LIFE TABLE ANALYSIS FOR COMPLEX SURVEY DATA

by

Kevin Francis O'Brien

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1337
April 1981
LIFE TABLE ANALYSIS FOR COMPLEX SURVEY DATA

by

Kevin Francis O'Brien

A Dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health

Chapel Hill
1981

Approved by:

Advisor

Reader

Reader
ABSTRACT

KEVIN FRANCIS O'BRIEN. Life Table Analysis for Complex Survey Data. (Under the direction of C. M. SUCHINDRAN)

This research is concerned with developing a methodology for performing a life table analysis for data obtained in a complex sample survey.

Four methods for estimating the variance of a function estimated using complex survey data are reviewed. These four methods are then applied to estimates of the conditional probabilities of an event and the associated survivorship probabilities, and suggestions are made as to the 'best' variance estimation method to use.

A test statistic analogous to the Mantel-Haenszel test is developed for comparing survivorship probabilities between two or more domains of interest for complex survey data. Use of the test statistic is illustrated for two data sets. Values of the test statistic based on the complex survey design are compared to values obtained using the Mantel-Haenszel test statistic assuming a simple random sample.

Life table regression models are reviewed and a model is developed which can be used with complex survey data. This brings to discussion the issue of likelihood based inference and estimation in complex surveys. Arguments based upon superpopulation models are given which support the use of likelihood methods for complex survey data. Variance estimates for the maximum likelihood estimators obtained from complex survey data are developed, and examples illustrating the given approach are given.
ACKNOWLEDGMENTS

It has been a great privilege to have been a student in the Department of Biostatistics at the University of North Carolina at Chapel Hill. I would like to thank the faculty, staff, and students for making my graduate studies so enjoyable.

I am especially grateful to my advisor, Dr. C. M. Suchindran, for his guidance and quick response to questions and problems presented during this research. The other members of my committee, Drs. H. Bradley Wells, Kerry L. Lee, William D. Kalsbeek, and Jack Kasarda were most helpful with their comments, encouragement, and support.

I would like to thank both the Department of Biostatistics and the Carolina Population Center for their financial support during my graduate training. The U.S. Bureau of the Census also provided support through Joint Statistical Agreements JSA 79-16 and 80-19.

Data for this research were provided by the kind generosity of The World Fertility Survey Program, London, and by the government of Sri Lanka.

A special thanks to Bea Parker for her skillful typing and preparation of this dissertation.

Finally, a thanks to my wife, Sharon, for her constant encouragement and support.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>ii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF GRAPHS</td>
<td>x</td>
</tr>
<tr>
<td>I</td>
<td>INTRODUCTION AND REVIEW OF THE LITERATURE</td>
</tr>
<tr>
<td>1.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>1.2</td>
<td>The Nature, Purpose, and Types of Life Tables</td>
</tr>
<tr>
<td>1.3</td>
<td>Notation and Definitions</td>
</tr>
<tr>
<td>1.3.1</td>
<td>Actuarial/Demographic Life Table Notation</td>
</tr>
<tr>
<td>1.3.2</td>
<td>Notation from Survivorship Analysis</td>
</tr>
<tr>
<td>1.3.3</td>
<td>Interrelatedness of the Two Sets of Notations</td>
</tr>
<tr>
<td>1.4</td>
<td>Life Table Construction: A Review</td>
</tr>
<tr>
<td>1.4.1</td>
<td>Methods for Actuarial/Demographic Life Tables</td>
</tr>
<tr>
<td>1.4.2</td>
<td>Methods for Follow-up Study Life Tables</td>
</tr>
<tr>
<td>1.5</td>
<td>Estimates of the Survivorship Function</td>
</tr>
<tr>
<td>1.6</td>
<td>Distributions of the Life Table Functions</td>
</tr>
<tr>
<td>1.6.1</td>
<td>The Distribution of the Number of Survivors</td>
</tr>
<tr>
<td>1.6.2</td>
<td>Distribution of the Number of Deaths</td>
</tr>
<tr>
<td>1.6.3</td>
<td>The Distribution of Life Expectancy</td>
</tr>
<tr>
<td>1.7</td>
<td>Variance Estimates for $\mu_t$ and $\lambda(t)$</td>
</tr>
<tr>
<td>1.8</td>
<td>Outline of Subsequent Chapters</td>
</tr>
<tr>
<td>II</td>
<td>VARIANCE ESTIMATES OF LIFE TABLE FUNCTIONS FROM A COMPLEX PROBABILITY SAMPLE</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
</tr>
<tr>
<td>2.2</td>
<td>Problems With Variance Estimation</td>
</tr>
<tr>
<td>2.3</td>
<td>A Review of Variance Estimation Methods</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Balanced Repeated Replications</td>
</tr>
<tr>
<td>2.3.2</td>
<td>The Taylor Series Linearization</td>
</tr>
<tr>
<td>2.3.3</td>
<td>The Jackknife</td>
</tr>
</tbody>
</table>
# TABLE OF CONTENTS (Continued)

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4 Variance Estimates of the Life Table Functions</td>
<td>32</td>
</tr>
<tr>
<td>2.5 The Data Used for This Study</td>
<td>36</td>
</tr>
<tr>
<td>2.6 Results</td>
<td>41</td>
</tr>
<tr>
<td>2.7 Discussion of the Results</td>
<td>44</td>
</tr>
<tr>
<td><strong>III</strong> COMPARING SURVIVORSHIP CURVES</td>
<td>61</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>61</td>
</tr>
<tr>
<td>3.2 A Review of Methods for Comparing Survivorship Curves</td>
<td>61</td>
</tr>
<tr>
<td>3.3 A Review of Some Methods for the Analysis of Complex Survey Data</td>
<td>65</td>
</tr>
<tr>
<td>3.4 An Examination of the Life Table Variance-Covariance Structure</td>
<td>68</td>
</tr>
<tr>
<td>3.5 A Modified Mantel-Haenszel Statistic</td>
<td>77</td>
</tr>
<tr>
<td>3.6 Extension to More Than Two Populations</td>
<td>79</td>
</tr>
<tr>
<td>3.7 Results of Applying the Modified Mantel-Haenszel Test</td>
<td>82</td>
</tr>
<tr>
<td>3.8 Discussion</td>
<td>87</td>
</tr>
<tr>
<td><strong>IV</strong> LIFE TABLE REGRESSION ANALYSIS</td>
<td>98</td>
</tr>
<tr>
<td>4.1 Introduction</td>
<td>98</td>
</tr>
<tr>
<td>4.2 Life Table Regression Model</td>
<td>98</td>
</tr>
<tr>
<td>4.3 A Life Table Regression Model</td>
<td>106</td>
</tr>
<tr>
<td>4.4 An Approximate Solution for the Two Sample Problem</td>
<td>110</td>
</tr>
<tr>
<td>4.5 Approximate Solution to the r-Sample Problem</td>
<td>115</td>
</tr>
<tr>
<td>4.6 Application of the Method Developed in Section 4.4</td>
<td>116</td>
</tr>
<tr>
<td>4.7 Life Table Regression in Complex Probability Samples</td>
<td>118</td>
</tr>
<tr>
<td><strong>V</strong> MAXIMUM LIKELIHOOD ESTIMATION IN COMPLEX PROBABILITY SAMPLES</td>
<td>128</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>128</td>
</tr>
<tr>
<td>5.2 Use of Maximum Likelihood in Complex Probability Samples</td>
<td>128</td>
</tr>
<tr>
<td>5.3 Life Table Analysis and Superpopulations</td>
<td>129</td>
</tr>
<tr>
<td>5.4 Likelihood Construction for Survey Data</td>
<td>130</td>
</tr>
<tr>
<td>5.5 Examples Using Weighted Likelihoods</td>
<td>132</td>
</tr>
<tr>
<td>5.6 Variance Estimation for the Maximum Likelihood Estimator: The Single Parameter Case</td>
<td>135</td>
</tr>
<tr>
<td>5.7 Variance Estimation for the Maximum Likelihood Estimator: The Multiparameter Case</td>
<td>140</td>
</tr>
<tr>
<td>5.8 Application to Life Table Regression Problems</td>
<td>142</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>VI</td>
<td>SUMMARY AND SUGGESTIONS FOR FUTURE RESEARCH</td>
</tr>
<tr>
<td>6.1</td>
<td>Summary</td>
</tr>
<tr>
<td>6.2</td>
<td>SUGGESTIONS FOR FUTURE RESEARCH</td>
</tr>
<tr>
<td></td>
<td>BIBLIOGRAPHY</td>
</tr>
<tr>
<td></td>
<td>APPENDIX</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5.1</td>
<td>Some characteristics of the four subsamples formed from zones one and two of the 1975 World Fertility Survey of Sri Lanka</td>
<td>38</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Some characteristics of the three subsamples formed from the 1973 National Survey of Family Growth</td>
<td>40</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Applications demonstrating the performance of the proposed test statistic with respect to the method used for variance estimation</td>
<td>94</td>
</tr>
<tr>
<td>3.7.2</td>
<td>Applications demonstrating the performance of the proposed test statistic, $T_{CPS}$; Using data from the 1975 World Fertility Survey of Sri Lanka</td>
<td>95</td>
</tr>
<tr>
<td>3.7.3</td>
<td>Applications demonstrating the performance of the proposed test statistic, $T_{CPS}$; Using data from the 1975 World Fertility Survey of Sri Lanka</td>
<td>96</td>
</tr>
<tr>
<td>3.7.4</td>
<td>Applications demonstrating the performance of the proposed test statistic, $T_{CPS}$; Using data from the 1973 National Survey of Family Growth</td>
<td>97</td>
</tr>
<tr>
<td>4.1</td>
<td>Number of Adult Drosophila Melanogaster living and number dying and sex</td>
<td>124</td>
</tr>
<tr>
<td>4.2</td>
<td>Survival data for lung cancer patients on standard therapy and test therapy</td>
<td>125</td>
</tr>
<tr>
<td>4.3</td>
<td>Estimated regression parameters and their estimated standard errors by the method of estimation using data from Chiang (1968)</td>
<td>126</td>
</tr>
<tr>
<td>4.4</td>
<td>Estimated regression parameters and their estimated standard errors by the method of estimation using data from Holford (1976)</td>
<td>127</td>
</tr>
<tr>
<td>5.1</td>
<td>Parameter estimates by model and by sample design</td>
<td>144</td>
</tr>
<tr>
<td>5.2</td>
<td>Variance estimates by model and by sample design</td>
<td>145</td>
</tr>
<tr>
<td>5.3</td>
<td>Design effects for the logistic and life table regression models</td>
<td>146</td>
</tr>
</tbody>
</table>
## LIST OF TABLES (Continued)

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.1</td>
<td>Variances and design effects by time interval and by method of estimation for the conditional probabilities of subsample one taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>159</td>
</tr>
<tr>
<td>A.2</td>
<td>Variances and design effects by time interval for the conditional probabilities of subsample two taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>160</td>
</tr>
<tr>
<td>A.3</td>
<td>Variances and design effects by time interval for the conditional probabilities of subsample three taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>161</td>
</tr>
<tr>
<td>A.4</td>
<td>Variances and design effects by time interval for the conditional probabilities of subsample four taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>162</td>
</tr>
<tr>
<td>A.5</td>
<td>Variances and design effects by time interval for the conditional probabilities of subsample one taken from the 1973 National Survey of Family Growth</td>
<td>163</td>
</tr>
<tr>
<td>A.6</td>
<td>Variances and design effects by time interval for the conditional probabilities of subsample two taken from the 1973 National Survey of Family Growth</td>
<td>164</td>
</tr>
<tr>
<td>A.7</td>
<td>Variances and design effects by time interval for the conditional probabilities of subsample three taken from the 1973 National Survey of Family Growth</td>
<td>165</td>
</tr>
<tr>
<td>A.8</td>
<td>Variances and design effects by time interval for the survivorship probabilities of subsample one taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>166</td>
</tr>
<tr>
<td>A.9</td>
<td>Variances and design effects by time interval for the survivorship probabilities of subsample one taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>167</td>
</tr>
<tr>
<td>A.10</td>
<td>Variances and design effects by time interval for the survivorship probabilities of subsample three taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>168</td>
</tr>
<tr>
<td>A.11</td>
<td>Variances and design effects by time interval for the survivorship probabilities of subsample three taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>169</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>A.12</td>
<td>Variances and design effects by time interval for the survivorship probabilities of subsample three taken from the 1973 National Survey of Family Growth</td>
<td>170</td>
</tr>
<tr>
<td>A.13</td>
<td>Variances and design effects by time interval for the survivorship probabilities of subsample three taken from the 1973 National Survey of Family Growth</td>
<td>171</td>
</tr>
<tr>
<td>A.14</td>
<td>Variances and design effects by time interval for the survivorship probabilities of subsample three taken from the 1973 National Survey of Family Growth</td>
<td>172</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.1</td>
<td>A contingency table of the number of events and withdrawals by time interval</td>
<td>33</td>
</tr>
<tr>
<td>3.2.1</td>
<td>A table giving counts of those who were exposed to risk and who either survived or died in the i-th time interval by domain</td>
<td>62</td>
</tr>
<tr>
<td>4.1</td>
<td>Data for the Koch, Johnson, and Tolley Approach</td>
<td>119</td>
</tr>
</tbody>
</table>
## LIST OF GRAPHS

<table>
<thead>
<tr>
<th>Graph</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Design Effect by Time Interval for the Conditional Probabilities of Subsample One Taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>47</td>
</tr>
<tr>
<td>2.2</td>
<td>Design Effect by Time Interval for the Conditional Probabilities of Subsample Two Taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>48</td>
</tr>
<tr>
<td>2.3</td>
<td>Design Effects by Time Interval for the Conditional Probabilities of Subsample Three Taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>49</td>
</tr>
<tr>
<td>2.4</td>
<td>Design Effects by Time Interval for the Conditional Probabilities of Subsample Four Taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>50</td>
</tr>
<tr>
<td>2.5</td>
<td>Design Effects by Time Interval for the Conditional Probabilities of Subsample One Taken From the 1973 National Survey of Family Growth</td>
<td>51</td>
</tr>
<tr>
<td>2.6</td>
<td>Design Effects by Time Interval for the Conditional Probabilities of Subsample Two Taken from the 1973 National Survey of Family Growth</td>
<td>52</td>
</tr>
<tr>
<td>2.7</td>
<td>Design Effects by Time Interval for the Conditional Probabilities of Subpopulation Three Taken from the 1973 National Survey of Family Growth</td>
<td>53</td>
</tr>
<tr>
<td>2.8</td>
<td>Design Effects by Time Interval for the Survivorship Probabilities of Subsample One Taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>54</td>
</tr>
<tr>
<td>2.9</td>
<td>Design Effects by Time Interval for the Survivorship Probabilities of Subsample Two Taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>55</td>
</tr>
<tr>
<td>2.10</td>
<td>Design Effects by Time Interval for the Survivorship Probabilities of Subsample Three Taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>56</td>
</tr>
<tr>
<td>2.11</td>
<td>Design Effects by Time Interval for the Survivorship Probabilities of Subsample Four Taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>57</td>
</tr>
<tr>
<td>Graph</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>-----------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>2.12</td>
<td>Design Effects by Time Interval for the Survivorship Probabilities of Subsample One Taken from the 1973 National Survey of Family Growth</td>
<td>58</td>
</tr>
<tr>
<td>2.13</td>
<td>Design Effects by Time Interval for the Survivorship Probabilities of Subsample Two Taken from the 1973 National Survey of Family Growth</td>
<td>59</td>
</tr>
<tr>
<td>2.14</td>
<td>Design Effects by Time Interval for the Survivorship Probabilities of Subsample Three Taken from the 1973 National Survey of Family Growth</td>
<td>60</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Conditional Probabilities by Time for Subpopulations One and Two Taken from the 1975 World Fertility Survey of Sri Lanka</td>
<td>88</td>
</tr>
<tr>
<td>3.7.2</td>
<td>Survivorship Probabilities by Time for Subpopulations One and Two Taken from the World Fertility Survey of Sri Lanka</td>
<td>89</td>
</tr>
<tr>
<td>3.7.3</td>
<td>Conditional Probabilities by Time for Subpopulations One, Two and Three Taken from the World Fertility Survey of Sri Lanka</td>
<td>90</td>
</tr>
<tr>
<td>3.7.4</td>
<td>Survivorship Probabilities by Time for Subpopulations One, Two and Three Taken from the World Fertility Survey of Sri Lanka</td>
<td>91</td>
</tr>
<tr>
<td>3.7.5</td>
<td>Conditional Probabilities by Time for Subpopulations One and Two Taken from the 1973 National Survey of Family Growth</td>
<td>92</td>
</tr>
<tr>
<td>3.7.6</td>
<td>Survivorship Probabilities by Time for Subpopulations One and Two Taken from the 1973 National Survey of Family Growth</td>
<td>93</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION AND REVIEW OF THE LITERATURE

1.1. Introduction

The life table is one of the oldest statistical methods for measuring mortality or for the study of any event which has an associated waiting time. The life table is essentially a method which combines the experience of individuals from several age or time intervals into a statistical model depicting the occurrence of some event (death) over time (Shryock, Siegel, and Associates, 1971).

Currently much information is obtained using complex sample surveys. The terminology complex sample survey is used to refer to probability samples other than a simple random sample with replacement. Specifically, this type of survey involves several stages of selection, stratification and clustering. When analyzing life tables from complex survey data, several statistical problems are encountered. These problems include variance estimation for the life table functions, and the construction of appropriate statistical tests for comparing life tables. In this dissertation a methodology for life table analysis of complex survey data is presented.

1.2. The Nature, Purpose, and Types of Life Tables

This review of life tables will cover definitions and notations, techniques of life table construction and the distributions of the various life table functions.
An important distinction between life tables is determined by the data source. The data source dichotomizes life tables into the actuarial/demographic type constructed from vital statistics data, and the follow-up life table constructed from clinical and laboratory data or survey data. Life tables are also classified as current or cohort. The cohort life table records the experience of an actual cohort of individuals (persons with a common background experience). An example of such a life table is the generation life table which summarizes the mortality experience of an entire birth cohort. Most follow-up life tables are of the cohort form. This cohort nature is developed from the view that the individuals under study are followed from one point in time to another later point. The cohort status develops from the view of the individuals entering the study at the same point in time. The current life table, also called the periodic life table, summarizes the experience of a cross-section of many generations or cohorts. The terms unabridged and abridged (complete and incomplete) refer to the length of the age or time intervals in which the data are given. Unabridged generally refers to single units of time or age, and abridged refers to intervals greater than unity. Often when follow-up study life tables are discussed, the terms grouped and ungrouped data are used in place of abridged and unabridged (Shryock, Siegel, and Associates, 1971; Chiance, 1968).

1.3. Notation and Definitions

In this section some commonly used notation is presented. One difficulty is that the notation for the actuarial/demographic life table is different from the notation used in the study of survivorship by
statistics. Both sets of notation will be given, and the inter-
relationship between the two sets of notation will be provided.

1.3.1. Actuarial/Demographic Life Table Notation

The notation and definitions presented here are those given by
Shryock, Siegel, and Associates (1971). In this type of life table
the event of interest is death. Usually the life table is viewed as
the experience of an actual or synthetic cohort. The initial size of
the cohort is termed the radix. Definitions of the life table functions
based upon the variable \( t \) denoting age are as follows.

\[ [t, t+h): \] The interval between two exact ages. The length of
the interval is \( h = t+h-t \).

\[ h^q_t: \] The proportion of individuals alive at age \( t \) who will
die in the interval \([t, t+h)\). These values can also
be viewed as the conditional probability of dying in
the interval \([t, t+h)\).

\[ l_t: \] The number of individuals alive at the exact age \( t \).

\[ d_t: \] The number of individuals who die in the interval
\([t, t+h)\).

\[ l_t^t: \] The number of person years lived in the interval
\([t, t+h)\).

\[ T_t: \] The total number of person years lived after age \( t \).

\[ \bar{e}_t: \] The average remaining lifetime for an individual of
age \( t \).

Other life table functions not always given in the body of the life
table are:

\[ m_t = \frac{d_t}{l_t^t}: \] The life table death rate for the interval
\([t, t+h)\).

\[ p_t = 1 - h^q_t: \] The proportion of individuals alive at \( t \) who
do not die in the interval \([t, t+h)\).
1.3.2. Notation from Survivorship Analysis

A standard set of notation does not exist in the field of survivorship analysis. However, the notation given here is that which is most frequently used.

Here $T$ denotes a non-negative random variable denoting time of death, or more generally the time of the event of interest. The random variable $T$ is assumed to have an absolutely continuous probability distribution. Definitions of interest are:

The probability distribution function of $T$.

$$ F(t) = P(T \leq t) . $$

The probability density function of $T$.

$$ f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} = \frac{d}{dt} F(t) . $$

The survivorship function of $T$.

$$ S(t) = P(T > t) = 1 - F(t) . $$

The hazard rate function, or force of mortality

$$ \mu(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} $$

$$ = \frac{f(t)}{S(t)} = - \frac{d}{dt} \log(S(t)) . $$

Relationships which can be obtained from the above are:

$$ S(t) = \exp(- \int_0^t \mu(x)dx) , $$

and

$$ f(t) = \mu(t) \exp(- \int_0^t \mu(x)dx) . $$
1.3.3. Interrelatedness of the Two Sets of Notation

Suppose the radix is taken to be unity, then for the interval \([t, t+h)\) the following relationships hold.

1. \( \ell_t = \ell(t) = S(t) = P(T > t) \).

2. \( h^d_t = \ell(t) - \ell(t+h) = \int_0^h \mu(t+x) \ell(t+x) dx = \int_0^h f(t+x) dx \).

3. \( h^L_t = \int_0^h \ell(t+x) dx \).

4. \( h^m_t = \frac{h^L_t}{h^L_t} = \frac{\int_0^h \mu(t+x) \ell(t+x) dx}{\int_0^h \ell(t+x) dx} = \frac{\int_0^h f(t+x) dx}{\int_0^h S(t+x) dx} \).

5. \( h^q_t = h^d_t / \ell(t) = (S(t) - S(t+h))/S(t) = 1 - \frac{S(t+h)}{S(t)} \),

and

\( h^p_t = 1 - h^q_t = \frac{S(t+h)}{S(t)} \).

6. \( \ell_0 = \int_0^\infty \ell(x) dx \).

and

7. \( \ell_t = \int_t^{\infty} \ell(x) dx / \ell(t) \).

1.4. Life Table Construction: A Review

In the following the focus of attention will be the life table functions \( h^q_t \) and \( S(t) \). The reason for examining these functions is that all of the remaining life table functions can be derived from either one.
1.4.1. Methods for Actuarial/Demographic Life Tables

Numerous methods are available for constructing an actuarial/demographic life table. Some well known methods are those of Reed and Merrel (1939), Greville (1943), Chiang (1968, 1972), Fergany (1971), and Keyfitz and Frauenthal (1975).

When constructing an actuarial/demographic life table, two problems are encountered. These problems concern the estimation of $h^q_t$ and $h^l_t$. The focus here will be on $h^q_t$. The reason being that in many situations $h^l_t$ is not of particular interest, whereas $h^q_t$ and $\lambda(t)$ or $S(t)$ are usually of interest.

The work of Keyfitz and Frauenthal (1975) is presented first since the methods of Reed and Merrel (1939), and Greville (1943) can be derived as special cases of their approach.

Keyfitz and Frauenthal begin by assuming that the data available are the observed number of deaths and the mid-year population for each age interval. The data are denoted by the pair $(h^D_t, h^K_t)$, $t=1, \ldots, k$. The observed central death rate is obtained from each pair using

$$h^M_t = \frac{h^D_t}{h^K_t}.$$  

The central assumption of their approach is that the underlying mortality curve should be the same for both the observed population and the stationary life table population. Using this assumption they write

$$h^M_t = \frac{\int_0^h K(t+x)\mu(t+x)dx}{\int_0^h K(t+x)dx} \quad (1.4.1)$$

The use of $K(t)$ as opposed to $\lambda(t)$ is to place emphasis on the point
that the observed population and that of the life table are different. The underlying mortality, assumed common to both populations, is \( \mu(t) \).

From life table theory the following can be derived:

\[
\ell(t+h) = \ell(t) \exp\left(-\int_0^h \mu(t+x)dx\right). \tag{1.4.2}
\]

If \( h \) is small, equation (1.4.2) can be approximated by the Taylor's expansion about the point \( t + \frac{h}{2} \). Applying this technique and using several substitutions, Keyfitz and Frauenthal obtain

\[
\int_0^h \mu(t+x)dx \approx h \, M_t - \frac{h}{12h} K_t' \left( t + \frac{h}{2} \right) K_t' \left( t + \frac{h}{2} \right), \tag{1.4.3}
\]

where the prime denotes differentiation. The derivatives in equation (1.4.3) are approximated using the available data \( (hD_t, hK_t) \). Using such approximations, and substituting them into equation (1.4.3), the authors arrive at an expression for \( h^q_t \). This expression is

\[
h^q_t = 1 - \exp\left(-hM_t + \frac{h}{48h} (hK_t + hK_t)(hM_t + hM_t - hM_t)\right). \tag{1.4.4}
\]

Equation (1.4.4) is the major result of the Keyfitz and Frauenthal paper.

The result derived by T.N.E. Greville (1943) is

\[
\int_0^h \mu(t+x)dx = h \, m_t + \frac{h^3}{12} h_m^2 \left( \frac{d}{dt} \log(hm_t) \right). \tag{1.4.5}
\]

Greville's result can be obtained from that of Keyfitz and Frauenthal's as follows. The following substitutions are made into equation (1.4.3): \( \ell(t) \) for \( K(t) \) and the relationships

\[
\frac{d}{d(t+x)} \ell(t+x) = -\ell(t+x) \mu(t+x)
\]
ii. \( \frac{g(t+h/2)}{h} = \frac{1}{h} \)

iii. \( u(t+h/2) = h m_t \)

and

iv. \[ \frac{d}{dt} \left( \frac{h m_t}{m_t} \right) = \frac{d}{dt} \log(h m_t) \]

Greville's equation for \( h q_t \) is derived by expanding \( \log(h q_t) \), and is given by

\[ h q_t = \frac{1}{h + h m_t \left[ \frac{1}{2} + h \left( m_t - \frac{d}{dt} \log(m_t) \right) \right]} \quad (1.4.6) \]

Greville noted in his research that
\[ c = \frac{d}{dt} \log(h m_t) \]

was virtually constant through most age intervals and for most life tables. The constant \( c \) was found to usually lie in the interval (0.08, 0.104). Letting \( c = 0.096 \), and substituting into equation (1.4.5) produces the empirically derived estimator of \( h q_t \) found by Reed and Merrel (1939). This estimator is

\[ h q_t = 1 - \exp(h m_t + 0.008 h^3 m_t^2) \quad (1.4.7) \]

Fergany (1971) noted that if the intervals \([t, t+h)\) are relatively short allowing the force of mortality to be assumed constant, then the relationship

\[ h m_t = u(t) \]

can be obtained from equation (1.4.1). Fergany used this approximation to obtain the estimator

\[ h q_t = 1 - \exp(-h m_t) \quad (1.4.8) \]
This same result can be obtained by assuming that the population at risk remains constant during a short interval \([t, t+h]\). In many applications the researcher has little if any control over the interval length, and many investigators consider the typical five or ten year interval lengths too large to make the assumption of a constant force of mortality.

Chiang (1968, 1972) used the relationship

\[
h^L_t = h \lambda(t+h) + h^{a_t} d_t \tag{1.4.9}
\]

where

\[
h^{a_t} = \frac{\int_0^h x \lambda(t+x) \mu(t+x) dx}{\int_0^h \lambda(t+x) \mu(t+x) dx} \tag{1.4.10}
\]

Then by assuming that the observed death rate and the life table death rate were equal the estimator

\[
h^q_t = \frac{h \cdot M_t}{1 + (h \cdot h^{a_t}) M_t} \tag{1.4.11}
\]

is derived.

The only unknown factor on the right hand side of equation (1.4.11) is \(h^{a_t}\). Chiang provides estimates of these values obtained from vital registration data. His investigation indicated that the \(h^{a_t}\) are fairly constant over time and across populations. The generalization of this last finding can be questioned. Work by Coale and Demeny (1966) suggest that mortality patterns differ between geographic regions. Caution should then be used in the application of this method. Specifically, the set of \(h^{a_t}\) should be examined relative to the population under investigation. Chiang has provided a computing formula for the \(h^{a_t}\).
but this formula depends on a knowledge of the $h_{qt}$. Schoen (1978) studied the problem of estimating the $h_{at}$ from the available data. He developed a formula for calculating the $h_{at}$ based upon the observed central death rates, $h_{mt}$.

A final method to be discussed is that of reference to a standard life table. This method converts the $h_{mt}$ to the required $h_{qt}$ using the relationship between the $h_{mt}$ and $h_{qt}$ in an unabridged life table. The unabridged life table from which the relationship is obtained is called the standard life table. This approach is suitable only when the study population and the standard life table have a comparable mortality experience (Shryock, Siegel, and Associates, 1971). The method uses a conversion factor analogous to the $h_{at}$ used by Chiang (1968). The conversion factor is

$$h_{gt} = \frac{h}{h_{qt}} - \frac{1}{h_{mt}} , \tag{1.4.12}$$

where $h_{qt}$ and $h_{mt}$ are taken from the standard life table. It can be demonstrated that the quantity $h - h_{gt}$ represents the average amount of time lived in the interval $[t, \, t+h)$ by those who died in the interval. The equation for $h_{qt}$ is similar to that of (1.4.11) and is

$$h_{qt} = \frac{h \, h_{mt}}{1 + h_{qt} \, h_{mt}} . \tag{1.4.13}$$

Methods other than those presented here do exist, but those given are the most current and frequently used.
1.4.2. Methods for Follow-up Study Life Tables

The data for follow-up study life tables are obtained from a different source than those for the actuarial/demographic life table. This difference results in another type of problem. For the actuarial/demographic life table, the difficulty was how to pass from the observed mortality rates to the conditional probabilities. The problem in follow-up life tables is with incomplete data, but still centers on the $h_{qt}$. The term incomplete data refers to individuals for whom the period of follow-up is incomplete.

Three sources of incomplete data have been identified by Chiang (1968). These three sources are:

a. losses to follow-up: These are subjects for whom information should have been obtained, but for some reason removed themselves from the study.

b. withdrawals or censored observations: These are subjects who were admitted into the study between $t$ and $t+h$ years prior to the end of the study, and will therefore not have completed the entire interval at the end of the study.

c. deaths to competing risks: These are subjects who die from some other cause than the one of interest.

New notation is now given which will facilitate the incorporation of the incomplete data into the analysis. The notation given here is that used by Elandt-Johnson (1977).

Consider the period of follow-up as divided into $k$ time intervals of the form $[t_i, t_{i+1})$, $i=1,\ldots,k$, with $t_{i+1} - t_i = h_i$. Now define

$N_i$: The number of individuals alive at time $t_i$.

d$_i$: The number of all deaths in the interval $[t_i, t_{i+1})$.

c$_i$: The number of individuals due to withdraw in the interval $[t_i, t_{i+1})$.

d$^*_i$: The number of deaths which occur to the $c_i$ in the interval $[t_i, t_{i+1})$. 
\( W_i \): The number of the \( c_i \) who withdraw alive during the interval \([t_i, t_{i+1}]\).

Note that \((d_i - d'_i)\) is the number of deaths among the \((N_i - c_i)\) subjects who were not due for withdrawal from the interval. A common practice is to treat losses to follow-up as if they were withdrawals. This practice will be followed here. Thus, in the discussion which follows, \( W_i \) will refer to both those who withdrew and those who were lost to follow-up for an interval. It is implicitly assumed that those who were lost to follow-up were not undetected deaths. Deaths due to competing causes are not considered.

Most of the estimators which are presented make one of the following assumptions regarding the time of withdrawal.

i. All withdrawals take place at the midpoint of the interval.

ii. The distribution of withdrawals for a given interval is such that the expected value of the distribution of the time of withdrawal is the interval midpoint.

iii. The incomplete information can be ignored. The exposure time for a withdrawal is taken as the time up to one of the endpoints of the interval of withdrawal.

The estimators presented here are for grouped data: that is data for which the exact times of death and withdrawal are unknown. The information available is \((N_i, c_i, d_i, d'_i, W_i)\). Methods which do consider the exact times of the events will be given a brief discussion.

If there are no withdrawals in an interval, then the binomial estimator

\[
h_i q_{t_i} = \frac{d_i}{N_i}
\]

(1.4.14)

is used. This estimator also assumes that there is a constant
probability of death within an interval, and that the deaths are independent. Equation (1.4.14) will underestimate the probability of death if withdrawals occur. The underestimate occurs from the failure of the binomial estimator to consider the partial exposure of those who withdrew alive. This partial exposure requires that an adjustment be made to those exposed to risk in the interval. The actual number exposed will be less than the $N_i$ stated.

The following approach of Elandt-Johnson (1977) formalizes the problem of partial exposure. Let $T_{ij}$ be a random variable which represents the exact time of withdrawal for the $j$-th individual in the $i$-th interval. Then

$$O_{ij} = \frac{T_{ij} - t_i}{h_i}$$

is a random variable denoting the fraction of the interval for which the $j$-th person who withdrew alive is exposed. Hence,

$$O_i = \frac{W_i}{\sum_{j=1}^{W_i} O_{ij}}$$

is the total exposure of the $W_i$ individuals who withdrew alive in $[t_i, t_{i+1})$. The 'effective number exposed to risk' is then defined as

$$R_i = N_i - W_i + O_i = N_i - \sum_{j=1}^{W_i} (1 - O_{ij}) \quad (1.4.15)$$

The first estimator to be given is the actuarial estimator. This estimator is based upon the assumption (ii) of the previous section. Under this assumption

$$E(R_i) = E_i = N_i - \frac{1}{2} W_i$$
The actuarial estimator is

$$h_i^{q_t} = \frac{d_i}{E_i} = \frac{d_i}{N_i + \frac{1}{2}W_i}.$$  \hspace{1cm} (1.4.16)

This is the estimator considered by Berkson and Gage (1950) and by Cutler and Ederer (1958). If the distribution of time to death is assumed to be exponential and assumption (i) is used, then (1.4.16) is the maximum likelihood estimator (Elandt-Johnson, 1977). However, if assumption (ii) is made as well as the exponential distribution of time to death, then the following estimator of Chiang (1961) is obtained:

$$h_i^{q_t} = 1 - \frac{-d_i' + [d_i' + 4(2N_i - C_i)(2N_i + 1 + W_i)]^{1/2}}{2(2N_i - C_i)}. \hspace{1cm} (1.4.17)$$

This estimator becomes equal to the estimator (1.5.3) if $d_i' = 0$, because $W_i = C_i$ when $d_i' = 0$.

Elveback (1958) used assumption (ii) for withdrawals, but assumed a uniform distribution for the time to death. The resulting estimator using the method of maximum likelihood is

$$h_i^{q_t} = \frac{(2N_i + d_i - W_i) - [(2N_i + d_i - W_i)^2 - 8N_id_i]}{2N_i}. \hspace{1cm} (1.4.18)$$

Two other estimators discussed by Elandt-Johnson (1977) are

$$h_i^{q_t} = \frac{d_i}{N_i - \frac{1}{2}C_i}, \hspace{1cm} (1.4.19)$$

and

$$h_i^{q_t} = \frac{d_i - d_i'}{N_i - C_i}. \hspace{1cm} (1.4.20)$$

The estimator (1.4.17) will, except when $d_i' = 0$, over-estimate the conditional probabilities because the numerator will contain deaths to
those due to withdraw, but the denominator may be overly corrected for the partial exposure of those who withdrew alive. Equation (1.4.20) ignores withdrawals entirely.

Johnson (1976) has given approximate relationships between the estimators (1.4.16) and (1.4.18), and between (1.4.17) and (1.4.19). He shows that in most practical situations all four estimators will give very similar results. This finding of similarity has also been confirmed empirically by Elandt-Johnson (1977). Drolette (1975), Kuzma (1967), and Elandt-Johnson (1977) all provide discussion and examinations of the various estimators. These investigators have studied the bias for the estimators and all three report such bias as being small.

The previous estimators were all based on grouped data which did not consider the exact times of death and withdrawal. Estimators which use the information of exact times are now discussed.

Perhaps the most well known estimator for the conditional probability of an event, and the estimator for the survivorship probabilities is the Kaplan and Meier (1958) product limit estimator. This approach is based upon the creation of small time intervals \([t_i, t_{i+1})\) developed from the observed exact times of events, and the concept of a risk set. The risk set is the number of individuals at risk of death or some other event immediately prior to an observed event. In the previous notation this number was denoted by \(N_i\). Hence, if \(N_i\) individuals are at risk at time \(t_i-0\), and out of this \(N_i\), \(d_i\) die, then the Kaplan and Meier estimator is:

\[
h_{i|t_i} = \frac{d_i}{N_i}, \tag{1.4.21}
\]

which is the binomial estimator (1.4.14). The choice of intervals is
such that a death or withdrawal do not occur in the same interval. This construction of the intervals allows the simplicity of the estimator.

Other estimators based upon exact times of events are given by Kimball (1960), Little (1952) and Elveback (1958). Models based upon a known distributional form such as the Gompertz or Wiebull can also be employed for exact time data. Detailed discussion of using such model based approaches are given in Gross and Clark (1975) and Mann, Schafer and Singpurwalla (1974).

1.5. Estimates of the Survivorship Function

Recall that by definition the survivorship function is

\[ S(t) = S(t) = \prod_{j=1}^{t-h} (1 - h_{q_j}). \tag{1.5.1} \]

An estimate of \( S(t) \) is obtained by replacing \( h_{q_j} \) in equation (1.5.1) by its estimate. Estimates of \( h_{q_j} \) were given in section 1.4. Where a known parametric form of \( S(t) \) is assumed, the reader is referred to Gross and Clark (1975) for estimates. Throughout this dissertation the survivorship function will be denoted as \( S(t) \), \( S(t) \) or \( t^{P_0} \) depending upon the context.

1.6. Distribution of the Life Table Functions

In this section a brief discussion of the probability distributions of the life table functions \( S(t) \), \( h_{d_t} \) and \( e_t \) will be given.

The distributions given here are obtained under the assumption that the life table is a stationary population. This approach views the life table as a closed population with an equal number of births and deaths each year. The number of births and deaths each year is fixed and equals the radix \( e_0 \). Under this scheme the schedule of
probabilities, $h_{q_t}$, remains constant. An analogous situation is found in the study of Markov chains and invariant (stationary) distributions.

1.6.1. The Distribution of the Number of Survivors

The number of survivors at age $t$ is denoted by $l_t$. Although the $l_t$ are usually given for integral values of age, it is convenient to view $t$ as a continuous variate. The distribution of $l_t$ depends upon the force of mortality $u(t)$. The probability of surviving from birth to age $t$ is

$$ t^p_0 = \exp\left(- \int_0^t u(x)dx \right). $$

It can be demonstrated that the distribution of $l_t$ conditional on $l_0$ is binomial:

$$ P(l_t = c | l_0) = \binom{l_0}{c} t^p_0 (1-t^p_0)^{l_0-c}. \quad (1.6.1) $$

The expected value and variance are

$$ E(l_t | l_0) = l_0 t^p_0 $$

and

$$ \text{Var}(l_t | l_0) = l_0 t^p_0 (1-t^p_0). $$

The joint distribution of survivors, that is for the set $\{l_1, l_2, \ldots, l_k\}$ for a given age $k$ is

$$ P(l_1 = c_1, \ldots, l_k = c_k | l_0) = \prod_{i=0}^{k-1} \binom{c_i}{c_{i+1}} t^{c_{i+1}} (1-t^{c_{i+1}})^{c_i-c_{i+1}}, \quad (1.6.2) $$

$c_{i+1} = 0, 1, \ldots, c_i; c_0 = l_0$, (Chiang, 1968). This joint distribution is a product of several binomial distributions. The expected value of $l_t$, $t=1, \ldots, k$, is
\[ E(l_t | l_0) = l_0 \cdot t P_0, \]

and covariance for \( s \leq t, t = 1, \ldots, k \)

\[ \text{Cov}(l_t, l_s | l_0) = l_0 \cdot t P_0 (1 - s P_0). \]

1.6.2. Distribution of the Number of Deaths

The assumption of a stationary population gives

\[ \sum_{i=1}^{\omega} h_i d_i = l_0, \]

where \( d_t \) is the number of deaths in the interval \([t, t+h]\) and \( \omega \) is the final age. The probability of dying in the interval \([t, t+h]\) is \( t P_0 \cdot q_t \), which is the probability of surviving to age \( t \) and then dying in the specified interval. Under the conditions \( P_0 = 1 \), and \( q_\omega = 1 \), the joint distribution of the number of deaths is the multinomial distribution

\[ P(d_0 = c_0, \ldots, d_\omega = c_\omega | l_0) = \frac{l_0!}{c_0! \ldots c_\omega!} (P_0 \cdot q_0)^{c_0} \ldots (P_0 \cdot q_\omega)^{c_\omega}, \]

(Chiang, 1968). The expected values and covariances are:

\[ E(h_t | l_0) = l_0 \cdot t P_0 \cdot q_t, \]

and

\[ \text{Cov}(h_t, h_s | l_0) = -l_0 (t P_0 \cdot q_t) (s P_0 \cdot q_s), \]

for \( s \neq t \) and \( s, t = 1, \ldots, \omega \).

1.6.3. The Distribution of Life Expectancy

Define a continuous random variable \( Y \) representing future life time, and let \( y \) denote an observed value of \( Y \). For a person of age \( x \) the probability distribution function of \( Y \) is
\[ f(y) = \left[ \exp(-\int_x^{x+y} \mu(t) \, dt) \right] \lambda(x+y), \quad y \geq 0. \quad (1.6.4) \]

The expected value and variance of \( Y \) are

\[ E(Y|x) = e_x^Y = \int_0^\infty \exp(-\int_x^{x+y} \mu(t) \, dt) \, dy \]

and

\[ \text{Var}(Y|x) = \int_0^\infty (y - e_x^Y)^2 f(y) \, dy, \quad (\text{Chiang, 1968}). \]

Let \( Y_j, \, j = 1, \ldots, \ell_x \), denote the future lifetimes of \( \ell_x \) individuals. The \( Y_j, \, j = 1, \ldots, \ell_x \) are independently and identically distributed as \( (1.6.4) \).

The average future lifetime at age \( x \) is

\[ \bar{e}_x = \frac{1}{\ell_x} \sum_{j=1}^{\ell_x} Y_j, \]

and under the central limit theory (with the usual conditions on the moments) \( \bar{e}_x \) is normally distributed with mean, \( \bar{e}_x \), and variance \( \text{Var}(Y|x)/\ell_x \).

1.7. Variance Estimates for \( h_{qt} \) and \( \lambda(t) \)

Under the multinomial model for the distribution of deaths (section 1.6) the maximum likelihood estimator of \( h_{qt} \) is

\[ h_{qt}^* = 1 - \frac{\lambda(t+h)}{\lambda(t)}. \]

The variance of this estimator is

\[ \text{Var}(h_{qt}) = E(\frac{1}{\lambda(t)})h_{qt}(1-h_{qt}) \cdot \quad (1.7.1) \]

It can further be shown that the efficiency of this estimator with respect to the Cramer-Rao lower bound for variance is

\[ \left( \frac{1}{\ell_0 t_0} \right) / E(\frac{1}{\lambda(t)}) \quad (1.7.2) \]
Chiang (1968). Although (1.7.2) will generally not be unity, it can be shown that the estimator (1.7.1) is the minimum variance unbiased estimator.

The variance of the estimator for the survivorship function estimator, \( \hat{t}_0^o \), under the multinomial distribution is

\[
\text{Var}(\hat{t}_0^o) = \text{Var}\left(\sum_{j=0}^{t-h} (1-h^o_j) \right) = \sum_{j=0}^{t-h} \frac{h^p_j \cdot h^q_j}{\xi_j} + h^p_j - t^o_0^2. \tag{1.7.3}
\]

Equation (1.8.3) can be approximated as

\[
\text{Var}(\hat{t}_0^o) = t^o_0^2 \sum_{j=1}^{t-h} \frac{h^q_j}{h^p_j \xi_j}, \tag{1.7.4}
\]

(Greenwood, 1925).

1.8. Outline of Subsequent Chapters

In Chapter II statistical problems associated with the analysis of complex survey data are reviewed. In particular, the difficulty of computing variances for functions of the data are discussed. Methods for obtaining variance estimates are given and these methods are examined with respect to the life table functions \( h^p_t \) and \( t^o_0 \).

Chapter III reviews methods used for comparing the survivorship curves between two or more life tables. Using the variance estimators given in Chapter II, a modified Mantel-Haenszel (1958) test statistic for comparing the survivorship of life tables constructed from complex survey data is developed.

The fourth chapter deals with life table regression models developed by various authors. In this chapter a modification to one of these models is outlined. Problems of using life table regression models in complex survey data are discussed.
Chapter V continues the discussion begun in Chapter IV concerning life table regression in complex survey data. Arguments are presented under which the method of maximum likelihood estimation can be used for complex survey data. Variance estimates for the maximum likelihood estimators from complex survey data are given.

In Chapter VI a summary of the research of this dissertation is given, and suggestions for further research on related topics are made.
CHAPTER II

VARIANCE ESTIMATES OF LIFE TABLE FUNCTIONS
FROM A COMPLEX PROBABILITY SAMPLE

2.1. Introduction

Obtaining measures of precision for parameters estimated from survey data is an important aspect of a statistician's work. The use of complex probability samples has in some situations made this task difficult. The term complex probability sample is used as a name for those samples that employ several stages of selection, clustering, and stratification. The term complex survey will also be used for describing such samples. Analytic expressions for variances for only slightly complicated survey designs can become quite involved as McCarthy (1966) has shown. For many functions of the data, expressions for an exact variance cannot be written (Freeman, 1975; Kish and Frankel, 1974).

The estimates of the conditional probabilities of an event $h_0 t', h_{pt}$ and the survivorship probabilities $t_P$, or $S(t)$ in a life table, described in Chapter I, are non-linear functions of the sample data. Exact variances for such functions cannot be given. The term non-linear is used to describe any function $g(\cdot)$ of the sample data such that if $x$ and $y$ are data, then

1. $g(x+y) \neq g(x) + g(y)$

and 2. $g(cx) \neq cg(x)$, for some constant $c$.

The relationships (1) and (2) are equalities for linear functions.
The purpose of this chapter is to examine several methods for computing variances for random variables generated by a complex survey. These methods will then be used to obtain variance estimates of life table functions.

Four methods of variance estimation are examined. The four methods are: balanced repeated replications, or balanced half samples (BRR); the Taylor's series linearization (TL); the jackknife repeated replication (JRR); and a second jackknife method (JK).

In the following sections each of the four methods will be reviewed, and applied to data. The data sets used are the 1973 National Survey of Family Growth (NSFG) (French, 1978) and the 1975 World Fertility Survey of Sri Lanka (WFS) carried out by the Department of Census and Statistics, Sri Lanka, in association with the World Fertility Survey, London, and the International Institute of Statistics.

Comparison of the four methods regarding their use for computing variances of life table functions will be made on the basis of the results obtained from the applications, and a set of criteria. The specified criteria are:

1. The ease with which the method can be implemented, and the computing cost.
2. The necessary assumptions required for using the method.
3. The theoretical properties of the variance estimate.

2.2. Problems With Variance Estimation

Statisticians have always faced the problem of estimating the variance of a function of a random variable even when the data were obtained through simple random sampling. However, a new facet to the problem is introduced when the data are from a complex probability
sample. In the case of a complex probability sample the statistician must address not only the issue of estimation, but also must consider how the sampling design effects the process of estimation.

The difficulty is that in most complex probability samples the classical theory of statistical estimation and of inference do not apply. Most of statistical theory relies upon the simple random sample with replacement, and does not easily generalize to the case where a dependence relationship exists among the observations as in a multi-stage cluster sample.

The assumption of a simple random sample with replacement (SRSWR) for data with a covariance among the observations can often lead to incorrect estimates and inference statements (Walsh, 1947; Rao and Scott, 1979), the most common errors being an exaggerated level of significance for statistical tests, and underestimates of the variance.

A simple illustration of this problem is as follows: Suppose that an analyst has a probability sample denoted by S, with observations on units given by \( x_j \), \( j=1,\ldots,n \) which have a non-zero covariance. If the analyst sums the observations, then the true variance of \( \sum_{j=1}^{n} x_j \) is

\[
\text{Var}(\sum_{j=1}^{n} x_j) = \sum_{j=1}^{n} \text{Var}(x_j) + 2 \sum_{i>j}^{n} \text{Cov}(x_i, x_j).
\]  

(2.2.1)

Whereas, if the observations are assumed independent, then the variance is

\[
\text{Var}(\sum_{j=1}^{n} x_j) = \sum_{j=1}^{n} \text{Var}(x_j).
\]  

(2.2.2)

It is easily seen that any statistical inference based upon the independence assumption will result in a type one error either larger
or smaller than those specified by the analyst, and that the variance assuming independence will be incorrect by a factor given by the covariance term on the right hand side of (2.2.1).

As stated, the most common error is an exaggerated nominal significance level (type one error), and an underestimate of the variance. This set of problems results because the covariances given in (2.2.1) are positive in most situations.

A word needs to be said here about the nature of the covariance term in equation (2.2.1). This covariance term for survey data gathered by a probability sample is usually viewed as a measure of homogeneity among the sample observations within clusters (see Kish, 1965). The covariance is usually positive since the observations within a given cluster are generally more alike than observations selected at random from the entire population.

The critical problem which the researcher faces is how to obtain an estimate of the variance which will incorporate the effects induced by the selection procedure. For those estimates which are linear functions of the sample data as defined in section (2.1), for example means, and totals, an exact variance estimate incorporating the sampling design's effect can be found. However, for functions of the sample data which are non-linear, for example ratios, and regression coefficients, exact variances cannot be computed. For the non-linear case, some approximation must be used.

2.3. A Review of Variance Estimation Methods

The four methods mentioned in section (2.1) will be reviewed. These four methods are balanced repeated replication or balanced half-
samples (BRR), the Taylor series linearization (TL), the jackknife repeated replication (JRR), and a second jackknife (JK).

2.3.1. Balanced Repeated Replications

The BRR was developed by McCarthy (1966), and is related to the concept of repeated sampling. The BRR is a design dependent method, since it requires a sampling scheme of two independent units per stratum. This requirement is not a serious drawback, because most samples can be made to conform to this design by some method of re-stratification. The term re-stratification is used to mean any reorganization of the original sample data to achieve the two unit per stratum design. In some cases this will involve the pairing of units from strata with more than two units, and in other cases it will consist of collapsing the original strata when only one unit is in a stratum. Some care must be exercised when re-stratifying a sample, because if the measurements of the variable under study differ widely from stratum to stratum, then a serious bias can be introduced into the variance (Stanek and Lemeshow, 1978).

The BRR utilizes a number of half-samples consisting of one of the two units per stratum. The statistic of interest is then evaluated for each of the half-samples, and the average squared deviation of the half-sample estimates from the estimate based upon the entire sample is used as an estimate of the variance.

If we assume that there are \( H \) strata of two primary sampling units (PSU) per stratum, then the number of independent half-samples is \( 2^H - 1 \), which can be very large and unmanageable. McCarthy (1966) investigated using subsets of the set of all possible half-samples, and noted that randomly selecting the half-samples resulted in a considerable
'between-strata-component' to the variance. These between-strata-components did have different signs and cancel one another when all the possible half-samples were used. Using this information McCarthy then devised an efficient technique to choose a subset of the half-samples. This technique is based upon a matrix with orthogonal columns whose elements are either 0 or 1. A method for constructing such matrices is described by Plackett and Burman (1943). The rows of the matrix identify a given half-sample while the columns refer to strata. It is from the orthogonal matrix that the term 'balanced' is derived, for in each subset of half-samples each of the two PSUs within a given stratum appear an equal number of times. This balance causes the between-strata-components of variance to cancel. In order to insure such a balance the number of half-samples, denoted by \( L \), must be a multiple of 4 which is greater than or equal to \( H \), the number of strata.

As an example, if for a particular sample, there were five strata each with two PSUs, and the element 1 in the matrix represents inclusion of the first PSU of a stratum, and 0 inclusion of the second PSU, then an 8x8 Plackett and Burman matrix would be necessary since 8 is the smallest number which is a multiple of 4 which is greater than 5. From this matrix we would pick any set of 5 columns. The elements of the matrix would specify which PSUs to include in each half-sample. McCarthy (1966) gives several examples in his paper.

Kish and Frankel (1974) and Bean (1975) have carried out considerable empirical research of the BRR technique, and Krewski and Rao (1980) have investigated the theoretical aspects of this method.

A computing formula for the BRR as given by Kish and Frankel is

\[
\text{Var}(g(s)) = (2L)^{-1} \sum_{m=1}^{L} \{(g(H_m) - g(S))^2 + (g(C_m) - g(S))^2\} \quad (2.3.1)
\]
where
\[ g(S) : \text{is the function of the sample data } S \text{ for which we want} \]
\[ \text{a variance estimate.} \]
\[ H_m : \text{is the } m\text{-th half-sample.} \]
\[ C_m : \text{is the complement half-sample of those PSUs not in } H_m. \]

A sampling fraction can be introduced into equation (2.3.1) if necessary.

Kish and Frankel and Bean point out that both terms on the right hand side of (2.3.1) could be used individually as an estimate of the variance, but they suggest using both to add stability to the estimator.

A part of the computation for BRR not always stated is the adjustment of the sample weights for each half-sample in order that they sum to some specified total. Simmons and Baird (1968) give some empirical evidence, for ratios, that such an adjustment is not always necessary. However, Lemeshow (1979) has demonstrated using simulated data that if the statistic of interest varies widely from stratum to stratum, or variances are required for small domains within the sample, then not adjusting the weights can lead to more bias and variability in the BRR variance estimate. In this paper the weights will not be adjusted for each half-sample, since for the most part the statistics of interest are a type of ratio.

Empirical evidence suggests that the BRR performs well (Kish and Frankel, 1970, 1974; and Bean, 1975). Recent work by Krewski and Rao (1980) has established the consistency of the BRR variance estimator within the context of a sequence of finite populations \( \{P_H\} \) with the number of strata \( H \to \infty \).
2.3.2. The Taylor Series Linearization

The Taylor's series has long been used as a method for computing the variance of a function of a random variable (Kendall and Stuart, 1963, Vol. 1). Use of the Taylor's linearization (TL) for computing the variances of functions of data from complex probability samples has been given attention by Kish and Frankel (1974), Woodruff (1971), Tepping (1968), and Shah (1977). The TL is frequently used for computing the variance of a ratio (viz. Kish, 1965 or Cochran, 1963).

In a simplified form, for a random variable $X$ and a continuous function $g(X)$, $g$ is expanded in a Taylor series about the point $E(X) = \mu$ as

$$g(X) = g(\mu) + g'(\mu)(X-\mu),$$

prime denoting differentiation with respect to $X$. It can be shown that

$$E(g(X)) = g(\mu),$$

and

$$\text{Var}(g(X)) = g'(\mu)^2 \text{Var}(X). \quad (2.3.2)$$

The equation (2.3.2) points out two necessary requirements for using this method. The first is the tractability of the derivatives. Certain functions like multiple and partial correlations are difficult at this point. However, in this case approximations are available (Woodruff and Causey, 1976; and Tepping, 1968). The second requirement is an estimate of $\text{Var}(X)$, preferably unbiased. In many cases, $X$ is an estimated total based upon the weighted sample observations.

A more general situation occurs when $g$ has a vector argument. For this case equation (2.3.2) becomes

$$\text{Var}(g(X)) = \sum_{r=1}^{q} \left( \frac{\partial g}{\partial x_r} \right)^2 \text{Var}(X_r) + 2 \sum_{r<s}^{q} \left( \frac{\partial g}{\partial x_r} \frac{\partial g}{\partial x_s} \right) \text{Cov}(X_r,X_s). \quad (2.3.3)$$
The derivatives in equation (2.3.3) are evaluated at \( E(\chi) \), and \( q \) is the dimension of \( \chi \). A difficulty with this approach is the covariance term on the right hand side. If \( q \) is large, there will be a large number of covariance terms the exact number being \( \binom{q}{2} \). Woodruff (1971) has developed an algorithm for computing the variance of \( g \) without directly computing the variance and covariance terms in (2.3.3).

Shah (1977), Bean (1975), and Kish and Frankel (1974) have studied the TL empirically and found it to perform well. Krewski and Rao (1980) have shown that the TL provides consistent estimates for most multi-stage designs.

Kish and Frankel (1974) give the following formula for the TL for the two unit per stratum design.

\[
\text{Var}(g(s)) = \sum_{h=1}^{H} \left\{ \frac{8}{\sum_{r=1}^{n_{h1}} \sum_{j=1}^{n_{h2}} \left( \frac{\partial g}{\partial x_{hrji}} \right)^2} \right\} \cdot \frac{8}{\sum_{r=1}^{n_{h2}} \sum_{j=1}^{n_{h2}} \left( \frac{\partial g}{\partial x_{hrj2}} \right)^2} \right. \cdot \frac{8}{\sum_{r=1}^{n_{h1}} \sum_{j=1}^{n_{h1}} \left( \frac{\partial g}{\partial x_{hrj1}} \right)^2} .
\]  

(2.3.4)

For this equation:

\( S \) represents the sample data which in this case will be a \( q \) vector of sums over strata and individuals;

\( H \) is the number of strata in the sample;

\( n_{h1} \) and \( n_{h2} \) are the number of elementary observations in the two primary units per stratum;

and \( x_{hrji} \) is the value of the \( r \)-th variable \((r=1, \ldots, q)\) for the \( i \)-th \((i=1,2)\) PSU within the \( h \)-th stratum for the \( j \)-th elementary unit.
2.3.3. The Jackknife

The jackknife is an ad hoc statistical method which has proved useful for a variety of problems. Originally introduced as a technique for reducing bias in an estimator by Quenoulli (1956), the jackknife has since become an invaluable method for estimating variances and setting confidence limits (Miller, 1968, 1974).

Two forms of the jackknife have been used for estimating variances in complex surveys. The first was given by McCarthy (1966) will be denoted in this paper by JK. The second jackknife is due to Kish and Frankel (1974) and is called the jackknife repeated replication (JRR). The JK and JRR were developed for designs with two PSUs per stratum. However, the jackknife concept is quite general and could be applied to other sampling designs although specific formula for such designs have not appeared in the literature.

The procedure for using the JRR is to form subsamples by dropping one PSU from a specified stratum. But the PSU which remains has its observations counted twice. This process is followed until all PSUs have been excluded once. The result is the creation of $2H$ subsamples since there are $H$ strata and $2$ PSUs per stratum.

For each subsample the statistic of interest, $g$, is evaluated. The JRR variance estimate is

$$\text{Var}(g(S)) = (2H)^{-1} \sum_{h=1}^{H} \sum_{i=1}^{2} (g(S_{hi}) - g(S))^2,$$  \hspace{1cm} (2.3.5)

where

$g(S)$ is the function evaluated for all the sample data,

and $g(S_{hi})$ is the function evaluated for the subsample formed by dropping the $i$-th PSU from the $h$-th stratum.
The JK is developed through the use of pseudo-values. These
pseudo-values, \( g_{-hi} \) are formed, using the above notation, as
\[
g_{-hi} = 2H g(S) - (2H-1)g(S_{hi}) .
\]
Letting
\[
g^* = (2H)^{-1} \sum_{h=1}^{H} \sum_{i=1}^{2} g_{-hi}
\]
the variance estimate for \( g(S) \) is
\[
\text{Var}(g(S)) = (2H)^{-1} \sum_{h=1}^{H} \sum_{i=1}^{2} (g(S_{hi})-g^*)^2 . \tag{2.3.6}
\]

The JRR and JK have been studied by Kish and Frankel (1974) and
theoretically by Krewski and Rao (1980).

Kish and Frankel (1974) have studied the JRR, BRR and TL for
ratios, means, zero order correlations and regression coefficients.
Their investigation showed that the three methods gave similar results.
Independent work by Bean (1975) has confirmed the results. All of the
authors noted that the TL was more stable than either the BRR or JRR,
but the BRR performed better at setting confidence intervals than the
TL or JRR. These findings have also been reported by Lemeshow and
Levey (1978).

A variety of investigators have found that for several commonly
used statistics the four methods, BRR, JRR, TL, and JK, provide similar
results. However, there are no empirical or theoretical works indicat-
ing how the methods will perform for life table functions.

2.4. Variance Estimates of the Life Table Functions

The purpose of this chapter is to examine the performance of the
BRR, JRR, JK, and TL for computing variances of the life table functions.
the conditional probabilities $h_t^q$, $h_t^p$, and the survivorship probabilities $q_0(t)$ or $S(t)$.

The discussion will be confined to life tables constructed from retrospective data, since this is the type of data most frequently encountered for life table computations from survey data. The estimation process for the life table will be outlined to provide a basis for better understanding the life table methods and problems encountered with complex survey data.

The first step for constructing a life table from retrospective data is to obtain a contingency table which specifies how many events, such as death or birth, and how many withdrawals or censored observations occurred within each time interval $[t, t+h)$. A typical contingency table is that given by Figure 2.4.1. In Figure 2.4.1 the variable denoting time takes on the values $t=1, \ldots, k$ and $h=1$.

Figure 2.4.1: A contingency table of the number of events and withdrawals by time interval.

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Number of Events</th>
<th>Number of Withdrawals</th>
</tr>
</thead>
<tbody>
<tr>
<td>$[0,1)$</td>
<td>$E_1$</td>
<td>$W_1$</td>
</tr>
<tr>
<td></td>
<td>$\cdot$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td></td>
<td>$\cdot$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td></td>
<td>$\cdot$</td>
<td>$\cdot$</td>
</tr>
<tr>
<td>$[k-1,k)$</td>
<td>$E_k$</td>
<td>$W_k$</td>
</tr>
</tbody>
</table>
The entries in Figure 2.4.1 are obtained by summing and possibly weighting the sample observations. Suppose that there are $n$ sample observations indexed by $j=1,\ldots,n$, and define the following indicator variables:

$$I(E,t,j) = \begin{cases} 
1 & \text{if the } j\text{-th individual experiences the event in } [t,t+h) \\
0 & \text{otherwise}
\end{cases}$$

$$I(W,t,j) = \begin{cases} 
1 & \text{if the } j\text{-th individual is censored in } [t,t+h) \\
0 & \text{otherwise}
\end{cases}$$

Also, let $w_j$ denote the sample weight associated with the $j$-th individual in the sample, often

$$w_j = \Pr(\text{j-th unit is selected})^{-1}.$$  

Using these definitions

$$E_t = \sum_{j=1}^{n} I(E,t,j)w_j$$

and

$$W_t = \sum_{j=1}^{n} I(W,t,j)w_j.$$  

The second step in constructing a life table from survey data is to compute the number of individuals exposed to risk at the beginning of each time interval. These numbers denoted by $N_t$ are computed as

$$N_t = \sum_{s=t}^{k} (E_s + W_s).$$

Using the actuarial estimator for the conditional probability of an event (see equation 1.4.16, chapter I) the 'effective exposed to risk' is computed

$$R_t = N_t - \frac{1}{2} W_t.$$
and the conditional probability of an event is given by

\[ h^q_t = E_t / R_t. \]

In terms of the sample data the conditional probability is

\[ h^q_t = \frac{\sum_{j=1}^{n} I(E_t, t, j) w_j}{\sum_{s=t}^{k} \sum_{j=1}^{n} (I(E_t, t, j) + \frac{1}{2} I(W_t, t, j)) w_j}. \]  

(2.4.1)

The survivorship probabilities are found using equation 1.6.1 of chapter I; specifically

\[ t^p_0 = S(t) = \prod_{s=0}^{t-h} (1 - h^q_s). \]

If each time interval is viewed as a domain, then the \( h^q_t \) can be seen as a type of ratio estimator, and \( t^p_0 \) or \( S(t) \) are then products of such ratios. The life table functions are non-linear functions of the sample data as defined in section 2.1 of this chapter.

Application of the four variance estimation procedures to the above life table functions \( h^q_t \) and \( t^p_0 \) is now outlined. For the BRR the appropriate Plackett-Burman matrix which identifies the half-samples must be determined. When this matrix is obtained, then a life table is constructed for each half-sample, and for the entire set of data. The variance estimates are found using equation 2.3.1.

A similar procedure is used for the JRR and JK variance estimates. Here a life table is constructed for each subsample formed by removing a specified PSU from a given stratum. Formulae 2.3.5 and 2.3.6 are then used for the JRR and JK respectively. The JK requires the computation of pseudo-values as defined in section 2.3.3.
The TL variance estimates are found by computing $E_t, W_t,$ and $N_t$ for each PSU within each stratum. These PSU totals are then substituted into equation 2.3.4. The appropriate partial derivatives must also be computed and substituted into equation 2.3.4.

Each of the four methods, BRR, JRR, JK, and TL, requires extensive manipulation of the sample data, and the use of a computer is necessary. The programs used for the application of the four methods were written specifically for the problems examined here. General programs for the BRR and TL are available from various authors. A reference containing the source of such programs is Namboodiri (1978).

2.5. **The Data Used for This Study**

Two sets of data were used for examining the four variance estimation techniques. The two data sets were the 1975 World Fertility Survey of Sri Lanka (WFS) and the 1973 National Survey of Family Growth (NSFG).

Because interest is often directed at various subgroups, domains, or subpopulations within the survey data, the two data sets were divided into several smaller data sets. The smaller data sets were used to examine how the four methods of estimation performed in smaller domains or subpopulations.

The WFS survey of Sri Lanka consisted of 6,810 ever-married women in 606 PSUs which were contained in 17 primary strata defined on the basis of rural, urban, or estate type dwelling. The sample was then re-stratified by the WFS survey team into a design consisting of 303 strata each with two PSUs. This re-stratification was accomplished by pairing consecutively drawn PSUs in the sample. The original method of
selecting the PSUs, census blocks in this case, was systematic sampling from an available file of the census blocks. The pairing of consecutively drawn PSUs was performed in this case because it resulted in strata within which the PSUs were similar: due to the file being in geographic/location order.

Within the country of Sri Lanka, six domains of interest were identified by the survey team. These domains were termed zones, and were defined using regional and socio-economic characteristics of the zones. The original sample design can be viewed as six geographic/regional zones within which strata were defined by the type of dwelling place, urban, rural, or estate. Within the strata the PSUs or census blocks, were chosen systematically; within the census blocks the households for the sample were also selected systematically. All ever-married women within a household were interviewed.

The performance of the four methods, BRR, JRR, JK, and TL has been studied for ratios based upon the entire sample data, interest here is focused on the performance of the four methods for domains or sub-populations of the data. To examine the performance of the four methods in these smaller subsamples, only two geographic zones from the WFS were used. Zone one contained the capital city of Colombo, and zone two comprised the land area surrounding the capital city. Both zones are located in the southwest of Sri Lanka along the coast. Zone one consisted of four original strata and 98 PSUs. The 98 PSUs were later re-stratified into 49 psuedo-strata. Zone two contained three original strata, and the PSUs in this zone were also re-stratified.

Using these two zones four subsamples were formed. These subsamples were:
Subsample one: all four original strata of zone one, and is identical with zone one.

Subsample two: the original strata one and two of zone one, and the original strata one of zone two. These were all urban strata.

Subsample three: the original stratum one of zone one.

Subsample four: the original stratum two of zone one.

The four subsamples contained mostly urban population. Table 2.5.1 provides information regarding the subsamples.

Table 2.5.1. Some characteristics of the four subsamples formed from zones one and two of the 1975 World Fertility Survey of Sri Lanka

<table>
<thead>
<tr>
<th>Subsample</th>
<th>Number of Elementary Units</th>
<th>Number of PSUs</th>
<th>Number of Original Strata</th>
<th>Number of Pseudo-Strata</th>
<th>Spread of Sample Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>927</td>
<td>98</td>
<td>4</td>
<td>49</td>
<td>439-497</td>
</tr>
<tr>
<td>Two</td>
<td>575</td>
<td>64</td>
<td>3</td>
<td>32</td>
<td>443-1743</td>
</tr>
<tr>
<td>Three</td>
<td>275</td>
<td>28</td>
<td>1</td>
<td>14</td>
<td>443</td>
</tr>
<tr>
<td>Four</td>
<td>80</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>480</td>
</tr>
</tbody>
</table>

Although the selection process for the Sri Lanka WFS was probability proportional to size systematic sampling (Kish, 1965), for the purposes of this analysis, it will be assumed that the PSUs were selected independently within a pseudo-strata.

The event of interest for the WFS subsamples was the time from marriage until first birth. Women with pre-marital births were excluded from the analysis.

The 1973 NSFG is a five stage cluster sample with counties and independent cities comprising the PSUs. The sample contained 9,797
women aged 15-49 in 103 PSUs. The PSUs were re-stratified into 48 psuedo-strata with each stratum containing two PSUs. This re-stratification was accomplished as follows. Seven of the 48 strata were self-representing; that is, they came into the sample with certainty. Within these seven strata the data were combined into two psuedo PSUs. The other 41 strata were formed from one or more of the original PSUs, but all the original PSUs used to construct the psuedo-strata came from the same locality (see French, 1978 for more details). The final design available for analysis was that of 48 strata each containing two PSUs. However, neither the PSUs nor the strata of this final design were those of the original sample design used to actually obtain the data.

To investigate the behavior of the four variance computing methods the NSFG was divided into three subsamples. Two of these subsamples were identified on the basis of the black and white populations. Persons whose race was coded as other than black or white were not included in the study. The third subsample was formed by combining the white and black populations. The three subsamples created from the NSFG for this study were:

Subsample one: both the black and white populations.

Subsample two: the black population.

Subsample three: the white population.

Aspects of the three subsamples are given in Table 2.5.2.

The event of interest for the NSFG was the time from first marriage to the time when a couple stopped living together. Women who were never married were not included in the analysis.
Table 2.5.2: Some characteristics of the three subsamples formed from the 1973 National Survey of Family Growth.

<table>
<thead>
<tr>
<th>Subsample</th>
<th>Number of Elementary Units</th>
<th>Number of PSUs</th>
<th>Number of Pseudo-Strata</th>
<th>Spread of Sample Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>8998</td>
<td>96</td>
<td>48</td>
<td>344-14996</td>
</tr>
<tr>
<td>Two</td>
<td>3211</td>
<td>96</td>
<td>48</td>
<td>344-11323</td>
</tr>
<tr>
<td>Three</td>
<td>5787</td>
<td>96</td>
<td>48</td>
<td>526-14996</td>
</tr>
</tbody>
</table>

The two national data sets were quite different regarding the original sample design and with respect to the final two PSU per stratum design. The spread of the sample weights, the values that the sample weights could take within that spread, PSU size and variability of PSU size were different for the two samples. The NSFG had a large spread of sample weights and large number of sample weight values, whereas the WFS had a small spread of sample weights and a small number of sample weight values within that spread. PSU size for the NSFG was moderately large, and size was similar for the 96 PSUs. The PSUs for the WFS were relatively small and greatly varied in size. These differences between the two data sets also existed between the various subsamples formed from each. Certain of the differences can be noted by comparing Tables 2.5.1 and 2.5.2.
2.6. Results

The results are presented in four parts. The first and second parts give the results for the conditional probabilities, $h_q^t$ and $h_P^t$, for the WFS and NSFG respectively. The third and fourth parts present the results for the survivorship probabilities, $t_{P0}$ or $S(t)$, for the two data sets. Graphs are used for presenting the results, however, tables containing the results are given in an appendix to this dissertation. The results are the computed variances obtained using the BRR, JRR, JK, and TL, and the design effects associated with each of the four methods. Design effect here refers to the ratio of the variance as computed under the actual sample design (in this case the two PSU per stratum design) to the variance as computed using the assumption of a simple random sample with the same number of elementary units. This ratio is formally given by Kish (1965) as

$$DEFF = \frac{\text{Variance under the actual sample design}}{\text{Variance assuming a simple random sample}}.$$  \hspace{1cm} (2.6.1)

Graphs 2.1 - 2.4 present the results of applying the four variance estimation techniques to the WFS subsamples to obtain variance estimates for the conditional probabilities. All graphs show that the four variance estimation methods give design effects which fluctuate around one, and that the four methods generally agree. In Graphs 2.1 - 2.3 each of the four methods of estimation shows the same pattern of fluctuation about unity, but in Graph 2.4 the methods are in disagreement except for the JRR and TL. Table 2.5.1 shows that this subsample contains only four pseudo-strata and eight PSUs. These findings indicate that the four methods are sensitive to the size of the data set with respect to the number of strata and number of PSUs. The results also
point to possible inconsistencies in the four methods when the sub-
samples are fairly small. Again, by small the reference is to the
number of strata and PSUs. However, the number of elementary units
also may have some effect. Having a small number of elementary units
may cause variation in PSU and stratum size which could effect the
variance estimation methods.

The results for the conditional probabilities for the NSFG
subsamples are presented in Graphs 2.5 - 2.7. These graphs show good
agreement among the four methods, and design effects are generally
greater than one. Graphs 2.5 and 2.7 show the JK method to sharply
deviate from the pattern in the variances followed by the BRR, JRR,
and TL. This deviation is found for the time interval [2,3). The
reason for this deviation was not apparent when examining the data and
the JK method. Another interesting finding is the large difference in
the design effects for the blacks and whites. For the whites the
design effects are approximately one, whereas for the blacks they are
near four or five. One explanation for such a finding lies in the
variability in the sizes of the PSUs for the two races. The whites had
quite homogeneous PSU size while PSU size varied greatly for blacks.
Another possibility is that the two races may differ with respect to
the variability of the event under study.

A general statement concerning the four methods of variance esti-
mation is that the four methods tend to agree, but on the average the
BRR will provide larger variances than the other three methods. Among
the JRR, TL, and JK, no consistent ranking regarding the magnitude of
the variances was found. Kish and Frankel (1974) and others have noted
similar results, and have pointed to a positive bias associated with
the BRR as a reason for the larger variance estimate.

An important finding regarding the conditional probabilities is a caveat concerning the use of the four variance estimation methods in small data sets. The number of PSUs and strata may need to be relatively large for the methods to provide useful variance estimates. The number of elementary units in the data set may also be important, particularly the distribution of the elementary units over the strata and PSUs. It is well established that variable PSU size can affect variance estimates (Hansen, Hurwitz, and Madow, 1953). For an overall sample such variability can be controlled, but for domains of interest such variability may not be practical to control. Lemeshow (1979) and Stanek and Lemeshow (1977) have pointed out that the BRR can be adversely affected by variability in the statistic of interest across the strata and PSUs.

The results of applying the four variance estimation methods to the survivorship probabilities are given in Graphs 2.8 - 2.14. The findings are similar to those obtained for the conditional probabilities. Again, instability and a lack of agreement among the four methods in the small WFS data set subsample four, occurred (see Graph 2.11).

The plots for the NSFG present an interesting result. The design effects for both the BRR and JK exceed unity while the design effects for the JRR and TL fluctuate near unity (Graphs 2.12 and 2.14). Graph 2.12 is that for the white subpopulation and Graph 2.14 is for the combined black and white subpopulations. Since Graph 2.13, the results for blacks alone, does not show this discrepancy between the methods, the source for the discrepancy is within the subpopulation of whites. Exactly what is causing the BRR and JK to provide larger variances than
the JRR and TL methods is unknown. The finding does point out that relatively large differences can be observed among the methods for fairly large data sets.

A second finding is that the design effects for the domain of blacks are between 4.0-6.0 while those for the domain of whites are roughly between 1.0-2.0. This finding is similar to that reported for the conditional probabilities. A possible explanation for this result which was not mentioned before is that the blacks may be more homogeneous within clusters (PSUs) with respect to the behavior under investigation.

2.7. Discussion of the Results

The results of the experiments using the NSFG and WFS indicate that for relatively large domains or subsamples the four methods of variance estimation give estimates that in most situations agree, but some disagreement was noticed particularly in the small domains.

Of practical consideration at this point is which, if any, of the four methods is best for estimating the variances of the life table functions $h_t$ and $S(t)$. Three criteria for judging the four methods were given in section 2.1. Briefly, these criteria were: ease of implementation, required assumptions, and properties of the estimates. Regarding the final criterion, Krewski and Rao (1980) have shown that all four methods are consistent estimators and asymptotically equivalent.

Considering the ease of implementation favors the JRR. A ranking of difficulty (least → most) being: JRR, JK, TL, BRR. The jackknife methods do not require an orthogonal design matrix as does the BRR, nor do they require derivatives as does the TL. The JRR has the advantage
over the JK of not requiring the computation of the psuedo-values. Since the derivatives in this case are simple, the TL was easier to use than the BRR. Furthermore, if the BRR must have the sample weights recomputed for each half sample, then this method is at a disadvantage both in programming required and cost of using the program.

Concerning assumptions required, the JRR, JK, and BRR cover a larger class of functions since they do not require differentiability. However, the derivatives are simple for the life table functions studied. The JRR and BRR are design restricted in that the methods are only applicable to those designs with two PSUs per stratum. The TL and JK are not so restricted, but the only application of the JK in the literature is for the two PSU per stratum design. Application to other designs would require development. Here the TL has an advantage.

Up to this point the support for any one method is not overwhelming. However, a re-examination of Graphs 2.1 - 2.14 suggests that the Taylors linearization is the most practical. Study of the graphs shows the JK method frequently deviates from the pattern followed by the other three methods. Graphs 2.5 and 2.12 give evidence that the BRR might also give problems. Graph 2.5 shows the results for the $h^q_t$ for the NSFG domain of whites. In this graph the BRR, JRR, and TL are in good agreement, but the JK deviates from the other three. If the survivorship probabilities, which are functions of the $q^r_t$, are studied for the same domain, the BRR and JK both deviate sharply from the pattern of the TL and JRR. On the basis of the results for the $q^r_t$ of the same domain such a deviation by the BRR is unexpected. Similarly, in Graph 2.2 all four methods agree for the $q^r_t$, but when the survivorship probabilities are examined, Graph 2.9, both the JRR and JK behave
unexpectedly in the final time interval. Examination of Graphs 2.4 and 2.11 show a large amount of fluctuation in the BRR and JK methods as compared to the TL and JRR.

The previous examination indicates that the TL is more stable than the other three techniques. This result was also reported by Bean (1975) and Kish and Frankel (1974). This finding in conjunction with the TL's applicability in a wide range of sample designs, and the relatively simple derivatives which are associated with the life table functions suggests that the Taylor's linearization is the 'best' method of the four. This claim is tentative and does require further investigation using other sets of data.
Graph 2.1: Design Effects by Time Interval for the Conditional Probabilities of Subsample One Taken from the 1975 World Fertility Survey of Sri Lanka.

Legend:
- △ - BRR
- ✗ - TL
- ☐ - JRR
- ○ - JK

Design Effect

TIME INTERVAL IN YEARS

[0,1] [1,2] [2,3] [3,4]

Source: Table 1 of the Appendix.
Graph 2.2: Design Effect by Time Interval for the Conditional Probabilities of Subsample Two Taken from the 1975 World Fertility Survey of Sri Lanka.

Source: Table 2 of the Appendix.
Graph 2.3: Design Effects by Time Interval for the Conditional Probabilities of Subsample Three Taken from the 1975 World Fertility Survey of Sri Lanka.

Source: Table 3 of the Appendix.
Graph 2.4: Design Effects by Time Interval for the Conditional Probabilities of Subsample Four Taken from the 1975 World Fertility Survey of Sri Lanka.

Source: Table 4 of the Appendix.
Graph 2.5: Design Effects by Time Interval for the Conditional Probabilities of Subsample One Taken from the 1973 National Survey of Family Growth.

Source: Table 5 of the Appendix.
Graph 2.6: Design Effects by Time Interval for the Conditional Probabilities of Subsample Two Taken from the 1973 National Survey of Family Growth.

LEGEND

△ - BRR
X - TL
□ - JRR
○ - JK

Source: Table 6 of the Appendix.
Graph 2.7: Design Effects by Time Interval for the Conditional Probabilities of Subpopulation Three Taken from the 1973 National Survey of Family Growth.

Source: Table 7 of the Appendix.
Graph 2.8: Design Effects by Time Interval for the Survivorship Probabilities of Subsample One Taken from the 1975 World Fertility Survey of Sri Lanka.

Source: Table 8 of the Appendix.
Graph 2.9: Design Effects by Time Interval for the Survivorship Probabilities of Subsample Two Taken from the 1975 World Fertility Survey of Sri Lanka.

Source: Table 9 of the Appendix.
Graph 2.10: Design Effects by Time Interval for the Survivorship Probabilities of Subsample Three Taken from the 1975 World Fertility Survey of Sri Lanka.

Source: Table 10 of the Appendix.
Graph 2.11: Design Effects by Time Interval for the Survivorship Probabilities of Subsample Four Taken from the 1975 World Fertility Survey of Sri Lanka.

Source: Table 11 of the Appendix
Graph 2.12: Design Effects by Time Interval for the Survivorship Probabilities of Subsample One Taken from the 1973 National Survey of Family Growth.

Source: Table 12 of the Appendix.
Graph 2.13: Design Effects by Time Interval for the Survivorship Probabilities of Subsample Two Taken from the 1973 National Survey of Family Growth.

Source: Table 13 of the Appendix.
Graph 2.14: Design Effects by Time Interval for the Survivorship Probabilities of Subsample Three Taken from the 1973 National Survey of Family Growth.

LEGEND

• - BRR
× - TL
□ - JRR
○ - JK

Source: Table 14 of the Appendix.
CHAPTER III
COMPARING SURVIVORSHIP CURVES

3.1. Introduction

Methods for obtaining variance estimates of the life table functions were examined in Chapter II. Frequently, the investigator is not only interested in the variance estimates as measures of precision, but also in using the variances for inferential purposes. A question of interest with respect to life tables is whether two or more groups, domains, or subpopulations have similar survival probabilities.

The purpose of this chapter is to briefly review some methods used for comparing survivorship between various populations when the data are obtained by simple random sampling with or without replacement. A modification to one of the methods reviewed is then proposed which will allow the method to be used for data gathered by a complex survey design.

3.2. A Review of Methods for Comparing Survivorship Curves

This review is not intended to be comprehensive; rather the focus is on a few methods which are well known and commonly used. Perhaps the most well known and widely applied statistical test for comparing survivorship between two populations is the Mantel-Haenszel test. This test was developed by Mantel and Haenszel in 1959 for combining information from a set of independent 2x2 or four-fold contingency tables. An equivalent test was given by Cochran (1954). The test was later extended by Mantel (1963, 1966) to the comparison of survivorship
between two groups.

The procedure of Mantel and Haenszel (MH) provides a statistical test of partial association across a set of four-fold tables. The entries in such a four-fold table, when comparing two life tables, are the counts of those individuals of each domain who were exposed to risk and either died or survived a given time interval. Thus, if the data are given in K time intervals, there will be K 2x2 tables. An example of such a four-fold table is given by Figure 3.2.1.

Figure 3.2.1: A table giving the counts of those who were exposed to risk and who either survived or died in the i-th time interval by domain.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Died</th>
<th>Survived</th>
<th>Exposed To Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>D_{1i}</td>
<td>S_{1i}</td>
<td>N_{1i}</td>
</tr>
<tr>
<td>Two</td>
<td>D_{2i}</td>
<td>S_{2i}</td>
<td>N_{2i}</td>
</tr>
<tr>
<td>Total</td>
<td>D_{i}</td>
<td>S_{i}</td>
<td>N_{i}</td>
</tr>
</tbody>
</table>

The following development of the Mantel-Haenszel test is due to Fleiss (1973). A measure of association

\[ d_i = \frac{p_{1i}-p_{2i}}{p_i q_i} \]  

(3.2.1)

is formed for each table, where

\[ p_{1i} = \frac{S_{1i}}{N_{1i}} \]
\[ p_{2i} = \frac{S_{2i}}{N_{2i}} \]
\[ p_i = \frac{S_i}{N_i} \]

and \[ q_i = 1-p_i \].
This measure of association is called the standardized difference. The standardized difference has been discussed by Fleiss (1970) and Yates (1955). Yates has pointed out that the use of $d_i$ results in a difference analogous to logits, which may be expected to be reasonably constant over a wide range of $p_i$. Although the $p_i$ may vary from table to table, the $d_i$ should remain relatively stable as long as the difference $(p_{1i} - p_{2i})$ does not change sign. This stability is quite useful for comparing survivorship between groups, since as time passes, the probabilities of survival usually change. However, over time the difference in survival between two populations may remain constant due to a proportional hazards relationship. The use of the $d_i$ allows the hypothesis of a constant difference in survivorship to be expressed as a function of the conditional probabilities of survival. Furthermore, Birch (1964) has demonstrated that the power of the Mantel-Haenszel test is maximized when the difference between the domains being compared remains constant across the set of four-fold tables. The standardized difference was modified slightly by Mantel and Haenszel (1959), and for their test they used

$$d_i = \frac{N_i - 1}{N_i} \frac{p_{1i} - p_{2i}}{p_i q_i}.$$ 

The $d_i$ are used in a weighted average

$$\bar{d} = \frac{K}{\sum_{i=1}^{K} \omega_i} \frac{\sum_{i=1}^{K} \omega_i d_i}{\sum_{i=1}^{K} \omega_i}$$ \hspace{1cm} (3.2.2)

where

$$\omega_i = \text{Var}(d_i)^{-1}.$$ 

These variances are obtained under hypergeometric sampling resulting from the assumption of fixed marginal totals for each four-fold table.
It can be shown that

\[ \text{Var}(\bar{d}) = \left( \sum_{i=1}^{K} \omega_i \right)^{-1}. \] (3.2.3)

The MH test statistic is given by

\[ Q_{\text{MH}} = \bar{d} (\text{Var}(d))^{-1/2}. \] (3.2.4.a)

This test statistic, under the null hypothesis of no difference in survivorship between the two populations, has asymptotically a chi-square distribution with one degree of freedom.

A more common form of the Mantel-Haenszel test is

\[ Q_{\text{MH}} = \left( \sum_{i=1}^{K} D_{1i} \right) - \left( \sum_{i=1}^{K} E(D_{1i}) \right)^2 / \text{Var} \left( \sum_{i=1}^{K} D_{1i} \right). \] (3.2.4.b)

The expected values and variances are computed under a hypergeometric model assuming fixed marginal totals for each four-fold table.

The Mantel-Haenszel test assumes that any three-way interactions between the event under study, the time intervals and domains are negligible. This assumption is equivalent to assuming that any association between the event under study and the domains is constant across the time intervals. For a life table analysis, this assumption is satisfied if the hazard rates for any two of the populations being compared have a constant proportionality over the time intervals. The MH test is still valid even if the assumption is not met; however, Birch (1964) has shown that when the assumption is not valid, a loss of power may result.

Bishop, Fienberg, and Holland (1975) point out that the MH test is equivalent to a test of first order interaction in a log-linear model. They mention that the log-linear approach may provide a better framework...
since within this structure, several orders of interaction may be tested.

Tests directed at partial association have been examined by Landis, Heyman, and Koch (1978) within a weighted least squares approach. The application of weighted least squares to life table problems has been demonstrated by Koch, Johnson, and Tolley (1972). This approach has the advantage that hypothesis testing is done within a linear model framework based upon the conditional probabilities of an event. The approach is easily understood and can be applied to several populations.

The final method mentioned here for comparing survivorship curves is the log-rank test of Peto and Peto (1972) which is equivalent to the Mantel-Haenszel test. Peto (1972) has demonstrated that when censoring is equal in both samples, the log-rank test is the locally most powerful rank test. Peto and Pike (1973) have shown that the test statistic (3.2.4.b) can be approximated by

\[ Q_{LR} = \left( \sum_{i=1}^{K} D_{1i} - \sum_{i=1}^{K} E(D_{1i}) \right)^2 / \sum_{i=1}^{K} E(D_{1i}), \]  

(3.2.5)

where in the denominator the sum of the expected values is used in place of the sum of the variances. Such an approximation can be derived under the assumption of Poisson sampling. This approximation results in a conservative test. The log-rank test is also applicable when more than two populations are being compared (Mantel, 1966).

3.3. A Review of Some Methods for the Analysis of Complex Survey Data

All of the above mentioned tests require data from a simple random sample or stratified random sample, and all the tests assume the
observations are independent. The use of such methods when the independence assumption is invalid can result in an incorrect inference. This problem has been illustrated by Walsh (1947), Cowan and Binder (1978), and Felli gi (1978).

Because of the problems posed to statistical analysis by data from complex probability samples, during the past few years a considerable amount of research has been devoted to illustrate and solve these problems. Particular interest has been given to the analysis of categorical data. This research has resulted in modifications to several tests of independence and goodness of fit. The work of Rao and Scott (1979a, 1979b) serves as an excellent example.

Three separate, though not necessarily competing, approaches for the analysis of a $r \times c$ contingency table constructed from complex survey data can be identified. The first approach requires a consistent estimator of the variance-covariance matrix $\hat{\Sigma}$ with dimensions $(r-1) \times (c-1)$ of the vector $\hat{P}$ of cell proportions. The dimensions are reduced to provide a variance-covariance matrix with full rank. The test statistic is

$$Q = (\hat{P} - \hat{P}_0)' \hat{\Sigma}^{-1}(\hat{P} - \hat{P}_0)$$  \hspace{1cm} (3.3.1)

where $\hat{P}_0$ denotes the values of $\hat{P}$ (the population vector of proportions) specified by the null hypothesis. Under the null hypothesis, and under an appropriate central limit theorem, $Q$ is approximately distributed as a central chi-square with $(r-1) \times (c-1)$ degrees of freedom. Suitable central limit theorems have been given by: Madow (1948), Hajeck (1964), Fuller (1975), Schuster and Downing (1976), Rosen (1972), and Krewski and Rao (1980). Limit theorems establishing the consistency of $\hat{\Sigma}$ obtained by any of the four methods (balanced repeated replications,
jackknife repeated replications, jackknife, and Taylor series linearization) discussed in Chapter II have been established by Krewski and Rao (1980). This approach has been illustrated by Koch and Lemeshow (1972), Tomberlin (1979), and Schuster and Downing (1976).

A second development which closely parallels the Wald statistic approach is that given by Nathan (1973, 1975). Nathan derives approximations for the covariances and combines these with the method of balanced repeated replications to obtain a Hotelling's $T^2$ statistic as a test of independence in the $r \times c$ contingency table.

A fairly recent approach to tests of independence and goodness of fit for contingency tables from complex probability samples is that of Rao and Scott (1979a, 1979b) and Felliigi (1978). Rao and Scott write the typical chi-square test as a quadratic form, and then expand this quadratic form in terms of the eigenvalues of the variance-covariance matrix obtained under the sample design. The eigenvalues are then shown to be a type of generalized design effects. These eigenvalues can then be used to adjust the usual chi-square test for the survey design. The interesting facet of their approach is that the variance-covariance matrix need not be fully known to obtain the needed adjustment values for an asymptotically correct test. In cases where the dimensions of the contingency table are large, this approach could prove to be quite useful. This approach also allows several tests to be conducted with a single set of correction or adjustment factors. Hence, when publishing results, inclusion of these factors would greatly benefit those who perform secondary analyses and do not have the original data to obtain a variance-covariance matrix.
The methods reviewed for categorical data from a complex probability sample will in most cases provide a correct analysis for a single contingency table. The problem of interest here concerns several contingency tables, typically (2×2) or four-fold tables, and how to combine the information in each of the tables into a summary statistical test for association across the set of tables.

Several approaches could be taken to provide an adequate solution to the above problem. The log-linear model suggested by Tomberlin (1979) using the weighted least squares estimation given by Grizzle, Starmer, and Koch (1969), or the linear model based upon the cell proportions, using the same weighted least squares, of Koch, Johnson, and Tolley (1972) could be employed. The approach here will be to slightly modify the Mantel-Haenszel test, since it is a well known and widely used method. Before giving this modification we will turn to a discussion of the variance-covariance structure for the life table problem.

3.4. **An Examination of the Life Table Variance-Covariance Structure**

The intent here is to present a simplified discussion which demonstrates where the life table variance-covariance structure is derived. For this purpose we assume a two stage cluster sample. The selection of a equal size clusters is made by random sampling with replacement from a total of A equal size clusters in the population. The choice of b elementary units out of B units within each selected cluster is made by simple random sampling. The design is self-weighting, and the total sample size is n=ab.

Discussion will be focused on a given time interval and concerned with the proportion of elementary units having experienced some event E, in each of two domains. The interest is in comparing the two domains
with respect to the occurrence of the event. A third domain or sub-population is also defined. This third domain consists of all individuals who are not included in either of the two domains of interest. The three domains are assumed mutually exclusive.

Let

\[
I(D_1, E, j) = \begin{cases} 
1 & \text{if the } j\text{-th unit is in domain one and has experienced the event } E. \\
0 & \text{otherwise.} 
\end{cases}
\]

\[
I(D_2, E, j) = \begin{cases} 
1 & \text{if the } j\text{-th unit is in domain two and has experienced the event } E. \\
0 & \text{otherwise.} 
\end{cases}
\]

\(j=1, \ldots, N=AB\). The variances of \(I(D_1, E, j)\) and \(I(D_2, E, j)\) can be written as

\[
\text{Var}(I(D_1, E, j)) = \sigma_{11}^2 + \sigma_{12}^2
\]

\[
\text{Var}(I(D_2, E, j)) = \sigma_{21}^2 + \sigma_{22}^2
\]

where \(\sigma_{11}^2\) and \(\sigma_{21}^2\) represent the between cluster variation and \(\sigma_{12}^2\) and \(\sigma_{22}^2\) represent the within cluster variation. Under the null hypothesis that there is no difference between the proportion, \(p_1\), who experience the event in domain one and the proportion, \(p_2\), who experience the event in domain two, the following assumption regarding the variances is made:

\[
\sigma_{11}^2 = \sigma_{21}^2 = \sigma_1^2
\]

\[
\sigma_{12}^2 = \sigma_{22}^2 = \sigma_2^2
\]

This is a fairly strong assumption. However, it should be kept in mind that this unrealistic approach is taken only to demonstrate the characteristics of the life table variance-covariance structure for a given time interval.
The following notation and definitions are used in the development of the variance-covariance. Let

$$A = \{A_1, \ldots, A_a\}$$

denote the set of primary sampling units or clusters in the sample, and let

$$U = \{U_1, \ldots, U_a\}$$

denote the collection of sets whose elements are the elementary sample units $u_{ijr} : u_{ijr} \in U_i$, $i=1, \ldots, a$. The index $i$ denotes PSU; $j$ denotes individual units in the sample within a cluster ($j=1, \ldots, b$); and $r$ is used to index the domain of interest ($r=1,2,3$).

The elementary units, $u_{ijr}$, can further be separated into sets based upon both domain and PSU membership which are mutually exclusive. For example, the set $D_{ir}$ is the set which contains elementary units from the $i$-th PSU which are in the $r$-th domain. The collection of sets $D_{ir}$, $i=1, \ldots, a$; $r=1,2,3$ induce a partition on the set of PSUs which contain members of only one domain, any two domains, or all three domains. The cardinality of this partition for three domains is 7. In general, for $R$ domains, the number of identifiable partition sets is

$$c = \sum_{r=1}^{R} \binom{R}{r} = \sum_{r=1}^{R} \frac{R!}{r!(R-r)!} .$$

The partition of the PSUs defines a collection of sets

$$A = \{A_{100}, A_{120}, A_{003}, A_{120}, A_{130}, A_{023}, A_{123}\} ,$$

where the subscript denotes which domains are represented. For example, $A_{023}$ is the set of PSUs containing elementary units of both domain two and domain three, but not domain one. For the purpose of referring to these sets, let $\ell$ take values in the label set
L = \{100,020,003,120,103,023,123\}. Let
\[\alpha = \{\alpha_{100}, \alpha_{020}, \alpha_{003}, \alpha_{120}, \alpha_{103}, \alpha_{023}, \alpha_{123}\}\]

where
\[\alpha_{x} = \frac{1}{a_x}\] (The number of PSUs in \(A_x\))
denote a set of proportions associated with the PSU set A. There is
defined then a partition of PSUs based upon domain membership of the
elementary units, and a set of proportions of PSUs in each partition set.

A partition of the elementary sample units within a PSU based upon
domain membership of the units is now given. This partition is denoted by
\[B = \{B_{1100}, B_{1120}, B_{1103}, B_{1123}, B_{2020}, B_{2120}, B_{2023}, B_{2123}, B_{3003}, B_{3103}, B_{3023}, B_{3123}\}.

The first subscript associated with each member of the partition
indexes which domain is referred to while the second, third, and fourth
indexes designate the particular PSU partition. For example,
\[B_{1100}\] is the set of elementary units in domain one given a
PSU from \(A_{100}\). The set \(A_{100}\) only contains members of
domain one.

\[B_{1120}\] is the set of elementary units in domain one given a
PSU from \(A_{120}\). The set \(A_{120}\) contains members of
domain one and domain two.

\[B_{1103}\] is the set of elementary units in domain one given a
PSU from \(A_{103}\). The set \(A_{103}\) contains members of
domain one and domain three.

\[B_{1123}\] is the set of elementary units in domain one given a
PSU from \(A_{123}\). The set \(A_{123}\) contains members of
domain one, domain two, and domain three.
Analogous to the set $\alpha$ of proportions associated with the PSU partition there is a set of proportions associated with the set $B$. Again, letting $\lambda$ take on values in the index set $L = \{100, 020, 003, 120, 103, 023, 123\}$. The set of proportions of elementary units in a given domain from a given PSU partition set is

$$B = \{b_{1100}, b_{1120}, b_{1103}, b_{1123}, b_{2020}, b_{2120}, b_{2023}, b_{2123}, b_{3003}, b_{3103}, b_{3023}, b_{3123}\},$$

where

$$b_{r\lambda} = \frac{1}{n_{\alpha\lambda}} \left( \text{number of elementary units in each PSU of the set} \begin{array}{c} \lambda \\ \text{which belong to domain } r \end{array} \right)_{r=1,2,3}.$$

The following relationships are obtained from the sets $\alpha$ and $B$.

$$\sum_{\lambda \in L} \alpha_{\lambda} = 1$$

and

$$b_{1100} = b_{2020} = b_{3003} = 1$$
$$b_{1120} + b_{2120} = 1$$
$$b_{1103} + b_{3103} = 1$$
$$b_{2023} + b_{3023} = 1$$
$$b_{1123} + b_{2123} + b_{3123} = 1.$$

From the above specifications, the number of elementary units in domain one or domain two is

$$n_{12} = ab(\alpha_{100} + \alpha_{020} + \alpha_{120} (b_{1120} + b_{2120}) + \alpha_{103} b_{1103} + \alpha_{023} b_{2023} + \alpha_{123} (b_{1123} + b_{2123})).$$

Estimates of the proportions $p_1$ and $p_2$ can be obtained through the use of three indicator functions. These three indicator functions are now specified. The three functions are:
\[ I(B_{r}, j) = \begin{cases} 
1 & \text{if unit } j \text{ is in the set } B_{r}, \ r=1,2,3, \ l \in L \\
0 & \text{otherwise} 
\end{cases} \]

\[ I(\mathcal{D}_{r}, j) = \begin{cases} 
1 & \text{if unit } j \text{ is a member of the } r\text{-th domain} \\
0 & \text{otherwise} 
\end{cases} \]

and

\[ I(E, j) = \begin{cases} 
1 & \text{if unit } j \text{ has the characteristic of interest} \\
0 & \text{otherwise} 
\end{cases} . \]

The proportions of interest are computed as:

\[ \hat{p}_1 = \frac{\sum_{j=1}^{n} I(E, j) I(\mathcal{D}_{1}, j)}{\sum_{j=1}^{n} I(\mathcal{D}_{1}, j)} \]  \hspace{1cm} (3.4.3)

\[ \hat{p}_2 = \frac{\sum_{j=1}^{n} I(E, j) I(\mathcal{D}_{2}, j)}{\sum_{j=1}^{n} I(\mathcal{D}_{2}, j)} \]  \hspace{1cm} (3.4.4)

Each of the proportions can be viewed as a weighted sum of several proportions. The component proportions being defined on the applicable sets in \( B \). This weighted sum for \( p_1 \) is:

\[ \hat{p}_1 = \gamma_{1100} \left( \sum_{j=1}^{n} I(B_{1100}, j)I(E, j)/\sum_{j=1}^{n} I(B_{1100}, j) \right) \]

\[ + \gamma_{1120} \left( \sum_{j=1}^{n} I(B_{1120}, j)I(E, j)/\sum_{j=1}^{n} I(B_{1120}, j) \right) \]

\[ + \gamma_{1103} \left( \sum_{j=1}^{n} I(B_{1103}, j)I(E, j)/\sum_{j=1}^{n} I(B_{1103}, j) \right) \]

\[ + \gamma_{1123} \left( \sum_{j=1}^{n} I(B_{1123}, j)I(E, j)/\sum_{j=1}^{n} I(B_{1123}, j) \right) \]

\[ = \gamma_{1100} \hat{p}_{1100} + \gamma_{1120} \hat{p}_{1120} + \gamma_{1103} \hat{p}_{1103} + \gamma_{1123} \hat{p}_{1123} . \]  \hspace{1cm} (3.4.5)
The $\gamma_{1x}, \forall x \in \mathcal{L}$ have the following values:

\[
\begin{align*}
\gamma_{1100} &= \frac{\alpha_{100}}{\alpha_{100} + \alpha_{120}^{0} \beta_{1120} + \alpha_{103}^{0} \beta_{1103} + \alpha_{123}^{0} \beta_{1123}} \\
\gamma_{1120} &= \frac{\alpha_{120}^{0} \beta_{1120}}{\alpha_{100} + \alpha_{120}^{0} \beta_{1120} + \alpha_{103}^{0} \beta_{1103} + \alpha_{123}^{0} \beta_{1123}} \\
\gamma_{1103} &= \frac{\alpha_{103}^{0} \beta_{1103}}{\alpha_{100} + \alpha_{120}^{0} \beta_{1120} + \alpha_{103}^{0} \beta_{1103} + \alpha_{123}^{0} \beta_{1123}} \\
\gamma_{1123} &= \frac{\alpha_{123}^{0} \beta_{1123}}{\alpha_{100} + \alpha_{120}^{0} \beta_{1120} + \alpha_{103}^{0} \beta_{1103} + \alpha_{123}^{0} \beta_{1123}}
\end{align*}
\]

Similarly, the weights

\[
\begin{align*}
\gamma_{2020} &= \frac{\alpha_{020}}{\alpha_{020} + \alpha_{120}^{0} \beta_{2120} + \alpha_{023}^{0} \beta_{2023} + \alpha_{123}^{0} \beta_{2123}} \\
\gamma_{2120} &= \frac{\alpha_{120}^{0} \beta_{2120}}{\alpha_{020} + \alpha_{120}^{0} \beta_{2120} + \alpha_{023}^{0} \beta_{2023} + \alpha_{123}^{0} \beta_{2123}} \\
\gamma_{2023} &= \frac{\alpha_{023}^{0} \beta_{2023}}{\alpha_{020} + \alpha_{120}^{0} \beta_{2120} + \alpha_{023}^{0} \beta_{2023} + \alpha_{123}^{0} \beta_{2123}} \\
\gamma_{2123} &= \frac{\alpha_{123}^{0} \beta_{2123}}{\alpha_{020} + \alpha_{120}^{0} \beta_{2120} + \alpha_{023}^{0} \beta_{2023} + \alpha_{123}^{0} \beta_{2123}}
\end{align*}
\]

can be defined, and $\hat{p}_2$ written as

\[
\hat{p}_2 = \gamma_{2020} \hat{p}_{2020} + \gamma_{2120} \hat{p}_{2120} + \gamma_{2023} \hat{p}_{2023} + \gamma_{2123} \hat{p}_{2123}. \tag{3.4.6}
\]

The weights $\gamma_{1x}, \forall x \in \mathcal{L}$ are the proportions of elementary units in domain one or two that each of the $B_{1x}$ or $B_{2x}$ sets comprise.

Earlier it was assumed under the null hypothesis, $p_1 = p_2$, that the between PSU components of variance were equal, and that the within PSU components of variance were equal for the two domains, one and two.

The variances of the estimates $\hat{p}_1$ and $\hat{p}_2$ given in terms of the component PSU and domain sets $B$. The variance of $\hat{p}_1$ is approximately given by:
\[ \text{Var}(\hat{p}_1) = \gamma_{1103}^2 \left( \frac{\sigma_1^2}{\alpha_{103a}} + \frac{\sigma_2^2}{\alpha_{100ab}} \right) + \gamma_{1120}^2 \left( \frac{\sigma_1^2}{\alpha_{120a}} + \frac{\sigma_2^2}{\alpha_{120b}1120ab} \right) \\
\quad + \gamma_{1103}^2 \left( \frac{\sigma_1^2}{\alpha_{103a}} + \frac{\sigma_2^2}{\alpha_{103b}1103ab} \right) + \gamma_{1123}^2 \left( \frac{\sigma_1^2}{\alpha_{123a}} + \frac{\sigma_2^2}{\alpha_{123b}1123ab} \right) \]

and for \( \hat{p}_2 \):

\[ \text{Var}(\hat{p}_2) = \gamma_{2020}^2 \left( \frac{\sigma_1^2}{\alpha_{020a}} + \frac{\sigma_2^2}{\alpha_{020ab}} \right) + \gamma_{2120}^2 \left( \frac{\sigma_1^2}{\alpha_{120a}} + \frac{\sigma_2^2}{\alpha_{120b}2120ab} \right) \\
\quad + \gamma_{2023}^2 \left( \frac{\sigma_1^2}{\alpha_{023a}} + \frac{\sigma_2^2}{\alpha_{023b}2023ab} \right) + \gamma_{2123}^2 \left( \frac{\sigma_1^2}{\alpha_{123a}} + \frac{\sigma_2^2}{\alpha_{123b}2123ab} \right) \]

These variances assume the \( \gamma_{1l} \) and \( \gamma_{2l} \), \( l \in L \) are constant. The actual concern is the variance of the difference

\[ \hat{d} = \hat{p}_1 - \hat{p}_2 \]

This variance is formally written as

\[ \text{Var}(\hat{d}) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) - 2 \text{Cov}(\hat{p}_1, \hat{p}_2) \]

The components \( \text{Var}(\hat{p}_1) \) and \( \text{Var}(\hat{p}_2) \) have been discussed and specified. The real concern here is with the covariance term in equation (3.4.9), and attention is now given to identifying the possible sources of such a covariance term.

The covariance in equation (3.4.9) can be rewritten as:

\[ \text{Cov}(\hat{p}_1, \hat{p}_2) = \text{Cov}(\gamma_{1100}\hat{p}_{1100} + \gamma_{1120}\hat{p}_{1120} + \gamma_{1103}\hat{p}_{1103} + \gamma_{1123}\hat{p}_{1123}, \gamma_{2020}\hat{p}_{2020} + \gamma_{2120}\hat{p}_{2120} + \gamma_{2023}\hat{p}_{2023} + \gamma_{2123}\hat{p}_{2123}) \]
\[= \gamma_{1100'2020} \text{Cov}(\hat{p}_{1100}, \hat{p}_{2020}) = 0 \]

\[+ \gamma_{1100'2120} \text{Cov}(\hat{p}_{1100}, \hat{p}_{2120}) = 0 \]

\[+ \gamma_{1100'2023} \text{Cov}(\hat{p}_{1100}, \hat{p}_{2023}) = 0 \]

\[+ \gamma_{1100'2123} \text{Cov}(\hat{p}_{1100}, \hat{p}_{2123}) = 0 \]

\[+ \gamma_{1120'2020} \text{Cov}(\hat{p}_{1120}, \hat{p}_{2020}) = 0 \]

\[+ \gamma_{1120'2120} \text{Cov}(\hat{p}_{1120}, \hat{p}_{2120}) = \gamma_{1120'2120}^2 \left( \frac{\sigma_1^2}{\alpha_{120}^a} + \frac{\sigma_2^2}{\alpha_{120}^b \alpha_{120}^{ab}} \right)^{1/2} \]

\[\left( \frac{\sigma_1^2}{\alpha_{120}^a} + \frac{\sigma_2^2}{\alpha_{120}^b \alpha_{120}^{ab}} \right)^{1/2} \cdot \hat{p}_{1120} \hat{p}_{2120} \]

\[+ \gamma_{1120'2023} \text{Cov}(\hat{p}_{1120}, \hat{p}_{2023}) = 0 \]

\[+ \gamma_{1120'2123} \text{Cov}(\hat{p}_{1120}, \hat{p}_{2123}) = 0 \]

\[+ \gamma_{1103'2020} \text{Cov}(\hat{p}_{1103}, \hat{p}_{2020}) = 0 \]

\[+ \gamma_{1103'2120} \text{Cov}(\hat{p}_{1103}, \hat{p}_{2120}) = 0 \]

\[+ \gamma_{1103'2023} \text{Cov}(\hat{p}_{1103}, \hat{p}_{2023}) = 0 \]

\[+ \gamma_{1103'2123} \text{Cov}(\hat{p}_{1103}, \hat{p}_{2123}) = 0 \]

\[+ \gamma_{1123'2020} \text{Cov}(\hat{p}_{1123}, \hat{p}_{2020}) = 0 \]

\[+ \gamma_{1123'2120} \text{Cov}(\hat{p}_{1123}, \hat{p}_{2120}) = 0 \]

\[+ \gamma_{1123'2023} \text{Cov}(\hat{p}_{1123}, \hat{p}_{2023}) = 0 \]

\[+ \gamma_{1123'2123} \text{Cov}(\hat{p}_{1123}, \hat{p}_{2123}) = \gamma_{1123'2123}^2 \left( \frac{\sigma_1^2}{\alpha_{123}^a} + \frac{\sigma_2^2}{\alpha_{123}^b \alpha_{123}^{ab}} \right)^{1/2} \]

\[\left( \frac{\sigma_1^2}{\alpha_{123}^a} + \frac{\sigma_2^2}{\alpha_{123}^b \alpha_{123}^{ab}} \right)^{1/2} \cdot \hat{p}_{1123} \hat{p}_{2123} \]

\[(3.4.10)\]
The covariance can be rewritten as:

\[
\text{Cov}(\hat{p}_1, \hat{p}_2) = \gamma_{1120}^2 \gamma_{2120}^2 \text{Var}(\hat{p}_{1120})^{1/2} \text{Var}(\hat{p}_{2120})^{1/2} \rho_{\hat{p}_1, \hat{p}_2} \\
+ \gamma_{1123}^2 \gamma_{2123}^2 \text{Var}(\hat{p}_{1123})^{1/2} \text{Var}(\hat{p}_{2123})^{1/2} \rho_{\hat{p}_1, \hat{p}_2} \quad (3.4.11)
\]

assuming \( \rho_{\hat{p}_{1120}, \hat{p}_{2120}} = \rho_{\hat{p}_{1123}, \hat{p}_{2123}} = \rho_{\hat{p}_1, \hat{p}_2} \). It is apparent that this covariance is non-zero only if \( \rho_{\hat{p}_1, \hat{p}_2} > 0 \). The interpretation given to \( \rho_{\hat{p}_1, \hat{p}_2} \) is that \( \hat{p}_1 \) and \( \hat{p}_2 \) tend to vary similarly among repeated samples.

3.5. **A Modified Mantel-Haenszel Statistic**

The development of the Mantel-Haenszel test statistic given by Fleiss (1973) will be used to derive a similar test statistic for complex survey data. The key modification is the substitution of design based variances and covariances into the Fleiss approach. However, before going to this modification certain other aspects of the test need to be given.

The first point concerns the entries of counts given in Figure 3.2.1 in section 3.2. For complex surveys these entries will usually be weighted sums. As an example, if \( D_{1i} \) represents the first domain or subpopulation, \( E \) denotes the event of interest, and \( \omega_j \) denotes the sampling weight for the \( j \)-th individual, the the number of events in domain one for the \( i \)-th time interval is

\[
D_{1i} = \sum_{j=1}^{n} I(D_{1i}, E, j, i) \omega_j , 
\]

(3.5.1)

where
\[ I(D_{ij}, E_{ij}, j, i) = \begin{cases} 
1 & \text{if the j-th individual is in domain one and experiences the event in the i-th time interval} \\
0 & \text{otherwise.} 
\end{cases} \]

Similar constructions can be used to obtain the other counts in Figure 3.2.1 in section 3.2. The proportions \( p_{1i} \) and \( p_{2i} \) are computed from the entries as before.

The second point concerns the relationships between the set of four-fold tables. The MH test was designed for independent tables, but Mantel (1966) has demonstrated that the method is valid for life table situations where the four-fold tables are not necessarily independent. Mantel's demonstration rests on the assumption that the marginals for all tables are fixed. This same assumption can be applied to the situation of complex probability sample data. The main point here is that the survey design and clustering affects events and relationships within a given table not across tables, or time intervals.

The test for complex survey data can now be developed following the approach of Fleiss. First we construct the standardized difference as

\[ d_i = (p_{1i} - p_{2i}) / \sqrt{p_{1i}q_i} \quad (3.5.2) \]

To form the weighted average \( \bar{d} \) as before in section 3.2 we need \( \text{Var}(d_i) \), \( i = 1, \ldots, k \). These variances are computed using the variance-covariance matrix of \( (p_{1i}, p_{2i}) \) given by

\[ V_i = \begin{bmatrix} \text{Var}(p_{1i}) & \text{Cov}(p_{1i}, p_{2i}) \\ \text{Cov}(p_{1i}, p_{2i}) & \text{Var}(p_{2i}) \end{bmatrix} \quad (3.5.3) \]
Here, for complex survey data this matrix is obtained by using one of
the four methods (BRR, JK, JRR, TL) given in the second chapter. The
covariance term in (3.5.3) is similar to that discussed in section 3.4.
Using $V_i$, and assuming $p_i q_i$ as constant

$$\text{Var}(d_i) = \text{Var} \left( \frac{P_{1i} - P_{2i}}{p_i q_i} \right)$$

$$= (p_i q_i)^{-2} \{ \text{Var}(p_{1i}) + \text{Var}(p_{2i}) - 2 \text{Cov}(p_{1i}, p_{2i}) \}. \quad (3.5.4)$$

Let

$$\bar{d} = \frac{1}{K} \sum_{i=1}^{K} \omega_i d_i / \sum_{i=1}^{K} \omega_i$$

where $\omega_i = \text{Var}(d_i)^{-1}$. Then the test statistic is given by

$$Q_{CPS} = \bar{d} (\text{Var}(\bar{d}))^{-1} \bar{d}, \quad (3.5.5)$$

the CPS denoting complex probability sample. Under the null hypothesis
this test statistic has asymptotically a chi-square distribution with
one degree of freedom. The distribution follows from an appropriate
central limit theory such as that given by Schuster and Downing (1976),
and from the consistency of the variance estimate, established by
Krewski and Rao (1980).

3.6. Extension to More Than Two Populations

A generalization of the Mantel-Haenszel test and an equivalent test
developed by Cochran (1954) to more than two populations or domains has
been given by Hopkins and Gross (1971), by Landis, Heyman, and Koch
(1978), and by Mantel (1963).

The method used by Hopkins and Gross is to calculate a $\bar{d}$ for each
possible pairing of $r$ populations, and then establish a critical value
for the tests using the Bonferonni multiple comparison approach. The Landis, Heyman, and Koch method is similar, but uses weighted least squares estimation and a linear models framework.

The r-sample problem is now given within the structure of section 3.5. For the i-th interval we have observed

$$z_i = (p_{i1}, \ldots, p_{ir})$$

and the associated variance-covariance matrix

$$V_i = \begin{bmatrix}
\sigma^2_1 & \sigma_{12} & \cdots & \sigma_{1r} \\
\sigma_{12} & \sigma^2_2 & \cdots & \cdot \\
\cdot & \cdot & \ddots & \cdot \\
\sigma_{1r} & \cdot & \cdots & \sigma^2_r
\end{bmatrix}$$

This variance-covariance matrix is based on the design and

$$\sigma^2_h = \text{Var}(p_{hi}) \quad h=1,\ldots,r$$

and

$$\sigma_{hh'} = \text{Cov}(p_{hi}, p_{hi'}) \quad h \neq h' = 1,\ldots,r$$

A general test of 'homogeneity' for the r populations can be constructed by using a contrast matrix

$$C = \begin{bmatrix}
1, 0, 0, \ldots, -1 \\
0, 1, 0, \ldots, -1 \\
0, 0, 1, \ldots, -1 \\
\cdot \\
\cdot \\
0, 0, 0, \ldots, 1, -1
\end{bmatrix}$$

where C has dimension (r-1) \times r, and then forming

$$D_i = C z_i$$
with variance-covariance:

\[ V_{D_i} = C \Sigma_i C'. \]

The standardized differences for each interval \( i = 1, \ldots, k \) are formed using a diagonal matrix \( A_i \) whose diagonal elements are the quantity \((p_i q_i)^{-1}\) based upon the marginals of the \( i \)-th contingency table. The matrix \( A_i \) has dimension \((r-1) \times (r-1)\), and the diagonal elements are all the same. Using \( A_i \), define

\[ \Lambda_i = A_i D_i \]

with variance-covariance

\[ V_{\Lambda_i} = A_i V_{D_i} A_i' = A_i C \Sigma_i C' A_i'. \]

Following the two population problem, the weighted averages of the standardized differences are computed. This computation is given by

\[ \Lambda = \left[ \sum_{i=1}^{K} \frac{V_{\Lambda_i}^{-1}}{\Lambda_i} \right] \sum_{i=1}^{K} \frac{V_{\Lambda_i}^{-1}}{\Lambda_i} \Lambda_i \quad (3.6.1) \]

which is the multivariate analogue to \( \overline{d} \) of equation (3.2.2). Denoting the variance-covariance of \( \Lambda \) by \( V \), the test statistic is

\[ Q_{MCPS} = \Lambda' V^{-1} \Lambda \quad (3.6.2) \]

Using central limit theory and the consistency of the variance-covariance estimate, the test statistic \( Q_{MCPS} \) has a chi-square distribution with \((r-1)\) degrees of freedom. If the above test should be significant, the multiple comparison approach of Hopkins and Gross (1971) could be implemented to find which populations or domains are different.
3.7. Results of Applying the Modified Mantel-Haenszel Test

Three tests will be compared using the data sets discussed in section 2. of Chapter II. The three tests are:

1. The Mantel-Haenszel test computed as given in formula (3.2.4.b) using the hypergeometric based expected values and variances. This test will be denoted as $Q_{MH}$.

2. A test similar to $Q_{MH}$ except that the variances are computed using:

$$\text{Var}(d_i) = (p_1 q_1)^{-2} \left( \text{Var}(p_{1i}) + \text{Var}(p_{2i}) \right)$$

where

$$\text{Var}(p_{hi}) = \frac{p_{hi} q_{hi}}{N_{hi}}; \quad h = 1, 2.$$

The formula for this test is given by equation (3.2.4.a) and the test will be denoted by $Q_{SRS}$.

3. The third test will be that given by equation (3.5.5) and is the same as (2) above except that the variances and covariances are computed with the sample design considered. This third test is given as $Q_{CPS}$.

To examine the performance of the tests the two data sets discussed in Chapter II are used. The first of these data sets is the 1973 National Survey of Family Growth. This data set contains two main domains which are the black and white color groups. These two domains will be compared with respect to the time from first marriage to separation of the couple. The term separation here is defined as when the couple stopped living together. Four domains based upon geographic zones are used from the World Fertility Survey of Sri Lanka 1975 (see Chapter II, section 2.5). The event under study for the World Fertility Survey is the time to the birth of the first child from the date of a woman's first marriage.
Aside from studying the performance of the test statistics, the four methods for computing the design based variance-covariance (balanced repeated replications, jackknife repeated replication, Taylor series linearization, and the jackknife) will be examined to see if there is a substantial difference among the four methods regarding the value of the test statistic and the inference obtained.

To examine the sensitivity of the proposed tests, several changes to the observed proportions were made. The changes were arrived at by adding or subtracting a vector of constants from the observed vector of proportions. This addition or subtraction would change the differences between the proportions of the domains being compared but not affect the variability. The choice of the vector of constants was arbitrary but served the purposes of this research. Mantel (1966) has shown that the Mantel-Haenszel procedure is more sensitive to differences in survival in the beginning time intervals than in the later time intervals. Such a sensitivity is important since as time passes the two survivorship curves must become arbitrarily close as they approach zero. A simple explanation of this sensitivity is found by examining the derivation of the Mantel-Haenszel test given by Fleiss (1973). Fleiss shows that the standardized differences, equation (3.2.1), are weighted by the inverse of its variance when calculating the Mantel-Haenszel test statistic. In life testing situations, an initial set of cohorts or assumed cohorts start at time zero and these are then decremented as time passes. Thus, in the earlier time intervals there are more individuals, and differences are measured more precisely. The increased precision is reflected by the smaller variance, which when the inverse is taken, gives a larger weight to
the differences in the earlier time intervals.

The vector of observed proportions for a given domain consists of the set of estimated conditional probabilities of survival for the stated time intervals. Constants added or subtracted from the observed proportions changed the differences between the observed proportions in either: the first time interval; the first and second time intervals; the first, second, or third time intervals; or the final two time intervals. Those time intervals which were not changed were left as observed.

To aid in the general study of the tests, six graphs were constructed to provide a visual display of the relationships between the subpopulations or domains. Both the conditional probabilities and the survivorship probabilities have been plotted. These plots are given in graphs 3.7.1-3.7.6.

The first four graphs concern the World Fertility Survey Data. As can be seen in Graphs 3.7.1-3.7.4, the populations being compared do not differ a great deal, or if differences do occur, they change with time. Graphs 3.7.3 and 3.7.4 are examples where the relationships between the domains change with time. This is reflected by the lines in the graphs crossing. A fluctuation of this type, if real, will decrease the power of the Mantel-Haenszel type test statistics, since maximum power is obtained when the relationships are constant in a given direction (Birch, 1964).

The results of the application of the tests $Q_{MH}$, $Q_{SRS}$, and $Q_{CPS}$ to the data are given in Tables 3.7.1-3.7.4. Table 3.7.1 presents the findings for differences in the value of the test $Q_{CPS}$ among the four methods of variance estimation: BRR, JRR, JK, and TL of Chapter II.
Table 3.7.1 shows that the numerical value of the test statistic will vary with the method of variance estimation used, but the inference obtained is the same for all four methods of variance estimation. Krewski and Rao (1980) have shown all four estimation procedures provide consistent estimates. Hence, tests based on any of the four methods of variance estimation will be asymptotically equivalent. An interesting result is found in Table 3.7.1. This result is observed when comparing the values of the test statistic \( Q_{\text{SRS}} \) and \( Q_{\text{CPS}} \). For the WFS subpopulations, \( Q_{\text{SRS}} \) has a smaller numerical value than \( Q_{\text{CPS}} \). Except for computations based on the TL and JK variance estimates, the same result holds for the NSFG data set. The point is that this is the opposite of what would be expected. Use of simple random sampling with replacement assumptions generally results in underestimation of the variance, and an exaggerated type one error which results from an inflated test statistic value. An explanation for this result can be found by examining the design effects for certain time intervals. For certain methods of variance estimation and time intervals, the design effects are less than one (see Chapter II). The standardized differences in the case where the design effects are less than unity receive a greater weight, thus increasing the value of the test statistic for the design based variances.

Tables 3.7.2-3.7.4 give the results of the three tests \( Q_{\text{MH}} \), \( Q_{\text{SRS}} \), and \( Q_{\text{CPS}} \) for the data sets and changed proportions. Again, it is noticed that \( Q_{\text{CPS}} \) has a greater value for certain sets of proportions than either \( Q_{\text{MH}} \) or \( Q_{\text{SRS}} \). As before, this finding may be due to the design effects for certain intervals being less than unity. A general
finding is that the test statistics $Q_{MH}$ and $Q_{SRS}$ provide numerically larger values than the test statistic $Q_{CPS}$. This result may be due to the underestimation of the variances used in $Q_{MH}$ and $Q_{SRS}$, since these tests ignore the survey design. Also, in Tables 3.7.3 and 3.7.4, the test statistics $Q_{MH}$ and $Q_{SRS}$ gave values which were statistically significant ($p < .05$) for the observed set of proportions, and the set of changed proportions denoted by III. The test statistic $Q_{CPS}$ showed significance in only one of these comparisons; that for set III in Table 3.7.4.

All three tests appeared more sensitive to the changes (increased differences) made in the earlier time intervals; Tables 3.7.2, 3.7.3, and 3.7.4). An exception to this occurs in Table 3.7.4 regarding $Q_{CPS}$ for changes made in the first two time intervals versus changes made in the final two time intervals (viz. changes III and IV). One possible explanation for this exception is that the design effects are less than unity in the final two time intervals, [3,4] and [4,5] (see Chapter II). Changes made in the last two intervals also causes the data to follow the proportional hazards model. This relationship among the proportions can be noted by examining Graph 3.7.5, and then making the change IV of Table 3.7.4 to the data. Thus, although the test statistics $Q_{SRS}$ and $Q_{MH}$ usually may be more sensitive to differences in the earlier time intervals, the test statistic $Q_{CPS}$ will not always follow this pattern. This result is due to $Q_{CPS}$ having a numerical value which depends upon the design effects of the individual time intervals which may not be directly associated with the number of individuals or observations in the given interval, which in simple random sampling schemes greatly influences the precision of the estimates.
3.8. Discussion

The results of this chapter provide reason to believe that the proposed modification of the Mantel-Haenszel statistic to accommodate data from a complex probability sample will perform well. We note, however, that the results here are in no way conclusive, and that further study and application to other data sets is necessary.

An important finding is that the proposed test does appear to be sensitive to differences in the conditional probabilities, but that the sensitivity can depend on the design effects associated with the given time intervals. This finding is in contrast to that for the usual Mantel-Haenszel test which is usually sensitive to differences in the earlier or beginning time intervals.
Graph 3.7.1: Conditional Probabilities by Time for Subpopulations One and Two Taken from the 1975 World Fertility Survey of Sri Lanka.

Source: Tables 1 and 2 of the Appendix.
Graph 3.7.2: Survivorship Probabilities by Time for Subpopulations One and Two Taken from the World Fertility Survey of Sri Lanka

Source: Tables 8 and 9 of the Appendix.
Graph 3.7.3: Conditional Probabilities by Time for Subpopulations One, Two and Three Taken from the World Fertility Survey of Sri Lanka.

Source: Tables 2, 3, and 4 of the Appendix.
Graph 3.7.4: Survivorship Probabilities by Time for Subpopulations One, Two, and Three Taken from the World Fertility Survey of Sri Lanka.

LEGEND
- One
- Two
- Three

Source: Tables 9, 10, and 11 of the Appendix.
Graph 3.7.5: Conditional Probabilities by Time for Supopulations One and Two Taken from the 1973 National Survey of Family Growth

Source: Tables 5 and 6 of the Appendix.
Graph 3.7.6: Survivorship Probabilities by Time for Subpopulations One and Two Taken from the 1973 National Survey of Family Growth.

Source: Tables 12 and 13 of the Appendix.
Table 3.7.1: Applications Demonstrating the Performance of the Proposed Test Statistic with Respect to the Method Used for Variance Estimation

<table>
<thead>
<tr>
<th>SAMPLE/COMPARISON</th>
<th>METHOD OF VARIANCE ESTIMATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRS</td>
</tr>
<tr>
<td>WFS</td>
<td></td>
</tr>
<tr>
<td>Subpopulations 1 vs 2</td>
<td>0.67</td>
</tr>
<tr>
<td>Subpopulations 3 vs 4</td>
<td>0.12</td>
</tr>
<tr>
<td>NSFG</td>
<td></td>
</tr>
<tr>
<td>Blacks vs Whites</td>
<td>0.72</td>
</tr>
</tbody>
</table>
Table 3.7.2: Applications Demonstrating the Performance of the Proposed Test Statistic, Tcps; Using Data from the 1975 World Fertility Survey of Sri Lanka

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>SUBSAMPLE ONE</th>
<th>SUBSAMPLE TWO</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,1)</td>
<td>0.6318</td>
<td>0.6563</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.4024</td>
<td>0.3504</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.5571</td>
<td>0.4551</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.7183</td>
<td>0.7532</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.1610</td>
<td>0.1801</td>
</tr>
</tbody>
</table>

CHANGES MADE TO THE OBSERVED PROPORTIONS

I: (.08,0,0,0,0) added to subsample two.
II: (.05,.05,.05,0,0) added to subsample two.
III: (.1,.1,0,0,0) added to subsample two.
IV: (0,0,0,.1,.1) added to subsample two.
V: (.1,.08,.05,0,0) added to subsample two.

VALUES OF THE TEST STATISTICS

<table>
<thead>
<tr>
<th>TEST</th>
<th>OBSERVED</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>QMH</td>
<td>0.44</td>
<td>2.50</td>
<td>4.42</td>
<td>14.42</td>
<td>0.03</td>
<td>14.17</td>
</tr>
<tr>
<td>Qs SRS</td>
<td>0.67</td>
<td>2.14</td>
<td>3.38</td>
<td>13.98</td>
<td>0.46</td>
<td>13.76</td>
</tr>
<tr>
<td>Qcps</td>
<td>0.92</td>
<td>0.56</td>
<td>2.37</td>
<td>8.84</td>
<td>0.27</td>
<td>8.61</td>
</tr>
</tbody>
</table>
Table 3.7.3: Applications Demonstrating the Performance of the Proposed Test Statistic, $T_{CPS}$: Using Data from the 1975 World Fertility Survey of Sri Lanka

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>SUBSAMPLE THREE</th>
<th>SUBSAMPLE FOUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,1)</td>
<td>0.6096</td>
<td>0.6025</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.3851</td>
<td>0.4286</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.4876</td>
<td>0.6552</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.6207</td>
<td>0.5556</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.8889</td>
<td>0.6000</td>
</tr>
</tbody>
</table>

CHANGES MADE TO THE OBSERVED PROPORTIONS

I: (.08,0,0,0,0) added to subsample four.

II: (.05,.05,.05,0,0) added to subsample four.

III: (.1,.1,0,0,0) added to subsample four.

IV: (0,0,.1,.1) added to subsample four.

V: (.1,.08,.05,0,0) added to subsample four.

VALUES OF THE TEST STATISTICS

<table>
<thead>
<tr>
<th>SET OF PROPORTIONS</th>
<th>TEST</th>
<th>OBSERVED</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q_{MH}$</td>
<td>0.08</td>
<td>1.60</td>
<td>2.12</td>
<td>4.81</td>
<td>0.48</td>
<td>4.44</td>
</tr>
<tr>
<td></td>
<td>$Q_{SRS}$</td>
<td>0.12</td>
<td>1.50</td>
<td>2.04</td>
<td>4.50</td>
<td>0.48</td>
<td>4.20</td>
</tr>
<tr>
<td></td>
<td>$Q_{CPS}$</td>
<td>0.26</td>
<td>1.41</td>
<td>2.26</td>
<td>2.78</td>
<td>0.98</td>
<td>3.22</td>
</tr>
</tbody>
</table>
Table 3.7.4: Applications Demonstrating the Performance of the Proposed Test Statistic, $T_{CPS}$; Using Data from the 1973 National Survey of Family Growth

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>BLACKS</th>
<th>WHITES</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,1)</td>
<td>0.9816</td>
<td>0.9630</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.9826</td>
<td>0.9750</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.9836</td>
<td>0.9783</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.9829</td>
<td>0.9854</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.9903</td>
<td>0.9944</td>
</tr>
</tbody>
</table>

CHANGES MADE TO THE OBSERVED PROPORTIONS

I: (.08,0,0,0,0) subtracted from the black proportions.
II: (.05,.05,.05,0,0) subtracted from the black proportions.
III: (.1,.1,0,0,0) subtracted from the black proportions.
IV: (0,0,0,1,.1) subtracted from the black proportions.
V: (.1,.08,.05,0,0) subtracted from the black proportions.

VALUES OF THE TEST STATISTICS

<table>
<thead>
<tr>
<th>TEST</th>
<th>OBSERVED</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{MH}$</td>
<td>18.18</td>
<td>1375</td>
<td>6192</td>
<td>1720</td>
<td>963</td>
<td>1579</td>
</tr>
<tr>
<td>$Q_{SRS}$</td>
<td>12.64</td>
<td>1061</td>
<td>5624</td>
<td>1542</td>
<td>1088</td>
<td>1209</td>
</tr>
<tr>
<td>$Q_{CPS}$</td>
<td>0.72</td>
<td>128</td>
<td>5586</td>
<td>1414</td>
<td>2296</td>
<td>307</td>
</tr>
</tbody>
</table>
CHAPTER IV
LIFE TABLE REGRESSION ANALYSIS

4.1. Introduction

In this chapter a brief review of some currently used approaches to life table regression analysis is given. A slight modification to these approaches is developed which allows for an explicit, although approximate, solution for the regression coefficients. This approximate solution is available only when the covariates are of the (0-1) indicator type. Results of applying the new approach and the approximations are given and compared to results obtained using existing methods. A discussion of the problems associated with the use of life table regression models in complex probability samples is given, and possible solutions outlined.

4.2. Life Table Regression Models

The primary need for life table regression models is that often a researcher observes concomitant information on subjects in addition to their survival history. The question then arises whether this auxiliary information has any bearing on the time to the occurrence of the event under investigation. For example, in clinical trials involving several treatments, dose levels, and background information on each individual, it is of interest to know if the treatments differ, what effect dose may have, and how might an individual's background information be associated with differences in treatment and dose effects.
Without life table regression methods, the approach has been to construct several life tables from the various dose, treatment, and covariate combinations. This may be a workable solution for large data sets, but for small sets of data, some structure needs to be imposed to overcome the problem of small numbers or empty cells for various combinations of the treatments, etc. The regression approach provides such a structure.

To proceed with the discussion, some notation needs to be given. In the following we let $\beta$ denote a $p \times 1$ vector of regression coefficients requiring estimation and $Z_j$ will denote a $1 \times p$ vector of fixed covariates measured at the start of the investigation or when the $j$-th individual enters the study: $j=1, \ldots, N$. The random variable specifying time to the event will be denoted by $T$; arbitrary time points, or observed values of $T$, will be given by $t$ or $t_i$. The sample data in its entirety will be denoted by $S$.

To facilitate the discussion the event of interest will be death, and the hazard rate or instantaneous death rate will be given by $\lambda(t; Z_j)$ to indicate the dependence on the covariate values.

The first model discussed is that of Fiegel and Zelen (1965). Their model was based upon an exponential distribution of time to death for each individual. The density function of $T$ for the $j$-th individual is

$$f_j(t) = \begin{cases} \lambda_j \exp(-\lambda_j t), & t \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (4.2.1)$$

The authors gave their model in terms of the expected value of the random variable $T$:

$$E_j(T) = \frac{1}{\lambda_j} = \alpha + \beta Z_j \quad (4.2.2)$$
A likelihood was constructed from this model and the parameters \( \alpha \) and \( \beta \) obtained iteratively. Two alternative expressions for (4.2.2) given by Fiegel and Zelen are:

\[
E_j(T) = (\alpha + \beta Z_j)^{-1}
\]  \hspace{1cm} (4.2.3a)

and

\[
E_j(T) = \alpha \exp(\beta Z_j).
\]  \hspace{1cm} (4.2.3b)

The expression \( \exp(\beta Z_j) \) for incorporating the covariates has become widely used. Kalbfleisch and Prentice (1980) have recommended this form since it places no restrictions on the values the regression parameters might assume. Some restrictions are necessary in 4.2.2 and 4.2.3a to ensure that the hazard rate is greater than or equal to zero.

Glasser (1967) gives the functional form \( \exp(Z_j \beta) \) and discussed extending the analysis to include several covariates. Zippin and Armitage (1966) used the model (4.2.2) and included three covariates.

In 1972, D. R. Cox presented a paper before the Royal Statistical Society based upon the hazard model

\[
\lambda(t; Z_j) = \lambda_0(t) \exp(Z_j \beta),
\]  \hspace{1cm} (4.2.4)

where \( \lambda_0(t) \) is an underlying hazard rate whose form is unspecified. This model is a generalization of that given by equation (4.2.3b) because it makes no assumption regarding the underlying failure time distribution. The Cox model can also be viewed as an extension of the Kaplan-Meier (1958) product limit approach to estimation of the survivorship probabilities. The assumption of fixed covariates in model (4.2.4) leads to the proportional hazards assumption. This assumption states that the hazard rates for any two individuals are in a constant proportion over time: that the hazard rates are parallel, and the
survivorship curves do not cross one another. Kalbfleisch (1978) gives a more general and formal definition of the proportional hazards assumption. The relationship is derived from the following specifications.
Suppose \( U = Z + V \) where \( V \) is distributed as an extreme value type I variate. Then the density function of \( U \) is
\[
f_U(u) = \exp(-u - e^{-u}).
\]
Given a monotone increasing differentiable function \( g(\cdot) \) a third random variable is defined as
\[
Y = g^{-1}(U),
\]
which has density function
\[
f_Y(y) = \exp(-(g(y) - Z \beta) - \exp(g(y) - Z \beta))g'(y).
\]
If the random variable for time to failure, \( T \), is given by
\[
T = e^Y,
\]
then the density function of \( T \) is derived as
\[
f_T(t) = \lambda(t) \exp(-Z \beta) \exp(-\exp(-Z \beta) \int_0^t \lambda(u) \, du), \quad 0 < t < \infty
\]
where \( \lambda(t) = \exp(g(\log t))g'((\log t) / t) \) is a arbitrary hazard function.
This is exactly the density associated with the Cox proportional hazard model (4.2.4).

Regarding estimation, Cox treated \( \lambda_0(t) \) as a nuisance function and through careful consideration of the risk set, \( R(t_i) \), (the set of subjects or individuals at risk immediately prior to each observed failure time \( t_i \)), he developed a likelihood independent of the underlying hazard. The likelihood derived by Cox, assuming no ties, is
\[ L(\beta; s) = \prod_{j=1}^{N} \frac{\exp(Z_j \beta)}{\sum_{m \in R(t_j)} \exp(Z_m \beta)} . \] (4.2.5)

This likelihood is the product of the probabilities that individual \( j \) dies at \( t_j \), given the failure times \( t_1, \ldots, t_N \), and the risk set \( R(t_i) \) available at time \( t_i: i=1, \ldots, N \). For discrete failure times and numerous ties, Cox proposed a logistic model

\[ \frac{\lambda(t; Z_j)}{1 - \lambda(t; Z_j)} = \exp(Z_j \beta) \frac{\lambda_0(t)}{1 - \lambda_0(t)} . \] (4.2.6)

The likelihood Cox used aroused considerable discussion and concern, and his paper touched off renewed interest in life table regression problems and survivorship analysis in general.

Many papers which immediately followed were primarily concerned with establishing the exact nature of Cox's likelihood and modifying certain assumptions allowing easier estimation of the survivorship function or a more adequate method for handling tied and grouped data.

The first modification was suggested by Breslow (1972). Consider the sample to consist of \( K \) failure times. Breslow proposed that \( \lambda_0(t) \) be assumed constant between observed failure times. If we denote the observed failure times by \( t_1, \ldots, t_K \), then Breslow's assumption implies

\[ \lambda_0(t) = \lambda_i, \quad i=1, \ldots, K \]

for \( t \in [t_i, t_{i+1}) \). A piecewise exponential distribution of failure time is obtained using this modification.

Cox had originally assumed \( \lambda_0(t) \) to be zero except at the observed failure times. He used this assumption to facilitate estimation of the survivorship curve. The implementation of Breslow's assumption greatly
eased this estimation problem. In a 1974 paper Breslow demonstrated that the parameter estimates obtained using his and Cox's assumptions regarding \( \lambda_0(t) \) differed very little.

A thorough examination of Cox's paper was given by Kalbfleisch and Prentice (1973). In their paper, the authors provide a derivation of Cox's model based upon the marginal distribution of the ranks associated with the failure times. Cox (1975) does provide justification of his approach and derivation in a paper on partial likelihood. Kalbfleisch and Prentice also give a discrete time version of the proportional hazards model. This discrete time model was given as a substitute for the logistic approach suggested by Cox. The Kalbfleisch and Prentice discrete model is

\[
\lambda(t; z) dt = 1 - (1 - \lambda_0(t) dt)^{\exp(Z_j \beta)}.
\] (4.2.7)

In this model \((1 - \lambda_0(t) dt)\) is the conditional probability of survival for a small interval of time \([t, t+dt]\). The relationship given by (4.2.7) becomes that given by equation (4.2.4) when time becomes continuous.

Kalbfleisch and Prentice also point out that the discrete time approach used by Cox, (4.2.6), results in an inference statement concerning the regression parameter in a logistic regression rather than in the hazard model (4.2.4). They do point out however that for short intervals the logistic model provides an adequate first approximation to (4.2.4) which is as Cox had indicated in his 1972 paper.

A more serious problem noted by Kalbfleisch and Prentice dealt with Cox's method for estimating the survivorship function. Cox used his logistic analog of equation (4.2.4) to obtain the estimates.
Kalbfleisch and Prentice state that the estimates obtained in this fashion relate to each other only through the logistic model; continuous survivorship estimates based upon (4.2.4) cannot be made arbitrarily close to those obtained using Cox's approach; thus, the estimates obtained using Cox's method are not a true supremum of the likelihood under the model given by (4.2.4). The authors do point out that use of the discrete model (4.2.7) does result in a legitimate supremum, and when $\beta = 0$, this approach is identical to that of Kaplan and Meier (1958).

Holford (1976) introduces changes into the Cox model which allow the use of the life table regression analysis in actuarial type life tables where the follow-up time is divided into previously stated intervals. Holford followed Breslow's initiative and assumed $\lambda_0(t) = \lambda_i$ for $t \in [t_i, t_{i+1})$. This model, as well as Breslow's, is:

$$\lambda(t; Z_j) = \lambda_i \exp(Z_j \beta), \ t [t_i, t_{i+1}).$$

(4.2.8)

The likelihood obtained using this model is a product of the individual likelihoods for each time interval and is piecewise in nature. Holford's paper also provides some interesting relationships between hypotheses tests based upon the log likelihood for model (4.2.8) and the partition into components of the chi-square test of homogeneity and association (see Fleiss, 1973).

Thompson (1977) points out that the best method for handling numerous ties is to use grouping rather than a discrete time model. Thompson develops a logistic model based upon (4.2.4) specifically for grouped data. He demonstrates that when the grouping intervals become arbitrarily small, his model and the continuous model (4.2.4) become the same.
The previous discussion has not touched upon time dependent covariates; fixed covariates have been assumed leading to the proportional hazards assumption.

The incorporation of time dependent covariates allows more flexibility in the analysis. The use of time dependent covariates allows the ratio of the hazard rate to change over time, and the survivorship curves can cross. However, if time dependent covariates are included in an analysis, then the arguments of Kalbfleisch and Prentice (deriving Cox's likelihood from the marginal distribution of the rank statistic associated with the failure times) no longer holds. Cox's development based upon the partial likelihood holds in general, for both fixed and time dependent covariates. Several models developed specifically for time dependent covariates will now be given.

A general model which forces only a mild parametric structure on the underlying hazard rate has been given by Taulbee (1976). Taulbee uses the generalized Rayleigh distribution to specify the time to failure. This distribution has a hazard rate which is a polynomial in $t$, the time to failure. The hazard is combined with a function of the covariates which also may be a function of time. The hazard model for this approach is:

$$
\lambda_m(t;Z_j) = \sum_{k=0}^{m} \lambda_k h(Z_j, \beta_k) t^k.
$$

(4.2.9)

In this equation $m$ is the degree of the polynomial in time which specifies the underlying hazard, and $h(Z_j, \beta_k)$ is a function incorporating the covariates. For example, $h(Z_j, \beta_k) = \exp(Z_j \beta_k)$. The model (4.2.9) allows considerable flexibility regarding forms that the survivorship curves and hazard rates may assume. Estimation of the
survivorship curve is facilitated by the parametric assumption on the underlying hazard. However, because of computational difficulty, this model will only prove useful when both \( m \) and \( p \), the polynomial degree and dimension of \( \mathbb{R}^k \) respectively, are small.

Other models for time dependent covariates have been given by Peduzzi, Holford, and Hardy (1977). This model is a generalization of Holford's (1976) model to incorporate time dependent covariates. Crowley and Hu (1974) give an example of the Cox model to heart transplant data with time dependent covariates.

The models presented in this section were based upon prospective cohort data, although retrospective data could be used. Many of the models depend upon known exact times of events like death and censoring. Data with ties or grouped into intervals with unknown times of events do cause estimation problems. The above models were also formulated for small to moderate size data sets. In large data sets, like those available from national surveys or disease registries, the models could be difficult to use.

In the following section a model based upon the discrete hazard model of Kalbfleisch and Prentice (1973) is given which attempts to provide more easily obtained parameter estimates, particularly in large data sets. The proposed model might also be useful for cross-sectional data where survival status is available for various age groups or several cohorts.

4.3. A Life Table Regression Model

The model developed in this section is primarily for grouped survival data and large data sets. The observations available to the analyst are the number of individuals who experience an event in a
given time interval \([t_i, t_{i+1}]\), \(i=1,\ldots,K\); the number of individuals exposed to risk at the start of each interval; and the number of censored observations for each interval. The term censored is used in the following sense: Suppose all individuals are followed or observed for at least \(\tau\) units of time; at the close of the study some percentage of the individuals will not have experienced the event. The individuals not experiencing the event by the end of the study are termed censored. (See Chapter I.) The data for each time interval can be regarded as a 3-tuple: \((N_i, D_i, W_i)\), \(i=1,\ldots,K\), denoting the number of exposed to risk, the number experiencing events, and the number of censored observations in the \(i\)-th interval. The data can also consist of a set of indicator variables specifying whether an individual survived, died, or was censored for each interval.

The model for the hazard rate underlying the occurrence of the event is assumed to be that of equation (4.2.4); events are assumed to occur independently among the individuals.

Under the assumed hazard model the conditional probability of the \(j\)-th individual surviving the \(i\)-th interval is:

\[
p_{ij} = \exp(-\exp(Z_{ij}\beta) \int_{t_i}^{t_{i+1}} \lambda_0(t)\,dt) . \tag{4.3.1}
\]

Allowing the underlying hazard, \(\lambda_0(t)\), to remain arbitrary, and viewing \(\exp(Z_{ij}\beta)\) as a proportional risk factor for the \(j\)-th individual, equation (4.3.1) can be written as

\[
p_{ij} = p_i \exp(Z_{ij}\beta) , \tag{4.3.2}
\]

where

\[
p_i = \exp(-\int_{t_i}^{t_{i+1}} \lambda_0(t)\,dt) .
\]
Following the suggestion of Prentice and Gloeckler (1978), it is assumed that the likelihood contribution for an individual who is censored in an interval is $p_{ij}^{1/2}$. This assumes that censoring occurs at random within an interval.

The likelihood for the data under the assumed hazard model is

$$ L(p, \beta; S) = \prod_{i=1}^{K} \prod_{j=1}^{N_i} p_i^{\psi_{ij}} (1-p_i)^{\delta_{ij}} \gamma_j^{\alpha_{ij}} p_i^{1/2}, \quad (4.3.2) $$

where: $\gamma_j = \exp(Z_j \beta)$,

$$ \psi_{ij} = \begin{cases} 1 & \text{if the } j\text{-th individual survived the } i\text{-th time interval.} \\ 0 & \text{otherwise} \end{cases} $$

$$ \delta_{ij} = \begin{cases} 1 & \text{if the } j\text{-th individual died in the } i\text{-th time interval.} \\ 0 & \text{otherwise} \end{cases} $$

$$ \alpha_{ij} = \begin{cases} 1 & \text{if the } j\text{-th individual was censored in the } i\text{-th time interval} \\ 0 & \text{otherwise} \end{cases} $$

$p = (p_1, \ldots, p_K)$

and $\psi_{ij} + \delta_{ij} + \alpha_{ij} = 1, \forall i, j$.

This likelihood is a reformulation of that given by Kalbfleisch and Prentice (1973). The modification is based on the use of three indicator variables and the piecewise nature of the data. Using this likelihood, the maximum likelihood estimators of $\beta$ and $p$ are obtained by setting the partial derivatives of the log-likelihood

$$ L(p, \beta; S) = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \psi_{ij} \gamma_j \log p_i + \sum_{i=1}^{K} \sum_{j=1}^{N_i} \delta_{ij} \log(1-p_i) $$

$$ + \sum_{i=1}^{K} \sum_{j=1}^{N_i} \frac{1}{2} \alpha_{ij} \gamma_j \log p_i $$

$$ (4.3.4) $$
to zero and solving the likelihood equations simultaneously. The partial derivatives are:

\[
\frac{\partial L}{\partial p_i} = \sum_{j=1}^{N_i} \left[ \frac{Y_j}{p_i(1-p_i)} \right] F_{ij},
\]

(4.3.5)

and

\[
\frac{\partial L}{\partial \beta} = \sum_{i=1}^{K} \sum_{j=1}^{N_i} \left[ \frac{Z_{ij} \delta_j \log p_i}{(1-p_i)} \right] F_{ij}.
\]

(4.3.6)

The term

\[
F_{ij} = (\psi_{ij} + \frac{1}{2} \alpha_{ij}) - (\psi_{ij} + \frac{1}{2} \alpha_{ij} + \delta_{ij}) p_i.
\]

It is interesting to note that both (4.3.5) and (4.3.6) are sums of weighted differences: The differences given by the \( F_{ij} \) are, with respect to the given model, of the form:

(Survivors - Expected Survivors).

To demonstrate the above claim, assume \( \beta = 0 \). This assumption allows us to rewrite (4.3.5) as

\[
\sum_{j=1}^{N_i} (p_i(1-p_i))^{-1} [(\psi_{ij} + \frac{1}{2} \alpha_{ij}) - (\psi_{ij} + \frac{1}{2} \alpha_{ij} + \delta_{ij})(p_i)] .
\]

Summation over \( j \) gives

\[
(p_i(1-p_i))^{-1} [(S_i + \frac{1}{2} W_i) - (S_i + \frac{1}{2} W_i + D_i)p_i].
\]

The maximum likelihood estimate for \( p_i \) is then

\[
p_i = (S_i + \frac{1}{2} W_i)/(S_i + \frac{1}{2} W_i + D_i).
\]

Under the assumed model, when \( \beta = 0 \), the conditional probability of survival is estimated by the above. This estimator implies that the numerator gives the number of survivors and the denominator the number
exposed to risk: Further, the denominator is the same 'effective-exposed-to-risk' defined for the actuarial estimator of the conditional probabilities by equation (1.5.2) of Chapter I. Hence,

\[(S_i + \frac{1}{2} W_i) - (S_i + \frac{1}{2} W_i + D_i)p_i\]

is of the form

\[(\text{Survivors}) - (\text{Expected Survivors}).\]

When $\beta \neq 0$, we are essentially taking the differences at the individual rather than the group level given above. The weights are the bracketed terms in (4.3.5) and (4.3.6). Estimates for $p$ and $\beta$ are obtained by some iterative algorithm.

Up to this point, the approach here offers no easier method for obtaining parameter estimates than the model of section 4.2. Unfortunately, for continuous covariates, this is the case, but when (0-1) indicator type covariates are used, an approximate solution for $\beta$ can be obtained. This approximation will now be presented.

4.4. An Approximate Solution for the Two Sample Problem

For large samples and (0,1) type covariates, the likelihood (4.3.4) offers an explicit solution for $\beta$ and $p$. In this section the two sample problem will be used to illustrate the procedure. Application to the $r$-sample case will be given in a later section.

The parameter $\beta$ is a scalar in the two sample case, and the covariate $Z_j$ takes on values 0 or 1 indicating whether $j$-th individual is in population one or population two. The likelihood (4.3.4) can be factored into two parts in this case. The first part represents the contribution to the likelihood made by individuals in population one, and the second part, the contribution of those in population two.
Using the above statements concerning the regression parameter and covariates in conjunction with the proportional hazards model gives the following relationship between the conditional probabilities of the two populations:

\[ p_{2i} = p_{1i} e^\beta, \quad i=1, \ldots, K. \quad (4.4.1) \]

The relationship (4.4.1) views population one in the sense of a standard with hazard rate \( \lambda_0(t) \). The hazard for population two is \( \lambda_0(t) \exp(\beta) \).

One approximation for \( \beta \) is arrived at through equation (4.4.1). The rationale for such an approximation is as follows. Let \( N_{mi}, S_{mi}, D_{mi}, W_{mi} \) denote the numbers: at risk, surviving, dying, and censored, respectively, for the \( i \)-th time interval, \( m \)-th population. Using this notation and the relationship (4.4.1), the log-likelihood (4.3.4) becomes:

\[
L(p_1, \beta; \mathcal{S}) = \sum_{i=1}^{K} \left\{ S_{2i} e^\beta \log p_{1i} + D_{2i} \log(1-p_{1i}) + \frac{1}{2} W_{2i} \log p_{1i} + S_{1i} \log p_{1i} + D_{1i} \log(1-p_{1i}) - \frac{1}{2} W_{1i} \log p_{1i} \right\}. \quad (4.4.2)
\]

The partial derivative of \( L(p_1, \beta; \mathcal{S}) \) with respect to \( \beta \) is

\[
\frac{\partial L}{\partial \beta} = \sum_{i=1}^{K} \left\{ (S_{2i} + \frac{W_{2i}}{2} - (S_{2i} + \frac{W_{2i}}{2} + D_{2i}) p_{1i} e^\beta) \frac{e^\beta \log p_{1i}}{1-p_{1i}} \right\}. \quad (4.4.3)
\]

Suppose that \( 0<p_i<1 \), and that \( \beta \) can change or take on a different value in each time interval. Let \( \beta_i, \quad i=1, \ldots, K \) denote the \( K \) values. Substituting the \( \beta_i \) into the likelihood (4.4.2), taking the derivative and setting it to zero yields
\[ \hat{\beta}_i = \log(-\log \left( \frac{S_{2i} + \frac{W_{2i}}{2}}{S_{2i} + \frac{W_{2i}}{2} + D_{2i}} \right)) - \log(-\log(p_{1i})). \] (4.4.4)

A reconsideration of the likelihood shows that it could have been written as

\[ \mathcal{L}(p_{1i}, p_{2i}; S) = \sum_{i=1}^{K} S_{2i} \log(p_{2i}) + \frac{W_{2i}}{2} \log(p_{2i}) + D_{2i} \log(1-p_{2i}) + S_{1i} \log(p_{1i}) + \frac{W_{1i}}{2} \log(p_{1i}) + D_{1i} \log(1-p_{1i}). \] (4.4.5)

The maximum likelihood estimates for the \( p_{mi} \), \( m = 1, 2 \), \( i = 1, \ldots, K \) are

\[ \hat{p}_{mi} = \frac{S_{mi} + \frac{W_{mi}}{2}}{S_{mi} + \frac{W_{mi}}{2} + D_{mi}}. \] (4.4.6)

From (4.4.6) and (4.4.2), it can be inferred that the \( \hat{\beta}_i \) are approximately the log-log differences of the conditional probabilities. The approximation to \( \beta \), the constant risk factor, can be found through averaging the \( \hat{\beta}_i \). Edwards (1972) gives a theorem stating that an approximate maximum likelihood estimator for combined samples can be obtained by using a weighted average of the maximum likelihood estimates from the individual samples. The weights in this case are the observed informations: The negatives of the second partial derivatives of the log-likelihood; evaluated at the maximum likelihood estimate. Edwards shows that the combined sample information is approximately equal to the sum of the observed individual informations.

The equations for the \( i \)-th information associated with \( \hat{\beta}_i \), \( i = 1, \ldots, K \) is:
\[
\frac{\partial^2}{\partial \beta_i^2} = A_{2i} e^{\beta_i} p_{1i} \log p_{1i} \left[ \frac{\log p_{1i}}{(1-p_{1i})^2} - \frac{1}{(1-p_{1i})} \right], \tag{4.4.7}
\]

where: \( A_{2i} = (S_{2i} + \frac{1}{2} W_{2i}) - (S_{2i} + \frac{1}{2} W_{2i} + D_{2i}) \).

The weighted average in this case is somewhat complex and will not be given. A quick estimate could be obtained by using an unweighted average. This approach also allows an examination of the variability between the \( \hat{\beta}_i \) and can be used to check the validity of a constant risk factor.

A second approximation to \( \beta \) can be obtained through the following:

\[
p_{1i} e^{\beta} = 1 - e^{\beta} q_{1i}, \tag{4.4.8}
\]

with \( q_{1i} = 1-p_{1i} \). Using (4.4.8) gives

\[
\hat{\beta} = \log \left\{ \frac{\sum_{i=1}^{K} \frac{D_{2i}}{W_{2i}}}{\sum_{i=1}^{K} q_{1i}(S_{2i} + \frac{W_{2i}}{2} + D_{2i})} \right\}. \tag{4.4.9}
\]

The \( q_{1i} \) must be obtained using (4.4.6). Here \( \hat{\beta} \) is the log of the ratio of the observed events in population two to the events expected in population two if the schedule of probabilities of an event in the first population held in population two. This ratio is essentially the log of the standardized mortality ratio (see Shryock and Siegel, 1971).

An outline for obtaining an approximation to \( \beta \) using the above procedures is as follows:

1. For each of the two populations obtain the conditional probabilities using (4.4.6).
2. Estimate the $\hat{\beta}_i$ using the probabilities found in step 1 by forming the log-log differences.

3. Examine the $\hat{\beta}_i$ for the plausibility of a single constant risk factor.

4. Assuming a single risk factor can be used, estimate $\beta$ using either the average $\hat{\beta}_i$ or equation (4.4.9).

Asymptotic variances can be obtained using the Taylor's linearization and the binomial variances associated with the conditional probabilities. Using this approach:

$$\text{Var}(\hat{\beta}_i) = \sum_{i=1}^{2} (p_{mi}\log(p_{mi}))^{-2}\text{Var}(p_{mi}), \quad (4.4.10)$$

with

$$\text{Var}(p_{mi}) = p_{mi}(1-p_{mi})/(N_{2i} - \frac{1}{2}W_{2i}).$$

If $\hat{\beta}$ is obtained by a weighted average of the $\hat{\beta}_i$, and letting $\lambda_i$ denote the weights, then:

$$\text{Var}(\hat{\beta}) = \sum_{i=1}^{K} \lambda_i^2\text{Var}(\hat{\beta}_i)/\left(\sum_{i=1}^{K} \lambda_i\right)^2. \quad (4.4.11)$$

The approximate variance for equation (4.4.9) is

$$\text{Var}(\hat{\beta}) = \sum_{i=1}^{K} \sum_{i=1}^{K} D_{2i} N_{2i}^2 (q_{2i}(1-q_{2i}))$$

$$+ \left\{ \sum_{i=1}^{K} q_{1i} N_{2i}^i \right\} N_{2i}^2 q_{1i}(1-q_{1i})/N_{1i}^i. \quad (4.4.12)$$

The $N_{2i}^i$ and $N_{1i}^i$ are the initial numbers exposed to risk (see Elandt-Johnson, 1977) and are given by

$$N_{mi}^i = N_{mi} - \frac{1}{2}W_{mi}.$$
We now turn to a brief discussion of the above procedure when r-samples are available. Examples of applying the above methods follows the r-sample problem.

4.5. Approximate Solution to the r-sample Problem

The procedure of section 4.4 readily generalizes to the case of r-samples. We assume that there are r-treatment groups or other such groups or populations under study; for convenience let group one be a control or standard in some sense. Using the proportional hazards model (r-1) relationships similar to (4.4.1) can be constructed of the form:

\[ p_{mi} = p_{1i}^{\beta_m} , \quad m=2, \ldots, r . \]  

(4.5.1)

Estimates of the \( \hat{\beta}_m \) can be obtained in a manner similar to that given in section 4.4: Using either a weighted average of the \( \hat{\beta}_{mi} \) associated with each interval \( i=1, \ldots, k \) or using formula (4.4.9).

Since the groups or populations under study are assumed independent the variance-covariance matrix \( \Sigma \) of the vector \( \hat{\beta} = (\hat{\beta}_2, \ldots, \hat{\beta}_r) \) will be diagonal. The elements along the main diagonal can be obtained using either equation (4.4.11) or (4.4.12). Comparisons can be made using hypotheses of the form

\[ H_0: \; C\hat{\beta} = 0 , \]

where \( C \) is an indicator contrast matrix specifying which elements of \( \hat{\beta} \) are to be considered. The appropriate test statistic is given by

\[ Q = (C\hat{\beta})' [C\Sigma C]^{-1} (C\hat{\beta}) . \]  

(4.5.2)

\( Q \) is distributed asymptotically as a chi-square variate with \( q = \text{rank}(C) \) degrees of freedom.
The methods of sections 4.4 and 4.5, although not a perfect solution to the problem of obtaining quick and easy estimates of the parameters in a life table regression, do afford some savings for (0-1) type covariates. The given methods also allow for an examination of the proportional hazards assumption as well as a convenient summary statistic, $\hat{\beta}$, instead of an entire set of life tables. If all that is available are the several life tables, then the given procedure will provide estimates of the regression coefficients without the original data.

In the following section we provide some results of applying the given procedure of the last two sections. We then turn our attention back to complex probability samples and life table regressions.

4.6. Application of the Method Developed in Section 4.4

Two sets of data were used for examining the proposed approach. The first of these data sets consists of survival data on Drosophila melanogaster by age and sex: taken from Chiang (1968, page 216). The second set of data is that of Prentice (1973) as presented by Holford (1976). This second data set concerns two treatments for lung cancer patients. These data sets are given in Tables 4.1 and 4.2.

For each of the data sets the methods of Kalbfleisch and Prentice (1973), Holford (1976), and those given in this paper were used to estimate $\hat{\beta}$. Maximum likelihood estimation was carried out using the FORTRAN subroutine MAXLIK (Kaplan and Elston, 1972). Tables 4.3 and 4.4 present the results.

As can be seen in both Tables 4.3 and 4.4, the values of the regression coefficients and the test statistics vary considerably in some cases, but the inferences drawn are the same for all the approaches.
The variability in the estimates and test values may be due to differences in assumptions between the approaches, and in some cases we are using approximations. Holford's method assumes that all events occur at the midpoint of an interval; the model of Kalbfleisch and Prentice does not use the incomplete information of censored observations in an interval; instead, for individuals who were censored, all that is assumed known is that they survived to the beginning of the interval under consideration. For example, if an individual was censored in the i-th interval \([t_i, t_{i+1})\), the information used is the knowledge of survival to \(t_i\); hence this individual would not be in the risk set associated with the i-th interval. The approach given by this paper is to assume that withdrawals or censoring occurred uniformly within an interval. Other reasons for differences between the Kalbfleisch and Prentice approach and that proposed in this paper may be due to a reparameterization of the Kalbfleisch and Prentice model suggested by Prentice and Gloeckler (1978). The suggested reparameterization involves the use of \(\log(-\log(p_{ij}))\) to remove any range restrictions on the parameter values; the reparameterization might also give the likelihood a shape closer to that of the multinormal distribution to obtain better convergence properties.

The results, though not definitive nor exhaustive, do indicate that the likelihood (4.4.4) can be used to obtain parameter estimates and inferences similar to those given by other approaches. An advantage of (4.4.4) is that both types of covariates, continuous and discrete, can be used; in the case of (0-1) covariates however, the approach given here allows an approximation to the regression coefficients which other methods do not. The results indicate that this approximation
works well, at least for the data sets used; however, moderate to large data sets are required.

The proposed method is in no way designed to replace the other methods, but rather to augment them in some situations; for example, to obtain initial estimates. The safest approach, as usual, is to examine the given problem and choose a method based on this examination.

4.7. Life Table Regression in Complex Probability Samples

A problem confronts the investigator who wants to apply the likelihood methods, given in the previous sections, to data gathered by a complex probability sample. The difficulty centers on how the likelihood is constructed, and how to encompass the survey design into this construction. Two solutions to this problem are given in this section, and in the following chapter, the problem of maximum likelihood in complex probability samples is more thoroughly examined.

One approach for using the hazard model (4.2.4) of Cox (1972) is through the weighted least squares method of Grizzle, Starmer, and Koch (1969). An example of using this technique for survival data is found in Koch, Johnson, and Tolley (1972). The data necessary for this approach is a contingency table of the form given in Figure 4.1. The populations \(1, \ldots, r\) are defined by cross-classification of the various levels of the covariates. The covariates are categorical in nature, or in the case of continuous covariates, a categorization has been imposed. The analytic posture taken by this approach is that interest is in differences between the various populations; not in an assumed underlying stochastic process which might have generated the observed data. The analysis is at the group or macro level rather than the individual or micro level.
Figure 4.1: Data for the Koch, Johnson and Tolley Approach

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Population One</th>
<th>...</th>
<th>Population r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exposed</td>
<td>Died</td>
<td>Withdrew</td>
</tr>
<tr>
<td>([t_1,t_2)]</td>
<td>(N_{11})</td>
<td>(D_{11})</td>
<td>(W_{11})</td>
</tr>
<tr>
<td>([t_2,t_3)]</td>
<td>(N_{12})</td>
<td>(D_{12})</td>
<td>(W_{12})</td>
</tr>
<tr>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
<td>(\vdots)</td>
</tr>
<tr>
<td>([t_K,t_{K+1})]</td>
<td>(N_{1K})</td>
<td>(D_{1K})</td>
<td>(W_{1K})</td>
</tr>
</tbody>
</table>

The conditional probability of survival

\[ p_{mi} = \Pr(\text{Individuals in the } m\text{-th group survive the } i\text{-th interval survival of the individuals to the beginning of the } i\text{-th interval}) \]

\(i = 1, \ldots, K\) and \(m = 1, \ldots, r\). Letting

\[ p' = (p_{11}, \ldots, p_{1K}, \ldots, p_{r1}, \ldots, p_{rK}) \]

denote the vector of these conditional probabilities, a linear model is fit to the \(\log(-\log)\) transformation of the vector \(p\). This approach is similar to that given in sections 4.4 and 4.5. The model here is

\[ F(p) = \log(-\log(p)) = \chi^T \beta \quad (4.7.1) \]

The matrix \(\chi\) is a design matrix of contrasts between the various populations.

For example, suppose there is survival information for \(r=4\) subpopulations or domains, and \(K=2\) time intervals. Let the domains be defined by the cross-classification of two age groups, \([0,25); [25,45]\), with two racial classifications, black and white. The domains or subpopulations are:
\[
\begin{align*}
r=1 &: \text{Black, } [0,25) \\
r=2 &: \text{Black, } [25,45) \\
r=3 &: \text{White, } [0,25) \\
r=4 &: \text{White, } [25,45) .
\end{align*}
\]

Associated with each subpopulation are two conditional probabilities denoted by \( p_{ri} \), \( i=1,2 \). The vector of such probabilities for the above four subpopulations is

\[
p' = (p_{11}, p_{12}, p_{21}, p_{22}, p_{31}, p_{32}, p_{41}, p_{44}) .
\]

The conditional probabilities under the Cox (1972) proportional hazards model

\[
\lambda(t; x) = \lambda_0(t) \exp(x_{\beta})
\]

are defined by the equation

\[
p_{ri} = \exp(-\exp(x_{\beta}) \int_{t_i}^{t_{i+1}} \lambda_0(u) du) .
\]

Note that here \( x \) is a design matrix and is substituted in place of \( Z \) which denoted a vector of covariables.

The log(-log) transformation of the \( p_{ri} \) gives a relationship which is linear in the covariates. The equation

\[
\log(-\log(p_{ri})) = \log(\int_{t_i}^{t_{i+1}} \lambda_0(u) du) + x_{\beta}
\]

specifies this relationship for the \( r \)-th domain and the \( i \)-th time interval. Using the Grizzle, Starmer, and Koch (1969) methodology, a linear model is fit to the log-log transformed conditional probabilities of survival. A variety of linear models could be used to analyze the data of this problem. One such model, which will be discussed here, is a hierarchical nested design. Such a model can be constructed so that
parameter estimates of factor effects can be based upon within time
interval comparisons. A model of this type is congruent with an
approach to the data where the proportional hazards relationship is not
assumed to hold. Hypotheses based upon contrast matrices can be used
to examine factor effects and the proportional hazards assumption. For
the example given, such a model is:

\[
F(p) = \log\log \begin{bmatrix}
P_{11} & 1 & 0 & 1 & 0 & 1 & 0 \\
P_{12} & 0 & 1 & 0 & 1 & 0 & 1 \\
P_{21} & 1 & 0 & -1 & 0 & -1 & 0 \\
P_{22} & 0 & 1 & 0 & -1 & 0 & -1 \\
P_{31} & -1 & 0 & 1 & 0 & -1 & 0 \\
P_{32} & 0 & -1 & 0 & 1 & 0 & -1 \\
P_{41} & -1 & 0 & -1 & 0 & 1 & 0 \\
P_{42} & 0 & -1 & 0 & -1 & 0 & 1 \\
\end{bmatrix}
\times \begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\beta_4 \\
\beta_5 \\
\beta_6 \\
\end{bmatrix}
\]  \hspace{1cm} (4.7.2)

The regression parameters \( \beta_1, \ldots, \beta_6 \) have the following interpretation:

- \( \beta_1 \) - is a race effect within time interval one.
- \( \beta_2 \) - is a race effect within time interval two.
- \( \beta_3 \) - is an age effect within time interval one.
- \( \beta_4 \) - is an age effect within time interval two.
- \( \beta_5 \) - is the age x race interaction within time interval one.
- \( \beta_6 \) - is the age x race interaction within time interval two.

Hypotheses tests of the form \( C\beta = 0 \) can be given to test if the
various factors have non-zero effects. For example,

\[
C = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
\]
can be employed to test for a race effect, and

\[ C = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \]

and

\[ C = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \]

can be used to test for age effects and age x race interactions, respectively. An overall test of the proportional hazards model is given by

\[ C = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} \]

A major drawback of such models is the unmanageable size of the design matrix when a large number of subpopulations and time intervals are being studied. The above approach does offer a great deal of flexibility with respect to model fitting and hypothesis testing.

The variance-covariance matrix for the data given in Figure 4.1 must be estimated using one of the four variance estimation procedures presented in Chapter II. The structure of the covariance can be assumed to be that specified in Chapter III, section 3.4; that is, the variance-covariance will be a block diagonal matrix. Estimation of the regression coefficients is carried out by entering the vector \( \beta \) and the variance-covariance matrix into the computer program GENCAT (Landis, Stanish, Freeman, and Koch, 1976). This program uses the method of weighted least squares for estimation, and Wald (1943) test statistics (see Grizzle, Starmer, and Koch, 1969).
A second solution to the problem is found using the methods of sections 4.4 and 4.5. Again, the variance-covariance estimate of \( p \) must be obtained using one of the four methods given in Chapter II. This approach and that outlined immediately above are essentially the same.

Problems with the two solutions given above are

1. Both require relatively large data sets and can be used only with categorical type covariates.

2. Both methods focus on group behavior, not that of an individual.

The requirement of a large data set cannot always be satisfied, particularly for various subpopulations. Small cell counts or zero cell counts are often encountered which will not allow the log(-log) transformation. To overcome such problems a model imposing more structure on the data is necessary.

If a continuous covariate is available, there is usually some loss of information by categorizing, and often relevant categories cannot be defined.

One solution to some of these problems is through the use of the life table regression models and the method of maximum likelihood estimation. Chapter V outlines a possible approach for using these techniques in complex probability samples.
Table 4.1: Number of Adult Drosophila Melanogaster Living and Number Dying and Sex\(^{(1)}\)

<table>
<thead>
<tr>
<th>DAYS</th>
<th>N(_1i)</th>
<th>D(_1i)</th>
<th>P(_1i)</th>
<th>N(_2i)</th>
<th>D(_2i)</th>
<th>P(_2i)</th>
<th>(\hat{\beta}_i) (^{(2)})</th>
<th>SE((\hat{\beta}_i)) (^{(3)})</th>
<th>(\hat{\beta}) (\text{SE}(\hat{\beta}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5</td>
<td>270</td>
<td>2</td>
<td>0.9926</td>
<td>275</td>
<td>4</td>
<td>0.9855</td>
<td>0.6784</td>
<td>0.8660</td>
<td>0.78</td>
</tr>
<tr>
<td>5-10</td>
<td>268</td>
<td>4</td>
<td>0.9851</td>
<td>271</td>
<td>7</td>
<td>0.9742</td>
<td>0.5540</td>
<td>0.6268</td>
<td>0.88</td>
</tr>
<tr>
<td>10-15</td>
<td>264</td>
<td>3</td>
<td>0.9886</td>
<td>264</td>
<td>3</td>
<td>0.9886</td>
<td>0.0000</td>
<td>0.8165</td>
<td>0.00</td>
</tr>
<tr>
<td>15-20</td>
<td>261</td>
<td>7</td>
<td>0.9732</td>
<td>261</td>
<td>7</td>
<td>0.9732</td>
<td>0.0000</td>
<td>0.5345</td>
<td>0.00</td>
</tr>
<tr>
<td>20-25</td>
<td>254</td>
<td>3</td>
<td>0.9882</td>
<td>254</td>
<td>13</td>
<td>0.9488</td>
<td>1.4866</td>
<td>0.6405</td>
<td>2.32</td>
</tr>
<tr>
<td>25-30</td>
<td>251</td>
<td>3</td>
<td>0.9880</td>
<td>241</td>
<td>22</td>
<td>0.9087</td>
<td>2.0746</td>
<td>0.6155</td>
<td>3.37</td>
</tr>
<tr>
<td>30-35</td>
<td>248</td>
<td>16</td>
<td>0.9355</td>
<td>219</td>
<td>31</td>
<td>0.8584</td>
<td>0.8279</td>
<td>0.3079</td>
<td>2.69</td>
</tr>
<tr>
<td>35-40</td>
<td>232</td>
<td>66</td>
<td>0.7155</td>
<td>188</td>
<td>68</td>
<td>0.6383</td>
<td>0.2935</td>
<td>0.1738</td>
<td>1.69</td>
</tr>
<tr>
<td>40-45</td>
<td>166</td>
<td>36</td>
<td>0.7831</td>
<td>120</td>
<td>51</td>
<td>0.5750</td>
<td>0.8170</td>
<td>0.2191</td>
<td>3.73</td>
</tr>
<tr>
<td>45-50</td>
<td>130</td>
<td>54</td>
<td>0.5846</td>
<td>69</td>
<td>38</td>
<td>0.4493</td>
<td>0.3991</td>
<td>0.2161</td>
<td>1.85</td>
</tr>
</tbody>
</table>

\(^{(1)}\)Data are from Chiang (1968), page 216.

\(^{(2)}\)Computed using equation (4.4.10).

\(^{(3)}\)Computed using the linear Taylor's approximation.
<table>
<thead>
<tr>
<th>DAYS</th>
<th>$N_{1i}$</th>
<th>$D_{1i}$</th>
<th>$W_{1i}$</th>
<th>$p_{1i}$</th>
<th>$N_{2i}$</th>
<th>$D_{2i}$</th>
<th>$W_{2i}$</th>
<th>$p_{2i}$</th>
<th>$\hat{\beta}_i$ (2)</th>
<th>SE($\hat{\beta}_i$) (3)</th>
<th>$\frac{\hat{B}}{SE(\hat{B})}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-20</td>
<td>64</td>
<td>15</td>
<td>0</td>
<td>0.7656</td>
<td>60</td>
<td>14</td>
<td>0</td>
<td>0.7667</td>
<td>-0.0051</td>
<td>0.3726</td>
<td>-0.01</td>
</tr>
<tr>
<td>24-40</td>
<td>49</td>
<td>6</td>
<td>1</td>
<td>0.8763</td>
<td>46</td>
<td>11</td>
<td>0</td>
<td>0.7609</td>
<td>0.7273</td>
<td>0.5066</td>
<td>1.44</td>
</tr>
<tr>
<td>41-60</td>
<td>42</td>
<td>7</td>
<td>0</td>
<td>0.8333</td>
<td>35</td>
<td>10</td>
<td>0</td>
<td>0.7143</td>
<td>0.6127</td>
<td>0.4942</td>
<td>1.24</td>
</tr>
<tr>
<td>61-80</td>
<td>35</td>
<td>2</td>
<td>0</td>
<td>0.9428</td>
<td>25</td>
<td>4</td>
<td>0</td>
<td>0.8400</td>
<td>1.0862</td>
<td>0.8665</td>
<td>1.25</td>
</tr>
<tr>
<td>81-100</td>
<td>33</td>
<td>4</td>
<td>2</td>
<td>0.8750</td>
<td>21</td>
<td>6</td>
<td>2</td>
<td>0.7000</td>
<td>0.9825</td>
<td>0.6350</td>
<td>1.54</td>
</tr>
<tr>
<td>101-150</td>
<td>27</td>
<td>12</td>
<td>1</td>
<td>0.5472</td>
<td>13</td>
<td>5</td>
<td>1</td>
<td>0.6000</td>
<td>-0.1659</td>
<td>0.5299</td>
<td>-0.31</td>
</tr>
<tr>
<td>151-200</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>0.4815</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0.7143</td>
<td>-0.7757</td>
<td>0.8054</td>
<td>-0.96</td>
</tr>
<tr>
<td>201-250</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>0.5000</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0.3333</td>
<td>-0.4606</td>
<td>0.8236</td>
<td>-0.56</td>
</tr>
</tbody>
</table>

1 Data are from Holford (1976).

2 Computed using equation (4.4.10).

3 Computed using the linear Taylor's approximation.
<table>
<thead>
<tr>
<th>METHOD OF ESTIMATION</th>
<th>$\hat{\beta}$</th>
<th>$\hat{SE}(\hat{\beta})$</th>
<th>$\frac{\hat{\beta}}{\hat{SE}(\hat{\beta})}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holford</td>
<td>0.2679</td>
<td>0.0967</td>
<td>2.77</td>
</tr>
<tr>
<td>Kalbfleisch and Prentice</td>
<td>0.3841</td>
<td>0.1102</td>
<td>3.47</td>
</tr>
<tr>
<td>Likelihood (4.4.3)</td>
<td>0.5888</td>
<td>0.0977</td>
<td>6.03</td>
</tr>
<tr>
<td>Approximation (4.4.4)</td>
<td>0.5555</td>
<td>0.0983</td>
<td>5.65</td>
</tr>
<tr>
<td>variances of weights</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation (4.4.4)</td>
<td>0.7131</td>
<td>0.1761</td>
<td>4.05</td>
</tr>
<tr>
<td>unweighted average</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Approximation (4.4.9)</td>
<td>0.5256</td>
<td>0.0841</td>
<td>6.26</td>
</tr>
</tbody>
</table>
Table 4.4: Estimated Regression Parameters and Their Estimated Standard Errors by the Method of Estimation Using Data from Holford (1976)

<table>
<thead>
<tr>
<th>METHOD OF ESTIMATION</th>
<th>$\hat{\beta}$</th>
<th>$SE(\hat{\beta})$</th>
<th>$\frac{\hat{\beta}}{SE(\hat{\beta})}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holford</td>
<td>0.2436</td>
<td>0.2000</td>
<td>1.22</td>
</tr>
<tr>
<td>Kalbfleisch and Prentice</td>
<td>0.3026</td>
<td>0.1940</td>
<td>1.56</td>
</tr>
<tr>
<td>Likelihood (4.4.3)</td>
<td>0.3040</td>
<td>0.1925</td>
<td>1.58</td>
</tr>
<tr>
<td>Approximation (4.4.4) variances of weights</td>
<td>0.2628</td>
<td>0.1711</td>
<td>1.53</td>
</tr>
<tr>
<td>Approximation (4.4.4) unweighted average</td>
<td>0.3653</td>
<td>0.2305</td>
<td>1.58</td>
</tr>
<tr>
<td>Approximation (4.4.9)</td>
<td>0.2924</td>
<td>0.1657</td>
<td>1.76</td>
</tr>
</tbody>
</table>
CHAPTER V

MAXIMUM LIKELIHOOD ESTIMATION IN
COMPLEX PROBABILITY SAMPLES

5.1. Introduction

The method of maximum likelihood estimation is applicable to data obtained by a simple random sample with replacement from a population with a known or assumed structure. The ability to use maximum likeli-
hood estimation in complex probability samples would allow the use of analytic methods like the life table regression methods presented in Chapter IV. The purpose of this chapter is to explore the use of maximum likelihood estimation in complex probability samples.

5.2. Use of Maximum Likelihood in Complex Probability Samples

Use of maximum likelihood estimation for survey data has been discussed by Basu (1978). Basu points out that if one adopts the approach that the sampling design generates the randomness in the data, then the maximum likelihood method fails to be very informative. The point is that the likelihood is not failing but merely reflecting what we already know. The central issue is the conceptual framework which is used for establishing the probabilistic or stochastic aspects of the data. Did the sampling design impose the probability structure, or was a sample taken from a stochastic process? Cassel, Särndal and Wretman (1977) address this issue as the choice between the fixed population approach or the superpopulation approach. They give the following definitions:
i. The fixed population approach: with each population unit is associated a fixed but unknown real number, that is, the value of the variable under study.

ii. The superpopulation approach: with each unit is associated a random variable for which a stochastic structure is specified. The actual value associated with a population unit is treated as the outcome of this random variable.

The superpopulation approach as stated allows quite readily for the use of likelihood methods, since it assumes that the observed data arose from some underlying stochastic process. When using the superpopulation approach, the sampling design is viewed only as a data gathering technique. The approach does not totally ignore the sampling design but how the design is to be incorporated into the approach is unspecified.

A point of controversy centers on the degree to which the sampling design should be incorporated into the estimation process. Some individuals have advocated total disregard for the sampling design in estimation, while others have maintained a view that the sampling design should govern the estimation process. The approach taken in this dissertation is that the superpopulation approach is acceptable, but that certain aspects of the design may effect inferences drawn from the data.

5.3. **Life Table Analysis and Superpopulations**

The life table is essentially a model depicting the occurrence of an event which has an associated waiting time. It is assumed that underlying the occurrence of the event is a stochastic process. For example, a model for the life table is the pure death process. However,
the exact nature of the underlying process can be left unspecified (Chiang, 1968). With this view of the life table, the superpopulation approach to the analysis of survey data is applicable.

5.4. **Likelihood Construction for Survey Data**

In this section arguments are given which allow for the construction of the likelihood under a given model for a variety of probability sampling designs.

The following notation and assumptions will be used for constructing a likelihood for survey data. Assume a probability sample of \( n \) individuals from a population of size \( N \). Let \( X_j \) denote a response of interest for the \( j \)-th individual in the sample, and let \( Z_j \) denote a vector of covariates measured on the \( j \)-th individual in the sample, \( j=1,...,n \). The data are denoted by \( S = (X,Z) \) where both \( X \) and \( Z \) are matrices in which the \( j \)-th row corresponds to the information for the \( j \)-th individual. For example, \( X_j \) could contain information on survival, death, or withdrawal for each time interval, and \( Z_j \) could be a set of variables such as age, sex, diet, and smoking history. Let \( \pi_j \) be the probability of selection and \( \omega_j \) the sample weight for the \( j \)-th individual, \( j=1,...,n \). Furthermore, it is assumed that any labels associated with the data are uninformative.

Two key assumptions for writing the likelihood are now given. The first is a relationship regarding the sample weights, and is

\[
\sum_{j=1}^{n} \omega_j = N. \tag{5.4.1}
\]

Equation (5.4.1) implies that the sample weights impart information as to the frequency with which an elementary unit with a specified set of
characteristics, not necessarily \( z_j \), appears in the population. Sampling designs which use selection probabilities proportional to size (Kish, 1965) satisfy equation (5.4.1). However, selection probabilities need not be proportional to size. The requirement is that the sampling weights satisfy equation (5.4.1).

The second key assumption for writing a likelihood based upon a given model is that the events or responses, denoted by \( x_j \), \( j = 1, \ldots, N \), occur independently among the population members, and that the events be independent of the sample design. These two assumptions allow the likelihood of the sample data to be written as a product of the individual probabilities of events as specified by the superpopulation model.

Suppose the superpopulation model which is assumed to govern the occurrence of \( x_j \) is denoted by \( f(x_j; \theta, z_j) \), where \( \theta \) is a \( p \)-dimensional vector of parameters associated with the process. Interest is in estimating \( \theta \) or a subset of the elements in \( \theta \) using the sample data \( S = (x, z) \). The likelihood for the \( j \)-th individual is

\[
L(\theta; x_j, z_j) = f(x_j; \theta, z_j)^{w_j} .
\]

The assumptions regarding the sampling weights and independence of the events allow the total likelihood for the sample data to be written as

\[
L(\theta; s) = \prod_{j=1}^{n} f(x_j; \theta, z_j)^{w_j} .
\]

Equation (5.4.3) will be referred to as a weighted likelihood. The log-likelihood is then given by

\[
L(\theta; s) = \sum_{j=1}^{n} w_j \log(f(x_j; \theta, z_j)) .
\]
The score statistic for the $i$-th parameter $\theta_i$, $i=1,...,p$ is defined as

$$\frac{dL(\theta:s)}{d\theta_i} = \sum_{j=1}^{n} \omega_j \frac{d}{d\theta_i} \log(f(x_j;\theta,z_j)) . \tag{5.4.5}$$

The estimate of $\theta$, denoted by $\hat{\theta}$, is found by simultaneously solving the $p$ equations of the form

$$\frac{dL(\theta:s)}{d\theta_i} = 0 , \ i = 1,...,p . \tag{5.4.6}$$

5.5. **Examples Using Weighted Likelihoods**

The following examples illustrate the use of the weighted likelihood.

**Example 1: Exponential Population**

$$f(x; \theta) = \begin{cases} \theta \exp(-\theta x), & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

using equations (5.4.4 - 5.4.6) gives

$$L(\theta:x) = \sum_{j=1}^{n} \left( \omega_j \log(\theta) - \omega_j \theta x_j \right) ,$$

$$\frac{dL}{d\theta} = \sum_{j=1}^{n} \frac{\omega_j}{\theta} - \sum_{j=1}^{n} \omega_j x_j .$$

The solution for $\hat{\theta}$ is

$$\hat{\theta} = \frac{\sum_{j=1}^{n} \omega_j}{\sum_{j=1}^{n} \omega_j x_j} .$$
Example 2: Normal Population \((\mu, \sigma^2), \theta = (\mu, \sigma^2)\)

\[
f(x; \theta) = \begin{cases} 
(2\pi\sigma^2)^{1/2} \exp\left(-\frac{1}{2\sigma^2} (x-\mu)^2\right), & -\infty < x < \infty \\
0, & \text{otherwise}
\end{cases}
\]

\[
L(\theta; x) = \sum_{j=1}^{n} \left[ -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (x_j - \mu)^2 \right] \omega_j
\]

\[
\frac{\partial L}{\partial \sigma^2} = \sum_{j=1}^{n} \left[ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} (x_j - \mu)^2 \right] \omega_j \tag{5.5.1}
\]

and

\[
\frac{\partial L}{\partial \mu} = -\sum_{j=1}^{n} \frac{\omega_j}{\sigma^2} (x_j - \mu). \tag{5.5.2}
\]

Setting (5.5.2) equal to zero and solving for \(\hat{\mu}\) gives

\[
\hat{\mu} = \frac{\sum_{j=1}^{n} \omega_j x_j}{\sum_{j=1}^{n} \omega_j}. \tag{5.5.3}
\]

Substituting (5.5.3) into (5.5.1), setting (5.5.1) to zero, and solving for \(\hat{\sigma}^2\) gives

\[
\hat{\sigma}^2 = \frac{\sum_{j=1}^{n} \omega_j (x_j - \hat{\mu})^2}{\sum_{j=1}^{n} \omega_j} \tag{5.5.4}
\]

Example 3: Normal Regression

Assume \(\sigma^2\) known, \(\mu = \alpha + \beta(z - \bar{z})\), and \(\theta = (\alpha, \beta)\).

\[
f(x; \theta, z) = \begin{cases} 
(2\pi\sigma^2)^{1/2} \exp\left(-\frac{1}{2\sigma^2} (x-\alpha - \beta(z - \bar{z}))^2\right), & -\infty < x < \infty \\
0, & \text{otherwise}
\end{cases}
\]

\[
L(\theta; S) = \prod_{j=1}^{n} \omega_j (x_j - \alpha - \beta(z_j - \bar{z}))^2. \tag{5.5.5}
\]
Solving the resulting likelihood equations for \( \hat{\alpha} \) and \( \hat{\beta} \) gives:

\[
\hat{\beta} = \frac{\left( \sum_{j=1}^{n} \omega_j \frac{x_j z_j - \sum_{j=1}^{n} \omega_j z_j}{\sum_{j=1}^{n} \omega_j} \right)}{\sum_{j=1}^{n} \omega_j (z_j - \bar{z})^2}
\]  

(5.5.6)

and

\[
\hat{\alpha} = \frac{\sum_{j=1}^{n} \omega_j x_j}{\sum_{j=1}^{n} \omega_j} .
\]

(5.5.7)

Example 4: Logistic Regression Model

In this case, \( \theta = \beta = (\beta_1, \ldots, \beta_p) \), and \( x_j = 0 \) or \( 1 \).

\[
L(\theta; S) = \frac{1}{n} \left( \sum_{j=1}^{n} \omega_j x_j z_j \beta - \sum_{j=1}^{n} \omega_j \log \{ 1 + \exp (z_j \beta) \} \right) .
\]

The score with respect to \( \beta_\lambda \), \( \lambda = 1, \ldots, p \) is

\[
\frac{\partial L}{\partial \beta_\lambda} = \frac{\sum_{j=1}^{n} \omega_j x_j z_j \beta - \sum_{j=1}^{n} \omega_j z_j \beta}{1 + \exp (z_j \beta)} .
\]

(5.5.7)

Estimates of \( \beta \) require an iterative procedure unless \( p = 1 \), and \( z_j \) is an indicator variable taking values 0 or 1.

Example 5: Single Parameter Exponential Family

\[
f(x; \theta) = \exp(A(\theta)B(x) + C(x) + D(\theta)) ,
\]

\[
L(\theta; x) = \left( \sum_{j=1}^{n} \omega_j (A(\theta)B(x_j) + C(x_j) + D(\theta)) \right) .
\]

The score statistic is:

\[
\frac{\partial L}{\partial \theta} = \sum_{j=1}^{n} \omega_j \frac{d}{d\theta} A(\theta)B(x_j) + \sum_{j=1}^{n} \omega_j \frac{d}{d\theta} D(\theta) .
\]

(5.5.8)
It is observed that any solution to (5.5.8) will involve a weighted sum of the statistic \( B(x) \) and the sum of the sample weights. A similar result can be shown for the multi-parameter exponential family.

The point to be emphasized is that the likelihood specified by equation (5.4.3) provides reasonable estimators, at least in the examples presented. The results for the exponential family indicate that the approach may work for a wide selection of models.

The estimates found for the examples are a type of Horvitz-Thompson estimators (Horvitz and Thompson, 1952). The Horvitz-Thompson estimator is discussed by Cassel, Särndal and Wretman (1977). They show that for a wide variety of sample designs the Horvitz-Thompson estimator given by

\[
    t_{HT} = \frac{\sum_{j=1}^{n} \omega_j x_j}{N}
\]

(5.5.9)

has certain optimal properties. For certain classes of sample designs \( t_{HT} \) is the unique hyperadmisable estimator, is design unbiased and has minimum variance (Cassel, Särndal and Wretman, 1977).

5.6. Variance Estimation for the Maximum Likelihood Estimator: The Single Parameter Case

The approach to the data here has assumed that the stochastic process (superpopulation model) underlying the data is independent of the sampling design. However, the design may affect the precision of the parameter estimates. How the sample design can be introduced into the estimation procedure is now examined. To aid this discussion, a review of certain aspects related to maximum likelihood estimation in simple random samples with replacement is given.
For the simple random sample with replacement, the variance of the maximum likelihood estimator is given by

\[
\text{Var}(\hat{\theta}) = \left[ \mathbb{E}\left( \frac{\partial^2 \ell(\theta; S)}{\partial \theta^2} \right) \right]^{-1}.
\]  

(5.6.1)

Often the expected value in (5.6.1) is too difficult to obtain, and the second partial derivative of the log-likelihood evaluated at \( \hat{\theta} \) is used for estimating \( \text{Var}(\hat{\theta}) \). A problem with equation (5.6.1) is that it disregards the design of the sample. As pointed out in Chapter II, disregard for the sample design can result in underestimation of the variance. The problem is one of obtaining a variance estimate for a function of the sample observations, analogous to the problem discussed in Chapter II. The current problem can be divided into two cases. The first is where the estimator has an explicit solution in terms of the sample data (see example 1, section 5.5). The second case is where an iterative solution to the likelihood equations must be employed. An example is that given by the logistic regression model (see example 4, section 5.5).

Where an explicit solution exists, the methods of balanced repeated replications (BRR), jackknifing (JRR, JK), or the Taylor's linearization (TL) can be directly applied to the data. Such methods could also be used in conjunction with an iterative estimation procedure as well, but could prove costly or cumbersome.

One possible solution to the problem of estimating a variance when an iterative parameter estimate is necessary is through the use of the score statistic

\[
S(\theta) = \frac{\partial \ell(\theta; S)}{\partial \theta}.
\]  

(5.6.3)
Two steps are necessary to obtain the $\text{Var}(\hat{\theta})$ using (5.6.3). The first step involves obtaining $\hat{\theta}$ based upon all the sample data. The second step is to obtain $\text{Var}(\hat{\theta})$ using any of the four variance estimation procedures of Chapter II (BRR, JRR, JK, or TL).

The following relationships can then be used to obtain an estimate of $\text{Var}(\hat{\theta})$. Assume a continuous random variable $X$, and a continuous function of $X$, $f$, which is bijective and twice differentiable. Now, define a new random variable

$$Y = f(X).$$

Let

$$E(X) = \mu_X$$

and

$$\text{Var}(X) = \sigma_X^2.$$

Then for large samples

$$E(Y) = f(\mu_X)$$

and

$$\text{Var}(Y) = \left(\frac{d}{dx} f \bigg|_{x=\mu_X}\right)^2 \sigma_X^2. \quad (5.6.4)$$

If for any $x$ in the domain of $f$ or, less restrictively, for certain neighborhoods of $\mu_X$

$$\frac{d}{dx} f \neq 0$$

then

$$x = f^{-1}(y)$$

and

$$\text{Var}(x) = \left[\frac{d}{dy} f^{-1} \bigg|_{y=E(Y)}\right]^2 \text{Var}(Y). \quad (5.6.5)$$
From calculus

\[ \frac{d}{dy} f^{-1} = \left[ \frac{d}{dx} f \right]^{-1}. \]  

(5.6.6)

Making the following substitutions

\[ \hat{\theta} = \chi \]
\[ S(\hat{\theta}) = f(\chi) = Y , \]

then from (5.6.5)

\[ \text{Var}(\hat{\theta}) = \left[ \frac{\partial S(\hat{\theta})}{\partial \theta} \right]^{-2} \text{Var}(S(\hat{\theta})) . \]  

(5.6.7)

The result is an expression which gives the variance of the maximum likelihood estimator in terms of the variance of the score statistic. The conditions placed upon \( f \), used for developing (5.6.7) are no more restrictive than the regularity conditions assumed when using maximum likelihood estimation.

For data from a simple random sample with replacement, the variance of the score can be written as

\[ \text{Var}(S(\hat{\theta})) = -\frac{\partial S(\hat{\theta})}{\partial \theta} \bigg|_{\theta = \hat{\theta}} . \]  

(5.6.8)

(Kendall and Stuart; Vol. 2, 1963).

Equation (5.6.8) is termed the observed information. If the expected value of the right hand side of (5.6.8) is used then the term information is employed. In large samples the observed information and information are approximately equal. Substituting equation (5.6.8) into (5.6.7) gives
\[
\text{Var}(\hat{\theta}) = \left[ \frac{\partial S(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \right]^2 \left[ \frac{\partial S(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \right]^{-1}
\]

(5.6.9)

which is the correct asymptotic variance for \( \hat{\theta} \).

Next it is shown that the use of equation (5.6.7) in complex sample survey results in inflating equation (5.6.9) by the design effect (DEFF) associated with the score and its variance.

Recall from Chapter II that

\[
\text{DEFF} = \frac{\text{Variance with respect to the true design}}{\text{Variance based upon a simple random sample of the same size}}
\]

(5.6.10)

with this definition of the DEFF, note that for a complex sample survey, equation (5.6.7) can be written as

\[
\text{Var}(\hat{\theta}) = \left[ \frac{\partial S(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \right]^2 \text{Var}_{CSS}(S(\hat{\theta}))
\]

\[
= \left\{ \left[ \frac{\partial S(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \right]^{-1} \text{Var}_{CSS}(S(\hat{\theta})) \right\} \left[ \frac{\partial S(\theta)}{\partial \theta} \bigg|_{\theta = \hat{\theta}} \right]^{-1}
\]

\[
= (\text{DEFF}) \text{Var}_{SRS}(\hat{\theta}), \quad (5.6.11)
\]

where \( \text{Var}_{CSS}(S(\hat{\theta})) \) denotes the variance of \( S(\hat{\theta}) \) for the complex sample survey, and \( \text{Var}_{SRS}(\hat{\theta}) \) denotes the variance of \( \hat{\theta} \) for a simple random sample. Thus, equation (5.6.7) has some intuitive appeal since for simple random samples, it provides the correct asymptotic variance, and for complex probability samples, it results in a relationship between the design effect and the simple random sample variance.
5.7. Variance Estimation for the Maximum Likelihood Estimator: The Multiparameter Case

When several parameters are involved, the estimation process becomes slightly more complex than that for the single parameter. However, the increase in complexity occurs only in the case when iterative methods must be used to obtain parameter estimates. The case where an explicit solution to the likelihood equations exists can be dealt with by any of: BRR, JRR, JK, or TL methods.

If the parameter values must be obtained iteratively, then the estimation process is more complicated. In this situation the variances and covariances can be obtained using a system of linear equations developed below.

The i-th score statistic associated with the i-th parameter \( \theta_i; i = 1, \ldots, p \) is defined as

\[
S_i(\theta) = \frac{\partial L(\theta; S)}{\partial \theta_i}.
\]  

(5.7.1)

The important characteristic of (5.7.1) is that it is not a function of \( \theta_i \) alone, but of the entire vector \( \theta \).

The matrix notation given below facilitates the development of the variance estimator. Let

\[
S(\hat{\theta})' = (S_1(\hat{\theta}), S_2(\hat{\theta}), \ldots, S_p(\hat{\theta}))
\]

be the vector of score statistics, where the prime denotes the transpose operation. Furthermore, let

\[
V(S(\hat{\theta}))
\]
denote the variance-covariance for \( S(\hat{\theta}) \), and

\[
V(\hat{\theta})
\]
the variance-covariance of the vector \( \hat{\theta} \). Also, let the matrix of partial derivatives of \( S(\hat{\theta}) \) be given by

\[
H(\hat{\theta}).
\]

With this notation, the large sample variance-covariance of \( S(\hat{\theta}) \) is

\[
V(S(\hat{\theta})) = H(\hat{\theta})V(\hat{\theta})H(\hat{\theta})'.
\]  

(5.7.2)

Assuming no functional relationships which might cause \( H(\hat{\theta}) \) to be singular, the variance-covariance of \( \hat{\theta} \) is given by

\[
V(\hat{\theta}) = H(\hat{\theta})^{-1} V(S(\hat{\theta})) [H(\hat{\theta})']^{-1}.
\]  

(5.7.3)

The matrix \( V(S(\hat{\theta})) \) can be found using

\[
\hat{V}(S(\hat{\theta})) = \frac{1}{L} \sum_{x=1}^{L} (S(\hat{\theta})_{x} - S(\hat{\theta}))(S(\hat{\theta})_{x} - S(\hat{\theta})).
\]  

(5.7.4)

where \( S(\hat{\theta})_{x} \) is the value of the vector \( S(\hat{\theta}) \) evaluated at the \( x \)-th BRR half-sample or jackknifed sample; \( x=1, \ldots , L \). The Taylor's linearization might also be used.

Design effects associated with equation (5.7.4) can be found by dividing the design based variances by the simple random sample variances based on the same sample size. A generalized design effect can be obtained using

\[
G = \left[ \frac{\text{Det}(V_{CSS})}{\text{Det}(V_{SRS})} \right]^{1/p},
\]  

(5.7.5)

where \( \text{Det}(V_{CSS}) \) and \( \text{Det}(V_{SRS}) \) represent the determinants of the variance-covariance matrices. Equation (5.7.5) implies that all parameters have the same design effect which may not be true. Rao and Scott (1979) give another approach to design effects using the
characteristic roots of the matrix

\[ \mathcal{D} = \mathcal{V}_{CSS} \mathcal{V}_{SRS}^{-1} \]  

(5.7.6)

They call the characteristic roots of \( \mathcal{D} \) generalized design effects.

5.8. **Application to Life Table Regression Problems**

Use of the methods developed in sections 5.5-5.7 are directly applicable to the problems of life table regression which generally require iterative solutions for the model parameters. As an example the methods of this chapter are applied to the likelihood given in Chapter IV by equation (4.3.2). Here, however, the sampling weights \( \omega_j \) are included, and it is assumed that withdrawals do not occur. The likelihood is

\[ L(\hat{\beta};S) = \prod_{i=1}^{k} \prod_{j=1}^{n_i} p_i^{e_j} \psi_j \omega_j \exp(Z_j'\hat{\beta}) \exp(Z_j'\beta) \delta_j \omega_j, \]  

(5.8.1)

and is based upon the hazard model

\[ \lambda(t) = \lambda_0(t) \exp(Z_t'\beta). \]  

(5.8.2)

The log-likelihood is

\[ L(\hat{\beta};S) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\{ \psi_j \omega_j \exp(Z_j'\hat{\beta}) \log(p_i) \right. \]  

\[ \left. + \delta_j \omega_j \log(1-p_i) \right\} \exp(Z_j'\beta). \]  

(5.8.3)

The score statistics in terms of \( \beta_\xi, \xi=1,\ldots,p \) and \( p_i, i=1,\ldots,k \) are
\[ \frac{\partial L}{\partial \beta_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[ \psi_{ij} Z_{ij} \frac{\exp(Z_{ij}^t \beta)}{1 - p_{ij}} \right] \] 

and

\[ \frac{\partial L}{\partial p_{ij}} = \sum_{j=1}^{n_i} \frac{\psi_{ij} \exp(Z_{ij}^t \beta)}{p_{ij} \exp(Z_{ij}^t \beta)} - \frac{\delta_{ij} \exp(Z_{ij}^t \beta) p_{ij}}{1 - p_{ij}} \]  

(5.8.4)

(5.8.5)

For this model an iterative scheme is necessary to solve for \( \beta \) and the \( p_{ij} \), \( i=1, \ldots, k \). Variances can be estimated using the method given in section 5.7. However, if the covariates are indicator type variables, and the \( p_{ij} \) are obtained using the expression (4.4.6), then the variances of the \( \beta \) can be solved for in terms of the variances of the \( p_{ij} \) (see section 4.4).

The data used in the application consists of 2,297 black women married at least five years. The data were taken from the 1973 National Survey of Family Growth.

Three regression models were used for examining the relationship between age at first marriage and the probability of separation. The three methods were: a simple linear regression; a logistic regression; and a life table regression.

Let \( Y \) be an indicator random variable defined as

\[ Y = \begin{cases} 1 & \text{If a couple separated} \\ 0 & \text{otherwise} \end{cases} \]

Let \( x \) denote the age at first marriage. The three models are
i. Linear Regression

\[ E(Y) = \alpha + \beta x \]

ii. Logistic Regression

\[ E(Y) = \left(1 + \exp(-(\alpha + \beta x))\right)^{-1} \]

iii. Life Table Regression

\[ E(Y) = 1 - \exp(-\alpha \exp(\beta x)) \]

The results of applying these three methods to the data are given in Tables 5.1, 5.2, and 5.3. All three models were applied using the weighted likelihood approach of section 5.4. The models were also applied under the assumption of a simple random sample with replacement. Variances for the parameters were estimated under the simple random sample assumption and by using the BRR variance estimation method of Chapter II.

Table 5.1: Parameter Estimates by Model and by Sample Design

| MODEL                        | PARAMETER |            | P(Y=1|X=0) | P(Y=1|X=3) |
|------------------------------|-----------|------------|----------|-----------|
|                              | \(\alpha\) | \(\beta\) |          |           |
| Linear Regression            |           |            |          |           |
| Simple Random Sample         | 0.5707    | -0.0456    | 0.5707   | 0.4339    |
| Complex Design               | 0.5643    | -0.0387    | 0.5643   | 0.4482    |
| Logistic Regression          |           |            |          |           |
| Simple Random Sample         | 0.3046    | -1.920     | 0.5756   | 0.4326    |
| Complex Design               | 0.2701    | -0.1603    | 0.5670   | 0.4475    |
| Life Table Regression        |           |            |          |           |
| Simple Random Sample         | 0.4807    | 0.0482     | 0.3816   | 0.4267    |
| Complex Design               | 0.4842    | 0.0440     | 0.3838   | 0.4245    |
Table 5.1 gives the estimated parameters for the three models. The findings for these data indicate that including the weights or excluding them has little effect upon the estimates obtained. A similar result was found in Chapter II for the conditional probabilities and the survivorship probabilities. All three models give essentially the same estimate of the probability of separation.

In Table 5.2 the variance and covariance estimates for the parameters under the two sampling and estimation processes are given.

Table 5.2: Variance Estimates by Model and by Sample Design

<table>
<thead>
<tr>
<th>MODEL</th>
<th>ESTIMATE</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Var(α)</td>
<td>Var(β)</td>
<td>Cov(α,β)</td>
</tr>
<tr>
<td>Linear Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Random Sample</td>
<td>6.44 × 10^{-4}</td>
<td>4.84 × 10^{-5}</td>
<td>-1.61 × 10^{-4}</td>
</tr>
<tr>
<td>Complex Design</td>
<td>1.58 × 10^{-3}</td>
<td>9.37 × 10^{-5}</td>
<td>-1.05 × 10^{-4}</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Random Sample</td>
<td>1.13 × 10^{-2}</td>
<td>8.86 × 10^{-4}</td>
<td>-2.89 × 10^{-3}</td>
</tr>
<tr>
<td>Complex Design</td>
<td>3.25 × 10^{-2}</td>
<td>5.01 × 10^{-3}</td>
<td>-1.26 × 10^{-2}</td>
</tr>
<tr>
<td>Life Table Regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple Random Sample</td>
<td>2.15 × 10^{-3}</td>
<td>6.07 × 10^{-4}</td>
<td>-1.07 × 10^{-3}</td>
</tr>
<tr>
<td>Complex Design</td>
<td>6.23 × 10^{-3}</td>
<td>1.12 × 10^{-3}</td>
<td>-2.61 × 10^{-3}</td>
</tr>
</tbody>
</table>

Table 5.3 provides the design effects. The design effects for the two parameters for the three models are all greater than one. The square root of the ratio of the determinants of the variance-covariance matrices (see equation 5.7.6) also are greater than unity for the three procedures.
Table 5.3: Design Effects for the Logistic and Life Table Regression Models

<table>
<thead>
<tr>
<th>MODEL</th>
<th>$\alpha$ (a)</th>
<th>$\beta$ (a)</th>
<th>$G$ (b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>2.46</td>
<td>1.94</td>
<td>2.72</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>2.88</td>
<td>5.65</td>
<td>1.55</td>
</tr>
<tr>
<td>Life Table Regression</td>
<td>2.90</td>
<td>1.86</td>
<td>1.09</td>
</tr>
</tbody>
</table>

\(^a\)Computed using equation (5.6.10).
\(^b\)Computed using equation (5.7.5).

The design effects found in the above examples are in agreement with those reported for these data in Chapter II. The results do indicate that the approach given in this chapter may provide reasonable parameter and variance estimates. Although the results do indicate the methods perform well, more examples and more data sets need to be studied. Simulations might also be used for investigating the method.
CHAPTER VI

SUMMARY AND SUGGESTIONS FOR FUTURE RESEARCH

6.1. Summary

The research presented in the previous chapters was motivated by a need for a life table methodology for data from a complex probability sample.

The first chapter introduced the reader to the field of life table and survivorship analysis by reviewing the various types of life tables, the notation most commonly used, computing formulae, and the problems associated with obtaining estimates of the conditional probabilities. The probability distributions of several life table functions were also given.

In Chapter II problems encountered when working with data from complex probability samples were discussed, and in particular the difficulty of obtaining variance estimates for non-linear functions of the sample data was discussed. An examination of the conditional probability of an event, and the survivorship probabilities showed these functions to be non-linear. Methods for computing variance estimates for complex survey data were reviewed, and their application to the above mentioned life table functions discussed. Examples of using the variance estimation techniques of: balanced repeated replications (BRR), the jackknife (JRR or JK), and the Taylor linearization (TL) were given for the conditional probabilities and for the survivorship probabilities. The applications were studied in an attempt to see
if one of these methods was superior for providing variance estimates for life table functions estimated from complex survey data. The results from applying the BRR, JRR, JK, and TL indicated that the TL might be the 'best' overall method. The choice of TL was made based upon necessary assumptions, ease of obtaining the estimate and behavior of the estimate. It was noted that all four methods suffered a loss of stability when the sample size with respect to strata and primary sampling units was small. These results indicate that caution should be used when such methods are used at the domain level.

Chapter III began with a review of statistical tests for comparing the survivorship between two or more life tables. Particular attention was given to the tests developed by Cochran (1954) and Mantel and Haenszel (1959). The problems encountered with making comparisons between life tables constructed from complex survey data were presented, and several methods for performing tests of independence and association for categorical data were reviewed. A modification of the test of Mantel and Haenszel was developed, and a conceptual framework for specifying the structure of the variance-covariance matrix for the modified test was given. The test developed was based upon the work of Wald (1943) and used the variance-covariance estimation methods presented in Chapter II. Performance of the test was examined using data from the National Survey of Family Growth, 1973, and data from the 1975 World Fertility Survey of Sri Lanka.

Techniques for life table regression analysis were reviewed in Chapter IV, and a modification based upon the hazard model of Cox (1972) was given which allowed for an explicit solution to the likelihood equations for the regression coefficients \( \beta \). The explicit solution was
only available when the covariates for the analysis were (0,1) indicators. This new technique was then compared with several other methods for two sets of data. The results of the comparison gave evidence that the approximations obtained using the new approach provided an inference similar to that obtained using the established methods. Discussion then turned to application of the life table regression methods to data from a complex probability sample, and the difficulties of such an application. The central issue of applying such methods centered on the use of maximum likelihood estimation in complex probability samples. This issue was taken up in Chapter V.

The fifth chapter considered the problems and arguments surrounding the use of likelihood based inference and estimation in complex probability samples. Using the superpopulation approach arguments were presented providing a framework in which maximum likelihood estimation could be used with data from a complex sample survey. Several examples demonstrating the form of the estimator using the proposed methodology were given. The major thrust of these examples was to illustrate that in many cases the estimator obtained was a Horvitz-Thompson form of estimator. The question of variance estimates for maximum likelihood estimators was examined and variance estimates based upon the methods of Chapter II were proposed. Applications were performed and the results discussed.

6.2. Suggestions for Future Research

The research in this dissertation was an attempt to provide a methodology for a life table analysis of complex sample survey data. Continuing research in this area is necessary and the following four problems are proposed for future study.
i. Variance estimates were studied for only two life table functions $q_t$ and $l(t)$. Other life table functions like $\delta_0$ need to be studied.

ii. The test developed in Chapter III should be formally studied. The distributional properties formally established, power against various alternatives examined, and the behavior of the test for various sample designs and patterns of design effects explored.

iii. Regression models for vital registration data would be useful. Since data from such registries are the central death rates, $h_t$, models based upon these rates might be developed. Also, the properties of the proposed estimators of the regression parameters given in Chapter IV need to be studied.

iv. Further study of maximum likelihood methods for complex probability samples is necessary. Data sets other than those used in this dissertation should be used and simulations could be employed to study the performance of the methodology for known parameter and variance parameters.
BIBLIOGRAPHY


Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. Transactions of The American Mathematical Society 54, 426-482.


### Table A.1

VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL AND BY METHOD OF ESTIMATION FOR THE CONDITIONAL PROBABILITIES OF SUBSAMPLE ONE TAKEN FROM THE 1975 WORLD FERTILITY SURVEY OF SRI LANKA

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Conditional Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SRS</td>
</tr>
<tr>
<td>[0,1)</td>
<td>0.6318</td>
<td>2.57</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.4024</td>
<td>4.43</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.5771</td>
<td>11.63</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.7183</td>
<td>17.42</td>
</tr>
</tbody>
</table>

### Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[0,1)</td>
<td>0.79</td>
</tr>
<tr>
<td>[1,2)</td>
<td>1.45</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.98</td>
</tr>
<tr>
<td>[3,4)</td>
<td>1.18</td>
</tr>
</tbody>
</table>

1Estimates need to be multiplied by $10^{-4}$. 
TABLE A.2

VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE CONDITIONAL PROBABILITIES OF SUBSAMPLE TWO TAKEN FROM THE 1975 WORLD FERTILITY SURVEY OF SRI LANKA

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Conditional Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SRS</td>
</tr>
<tr>
<td>[0,1)</td>
<td>0.6563</td>
<td>4.03</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.3504</td>
<td>6.57</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.4551</td>
<td>20.51</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.7532</td>
<td>36.96</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.7526</td>
<td>57.04</td>
</tr>
</tbody>
</table>

Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[0,1)</td>
<td>2.85</td>
</tr>
<tr>
<td>[1,2)</td>
<td>1.25</td>
</tr>
<tr>
<td>[2,3)</td>
<td>1.29</td>
</tr>
<tr>
<td>[3,4)</td>
<td>1.21</td>
</tr>
<tr>
<td>[4,5)</td>
<td>1.69</td>
</tr>
</tbody>
</table>

1Estimates need to be multiplied by $10^{-4}$. 
TABLE A.3
VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE CONDITIONAL PROBABILITIES OF SUBSAMPLE THREE TAKEN FROM THE 1975 WORLD FERTILITY SURVEY OF SRI LANKA

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Conditional Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SRS</td>
</tr>
<tr>
<td>[0,1)</td>
<td>0.6096</td>
<td>8.76</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.3851</td>
<td>14.71</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.4876</td>
<td>41.30</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.6207</td>
<td>81.18</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.8889</td>
<td>54.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[0,1)</td>
<td>0.96</td>
</tr>
<tr>
<td>[1,2)</td>
<td>1.64</td>
</tr>
<tr>
<td>[2,3)</td>
<td>1.18</td>
</tr>
<tr>
<td>[3,4)</td>
<td>1.33</td>
</tr>
<tr>
<td>[4,5)</td>
<td>1.12</td>
</tr>
</tbody>
</table>

¹Estimates need to be multiplied by $10^{-4}$. 
TABLE A.4

VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE CONDITIONAL PROBABILITIES OF SUBSAMPLE FOUR TAKEN FROM THE 1975 WORLD FERTILITY SURVEY OF SRI LANKA

Variance Estimates

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Conditional Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SRS</td>
</tr>
<tr>
<td>[0,1)</td>
<td>0.6025</td>
<td>3.07</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.4286</td>
<td>5.83</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.6552</td>
<td>15.58</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.5556</td>
<td>27.44</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.6000</td>
<td>48.00</td>
</tr>
</tbody>
</table>

Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[0,1)</td>
<td>1.06</td>
</tr>
<tr>
<td>[1,2)</td>
<td>2.56</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.63</td>
</tr>
<tr>
<td>[3,4)</td>
<td>2.21</td>
</tr>
<tr>
<td>[4,5)</td>
<td>3.00</td>
</tr>
</tbody>
</table>

1 Estimates need to be multiplied by $10^{-3}$. 
### TABLE A.5
VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE CONDITIONAL PROBABILITIES OF SUBSAMPLE ONE TAKEN FROM THE 1973 NATIONAL SURVEY OF FAMILY GROWTH

#### Variance Estimates

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Conditional Probability</th>
<th>SRS</th>
<th>BRR</th>
<th>TL</th>
<th>JRR</th>
<th>JK</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,1)</td>
<td>0.9816</td>
<td>3.22</td>
<td>2.84</td>
<td>2.69</td>
<td>2.99</td>
<td>3.26</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.9826</td>
<td>3.25</td>
<td>3.60</td>
<td>3.24</td>
<td>3.24</td>
<td>3.72</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.9836</td>
<td>3.19</td>
<td>3.32</td>
<td>2.86</td>
<td>2.73</td>
<td>5.06</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.9829</td>
<td>3.61</td>
<td>4.98</td>
<td>4.36</td>
<td>4.58</td>
<td>5.49</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.9903</td>
<td>2.29</td>
<td>2.29</td>
<td>2.24</td>
<td>2.28</td>
<td>2.25</td>
</tr>
</tbody>
</table>

#### Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[0,1)</td>
<td>0.88</td>
</tr>
<tr>
<td>[1,2)</td>
<td>1.11</td>
</tr>
<tr>
<td>[2,3)</td>
<td>1.04</td>
</tr>
<tr>
<td>[3,4)</td>
<td>1.38</td>
</tr>
<tr>
<td>[4,5)</td>
<td>1.00</td>
</tr>
</tbody>
</table>

1Estimates need to be multiplied by $10^{-6}$. 
### TABLE A.6

**VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE CONDITIONAL PROBABILITIES OF SUBSAMPLE TWO TAKEN FROM THE 1973 NATIONAL SURVEY OF FAMILY GROWTH**

#### Variance Estimates

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Conditional Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,1)</td>
<td>0.9630</td>
<td>SRS 47.80 50.30 48.80 46.70</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.9750</td>
<td>7.57 31.50 33.30 32.70 31.20</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.9783</td>
<td>7.97 10.00 11.50 9.86 11.50</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.9854</td>
<td>5.32 14.30 14.80 14.60 14.60</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.9944</td>
<td>3.67 2.00 1.47 2.11 2.17</td>
</tr>
</tbody>
</table>

#### Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR  TL  JRR  JK</td>
</tr>
<tr>
<td>[0,1)</td>
<td>5.63 5.92 5.74 5.50</td>
</tr>
<tr>
<td>[1,2)</td>
<td>4.16 4.39 4.31 4.12</td>
</tr>
<tr>
<td>[2,3)</td>
<td>1.25 1.44 1.24 1.44</td>
</tr>
<tr>
<td>[3,4)</td>
<td>2.68 2.78 2.74 2.74</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.54 0.40 0.57 0.59</td>
</tr>
</tbody>
</table>

1^Estimates need to be multiplied by 10^-6.
### Table A.7

**VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE CONDITIONAL PROBABILITIES OF SUBSAMPLE THREE TAKEN FROM THE 1973 NATIONAL SURVEY OF FAMILY GROWTH**

**Variance Estimates**

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Conditional Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SRS</td>
</tr>
<tr>
<td>[0,1)</td>
<td>0.9797</td>
<td>2.42</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.9818</td>
<td>2.30</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.9830</td>
<td>2.33</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.9832</td>
<td>2.17</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.9907</td>
<td>1.41</td>
</tr>
</tbody>
</table>

**Design Effects**

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[0,1)</td>
<td>1.30</td>
</tr>
<tr>
<td>[1,2)</td>
<td>1.38</td>
</tr>
<tr>
<td>[2,3)</td>
<td>1.21</td>
</tr>
<tr>
<td>[3,4)</td>
<td>1.99</td>
</tr>
<tr>
<td>[4,5)</td>
<td>1.41</td>
</tr>
</tbody>
</table>

*Estimates need to be multiplied by $10^{-6}$.*/
### TABLE A.8

VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE SURVIVORSHIP PROBABILITIES OF SUBSAMPLE ONE TAKEN FROM THE 1975 WORLD FERTILITY SURVEY OF SRI LANKA

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Survivorship Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SRS</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.6318</td>
<td>2.57</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.2543</td>
<td>2.17</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.1467</td>
<td>1.44</td>
</tr>
<tr>
<td>[4,ω)</td>
<td>0.1054</td>
<td>1.13</td>
</tr>
</tbody>
</table>

#### Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.79</td>
</tr>
<tr>
<td>[2,3)</td>
<td>1.23</td>
</tr>
<tr>
<td>[3,4)</td>
<td>1.18</td>
</tr>
<tr>
<td>[4,ω)</td>
<td>0.99</td>
</tr>
</tbody>
</table>

1Estimates need to be multiplied by 10^-4.
TABLE A.9

VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE SURVIVORSHIP PROBABILITIES OF SUBSAMPLE ONE TAKEN FROM THE 1975 WORLD FERTILITY SURVEY OF SRI LANKA

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Survivorship Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,2)</td>
<td>0.6563</td>
<td>SRS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5.40</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.2300</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.32</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.1046</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.77</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.0787</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.41</td>
</tr>
<tr>
<td>[5,ω)</td>
<td>0.0592</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.15</td>
</tr>
</tbody>
</table>

Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[1,2)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.01</td>
</tr>
<tr>
<td>[2,3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.20</td>
</tr>
<tr>
<td>[3,4)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.21</td>
</tr>
<tr>
<td>[4,5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.60</td>
</tr>
<tr>
<td>[5,ω)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.25</td>
</tr>
</tbody>
</table>

1Estimates need to be multiplied by 10^-4.
### TABLE A.10

VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE SURVIVORSHIP
PROBABILITIES OF SUBSAMPLE THREE TAKEN FROM THE
1975 WORLD FERTILITY SURVEY OF SRI LANKA

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Survivorship Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRS</td>
<td>BRR</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.6096</td>
<td>8.76</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.2348</td>
<td>6.76</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.1145</td>
<td>3.88</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.0710</td>
<td>2.56</td>
</tr>
<tr>
<td>[5,ω)</td>
<td>0.0632</td>
<td>2.30</td>
</tr>
</tbody>
</table>

### Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.97</td>
</tr>
<tr>
<td>[2,3)</td>
<td>1.17</td>
</tr>
<tr>
<td>[3,4)</td>
<td>1.20</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.78</td>
</tr>
<tr>
<td>[5,ω)</td>
<td>0.68</td>
</tr>
</tbody>
</table>

1 Estimates need to be multiplied by 10\(^{-4}\).
### TABLE A.11

VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE SURVIVORSHIP PROBABILITIES OF SUBSAMPLE THREE TAKEN FROM THE 1975 WORLD FERTILITY SURVEY OF SRI LANKA

#### Variance Estimates

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Survivorship Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SRS</td>
<td>BRR</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.6025</td>
<td>30.70</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.2582</td>
<td>26.81</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.1692</td>
<td>21.90</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.0940</td>
<td>14.61</td>
</tr>
<tr>
<td>[5,ω)</td>
<td>0.0564</td>
<td>9.50</td>
</tr>
</tbody>
</table>

#### Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[1,2)</td>
<td>1.06</td>
</tr>
<tr>
<td>[2,3)</td>
<td>2.54</td>
</tr>
<tr>
<td>[3,4)</td>
<td>2.84</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.89</td>
</tr>
<tr>
<td>[5,ω)</td>
<td>2.93</td>
</tr>
</tbody>
</table>

1Estimates need to be multiplied by 10⁻⁴.
### TABLE A.12

**VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE SURVIVORSHIP PROBABILITIES OF SUBSAMPLE THREE TAKEN FROM THE 1973 NATIONAL SURVEY OF FAMILY GROWTH**


<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Survivorship Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SRS</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.9816</td>
<td>3.22</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.9645</td>
<td>6.24</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.9487</td>
<td>9.00</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.9325</td>
<td>11.90</td>
</tr>
<tr>
<td>[5,ω)</td>
<td>0.9234</td>
<td>13.70</td>
</tr>
</tbody>
</table>

### Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.88</td>
</tr>
<tr>
<td>[2,3)</td>
<td>1.22</td>
</tr>
<tr>
<td>[3,4)</td>
<td>1.21</td>
</tr>
<tr>
<td>[4,5)</td>
<td>1.57</td>
</tr>
<tr>
<td>[5,ω)</td>
<td>1.88</td>
</tr>
</tbody>
</table>

1Estimates need to be multiplied by 10^{-6}. 
### TABLE A.13

VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE SURVIVORSHIP PROBABILITIES OF SUBSAMPLE THREE TAKEN FROM THE 1973 NATIONAL SURVEY OF FAMILY GROWTH

#### Variance Estimates

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Survivorship Probability</th>
<th>SRS</th>
<th>BRR</th>
<th>TL</th>
<th>JRR</th>
<th>JK</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,2)</td>
<td>0.9630</td>
<td>8.49</td>
<td>47.80</td>
<td>50.30</td>
<td>48.80</td>
<td>46.70</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.9000</td>
<td>15.20</td>
<td>78.10</td>
<td>78.70</td>
<td>77.80</td>
<td>84.50</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.9186</td>
<td>21.80</td>
<td>83.10</td>
<td>85.40</td>
<td>80.60</td>
<td>85.20</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.9051</td>
<td>25.80</td>
<td>89.60</td>
<td>95.50</td>
<td>82.90</td>
<td>90.60</td>
</tr>
<tr>
<td>[5,( \omega ))</td>
<td>0.9000</td>
<td>28.40</td>
<td>93.60</td>
<td>95.60</td>
<td>84.90</td>
<td>91.60</td>
</tr>
</tbody>
</table>

#### Design Effects

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[1,2)</td>
<td>5.63</td>
</tr>
<tr>
<td>[2,3)</td>
<td>5.14</td>
</tr>
<tr>
<td>[3,4)</td>
<td>3.81</td>
</tr>
<tr>
<td>[4,5)</td>
<td>3.47</td>
</tr>
<tr>
<td>[5,( \omega ))</td>
<td>3.30</td>
</tr>
</tbody>
</table>

\(^1\)Estimates need to be multiplied by 10^{-6}.
**TABLE A.14**

VARIANCES AND DESIGN EFFECTS BY TIME INTERVAL FOR THE SURVIVORSHIP PROBABILITIES OF SUBSAMPLE THREE TAKEN FROM THE 1973 NATIONAL SURVEY OF FAMILY GROWTH

**Variance Estimates**

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Estimated Survivorship Probability</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SRS</td>
</tr>
<tr>
<td>[1,2)</td>
<td>0.9797</td>
<td>2.42</td>
</tr>
<tr>
<td>[2,3)</td>
<td>0.9618</td>
<td>4.53</td>
</tr>
<tr>
<td>[3,4)</td>
<td>0.9455</td>
<td>6.51</td>
</tr>
<tr>
<td>[4,5)</td>
<td>0.9296</td>
<td>8.23</td>
</tr>
<tr>
<td>[5,∞)</td>
<td>0.9210</td>
<td>9.29</td>
</tr>
</tbody>
</table>

**Design Effects**

<table>
<thead>
<tr>
<th>Time Interval In Years</th>
<th>Method of Variance Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BRR</td>
</tr>
<tr>
<td>[1,2)</td>
<td>1.26</td>
</tr>
<tr>
<td>[2,3)</td>
<td>1.74</td>
</tr>
<tr>
<td>[3,4)</td>
<td>1.80</td>
</tr>
<tr>
<td>[4,5)</td>
<td>2.36</td>
</tr>
<tr>
<td>[5,∞)</td>
<td>2.72</td>
</tr>
</tbody>
</table>

1Estimates need to be multiplied by $10^{-6}$. 