^{\frac{\lambda}{2}} \frac{x^{M-1} e^{-x}}{\Gamma(M)} dx + \frac{\hat{\lambda}}{2M} \int_0^{\frac{\hat{\lambda}}{2}} \frac{x^{M-1} e^{-x}}{\Gamma(M)} dx$$

$$\leq \frac{\lambda}{2M} + \frac{\hat{\lambda}}{2M},$$

where the $p(j; \lambda)$ is the probability distribution function for Poisson distribution with parameter $\lambda$.

Hence, let

$$M = \frac{2(\lambda + \hat{\lambda})}{\delta},$$

then inequality (18.4) holds.
For the first part of (18.3), if $M$ is large, we have

$$\sum_{j=0}^{M} \left[ e^{-\frac{j}{2}} \cdot \left( \frac{\hat{\lambda}}{2} \right)^{\frac{j}{2}} - e^{-\frac{j}{2}} \cdot \left( \frac{\hat{\lambda}}{2} \right)^{\frac{j}{2}} \right] \cdot g(i)$$

$$\leq \sum_{j=0}^{M} \left[ e^{-\frac{j}{2}} \cdot \left( \frac{\lambda}{2} \right)^{\frac{j}{2}} + e^{-\frac{j}{2}} \cdot \left( \frac{\hat{\lambda}}{2} \right)^{\frac{j}{2}} \right]$$

$$= \sum_{j=0}^{M} \left[ e^{-\frac{j}{2}} \cdot \left( \frac{\hat{\lambda}}{2} \right)^{\frac{j}{2}} \right] + \sum_{j=0}^{M} \left[ e^{-\frac{j}{2}} \cdot \left( \frac{\lambda}{2} \right)^{\frac{j}{2}} \right]$$

$$= \frac{\Gamma(M + 1, \frac{\lambda}{2}) + \Gamma(M + 1, \frac{\hat{\lambda}}{2})}{M!}$$

$$= \int_{\frac{\lambda}{2}}^{+\infty} \frac{t^{M-1}}{\Gamma(M+2)} \cdot e^{-t} \cdot dt + \int_{\frac{\hat{\lambda}}{2}}^{+\infty} \frac{t^{M-1}}{\Gamma(M+2)} \cdot e^{-t} \cdot dt$$

$$\leq \frac{2(M+1)}{\lambda} + \frac{2(M+1)}{\hat{\lambda}}$$

$$= 2 \left( \frac{1}{\lambda} + \frac{1}{\hat{\lambda}} \right) \left( \frac{2}{\delta} (\lambda + \hat{\lambda}) + 1 \right)$$

From above calculation, we can see

$$\sum_{j=0}^{\infty} \left[ 1 - e^{-\frac{j}{2}} \cdot \left( 1 + \frac{\hat{\epsilon}_0}{n} \right)^{\frac{j}{2}} \right] \cdot \left( e^{-\frac{j}{2}} \cdot \left( \frac{\hat{\lambda}}{2} \right)^{\frac{j}{2}} \right) \cdot g(j)$$

$$\leq \frac{\delta}{4} + 2 \left( \frac{1}{\lambda} + \frac{1}{\hat{\lambda}} \right) \left( \frac{2}{\delta} (\lambda + \hat{\lambda}) + 1 \right).$$
Since the Taylor expansion of $e^{-\frac{\hat{n}c}{T}}$ at 0 is

$$e^{-\frac{\hat{n}c}{T}} = 1 - \frac{\hat{n}0c}{2} + \frac{(\hat{n}0c)^2}{8} + O(\hat{n}0^3).$$

Then

$$\sum_{j=0}^{\infty} \left[ e^{-\frac{\hat{n}c}{T}} \cdot \left( 1 + \frac{\hat{n}0j}{n} \right) - 1 \right] \cdot \left( e^{-\frac{nc}{T}} \cdot \left( \frac{n\hat{n}}{2} \right)^j \cdot \frac{\gamma(j + \frac{k-1}{2}, \chi^2_{(k-1)}(\alpha))}{\Gamma(j + \frac{k-1}{2})} \right) = O(\hat{n}0) \cdot \left( \frac{n\hat{n}}{2} \right)^j \cdot \frac{\gamma(j + \frac{k-1}{2}, \chi^2_{(k-1)}(\alpha))}{\Gamma(j + \frac{k-1}{2})}.$$

Now we get $\text{Power}(\hat{n}) = 1 - \beta + O(\hat{n}0).$
Bibliography


Kuznetsov, S. O., and Obiedkov, S. A. Algorithms for the construction of concept lattices and their diagram graphs. *Principles of Data Mining*


