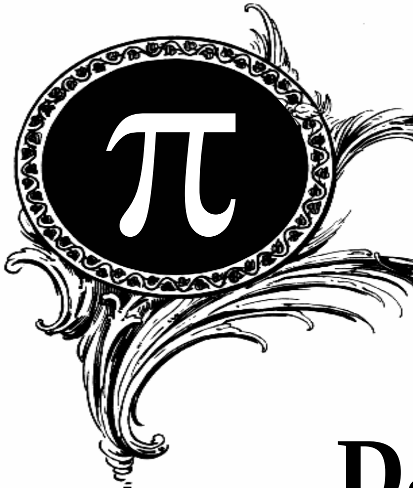


A Handbook of Categorical Data Analysis in Health Science Research

Bandit Thinkhamrop, Ph.D.

A Handbook of Categorical Data Analysis in Health Science Research

Bandit Thinkhamrop, Ph.D. (Statistics)
Associate Professor
Department of Biostatistics and Demography
Faculty of Public Health,
Khon Kaen University
THAILAND

Title	**:**	**A Handbook of Categorical Data Analysis** **in Health Science Research**
Author	**:**	**Bandit Thinkhamrop**
Publishing	**:**	**ISBN : 974-654-094-7** **January 2001** **Number of copies : 500**
Publisher	**:**	**Khon Kaen University Press** **Khon Kaen, Thailand**

PREFACE

This handbook is primarily designed for health-related Master degree students, in particular, Master of Public Health in Biostatistics at Khon Kaen University, Thailand. However, it could be used as a practical guide for health science researchers. It is also suitable as a review for PhD candidates, i.e., Doctor of Philosophy in Public Health at Khon Kaen University which started in 2001. I summarized some important concepts for each topic presented in each chapter. However this handbook is really not self-contained. Details for each topic can be found in the corresponding references given at the end of each chapter. I have tried to avoid mathematical notations as possible. Practical approaches for each type of problems were illustrated through examples. The example data are mostly adapted from many books that were related to categorical data analysis in which their authors used them to illustrate concepts of statistical methods. Most of them were difficult to follow for students who had limited mathematical and statistical background. Here I tried to provide a complete analysis as it should be done in the real world when we analyze the data. All examples were organized so that they are easy to follow and logical. Readers can also examine further approaches to the same problems by other authors that were given in each example. Advanced readers, in particular, students in Master of Public Health in Biostatistics are expect explore further in other related books the theoretical grounds of each statistical methods. I did not repeat those in this book but provided the references. All of these references are very specific - i.e., page numbers were given. By these methods, I hope readers can gain more insightful in the statistical methods presented in this book although what I had summarized in this book are also sufficient to understand the practical approaches that had

been followed. Practicing exploring further references is believed to serve as a basis of getting update to the most recent advanced knowledge in the future.

The approaches of data analysis in this book emphasize in that each problem need to be analyzed under a sufficient knowledge of the underlying research questions. Then showing that descriptive statistics are important and useful. The inferential components of statistics has to be provided both estimation (ie., confidence intervals) and test hypothesis (i.e., p-values). Conclusions need to be based mainly on the estimation than the test hypothesis. Thus each of most of the examples consisted of four components - i) describing the proportions; ii) estimating measure of effect; iii) testing the hypothesis; and iv) summary findings.

This book started with describing an overview of categorical data analysis in Chapter 1 in which some simple (univariate) analyses were discussed. Chapters 2 - 5 involve bivariate analysis in several situations. Chapters 6 - 7 related to multivariable analysis. References and exercises were provided at the end of each chapter. Readers are encouraged to try doing the exercise then compare with the detailed answers given in the appendix at the end of the book.

I tried to limit the statistical software used in this book so that readers can gain concepts underlying the analysis rather than the software commands. Stata is my choice because it covers a wide range of statistical methods presented in this book yet small and affordable. Readers can easily access more information about Stata via the internet at http://www.stata.com. Commands are in bold letters following a dot. The results are displayed in letters smaller than and different fonts from the main texts. This is to enable readers repeat the analysis.

I thank Professor Annette Dobson, The University of Newcastle, Australia for her great contribution, even not directly, in this work. I thank Dr Diane O'Cornel, who was my teachers in this subjects when I was studying at The University of Newcastle. This book was very much influenced by her. Study modules of Categorical Data Analysis prepared by Centre of Clinical Epidemiology and Biostatistics (CCEB), The University of Newcastle, Australia in 1993 were also plays a large role in organizing the chapters of this book. Special thanks are to staff of the Department of Biostatistics and Demography for their encouragement, comments, and suggestions. I also thank to authors of several books from which I have learned greatly; in particular, books by Agresti (1990), Fleiss (1981), Feinsberg (1980), Everitte (1977), Kleinbaum (1994), and Altman (1991). Finally I thank all batches of my students since 1996 - those who took a Master of Public Health in Biostatistics for their comments on the earlier drafts, especially the batch in 1999.

If there is some mistakes that could be found in this book, it is solely my false. I would appreciate any suggestions and very much welcome any comments.

Bandit Thinkhamrop
January 2001

คำปรารภ

แรงดลใจในการเขียนหนังสือเล่มนี้มาจากปัญหาของนักศึกษาในทำความเข้าใจวิธีการทางสถิติเพื่อวิเคราะห์ข้อมูลการวิจัย และการเรียนการสอนวิชา 516 707 Analysis of Categorical Data สำหรับนักศึกษาปริญญาโทหลักสูตรสาธารณสุขศาสตรมหาบัณฑิต สาขาชีวสถิติ ซึ่งผู้เขียนเป็นผู้รับผิดชอบมาตั้งแต่ปี 2536 (เป็นหลักสูตรที่มีการเรียนการสอนเป็นภาษาอังกฤษ และเป็นเหตุผลของการเขียนหนังสือเล่มนี้เป็นภาษาอังกฤษ) และหลักสูตร 516 701 Biostatistics for Medical Science and Health Research สำหรับนักศึกษาปริญญาโทหลายสาขาทางวิทยาศาสตร์สุขภาพ ที่พบว่าตำราที่เกี่ยวกับการวิเคราะห์ข้อมูลแจงนับที่มีอยู่ในปัจจุบันส่วนมากเน้นหนักทางด้านทฤษฎี เต็มไปด้วยสูตรทางคณิตศาสตร์ หรือที่พยายามทำให้ง่ายขึ้นก็มีเนื้อหาแยกเป็นส่วนๆ ยากแก่การประสานเชื่อมโยงระหว่างเนื้อหา และที่ยากยิ่งกว่าคือการเชื่อมโยงเนื้อหาทางทฤษฎีกับโลกความเป็นจริง คือไม่เพียงวิเคราะห์ข้อมูล แต่ต้องเขียนออกมาเป็นรายงานสรุปผลด้วยสภาพเหล่านี้ ทำให้นักศึกษาหรือบุคลากรทางด้านการแพทย์และสาธารณสุข ซึ่งมีพื้นฐานทางด้านคณิตศาสตร์และสถิติที่จำกัดนั้น เข้าใจเนื้อหาได้ยาก หรือหมดความพยายามที่จะศึกษา นอกจากนั้น 7 คณะในสาขาวิทยาศาสตร์สุขภาพยังได้ร่วมกันเปิดหลักสูตรนานาชาติระดับปริญญาเอก Doctor of Philosophy (Public Health) ซึ่งรับนักศึกษารุ่นแรกในปี 2544 นี้ ทั้งหมดเหล่านี้เป็นแรงผลักดันที่สำคัญให้เกิดหนังสือเล่มนี้

ในแต่ละบทของหนังสือเล่มนี้ เริ่มจากการสรุปแนวคิดที่สำคัญ โดยได้พยายามหลีกเลี่ยงสูตรทางคณิตศาสตร์ ยกเว้นที่มีความจำเป็นอย่างยิ่งต่อการทำความเข้าใจเนื้อหาซึ่งจำกัดให้เหลือน้อยที่สุด จากนั้นใช้ตัวอย่างเป็นตัวเดินเรื่อง

ตัวอย่างส่วนมากอ้างอิงมาจากตำราของนักสถิติชั้นนำของโลกที่เกี่ยวกับการวิเคราะห์ข้อมูลแจงนับพร้อมกับได้ดัดแปลงเพื่อให้ง่ายต่อการเข้าใจและสอดคล้องกับปัญหาในประเทศไทย ในขณะที่ผู้แต่งที่ได้อ้างอิงไว้นั้นใช้ในการแสดงตัวอย่างการคำนวณเป็นหลัก ในหนังสือเล่มนี้ ตัวอย่างเป็นมากกว่าโจทย์เพื่อแสดงการคำนวณ เช่นมีการชี้ให้เห็นว่าคำถามการวิจัยและประเภทของการวิจัยมีความสำคัญต่อการวิเคราะห์ข้อมูล มีการวิเคราะห์ข้อมูลโดยใช้วิธีการทางสถิติที่กล่าวถึงในบทนั้น มีการวิเคราะห์ที่มีให้ครบองค์ประกอบทั้งสถิติเชิงพรรณนาและการอนุมานทางสถิติ และมีการแปลความหมายและนำเสนอผลการวิเคราะห์ในรูปแบบที่ถูกต้องตามทฤษฎีและเป็นที่ถือปฏิบัติกันทั่วไปในวารสารวิชาการต่างๆ การวิเคราะห์ข้อมูลแสดงให้เห็นโดยการใช้คอมพิวเตอร์ ซึ่งเป็นวิธีการที่ทำในชีวิตจริง การคำนวณจึงดูไม่ยุ่งยาก ผู้อ่านสามารถค้นคว้าเพิ่มเติมถึงรากเหง้าการวิเคราะห์ รวมถึงสูตรที่ใช้ในการคำนวณ โดยค้นคว้ารายการเอกสารอ้างอิงที่ให้ไว้อย่างจำเพาะถึงเลขหน้าในหนังสือที่อ้างอิงนั้น

ความสำคัญอื่นๆ มีกล่าวแล้วใน Preface ที่เสนอไว้ก่อนหน้านี้ อนึ่ง การเขียนหนังสือเล่มนี้เป็นภาษาอังกฤษ พึงเป็นส่วนสำคัญในการวางรากฐานการเรียนรู้ที่ดีของผู้อ่าน เนื่องจากวิชาการทางด้านนี้ มีการพัฒนารุดหน้าอย่างไม่หยุดยั้งและรวดเร็วตามเทคโนโลยี และล้วนเป็นภาษาอังกฤษ

ผู้เขียนขอน้อมรับคำแนะนำปรับปรุงแก้ไขหนังสือเล่มนี้ด้วยความยินดิยิ่ง เพื่อยังประโยชน์แก่สังคมแห่งการเรียนรู้ และการพัฒนาองค์ความรู้ด้านวิทยาศาสตร์สุขภาพต่อไป

บัณฑิต ถิ่นคำรพ

มกราคม 2544

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1

An Overview

Chapter Objectives

After completing this chapter, readers should be able to:

- **describe goals of data analysis;**
- **specify components of statistics needed for reporting of health science research;**
- **describe type of variables and research with categorical outcome;**
- **describe general concepts of categorical data analysis;**
- **specify appropriate statistical methods for analysis of research with categorical outcome in relation to type of dependent and independent variables;**
- **calculate the point estimate of proportions for a dichotomous outcome and their confidence intervals;**
- **test a hypothesis for single proportion; and**
- **estimate proportions for a polytomous or ordinal outcome.**

1.1 Ultimate goal of data analysis

A key concept underlying the subject of statistics is “variability”. Statistical methods can help us to explain variation observed in the data being collected. By explaining such variation we interpret the results. Reliable results depend upon an appropriate research design. If the design of the study is unacceptable, the research is rather useless no matter how well the data were analyzed.

Statistics is a curious amalgam of mathematics, logic and judgement (Altman, 1991). The logical process and judgement are more difficult than mathematics. These involve careful thought about the topic under investigation, the principles of research methodology, the concepts underlying statistical methods used, and interpretation of the results. Thus, in data analysis, we cannot just looking solely at the data - dumping into the computer and take the outputs.

The ultimate goal of most of health research were to obtain body of knowledge regarding the study health events. Statistics thinking can contribute to every stage of the study. The body of knowledge in the sense of statistics is the ultimate outcome that answers the research question(s). Since we mainly aim to obtain body of knowledge that can be generallizeable, statistics should include both descriptive and inferential components of statistics. The descriptive component is to describe the study sample whereas the inferential component involves using information obtained from a sample to describe a larger population.

As mentioned that the body of knowledge is universal in nature, the inferential component of statistics should be presented. There are two sub-components within this

component - estimation and hypothesis testing. The estimation is presented as the confidence intervals whereas the hypothesis testing is presented as p-value. However, a large number of researches were misguided to use solely p-value for drawing conclusion from. Overemphasizing use of p-value (or the most popular term is the significant test) is rather misleading. Recent approach advocates use of confidence intervals followed by the p-value. A good readings for interpretation of confidence intervals is given by Guyatt (1995).

1.2 Research with categorical outcome

In planning for the research (i.e., preparing the research proposal), study variables should be clearly defined. We can know from that at least what is the outcome (or response or dependent variable) and what is (are) the independent (or study factors or explanatory variables). In most cases, there is only one outcome and several explanatory variables in a study. These variables need to be classified in to at least two main types - categorical or continuous. Knowing the types of variable will lead to appropriately choosing statistical methods for further analysis.

This book focused on a categorical outcome. Categorical data could be one of the following types of data.

1.2.1 Nominal data: There could be only two possible values of such variable such as DEAD (dead or alive), CURED (cured or not cured), TEST (positive, negative), PAIN (yes or no), etc. This type of data is called dichotomous. If there are three or more possible values, it is called polytomous such as DELIVERY (vaginal, caesarian section, or others). Note that capital letters are to indicate variables' name.

1.2.2 Ordinal data: It is the polytomous data that can be ranked such as SYMPTOM (severe, moderate, mild).

1.2.3 Count data: It is a discrete quantity such as INJURY recorded as number of episode of injuries per period, EPILEPSY recorded as number of epileptic attack per two weeks, etc.

Note that continuous outcome could be grouped then this can be analyzed as categorical data. However this practice is not recommended as it thrown away some information and thus considered less efficient than being analyzed as its original continuous data. On the contrary, ordinal and count outcome could also be analyzed as if they are continuous. However, this approach is acceptable in some certain circumstances. It is also often that some higher level of outcome are collapsed so that it can be less level such as birth weight in grams are grouped into 3 grouped - low, normal, and high, and it can then be collapsed into two groups - normal and abnormal. This approach also needs a careful though (Stromberg, 1996).

1.3 An overview of categorical data analysis

Once data had been collected, we need to summarize it before further analysis. This serve as the tools for both determine the distribution of data and also to describe the characteristics of the study sample and estimate statistics of interest.

The analysis of categorical data generally involves the proportion of "successes" in a given population. This may consist of estimating a single parameter, comparing two parameters, or investigating the potential relationship between two or more categorical variables.

Aside from type of the data, research design is also an important criterion for determining appropriate statistical methods. Some approaches for the data analysis were summarized in Table 1.1. These approaches are limited only to common type of research where there was only one outcome and several explanatory variables. If only an outcome was analyzed and all explanatory variables were just for

describing the study sample, it is termed a *univariate* analysis (Chapter 1 sections 1.4, 1.5, and 1.6). If the outcome was analyzed with only one explanatory variable at a time, it is *bivariate* analysis (Chapters 2, 3, 4, and 5). If the outcome was analyzed with several explanatory variables at the same time, it is a *multivariable* analysis (Chapters 6, and 7). We will not cover *multivariate* analysis where more than one outcome was analyzed at a time (Kleinbaum et al., 1998; page 1 has a discussion regarding multivariable and multivariate analysis). Chapter 8 presents special issues related to analysis of categorical data that were not mentioned in the remaining chapters.

Table 1.1 summary of approaches comomly used for analysis of a categorical outcome.

Independent (exploratory) variable(s)	A dependent variable (An outcome)		
	Two categories(dichotomous)	Three categories or more (polytomous)	Three categories or more (ordinal)
1. None	Estimating proportion (Chapter 1) next section	Estimating proportions (Chapter 1) next section	Estimating proportions (Chapter 1) next section
2. One variable 2.1 Two categories(dichotomous)	2-by-2 Table (Chapter 2)	2-by-C Table (Chapter 3)	2-by-C Table (Chapter 3)
2.2 Three categories or more(polytomous)	2-by-C Table (Chapter 3)	R-by-C Table (Chapter 4)	R-by-C Table (Chapter 4)
2.3 Three categories or more(ordinal)	2-by-C Table (Chapter 3)	R-by-C Table (Chapter 4)	R-by-C Table (Chapter 4)
2.4 Continuous data	Logistic regression (Chapter 6)	Multinomial logistic regression (Chapter 6)	Ordered logistic regression (Chapter 6)
3. More than one variables 3.1 All are categorical	Logistic regression (Chapter 6), or Log-linear model (Chapter 7)	Multinomial logistic regression (Chapter 6), or Log-linear model (Chapter 7)	Ordered logistic regression (Chapter 6), or Log-linear model (Chapter 7)

3.2 All are continuous	**Logistic regression (Chapter 6)**	**Multinomial logistic regression (Chapter 6)**	**Ordered logistic regression (Chapter 6)**
3.3 Mixed (categorical and continuous)	**Logistic regression (Chapter 6)**	**Multinomial logistic regression (Chapter 6)**	**Ordered logistic regression (Chapter 6)**
Repeated measurement of a categorical outcome	**Matched 2-by-2 Table (Chapter 2) or GEE (Chapter 6)**	**Squared Table (Chapter 5) or GEE (Chapter 6)**	**Squared Table (Chapter 5) or GEE (Chapter 6)**
Outcome as a count data	**Poisson regression model (Chapter 6)**		

1.4 Estimating proportions for a dichotomous outcome

In many health researches, we randomly selected a sample of n subjects to determine a number of x subjects who represent one of two outcomes so that a statistic "proportion", denoted by p, can be estimated as p = x / n to summarize the data. For example, a total of 400 children were randomly selected from a community to determine measles vaccine coverage, 320 of them were reported vaccinated. Thus the proportion is 0.8 or 80% which is the vaccine coverage. In this case x follows a binomial distribution. A suggested reading for this distribution is in Altman (1991); page 63 - 66 and 68 - 70. The same author also provided a readable detail, formula and a work example, on obtaining confidence intervals for one proportion on page 230.

Here we consider "vaccination" the dichotomous outcome since it has two possible categories - vaccinated or non-vaccinated. All other variables could be also collected but just for describing the study samples - not for comparing such outcome by groups of these variables. In other words, there is no explanatory variable of interest. Richardson (1994) termed this a 2-by-1 Table as opposed to 2-by-2 Table where there is a dichotomous explanatory variable.

The above example can be calculated using an immediate "ci" command of STATA (see StataCorp., 1999; Volume 1: A-G

page 194-200) requesting for an estimated proportion and binomial exact confidence intervals as shown below.

```
. cii 400 320

                                                         -- Binomial Exact --
Variable |      Obs          Mean    Std. Err.       [95% Conf. Interval]
---------+-------------------------------------------------------------
         |      400            .8         .02         .7573914    .8381042
```

Now let's use a data set. The following data set will be used throughout the book to avoid confusion that may caused by several data sets. We will refer to this data set "The Example Data Set". It was available in the internet which can be downloaded directly at the following address:

http:/bandit.mykku.net

The six variables (Table 1.2) denoted by V1, V2, …, and V6 were modified to suite the topics being discussed. Note that "id" stands for the identification number of individual record.

Table 1.2 Summary of the example data set

id	V1	V2	V3	V4	V5	V6
1.	1	1	0	2600	30	0
2.	1	1	0	2900	29	1
3.	1	1	0	3100	25	0
4.	1	1	0	3000	21	0
5.	1	1	0	2600	19	0

--- 457 records were skipped ---

463.	0	0	0	2600	30	0
464.	0	1	0	3500	30	0
465.	0	1	0	3200	22	1

Example 1.1

A hypothetical scenario of the following data set is that it is from a cross-sectional study was conducted among 465 women who have had delivered their children 1 to 6 months before the study was started (Table 1.2). It aimed to determine prevalence of neonatal death.

Table 1.3 Summary of the variables for the example data set in Example 1.1.

Variable names	Descriptions	Values
V1	Dead within the first month of life	1 = Dead 0 = Alive
V2	Gender	1 = Male 0 = Female
V3	Mother attending antenatal care during pregnancy	1 = Yes 0 = No
V4	Birth weight	Weight in grams
V5	Mother's age	Age in years
V6	Place of birth	0 = Hospital 1 = Health center 2 = Home 3 = Roadside (During travelling)

Preview: V1 is an outcome, the remaining variables are to describe characteristics of the children. We will focus here only on analysis of the main outcome. An example of complete analysis was demonstrated at the end of Chapter 10.

Steps for the data analysis with Stata :

1. Open the example data set in Stata using the “use” command.

```
. use example.dta, clear
```

2. Examine the data using “summarize” command (see StataCorp., 1999; Volume 4: Su-Z page 1-7).

```
. su

Variable |     Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
      id |     465         233   134.3782          1        465
      v1 |     465    .1397849   .3471372          0          1
      v2 |     465    .5182796   .5002039          0          1
      v3 |     465    .0752688   .2641087          0          1
      v4 |     465    3010.695   437.7349       1850       4000
      v5 |     465    25.52473   5.362298         17         42
      v6 |     465     .255914   .5882217          0          3
```

3. Obtain the frequency, the estimated proportion, and the confidence intervals of neonatal dead using the following two commands, i.e. "tab" (see StataCorp., 1999; Volume 4: Su-Z page 144-152) and "ci" (see StataCorp., 1999; Volume 1: A-G page 194-200).

```
. tab v1

         V1 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        400       86.02       86.02
          1 |         65       13.98      100.00
------------+-----------------------------------
      Total |        465      100.00


. ci v1

Variable |     Obs         Mean    Std. Err.       [95% Conf. Interval]
---------+-------------------------------------------------------------
      v1 |     465     .1397849    .0160981        .1081507    .1714192
```

4. Summarize findings:

Among a total of 465 children, 65 died within the first month of life. The prevalence of neonatal dead was 14.0% (95%CI: 10.8% to 17.1%).

1.5 Test hypothesis for a proportion

So far we have done both the descriptive (i.e., the estimated prevalence of 14.0%) and inferential components (i.e., the 95%CI) of statistics. For the inferential component, there could be a hypothesis testing if the study also aim to compare the prevalence in the study area to that of another area or other standard value. Altman (1991); page 230-231, provide a good summary on the formula and the working example. For example, the Ministry of Public Health set the goal to reduce the prevalence to be 5.0%. The investigators aim to test if their finding different from 0.5%. Of course, the observed prevalence of 14.0% is clearly different from the null value of 5.0%. But whether this difference is due to chance or not is the question that needs a test hypothesis. The p-value obtained from the test is the probability of having observed the prevalence of 0.14 or more when the true prevalence is 0.05. A good practical guide for interpretation of p-value is given by Altman (1991); page 167. The following "prtest" Stata command (see StataCorp., 1999; Volume 3: P-St page 85-88) provides the calculation.

```
. prtest v1 = 0.05

One-sample test of proportion                    v1: Number of obs =      465

Variable |       Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+------------------------------------------------------------------
      v1 |  .1397849    .0160808   8.69267   0.0000       .1082672    .1713027

                         Ho: proportion(v1) = .05

     Ha: v1 < .05            Ha: v1 ~= .05              Ha: v1 > .05
          z =  8.883              z =  8.883                  z =  8.883
      P < z = 1.0000         P > |z| = 0.0000            P > z = 0.0000
```

Alternatively, we can use the immediate form of the “prtest” command as follows:

```
. prtesti 465 0.14 0.05

One-sample test of proportion                      x: Number of obs =      465

Variable |       Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+------------------------------------------------------------------
       x |        .14    .0160911   8.70044   0.0000       .1084619    .1715381

                          Ho: proportion(x) = .05

       Ha: x < .05              Ha: x ~= .05              Ha: x > .05
        z =  8.905                z =  8.905               z =  8.905
      P < z = 1.0000          P > |z| = 0.0000         P > z = 0.0000
```

For small sample, the exact binomial probability test should be used. Richardson (1994); page 129, suggested that it should be used routinely in the analysis of 2-by-1 Tables that are derived from fewer than 100 subjects. The "bitest" command of Stata (see StataCorp., 1999; Volume 1: A-G page 138-141) calculates the exact p-value for this test.

```
. bitest v1 = 0.05

Variable |        N   Observed k   Expected k   Assumed p   Observed p
---------+------------------------------------------------------------
      v1 |      465          65        23.25       0.05000      0.13978

  Pr(k >= 65) = 0.000000  (one-sided test)
  Pr(k <= 65) = 1.000000  (one-sided test)
  Pr(k >= 65) = 0.000000  (two-sided test)

  Note: Lower tail of two-sided p-value is empty.
```

Alternatively, we can use the immediate form of the "bitest" command as follows:

```
. bitesti 465 65 0.05

        N   Observed k   Expected k   Assumed p   Observed p
      465          65        23.25       0.05000      0.13978

  Pr(k >= 65) = 0.000000  (one-sided test)
  Pr(k <= 65) = 1.000000  (one-sided test)
  Pr(k >= 65) = 0.000000  (two-sided test)

  Note: Lower tail of two-sided p-value is empty.
```

For large study such as this example, the results from using asymptotic methods (i.e., z-test provided by the "prtest"

command) and exact test (i.e., Binomial exact probability test provided by the “bitest” command) are identical.

Since p-value < 0.001, the null hypothesis is rejected and concluded that the prevalence of 14.0% is statistically significant different from 5.0% to which the Ministry of Public Health aimed to reduce. (Note that we will never quote p-value = 0.000000 for our report since this means that it is impossible which is not true, at least one study could have happened - the study being analyzed here!)

1.6 Estimating proportions for a polytomous or ordinal outcome

Suppose we now have another cross-sectional study where V6 is an outcome. The variable has 4 levels (Table 1.2). Think of these four outcomes as delivery at “hospital”, “health center”, “home”, and “road side while travelling”. Even though the outcome are coded 0, 1, 2, 3, and 4 the numerical values are arbitrary. There was no natural ordering by place of delivery. The binomial distribution cannot be assumed but this data has a multinomial distribution. Definition of this distribution is given by Agresti (1990); page 38 - 39. For additional details of calculation of confidence intervals, see Goodman (1965).

For this example, first we can estimate the proportions using “svyprop” command (see StataCorp., 1999; Volume 4: Su-Z page 18-30) then using “display” command to calculate the 95% confidence intervals. The formula for such calculation is “Estimated proportion ± 1.96(Standard Error)”.

```
. svyprop v6

pweight:  <none>                                Number of obs      =       465
Strata:   <one>                                 Number of strata   =         1
PSU:      <observations>                        Number of PSUs     =       465
                                                Population size    =       465
Survey proportions estimation
```

```
      v6      _Obs   _EstProp    _StdErr
       0       375   0.806452   0.018341
       1        68   0.146237   0.016404
       2        15   0.032258   0.008202
       3         7   0.015054   0.005653

. disp 0.806452 - 1.96 *  0.018341 , 0.806452 + 1.96 *  0.018341
.77050364 .84240036

. disp 0.146237 - 1.96 *  0.016404 , 0.146237 + 1.96 *  0.016404
.11408516 .17838884

. disp 0.032258 - 1.96 *  0.008202 , 0.032258 + 1.96 *  0.008202
. 01618208 .04833392

. disp 0.015054 - 1.96 *  0.005653 , 0.015054 + 1.96 *  0.005653
.00397412 .02613388
```

The findings can be summarized as follows:

> **The cross-sectional study involved 465 subjects. The proportions of those who delivered at the hospital was 80.6% (95%CI: 77.0% to 84.2%), at health center was 14.6% (95%CI: 11.4% to 17.8%), at home was 3.2% (95%CI: 1.6% to 4.8%), and at the roadside while travelling was 1.5% (95%CI: 0.4% to 2.6%).**

The test hypothesis for this type of outcome in one group is uncommon. However, recent approaches emphasize estimation as had been presented. Polytomous and ordinal outcomes will be dealt with in more details in Chapter 3 - 6.

Chapter references

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

Altman, D.G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

Goodman, L.A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics.* 7:247-254.

Guyatt, G., Jaeschke, R., Heddle, N., Cook, D., Shannon, H., and Walter S. (1995). Interpreting study results: confidence intervals. *Canadian Medical Association Journal.* 152:169-173.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E, and Nizam, A. (1998). *Applied regression analysis and other multivariable methods*. Pacific Grove: Duxbury Press.

Richardson, J.T.E. (1994). The analysis of 2×1 and 2×2 contingency tables: an historical review. *Statistical Methods in Medical Research*. 3:107-133.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

Stromberg, U. (1996). Collapsing ordered outcome categories: a note of concern. Am. J. Epidemiology 144(4): 421-424.

Chapter 2

Analysis of 2-by-2 Tables

Chapter Objectives

After completing this chapter, readers should be able to:

- **state the null hypotheses and perform appropriate tests of these hypotheses for 2-by-2 tables formed by the cross-classification of two dichotomous variables from cross-sectional studies, prospective (cohort or experimental) studies, and retrospective studies;**
- **describe appropriate proportions and calculate measures of association for 2-by-2 tables and corresponding 95% confidence intervals;**
- **Analyze data collected from a matched pairs study;**
- **define and calculate sensitivity, specificity, negative and positive predictive values, and likelihood ratios for assessing performance of a diagnostic test;**
- **perform stratified analysis and interpret the results; and**
- **define the concepts and be able to detect confounding and interaction.**

Contents

2.1 Introduction

The 2-by-2 or four-fold Table is formed by the cross-classification of two dichotomous variables. Practically, one variable is an outcome and another variable is an independent variable. In this sense, we are dealing with two proportions - proportion of an event (eg. disease) for each of the two groups of an explanatory variable (eg. study factor).

Generally, this analysis serves as a good explanatory tool for the more complicated one that were discussed in Chapter 6 onward. However in some experimental study such as clinical trials, this approach can be the ultimate analysis from which the conclusion was drawn. For example, the efficacy of a treatment in curing a disease was assessed and effects of all other variables such as characteristics of patients and disease severity were controlled for by randomization technique.

This chapter presents systematic approaches for analyzing a dichotomous outcome with a dichotomous explanatory variable for various types of study designs. The notation bellow (Table 2.1) will be used throughout.

Table 2.1 Notation of a 2-by-2 Table displaying cell frequencies

		Variable 1 (Outcome)		
		1	**2**	**Total**
Variable 2 ***(Independent variable)***	**1**	n_{11}	n_{12}	n_{1+}
	2	n_{21}	n_{22}	n_{2+}
Total		n_{+1}	n_{+2}	n_{++}

In general, Variable 1 is an outcome while Variable 2 is an independent variable. However, the format of the table can be exchangeable, especially the computer output. Being able to classify which variable is the outcome is of great benefit in helping us locates appropriate cell frequencies and other statistics.

The frequencies in the above table can be generated using three different study designs. That is, cross-sectional study, cohort study, and case-control study. Details for each study can be found in several books, a readable one is Altman (1991); page 91-103.

Section 2.2 described the cross-sectional study where the grand total (i.e., n_{++}) is fixed. Section 2.3 described the cohort study where the row total (i.e., n_{1+} and n_{2+}) is fixed. The clinical trial mentioned above can be classified as a cohort study as they are both prospective studies. Section 2.4 described the case-control study where the column total (i.e., n_{+1} and n_{+2}) is fixed. If either the outcome or the independence

variable was matched by another variable, it became a matched 2-by-2 Table presented in Section 2.5. The diagnostic test is a special type of analyzing a 2-by-2 Table presented in section 2.6. The statistical analysis appropriate to each of these and the corresponding interpretation will be described.

Since the analysis of categorical variable involves proportion, we denote small letter "p" as the sample proportion and the Greek letter "π" as the population proportion. Below is the table displaying the proportions for each cell - Table 2.2 is for the sample proportion and Table 2.3 is for the population.

Table 2.2 Notation of a 2-by-2 Table displaying the population proportions from which the sample was drawn.

		Variable 1 (Outcome)		
		1	**2**	**Total**
Variable 2 **(*Independent Variable)***	**1**	p_{11}	p_{12}	p_{1+}
	2	p_{21}	p_{22}	p_{2+}
Total		p_{+1}	p_{+2}	p_{++}

Table 2.3 Notation of a 2-by-2 Table displaying the population proportions from which the sample was drawn.

		Variable 1 (Outcome)		
		1	2	Total
Variable 2 (*Independent* *Variable)*	1	π_{11}	π_{12}	π_{1+}
	2	π_{21}	π_{22}	π_{2+}
Total		π_{+1}	π_{+2}	π_{++}

2.2 Cross-sectional study

Select a total of n_{++} subjects from a large population and then classify each subject on two dichotomous variables. Only the total sample size n_{++} can be specified in advance and it is said to be fixed, i.e., the grand total is fixed. The four cell frequencies n_{11}, n_{12}, n_{21}, and n_{22} are random variables.

Example 2.1

The following is a hypothetical data (CCEB, 1993) to determine if there is an association between gender and smoking. A sample of 100 people were interviewed for their smoking status. They were then crossed-classified according to gender and smoking status as follows:

Table 2.4 Number of smoking status by gender - data for example 2.1

		Smoker		
		Yes	No	
Gender	Male	18	37	55
	Female	15	30	45
		33	67	100

Ex 2.1-1 Describing the proportions

In reference to the notation in Table 2.2, the appropriate proportion for the cross-sectional survey, where the grand total is fixed, is $p_{ij} = n_{ij} / n_{++}$ where i = 1, 2 and j = 1, 2. For example, the proportion of male who smoked can be calculated by 18/100. However, these proportions are difficult to interpret. For the purpose of describing the proportions, therefore, we need to assume the groups under the independent variable known in advance, i.e., row total fixed. Therefore, the proportion that will be used for describing this data can be calculated as follows:

The proportion of male who smoked : p_1 = 18/55 = 0.327
The proportion of female who smoked : p_2 = 15/45 = 0.333

Note that the proportions of non-smoker for male and female were not presented since they are completely determined by that of the smoker.

Ex 2.1-2 Estimating measure of effect

In a cross-sectional study, the appropriate measure of effect is the odds ratio (OR) although the risk ratio or relative risk (RR) can be used in some certain conditions. These are indices of comparison between two proportions relatively. The absolute difference between the two proportions (risk difference or RD) is rarely used. See Agresti (1990); page 13-16 for more details about the three measures. For summary of formulae and working examples of calculating RR and RR, see Altman (1991); page 266-270. We use a single "csi" command of Stata (see StataCorp., 1999; Volume 1: A-G page 382-384) to do all these as follows. In ovals are RR and OR and their confidence intervals.

```
. csi 18 15 37 30, or

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        18          15  |         33
        Noncases |        37          30  |         67
-----------------+------------------------+----------
           Total |        55          45  |        100
                 |                        |
            Risk |  .3272727    .3333333  |        .33
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |        -.0060606       | -.1913915    .1792703
      Risk ratio |         .9818182       |  .5604726    1.719918
 Prev. frac. ex. |         .0181818       | -.7199179    .4395274
 Prev. frac. pop |              .01       |
      Odds ratio |          .972973       |  .4244507    2.228504  (Cornfield)
                 +-----------------------------------------------
                            chi2(1) =     0.00  Pr>chi2 = 0.9489
```

Ex 2.1-3 Testing the hypothesis

By the definition of independence, characteristics 1 and 2 and independent if each joint proportion π_{11}, π_{12}, π_{21} , π_{22} is the product of the two corresponding total or marginal proportions, ie,

$$H_0 : \pi_{ij} = \pi_{i+} \; \pi_{+j} \qquad i = 1,2; \; j = 1, 2.$$

This is the hypothesis of independence. This form is a specific hypothesis.

For a general hypothesis, we can state that

H_0 : There is no association between gender and smoking

We need to determine how close the π_{ij} are to the expected values $\pi_{i+}\pi_{+j}$.

Since only the total sample size n_{++} is fixed, the are observations from a multinomial distribution with sample size n_{++} and cell probabilities $\{\pi_{ij}\}$. Details can be found in Agresti (1990); page 39-39. Altman (1991) provided a summary of formula and an example on page 250-252.

We test the hypothesis of independence using the Pearson chi-square statistic

$$\chi^2 = \sum_i \sum_j \left[n_{ij} - \frac{n_{i+}n_{+j}}{n++} \right]^2 / \frac{n_{i+}n_{+j}}{n_{++}}$$

$$\chi^2 = \frac{n_{++}(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}}$$

This value is compared with a table for the chi-square distribution with 1 *df*.

Note that this formula for χ^2 for a 2-by-2 table does not require us to calculate expected values for the individual cells. Thus always calculate the smallest expected value for the table (using the smallest row total and smallest column total). If it

is greater than 5 then go ahead and calculate χ^2 . We use Stata do the calculation.

First we calculated for a smallest expected frequency as follows:

```
. display (33 * 45) / 100
14.85
```

It is 14.85 which is larger than 5, then Pearson’s chi-square is appropriate. Then we use the immediate form of “tabulate” command of Stata (see StataCorp., 1999; Volume 4: Su-Z, page 157-174) with the option “chi2” to obtain such test statistics as follows:

```
. tabi 18 37 \ 15 30, chi2

           |          col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |        18         37 |        55
         2 |        15         30 |        45
-----------+----------------------+----------
     Total |        33         67 |       100

          Pearson chi2(1) =   0.0041   Pr = 0.949
```

Chi-square of 0.0041 with 1 degree of freedom gives p-value = 0.949.

Ex 2.1-4 Summary findings

This cross-sectional study involved 100 people. Among a total of 55 males, 32.7% were smoked whereas among 45 females, 33.3 % were smoked. The two proportions were more or less the same (OR = 1.0, 95%CI: 0.4 to 2.2) and was not statistically significant (p-value = 0.949).

Note that we can obtain all statistics needed for the above summaries with a single command that used in Ex1-2. From the output, we can get the cell frequencies, appropriate

proportions, odds ratio and its confidence intervals, and p-values of chi-square test.

2.3 Prospective (cohort or experimental) study

Select n_{1+} subjects who are classified as level 1 of variable 2 (independent variable) and n_{2+} subjects who are classified as level 2. Then classify them according to variable 1 (outcome). Here, one set of margins (n_{1+} and n_{2+}) is fixed in advance (i.e., row totals are fixed) - the cell frequencies n_{11} and n_{21} are random variables.

Example 2.2

Data were taken from CCEB (1993)

Observational study: Identify 80 people with hypertension (exposed) and 70 normotension (unexposed), then classify them according to whether or not they died after 10 years (Exposure → Outcome).

Experimental study: A randomized controlled trial with 2 treatments and a dichotomous outcome (dead or alive).

Table 2.5 Number of outcome by treatment - data for example 2.2

		Outcome		
		Dead	Alive	
Treatment	1	32	48	80
or				
Exposure	2	14	56	70
		46	104	150

Ex 2.2-1 Describing the proportions

The proportion of interest are :

$p_1 = \frac{32}{80} = 40\%$ of people in treatment group 1 (exposed) died,

and

$p_2 = \frac{14}{70} = 20\%$ of people in treatment group 2 (unexposed) died.

(See also the Stata output, in the square, in the next section - Ex 2.2-2)

Ex 2.2-2 Estimating measure of effect

Both the RR and RD are appropriate for a prospective study. However the RR does not take baseline risk into account and can therefore be misleading for an experimental study such as a clinical trial (Jaeschke et al. 1995). Thus RD is most appropriate for this type of study. The RR is appropriate for an etiological study which were mostly designed as an observation study. Formula and work example can be found in Altman (1991); page 233 for RD and page 266-268 for RR. A single command that has been used in section Ex 2.1-2 provides all these (in the oval) as shown below:

```
. csi 32 14 48 56

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        32          14  |         46
        Noncases |        48          56  |        104
-----------------+------------------------+----------
           Total |        80          70  |        150
                 |                        |
            Risk |        .4          .2  |   .3066667
                 |                        |
                 |      Point estimate    |   [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |                .2      |   .0575049    .3424951
      Risk ratio |                 2      |   1.165562    3.431821
 Attr. frac. ex. |                .5      |   .1420448    .7086095
 Attr. frac. pop |          .3478261      |
                 +-------------------------------------------------
                               chi2(1) =     7.02  Pr>chi2 = 0.0080
```

RD = 0.2 (95%CI: 0.06 to 0.34) Death rate among the treatment group 1 are 20% higher than the treatment group 2. We are 95% sure that the risk difference would be between 6.0% to 34.0% (rounded from 5.7% to 34.2%).

RR = 2.0 (95%CI: 1.2 to 3.4) Patients in the treatment group 1 are 2 times more likely to die than those who were in the treatment group 1. We are 95% sure that the relative risk would be between 1.2 to 3.4.

Note: As the above which showed that if $p_1 = 0.4$ and $p_2 = 0.2$, the RD = 0.2 and RR = 2. Now, lets make a data from another study to see how the two measure of effect behaves by assuming $p_1 = .04$ and $p_2 = 0.02$. In this study, RD = 0.02 and RR = 2. The RR is exactly the same as the previous study whereas the RD dropped from 20% to 2%. Of course, the former study provided a convincing finding for adopting the treatment group 1 in replacement of the treatment group 2 whereas the later study provide a weak evidence irrespective of the p-value or significant results. This conclusion is based on RD - not RR. On the other hand, if the two studies were etiological study, they concluded the same messages that exposed to the factor are 2 time more likely to die that not exposed (see more details in Jaeschke et al. 1995).

Ex 2.2-3 Testing the hypothesis
General hypothesis:
For observational study

H_0 : There is no association between exposure and death

For experimental study
H_0 : The death rates of the two treatment groups are the same

Specific hypothesis:

H_0 : $\pi_{11} = \pi_{21}$, where

π_{11} is estimated by $p_{11} = \dfrac{n_{11}}{n_{1+}}$, **and**

π_{21} is estimated by $p_{21} = \dfrac{n_{21}}{n_{2+}}$.

This is the hypothesis of homogeneity compares the two binomial distributions implied by the assumptions. Altman (1991); page 234-235 and 250-253 provided formula and work examples.

For the analysis of this example, see the Stata output, at the last line, provided in the previous section (Ex 2.2-2).

So we reject H_0 and conclude that treatment group 2 is significantly better than treatment group 2 (p-value = 0.008).

Ex 2.2-3 Summary findings

Observational study: **A ten-year follow-up study of 80 people with hypertension, 40% died and 70 people with normotension, 20% died. Those who were hypertension were 2 times more likely to die than those who were normotension (95%CI: 1.2 to 3.4). This is statistically significant (p-value = 0.008). The findings suggested that hypertension is a significant predictor of death within ten years.**

Experimental study: **A randomized controlled trial with 2 treatments - 80 subjects in group 1 and 70 subjects in group 2, the death rate was 40% and 20% respectively. The death rate was 20% higher in group 1 than in group 2 (95%CI: 5.7% to 34.2%). This difference is statistically significant (p-value = 0.008). This suggested a better**

efficacy of treatment group 2 in preventing death.

2.4. Case-control study

Select n_{+1} with level 1of variable 1 and n_{+2} with level 2. Determine levels of variable 2. In this case n_{+1} , n_{+2} are set in advance (i.e., column totals are fixed) and n_{11} and n_{12} are random variables.

Example 2.3

A case-control study aimed to determine effect of smoking on lung cancer (CCEB, 1993). One hundred and seventy cases of lung cancer and 430 appropriate controls were chosen to find out whether each person was a smoker or non-smoker in the past. The data is given below.

Table 2.6 Number of lung cancer patients by smoking status - data for example 2.2

		Lung Cancer		
		Cases	**Controls**	
Smoker	**yes**	**160**	**320**	**480**
	No	**10**	**110**	**120**
		170	**430**	**600**

Ex 2.3-1 Describing the proportions

The proportion of interest are :

p₁ = $\frac{160}{170} = 94.1\%$ → **proportion of cases exposed, and**

$p_2 = \dfrac{320}{420} = 74.4\%$ → proportion of control exposed.

(See also the Stata output, in the square, in the next section - Ex 2.3-2)

Ex 2.3-2 Estimating measure of effect

Only the OR is appropriate for a case-control study. The RR cannot be used since it can only be estimated from a cross-sectional or a prospective study. The RR is the ratio of incidence (or prevalence) rates for those with and without the exposure whereas the incidence rates cannot be estimated from a retrospective (case-control) study. Additional details, formula, and work examples can be found in Altman (1991); page 268-270. A single "cci" command of Stata (see StataCorp., 1999; Volume 1: A-G, page 387-390) provides all these (in the oval) as shown below:

```
. cci 160 10 320 110

                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       160          10  |        170       0.9412
        Controls |       320         110  |        430       0.7442
-----------------+------------------------+----------------------
           Total |       480         120  |        600       0.8000
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |             5.5        |   2.831731    10.67328  (Cornfield)
 Attr. frac. ex. |         .8181818       |   .6468592    .9063081  (Cornfield)
 Attr. frac. pop |         .7700535       |
                 +-----------------------------------------------
                             chi2(1) =    29.55  Pr>chi2 = 0.0000
```

OR = 5.5 (95%CI: 2.8 to 10.7) The odds of smoking among cases is 5.5 times the corresponding odds among controls. Assuming the lung cancer is rare, this can be interpreted as the risk. That is,

those who smoked are 5.5 times more likely to develop lung cancer (95%CI: 2.8 to 10.7). This association is statistically significant (p-value < 0.001).

Note that we will never quote p-value = 0.0000 as suggested by last line of the output sine it means impossible which is not the case.

Ex 2.3-3 Testing the hypothesis

General hypothesis:

H_0 : There is no association between smoking and lung cancer,

or

H_0 : Proportion of cases exposed = proportion of controls exposed

Specific hypothesis:

H_0 : $\pi_{11} = \pi_{12}$, where

π_{11} is estimated by $p_{11} = \frac{n_{11}}{n_{+1}}$ **, and**

π_{12} is estimated by $p_{12} = \frac{n_{12}}{n_{+2}}$.

This is the hypothesis of homogeneity compares the two binomial distributions

For the analysis of this example, see the Stata output, at the last line, provided in the previous section (Ex 2.3-2).

So we reject H_0 and conclude smoking is statistically significantly associated with lung cancer (p-value < 0.001).

Ex 2.3-3 Summary findings

A total of 170 case, 94.1% were smokers as compared to 74.4% of 430 controls. The odds of smoking among cases was 5.5 times the corresponding odds among controls. If lung cancer is rare, this can be interpreted as the risk. That is, those who smoked are 5.5 times more likely to develop lung cancer (95%CI: 2.8 to 10.7). This association is statistically significant (p-value < 0.001). The findings suggested that smoking is a significant predictor of lung cancer.

2.5 Matched pairs data

So far we have covered 2-by-2 Table where the data was independent. In some situation, the investigator needs to control effect of extraneous variables on the association of the two variables - the independent variable and the outcome. One approach for such purpose is “matching” study subjects on one or more extraneous variables.

An example of matched data in case-control study is that case with the disease under study is matched with a control. Matching is based on certain criteria such as age, sex, race, etc. Each case and control subject is then classified according to the presence or absence of the study factor or exposure of interest. Matching is undertaken to increase the validity of the inferences by controlling for confounding factors (details discussed under the section of stratified analysis).

Matched data in prospective study without randomization could be done by that each subject with the risk factor present (eg, exposure to an agent) is matched with a control subject on the basis of certain matching criteria who does not have the factor of interest (eg, no exposure). After a specified follow-up period, each subject is classified according to the presence or absence of the response variable (eg, disease). For randomized design, each subject randomly drawn from the target population is paired on the basis of the matching criteria with another randomly selected subject from the target population. Within each pair, the two factor levels (eg, treatments) are randomly allocated to the two members of the pair using a suitable randomization procedure. After a specified follow-up period, each subject is classified according to the presence or absence of the response variable (eg, disease). Matching in controlled trials increases the precision of the comparisons among the treatments.

The correct analysis of a properly matched study retains the pairing. Details for analysis of matched study can be found in Fleiss (1981); page 113-137.

Example 2.4

A matched case-control study aimed to determine effect of smoking on lung cancer (CCEB, 1993). One hundred and seventy cases of lung cancer and 430 appropriate controls were chosen to find out whether each person was a smoker or non-smoker in the past. The data is given below.

Table 2.7 Number of lung cancer patients by smoking status - data for example 2.4

			Controls		
			Without Lung Cancer		
			Smoked	Not smoked	
Cases	*With Lung Cancer*	Smoked	160	320	480
		Not smoked	10	110	120
			170	430	600

Ex 2.4-1 Describing the proportions

The proportion of interest are :

$p_1 = \frac{480}{600} = 80.0\%$ ➡ **proportion of cases exposed, and**

$p_2 = \frac{170}{600} = 28.3\%$ ➡ **proportion of controls exposed.**

(See also the Stata output, in the square, in the next section - Ex 2.4-2)

Ex 2.4-2 Estimating measure of effect

Only the OR is appropriate for a matched case-control study. More details, formula, and work examples can be found in Fliess (1981); page 115-116. A single "mcci" command (see StataCorp., 1999; Volume 1: A-G, page 400-402) provides all these (in the oval) as shown below:

```
.  mcci 160 320 10 110

                 | Controls               |
Cases            |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
         Exposed |       160         320  |        480
       Unexposed |        10         110  |        120
-----------------+------------------------+----------
           Total |       170         430  |        600

McNemar's chi2(1) =    291.21        Pr>chi2 = 0.0000
Exact McNemar significance probability       = 0.0000

Proportion with factor
        Cases             .8
        Controls    .2833333      [95% conf. interval]
                    ---------     --------------------
        difference  .5166667       .4724295   .5609038
        ratio       2.823529       2.492651   3.198329
        rel. diff.  .7209302       .6771888   .7646717

        odds ratio        32       17.17507    67.3789   (exact)
```

OR = 32 (95%CI: 17.2 to 67.4)	**Those who smoked are 32 times more likely to develop lung cancer (95%CI: 17.2 to 67.4). This association is statistically significant (p-value < 0.001).**

Ex 2.4-3 Testing the hypothesis

General hypothesis:

H_0 : There is no association between smoking and lung cancer, or

H_0 : Proportion of cases exposed = proportion of controls exposed

Specific hypothesis:

H_0 : $\pi_{12} = \pi_{21}$, where

π_{12} is estimated by $p_{12} = \dfrac{n_{12}}{n_{12} + n_{21}}$**, and**

π_{21} is estimated by $p_{21} = \dfrac{n_{21}}{n_{12} + n_{21}}$

This is the hypothesis of homogeneity compares the two binomial distributions

For the analysis of this example, see the Stata output, at the line with bold italic letters, provided in the previous section (Ex 2.4-2). McNemar's chi-square can be used for this example since the sample is sufficiently large. (Large sample is defined as n_{12} + n_{21} > 20. If this is not hold, Exact McNemar significance probability test should be used.)

So we reject H_0 and conclude smoking is statistically significantly associated with lung cancer (p-value < 0.001).

Ex 2.4-3 Summary findings

A total of 480 cases of lung cancer, 80.0% were smokers as compared to 28.3% of 170 controls. The odds of smoking among cases was 32 times the corresponding odds among controls. If lung cancer is rare, this can be interpreted as the risk. That is, those who smoked are 32 times more likely to develop lung cancer (95%CI: 17.2 to 67.4). This association is statistically significant (p-value < 0.001). The findings suggested that smoking is a significant predictor of lung cancer.

Note: For matched prospective studies, a comprehensive guide is given by Altman (1991); page 235-241. Data analysis for this type of design can use the same Stata command as that was used in Ex 2.4-2. Proportions used for describing the sample and test of hypothesis can quoted and interpreted the same manner as that fore the matched case-control, except measure of effect where the difference between two proportions (RD) is more appropriate than OR.

2.6 The evaluation of a screening test

Diagnostic test is another form of the 2-by-2 Table that is obtained from a study, the aim of which is to evaluate a diagnostic test intended for use in a screening program. A recommended reading is Altman (1991); page 409-419. Below layouts the table.

Table 2.8 Notation for evaluation of a screening test

Result

		Gold Standard Test	
		Disease Status	
		D	$\bar{\mathbf{D}}$
Diagnostic Test Result	+	n_{11}	n_{12}
	-	n_{21}	n_{22}

Where positive test result (+) indicates the presence of disease.

Followings are the statistics need to be reported for this type of the study. Item 1 to 4 is the must. Items 5 may give further inside to the interpretation of the diagnostic test data. The last item is optional depending in whether or not the diagnostic test has more than 2 categories.

1. **Sensitivity= proportion of diseased who have a +ve test**

   which is estimated by $\dfrac{n_{11}}{n_{11}+n_{21}}$.

2. **Specificity = proportion of non-diseased who have a - ve test**

   which is estimated by $\dfrac{n_{22}}{n_{12}+n_{22}}$.

3. **Positive predictive Value (*PPV*) = proportion of those with a +ve test who have the disease. This is estimated by** $\dfrac{n_{11}}{n_{11}+n_{12}}$.

4. **Negative Predictive Value (*NPV*) = proportion of those with a -ve test who do not have the disease. This is estimated by** $\dfrac{n_{22}}{n_{21}+n_{22}}$.

Note: ***PPV*** **and** ***NPV*** **depend on the prevalence of the disease (Which may or may not be** $\dfrac{n_{11}+n_{21}}{n_{++}}$) **in the population.**

5. **Likelihood ratio positive (LRP) = the ratio of probability of getting that result if the patient truly had the condition of interest with the corresponding probability if they were healthy. This is estimated by sensitivity / (1 - specificity).**

6. **Receiver Operating Charateristic (ROC) curve is a method of measuring and comparing the accuracy of one or more variables at predicting whether each observation is a member of one of two groups/categories. The ROC curve plots the Sensitivity (True Positive rate) against 1-Specificity (False Positive rate). The larger the Area Under the ROC Curve, the better the variable is at predicting group membership. Thus this is appropriate for a single diagnostic test where there were many cut-off values and for the investigator to use for comparing two or more competing methods.**

Example 2.5

This data is taken from Fleiss (1981); page 6. Two thousands of people were undergone two tests - one is a gold standard

test and another is a new diagnostic test. This study aimed to evaluate performance of the test. Data is shown below.

Table 2.9 Number of test results by results from the gold standard - data for example 2.5

		Gold Standard Test	
		D+	**D-**
Diagnostic Test	+	**950**	**10**
	-	**50**	**990**
		1000	**1000**

Step 1: Create a data file in Stata by using the following 5 commands.

```
. tabi 950 10 \ 50 990, replace

           |          col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |       950         10 |       960
         2 |        50        990 |      1040
-----------+----------------------+----------
     Total |      1000       1000 |      2000

           Fisher's exact =                 0.000
   1-sided Fisher's exact =                 0.000

. rename col  gold
. rename row test
. recode gold 1=1 2=0
(2 changes made)


. recode test 1=1 2=0
(2 changes made)
```

Step 2: Calculate the diagnostic performance using 'diagtest' command, available at http://www/sata.com in STB-56 sbe36, as follows:

```
. diagtest  test gold [freq=pop]

           |         gold
      test |         0          1 |     Total
-----------+----------------------+----------
         0 |       990         50 |      1040
         1 |        10        950 |       960
-----------+----------------------+----------
     Total |      1000       1000 |      2000


True D defined as gold ~= 0                           [95% Conf. Inter.]
-------------------------------------------------------------------------
Sensitivity                     Pr( +| D)  95.00%      94.04%   95.96%
Specificity                     Pr( -|~D)  99.00%      98.56%   99.44%
Positive predictive value       Pr( D| +)  98.96%      98.51%   99.40%
Negative predictive value       Pr(~D| -)  95.19%      94.25%   96.13%
-------------------------------------------------------------------------
Prevalence                      Pr(D)      50.00%      47.81%   52.19%
-------------------------------------------------------------------------
```

We can also do that using 'roctab' command as follows:

```
. roctab  gold test  [freq=pop], table detail

           |         test
      gold |         0          1 |     Total
-----------+----------------------+----------
         0 |       990         10 |      1000
         1 |        50        950 |      1000
-----------+----------------------+----------
     Total |      1040        960 |      2000



Detailed report of Sensitivity and Specificity
------------------------------------------------------------------------------
                                           Correctly
Cut point     Sensitivity   Specificity   Classified          LR+          LR-
------------------------------------------------------------------------------
( >= 0 )           100.00%         0.00%       50.00%       1.0000
( >= 1 )            95.00%        99.00%       97.00%      95.0000       0.0505
( >  1 )             0.00%       100.00%       50.00%                    1.0000
------------------------------------------------------------------------------

                       ROC                    -Asymptotic Normal--
            Obs       Area     Std. Err.      [95% Conf. Interval]
         --------------------------------------------------------
           2000     0.9700       0.0038        0.96257     0.97743
```

By the 'roctab' command, we can get the 'Likelihood ratio test' and 'Area under ROC and its 95%CI'. This command is in STB52: sg120 which can be downloaded from http://www/sata.com.

Alternative ways:

First, we fit logistic regression model to the data using "logit" command (see StataCorp., 1999; Volume 2: H-O, page 228-239)

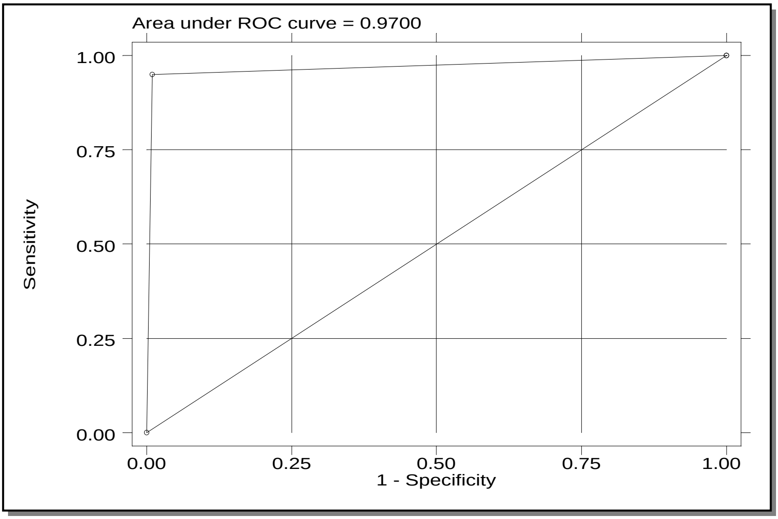
```
. logit gold test [freq=pop]

Iteration 0:   log likelihood = -1386.2944
Iteration 1:   log likelihood = -394.64103
Iteration 2:   log likelihood = -281.61583
Iteration 3:   log likelihood = -259.63674
Iteration 4:   log likelihood = -256.34487
Iteration 5:   log likelihood = -256.11912
Iteration 6:   log likelihood =  -256.1172

Logit estimates                                   Number of obs   =       2000
                                                  LR chi2(1)      =    2260.35
                                                  Prob > chi2     =     0.0000
Log likelihood =  -256.1172                       Pseudo R2       =     0.8153

------------------------------------------------------------------------------
    gold |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    test |   7.539559   .3493483     21.582   0.000       6.854849    8.224269
   _cons |  -2.985682   .1449486    -20.598   0.000      -3.269776   -2.701588
------------------------------------------------------------------------------
```

Second, we obtain the test performance (see StataCorp., 1999; Volume 2: H-O, page 212)

```
. lstat

Logistic model for gold

              -------- True --------
Classified |         D            ~D        Total
-----------+--------------------------+-----------
     +     |       950           10  |        960
     -     |        50          990  |       1040
-----------+--------------------------+-----------
   Total   |      1000         1000  |       2000

Classified + if predicted Pr(D) >= .5
True D defined as gold ~= 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   95.00%
Specificity                     Pr( -|~D)   99.00%
Positive predictive value       Pr( D| +)   98.96%
Negative predictive value       Pr(~D| -)   95.19%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    1.00%
False - rate for true D         Pr( -| D)    5.00%
False + rate for classified +   Pr(~D| +)    1.04%
False - rate for classified -   Pr( D| -)    4.81%
--------------------------------------------------
Correctly classified                        97.00%
--------------------------------------------------
```

Third, we can obtain ROC curve (Note that this is just for illustration use of the Stata command - not appropriate for this example data since the test is dichotomous where is no other choice of cut-off value, see StataCorp., 1999; Volume 2: H-O, page 213)

```
. lroc

Logistic model for gold

number of observations =     2000
area under ROC curve   =   0.9700
```



Note:

i) The 95% confidence intervals for sensitivity, specificity, PPV, and NPV should always be reported. Presentation the confidence intervals for these statistics had been advocated by Harper and Reeves (1999).

ii) To determine an optimal cut-off value, use "lsens" command. The probability for the optimal cut-off value refers to the ordinate at the horizontal axis of the graph

corresponding to the two graphs cross each other. Then use the command "lstat, cutoff( )". The bracket in the option of this command is for the probability mentioned earlier.

iii) To compare two or more diagnostic tests, use "nproc" command. This free program is an automatic do file of Stata that can be download from http://www/sata.com. This command calculates nonparametric area under ROC curve and standard errors for ROC curves for each test. Another useful program is 'roccomp' in STB52: sg120 which can be downloaded from http://www/sata.com as well.

2.7 Stratified analysis

This methods is to adjust or control for the effects of *extraneous* variables, *nuisance* factors, *confounding* variables or *covariables* when assessing the relationship between a dichotomous exposure variable (eg, smoker - yes/no) and a dichotomous outcome (lung cancer - yes/no). In fact, there are several methods for the adjustment (see more details in Chapter 10). For the stratification methods, it can be referred to both pre and post data collection. Pre-stratification randomization is used for controlling effects of extraneous variables in the design stage (before data collection) whereas the post-stratification is a statistical method to do the same purpose in the analysis stage (after data collection).

In general, it is know as stratified analysis. It is performed after data has been collected - thus in experimental studies such as clinical trials it is known as the post-randomization stratification. Theory and examples are best described in Kleinbaum, Kupper, and Morgenstern (1986), page 321-376. A simpler one is in Fleiss (1981); page 160-187.

This involves the formation of similar subgroups (or strata) determined by the levels of the extraneous variable(s). The association between the risk factor and the response variable can be examined within strata or summarized across strata.

This approach is an essential step for the complicated modeling approach to be discussed next. Although its major role is for exploratory data analysis (EDA), stratified analysis can be a final and valid method for a well-designed study. EDA serves as not only a tool for assessing roles of the extraneous variables to enable investigators to make decision as to how the variables will be fitted in the model but also a screening tools for candidates from several variables in hand to be entered into the model. A practical steps, partly modified from Kleinbaum, Kupper, and Morgenstern (1986); page 321-322 and Kleinbaum (1994), involved the following seven steps:

2.7.1 Obtain a measure of association (e.g., relative risk or odds ratio as appropriate) quantifying association between the exposure of interest and the outcome.

2.7.2 Categorize each of the extraneous variables to be controlled. The categorization could be a combination of two or more variables so that more than one extraneous variable could be controlled for their effect at a time.

2.7.3 For the categories defined in step 1, organize the study subjects into combination of categories of each control variable - i.e., cross-tabulate the exposure of interest with the outcome for each group of the extraneous variable. These combinations are called "strata".

2.7.4 Carry out simple analysis within each stratum, using a Mantel-Haenszel χ^2 test for association and an measure of association (e.g., relative risk or odds ratio)

appropriate for the designed used. By this methods, we have stratum-specific measure of association (e.g., OR$_1$, OR$_2$, …, OR$_k$, one for each categories of the k level of extraneous variable).

2.7.5 Carried out a test of homogeneity of the measure of association across stratum (e.g., Woolf's test).

2.7.6 Assessing role of the extraneous variable whether or not it is an effect modifier. Determine if there is an interaction effect If the test of homogeneity in #2.5 suggest a significant different (p-value < 0.05), the interaction effect is existed. The extraneous variable is said to be an effect modifier. If the p-value ≥ 0.05, we can only say that the interaction effect cannot be detected - there may be or may be not. Since it has been known that the test for this effect lack of power, one recommendation would be that investigator should judged about interaction effect based on the magnitude of different of measure of association across stratum. If the difference was considerably clinically or socially important, then we conclude that there was an interaction effect. That is, the association between the exposure of interest and the outcome depend on level of the extraneous variable. Then we report the stratum-specific measure of association and their 95% confidence intervals. The analysis is complete at this step except there was no interaction effect that we need to proceed the next step.

2.7.7 Assessing role of the extraneous variable whether or not it is a confounder. This step is needed only if there was no interaction effect. It involves accumulating information over the strata to obtain the (summary) measure of association - the one that adjusted for effect of the extraneous variable. Comparing the adjusted measure of association with the crude one obtained at

the first step. If they are considerably different and the difference is clinically or socially meaningful, then we conclude that the extraneous variable is a confounder of the association between the exposure of interest and the outcome. In this case, we need to report the adjusted measure of association and its 95% confidence intervals. It they are more or less the same, then the extraneous variable plays no role in the association between the exposure of interest and the outcome. In this case, reporting the crude or adjusted measure of association make no difference since they are similar. However, the adjusted one is preferred since it has been taken into account for effect of the extraneous variable. Kleinbaum (1994) suggested that the one with a narrow confidence intervals is preferred since it is more precise estimates.

Example 2.6

The following example used the example data set described in Example 1.1. Here the descriptions of the variable lists are slightly different from that in Table 1.2. The investigator wanted to examine the effect of V3 (ANC - mother attending antenatal care during pregnancy) on V1 (DEAD - dead within the first month of life) controlling for the effect of V2 (SMK - parents smoking).

Since this is a cross-sectional study, we will use OR as a measure of effect. Thus the following Stata commands will be "cc" - abbreviated from case-control rather than "cs" - abbreviated from cohort study. The different is that the former provides OR and the later provides RR. Note that we have used these commands, but in the immediate form, in the previous section on analyzing data from the three designs - cross-sectional, prospective, and case-control studies. Thus

Step 1 of this example also serves as the example of using these commands for a data set.

Step 1: Performing a crude analysis to examine the association between V3 (ANC) on V1 (DEAD)

```
. use example.dta

. cc v1 v3
```

```
                 | V3                       |              Proportion
                 |   Exposed   Unexposed  |     Total     Exposed
-----------------+------------------------+----------------------
           Cases |        21          44  |        65      0.3231
        Controls |        14         386  |       400      0.0350
-----------------+------------------------+----------------------
           Total |        35         430  |       465      0.0753
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         13.15909       |  6.309044    27.44195  (Cornfield)
 Attr. frac. ex. |         .9240069       |  .8414974    .9635594  (Cornfield)
 Attr. frac. pop |         .2985253       |
                 +-----------------------------------------------
                              chi2(1) =    66.67  Pr>chi2 = 0.0000
```

Children whose mothers attended ANC were 13.2 times more likely to die within the first month of life than those whose mothers did not. This magnitude of association ignored effects of other variables. At this stage, we obtained OR$_{\text{crude}}$ = 13.2.

Step 2: Examining the association between V3 (ANC) on V1 (DEAD) within each stratum of V2 (SMK)

```
. cc v1 v3 if v2 = = 0
```

```
                 | V3                       |              Proportion
                 |   Exposed   Unexposed  |     Total     Exposed
-----------------+------------------------+----------------------
           Cases |         2          26  |        28      0.0714
        Controls |         6         190  |       196      0.0306
-----------------+------------------------+----------------------
           Total |         8         216  |       224      0.0357
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         2.435897       |         0    11.25226  (Cornfield)
 Attr. frac. ex. |         .5894737       |         .     .911129  (Cornfield)
 Attr. frac. pop |         .0421053       |
                 +-----------------------------------------------
                              chi2(1) =     1.19  Pr>chi2 = 0.2763
```

```
. cc v1 v3 if v2 = = 1

                 | V3                     |
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        19          18  |         37      0.5135
        Controls |         8         196  |        204      0.0392
-----------------+------------------------+------------------------
           Total |        27         214  |        241      0.1120
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         25.86111       |   10.08458    66.21777  (Cornfield)
 Attr. frac. ex. |         .9613319       |   .9008387    .9848983  (Cornfield)
 Attr. frac. pop |         .4936569       |
                 +-------------------------------------------------
                               chi2(1) =    70.82  Pr>chi2 = 0.0000
```

At this step, we obtain OR describing association between ANC and DEAD for each group of SMK. That is, OR_{Smoked} = 25.9 and $OR_{Not\ smoked}$ = 2.4. In practice, we need not to do this since the command used in the next step. This is for illustration and displaying the data in two separate tables (bold italic letters in the square).

Step 3: Performing a stratified analysis to examine the association between V3 (ANC) on V1 (DEAD) adjusted for the effect of V2 (SMK)

```
. cc v1 v3, by(v2)

              V2 |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   2.435897            0   11.25226      .6964286 (Cornfield)
               1 |   25.86111     10.08458   66.21777      .5975104 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   13.15909     6.309044   27.44195               (Cornfield)
    M-H combined |   13.25311     6.309988     27.836
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     5.91  Pr>chi2 = 0.0150

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =     63.22
                                                Pr>chi2 =    0.0000
```

In practice, we need only this command for stratified analysis since it provides all statistics needed. We will summary only the necessary ones - the four components, as follows:

1) The crude measure of effect

OR_{crude}	**=**	**13.2**

2) The stratum-specific measure of effect

OR_1	**=**	**2.4**
OR_2	**=**	**25.9**

3) The adjusted measure of effect

$OR_{adjusted}$	**=**	**13.3**

4) Test of homogeneity of OR across stratum

p-value	**=**	**0.015**

Following the steps described in 2.7.1 to 2.7.7, we conclude that there is a significant interaction effect of SMK on the association between ANC and DEAD (p-value = 0.015). Thus the adjusted measure of effect ($OR_{adjusted}$ = 13.3) is less useful. The stratum-specific measure of effects was then more appropriate.

Step 4: Summary findings

Ignoring effects of parent smoking status, children whose mothers attended ANC were 13.2 times more likely to die within the first month of life than those whose mothers did not. There is a significant interaction effect of parent smoking on the association between mother attending ANC and dead of children (p-value = 0.015). That is, the effect of mother attending ANC on dead of children depended on whether or not their parent smoked. For smoker parents, children whose mothers attended ANC were 25.9 times more likely to die within the first month of life than those whose mothers did not (95%CI: 10.1 to 66.2). For non-smoker parents, children whose mothers attended ANC were 2.4 times more likely to die within the first month of life than those whose mothers did not (95%CI: 0.0 to 11.3). Note that these confidence intervals

may not be valid due to small sample, thus exact confidence intervals are preferred.

Note:

1. **In the above example, both the crude and the adjusted measure of effects are not a valid measure of effect in quantifying the association between ANC and DEAD. However they should not totally be ignored in drawing the conclusion or at least they should be mentioned in the discussion section. For example in the above case, comparing the crude OR and the stratum-specific OR we feel that it is far more to believe the crude OR and that GENDER plays a large effects on the association under investigation. This is why the table presenting the results (see Chapter 10) includes both the crude and adjusted measure of effects.**

2. **The presented analysis is adjusted for effect of only one extraneous variable while, in the real world, children death is likely to be affected by several variables. Thus conclusion drawn from this should be very caution about lacking of controlling for effects of several other factors. The most efficient analysis will be discussed in Chapter 6.**

3. **The above example is for observational studies. For experimental studies such as clinical trials, however, we are interested in the RD rather than RR or OR. Followings are some useful Stata commands of doing these. Here we assume the example data is from a clinical trial where V1 is a treatment outcome (1=cured, 0=not cued) and V2 is a treatment (1=drug A, 0=drug B). The investigator randomly allocated the patients into each treatment using stratified block randomization where the stratified variable is V3 which is age group (1= old, 0=young). The trial aims to determine the efficacy of drug A as compared to the**

standard drug B. The first two commands are the crude analysis providing identical results, showing how cured rates (in oval) for each treatment and the rate difference (in the squares) are presented in the outputs. The last command is to quantify magnitude of effect, taken into account of the effect of V3.

```
. cs v1 v2

                 | V2                     |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        37          28  |         65
        Noncases |       204         196  |        400
-----------------+------------------------+----------
           Total |       241         224  |        465
                 |                        |
            Risk |   .153527        .125  |   .1397849
                 |                        |
                 |      Point estimate    |   [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |          .028527       | -.0342996    .0913535
      Risk ratio |         1.228216       |   .778466    1.937803
 Attr. frac. ex. |         .1858108       | -.2845776    .4839517
 Attr. frac. pop |         .1057692       |
                 +-------------------------------------------------
                               chi2(1) =     0.79  Pr>chi2 = 0.3754
```

```
. prtest v1, by(v2)

Two-sample test of proportion                      0: Number of obs =      224
                                                   1: Number of obs =      241

------------------------------------------------------------------------------
Variable |       Mean    Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |       .125    .0220971   5.65685   0.0000        .0816905    .1683095
       1 |    .153527    .0232215   6.61141   0.0000        .1080137    .1990403
---------+--------------------------------------------------------------------
    diff |   -.028527    .0320549                          -.0913535    .0342996
         |  under Ho:    .0321831  -.886396   0.3754
------------------------------------------------------------------------------

               Ho: proportion(0) - proportion(1) = diff = 0

       Ha: diff < 0            Ha: diff ~= 0              Ha: diff > 0
        z = -0.886              z = -0.886                 z = -0.886
      P < z = 0.1877        P > |z| = 0.3754           P > z = 0.8123
```

```
. cs v1 v2, by(v3) istandard rd

               V3 |       RD       [95% Conf. Interval]        Weight
------------------+----------------------------------------------------
                0 |  -.0362582    -.0934065   .0208901           214
                1 |   .4537037     .1077276   .7996798            27
------------------+----------------------------------------------------
            Crude |    .028527    -.0342996   .0913535
I. Standardized |   .0186338    -.0452218   .0824894
```

Note that, ignoring effect of age, Drug A was 2.9% higher cured rate than Drug B. However, this effect was reverse in young age group. That is, Drug B was 3.6% higher cured rate than Drug A. On the other hand, among old age group, Drug A was 45.4% higher cured rate than Drug B. This suggested an interaction effect and the adjusted rate difference of 1.9% should be disregarded.

Chapter references

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

Altman, D.G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

CCEB (Centre for Clinical Epidemiology and Biostatistics). The University of Newcastle, Australia. (1993). *STAT402:Analysis of categorical data - 2-by-2 Table.* Newcastle: The University of Newcastle, NSW. Australia.

Davies, H.T.O., Crombie, I.K., Tavakoli, M. (1998). When can odds ratios mislead? *BMJ*. 316:989-991

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. 2nd edition. New York: John Willey & Sons.

Harper, R., and Reeves, B. (1999). Reporting of precision of estimates for diagnostic accuracy: a review. *BMJ*. 318: 1322-1323.

Jaeschke, R., Guyatt, G., Shannon, H., Walter, S. Cook, D. Heddle, N. (1995). Assessing the effects of treatment: measures of association . *Canadian Medical Association Journal*. 152: 351-357

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982). *Epidemiologic research: principles and qualitative methods*. London: Lifetime Learning Publications.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

Exercise

1. Daniel (1991); page 550 provide a problem that a group of 350 adults who participated in a health survey were asked whether or not they were on a diet. The responses by gender are given in the table below.

	Gender		
	Male	Female	Total
On diet	14	25	39
Not on diet	159	152	311
Total	173	177	350

Do these data suggest that being on a diet is dependent on gender ?

i) State what type of study this is and an appropriate null hypothesisgender?

ii) Test the null hypothesis.

iii) Calculate a measure of the association and a 95% confidence interval.

iv) Summarize your findings.

2. A retrospective study on deaths in all men aged 50-54 over a one month period indicated that of 35 men who died from cardiovascular disease (CVD), 5 were on a high salt diet before they died, whereas of 25 men who died from other causes 2 were on such a diet. Is there a relationship between dying from CVD and a high salt diet.?

3. The following data are from a cases-control study of oral contraceptive use in relation to myocardial infarction (MI) (Shapiro *et al*, 1979).

OC use	Cases	Controls
E+	29	135
E-	205	1607
Total	234	1742

i) State the null hypothesis.

ii) Perform a test appropriate to the null hypothesis.

iii) Calculate a measure of association and a 95% confidence interval.

iv) Summarize your findings.

4. **The following table presents data from a matched case-control study where E denotes exposure to the variable of interest and $\bar{E}$ denotes no exposure.**
   - **i) State the hypothesis being tested.**
   - **ii) Test this hypothesis.**
   - **iii) Calculate the odds ratio and interpret it.**
   - **iv) Obtain a 95% confidence interval for the odds ratio. What information is conveyed by this interval ?**
   - **v) Write a short report summarizing your findings.**

	Controls		
Cases	E+	E-	Total
E+	15	20	35
E-	5	60	65
Total	20	80	100

5. **Followings are adapted from Kleinbaum, Kupper, and Morgenstern (1982); page 363-365. A follow-up study on the utilisation of a vaccine at a large hospital. Six hundred mothers who delivered their babies at the hospital were assessed for their perception of vaccination and then followed for one year to find that whether or not their children were vaccinated. Perception of vaccination were assessed using series of questions and then classify mothers into two groups - positive receptive perception, i.e., perceived benefit of vaccination, and negative receptive perception, i.e., perceived no benefit of vaccination.**

56

Vaccination status of children was obtained from their Expanded Program on Immunization (EPI) cards. The following table summarizes the study children whose mothers were recruited in the study, by type of perception (R+ versus R-), vaccination status (V+ versus V-), parents living together (YES versus NO) and sex (M versus F.)

Parents living together	Sex	V+		V-		Total
		R+	R-	R+	R-	
Yes	Male	68	17	172	43	300
	Female	8	12	52	78	150
No	Male	1	4	9	36	50
	Female	81	9	9	1	100
Total		158	42	242	158	600

i) **Examine the relationship between vaccine receptive perception and vaccine acceptance ignoring the effects of parents living together and sex. State the null hypothesis being tested. Perform a test of significance and obtain a measure of the association. Calculate a 95% CI for this measure.**

ii) **Ignoring parents living together, does sex appear to be confounding the association between vaccine receptive perception and vaccine acceptance? Explain your answer.**

iii) **Ignoring sex, does parents living together appear to be confounding the association between vaccine receptive perception and vaccine acceptance? Explain your answer.**

iv) **Stratifying on both parents living together and sex simultaneously, how do the resulting stratum-specific measures of association compare with the crude estimate and the adjusted estimates based on controlling for sex and parents living together separately?**

v) **What conclusion can you draw about the effect of parents living together and sex on the observed relationship between vaccine receptive perception and vaccine acceptance?**

vi) **Based on your results discuss whether or not vaccine receptive perception is a determinant of vaccine acceptance.**

vii) **Summarize your findings**

Chapter 3

Analysis of 2-by-C Tables

Chapter Objectives

After completing this chapter, readers should be able to:

- **describe appropriate proportions and calculate measures of association for 2-by-C tables and corresponding 95% confidence intervals;**
- **test hypotheses appropriate to 2-by-C Tables;**
- **perform a test for trend in the proportions in a 2-by-C table where the column variable is ordinal and i) the column totals are fixed, and ii) the row totals are fixed; and**
- **interpret the results from the analysis.**

Contents

3.1 Introduction

So far we have covered analyzing a dichotomous outcome with a dichotomous independent variable. This chapter, we expand the type of the independent variable to be more than two categories. Tables where one variable (say variable 2 - see tables below) has more than two levels are a straightforward extension of our analysis of 2-by-2 tables. The general form of the test statistics remain the same. The different study designs again give rise to the same test statistic. An additional consideration is when variable 2 has ordered categories (such as severity of disease: none, mild, moderate, severe). The question of trend or dose response is a unique issue for this type of study. Below is a general form of the 2-by-C Table.

Table 3.1 Notation of observed data

		Variable 2				
		1	**2**	**...**	**C**	
Variable 1	**1**	n_{11}	n_{12}	**...**	n_{1C}	n_{1+}
	2	n_{21}	n_{22}	**...**	n_{2C}	n_{2+}
		n_{+1}	n_{+2}	**...**	n_{+C}	n_{++}

Several analytical methods exist to account for the quantitative nature of the categories to improve the chi-square test. An excellent comprehensive and readable review of theories and practical examples was given by Altman (1991); page 259-265. At the Stata web site, there is a frequently asked questions by Sribney (1999) comparing several methods implemented by several software regarding test for trend.

Followings we will discuss two main types of the 2-by-C Table based on type of Variable 2 in reference to Table 3.1. That is, the section of nominal and ordinal variable. Within each section, there was 2 type of studies based on that whether row or column is fixed. For small sample, assuming both row and column total fixed, the exact method is appropriate and it was described in Chapter 8.

3.2 Nominal variable

3.2.2 The row and column totals are fixed.

This type of data is from a cross-sectional design in which n_{++} individuals are chosen. The frequencies (n_{ij} *where* $i = 1, 2; j = 1, ..., c$) follow a full multinomial distribution. The null hypothesis is that the row variable and column variable are independent (see Table 3.2 for notation of the population proportions).

$$H_0 : \pi_{11} = \pi_{i+} \ \pi_{+j} \ \text{where} \sum_i \sum_j \pi_{ij} = 1 \qquad i = 1, 2$$

$$j = 1, ..., c$$

Table 3.2 Notation of population proportion

		Variable 2				
		1	2	...	c	
Variable 1	1	π_{11}	π_{12}	...	π_{1c}	
	2	π_{21}	π_{22}	...	π_{2c}	

Example 3.1

Five hundred and eleven subjects were recruited to examine the relationship between social class (3 levels) and major

depression (2 levels). The results are summarized in the following table.

Table 3.3 Number of depressive patients by social class - data for example 3.1

Depressed	Social class			Total
	Lower	Middle	Upper	
Yes	4	13	5	22
No	159	212	118	489
Total	163	225	123	511

Ex 3.1-1 Describing the proportions

For the purpose of describing the proportions, we assume the groups under the independent variable known in advance, i.e., column totals are fixed. Therefore, the proportion that will be used for describing this data can be calculated as follows:

The proportion of the lower class who depressed :

$p_1 = 4/163 = 0.024$

The proportion of the middle class who depressed :

$p_2 = 13/225 = 0.058$

The proportion of the upper class who depressed :

$p_2 = 5/123 = 0.041$

Considering the output in Ex3.1-3, the proportions reported here are in the square.

Ex 3.1-2 Estimating measure of effect

For the 2-by-C Table, we can use "local" odds ratio as the measure of effect. By choosing appropriate reference category, we can calculate OR for other categories compared with the

reference. Here we choose the lower class as it gave the lowest proportion of depressed (see EX3.1-1 shown above). For "cci" command see StataCorp. (1999), Volume 1: A-G, page 387-389.

```
. cci 13   212  4  159
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        13         212  |        225      0.0578
        Controls |         4         159  |        163      0.0245
-----------------+------------------------+----------------------
           Total |        17         371  |        388      0.0438
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         2.4375         |   .819265     7.23088  (Cornfield)
 Attr. frac. ex. |        .5897436        | -.2206063    .8617042  (Cornfield)
 Attr. frac. pop |        .0340741        |
                 +-----------------------------------------------
                             chi2(1) =     2.49  Pr>chi2 = 0.1144

. cci  5  118  4  159
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |         5         118  |        123      0.0407
        Controls |         4         159  |        163      0.0245
-----------------+------------------------+----------------------
           Total |         9         277  |        286      0.0315
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |        1.684322        |  .4784226    5.921692  (Cornfield)
 Attr. frac. ex. |        .4062893        | -1.090202    .8311294  (Cornfield)
 Attr. frac. pop |        .0165158        |
                 +-----------------------------------------------
                             chi2(1) =     0.60  Pr>chi2 = 0.4397
```

Ex 3.1-3 Testing the hypothesis

By the definition of independence, characteristics 1 and 2 and independent if each joint proportion π_{11}, π_{12}, π_{21} , π_{22} is the product of the two corresponding total or marginal proportions, ie,

H_0 : major depression is independent of social class

$$H_0 \ : \ \pi_{11} = \pi_{i+}\ \pi_{+j} \quad \text{where} \qquad \sum_i \sum_j \pi_{ij} = 1;\ i = 1, 2$$

$$j = 1, ..., c$$

This is the hypothesis of independence. This form is a specific hypothesis.

For a general hypothesis, we can state that

H_0 : There is no association between social class and depressive disorder

For "tabi" command see StataCorp., (1999), Volume 4: Su-Z, page 144-152.

```
. tabi 4   13   5 \ 159  212  118, col chi2

           |               col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |         4         13          5 |        22
           |      2.45       5.78       4.07 |      4.31
-----------+---------------------------------+----------
         2 |       159        212        118 |       489
           |     97.55      94.22      95.93 |     95.69
-----------+---------------------------------+----------
     Total |       163        225        123 |       511
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =   2.5573   Pr = 0.278
```

Chi-square of 2.56 with 2 degree of freedom gives p-value = 0.278. The null hypothesis is not rejected. We have no sufficient information to conclude that there is an association between social class and depressive disorder.

Ex 3.1-4 Summary findings

This cross-sectional study involved 511 people. The lower class people had a lowest proportion of being depressed. That is, among a total of 163 who were the lower class, 2.5% were depressed whereas among 225 who were the middle class, 5.8 % were depressed and 118 who were the upper class, 4.1% were depressed. The middle class was 2.4 times more likely to be depressed than the lower class (95%CI: 0.8 to 7.2) while the upper class was 1.7 times more likely to be depressed than the

lower class (95%CI: 0.5 to 5.9). However, these were not statistically significant (p-value = 0.278).

3.2.2 The column totals are fixed.

Table 3.4 Notation of population proportion in which the column totals are fixed

	Variable 2				
Variable 1	**1**	**2**	**...**	***c***	**Total**
1	π_1	π_2	**...**	π_c	
2	$1 - \pi_1$	$1 - \pi_2$	**...**	$1 - \pi_c$	
Total	**1**	**1**		**1**	

This is equivalent to choosing n_{+1} of type 1 in variables 2, n_{+2} of type 2, n_{+c} of type *c*. We are interested in the proportion that fall into level 1 of variable 1.

The null hypothesis is expressed in terms of homogeneity of the probabilities $\pi_1, \ldots, \pi_c$.

$$H_0 : \pi_1 = \pi_2 = \ldots = \pi_c = \pi$$

n_{ij} is binomial with parameters n_{+j} and π_j thus $E[n_{ij}] = n_{+j}\,\pi_j$. So the expected values are the same as those obtained under the hypothesis of independence.

The statistical approach for this type of study is similar to what has been performed in the above example (Ex 3.1).

Another approach given by Fleiss (1981); page 138 - 143 considered the above situation as the problem concerning comparison of a number of proportions. In this case, it can be called the R-by-2 Tables.

3.3 Ordinal variable

3.3.1 Column totals are fixed

Suppose in a 2-by-C table that the column variable is ordinal. A question of interest is whether there is a trend in the proportions falling into the first (or second) row across levels of the column variable.

In general the groups represented by the column variable may correspond to different values of a quantitative variable such as age or they may correspond to qualitative categories such as severity of disease, which can be ordered, but not necessarily assigned as numerical value. One might ask whether there is a significant trend in the proportion falling into the first row from group 1 to group C.

Assign a quantitative variable x to the groups. The variable x takes the value x_1,..., x_c. For example x may take the integer values 1, ..., C or values corresponding to the group defined by the categories. The table can be displayed as follows:

Table 3.5 Notations for R-by-C Tables where column totals are fixed

Group	1	2	...	c	
x	x_1	x_2	...	x_c	Total
Positive	n_{11}	n_{12}	...	n_{1c}	n_{1+}
Negative	n_{21}	n_{22}	...	n_{2c}	n_{2+}
Total	n_{+1}	n_{+2}	...	n_{+c}	n_{++}
Proportio n Positive	p_1	p_2	...	p_c	$p = \frac{n_{1+}}{n_{++}}$

The numerator of the χ^2 statistic is $\sum_{j=1}^{c} n_{+j}(p_j - p)^2$ which is a weighted sum of squares of the p_j about the mean p. It also turns out to be a straightforward sum of square (SS), between groups, of a variable (y) taking the value 1 for each individual classified as positive and 0 for each individual classified as negative. This SS can be divided as in ANOVA and regression into the SS due to the regression of y on x and a SS due to departures from linear regression. If there is a trend of p_j with x_j we might find the first SS (regression) to be greater than the second SS. Dividing the portion of the SS due to regression by p(1-p) gives us a chi-square statistic with 1 df which is part of the overall chi-square statistic and is particularly sensitive to trend. The formula for this χ^2 is

$$\chi_1^2 = \frac{n_{++}\left(n_{++}\sum_j n_{1j}x_j - n_{1+}\sum n_{+j}x_j\right)^2}{n_{1+}(n_{++} - n_{1+})\left\{n_{++}\sum n_{+j}x_j^2 - \left(\sum n_{+j}x_j\right)^2\right\}}$$

The difference between χ_{c-1}^2 and χ_1^2 may be regarded as a χ^2 statistic with (c-2) df testing departures from linear regression of p_j on x_j. These chi-square tests are approximate, but the approximation is likely to be adequate if only a small proportion of the expected frequencies are less than 5.

Example 3.2

The data is taken from Holmes and Williams (1954) cited by Agresti (1990); page 297. Children on each tonsil size were classified on whether they are carriers of the pathogenic virus as follows.

Table 3.6 Number of tonsilitis patients by type of carriers - data for example 3.2

	Tonsils			
	Present/Not Enlarged	Enlarged	Greatly Enlarged	Total
Carrier	19	29	24	72
Non-Carrier	497	560	269	1326
Total	516	589	293	1398

Ex 3.2-1 Describing the proportions

Clearly this data represents a column total fixed table. The proportion of carriers for not enlarge is .0368 or 3.7%, for enlarged is .0492 or 4.9%, and for greatly enlarge is .0819 or 8.2%. They were shown in the rectangular of the "tabi" command of Stata shown below.

```
. tabi 19 29 24 \ 497 560 269, col chi2

           |                col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        19         29         24 |        72
           |      3.68       4.92       8.19 |      5.15
-----------+---------------------------------+----------
         2 |       497        560        269 |      1326
           |     96.32      95.08      91.81 |     94.85
-----------+---------------------------------+----------
     Total |       516        589        293 |      1398
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =   7.8848   Pr = 0.019
```

Ex 3.2-2 Estimating measure of effect

As mentioned in EX3.1-2, we can use "local" odds ratio as the measure of effect. By choosing not enlarged tonsil as the

reference category, we can then calculate OR for enlarged (1.4) and greatly enlarged (2.3) as follows:

```
. cci 29 19 560 497

                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        29          19  |         48       0.6042
        Controls |       560         497  |       1057       0.5298
-----------------+------------------------+----------------------
           Total |       589         516  |       1105       0.5330
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         1.354605       |  .7556637    2.427803  (Cornfield)
 Attr. frac. ex. |         .2617776       | -.3233399     .588105  (Cornfield)
 Attr. frac. pop |         .1581573       |
                 +-----------------------------------------------
                              chi2(1) =     1.02  Pr>chi2 = 0.3125
```

```
. cci 24 19 269 497

                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        24          19  |         43       0.5581
        Controls |       269         497  |        766       0.3512
-----------------+------------------------+----------------------
           Total |       293         516  |        809       0.3622
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |          2.33379       |  1.265725    4.302563  (Cornfield)
 Attr. frac. ex. |         .5715124       |  .2099393    .7675804  (Cornfield)
 Attr. frac. pop |         .3189837       |
                 +-----------------------------------------------
                              chi2(1) =     7.55  Pr>chi2 = 0.0060
```

Ex 3.2-3 Testing the hypothesis

The three steps for hypothesis testing is as follows:

Step 1 Overall test for association

The overall χ^2_{c-1} is 7.88 with C-1 = 3-2 = 2 *df* as shown in the oval of the above output from Stata (Ex. 3.2-1).

Step 2 Test for linear trend

The chi-square test for trend χ_1^2 is 7.19 with 1 *df* corresponding to p-value = 0.007. Using Stata to get these results needs a few commands. First we need to obtain a data file. By using "tabi" with "replace option, we can have a summary form of the file as listed using the "list" command as shown below.

```
. tabi 19 29 24 \ 497 560 269, replace

           |               col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        19         29         24 |        72
         2 |       497        560        269 |      1326
-----------+---------------------------------+----------
     Total |       516        589        293 |      1398

          Pearson chi2(2) =   7.8848   Pr = 0.019


. list

          row       col           pop
  1.        1         1            19
  2.        1         2            29
  3.        1         3            24
  4.        2         1           497
  5.        2         2           560
  6.        2         3           269
```

Note that "row" variable represents carrier status and "col" variable represents tonsil size. The "pop" variable is the frequency of each combination of "row" and "col" variables. The "tabodds" command (see StataCorp., (1999), Volume 1: A-G, page 396-397) can be used for testing for linear trend. But we need to have the dependent dichotomous variable coded as 0 and 1. Thus we first use "recode" command (see StataCorp., (1999), Volume 3: P-St, page 136) for that purpose and followed by the "tabodds" command as follows:

```
. recode row 2=0
(3 changes made)
```

```
. tabodds row col [freq=pop]

------------+------------------------------------------------------------
        col |      cases     controls       odds      [95% Conf. Interval]
------------+------------------------------------------------------------
          1 |         19          497    0.03823        0.02418   0.06045
          2 |         29          560    0.05179        0.03565   0.07522
          3 |         24          269    0.08922        0.05877   0.13546
------------+------------------------------------------------------------
Test of homogeneity (equal odds): chi2(2)  =     7.88
                                  Pr>chi2  =   0.0195

Score test for trend of odds:     chi2(1)  =     7.19
                                  Pr>chi2  =   0.0073
```

Alternatively, one could use the "nptrend" command (see StataCorp., 1999, Volume 2: H-O, page 465-468) for performing test for trend.

Followings are to show how the results above are achieved. In the absence of scores for tonsil size we will use -1, 0, 1.

$$\therefore \quad \chi_1^2 \quad = \quad \frac{1398\left[1398(-19+24)-72(-516+293)\right]^2}{72(1326)\left[1398(849+293)-(-516+293)^2\right]}$$

$$= \quad \frac{1398\left[1398\times 5+72\times 223\right]^2}{72\times 1326\left[1398\times 809-(-223)^2\right]}$$

$$= \quad \mathbf{7.19}$$

Step 3 Test for departure from linear trend
The test for departures from a linear trend is

$$\chi_{c-1}^2 \text{ - } \chi_1^2 \;=\; \mathbf{7.88\ -\ 7.19\ =\ 0.69}$$

which is clearly non-significant compared with a chi-square distribution with (C-1) - 1 = (3-1) - 1 = 1 *df* which yields p-value = 0.406. The "disp chiprob()" command shown below can be used to obtain the chi-square probability.

```
. disp chiprob(1, 0.69)
.40616438
```

Thus there is a definite trend which may result in approximately equal increases in the proportion of carriers with increasing tonsil enlargement. In other words, almost all relationship between carrier status and tonsil size was explained by the linear trend.

Ex 3.2-4 Summary findings

Among the three groups based on tonsil size - 516 not enlarged, 589 enlarged, and 293 greatly enlarged, the proportion of the virus carriers were 3.7%, 4.9%, 8.2% respectively. That is, the proportion of the virus carrier increase as the tonsil size increased. Those who had enlarged tonsil size were 1.4 times as likely to be carriers of the virus as those who had not enlarged tonsil size (95%CI: 0.8 to 2.4). Likewise, those who had enlarged tonsil size were 2.3 times as likely to be carriers of the virus as those who had not enlarged tonsil size (95%CI: 1.3 to 4.3). There is a definite trend which may result in approximately equal increases in the proportion of carriers with increasing tonsil enlargement (The overall chi-square test for association = 7.88, p-value = 0.020; the chi-square test for linear trend = 7.19, p-value = 0.007, and the test for departure from linear trend = 0.69, p-value = 0.406.).

3.3.2 Row totals are fixed

Now we demonstrate the approach for that the row totals are fixed and the column variable is ordinal.

Example 3.3

An experiment on the use of sulfones and streptomycin drugs in the treatment of leprosy from Cochran (1954) cited by Agresti (1990); page 101, leprosy patients with different severity levels of disease (little or much infiltration) were graded according to their improvement after treatment as follows.

Table 3.7 Number of leprosy patients for each degree of infiltration by level of changes in health - data for example 3.3

Degree of Infiltration	Changes in Health					Total
	Marked	*Moderate*	*Slight*	*Stationary*	*Worse*	
Little	11	27	**42**	53	11	**144**
Much	7	15	**16**	13	1	**52**
Total	**18**	**42**	**58**	**66**	**12**	**196**

There are two questions of interest:

i) Are the two groups (little *vs* much) homogeneous in their degree of improvement?

ii) Is the degree of improvement the same in the two groups?

Ex 3.3-1 Describing the proportions
Followings are the row proportions.

Little	**0.076**	**0.188**	**0.292**	**0.368**	**0.076**	**1.0**
Much	**0.135**	**0.288**	**0.308**	**0.250**	**0.019**	**1.0**

We can obtain these proportions using "tabi" command as follows:

```
. tabi 11 27 42 53 11 \ 7 15 16 13 1, row chi2
```

```
           |                          col
       row |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
         1 |        11         27         42         53         11 |       144
           |      7.64      18.75      29.17      36.81       7.64 |    100.00
-----------+-------------------------------------------------------+----------
         2 |         7         15         16         13          1 |        52
           |     13.46      28.85      30.77      25.00       1.92 |    100.00
-----------+-------------------------------------------------------+----------
     Total |        18         42         58         66         12 |       196
           |      9.18      21.43      29.59      33.67       6.12 |    100.00

          Pearson chi2(4) =   6.8807   Pr = 0.142
```

We can see that subjects with much infiltration are more likely to show an improvement than subjects with little infiltration. This suggests that the degree of improvement is differ. (The test hypothesis can provide how likely this difference could happen by chance - see Ex 3.3-3 below.)

Ex 3.3-2 Estimating measure of effect

Similar to the previous two examples, the "local" odds ratios might be used as the measure of association for the 2-by-C Table where the row total is fixed. Additionally, we can think of this problem as comparing continuous outcome between the two groups. One way of examining this is to score the categories of improvement and compare the mean scores across the two groups. One scoring scheme is to grade the responses 5, 4, 3, 2, and 1 corresponding to marked improvement through to a worsening of infiltration. The mean scores are given by

$$f_i = \sum_{j=1}^{c} x_j \pi_{ij} \qquad \boldsymbol{i = 1,2}$$

where x_j is the score corresponding to categories *j*. The higher score, the more the improvement.

```
. tabi 11 53 42 27 11 \ 1 13 16 15 7, replace

           |                          col
       row |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
         1 |        11         53         42         27         11 |       144 
         2 |         1         13         16         15          7 |        52 
-----------+-------------------------------------------------------+----------
     Total |        12         66         58         42         18 |       196 

          Pearson chi2(4) =   6.8807   Pr = 0.142


. list

          row       col           pop
  1.        1         1            11
  2.        1         2            53
  3.        1         3            42
  4.          1           4             27
  5.          1           5             11
  6.          2           1              1
  7.          2           2             13
  8.          2           3             16
  9.          2           4             15
 10.          2           5              7
```

```
. expand pop
(186 observations created)
```

At this stage we have a data file of 196 records where "row" variable refers to infiltration groups (i.e., 1= Little, 2=Much) and "col" is the degree of improvement (i.e., 1=Marked, 2= Moderate, 3=Slight, 4=Stationary, and 5=Worse).

To estimate f_j use

$$\hat{f}_i = \sum_j x_j p_{ij}$$

where $p_{ij} = \frac{n_{ij}}{n_{i+}}$ **represents the proportion in row *i* falling into level *j* of variable 2.**

For simplicity of calculation, another scoring scheme is to grade the responses 3, 2, 1, 0, and -1 corresponding to marked improvement through to a worsening of infiltration. Thus the mean score for each group can be calculated as follows:

$\hat{f}_1$ **= 3(.076) + 2(.188) + 1(.292) - .076 = 0.819**

$\hat{f}_2$ **= 3(.135) + 2(.288) + 1(.308) - (.019) = 1.269**

The difference of mean score between the two groups is $\hat{f}_2$ **-** $\hat{f}_1$ **= 0.819 - 1.269 = 0.45. We can also think of this as using two-sample t-test as follows.**

```
. replace row = 0 if row == 2
(52 real changes made)
```

The above command is needed for coding the dependent variable to be 0 and 1 so that the difference is not negative and it is necessary for further analysis using "tabodds" command discussed in EX3.3-3 below.

We now use "ttest" command (see StataCorp., (1999), Volume 4: Su-Z, page 225-232) to estimate the difference of mean score between the two group.

```
. ttest col, by(row)
```

```
Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      52    3.269231     .145613    1.050031      2.9769    3.561561
       1 |     144    2.819444    .0890535    1.068643    2.643413    2.995476
---------+--------------------------------------------------------------------
combined |     196    2.938776    .0771119    1.079566    2.786695    3.090856
---------+--------------------------------------------------------------------
    diff |            .4497863    .1721066                .1103461    .7892265
------------------------------------------------------------------------------
Degrees of freedom: 194

                      Ho: mean(0) - mean(1) = diff = 0

     Ha: diff < 0               Ha: diff ~= 0              Ha: diff > 0
       t =   2.6134               t =   2.6134              t =   2.6134
   P < t =   0.9952         P > |t| =   0.0097          P > t =   0.0048
```

The difference of the mean score is 3.269231 - 2.819444 = 0.45. The mean score for each group were different from that were obtained previously (i.e., $\hat{f}_2$ - $\hat{f}_1$ = 0.819 - 1.269) due to different scoring scheme while the difference is exactly the same (i.e., 0.45). This difference can be used as the magnitude of the effect. That is, the mean score of degree of improvement

of the "much infiltration" group was 0.045 greater than that of the " little infiltration" group (95%CI: 0.11 to 0.79).

Note that the score is arbitrary. Thus it is difficult to interpret. This approach is much useful in the situation where the C variable is quantitative such as diameter of leprosy wound classified as less the 10, 10 to 20, and 20 or more centimeters. Here in the current example it was a qualitative C variable. The local odds ratios could be used as the interpretation is straight forwards. Assign the "little infiltration" as a reference group, the local odds ratios for improvement were as follows:

```
  . cci    13    1    53   11

                                                    Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        13           1  |         14      0.9286
        Controls |        53          11  |         64      0.8281
-----------------+------------------------+----------------------
           Total |        66          12  |         78      0.8462
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         2.698113       |  .4015869           .  (Cornfield)
 Attr. frac. ex. |         .6293706       | -1.490121           .  (Cornfield)
 Attr. frac. pop |         .5844156       |
                 +-----------------------------------------------
                            chi2(1) =     0.89  Pr>chi2 = 0.3454
```

```
. cci    16    1    42   11

                                                    Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        16           1  |         17      0.9412
        Controls |        42          11  |         53      0.7925
-----------------+------------------------+----------------------
           Total |        58          12  |         70      0.8286
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         4.190476       |  .6288552           .  (Cornfield)
 Attr. frac. ex. |         .7613636       | -.5901912           .  (Cornfield)
 Attr. frac. pop |         .7165775       |
                 +-----------------------------------------------
                            chi2(1) =     2.00  Pr>chi2 = 0.1568
```

```
. cci    15    1    27   11
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        15           1  |         16      0.9375
        Controls |        27          11  |         38      0.7105
-----------------+------------------------+----------------------
           Total |        42          12  |         54      0.7778
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         6.111111       |  .8989672            .  (Cornfield)
 Attr. frac. ex. |         .8363636       | -.1123876            .  (Cornfield)
 Attr. frac. pop |         .7840909       |
                 +-----------------------------------------------
                               chi2(1) =     3.36  Pr>chi2 = 0.0670
```

```
. cci   7    1    11   11
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |         7           1  |          8      0.8750
        Controls |        11          11  |         22      0.5000
-----------------+------------------------+----------------------
           Total |        18          12  |         30      0.6000
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |              7         |  .9089896            .  (Cornfield)
 Attr. frac. ex. |         .8571429       | -.1001226            .  (Cornfield)
 Attr. frac. pop |              .75       |
                 +-----------------------------------------------
                               chi2(1) =     3.44  Pr>chi2 = 0.0637
```

Again, subjects with much infiltration are more likely to show an improvement than subjects with little infiltration. That is, the odds of improvement among "much infiltration" as compared to that of among "little infiltration" group was 2.7, 4.2, 6.1, and 7.0.

Ex 3.3-3 Testing the hypothesis

The null hypothesis of no difference is

$$H_0 : f_1 = f_2 = f$$

The estimate f_j use the formula shown above, for the variance of $\hat{f}$ we use

$$v\hat{a}r(\hat{f}_i) = \frac{\sum_j x_j^2 p_{ij} - \hat{f}_i^{\,2}}{n_{i+}}$$

The test statistic based on the Neyman chi-square is

$$Q = \frac{(\hat{f}_1 - \hat{f}_2)^2}{\left(v\hat{a}r(\hat{f}_1) + v\hat{a}r(\hat{f}_2)\right)}$$

which has a chi-square distribution with 1 df.
Followings are the example of calculating for the Neyman chi-square.

$v\hat{a}r(\ \hat{f}_1)$=$\frac{1}{144}$**{[9(.076) + 4(.188) + (.292) + (.076)] - .819^2}**
= 0.0079

$v\hat{a}r(\ \hat{f}_2)$**=0.0208**

The Neyman chi-square can be calculated as

$$\boldsymbol{Q} \quad = \quad \frac{(.819-1.269)^2}{(.0079+.0208)} \ = \ \mathbf{7.06}$$

Comparing this with chi-square distribution of 1 df leads us to reject *H*$_0$ and conclude that subjects with much infiltration show a greater degree of improvement (p-value = 0.007).

```
. disp chiprob(1, 7.06)
.0078824
```

Analyzed these data using the chi-square test for trend (that is, assuming that the column totals are fixed), the method gives χ^2_1 = 6.63 (p-value = 0.01 as shown below.

```
. tabodds row col
------------+-----------------------------------------------------------
        col |      cases     controls       odds      [95% Conf. Interval]
------------+-----------------------------------------------------------
          1 |         11            1   11.00000        1.42017  85.20081
          2 |         53           13    4.07692        2.22272   7.47791
          3 |         42           16    2.62500        1.47591   4.66874
          4 |         27           15    1.80000        0.95755   3.38365
          5 |         11            7    1.57143        0.60918   4.05364
------------+-----------------------------------------------------------
Test of homogeneity (equal odds): chi2(4)  =     6.85
                                  Pr>chi2  =   0.1443

Score test for trend of odds:     chi2(1)  =     6.63
                                  Pr>chi2  =   0.0100
```

Alternatively, the test for trend can be obtained by comparing the median (rank) score of tonsil enlargement between the two groups of degree of infiltration. The score of 1 to 5 can be best using Mann-Whitney-U-test. The command "ranksum" (see StataCorp., (1999), Volume 3: P-St, page 316-322)can handle this.

```
. ranksum col, by(row)
Two-sample Wilcoxon rank-sum (Mann-Whitney) test
     row |      obs    rank sum    expected
---------+---------------------------------
       0 |       52        5993        5122
       1 |      144       13313       14184
---------+---------------------------------
combined |      196       19306       19306

unadjusted variance   122928.00
adjustment for ties    -9209.14
                     ----------
adjusted variance     113718.86

Ho: col(row==0) = col(row==1)
             z =   2.583
    Prob > |z| =   0.0098
```

Since Z is equivalent to chi-square with one degree of freedom, thus given Z of 2.583, the equivalent χ_1^2 is 6.67, as computed below.

```
. disp 2.583^2
6.671889
```

This (χ_1^2 = 6.67) was slightly different from chi-square test for trend (χ_1^2= 6.63) obtained using the "tabodds" command as shown above. They lead to the same p-value of 0.01.

For the overall test for association (using Pearson's chi-square) is χ_{c-1}^2= 6.88 with 4 *df* (see the output in Ex.3.3-1 shown in the rectangular) which is non-significant (p-value = 0.14). These results are identical to those obtained using the second method described above if the Pearson chi-square is used (based on the variances of $\hat{f}_1$ and $\hat{f}_2$ calculated under H_0) instead of the Neyman chi-square as calculated here.

Test for departure from linear trend is 6.88 - 6.67 = 0.21. The degree of freedom is (C-1) - 1 = (5-1) - 1 = 3. The p-value is 0.976 as shown below.

```
. disp chiprob(3, 0.21)
.97595904
```

This suggested that almost all observed variation between the group of degree of infiltration were attributed to the linear trend in changes of degree of improvement.

Notes:

i) Ignoring the ordinality of the degree of improvement, the test suggested no sufficient evidence of the association (p-value = 0.14) whereas the test for trend (i.e., accounted for the ordering) suggested a strong evidence (p-value = 0.01). Thus it is necessary to consider the ordinal nature of the C variable.

ii) We can fail to reject H_0 of homogeneity but reject the hypothesis of no difference in degree of improvement. This will occur when one hypothesis is global with many *df* and the other is specific with few *df*. The consensus is that the more specific hypothesis is more sensitive and therefore more appropriate.

iii)Choice of scores is arbitrary - here we assumed that the levels of improvement were equally spaced. The Neyman chi-square will not change as long as this assumption is met regardless of the actual values of the scale. If the scale is not equally spaced the statistic will be affected. If the ordinal variable represents a continuous variable that has been categorized, then the midpoint of each interval defining the categories could be used as the scores.

Ex 3.3-4 Summary findings

Among 144 patients who had little infiltration, the proportion of marked improvement through to a worsening of infiltration were 7.6%, 18.8%, 29.2%, 36.8%, and 7.6% respectively. Whilst for the patients who had much infiltration, the corresponding proportion were 13.5%, 28.8%, 30.8%, 25.0%, and 1.9% respectively. The odds of improvement from worse to stationary, slightly, moderate, and marked improvement among "much infiltration" as compared to that of among "little infiltration" group was 2.7, 4.2, 6.1, and 7.0 respectively. The degree of improvement between the two groups was statistically significant (p-value = 0.01). Almost all observed variations between the group based on degree of infiltration were attributed to the linear trend in changes of degree of improvements. Overall chi-square test for association = 6.88 with 4 df, the chi-square-test for trend = 6.67 with 1 df, and thus the chi-square test for departure from linear trend = 0.21 with 3 df.

Chapter references

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

Altman, D.G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. 2nd edition. New York: John Willey & Sons.

Sribney, W. A. (1999). Comparison of different tests for trend. Stata Corporation, http://www.stata.com/support/faqs/stat/trend.html (connected at 12 December 1999).

StataCorp. (1999). *Stata statistical software: Release 6.0.* College Station. TX: Stata Corporation.

Exercise

1. Helmes and Fekken (1986) cited by Agresti (1990); page 72 reported the numbers of psychiatric patients by their diagnoses and by whether or not their treatment included drugs. The data are shown in the table.

Diagnosis	Drugs	No Drugs
Schizophrenia	105	8
Affective disorder	12	2
Neurosis	18	19
Personality	47	52
disorder	0	13
Special symptoms		

i) Test the hypothesis of independence between diagnosis and prescription of drugs.

ii) Comment on the appropriateness of the test you used.

iii) Summarise your findings.

2. Doll and Hill (1952) cited by Agresti (1990); page 31 presented data from a case-control study of lung cancer and tobacco smoking among patients in hospitals in several

English cities. The table compares male lung cancer patients with control patients having other diseases, according to the average number of cigarettes smoked daily over a ten-year period preceding the onset of the disease.

	Disease Group	
Av. No. cigarettes per day	**Lung Cancer Patients**	**Control Patients**
None	**7**	**61**
<5	**55**	**129**
5-14	**489**	**570**
15-24	**475**	**431**
25-49	**293**	**154**
50+	**38**	**12**

i) Is there an association between disease group and cigarette smoking?

ii) Perform a test of trend to determine whether or not lung cancer patients tend to smoke more than control patients.

iii) Calculate the odds ratio for each level of smoking using 'None' as the reference category. Comment on the results.

iv) Compute the odds ratios for each pair of adjacent levels of smoking. Comment on the pattern of association.

v) Comment on the choice of controls in this study. Is it likely to bias the results? Explain your answer.

Chapter 4

Analysis of R-by-C Tables

Chapter Objectives

After completing this chapter, readers should be able to:

- state the null hypothesis and statistical test appropriate to an R-by-C table in the case of i) row and column totals fixed; ii) row totals fixed; and iii) sample size only fixed;
- describe and interpret measures of association suitable for R-by-C tables when: i) the variables are nominal; and ii) the variables are ordinal;
- explore patterns of association in an R-by-C table using cell proportions, cell chi-square, and local odds ratios; and
- interpret the results from the analysis.

4.1 Introduction

An R-by-C Table refers to the contingency tables with more than two rows and two columns. The interpretation of the patterns of association is less clear and more detailed analysis may be necessary to decide where in the table any departures from independence arise. In addition, one or both variables may be ordered.

Below is a general form of the R-by-C Table.

Table 4.1 Notation of observed data

		Variable 2				
		1	2	...	c	
	1	n_{11}	n_{12}	...	n_{1c}	n_{1+}
	2	n_{21}	n_{22}	...	n_{2c}	n_{2+}
Variable 1	…	...	...	...	...	...
	r	n_{r1}	n_{r2}	...	n_{rc}	n_{r+}
		n_{+1}	n_{+2}	...	n_{+c}	n_{++}

Table 4.2 Notation of population proportions

		Variable 2				
		1	2	...	c	
	1	π_{11}	π_{12}	...	π_{1c}	π_{1+}
	2	π_{21}	π_{22}	...	π_{2c}	π_{2+}
Variable 1	…	...	...	...	...	...
	r	π_{r1}	π_{r2}	...	π_{rc}	π_{r+}
		π_{+1}	π_{+2}	...	π_{+c}	π_{++}

where n_{ij} are observed cell counts i = 1, ..., r and j = 1, ..., c.

[see details in Everitt (1977); page 16-21]

4.2 Measures of Association

A large number of measures of association for R-by-C tables have been proposed. Statistical packages print out a number of them. A section of Everitt (1977); page 56-66 describes some to them. Selvin (1995); page 273-288 provide an example and computer output using Stata. Followings are summaries of the most common use measure of associations.

4.2.1 Odds Ratios

Agresti (1990); page 18-19 described the odds ratio for R-by-C Tables quite comprehensive. Odds ratios are also useful for describing contingency tables larger than 2 × 2. Odds ratio for R-by-C tables can use each of the $\binom{r}{2}$ = *r*(*r* - 1)/2 pairs of rows in combination with each of the $\binom{c}{2}$ = *c*(*c* - 1)/2 pairs of columns. For rows *i* and *i'* and columns *j* and *j'*, the odds ratio $\pi_{ij}\pi_{i'j'} / \pi_{ij'}\ \pi_{i'j}$ uses four cells in a rectangular pattern. There are $\binom{r}{2}\binom{c}{2}$ odds ratios of this type.

However this set of odds ratios contains much redundant information. Consider the subset of (*r* - 1)(*c* - 1) 'local' odds ratios.

$$\psi_{ij} = \frac{\pi_{i,j}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}} \qquad \begin{array}{l} i = 1, \ldots, r-1 \\ j = 1, \ldots, c-1 \end{array}$$

These 'local' odds ratios use cells in adjacent rows and adjacent columns. These (*r*-1)(*c*-1) odds ratios determine all $\binom{r}{2}\binom{c}{2}$ odds ratios formal by pairs of rows and pairs of columns (see example below).

The construction for a minimal set of odds ratios is not unique. Another basic set is:

$$\psi_{ij} = \frac{\pi_{ij}\pi_{rc}}{\pi_{rj}\pi_{ic}} \qquad \begin{array}{l} i = 1,\dots,r-1 \\ j = 1,\dots,c-1 \end{array}$$

Each odds ratio uses the rectangular pattern of cells determined by rows *i* and *r* and columns *j* and *c*.

In summary for R-by-C tables, it is rarely possible to summarize association by a single number without some loss of information. However, summary indices can describe certain features of the association.

4.2.2 Summary measures of association

While there are more than one local odds ratios to indicate the degree of association for a R-by-C Table, a single measurement is needed to summarize as a global association. Some of these had been described in Agresti (1990); page 18-19, page 19-26. Selected measures of association which are commonly used, by each situation, are as follows:

4.2.2.1 One or both variables are nominal

i) Cramer's V

It is a transformation of chi-square statistics. It has <u>no probabilistic interpretation</u>. That is, we cannot express in words in terms of probability or errors in prediction (Everitt, 1977, page 63). For complete association, it may <u>not be equal to 1.0</u>. Thus it has a little useful. The formula and a computational example can be found in Selvin (1995); page 273.

ii) Lambda (λ)

It involves two estimate probabilities - one is the maximum probability of predicting that an observation belongs to a specific row and another is the maximum probability of predicting that an

observation belongs to a specific row given that the observation belongs to specific column. Everitt (1977); page 60-61 classify the lambda into two main categories - asymetric (λ_R and λ_C) and symmetric (λ) depending on that whether or not we have an explanatory and a dependent variable, i.e., row or column variable is given beforehand or either one, respectively. The value of λ is 0 if the row variable cannot be predicted from knowledge of the column variable. Thus the value of λ different from 0 indicates the degree of association. In other words, it is simply the proportional reduction in error. Formula and a work example is given by Selvin (1995); page 273-275.

4.2.2.2 Both of the variables are ordinal

i) Kendall's Tau statistics (τ)

It is a measure of correlation between two sets of rank data. This statistics have no obvious probabilistic interpretation. Its value may not ranged from -1 to 1. There are three type of Tau statistics - τ_a is not applicable for contingency table data; τ_b and τ_c may have the value ranged between -1 and 1 only in certain situation such as when the sample size is sufficiently large (Everitt, 1977, page 63).

ii) Gamma coefficient (γ)

It is a special case of rank correlation coefficient which has probabilistic interpretation. That is, its value ranged from -1 to 1 where 1 indicates a complete association that all data are on the main diagonal and -1 also indicates a complete association but that all data are on the other diagonal of the table. Value of γ near 0 indicates weak association (see Everitt, 1977, page 63). The γ is suitable for *qualitative ordinal variables* where the score is arbitrary such as

disease improvement, level of severity, etc. The coefficient is based on rank. Formula and a work example is given by Selvin (1995); page 276-278.

iii) Somer's d (d_{yx})

It is suitable for asymmetric situation where we have an explanatory variable and a dependent variable. This coefficient has similar interpretation to that of γ (see Everitt, 1977, page 63).

iv) Correlation coefficient (r)

This is Pearson's correlation coefficient. It has the properties similar to rank correlation but this not based on rank but on the meaningful numerical values. Thus it is suitable for *quantitative ordinal variable* such as age, blood pressure, etc. Formula and a work example is given by Selvin (1995); page 278-279.

Stata provides Crames'V, Gamma, and τ_b (see later in the example).

4.3 Test of Association

Followings are summaries of test statistics for R-by-C Tables. They lead to the same conclusion for large sample. For small sample, caution is needed in choosing the appropriate methods. Details can be found in Agresti (1990); page 47-49.

4.3.1 Both Sets of Margins Fixed

H_0 is the hypothesis of randomness
Distribution - multivariate hypergeometric

Test Statistic

- Large sample : $Q = \dfrac{(n_{++} - 1)}{n_{++}} \chi^2_p$

  where χ^2_p is the usual Pearson chi-square.

- Small sample : Freeman Halton Conditional Exact Test

4.3.2 Row Margins Fixed

H_0 is the hypothesis of row homogeneity

Distribution - product multinomial

Test Statistic - Pearson chi-square (χ^2_p) or Likelihood ratio test (G^2)

4.3.3 Sample Size Fixed

H_0 is the hypothesis of independence

Distribution - full multinomial

Test Statistic - Pearson chi-square (χ^2_p) or Likelihood ratio test (G^2)

Note that the three large sample models (Q, χ^2_p, and G^2)lead to the same expected cell frequencies and the same test statistics. The exact test can be found in Agresti (1991); page 59-67 and also the last chapter of this book.

Example 4.1

Hypothetical data of a study to determine the association between blood group and psychiatric disorder among 200 psychiatric patients at a hospital. The data are as follows:

Table 4.3 Number of psychiatric disorder patients by blood group - data for example 4.1

Psychiatric disorder	Blood group				Total
	A	B	AB	O	
Schizophrenia	7	25	20	28	80
Neurosis	12	20	16	10	58
Depressed	30	10	10	12	62
Total	49	55	46	50	200

Note that these data have no natural ordering although psychiatric disorder may be considered an ordinal variable. (The exercise at the end of this Chapter involves two ordinal variables.)

Ex 4.1-1 Describing the proportions

Assuming the column totals are fixed, the proportions of each type of psychiatric disorders for each group of blood group (shown below) suggest a rough meaningful magnitude and pattern of the association.

Table 4.4 Percentages of psychiatric disorder patients by blood group - from the data of example 4.1

Psychiatric disorder	Blood group			
	A	B	AB	O
Schizophrenia	8.8%	31.3%	25.0%	35.0%
Neurosis	20.7%	34.5%	27.6%	17.2%
Depressed	48.4%	16.1%	16.1%	19.4%
Total	24.5%	27.5%	23.0%	25.0%

The "tabi" command (see StataCorp., 1999, Volume 4: Su-Z, page 144-152) can be used to obtained these proportion as follows:

```
. tabi 7 25 20 28 \ 12 20 16 10 \ 30 10 10 12, row

           |                      col
       row |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |         7         25         20         28 |        80
           |      8.75      31.25      25.00      35.00 |    100.00
-----------+--------------------------------------------+----------
         2 |        12         20         16         10 |        58
           |     20.69      34.48      27.59      17.24 |    100.00
-----------+--------------------------------------------+----------
         3 |        30         10         10         12 |        62
           |     48.39      16.13      16.13      19.35 |    100.00
-----------+--------------------------------------------+----------
     Total |        49         55         46         50 |       200
           |     24.50      27.50      23.00      25.00 |    100.00
```

Cell chi-square were as follows:

Table 4.5 Cell chi-square of psychiatric disorder patients by blood group - from the data of example 4.1

Psychiatric disorder	**Blood group**			
	A	**B**	**AB**	**O**
Schizophrenia	***8.10***	**0.41**	**0.14**	**3.20**
Neurosis	**0.34**	**1.03**	**0.53**	**1.40**
Depressed	***14.44***	**2.92**	**1.27**	**0.79**

Comparing these with 1 degree of freedom, only the two cells (displayed in italic bold letters) are significant. This suggested that Blood group A with Schizophrenia or Depressed contribute greatly to the association between "blood group" and "psychiatric disorder".

Note : Example of calculating the cell chi-square for the first cell is $[(O- E)^2 / E] = \{[7- (49\times80/200)]^2\} / (49\times80/200)\} = 8.1$. The chi-square which greater than a critical value of 3.84 (i.e., χ^2_1 at $\alpha = 0.05$) is said to

be significant. That is, the observed frequency is different, beyond chance, from what would be expected if there was no association between the two variables.

Ex 4.1-2 Estimating measure of effect

Calculating "local" odds ratios for adjacent rows and columns for the 3-by-4 Table we can have 6 odds ratios altogether. It is not easy to interpret all these in words since they are almost another raw data. Thus odds ratios are not useful in this situation. However we will illustrate their calculations as follows:

$$OR_{11} = \frac{7\times 20}{12\times 25} = 0.47 \qquad OR_{12} = \frac{7\times 16}{12\times 20} = 0.47$$

$$OR_{13} = \frac{7\times 10}{12\times 28} = 0.21$$

$$OR_{21} = \frac{12\times 10}{30\times 20} = 0.20 \qquad OR_{22} = \frac{12\times 10}{30\times 16} = 0.25$$

$$OR_{23} = \frac{12\times 12}{30\times 10} = 0.48$$

Thus, in all case, those who have blood group A are less likely to get severe psychiatric disorder than those whose blood group is B, AB, or O.

Now consider a "single" summary measure of association. The following Stata command provides some of them.

```
. tabi 7 25 20 28 \ 12 20 16 10 \ 30 10 10 12, row all

           |                      col
       row |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |         7         25         20         28 |        80 
           |      8.75      31.25      25.00      35.00 |    100.00 
-----------+--------------------------------------------+----------
         2 |        12         20         16         10 |        58 
           |     20.69      34.48      27.59      17.24 |    100.00 
-----------+--------------------------------------------+----------
         3 |        30         10         10         12 |        62 
           |     48.39      16.13      16.13      19.35 |    100.00 
-----------+--------------------------------------------+----------
     Total |        49         55         46         50 |       200 
           |     24.50      27.50      23.00      25.00 |    100.00 

          Pearson chi2(6) =  34.5648   Pr = 0.000
 likelihood-ratio chi2(6) =  34.4582   Pr = 0.000
               Cramer's V =   0.2940
                    gamma =  -0.3707  ASE = 0.082
          Kendall's tau-b =  -0.2683  ASE = 0.061
```

From the above output, neither Gamma nor Kendall's tau-b can be used for this problem since these two measures of association regards the blood group and psychiatric disorder as continuous which is not appropriate. Only Cramer's V can be used. We can also use Lambda coefficient (λ) which did not provided by Stata. However it can be easily calculated.

The problem illustrates a situation where the row variable (i.e., Psychiatric disorder) can be predicted from knowledge of the column variable (i.e., Blood group). Therefore, the λ_c is an appropriate measure of association (see Selvin, 1995, page 274 for more details). This data gives λ_c = [(30+25+20+28)-80] / (200-80) = 0.19. The Cramer's V = 0.29. These are slightly different from 0 indicating a weak association.

Ex 4.1-3 Testing the hypothesis

Based on the above Stata output, the χ^2_6 = 34.56. Thus there is a statistically significantly association between blood group and psychiatric disorder (p-value < 0.001).

Ex 4.1-4 Summary findings

Those who have blood group A are less likely to get severe psychiatric disorder than those whose blood group is B, AB, or O (see the proportions and odds ratios shown earlier. There is a statistically significantly association between blood group and psychiatric disorder (p-value < 0.001). Blood group A with "Schizophrenia" or "Depressed" contributed greatly to the association. However the magnitude of such association is small. That is, knowing blood group can predict a little about whether or not people has psychiatric disorder (λ_c = 0.19). Cramer's V of 0.29 also suggested a weak association between the two variables.

Chapter references

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

Everitt, B.S. (1977). *The Analysis of Contingency Tables*. London: Chapman and Hall.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. 2nd edition. New York: John Willey & Sons.

Selvin S. (1995). *Practical biostatistical methods*. Belmont: Duxbury Press.

Senie, R.T., Rosen, P.P, Lesser, M.L., and Kinne, D.W. (1981). Breast self-examination and medical examination related to breast cancer stage. *Am J Public Health*. 71(6):583-90.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

Exercise

Senie et al (1981) studied factors related to breast self-examination (BSE) among 1216 women with breast cancer. The data for age group and frequency of BSE are given below.

Age	Frequency of BSE			Total
	Monthly	Occasional	Never	
< 45	91	90	51	232
45 - 59	150	200	155	505
60 +	109	198	172	479
Total	350	488	378	1216

i) Is frequency of BSE dependent of age?

ii) Describe the patterns of association between age and frequency of BSE.

Chapter 5

Analysis of Square Tables

Chapter Objectives

After completing this chapter, readers should be able to:

- **describe the properties of marginal homogeneity and symmetry for a K-by-K Table;**
- **identify appropriate tests for the hypotheses of marginal homogeneity and symmetry for a K-by-K Table;**
- **calculate a measure of agreement for two observers and test whether the observed agreement is better than that expected by chance for i) for a nominal classification scale and ii) for an ordinal classification scale;**
- **interpret the results from the analysis.**

Contents

5.1 Introduction

A square table is the contingency tables with more than two rows and two columns where the cell frequencies represent the number of pair for each combination of the independent variable. I can be referred to the K-by-K Table where K is the number of categories of the dependent variable. The data are obtained by that the same subjects are measured twice or a measurement is made on a paired of subject. Thus the number of rows and columns is automatically equal. It is a general form of the matched contingency tables. The matched 2-by-2 Table described in Chapter 2 is a special case of the square tables. In such case, the outcome is dichotomous. This Chapter involves polytomous outcome. Note that the R-by-C Table where R is equal to C such as a 3-by-3 Table as shown in the exercise of the previous Chapter cannot be considered a square table. The main distinction here is that the data is not from a matched study.

There are two main types of data for the square tables:

5.1.1 Observer agreement data - two observers each classify the same subjects using k- point categorical scale. Two questions are of interest :

i) Do the observers use the categories of the scale with the same frequency? This is *marginal homogeneity.*

ii) What is the extent of the *agreement* between the observers ?

5.1.2 Repeated measures data - patients classified on a k-point scale before and after treatment. The question of

interest is that whether there is improvement after treatment? So the data off the diagonal (indicating change) are concentrated in the appropriate triangle (lower left or upper right)? This is *symmetry*.

Below is a general form of the square Table.

Table 5.1 Notation of observed data

		Obsever 2				
		1	2	...	k	
	1	n_{11}	n_{12}	...	n_{1k}	n_{1+}
	2	n_{21}	n_{22}	...	n_{2k}	n_{2+}
Observer 1	…	...	...	...	...	...
	k	n_{k1}	n_{k2}	...	n_{kk}	n_{k+}
		n_{+1}	n_{+2}	...	n_{+k}	n_{++}

Table 5.2 Notation of population proportions

		Obsever 2				
		1	2	...	k	
	1	π_{11}	π_{12}	...	π_{1k}	π_{1+}
	2	π_{21}	π_{22}	...	π_{2k}	π_{2+}
Observer 1	…	...	...	...	...	...
	k	π_{k1}	π_{k2}	...	π_{kk}	π_{k+}
		π_{+1}	π_{+2}	...	π_{+k}	π_{++}

where n_{ij} are observed cell counts i = 1, ..., k and j = 1, ..., k.

5.2 Tests of Marginal Homogeneity and Symmetry

For the 2-by-2 Tables, the hypothesis of marginal homogeneity is

$$H_M : \pi_{i+} = \pi_{+i}, i = 1, 2.$$

Thus $\pi_{11} + \pi_{12} = \pi_{11} + \pi_{21}$. That is, $\pi_{12} = \pi_{21}$ and this is symmetry (H_S).

Similarly, $\pi_{21} + \pi_{22} = \pi_{12} + \pi_{22}$, ie $\pi_{21} = \pi_{12}$ which is the same as the above case. In other words, it lead to

$$H_S: \pi_{12} = \pi_{21}.$$

So $H_M \equiv H_S$. McNemar's chi-square test can be used to test this hypothesis

For the K-by-K Tables where K is greater than 2, marginal homogeneity can be addressed via symmetry and asymmetry. Total symmetry (or interchangeability) is given by

$$H_S : \pi_{ij} = \pi_{ji} \qquad \text{for } i, j = 1, ..., k$$

Homogeneity of marginal distributions is given by

$$H_M : \pi_{i+} = \pi_{+i} \qquad \text{for } i = 1, ..., k$$

Thus if there is symmetry, there has to be marginal homogeneity. But if there is marginal homogeneity, it is not necessary symmetry.

Symmetry refers to that the probability that an observation falls in the (i, j) cell of a square table is the same as the probability that it falls in the (j, i) cell of the table. It requires that the expected marginal total for any one row of the table, say the kth row, is the same as the expected marginal total for the corresponding kth column (Stasny and Bauer, 1990). The hypothesis of symmetry is thus very restrictive. A more useful concept is that of quasi-symmetry. The quasi-symmetry does

not require the equality of expected row and column totals. It is a useful method for studying marginal homogeneity. Symmetry is equivalent to quasi symmetry and marginal homogeneity holding simultaneously. The interpretation of the quasi-symmetry is that there is symmetry in the observed data once one has taken into account the difference in the row and column totals (Stasny and Bauer, 1990). Agresti (1990); page 347-365 provide details on the test for Marginal Homogeneity and Symmetry.

5.3 Measuring Agreement

Details of measuring agreement described in the whole Chapter 13 of Fleiss (1981); page 212-236 is recommended. Another approaches were given by Agresti (1990); page 365-373. Altman (1991); page 403-409 provided a practical guide of the analysis, reporting, and interpretation. Followings is summary of important issues.

When two (or more) observers are asked to allocate subjects to two or more categories, we may be interested in the level of agreement (or concordance) between the observers. By agreement we mean the extent to which the observers both allocated a subject to the *same* category. We are not interested in association (the degree to which one observer's ratings predict or are associated with the other observers) but the extent to which they are the same.

5.3.1 Nominal Scale : Responses on the main diagonal indicate agreement. If the observers act independently and allocate categories at random according to their marginal distributions, then the amount of agreement that 'can occur by chance'. We need a 'chance corrected measure of agreement' called "Kappa" statistics. The

kappa statistic measure of agreement is scaled to be 0 when the amount of agreement is what would be expected to be observed by chance and 1 when there is perfect agreement.

5.3.2 Ordinal Data : Kappa does not give any partial credit for disagreements at different levels. For example, disagreements that involve only one category compared to disagreements involving distant categories. Thus we need a weighted kappa.

For intermediate values, Landis and Kock (1977); page 165 suggest the following interpretations of agreement:

Below 0.0	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

Example 5.1

We use a hypothetical data set adapted from Selvin (1995); page 274 to illustrate the concepts of analysis the repeated measurement data. We suppose the data is from a study to determine effect of an intervention on patients' satisfaction. One hundred and fifty subjects were asked their level of satisfaction (1= dissatisfy, 2=neutral, and 3=satisfy) before and after implementation of the intervention. The data are shown below.

Table 5.3 Number of patients for each level of satisfaction before and after the intervention - data for example 5.1

Before	After			Total
	dissatisfy	neutral	satisfy	
dissatisfy	7	25	20	52
neutral	12	20	16	48
satisfy	30	10	10	50
Total	49	55	46	150

Ex 5.1-1 Describing the proportions

About one third of patients rated themselves similarly in dissatisfy, neutral, and satisfy to the standard procedure before the intervention (i.e., 34.7%, 32.0%, and 33.3% respectively) and after the intervention (i.e., 32.7%, 36.7%, and 30.7% respectively). Note than for "tabi" command see StataCorp. (1999), Volume 4: Su-Z, page 144-152.

```
. tabi 7 25 20 \ 12 20 16 \ 30 10 10, row col

           |               col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |         7         25         20 |        52
           |     13.46      48.08      38.46 |    100.00
           |     14.29      45.45      43.48 |     34.67
-----------+---------------------------------+----------
         2 |        12         20         16 |        48
           |     25.00      41.67      33.33 |    100.00
           |     24.49      36.36      34.78 |     32.00
-----------+---------------------------------+----------
         3 |        30         10         10 |        50
           |     60.00      20.00      20.00 |    100.00
           |     61.22      18.18      21.74 |     33.33
-----------+---------------------------------+----------
     Total |        49         55         46 |       150
           |     32.67      36.67      30.67 |    100.00
           |    100.00     100.00     100.00 |    100.00
```

Among a total of 150 patients, 25 (16.7%) change from the dissatisfy to the neutral response while there were only 12 (8.0%) changes in the opposite direction. For the change between dissatisfy and satisfy, 20 (13.3%) change from the dissatisfy to the satisfy response whereas there were 30 (20.0%) changes in the opposite direction. For the change between satisfy and neutral, 16 (10.7%) change from the neutral to the satisfy response whereas there were 10 (6.7%) changes in the opposite direction. If we consider the change between dissatisfy and satisfy the most important (i.e., more weight) we conclude that the intervention tend to reduce the patients' satisfaction.

Ex 5.1-2 Testing the hypothesis

There was a significant change in patients' satisfaction (symmetry test chi-square (3df) = 7.95; p-value = 0.047). This result is from asymptotic (i.e., large sample) test which is the same as that from the exact test. In a small sample size we need to quote the exact test results.

As indicated by the black arrow shown below, cell n_{12} and n_{21} contribute most to the symmetry chi-square. These correspond to changes between the dissatisfy and neutral categories. That is, among a total of 150 subjects, 25 (16.7%) change from the dissatisfy to the neutral response while there were only 12 (8.0%) changes in the opposite direction.

```
. tabi 7 25 20 \ 12 20 16 \ 30 10 10, replace

           |                col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |         7         25         20 |        52 
         2 |        12         20         16 |        48 
         3 |        30         10         10 |        50 
-----------+---------------------------------+----------
     Total |        49         55         46 |       150 

          Pearson chi2(4) =  27.1285   Pr = 0.000
```

```
. list

          row       col           pop
  1.        1         1             7
  2.        1         2            25
  3.        1         3            20
  4.        2         1            12
  5.        2         2            20
  6.        2         3            16
  7.        3         1            30
  8.        3         2            10
  9.        3         3            10
```

```
. symmetry row col [freq=pop], contrib mh trend exact

----------+--------------------------
          |            col
      row |   1      2      3    Total
----------+--------------------------
        1 |    7     25     20     52
        2 |   12     20     16     48
        3 |   30     10     10     50
          |
    Total |   49     55     46    150
----------+--------------------------

                   Contribution
                    to symmetry
   Cells            chi-squared
--------------    --------------
 n1_2 & n2_1          4.5676
 n1_3 & n3_1          2.0000
 n2_3 & n3_2          1.3846

                                        Chi-Squared    df      Prob>chi2
------------------------------------+--------------------------------
Symmetry (asymptotic)               |       7.95       3        0.0470
Marginal homogeneity (Stuart-Maxwell) |     0.80       2        0.6714
Marginal homogeneity (Bickenboller)   |     0.73       2        0.6937
Marginal homogeneity (no diagonals)   |     0.73       2        0.6949
------------------------------------+--------------------------------
Linear trend in the (log) RR        |       0.00       1        0.9508
------------------------------------+--------------------------------
Symmetry (exact significance probability)                       0.0473
```

Note that for "symmetry" command see StataCorp., (1999), Volume 4: Su-Z, page 112-119.

Ex 5.1-3 Summary findings

On average, about one third of patients rated themselves similarly in dissatisfy, neutral, and satisfy to the standard

procedure before the intervention (i.e., 34.7%, 32.0%, and 33.3% respectively) and after the intervention (i.e., 32.7%, 36.7%, and 30.7% respectively). Among a total of 150 patients, 25 (16.7%) change from the dissatisfy to the neutral response while there were only 12 (8.0%) changes in the opposite direction. For the change between dissatisfy and satisfy, 20 (13.3%) change from the dissatisfy to the satisfy response whereas there were 30 (20.0%) changes in the opposite direction. For the change between satisfy and neutral, 16 (10.7%) change from the neutral to the satisfy response whereas there were 10 (6.7%) changes in the opposite direction. There was a significant change in patients' satisfaction (symmetry test chi-square (3df) = 7.95; p-value = 0.047). However the largest contribution to the symmetry chi-square test was the changes between the dissatisfy and neutral. If we consider the change between dissatisfy and satisfy the most important (i.e., more weight) we conclude that the intervention tend to reduce the patients' satisfaction.

Example 5.2

We use a hypothetical data by supposing that the data is from a study to determine how close the results of the two laboratory technicians in classifying type of a pathogen. One hundred and fifty specimens were examined independently by each technician and classified it into the three type - A, B, or C. The data are shown below.

Table 5.4 Number of specimen in each type of classifications by two laboratory technicians - data for example 5.2

Technician 1	Technician 2			Total
	A	B	C	
A	30	20	10	60
B	12	25	16	53
C	7	10	20	37
Total	49	55	46	150

Ex 5.2-1 Describing the proportions

Technician 1 tended to classify the pathogen to type A (40.0%) and B (35.3%) more than type C (24.7%) while technician 2 classified the pathogen to the three types similarly, i.e., A (32.7%), B (36.7%) and C (30.7%).

```
. tabi 30 20 10 \ 12 25 16 \ 7 10 20, row col
```

```
           |               col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        30         20         10 |        60 
           |     50.00      33.33      16.67 |    100.00 
           |     61.22      36.36      21.74 |     40.00 
-----------+---------------------------------+----------
         2 |        12         25         16 |        53 
           |     22.64      47.17      30.19 |    100.00 
           |     24.49      45.45      34.78 |     35.33 
-----------+---------------------------------+----------
         3 |         7         10         20 |        37 
           |     18.92      27.03      54.05 |    100.00 
           |     14.29      18.18      43.48 |     24.67 
-----------+---------------------------------+----------
     Total |        49         55         46 |       150 
           |     32.67      36.67      30.67 |    100.00 
           |    100.00     100.00     100.00 |    100.00 
```

Ex 5.2-2 Estimating measure of effect

Observed agreement:

P_0 $= \sum_{i=1}^{k} n_{ii} / n_{++}$

= (30 + 25 + 20) / 150

= 0.5

Chance-expected agreement:

P_e $= \sum_{i=1}^{k} n_{+i} n_{i+} / n_{++}^{2}$

= [(49×60) + (55×53) + (46×37)] / 150^2

= 0.336

Chance-corrected agreement or kappa:

Kappa $= \frac{P_0 - P_e}{1 - P_e}$

= (0.50 - 0.33) / (1 - 0.33)

= 0.247

Stadard error of kappa :

SE(Kappa) $= \sqrt{\frac{P_0(1-P_0)}{n_{++}(1-P_e)^2}}$

$$= \sqrt{\frac{0.5(1-0.5)}{150(1-0.336)^2}}$$

$$= \mathbf{0.0615}$$

Therefore 95%CI (Kappa) = 0.247 ± (1.96×0.0615)

= 0.126 to 0.308

Tips: We can use Stata as a hand calculator for calculating the above statistics as follows:

```
. disp (30+25+20)/150
.5


. disp ((49*60)+(55*53)+(46*37)) / (150^2)
.33586667


. disp (0.50 - 0.336) / (1 - 0.336)
.24698795


. disp sqrt( (0.5*(1-0.5)) / (150*((1-0.336)^2)) )
.06148318


. disp 0.247 - 1.96*0.0615,  0.247 + 0.0615
.12646 .3085
```

Thus kappa is 0.25 (95%CI: 0.13 to 0.31). We conclude that the level of agreement achieved by the technicians is just fair (see the below outputs kappa statistics within the ovals).

Stata commands:
To obtain a data file, we type

```
. tabi 30 20 10 \ 12 25 16 \ 7 10 20, replace

           |                  col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        30         20         10 |        60 
         2 |        12         25         16 |        53 
         3 |         7         10         20 |        37 
-----------+---------------------------------+----------
     Total |        49         55         46 |       150 

          Pearson chi2(4) =  22.4418   Pr = 0.000
```

Then use "kap" command (see StataCorp., 1999, Volume 2: H-O, page 132-143) to estimate kappa statistics (in oval). The test statistics also provided (in rectangular and were discussed in the next section).

```
. kap row col [freq=pop], tab

           |                  col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        30         20         10 |        60 
         2 |        12         25         16 |        53 
         3 |         7         10         20 |        37 
-----------+---------------------------------+----------
     Total |        49         55         46 |       150 

             Expected
Agreement   Agreement     Kappa          Z        Pr>Z
-----------------------------------------------------
  50.00%      33.59%     0.2471       4.30      0.0000
```

Only the kappa statistics estimated by the above command is needed for the present study since the pathogen classification is nominal. For illustration purpose, if the scale A, B, and C are considered to be ordinal, we could estimate weighted kappa as well. Several type of weights can be applied. The following two commands are weight assigned automatically by Stata (see StataCorp, 1999; page 132-143 of Volumn 2 : H-O).

```
. kap row col [freq=pop], wgt(w)

Ratings weighted by:
   1.0000   0.5000   0.0000
   0.5000   1.0000   0.5000
   0.0000   0.5000   1.0000
```

```
             Expected
Agreement   Agreement     Kappa         Z           Pr>Z
  69.33%      56.63%     0.2929       4.65        0.0000
```

```
. kap row col [freq=pop], wgt(w2)

Ratings weighted by:
   1.0000   0.7500   0.0000
   0.7500   1.0000   0.7500
   0.0000   0.7500   1.0000

             Expected
Agreement   Agreement     Kappa         Z           Pr>Z
  79.00%      68.15%     0.3406       4.23        0.0000
```

Weight can be specify arbitrarily, we can define our own weight as follows:

```
. kapwgt weight 1 \ .8 1 \ 0 .8 1

. kap row col [freq=pop], wgt(weight)

Ratings weighted by:
   1.0000   0.8000   0.0000
   0.8000   1.0000   0.8000
   0.0000   0.8000   1.0000

             Expected
Agreement   Agreement     Kappa         Z           Pr>Z
  80.93%      70.46%     0.3546       4.08        0.0000
```

Ex 5.2-3 Testing the hypothesis

Taken the variance of the kappa calculated later in section Ex5.2-2, Z-test = 0.247 / 0.0615 = 4.02 corresponding to p-value = 0.00003

We can use Stata as a hand calculator for calculating the above statistics as follows:

```
. disp 0.247 / 0.0615
4.0162602

. disp 1-normprob(4.02)
.0000291
```

Thus we reject Ho and conclude that the level of agreement achieved by the technicians is statistically significantly better than that expected by chance (see also the above outputs p-values within the rectangular).

Ex 5.2-4 Summary findings

Technician 1 tended to classify the pathogen to type A (40.0%) and B (35.3%) more than type C (24.7%) while technician 2 classified the pathogen to the three types similarly, i.e., A (32.7%), B (36.7%) and C (30.7%). Kappa is 0.25 (95%CI: 0.13 to 0.31) suggesting the level of agreement achieved by the technicians is just fair. This level of agreement is statistically significantly better than that expected by chance (p-value < 0.001).

Chapter references

Agresti, A. (1990). ***Categorical data analysis*****. New York: John Wiley & Sons.**

Altman, D.G. (1991). ***Practical statistics for medical research*****. London: Chapman and Hall.**

Everitt, B.S. (1977). ***The Analysis of Contingency Tables*****. London: Chapman and Hall.**

Fleiss, J.L. (1981). ***Statistical methods for rates and proportions*****. 2nd edition. New York: John Willey & Sons.**

Landis, J.R. and Cock, G.G. (1977). The measurement of observer agreement for categorical data. ***Biometrics*****. 33:159-174.**

Stasny, E.A. and Bauer, H.R. (1990). Symmetry and quasi-symmetry: an example in modeling pairs of sounds from children's early speech. ***Stat. Med*****. 9:1143-1155.**

StataCorp. (1999). ***Stata statistical software: Release 6.0*****. College Station. TX: Stata Corporation.**

Exercise

Two anesthetists independently classified each of 45 patients as to their suitability for general anaesthetic. They used a 3-point categorical scale ranging from 1=entirely suitable to 3=unsuitable. The results are summarised in the following table :

	Technician 2			
Technician 1	**A**	**B**	**C**	**Total**
A	**15**	**3**	**0**	**18**
B	**1**	**12**	**8**	**21**
C	**0**	**0**	**6**	**6**
Total	**16**	**15**	**14**	**45**

a) Do the anaesthetists tend to use the categories of the scale in the same way?

b) Do the anaesthetists tend to agree on their classification?

c) Comment on the reliability of the classification scale they are using.

Chapter 6

Logistic Regression

Chapter Objectives

After completing this chapter, readers should be able to:

- **describe methods for dealing with effects from extraneous factors;**
- **describe statistical modeling approach for dealing with effects from extraneous factors;**
- **describe the logistic regression model;**
- **interpret the coefficients for the logistic regression model and calculate odds ratios and confidence intervals corresponding to independent variables included in the model;**
- **fit logistic regression models appropriate for situations where confounding or interactions are present;**
- **describe and perform appropriate model-fitting strategies;**
- **describe, perform, and interpret goodness-of-fit tests and diagnostics; and**
- **interpret the results from the analysis.**

Contents

6.1 Introduction

So far we have discussed the analytical methods concerning an outcome and one independent variable at a time. Most outcomes in medical science are usually caused by several factors. There might be also an inter-correlated among those factors. In quantifying the magnitude of association between a factor of interest and the outcome, researchers need to consider effects of the other “extraneous” factors. To do this, several methods had been proposed. This Chapter focused on a method called "logistic regression". Summaries of some other commonly used methods were also provided with a reference for advance readers. The first section of this chapter provided a brief method for dealing with effects of other variables. Then summary of key concepts of logistic regression was presented. The main part of the chapter was to demonstrate useful practical steps for such approach via an example. To do this we need to repeat some methods discussed in Chapter 2, readers are encourage to see section "stratified analysis" in that Chapter for more details. Note 2 at the end of this Chapter provided an example of how similar the results from stratified analysis compared to logistic regression.

6.2 Overview of methods for dealing with effects from extraneous factors

6.2.1 Controlling for effects of extraneous factors in the design stage

i) Randomization

- **A process for assigning patients to a test or control treatment that is free of selection bias (Meinert and Tonacia, 1986, page 90).**

- Applicable only in experimental study.
- Characteristics of the study subjects that are not measured or cannot be measured that could be predictive of the outcome of interest are randomly distributed across groups being compared. Thus they should be as similar as possible regarding factors that could affect the outcome - the potential confounders.
- Simple randomization may not guarantee comparability. Stratified block randomization could be used to assure comparability for a few variables, but the distribution with regards to others must be left to chance (Meinert and Tonacia, 1986, page 91 - 96). This is sometimes called "pre-randomization stratification whereas the stratified analysis is a post-randomization stratification".
- Although this has a major role in the design stage, ignoring method of randomization in the analysis could lead to an inefficient. Fleiss (1986) provided a good guideline for both the design and the appropriate analysis. For example, stratified block randomization should be analyzed using methods that accounted for stratification as described on pages 149 - 185.

ii) Restriction in the study design

- Specify narrow ranges of values for one or more extraneous variables as criteria for admissibility into the study - e.g., study only males or to ages between 40 to 50. "
- A homogeneous study group will provide a poor basis of generalization of the study results (Rothman and Greenland, 1998, page 145).

iii) Matching

- Each subject in one group was matched to one or more

specific subjects in the other group, so that all members of a pair or set are similar on the extraneous factors that could effect the outcome - e.g., match each male case to a male non-case.

- **The idea is to obtain an identical match on all variables except for the risk factor under investigation.**
- **In the analysis each matched set of subjects is regarded as a stratum and the same methods as for the analysis of stratified data can be applied.**
- **Might not practical if there were several extraneous variables to be controlled for.**

6.2.2 Controlling for extraneous factors in the analysis stage

i) Post hoc restriction or matching - rarely used, since usually involves discarding data

ii) Subgroup analysis

- **The investigator looks specifically at the intervention-control comparison within one or more particular subgroup rather than the overall comparison (Friedman, Furberg, and DeMets, 1996, page 304 - 306).**
- **It is used to answer the questions of this kind: "Among which group of the participants is the intervention most beneficial or harmful?".**
- **It is dangerous if subgrouping based on any outcome variable. Only baseline factors are appropriate for use in defining subgroups. (Meinert and Tonacia, 1986, page 194 provided a good example.)**
- **Generalization of findings could be limited.**
- **Data can apparently show one thing when they are in aggregate and show something quite different when**

they are disaggregated. This phenomenon is sometimes called Simpson's Paradox.

iii) Stratified analysis (Details has been discussed in Chapter 2)

- Involved pooling of the information over all strata if there is no interaction.
- Kleinbaum, Kupper, and Morgenstern (1986), page 321, suggested that stratification become worthwhile under the three following conditions: i) There are sufficient number in all strata, ii) An appropriate choice of control variables can be made., iii) An appropriate categorization scheme for each variable can be identified (ie., categories are meaningful and there is no residual confounding).
- Advantages: i) easy to understand and easy to interpret, ii) direct and logical strategies, iii) computationally simple, not require sophisticated software, iv) analyst usually sees intermediate data: e.g., stratum-specific effects of exposure, and iv) minimum assumption required.
- Limitations: i) difficult to deal with multiple potential confounders simultaneously: some strata may drop out of the analysis because a row or column total is zero, and ii) requires that all potential confounders be treated as categorical (discrete) variables, even if they are intrinsically continuous (e.g., age). If categories are too wide, can have residual confounding. If categories are too numerous, strata can again be lost from the analysis altogether, face problem of deciding how to form the categories, fairly simple dose-response relationship may be obscured.

iv) Multivariable analysis (This was discussed in the next section on "statistical modeling approach")

- In many cases, the methods at the design stage are impractical. It is sometime impossible in the situation where there were several prognostic factors for the outcome of interest. In this case, the methods at the analysis stage are preferred.

These methods have both advantages and disadvantages application. Summary of this issue was given by Kleinbaum, Kupper, and Morgenstern (1986), page 317.

6.3 Statistical modeling approach for dealing with effects from extraneous factors

This allows investigators control effects of several extraneous factors simultaneously at the same time. Ignoring effects of some important factors could lead to serious bias and misleading conclusion. We call this type of bias the "Simpson's paradox". Appleton, French, and Vanderpump (1996) demonstrated a very convincing example for this bias. Followings are some selected multivariable data analysis for dealing with effects from extraneous factors.

i) Logistic regression (details provided in the next section)

ii) Conditional logistic regression (the above discussions concerned unconditional logistic regression).

  - Used when study design involved matching of individuals into pairs or small sets or a study involved small sample size (Kleinbaum, 1994, page 105-108).

  - Kleinbaum, 1994, page 108 suggests a rule of thumb where the unconditional questionable and the conditional one preferred: 10 to 15 confounders and 10 to 15 product terms (We will discuss about the product term in the example at the next section).

- **Modeling strategies and interpretation of results are similar to those for unconditional logistic regression. Similar to the ordinary logistic regression, it yields measure of association as *"odds ratio"* associated with a particular explanatory variable of interest.**
- **An excellent introductory for conditional logistic regression was given by Kleinbaum, 1994, page 228-242. A simple and practical analysis was given by Rabe-Hesketh and Everitt (1998); page 145 - 147.**
- **Stata command for data analysis using this method is "clogit" (see StataCorp., 1999, Volume 1: A-G, page 201-261).**

iii) Cox's regression, or survival analysis under the proportional hazards model.

- **Used for studies where time to onset of the outcome is of interest. For example, the outcome is not "dead or alive" but "they can survive for how long". In this case, not all subjects can be followed until event occurred - some lost to follow-up, missing, withdrawals, or dead due to other cause. Survival analysis can handle this efficiently, and thus it is more efficient than logistic regression in this situation.**
- **Suitable with cohort studies (or randomized trials) using the "hazard rate" (roughly, incidence density for an individual) as the outcome.**
- **Also yields estimate of *"Hazard Ratio or Incidence Rate Ratio*", adjusted for potential confounding factors.**
- **Modeling strategies and interpretation of results are similar to those for logistic regression.**
- **A self-learning text by Kleinbaum (1996) provided a very good introductory.**

- **Stata commands for data analysis using this method are the "stset" command (see StataCorp., 1999, Volume 3: P-St, page 491-530) followed by "cox" (see StataCorp., 1999, Volume 1: A-G, page 264-269).**

iv) Multinomial logistic regression

- **An extension of logistic regression to accommodate polytomous (i.e.., more than two categories) outcome and the outcome have no nature ordering. For example, a study to determine factors affecting choice of health seeking behavior taken the value of "self treatment", "private clinic", and "public health care provider".**
- **One can estimate the *"relative risk ratio"* associated with a particular explanatory variable of interest from the model (StataCorp, 1999, page 403).**
- **Modeling strategies and interpretation of results are more difficult than those for ordinary logistic regression since there are multiple equation. Some investigators dichotomize the outcome so that ordinary logistic regression can be used. This practice might be acceptable provided that lost of information by such dichotomization is not obvious.**
- **For the introduction, see Hosmer and Lemeshow (1989); page 216 - 238.**
- **Stata command for data analysis using this method is "mlogit" (see StataCorp., 1999, Volume 2: H-O, page 379-412).**

v) Ordered logistic regression or proportional odds model

- **An extension of logistic regression as described above to accommodate ordinal (three or more categories that can be ranked) outcome. For** example, a study to determine factors affecting clinical outcome taken the value of "poor", "fair", "average", "good", and "excellent".

- The model yields *"probability"* of a subject having an outcome, associated with a particular explanatory variable of interest.
- Modeling strategies are similar to those for multinomial logistic regression but provide a mean to exploit the ordering information.
- Basic concepts and an example of data analysis can be found at Rabe-Hesketh and Everitt (1998); page 79 - 90.
- Stata command for data analysis using this method is "ologit" (see StataCorp., 1999, Volume 2: H-O, page 473-481).

vi) Poisson regression

- A regression model for a "Poisson" count outcome, i.e., a count of the number of occurrences of an event of interest, such as the number of case of a disease that occurred over a given follow-up time period.
- The natural measure of association is a *"rate ratio"* associated with a particular explanatory variable of interest.
- It expressed the log outcome (e.g., disease) rate as a linear function of a set of explanatory variables, the same principle as Log-linear model (see Agresti, 1990, page 130 - 152 and page 210 - 250).
- Modeling strategies and interpretation of results are similar to those for logistic regression.
- Details of this method was exclusively discussed in Kleinbaum, et al. (1998); page 689 - 709.

- Required some assumption about the distribution that need to be hold. An alternative methods if the data is severely sparse is the negative binomial regression (see StataCorp, 1999)
- Stata command for data analysis using this method is "poisson" (see StataCorp., 1999, Volume 3: P-St, page 25-34).

vii) Generalized linear model using Generalize Estimating Equations (GEEs)

- A statistical modeling suitable for any type of outcome with repeated measurements or other type of correlated data. It was proposed by Liang and Zegar (1986).
- Modeling strategies and interpretation of results are similar to those for logistic regression.
- The natural measure of association depend on type of the outcome and study design such as odds ratio or relative risk.
- An excellent example is given by Rabe-Hesketh and Everitt (1998); page 119 - 136.
- Stata commands for data analysis using this method are series of commands under the "xt" group of command (see StataCorp., 1999, Volume 4: Su-Z, page 317-359).

6.4 Logistics regression

The logistic regression model has become the standard method of analysis for the situation in which the relationship between a response (outcome) variable and one or more explanatory (predictors or covariates) variables is of interest and the outcome variable is categorical (taking on two or more possible values).

The goal of the analysis is to find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between the outcome variable and a set of independent variables.

For fuller details, a self-learning text by Kleinbaum (1994) is highly recommended. Followings are some important concepts.

Denoted "P" be probability of developing disease in a given individual (i.e, risk) and "x_1, x_2, ..., x_k" are several characteristics of an individual (e.g., gender - whether male or female, exposure - smoker or non-smokers, etc.) We can write

$$P = f(x_1, x_2, ..., x_k) \qquad (1)$$

That is, P is a function of the characteristics x_1, x_2, ..., x_k. But what is the nature of the function f( )? Lets take a simple linear model.

$$P = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \qquad (2)$$

where a, b_1, b_2,..., b_k are coefficients whose values are to be estimated from the data. By such estimation, we say "fit the model" to the data. There are several methods for the estimation procedure, almost all methods cannot be done without computer.

Each "b" coefficient represents the size of the effect of the corresponding "x" variable. It represents the change in "P" associated with a one-unit change in the corresponding "x".

The value "a" is a fitted constant (intercept), also estimated from the data, representing "P" for a person with $x_1 = x_2 = ... = x_k = 0$.

For some individuals, the right side of equation (2) may evaluate to less than 0, or to greater than 1. It is suitable for continuous

outcome. But for categorical outcome such as disease or non-disease, it needs to be between 0 and 1 so that it can be interpret as disease risks. That's the value of "P". The logistic model accomplishes this purpose.

$$P = \frac{1}{1+e^{(a+b_1x_1+...+b_kx_k)}} \qquad \textbf{(3)}$$

By equation (3), any values for a, b_1... b_k and x_1 ... x_k will yield a value of "P" between 0 and 1: i.e., a legitimate numerical value for disease risk. Some algebra shows that equation (3) can be solved for making "P" more interpretable as:

$$\log\left(\frac{P}{1-P}\right) = a + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

Where $\left(\frac{P}{1-P}\right)$ can be seen to be the <u>odds</u> of developing disease, and log$\left(\frac{P}{1-P}\right)$ is the <u>log odds</u> of developing disease, or the <u>logit</u> of P. (All logarithms are taken to the base e: i.e., natural logs.)

The model-fitting method then chooses the values for a, b_1 ... b_k that maximize agreement between the predicted value of P and the observed disease status of each subject.

One of the most useful properties of the logistic model is the interpretability of the b-coefficients. Say we are mainly interested in the characteristic x_1 (e.g., smoking), coded as follows:

x_1 = 1 if exposed, and
x_2 = 0 if not exposed

but that we must also consider another characteristic, x_2 (e.g., age) as a potential confounder. The logistic model is

$$\log\left(\frac{P}{1-P}\right) = a + b_1 x_1 + b_2 x_2$$

The effect of exposure to x_1, controlling for the effect of x_2, can be assessed by comparing disease risk in two persons who have different values of x_1 but the same value of x_2.

Denote "Pe" be the disease risk in the <u>e</u>xposed person and "Pu" be the disease risk in the <u>u</u>nexposed person. We get the log odds of exposed as:

$$\log\left(\frac{P_e}{1-P_e}\right) = a + b_1(1) + b_2 x_2 \qquad (4)$$

$$= a + b_1 + b_2 x_2$$

and we get the log odds of unexposed as:

$$\log\left(\frac{P_u}{1-P_u}\right) = a + b_1(0) + b_2 x_2 \qquad (5)$$

$$= a + 0 + b_2 x_2$$

Subtracting equation (5) from equation (4) we get :

$$\log\left(\frac{P_e}{1-P_e}\right) - \log\left(\frac{P_u}{1-P_u}\right) = b_1$$

Because of the properties of logarithms, this is the same as :

$$\log\left[\frac{P_e/(1-P_e)}{P_u/(1-P_u)}\right] = \log(\mathrm{OR}) = b_1$$

where OR= odds ratio. It is the odds of exposed divided by the odds of unexposed. By taking antilogs of both sides,

$$e^{\log(\mathrm{OR})} = \mathrm{OR} = e^{(b_1)} \qquad (6)$$

Hence the adjusted OR for the effect of exposure to x_1 on disease risk, controlling for the effect of x_2, is simply $e^{(b1)}$. This is sometimes also denoted exp(b_1), meaning "e to the b_1 power." One can also simply take e = 2.71828 and thus OR = $2.71828^{(b1)}$.

Equation (6) would also hold if there had been an arbitrary number of additional x's (e.g., x_3, x_4..., x_k) whose value had been the same for the two individuals being compared. Kleinbaum (1994) termed this "unspecified but fixed". Thus we call this OR the “adjusted OR” or “OR adjusted for <X>” where <X> is/are extraneous factors that we want to control for its/their effect(s).

If x_1 had been a continuous variable, then exp(b_1) would represent the adjusted OR for a <u>one-unit</u> change in x_1- e.g., a one-year increase in age.

From the model, one can test the statistical significance of each x-variable's contribution to the overall model, by determining whether the corresponding b-coefficient is statistically significantly different from zero. Confidence intervals can be obtained for the adjusted OR's, based on confidence limits for the corresponding b-coefficients, which the model-fitting method yields automatically. The extent to which x_2, say, confounds the association between x_2 and disease risk can be assessed by comparing the OR’s for x_1 in two models: one which includes x_2, and one which omits x_2. If these OR’s are similar, then x_2 is not an important confounder. This follows the standard practice of inferring whether confounding is present by comparing crude and adjusted measures of effect in stratified analysis. Test for modification of the effect of x_1 by x_2. Briefly, this is done by : i) creating a new variable, x_3 , as the product of x_1 and $x_{2;}$ ii) fitting a model which contains x_1, x_2, <u>and</u> x_3; iii) determining the size and statistical significance of the coefficient b_3, which reflects the magnitude of effect modification. The “pattern” of the relationships between an exposure and disease risk, by comparing the fit of alternative models using different ways of operationalizing exposure. For example, disease risk may increase or decrease linearly with exposure (a straight line graph), or exponentially (a S- or J-shaped curve), or it may be a U- shaped curve, etc. In the situation where the main aim is not for assessing the disease risk but for prediction, one can construct an receiver-operative characteristic (ROC) curve and determine the prognostic performance of the model. In such case, the confounding or interaction effect is not of interest. Thus investigator should be clear the main objective before fitting the model whether it is a risk assessment goal or a prediction goal (Kleinbaum, 1994). Model fitting strategies are quite different for each goals. This paper focused only on the risk assessment goal. For prediction goal, a good introductory reading was given by Kleinbaum, et al. (1998); page 386 - 403.

Example 6.1

The following example has been adapted from an unpublished study conducted in Indonesia (CCEB, 1993). Some modification to the data were made to enable experiencing most common steps of the analysis and using all necessary commands for the analysis. All the analysis was performed using Stata. Steps of the analysis involved univariate, bivariate or crude analysis, stratified analysis, and multivariable analysis which use logistic regression. The first three steps serve as an exploratory data analysis. The last step is the one from which the conclusion will be drawn. Below is the description of the data set to be used in this example.

A cross-sectional study was conducted among 465 women who have had delivered their children 1 to 6 months before the study was started. It aimed to determine the effect of antenatal care (ANC) on neonatal death. The mothers were randomly selected and interviewed using a structural questionnaire. The data file is available at *http://web.kku.ac.th/~bandit/data.* Below is the description of the variables.

Table 6.1 Description of the "Example data set"

Variable names	Descriptions	Values
DEAD	Dead within the first month of life	1 = Dead 0 = Alive
ANC	Mothers having antenatal care	1 = Yes 0 = No
SMK	Parents' smoking status	1 = Smoker 0 = Non-smoker
BWT	Birth weight	Weight in grams
MAGE	Mother's age	Age in years
PLACE	Place of birth	0 = Hospital 1 = Health center 2 = Home 3 = Road side (During travelling)

Preview of the problem:

Based on the research question that "Does ANC affect neonatal death?", we should know that this is the question for "risk assessment" where "ANC" is the "risk of interest". (Detailed discussion was provided in the next section.) On the contrary, if the research question is that "What is the best prediction model for neonatal death?" or more general "How neonatal dead is predicted?", it is the question for "prediction". Classifying the two different goals of analysis is necessary as mentioned above that modeling for risk assessment goal is different from for prediction goal.

The followings are steps commonly performed for most of data analysis. To be complete, there are computer outputs, using smaller and different letter fonts from the main text, inserted throughout the presentation. The lines in bold letter

following a "dot" before each outputs are the Stata commands so that one can repeat these and should get the same outputs. The outputs within ovals were to be quoted in the research report.

Step 1 Exploring the data and univariate analysis

To get familiar with the data set, we can display them in a listing form of data records as follows:

```
. list  dead anc smk bwt mage place

           dead      anc       smk      bwt       mage      place
  1.        1         1         0      2600        30         0
  2.        1         1         0      2900        29         1
  3.        1         1         0      3100        25         0
```

--- skip 460 records ---

```
464.        0         1         0      3500        30         0
465.        0         1         0      3200        22         1
```

Note that, to stop displaying all data records, we need to hold down the key <Ctrl> then press <Break> once.

The data file can also be summarized as follows:

```
. summarize
Variable |      Obs        Mean   Std. Dev.       Min        Max
---------+-----------------------------------------------------
    dead |      465    .1397849   .3471372          0          1
     anc |      465    .5182796   .5002039          0          1
     smk |      465    .0752688   .2641087          0          1
     bwt |      465    3010.695   437.7349       1850       4000
    mage |      465    25.52473   5.362298         17         42
   place |      465    .2408602   .5273217          0          3
```

At a first look at the data description and outputs from the two commands above, we should be able to classify type of variables. That is, "DEAD" is the dependent or response variable which is in nominal scale or two possible values. In other words, this study has a dichotomous outcome. The other five variables are the independent or explanatory variables.

Someone called these the (exposing) factors or predictors. Among these, "ANC" is the exposure of interest based on the research question. If the research objective changed from "to determine the effect of "ANC" on neonatal dead" to that "to determine factors affecting neonatal dead", there will be no exposure of interest. Classifying the two different type of explanatory variables is necessary for further analysis which is quite different from one another.

The real analysis begins with the univariate analysis - analyze one variable at a time. For this study, it is a cross-sectional study. We need to know the overall proportion, or more specifically - the prevalence, of neonatal dead. This is important for interpretation of odds ratio which is approximate the relative risk if the event is rare. Two simple Stata commands for this purpose are as follows:

```
. tab dead

       dead |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        400       86.02       86.02
          1 |         65       13.98      100.00
------------+-----------------------------------
      Total |        465      100.00
```

```
. ci dead

Variable |     Obs        Mean    Std. Err.       [95% Conf. Interval]
---------+-------------------------------------------------------------
    dead |     465     .1397849    .0160981        .1081507    .1714192
```

The prevalence of 0.13 is not too bad to assume to be rare event. Thus we can validly interpret the odds ratio as for the relative risk. The full format of reporting the above result is : "Among a total of 465 children, 65 died within one month of age. The neonatal dead rate was 14.0% (95%CI: 10.8% to 17.1%)".

Step 2 Bivariate (crude) analysis

This section is to determine, separately, the effect of each factor on DEAD ignoring the effect of other factors. This is an important steps for the study with several factors involved. It serve as a good tools for screening potential predictors to be the candidate to be entered into the initial model. As the rule of thumb, variables that have the p-value of 0.2 or lower will be considered to be the candidate. However, variables that have the p-value exceed 0.2 but were known to have an effect on the outcome were also considered to be the candidate. Commands and outputs were shown below. Some results from these outputs which are in the ovals will be reported in the table at the end of this paper.

Section 2.1 Crude effect of ANC on DEAD

ANC is a dichotomous predictor. Odds ratio is an appropriate measure of association since it is a cross-sectional study.

```
. cs dead anc, or

                 | anc                      |
                 |   Exposed   Unexposed    |      Total
-----------------+------------------------+------------
           Cases |        37          28    |         65
        Noncases |       204         196    |        400
-----------------+------------------------+------------
           Total |       241         224    |        465
                 |                          |
            Risk |   .153527        .125    |   .1397849
                 |                          |
                 |      Point estimate      |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .028527          |   -.0342996    .0913535
      Risk ratio |        1.228216          |     .778466    1.937803
 Attr. frac. ex. |        .1858108          |   -.2845776    .4839517
 Attr. frac. pop |        .1057692          |
      Odds ratio |        1.269608          |    .7512221    2.145309  (Cornfield)
                 +-------------------------------------------------
                               chi2(1) =     0.79  Pr>chi2 = 0.3754
```

Among a total of 241 children with ANC mothers, 15.4% of their children died whereas among 224 children with non-ANC mothers, 12.5% of their children died. Children whose mothers had ANC were 1.26 times more likely to die than

those had not (95%CI: 0.8 to 2.1). However, this is not statistically significant (p-value = 0.375).

Note that, based on the objective of the study, this is an exposure of interest. Although its crude effect yields p-value > 0.2, it has to be a candidate variable unless the research question will not be answered. Note also that this is a cross-sectional study and so the grand total only (465) is fixed. But the proportion reported here is assuming the column totals to be fixed. This is for simplicity of interpretation.

Section 2.2 Crude effect of SMK on DEAD

Similar to ANC, the SMK is a dichotomous predictor which can use the same command for analysis.

```
. cs dead smk, or

                 | smk                    |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        21          44  |         65
        Noncases |        14         386  |        400
-----------------+------------------------+----------
           Total |        35         430  |        465
                 |                        |
            Risk |        .6    .1023256  |   .1397849
                 |                        |
                 |      Point estimate    |   [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |          .4976744      |   .3328653    .6624835
      Risk ratio |          5.863636      |   3.972854    8.654289
 Attr. frac. ex. |          .8294574      |   .7482918    .8844504
 Attr. frac. pop |          .2679785      |
      Odds ratio |          13.15909      |   6.309044    27.44195  (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =    66.67  Pr>chi2 = 0.0000
```

Among 35 children with smoker parents, 60% of their children died as compared to the corresponding rate of 10.2% for 430 children with non-smoker parents. There is a statistically significant association between parent smoking and children dead (p-value < 0.001). That is, children whose

parents smoked were 13.2 times more likely to die than those whose parents did not smoke (95%CI: 6.3 to 27.4).

Note that this variable is undoubtedly a candidate variable for the initial model. Also the output gave p-value of 0.0000 but we cannot quote so which means impossible. Quoting this as p-value < 0.001 or even p-value < 0.01 is recommended.

Section 2.3 Crude effect of BWT on DEAD

BWT is a continuous variable. To be able to assess some trend of its effect and perform a stratified analysis in further steps, we need to categorize it. Based on empirical knowledge, we do that into three groups using the first four commands. Then we perform an overall test for association as well as the appropriate proportions to be quoted in the report. Finally we obtain the "local odds ratios" for each 2 by 2 table using appropriate reference category of BWT. We may choose "Lower than 2500 grams" as the reference group as it is easy to examine the trend.

```
. gen bwtg = .
(465 missing values generated)
```

```
. replace bwtg = 1 if bwt < 2500
(39 real changes made)
```

```
. replace bwtg = 2 if bwt >= 2500 & bwt <= 3000
(213 real changes made)
```

```
. replace bwtg = 3 if bwt > 3000
(213 real changes made)
```

```
. tab  bwtg dead, row chi2 exact

           |         dead
      bwtg |         0          1 |     Total
-----------+----------------------+----------
         1 |        27         12 |        39 
           |     69.23      30.77 |    100.00 
-----------+----------------------+----------
         2 |       175         38 |       213 
           |     82.16      17.84 |    100.00 
-----------+----------------------+----------
         3 |       198         15 |       213 
           |     92.96       7.04 |    100.00 
-----------+----------------------+----------
     Total |       400         65 |       465 
           |     86.02      13.98 |    100.00 

          Pearson chi2(2) =  20.3082   Pr = 0.000
           Fisher's exact =                 0.000
```

The next two commands are immediate commands for analysis of 2 by 2 table. One need to be very careful about how to enter the four cell frequencies so that the odds ratio is meaningful and remain the same as what appeared in the main table above.

```
. csi 38 12 175 27, or

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        38          12  |         50
        Noncases |       175          27  |        202
-----------------+------------------------+----------
           Total |       213          39  |        252
                 |                        |
            Risk |  .1784038    .3076923  |   .1984127
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |        -.1292886       | -.2829945    .0244174 
      Risk ratio |         .5798122       |  .3338618     1.00695 
 Prev. frac. ex. |         .4201878       |   -.00695    .6661382 
 Prev. frac. pop |         .3551587       |
      Odds ratio |         .4885714       |  .2294889    1.037412  (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =     3.46  Pr>chi2 = 0.0627
```

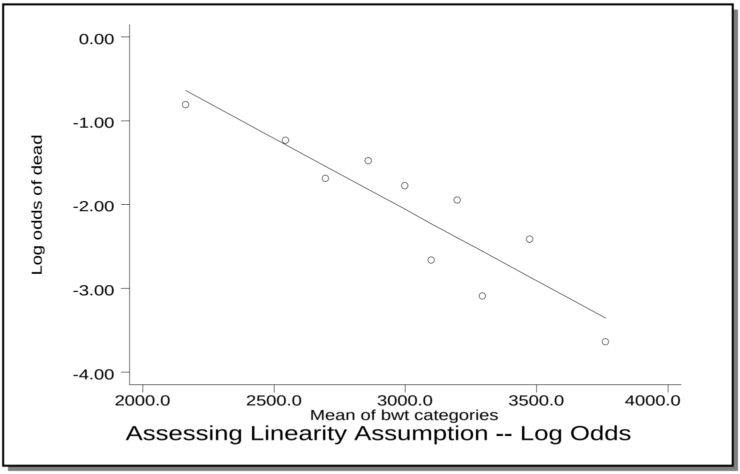
```
. csi 15 12 198 27, or
```

```
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        15          12  |         27
        Noncases |       198          27  |        225
-----------------+------------------------+----------
           Total |       213          39  |        252
                 |                        |
            Risk |  .0704225    .3076923  |   .1071429
                 |                        |
                 |      Point estimate    |   [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         -.2372698      |  -.386141   -.0883985
      Risk ratio |          .2288732      |  .1161831    .4508654
 Prev. frac. ex. |          .7711268      |  .5491346    .8838169
 Prev. frac. pop |          .6517857      |
      Odds ratio |          .1704545      |  .0730813    .3964905  (Cornfield)
                 +-----------------------------------------------
                            chi2(1) =    19.40  Pr>chi2 = 0.0000
```

Clearly, BWT is a candidate variable for entering the initial model. From the odds ratio for the three group of BWT on DEAD, they suggested no obvious departure from linear trend, i.e, the odds ratio decrease as the BWT increase (from 1 to 0.5 and then to 0.2 for BWT of <2500, 2500-3000, and >3001, respectively). Another useful command to examine linear trend is the "lintrend" command (see below for more details) as follows:

```
. lintrend dead bwt, groups(12) plot(log) xlab ylab
```

```
The proportion and log odds of dead by categories of bwt

  (Note: 12 bwt categories of equal sample size;
     Uses mean bwt value for each category)

      bwt        min        max       d    total     dead  logodds
   2162.8       1850       2400      12       39     0.31    -0.81
   2544.1       2500       2600      14       62     0.23    -1.23
   2695.3       2650       2700       5       32     0.16    -1.69
   2858.1       2750       2900       8       43     0.19    -1.48
   2998.0       2950       3000      11       76     0.14    -1.78
   3099.1       3060       3100       3       46     0.07    -2.66
   3196.9       3150       3200       4       32     0.12    -1.95
   3293.5       3250       3300       1       23     0.04    -3.09
   3473.7       3380       3500       6       73     0.08    -2.41
   3761.5       3600       4000       1       39     0.03    -3.64
```



The "lintrend" command is an batch file containing series of Stata commands, called an automatic do or "ado" file. The program was written by Garrett J. M. (3/96 STB Reprints Volume 5, pages 152-160) available at *http://www.stata.com*. It graphically examines the relationship between the log odds of a binary outcome by categories of an ordinal or interval independent variable. Similar to the previous approach, the graph suggested that there is a linear trend.

Knowing about linear relationship between the continuous exposure and the outcome enables analyst in making decision on whether the exposure will be entered into the model as continuous or categorical form. If it is linear, the exposure can be modeled as either continuous or categorical form. The former is the most efficient but difficult interpretation. The later is less efficient since it threw away some information resulting from categorization, but it is easy to interpret and more clinically meaningful. However this might not be practical for small sample since it could lead to several dummy variables for polytomous variable after such

categorization. For this example, if we decide to use the BWTG rather than the BWT, the BWTG has to be entered as the two dummy variables. On the contrary, the exposure needs to be categorized if there is non-linear relationship.

In this example, BWT can be entered as either the continuous or categorical variable. However, it is more clinically informative if we dichotomize it into “Low” and “Normal” birth weight. We can do so as follows:

. replace bwtg = .
(465 real changes made, 465 to missing)

. replace bwtg = 1 if bwt < 2500
(39 real changes made)

. replace bwtg = 0 if bwt >= 2500
(426 real changes made)

Note that we assigned 1 = Low birth weight and 0 = Normal birth weight.

```
. cs dead bwtg, or

                 | bwtg                   |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        12          53  |         65
        Noncases |        27         373  |        400
-----------------+------------------------+----------
           Total |        39         426  |        465
                 |                        |
            Risk |  .3076923    .1244131  |  .1397849
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .1832792       |  .0350754    .3314829
      Risk ratio |         2.473149       |  1.449993    4.218275
 Attr. frac. ex. |         .5956573       |  .3103413    .7629363
 Attr. frac. pop |         .1099675       |
      Odds ratio |         3.127883       |  1.513083     6.47943  (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =     9.98  Pr>chi2 = 0.0016
```

Section 2.4 Crude effect of MAGE on DEAD

MAGE is a continuous variable. The same approach for BWT can also be applied here. There is no obvious departures from linear trend (outputs not shown). In this case we can do either as dichotomous or contintinuous. We dichotomize it as it is clinically relevant, i.e, teenage pregnancy is at high risk (and the maximum age of 42 is not too bad to have a child!?).

```
. gen mageg = .
(465 missing values generated)

. replace mageg = 1 if mage < 20
(46 real changes made)

. replace mageg = 0 if mage >= 20
(419 real changes made)


. cs dead mageg, or

                 | mageg                  |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |         7          58  |         65
        Noncases |        39         361  |        400
-----------------+------------------------+----------
           Total |        46         419  |        465
                 |                        |
            Risk |  .1521739    .1384248  |   .1397849
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .0137491       | -.0951896    .1226878
      Risk ratio |         1.099325       |  .5336421    2.264657
 Attr. frac. ex. |         .0903512       | -.8739152     .558432
 Attr. frac. pop |         .0097301       |
      Odds ratio |         1.117153       |  .4874978    2.567507  (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =     0.07  Pr>chi2 = 0.7985
```

Note that, based on the above tables, MAGE can be ignored in model fitting. However, it is known to have a strong effect on pregnancy outcome. Thus we will consider as the candidate variable based on clinical grounds. In this study, it is also

justifiable since the number of variable is not many relative to the sample size.

Section 2.5 Crude effect of PLACE on DEAD

PLACE is a polytomous predictor. First we need an overall test for association as well as the appropriate proportions to be quoted in the report.

```
. tab place dead, row chi2 exact

           |         dead
     place |         0          1 |     Total
-----------+----------------------+----------
         0 |       337         38 |       375
           |     89.87      10.13 |    100.00
-----------+----------------------+----------
         1 |        47         21 |        68
           |     69.12      30.88 |    100.00
-----------+----------------------+----------
         2 |        11          4 |        15
           |     73.33      26.67 |    100.00
-----------+----------------------+----------
         3 |         5          2 |         7
           |     71.43      28.57 |    100.00
-----------+----------------------+----------
     Total |       400         65 |       465
           |     86.02      13.98 |    100.00

          Pearson chi2(3) =  24.0179   Pr = 0.000
           Fisher's exact =                 0.000
```

From the above result, there are four cells, highlighted in bold letters, with very small numbers. This could cause a problem in modeling. Aside the two categories can be collapsed without so much loss the information and still meaningful. Therefore we do that and obtain the new result as follows:

```
. replace place = 2 if place == 3
(7 real changes made)
```

```
. tab place dead, row chi2 exact

           |         dead
     place |         0          1 |     Total
-----------+----------------------+----------
         0 |       337         38 |       375
           |     89.87      10.13 |    100.00
-----------+----------------------+----------
         1 |        47         21 |        68
           |     69.12      30.88 |    100.00
-----------+----------------------+----------
         2 |        16          6 |        22
           |     72.73      27.27 |    100.00
-----------+----------------------+----------
     Total |       400         65 |       465
           |     86.02      13.98 |    100.00

          Pearson chi2(2) =  24.0035   Pr = 0.000
           Fisher's exact =                 0.000
```

Although there were a cell with only 6 children, we will keep on analysis this until we found problem at the next stage of analysis where we will consider collapsing the category again. Then we calculate the "local odds ratios" for each 2 by 2 table using appropriate reference category of PLACE. We may choose "delivery at the hospital" as the reference group as it is more relevant, ie. the lowest risk. The next two commands are immediate commands for analysis of 2 by 2 table. Again, be very careful about how to enter the four cell frequencies so that the odds ratio is meaningful and remain the same as what appeared in the main table above.

```
. csi 21 38 47 337, or

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        21          38  |         59
        Noncases |        47         337  |        384
-----------------+------------------------+----------
           Total |        68         375  |        443
                 |                        |
            Risk |  .3088235    .1013333  |   .1331828
                 |                        |
                 |      Point estimate    |   [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |          .2074902      |   .0935114     .321469
      Risk ratio |          3.047601      |   1.912134    4.857333
 Attr. frac. ex. |           .671873      |    .477024    .7941257
 Attr. frac. pop |          .2391412      |
      Odds ratio |          3.962486      |   2.156189    7.289677  (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =    21.47  Pr>chi2 = 0.0000
```

```
. csi 6 38 16 337, or

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |         6          38  |         44
        Noncases |        16         337  |        353
-----------------+------------------------+----------
           Total |        22         375  |        397
                 |                        |
            Risk |  .2727273    .1013333  |  .1108312
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .1713939       | -.0171971     .359985
      Risk ratio |         2.691388       |  1.276449     5.67478
 Attr. frac. ex. |         .6284444       |  .2165766    .8237817
 Attr. frac. pop |         .085697        |
      Odds ratio |         3.325658       |  1.268007    8.771802  (Cornfield)
                 +-----------------------------------------------
                            chi2(1) =     6.19  Pr>chi2 = 0.0128
```

Note that PLACE is also undoubtedly a candidates variable for entering to the initial model.

Step 3 Stratified analysis

Ideally, one should examine confounding and interaction effect using stratified analysis for all possible combination of explanatory variables. By this analysis, any joint effect of the variables could be detected and thus they can be entered into the initial model appropriately. More details about this matter could be found at Kleinbaum (1994); page 164 – 173.

For this study, there is an exposure of interest - ANC. Thus all stratified analysis will be mainly to assess the effect of extraneous variables on the association between ANC and DEAD. The "extraneous variable" is sometime called the "stratified variable" in this analysis. More attention should be made on interaction effect than confounding effect. Once any variables were in the model, their confounding effects were controlled. Whilst the interaction effects are the one that researchers try to discover and explain – not to control. If they exist, the terms to be put into the model are the product of the two or more variables or interaction terms – not a single variable or main effect. Thus we will look at first the p-value

of the test for homogeneity of odds ratios across stratum. If the p-value is 0.2 or less we will consider putting such interaction in the initial model for further model fitting. By this process, we identify three interaction terms- i) ANC*SMK; ii) ANC*MAGE; iii) ANC*PLACE. Detail outputs are shown below. The black arrows point to the p-values that were used for this purpose.

Section 3.1 Effect of SMK on the association between ANC and DEAD

```
 . cc dead anc, by(smk)


             smk |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   .6711146      .3590862   1.254787      11.85116 (Cornfield)
               1 |      7.125      1.297704   37.58284      .4571429 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608      .7512221   2.145309               (Cornfield)
    M-H combined |   .9108184      .5136778   1.615001
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     5.91  Pr>chi2 = 0.0151  ⇦

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.11
                                                Pr>chi2 =    0.7453
```

The following two commands are for obtaining proportions to be reported in the last table at the end of this paper.

```
. tab anc dead if smk == 0, row chi2 exact

           |         dead
       anc |         0          1 |     Total
-----------+----------------------+----------
         0 |       190         26 |       216
           |     87.96      12.04 |    100.00
-----------+----------------------+----------
         1 |       196         18 |       214
           |     91.59       8.41 |    100.00
-----------+----------------------+----------
     Total |       386         44 |       430
           |     89.77      10.23 |    100.00

          Pearson chi2(1) =   1.5385   Pr = 0.215
           Fisher's exact =                 0.265
   1-sided Fisher's exact =                 0.140
```

```
. tab anc dead if smk == 1, row chi2 exact

           |         dead
       anc |         0          1 |     Total
-----------+----------------------+----------
         0 |         6          2 |         8
           |     75.00      25.00 |    100.00
-----------+----------------------+----------
         1 |         8         19 |        27
           |     29.63      70.37 |    100.00
-----------+----------------------+----------
     Total |        14         21 |        35
           |     40.00      60.00 |    100.00

          Pearson chi2(1) =   5.2932   Pr = 0.021
           Fisher's exact =                 0.039
   1-sided Fisher's exact =                 0.030
```

Section 3.2 Effect of BWTG on the association between ANC and DEAD

```
. cc dead anc, by(bwtg)

            bwtg |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   1.339792     .7542532   2.379341      9.934272 (Cornfield)
               1 |        .49     .1209808    1.95487      2.564103 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   1.165453     .6827702   1.989367
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     1.62  Pr>chi2 = 0.2026

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.32
                                                Pr>chi2 =    0.5728
```

Section 3.3 Effect of MAGEG on the association between ANC and DEAD

```
. cc dead anc, by(mageg)

           mageg |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   1.599555     .9133475   2.800384      9.661098 (Cornfield)
               1 |   .1571429     .0305953   .8318563      3.043478 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   1.254014     .7442425   2.112957
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     5.93  Pr>chi2 = 0.0149

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.75
                                                Pr>chi2 =    0.3867
```

Section 3.4 Effect of PLACE on the association between ANC and DEAD

```
. cc dead anc, by(place)

           place |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   .7952381      .4090156   1.546609          9.52 (Cornfield)
               1 |       3.74       1.13241   12.16321      1.470588 (Cornfield)
               2 |   .7777778      .1316856   4.564086      1.227273 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608      .7512221   2.145309               (Cornfield)
    M-H combined |   1.147927      .6675961   1.973853
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(2) =     4.84  Pr>chi2 = 0.0888

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.25
                                                Pr>chi2 =    0.6185
```

Step 4 Multivariable analysis : Logistic regression

The first step is to prepare the variables in appropriate forms based on findings from the previous crude and stratified analysis. The first three following commands are to generate the interaction terms. For the "generate" command see StataCorp (1999); page 517-520 of Volumn 1 : A-G.

```
. gen a_smk = anc * smk
. gen a_mageg = anc * mageg
. gen a_place = anc * place
```

Section 4.1. The initial model – the full model

For details of the "logit" command see StataCorp (1999); page 228-239 of Volumn 2 : H-O.

```
. xi: logit dead  anc smk  bwtg mageg i.place  a_smk a_mageg
i.a_place
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)
```

```
i.a_place            Ia_pla_0-2   (naturally coded; Ia_pla_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -158.90781
Iteration 2:   log likelihood = -151.19391
Iteration 3:   log likelihood = -150.81363
Iteration 4:   log likelihood =  -150.8124
Iteration 5:   log likelihood =  -150.8124

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(10)     =      74.63
                                                  Prob > chi2     =     0.0000
Log likelihood =  -150.8124                       Pseudo R2       =     0.1983

------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |  -.5413629   .4338548     -1.248   0.212      -1.391703    .3089768
     smk |   .8913886   .9323986      0.956   0.339      -.9360792    2.718856
    bwtg |   1.117437   .4577921      2.441   0.015       .2201811    2.014693
   mageg |   1.439287   .6143028      2.343   0.019       .2352758    2.643299
Iplace_1 |   .5058782   .6178105      0.819   0.413      -.7050082    1.716765
Iplace_2 |   1.306483   .7715727      1.693   0.090      -.2057713    2.818738
   a_smk |   2.086607   1.073441      1.944   0.052      -.0172996    4.190513
 a_mageg |  -1.630821   1.032344     -1.580   0.114      -3.654178    .3925359
Ia_pla_1 |   .8395218   .8137985      1.032   0.302      -.7554939    2.434537
Ia_pla_2 |   .2971595   1.080756      0.275   0.783      -1.821084    2.415403
   _cons |   -2.38564   .2742555     -8.699   0.000      -2.923171   -1.848109
------------------------------------------------------------------------------
```

. lrtest, saving(0)

For details of the "lrtest" command see StataCorp (1999); page 246-250 of Volumn 2 : H-O.

Among all interaction terms (in italic bold letters), the ANC*PLACE will be removed due to the highest p-value (in oval). We need to remove both dummy variables of this term as it is the principle of hierarchical well-formatted model (see more detail in Kleinbaum, 1994, page 171 – 173).

Note that the “xi” before the command “logit” is to inform Stata that there are some polytomous variables in the model so that the “i.” before the polytomous variable can tell Stata create dummy variables automatically for those variables (see more details in StataCorp, 1999 Volume 4, page 306 - 314) . The “lrtest, saving(0)” command is for performing further likelihood ratio test to assessing the effect of deleting terms

from the model. The “saving(0)” option tell us that the estimation belongs to the full model.

Section 4.2. Model without ANC*PLACE

```
. xi: logit dead  anc smk  bwtg mageg i.place  a_smk a_mageg
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -159.76855
Iteration 2:   log likelihood = -151.72756
Iteration 3:   log likelihood = -151.36622
Iteration 4:   log likelihood = -151.36543

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(8)      =      73.52
                                                  Prob > chi2     =     0.0000
Log likelihood = -151.36543                       Pseudo R2       =     0.1954

------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |   -.310655    .355062     -0.875   0.382      -1.006564    .3852538
     smk |   .9825991   .9336054      1.052   0.293      -.8472338    2.812432
    bwtg |   1.040474   .4492353      2.316   0.021        .159989    1.920959
   mageg |    1.55312   .6067453      2.560   0.010       .3639209    2.742319
Iplace_1 |   .9748812   .3862543      2.524   0.012       .2178366    1.731926
Iplace_2 |   1.452425    .539014      2.695   0.007       .3959774    2.508873
   a_smk |   2.045493   1.077159      1.899   0.058      -.0657009    4.156686
 a_mageg |  -1.861715   1.004119     -1.854   0.064      -3.829751    .1063217
   _cons |  -2.487954   .2633255     -9.448   0.000      -3.004063   -1.971846
------------------------------------------------------------------------------

. lrtest, using(0)
Logit:  likelihood-ratio test                         chi2(2)     =       1.11
                                                      Prob > chi2 =     0.5752

. lrtest, saving(1)
```

The “lrtest, using(0)” command is for performing the likelihood ratio test to assessing the effect of deleting ANC*PLACE from the model. The “using(0)” option tell us that the test compared the current model against the full model which have previously been saved in 0. The test suggests that deleting the ANC*PLACE cause no effect to the model (p-value = 0.575). Now the next candidate term to be deleted is ANC*MAGEG.

Note that the likelihood ratio of this model is saved in 1.

Section 4.3. Model without ANC*MAGE

```
. xi: logit dead  anc smk  bwtg mageg i.place  a_smk
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -161.37036
Iteration 2:   log likelihood = -153.53573
Iteration 3:   log likelihood = -153.25674
Iteration 4:   log likelihood = -153.25615

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(7)      =      69.74
                                                  Prob > chi2     =     0.0000
Log likelihood = -153.25615                       Pseudo R2       =     0.1854

------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |  -.5590873   .3350725     -1.669   0.095      -1.215817    .0976429
     smk |   .8722993   .9317147      0.936   0.349      -.9538279    2.698427
    bwtg |   1.047556   .4440794      2.359   0.018       .1771763    1.917935
   mageg |   .7140317   .4804309      1.486   0.137      -.2275954    1.655659
Iplace_1 |   .9866509   .3859063      2.557   0.011       .2302884    1.743013
Iplace_2 |   1.478023    .540374      2.735   0.006       .4189098    2.537137
   a_smk |   2.246689   1.073616      2.093   0.036       .1424403    4.350939
   _cons |  -2.382853   .2488897     -9.574   0.000      -2.870668   -1.895038
------------------------------------------------------------------------------

. lrtest, using(1)
Logit:  likelihood-ratio test                         chi2(1)     =       3.78
                                                      Prob > chi2 =     0.0518

. lrtest, saving(2)
```

Again, the test suggests that deleting the ANC*MAGE cause no effect to the model (p-value = 0.052). Now the next candidate term to be deleted, based on p-value, is SMK. But we cannot delete it, based on the hierarchical well-formatted principle, since it is a component of a significant interaction term ANC*SMK. Thus there is only MAGEG that can be considered for deletion.

Section 4.4. Model without MAGEG

```
. xi: logit dead  anc smk  bwtg i.place  a_smk
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -162.26531
Iteration 2:   log likelihood = -154.50657
Iteration 3:   log likelihood = -154.26029
Iteration 4:   log likelihood =  -154.2599

Logit estimates                                   Number of obs   =        465
```

```
                                                  LR chi2(6)      =      67.73
                                                  Prob > chi2     =     0.0000
Log likelihood =  -154.2599                       Pseudo R2       =     0.1800

------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |  -.5157685   .3326557     -1.550   0.121      -1.167762    .1362246
     smk |   .7955996    .925203      0.860   0.390      -1.017765    2.608964
    bwtg |   1.093564   .4429316      2.469   0.014       .2254335    1.961694
Iplace_1 |   .8849724    .376117      2.353   0.019       .1477966    1.622148
Iplace_2 |   1.365092   .5319488      2.566   0.010       .3224913    2.407692
   a_smk |   2.266141   1.069409      2.119   0.034        .170138    4.362143
   _cons |  -2.295464   .2367904     -9.694   0.000      -2.759565   -1.831363
------------------------------------------------------------------------------

. lrtest, using(2)
Logit:  likelihood-ratio test                         chi2(1)     =       2.01
                                                      Prob > chi2 =     0.1565
```

Again, the test suggests that deleting the MAGEG cause no effect to the model (p-value = 0.156). We have no other terms that can be removed since they are all statistically significant predictors of DEAD. Thus the above model is the final model.

Step 5 Assessing model adequacy: test for goodness of fit of the model

```
. lfit

Logistic model for dead, goodness-of-fit test

       number of observations =       465
 number of covariate patterns =        16
              Pearson chi2(9) =        17.32
                  Prob > chi2 =         0.0440
```

The “lfit” command displays either the Pearson or Hosmer-Lemeshow goodness-of-fit tests. The Hosmer-Lemeshow test is preferred over the Pearson test when the number of observations per covariate pattern is small. This study, such numbers are sufficiently large. The test suggests that the model did not fit well with the data. For details of the "lfit" command see StataCorp (1999); page 209-211 of Volumn 2 : H-O.

Further analysis has been done to explore another type of the model. It was found that the model that fit well with the data is the one that did not categorize the BWT and MAGE. All commands and their outputs of fitting the model are listed in NOTE 1. The test of goodness-of-fit of the model yields p-value = 0.465. However the final model contains exactly the same variables as the above model where continuous variables were categorized. Comparing between the two models, the coefficients are very slightly different. Thus we choose the above model as it is more simple interpretation and informative.

Further assessment of the model can be done using methods proposed by Hosmer and Lemeshow (1989). The methods are mainly aim to detect the influence observation(s). That is, the one that causes unstable in model estimation. This can help improving the fit of the model and lead to a more valid model. Series of Stata commands facilitate this procedure (see more details in StataCorp, 1999 Volume 2, page 200-222).

Step 6 Obtaining measure of associations from the model

Odds ratios can be estimated using the command “logistic” as shown below. From the model, the odds ratio that can be obtained directly from the output are that of BWTG and PLACE (italic bold letters). For details of the "logistic" command see StataCorp (1999); page 201-226 of Volume 2 : H-O.

```
. xi: logistic dead  anc smk  bwtg i.place  a_smk
i.place             Iplace_0-2   (naturally coded; Iplace_0 omitted)

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(6)      =      67.73
                                                  Prob > chi2     =     0.0000
Log likelihood =  -154.2599                       Pseudo R2       =     0.1800

------------------------------------------------------------------------------
    dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |   .5970416   .1986093     -1.550   0.121       .3110624    1.145939
     smk |   2.215769   2.050036      0.860   0.390       .3614018    13.58497
    bwtg |   2.984892   1.322103      2.469   0.014       1.252866     7.11136
Iplace_1 |   2.422917   .9113004      2.353   0.019       1.159277    5.063957
Iplace_2 |   3.916082   2.083155      2.566   0.010       1.380563     11.1083
   a_smk |   9.642116   10.31136      2.119   0.034       1.185468    78.42503
------------------------------------------------------------------------------
```

In the present of an interaction effect which is a product of two variables, we need to estimate the odds ratios of one variable separately each level of the other variable. For the interaction term of ANC and SMK, we need to get the odds ratios of ANC on DEAD for each group of SMK. The effect of ANC among SMK = 0 is given by the odds ratio 0.60 for “anc” in the output above. The effect of ANC among SMK = 1 is given by the following commands.

```
. lincom anc + a_smk

 ( 1)  anc + a_smk = 0.0

------------------------------------------------------------------------------
    dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   5.756744   5.836508      1.726   0.084       .7891896    41.99257
------------------------------------------------------------------------------
```

For details of the "lincom" command see StataCorp (1999); page 179-185 of Volumn 2 : H-O.

Note that the combination of the two terms originally came from the principle of obtaining odds ratio from logistic regression model. The model can be written as follows:

Logit P(X) = a + b_1ANC + b_2SMK + b_3BWTG + b_4PLACE1
+ b_5PLACE2 + b_6ANC*SMK

We can use the corresponding coefficients in the output from "logit" command shown in Section 4.4 to replace the model as follows:

$$\text{Logit } P(X) = -2.29 - 0.52\text{ANC} + 0.80\text{SMK} + 1.10\text{BWTG} \\ + 0.90\text{PLACE1} + 1.36\text{PLACE2} + 2.27\text{ANC*SMK}$$

Given SMK = 0, the odds ratio of ANC can be estimated by comparing the two odds. That is, the odds of ANC = 1 divided by the odds of ANC = 0 (or simply odds of exposed divided by the odds of non-exposed). Since the coefficients are in log transformation, it is obtained by subtracting the two odds and then take the exponential (or anti-logs) of such results. Thus

$$\text{Logit } P(\text{SMK}=0, \text{ANC}=1) = a + b_1(1) + b_2(0) + b_3\text{BWTG} + b_4\text{PLACE1} \\ + b_5\text{PLACE2} + b_6(1)*(0)$$

$$\text{Logit } P(\text{SMK}=0, \text{ANC}=0) = a + b_1(0) + b_2(0) + b_3\text{BWTG} + b_4\text{PLACE1} \\ + b_5\text{PLACE2} + b_6(0)*(0)$$

Subtracting the second odds from the first odds given

$$\text{Log(Odds ratio)} = b_1 = -0.52$$

Thus Odds ratio = EXP(-0.52) = $2.7183^{(-0.52)}$ = 0.59

Similarly, given SMK = 1,

$$\text{Logit } P(\text{SMK}=1, \text{ANC}=1) = a + b_1(1) + b_2(1) + b_3\text{BWTG} + b_4\text{PLACE1} \\ + b_5\text{PLACE2} + b_6(1)*(1)$$

$$\textbf{Logit P(SMK=1, ANC=0)} = a + b_1(0) + b_2(1) + b_3\text{BWTG} + b_4\text{PLACE1} + b_5\text{PLACE2} + b_6(0)*(1)$$

Subtracting the second odds from the first odds given

Log(Odds ratio) = $b_1 + b_6 = -0.52 + 2.27 = 1.75$

Thus Odds ratio = EXP(1.75) = $2.7183^{(1.75)} = 5.7$

The "$b_1 + b_6$" is equivalent to the combination of ANC and SMK following the command "lincom" shown above.

Step 7 Summarize findings

Among a total of 465 children, 65 died within one month of age. The neonatal dead rate was 14.0% (95%CI: 10.8% to 17.1%). Children whose mothers had ANC were 1.3 times more likely to die than those had not (95%CI: 0.8 to 2.1, Table 6.1). However, this is not statistically significant (p-value = 0.375). Age of mothers at date of delivery was also not a significant predictor of children dead (p-value = 0.799) as there was similar proportions of dead among the two age group - teenage pregnancy and pregnancy at age of 20 years or more. The following three factors were statistically significant associated with dead: i) parent smoking (OR = 13.2; 95%CI: 6.3 to 27.4; p-value < 0.001), ii) low birth weight (OR = 3.1; 95%CI: 1.5 to 6.5; p-value = 0.002); and iii) place of delivery (p-value < 0.001) where delivered at health center were 4.0 times (95%CI: 2.2 to 7.3) and delivered at home or road side were 3.3 times (95%CI: 1.3 to 8.8) more likely to die than delivered at the hospital. However all these effects ignore the effect of other factors.

Table 6.2 Crude effect of each factor on neonatal dead

Factors	Number	Dead (%)	OR	95%CI	p-value
1. Attending ANC					0.375
Yes	241	15	1.3	0.8 to 2.1	
No	224	12	1.0		
2. Parents smoking during pregnancy					<0.001
Yes	35	60	13.2	6.3 to 27.4	
No	430	10	1.0		
3. Birth weight					0.002
< 2,500 grams	39	31	3.1	1.5 to 6.5	
2,500 grams or more	426	12	1.0		
4. Age of mothers at date of delivery					0.799
Less than 20 years	46	15	1.1	0.5 to 2.6	
20 or more	419	14	1.0		
5. Place of delivery					<0.001
Hospital	375	10	1.0		
Health center	68	31	4.0	2.2 to 7.3	
Home or Road side	22	27	3.3	1.3 to 8.8	

Taken into account of effects of other factors, there was a significant interaction effect (p-value = 0.034, Table 6.2). Parents smoking status was an effect modifier of the association between attending ANC and neonatal dead. That is, among children whose parents smoked, those whose mothers attending ANC was 5.8 times more likely to die within the first month of life than those whose mothers did not (95%CI: 0.8 to 42.0). On the contrary, among children whose parents did not smoke, the corresponding adjusted odds ratio (OR_{adj}) was 0.6 (95%CI: 0.3 to 1.1) suggesting a protective

effect of ANC. These effects have also been adjusted for the effect of birth weight and place of deliveries. Low birth weight has a significant risk effect on neonatal dead (OR_{adj} = 3.0; 95%CI: 1.2 to 7.1; p-value = 0.014). Similarly, place of deliveries has a significant risk effect on neonatal dead (p-value = 0.010). That is, comparing to delivered a baby at the hospital, those who delivered at health center has a higher risk to neonatal dead (OR_{adj} = 2.4; 95%CI: 1.2 to 5.1) and also at home or road side (OR_{adj} = 3.9; 95%CI: 1.4 to 11.1).

Table 6.3 Crude and adjusted odds ratio of each factors on neonatal dead

Factors	No	Dead (%)	Crude OR	Adjusted OR	95%CI	p-value
1. Attending ANC for each group of parents smoking status						**0.034**
1.1 Parents smokers						
Attending ANC	**27**	**70**	**7.1**	**5.8**	**0.8 to 42.0**	
Did not attending ANC	**8**	**25**	**1.0**	**1.0**		
1.2 Parent non-smokers						
Attending ANC	**214**	**8**	**0.7**	**0.6**	**0.3 to 1.1**	
Did not attending ANC	**216**	**12**	**1.0**	**1.0**		
2. Birth weight						**0.014**
< 2,500 grams	**39**	**31**	**3.1**	**3.0**	**1.2 to 7.1**	
2,500 grams or more	**426**	**12**	**1.0**	**1.0**		
3. Place of delivery						**0.010**
Hospital	**375**	**10**	**1.0**	**1.0**		
Health center	**68**	**31**	**4.0**	**2.4**	**1.2 to 5.1**	
Home or Road side	**22**	**27**	**3.3**	**3.9**	**1.4 to 11.1**	

NOTE 1: Model fitting without categorizing of all continuous variables showed that results were not obviously differ from the above approach where continuous variables were categorized. This suggested the conclusions that had been made above were robust. If this happen to be different, sources of the differences need to be investigated further. Choices of the model need to be carefully chosen. In fact, how each continuous variable will categorized must be decided in advance, i.e., before the data were collected to avoid bias. Followings are the commands and their outputs for this approach.

```
. gen a_mage = anc*mage

. xi: logit dead  anc smk  bwt mage i.place  a_smk a_mage
i.a_place
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)
i.a_place            Ia_pla_0-2   (naturally coded; Ia_pla_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -154.90817
Iteration 2:   log likelihood = -152.90179
Iteration 3:   log likelihood = -147.00808
Iteration 4:   log likelihood = -146.85182
Iteration 5:   log likelihood = -146.85158

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(10)     =      82.55
                                                  Prob > chi2     =     0.0000
Log likelihood = -146.85158                       Pseudo R2       =     0.2194

------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |  -2.467283   1.593321     -1.549   0.121      -5.590135    .6555685
     smk |   .2300021   .9311643      0.247   0.805      -1.595046    2.055051
     bwt |  -.0013569   .0003691     -3.677   0.000      -.0020803   -.0006336
    mage |  -.0853645   .0485749     -1.757   0.079      -.1805696    .0098406
Iplace_1 |    .375604   .6293903      0.597   0.551      -.8579782    1.609186
Iplace_2 |   2.055076   .8997447      2.284   0.022       .2916088    3.818543
   a_smk |   2.694181   1.094326      2.462   0.014       .5493412     4.83902
  a_mage |   .0687016   .0652431      1.053   0.292      -.0591726    .1965758
Ia_pla_1 |   1.024134   .8310937      1.232   0.218      -.6047794    2.653048
Ia_pla_2 |  -.2549567   1.242813     -0.205   0.837      -2.690825    2.180912
   _cons |   4.028138   1.555667      2.589   0.010       .9790869     7.07719
------------------------------------------------------------------------------

. lfit

Logistic model for dead, goodness-of-fit test

       number of observations =       465
 number of covariate patterns =       348
            Pearson chi2(337) =       349.92
                  Prob > chi2 =         0.3025
```

```
. lrtest, saving(0)

. xi: logit dead  anc smk  bwt mage i.place  a_smk a_mage
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -156.25447
Iteration 2:   log likelihood =  -148.2508
Iteration 3:   log likelihood = -147.74697
Iteration 4:   log likelihood = -147.74364
Iteration 5:   log likelihood = -147.74364

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(8)      =      80.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -147.74364                       Pseudo R2       =     0.2147
------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |  -2.366272   1.482909     -1.596   0.111      -5.272719    .5401757
     smk |   .3601289    .914142      0.394   0.694      -1.431556    2.151814
     bwt |  -.0012902   .0003632     -3.552   0.000      -.0020021   -.0005783
    mage |  -.0883421   .0461116     -1.916   0.055      -.1787191    .0020349
Iplace_1 |   .9379171   .3980015      2.357   0.018       .1578486    1.717986
Iplace_2 |   1.905634    .616415      3.091   0.002       .6974824    3.113785
   a_smk |   2.681957    1.08377      2.475   0.013       .5578071    4.806106
  a_mage |   .0713555   .0580151      1.230   0.219      -.0423521     .185063
   _cons |   3.823132   1.517522      2.519   0.012       .8488435     6.79742
------------------------------------------------------------------------------

. lrtest, using(0)
Logit:  likelihood-ratio test                         chi2(2)     =       1.78
                                                      Prob > chi2 =     0.4098
. lrtest, saving(1)

. xi: logit dead  anc smk  bwt mage i.place  a_smk
i.place              Iplace_0-2   (naturally coded; Iplace_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -156.79104
Iteration 2:   log likelihood = -148.99653
Iteration 3:   log likelihood = -148.51621
Iteration 4:   log likelihood = -148.51328
Iteration 5:   log likelihood = -148.51328

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(7)      =      79.23
                                                  Prob > chi2     =     0.0000
Log likelihood = -148.51328                       Pseudo R2       =     0.2106
------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |  -.5881127   .3379489     -1.740   0.082       -1.25048     .074255
     smk |   .3594272    .935828      0.384   0.701      -1.474762    2.193616
     bwt |  -.0013061   .0003619     -3.609   0.000      -.0020155   -.0005967
    mage |  -.0495717    .032095     -1.545   0.122      -.1124767    .0133334
Iplace_1 |   .9169965   .3977724      2.305   0.021       .1373769    1.696616
Iplace_2 |   1.916286   .6132308      3.125   0.002       .7143757    3.118196
   a_smk |    2.81268   1.105094      2.545   0.011       .6467354    4.978624
   _cons |   2.905163   1.301942      2.231   0.026        .353403    5.456923
------------------------------------------------------------------------------
. lrtest, using(1)
Logit:  likelihood-ratio test                         chi2(1)     =       1.54
                                                      Prob > chi2 =     0.2147
```

```
. lrtest, saving(2)

. xi: logit dead  anc smk  bwt i.place  a_smk
i.place             Iplace_0-2   (naturally coded; Iplace_0 omitted)

Iteration 0:   log likelihood =  -188.1264
Iteration 1:   log likelihood = -158.01447
Iteration 2:   log likelihood = -150.15686
Iteration 3:   log likelihood = -149.76281
Iteration 4:   log likelihood =  -149.7609
Iteration 5:   log likelihood =  -149.7609

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(6)      =      76.73
                                                  Prob > chi2     =     0.0000
Log likelihood =  -149.7609                       Pseudo R2       =     0.2039
------------------------------------------------------------------------------
    dead |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |  -.5391429   .3357003     -1.606   0.108      -1.197103    .1188177
     smk |   .4136563   .9396272      0.440   0.660      -1.427979    2.255292
     bwt |  -.0013371   .0003592     -3.722   0.000      -.0020411    -.000633
Iplace_1 |   .8139901   .3862701      2.107   0.035       .0569146    1.571066
Iplace_2 |   1.490551   .5456082      2.732   0.006       .4211785    2.559924
   a_smk |   2.554702   1.090557      2.343   0.019       .4172501    4.692153
   _cons |   1.774267   1.066874      1.663   0.096      -.3167668    3.865302
------------------------------------------------------------------------------
. lrtest, using(2)
Logit:  likelihood-ratio test                         chi2(1)     =       2.50
                                                      Prob > chi2 =     0.1142
. lfit

Logistic model for dead, goodness-of-fit test

       number of observations =       465
 number of covariate patterns =       121
             Pearson chi2(114) =       114.66
                   Prob > chi2 =         0.4650
```

This suggests the model fits reasonably well to the data.

NOTE 2 : Analysis results showing how stratified analysis is similar to logistic regression

Using the ANC data to examine the effect of ANC on DEAD, controlling for the effect of SMK

Note 2.1 Stratified analysis

```
. cc dead anc, by(smk)

             smk |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   .6711146      .3590862   1.254787      11.85116 (Cornfield)
               1 |      7.125      1.297704   37.58284      .4571429 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608      .7512221   2.145309               (Cornfield)
    M-H combined |   .9108184      .5136778   1.615001
-----------------+-------------------------------------------------
```

```
Test of homogeneity (M-H)      chi2(1) =     5.91  Pr>chi2 = 0.0151

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.11
                                                Pr>chi2 =    0.7453
```

Note 2.2 Logistic regression equivalent to the stratified analysis shown in Note 2.1, ignoring interaction effect.

```
. logistic dead anc smk

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(2)      =      45.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -165.47238                       Pseudo R2       =     0.1204

------------------------------------------------------------------------------
    dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |   .9077381   .2698217     -0.326   0.745       .5069255    1.625463
     smk |   13.52737   5.277243      6.677   0.000       6.297178      29.059
------------------------------------------------------------------------------
```

Note that the model without an interaction term, the odds ratio of ANC is the actually the Mantel-Haenszel odds ratio obtained in stratified analysis.

Note 2.3 Logistic regression equivalent to the stratified analysis in #1, incorporating interaction effect.

Note 2.3.1 Generate an interaction term between ANC and SMK

```
. gen x = smk*anc
```

Note 2.3.2 Fit the logistic regression model with the interaction term

```
. logistic dead anc smk x

Logit estimates                                   Number of obs   =        465
                                                  LR chi2(3)      =      52.05
                                                  Prob > chi2     =     0.0000
Log likelihood = -162.10316                       Pseudo R2       =     0.1383
```

```
------------------------------------------------------------------------------
    dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     anc |   .6711146   .2168253     -1.234   0.217       .3562775    1.264168
     smk |   2.435897   2.053089      1.056   0.291       .4669028    12.70842
       x |   10.61667   10.34062      2.425   0.015       1.573689     71.6238
------------------------------------------------------------------------------
```

Note that in the model with an interaction term, the odds ratio of ANC is the one given that SMK = 0. It is exactly the same as that obtained from stratified analysis in #1 and shown again below.

```
. cc dead anc, by(smk)

             smk |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   .6711146     .3590862   1.254787      11.85116 (Cornfield)
               1 |      7.125     1.297704   37.58284      .4571429 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.269608     .7512221   2.145309               (Cornfield)
    M-H combined |   .9108184     .5136778   1.615001
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     5.91  Pr>chi2 = 0.0151

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.11
                                                Pr>chi2 =    0.7453
```

Note 2.3.3 Obtain the odds ratio for each group of SMK

The odds ratio of ANC on DEAD for each group of SMK can be obtained from the linear combination of coefficient estimated by the logistic model. Such combination is ANC + ANC*SMK (or x in the output). For SMK = 0, the coefficient of ANC*SMK is also zero. The linear combination of the coefficient is ANC + 0 = ANC. Thus the odds ratio can be obtained directly from the output of logistic command as shown above. For SMK = 1, the coefficient of ANC*SMK is as it is. Thus the linear combination of the coefficient is ANC + ANC*SMK as given below.

```
. lincom anc + x

 ( 1)  anc + x = 0.0

------------------------------------------------------------------------------
    dead | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |      7.125   6.546829      2.137   0.033       1.176673    43.14337
------------------------------------------------------------------------------
```

It is exactly the same as that obtained from stratified analysis in #1 which was shown again below.

```
. cc dead anc, by(smk)

             smk |       OR       [95% Conf. Interval]    M-H Weight
-----------------+------------------------------------------------------
               0 |   .6711146      .3590862   1.254787      11.85116 (Cornfield)
               1 |      7.125     1.297704   37.58284      .4571429 (Cornfield)
-----------------+------------------------------------------------------
           Crude |   1.269608      .7512221   2.145309                (Cornfield)
    M-H combined |   .9108184      .5136778   1.615001
-----------------+------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     5.91  Pr>chi2 = 0.0151

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.11
                                                Pr>chi2 =    0.7453
```

Chapter references

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

CCEB (Centre for Clinical Epidemiology and Biostatistics). The University of Newcastle, Australia. (1993) *STAT402:Analysis of categorical data - Logistic regression.* Newcastle: The University of Newcastle, NSW. Australia.

Everitt, B.S. (1977). *The Analysis of Contingency Tables*. London: Chapman and Hall.

Fienberg, S.E. (1980). *The analysis of cross-classified categorical data*. 2nd edition. Cambridge: The MIT Press.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. 2nd edition. New York: John Willey & Sons.

Friedman, L.M., Furberg, S.C.D., and DeMets, D.L. (1996). *Fundamental of clinical trials*. 3rd edition. St.Louis: Mosby.

Hosmer, D.W. Lemeshow, S. (1989). *Applied Logistic Regression*. New York: John Wiley & Sons.

Kleinbaum, D.G. (1994). *Logistic Regression: A self-learning text*. New York Springer-Verlag.

Kleinbaum, D.G. (1996). *Survival analysis: A self-learning text*. New York: Springer-Verlag.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982). *Epidemiologic research: principles and qualitative methods*. London: Lifetime Learning Publications.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E, and Nizam, A. (1998). *Applied regression analysis and other multivariable methods*. Pacific Grove: Duxbury Press.

Liang, K-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* 73(l):13-22.

Meinert, C.L. and Tonascia, S. (1986). *Clinical trials, design, conduct, and analysis*. New York: Oxford University Press.

Rabe-Hesketh, S. and Everitt, B. (1998). *A handbook of statistical analysis using Stata*. London: Chapman & Hall/CRC.

Rothman, K.J. and Greenland, S. (1998). *Modern epidemiology*. 2nd edition. Philadelphia: Lippincott-Raven Publishers.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

Exercise

Data from a study of Morrison et.al. (1973) reprinted by CCEB (1993) relating to survival of breast cancer patients. The variables and categories are as follows:

- Degree of chronic inflammatory reaction
(1. Minimal, 2. Moderate-Severe)

- Age of diagnosis
(1. Under 50 years, 2. 50-69 years, 3. 70 or older)

- Nuclear grade
(1. Relative malignant appearance, 2. Relative benign appearance)

- Center where patient was diagnosed
(1. Tokyo, 2. Boston, 3. Glamorgan)

- Survival for three years
(1. No, 2. Yes)

Three-year survival of breast cancer patients according to two histologic criteria, age and diagnostic center

Diagnostic Center	Age	Survived	Minimal inflammation		Greater inflammation	
			Appearance		Appearance	
			Malignant	Benign	Malignant	Benign
Tokyo	< 50	No	9	7	4	3
		Yes	26	68	25	9
	50-69	No	9	9	11	2
		Yes	20	46	18	5
	70+	No	2	3	1	0
		Yes	1	6	5	1
Boston	< 50	No	6	7	6	0
		Yes	11	24	4	0
	50-69	No	8	20	3	2
		Yes	18	58	10	3
	70+	No	9	18	3	0
		Yes	15	26	1	1
Glamorgan	< 50	No	16	7	3	0
		Yes	16	20	8	1
	50-69	No	14	12	3	0
		Yes	27	39	10	4
	70+	No	3	7	3	0
		Yes	12	11	4	1

You are assigned to analyze the data and prepare a report.

Chapter 7

Log-linear Models

Chapter Objectives

After completing this chapter, readers should be able to:

- describe log-linear models for three-dimensional tables corresponding to different hypotheses;
- fit and interpret the results of log-linear models;
- identify the log-linear models which provide the best fit to a given set of data;
- interpret the results from the analysis

Contents

7.1 Introduction

In Chapters 2, 3, 4, and 5 we focused mainly on bivariate analysis - i.e., analysis the relationship between a response and a single explanatory variable. The tables are therefore two-way contingency tables. This Chapter turns to more complicated tables. When three categorical variables form a table, it became a three-way contingency table. For example, a sample of workers were classified by their gender (male, female), smoking status (smoke, non-smoke), and lung function test results (normal, abnormal). The more complicate table, the multi-way contingency tables in general, are also formed by the same manner. Log-linear model is used to determine relationship among several variables in a multi-dimensional contingency table. It is particularly useful in the situation where there was no particular variable is a response but all are in the same root. In the workers example, the counts of workers in each category is the dependent variable. The categorical variable used to classify the workers i.e., gender, smoking status, and lung function test results are independent variables. A log-linear model also provides a powerful tool to explore the possibility of combining variable(s) in a multi-way table to simpler forms and thus simplify analysis without distorting the relationships among the categorical variables.

7.2 Principles and type of log-linear models

Log-linear model is a linear model for the natural logarithm of the expected frequencies. For a two-way table the full model

with interaction will fit the data perfectly (i.e., it is a saturated model) since the number of cell frequencies is equal to the number of parameters in the model. The interaction term represents the association between the two variables. Followings are the reasons.

Recall the probability theory, two events A and B are independent when the probability of the joint occurrence is the product of the probabilities of each events. This can be expressed as

$$P(AB) = P(A)P(B)$$

The expected value under the null hypothesis of independence for a cell within a two-way contingency table is simply that this probability multiply by the sample size.

Now, if we take logarithm of this expression, we would get

$$\log [P(AB)] = \log[P(A)P(B)] = \log[P(A)] + \log[P(B)]$$

Thus, a quantity, say θ, that reflects the association between events A and B can be expressed as

$$\theta = \log [P(AB)] - \log[P(A)] + \log[P(B)]$$

That is, when there is an association between the two variables, the logarithm of the joint probability is not just a sum of the individual probabilities. The term "log[P(AB)]" is represented by the interaction term in a log-linear model.

This principle also applied for higher dimensional tables. In this book, we focus only on the three-way contingency table.

For a three-way contingency table, the saturated model is given by

$$\text{Log-frequency} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

Parameter u denotes the overall mean level of the logarithms of the frequency. When we fit the model, we estimate these parameters. Substituting the value of i, j, and k of variable 1, 2, and 3 respectively in the right hand side of the equation provide us a logarithm of the cell frequencies. Thus a cell frequency can be obtained by just take the exponential (or anti-logarithm) of such value (see Selvin, 1995; page 310). Some works with mathematics can get the odds ratio which is the direct measure of association for log-linear model. The $u_{1(i)}$ denotes the influence of variable "1", $u_{2(j)}$ denotes the influence of variable "2", and $u_{3(k)}$ denotes the influence of variable "3" in cell i, j, and k respectively. There are two two-way interaction terms. For example, $u_{12(ij)}$ represents the joint influence of two variables, i.e., variable "1" and "2". The term $u_{123(ijk)}$ is the three-way interaction. These interaction terms indicate pattern of association among the three variables. In this model, the estimated values are identical to the observed values which are the cell frequencies. It served as a starting point for comparison of the models that do not fit the data perfectly. It implies that the average of the pairwise measure of association becomes a less accurate assessment of statistical independence. Thus it is rather uninformative.

Followings are the other six possible types of models could be achieved.

7.2.1 No three-way interaction

$$\boxed{\text{Log-frequency} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}}$$

If the association between 2 of the variables differs in degree or direction across levels of the third then we have 3-way interaction, the measure of association between the first two variables in the 1st level of the third variable is the same as the corresponding measure within the kth level of the third variable OR the association between the first and the second variable does not differ across levels of the third variable.

Since the order among the variables is arbitrary, the hypothesis of no three-way interaction implies that the association between any pair of variables is the same at all levels of the remaining variable.

7.2.2 No three-way interaction and one two-way interaction absent

$$\boxed{\textbf{Log-frequency} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}}$$

Variables 1 and 2 are independent for every level of variable 3 but each is associated with variable 3. That is, variables 1 and 2 are conditionally independent, given the level of variable 3.

7.2.3 No three-way interaction and two two-way interactions absent

$$\boxed{\textbf{Log-frequency} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}}$$

It is called the partial independence - there is an association between two of the variables whilst the third is completely independent.

7.2.4 No three-way and two-way interaction

$$\boxed{\textbf{Log-frequency} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}}$$

It is called the mutual (or complete) independence - there are no associations of any kind between the three variables. If this model fits the data adequately then the differences between cell frequencies simply reflect differences between single variable marginal totals.

7.2.5 Non-comprehensive Models

If we continue to delete terms from the log-linear model so that there are fewer terms than in the complete independence model, the model would not include all 3 variables. This is a non-comprehensive model. If such a model fits the data adequately then one or more of the variables is (are) redundant and the dimensionality of the table can be reduced accordingly.

7.2.6 Collapsibility

A 3-way table may be collapsed over any variable that is independent of at least one of the remaining pair and the reduced table analyzed. That is if partial independence holds and the model

$$\boxed{\text{Log-frequency} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{23(jk)}}$$

provides an adequate fit to the data, the table could be collapsed over any one of the three variables to simplify the analysis. When only conditional independence holds, we have the model

$$\boxed{\text{Log-frequency} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{13(ik)} + u_{23(jk)}}$$

and care must be taken in deciding which variables are collapsible.

Below are a simple summary of the three categorical variables - variable A, B, and C, in a three-way contingency table (adapted from Selvin, 1995, page 304).

Table 7.1 Type of the models from a three-way contingency table

AB related?	*AC related?*	*BC related?*	*Then*
No	No	No	Complete independence
No	Yes	Yes	Conditional independence
Yes	No	Yes	Conditional independence
Yes	Yes	No	Conditional independence
No	No	Yes	Partial independence
No	Yes	No	Partial independence
Yes	No	No	Partial independence
Yes	Yes	Yes	No independence

7.3 Fitting Log-Linear Models and Parameter Estimation

Fitting particular log-linear models to the frequencies in a contingency table is equivalent to testing particular hypotheses about the table. Assessing the adequacy of a suggested model for the data involves finding estimates of the theoretical frequencies to be expected assuming the model is correct and comparing these with the observed values by means of the likelihood ratio statistic or Pearson's chi-square

statistic. The estimated expected values are obtained as functions of the relevant marginal as indicated or by iterative procedures.

The major advantages to fitting log-linear models is that we obtain estimates of the parameters which allow us to quantify the effects of various variables and interactions. Estimates of the parameters in the fitted model are obtained as functions of the ln e_{ijk}.

7.4 Response vs Explanatory Variables

So far we have considered all variables on equal footing - not distinguishing between outcome factors and explanatory factors. However for 3 variables we could have:

1) **3 response variables**
2) **2 response variables and 1 explanatory variable**
3) **1 response variable and 2 explanatory variables.**

To handle this situation we condition (ie, fix corresponding marginal totals) on the values of the explanatory variables.

For (1) only Poisson or multinomial models are appropriate. For (2) and (3) we use product multinomial models. For Example, in a case-control study, one variable is a response variable with marginal totals fixed by design. The log-linear model should: i) include interaction terms for each explanatory variable with the response variablen and; ii) include all explanatory variables with their main effects and higher way interactions - this is equivalent to conditioning on the marginal totals of the explanatory variables.

7.5 Selection of a Model

As a number of dimensions in a table increases, so does the number of possible models and the complexity of the models. In general complicated models involving large numbers of parameters tend to fit a set of data more closely than a simpler model. On the other hand a simple model is easy to interpret and is often preferred. Thus there needs to be a trade-off between the goodness-of-fit and simplicity.

7.5.1 Goodness-of-Fit Statistics

A non-significant chi-square is desirable (indicates a good fit).

1) $$\chi_p^2 = \sum \frac{(obs - exp)^2}{exp}$$

2) $$G^2 = 2\sum obs \ \ ln\left[\frac{obs}{exp}\right]$$

If the model fitted is correct (ie, H_0 is true) and the total sample size is large then χ_p^2 and G^2 have approximate chi-square distribution with degrees of freedom given by

***df* = # cells - # parameters fitted**

Under these conditions χ_p^2 and G^2 are asymptotically equivalent (rule of thumb: very large sample ≡ 10 × number of cells).

7.5.2 To Compare Models

It may be found that several models fit the data adequately and in general the preferred model will be the one with fewer parameters. However a test between rival models may be required in a situation where the research question interested in a particular interaction term. For example, the

research question might be "Do variable A and B differ across level of variable C?". In this case, we need only to compare two models. The first model is the model with all three two-way interactions (i.e., A*B, A*C, and B*C). The second model is the conditional independence model without the two two-way interactions corresponding to the research question (i.e., A*C, and B*C). For hierarchical models such a test may be obtained by subtracting the *G^2* values for the two models to assess the change in goodness of fit which results from adding further parameters. The difference in *G^2* is compared with a chi-square distribution with *df* equal to the difference in the degrees of freedom of the two models. This is equivalent to the likelihood ratio test for logistic regression described in Chapter 6.

7.5.3 Residuals

Once a preliminary model has been fitted, it is useful to make a cell-by-cell comparison of observed and expected frequencies. If the model fits poorly in some cells, this lack of fit may indicate associations and interactions that may be added to the model. This is particularly true when some variables are ordinal since the pattern of positive and negative deviations can indicate trends that are not well represented by the model.

Cell deviations may be measured through the standardized residuals.

$$R = \frac{n-\hat{e}}{\sqrt{\hat{e}}}$$ **where *n* is the observed count**

***e* is the fitted (expected) count**

The squared standardized residuals are components of the Pearson chi-square (see the output from the final model using the stata command “loglin” with “resid” option).

7.5.4 Useful Guide

A useful guide for model selection in searching for a simple but useful model to describe the relationship within the data from a three-way table is to start with fitting the model with $u_{123(ijk)} = 0$. The model is shown below.

Log-frequency = $u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$

Then fit further 3 conditional independence models which contain only two two-way interactions in each model and 3 partial independence models which contain only one two-way interactions in each model. Then fit the complete independence model with three variables without any interaction term. By these, we would have a total of 7 models. Each model we need G^2, degree of freedom, and p-value. We then choose the model that fit well to the data (i.e., p-value > 0.05) but less complicate (i.e., none of or fewer number of two-way interaction terms). Among all model that fit well to the data, we then assess the influence of the term(s) that had been removed from the more complicated model(s) by comparing the G^2 of the selected model with the those models, one at a time, using methods described in #7.5.1 and #7.5.2. The p-value < 0.05 indicates adding the term into the selected model improve the fit, thus the model with the new term added should be chosen. Again, there needs to be a trade-off between the goodness-of-fit and simplicity.

Example 7.1
The following data was adopted from Everitt (1977); page 95. The variables and their values are:

- Blood pressure: (1=Less than 127 mm Hg, 2=127-146, 3=147-166, and 4=167 or more)

- Serum cholesterol: (1=Less than 200 mg/100 cc, 2=200-219, 3=220-259, and 4=260 or more)
- Coronary heart disease: (1=Yes, and 2=No)

Table 7.2 Number of subjects by blood pressure level, heart disease status, and serum cholesterol level - data for example 7.1

		Serum cholesterol				
Blood Pressure		1	2	3	4	Total
With coronary heart disease (1)	1	2	3	3	4	12
	2	3	2	1	3	9
	3	8	11	6	6	31
	4	7	12	11	11	41
Without coronary heart disease (2)	1	117	121	47	22	307
	2	85	98	43	20	246
	3	119	209	68	43	439
	4	67	99	46	33	245
Overall total		408	555	225	142	1330

We will use Stata to fit the log-linear model. Command for loglinear model did not available in Stata version 6 which was used throughout this book. Judson D.J. had provided the program to be used in Stata. Readers can download the program from "http://www.stata.com/". It is located in the module "smv5.1" and the details were available in Stata Technical Buletin (STB) Reprints Vol 1, pages 139-152. This program need to use with Stata version 3, 4, or 5.

First of all, we enter the data to Stata using the following format.

chd	bp	chl	pop
1	1	1	2
1	1	2	3
1	1	3	3
1	1	4	4

chd	bp	chl	pop
1	2	1	3
1	2	2	2
1	2	3	1
1	2	4	3
1	3	1	8
1	3	2	11
1	3	3	6
1	3	4	6
1	4	1	7
1	4	2	12
1	4	3	11
1	4	4	11
2	1	1	117
2	1	2	121
2	1	3	47
2	1	4	22
2	2	1	85
2	2	2	98
2	2	3	43
2	2	4	20
2	3	1	119
2	3	2	209
2	3	3	68
2	3	4	43
2	4	1	67
2	4	2	99
2	4	3	46
2	4	4	33

Data from this study form a 2-by-4-by-4 Table. The analysis using log-linear modeling applied to these data yields the following results for all eight possible models:

Saturated Model:
log-frequency = CHD + BP + CHL + CHD*BP + CHD*CHL
+ BP*CHL + CHD*BP*CHL

1. Model:

log-frequency = CHD + BP + CHL + CHD*BP + CHD*CHL + BP*CHL

```
. loglin pop  chd bp chl, fit(chd, bp, chl, chd bp, chd chl, bp
chl)
Variable chd = A
Variable bp = B
Variable chl = C
Margins fit: chd, bp, chl, chd bp, chd chl, bp chl
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -78.489746
Iteration 1: Log Likelihood = -77.683594
Iteration 2: Log Likelihood = -77.676758

Poisson regression                                Number of obs    =      32
Goodness-of-fit chi2(9)    =     4.775            Model chi2(22)   =1639.451
Prob > chi2                =    0.8534            Prob > chi2      =  0.0000
Log Likelihood             =   -77.677            Pseudo R2        =  0.9134

------------------------------------------------------------------------------
     pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |   3.495052   .3489747     10.015   0.000       2.811074     4.17903
    AB22 |   .0913574   .4512986      0.202   0.840      -.7931716    .9758864
    AB23 |  -.5623114   .3508082     -1.603   0.109      -1.249883      .12526
    AB24 |  -1.342433   .3429665     -3.914   0.000      -2.014635   -.6702307
    AC22 |   .0384452    .303493      0.127   0.899      -.5563902    .6332806
    AC23 |  -.5872325   .3285031     -1.788   0.074      -1.231087    .0566217
    AC24 |  -1.203873   .3265999     -3.686   0.000      -1.843997   -.5637492
      B2 |   -.390569   .4606399     -0.848   0.397      -1.293407    .5122687
      B3 |   .6053894   .3613974      1.675   0.094      -.1029365    1.313715
      B4 |   .7873583   .3578944      2.200   0.028       .0858983    1.488818
    BC22 |   .0865808   .1945052      0.445   0.656      -.2946424     .467804
    BC23 |   .1759085   .2501767      0.703   0.482      -.3144289    .6662458
    BC24 |   .1846616   .3200576      0.577   0.564      -.4426398    .8119629
    BC32 |   .5090796   .1700817      2.993   0.003       .1757255    .8424336
    BC33 |   .3106371    .223539      1.390   0.165      -.1274912    .7487654
    BC34 |   .5236171   .2756988      1.899   0.058      -.0167426    1.063977
    BC42 |   .3671291   .1987235      1.847   0.065      -.0223618      .75662
    BC43 |   .5494598   .2463114      2.231   0.026       .0666982    1.032221
    BC44 |   .8502138   .2934776      2.897   0.004       .2750083    1.425419
      C2 |   .0038244   .3214342      0.012   0.991      -.6261751    .6338238
      C3 |  -.3031287   .3571926     -0.849   0.396      -1.003213    .3969559
      C4 |  -.3836104   .3751534     -1.023   0.307      -1.118898    .3516767
   _cons |   1.254175   .3508825      3.574   0.000       .5664583    1.941892
------------------------------------------------------------------------------
```

2. Model:

log-frequency = CHD + BP + CHL + CHD*BP + CHD*CHL

```
. loglin pop  chd bp chl, fit(chd, bp, chl, chd bp, chd chl)
Variable chd = A
Variable bp = B
Variable chl = C
Margins fit: chd, bp, chl, chd bp, chd chl
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.
```

```
Iteration 0: Log Likelihood = -88.375977
Iteration 1: Log Likelihood = -87.495605
Iteration 2: Log Likelihood = -87.489746

Poisson regression                                 Number of obs    =       32
Goodness-of-fit chi2(18)     =    24.401           Model chi2(13)   =1619.825
Prob > chi2                  =    0.1423           Prob > chi2      =  0.0000
Log Likelihood               =   -87.490           Pseudo R2        =  0.9025

------------------------------------------------------------------------------
     pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |   3.619369   .3572211     10.132   0.000       2.919229     4.31951
    AB22 |   .0661658   .4491846      0.147   0.883      -.8142198    .9465515
    AB23 |   -.591429   .3480325     -1.699   0.089       -1.27356    .0907022
    AB24 |  -1.454255   .3392087     -4.287   0.000      -2.119092    -.789418
    AC22 |   -.030277   .3003151     -0.101   0.920      -.6188837    .5583297
    AC23 |  -.6916755   .3241887     -2.134   0.033      -1.327074   -.0562773
    AC24 |  -1.372642   .3204974     -4.283   0.000      -2.000805   -.7444789
      B2 |   -.287682   .4409585     -0.652   0.514      -1.151945    .5765808
      B3 |   .9490806   .3399873      2.792   0.005       .2827177    1.615444
      B4 |   1.228665   .3282127      3.744   0.000       .5853803     1.87195
      C2 |   .3364722     .29277      1.149   0.250      -.2373465    .9102909
      C3 |   .0487902   .3124405      0.156   0.876      -.5635819    .6611622
      C4 |   .1823215    .302765      0.602   0.547      -.4110871      .77573
   _cons |   .9480394   .3501152      2.708   0.007       .2618263    1.634253
------------------------------------------------------------------------------
```

3. Model:
log-frequency = CHD + BP + CHL + CHD*BP + BP*CHL

```
. loglin pop  chd bp chl, fit(chd, bp, chl, chd bp,  bp chl)
Variable chd = A
Variable bp = B
Variable chl = C
Margins fit: chd, bp, chl, chd bp, bp chl
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -90.335449
Iteration 1: Log Likelihood = -87.391113
Iteration 2: Log Likelihood = -87.318848

Poisson regression                                 Number of obs    =       32
Goodness-of-fit chi2(12)     =    24.060           Model chi2(19)   =1620.167
Prob > chi2                  =    0.0200           Prob > chi2      =  0.0000
Log Likelihood               =   -87.319           Pseudo R2        =  0.9027

------------------------------------------------------------------------------
     pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |   3.241941   .2942619     11.017   0.000       2.665198    3.818684
    AB22 |   .0661659   .4491836      0.147   0.883      -.8142177    .9465496
    AB23 |  -.5914288   .3480317     -1.699   0.089      -1.273558    .0907007
    AB24 |  -1.454255   .3392078     -4.287   0.000       -2.11909   -.7894199
      B2 |  -.3655411   .4551419     -0.803   0.422      -1.257603    .5265207
      B3 |   .6266025    .355817      1.761   0.078       -.070786    1.323991
      B4 |   .8628062   .3507258      2.460   0.014       .1753963    1.550216
    BC22 |   .0866753   .1945032      0.446   0.656       -.294544    .4678946
    BC23 |    .173953   .2499885      0.696   0.487      -.3160155    .6639215
    BC24 |   .1791843    .318915      0.562   0.574      -.4458776    .8042462
    BC32 |   .5082825   .1699628      2.991   0.003       .1751616    .8414034
    BC33 |   .3269785   .2231387      1.465   0.143      -.1103654    .7643223
    BC34 |   .5686602   .2741296      2.074   0.038        .031376    1.105944
    BC42 |   .3643072   .1974599      1.845   0.065       -.022707    .7513215
    BC43 |   .6060868   .2438457      2.486   0.013        .128158    1.084016
    BC44 |   1.001152   .2882805      3.473   0.001       .4361323    1.566171
      C2 |   .0411581   .1283272      0.321   0.748      -.2103586    .2926748
      C3 |  -.8671005   .1685329     -5.145   0.000      -1.197419    -.536782
      C4 |  -1.521027    .216483     -7.026   0.000      -1.945326   -1.096728
   _cons |   1.498839   .2976597      5.035   0.000       .9154366    2.082241
------------------------------------------------------------------------------
```

4. Model:
log-frequency = CHD + BP + CHL + CHD*CHL + BP*CHL

```
. loglin pop  chd bp chl, fit(chd, bp, chl, chd  chl,  bp chl)
Variable chd = A
Variable bp = B
Variable chl = C
Margins fit: chd, bp, chl, chd chl, bp chl
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -92.934082
Iteration 1: Log Likelihood = -90.51709
Iteration 2: Log Likelihood = -90.491211

Poisson regression                                    Number of obs    =      32
Goodness-of-fit chi2(12)    =    30.404               Model chi2(19)   =1613.822
Prob > chi2                 =    0.0024               Prob > chi2      =  0.0000
Log Likelihood              =   -90.491               Pseudo R2        =  0.8992

------------------------------------------------------------------------------
     pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |   2.965273   .2292974     12.932   0.000       2.515858    3.414688
    AC22 |  -.0302771   .3003151     -0.101   0.920      -.6188838    .5583296
    AC23 |  -.6916755   .3241887     -2.134   0.033      -1.327074   -.0562773
    AC24 |  -1.372642   .3204974     -4.283   0.000      -2.000806    -.744479
      B2 |  -.3017864   .1405952     -2.146   0.032      -.5773479    -.026225
      B3 |   .0650637   .1275828      0.510   0.610       -.184994    .3151215
      B4 |  -.4750581   .1480435     -3.209   0.001       -.765218   -.1848982
    BC22 |   .0866753   .1945032      0.446   0.656      -.2945441    .4678946
    BC23 |   .1739532   .2499885      0.696   0.487      -.3160153    .6639218
    BC24 |   .1791842    .318915      0.562   0.574      -.4458778    .8042461
    BC32 |   .5082822   .1699628      2.991   0.003       .1751612    .8414031
    BC33 |   .3269786   .2231387      1.465   0.143      -.1103653    .7643224
    BC34 |   .5686601   .2741297      2.074   0.038       .0313758    1.105944
    BC42 |   .3643068   .1974599      1.845   0.065      -.0227074     .751321
    BC43 |   .6060866   .2438457      2.486   0.013       .1281577    1.084015
    BC44 |   1.001151   .2882806      3.473   0.001       .4361317    1.566171
      C2 |   .0699295   .3129367      0.223   0.823      -.5434151    .6832742
      C3 |  -.2231435   .3451478     -0.647   0.518      -.8996207    .4533337
      C4 |  -.2832652   .3592191     -0.789   0.430      -.9873216    .4207912
   _cons |   1.763588   .2365426      7.456   0.000       1.299974    2.227203
------------------------------------------------------------------------------
```

5. Model:

log-frequency = CHD + BP + CHL + CHD*BP

```
. loglin pop  chd bp chl, fit(chd, bp, chl, chd bp)
Variable chd = A
Variable bp = B
Variable chl = C
Margins fit: chd, bp, chl, chd bp
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -102.75049
Iteration 1: Log Likelihood = -99.592285
Iteration 2: Log Likelihood = -99.542969
Poisson regression                                    Number of obs    =      32
Goodness-of-fit chi2(21)    =    48.508               Model chi2(10)   =1595.719
Prob > chi2                 =    0.0006               Prob > chi2      =  0.0000
Log Likelihood              =   -99.543               Pseudo R2        =  0.8891
------------------------------------------------------------------------------
     pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |   3.241941   .2942627     11.017   0.000       2.665197    3.818685
    AB22 |   .0661658   .4491842      0.147   0.883      -.8142191    .9465508
    AB23 |  -.5914289   .3480323     -1.699   0.089       -1.27356    .0907019
    AB24 |  -1.454255   .3392085     -4.287   0.000      -2.119091   -.7894185
      B2 |   -.287682   .4409582     -0.652   0.514      -1.151944    .5765801
      B3 |   .9490806   .3399872      2.792   0.005        .282718    1.615443
      B4 |   1.228665   .3282125      3.744   0.000       .5853808     1.87195
      C2 |    .307701   .0652134      4.718   0.000       .1798852    .4355169
      C3 |  -.5951668   .0830387     -7.167   0.000      -.7579196   -.4324139
      C4 |   -1.05544   .0974332    -10.832   0.000      -1.246406   -.8644745
   _cons |    1.30324    .291603      4.469   0.000       .7317082    1.874771
------------------------------------------------------------------------------
```

6. Model:
log-frequency = CHD + BP + CHL + CHD*CHL

```
. loglin pop  chd bp chl, fit(chd, bp, chl, chd chl)
Variable chd = A
Variable bp = B
Variable chl = C
Margins fit: chd, bp, chl, chd chl
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -105.66895
Iteration 1: Log Likelihood = -102.74902
Iteration 2: Log Likelihood = -102.71582

Poisson regression                                  Number of obs    =      32
Goodness-of-fit chi2(21)    =    54.854             Model chi2(10)   =1589.373
Prob > chi2                 =    0.0001             Prob > chi2      =  0.0000
Log Likelihood              =  -102.716             Pseudo R2        =  0.8855

------------------------------------------------------------------------------
     pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |   2.965273   .2292974     12.932   0.000       2.515858    3.414688
    AC22 |  -.0302771   .3003151     -0.101   0.920      -.6188838    .5583296
    AC23 |  -.6916755   .3241887     -2.134   0.033      -1.327074   -.0562774
    AC24 |  -1.372642   .3204974     -4.283   0.000      -2.000806   -.7444789
      B2 |  -.2239275   .0840022     -2.666   0.008      -.3885687   -.0592862
      B3 |   .3875417   .0725428      5.342   0.000       .2453604    .5297229
      B4 |  -.1091992   .0814328     -1.341   0.180      -.2688045    .0504062
      C2 |   .3364723     .29277      1.149   0.250      -.2373464     .910291
      C3 |   .0487903   .3124404      0.156   0.876      -.5635818    .6611623
      C4 |   .1823216    .302765      0.602   0.547       -.411087    .7757302
   _cons |   1.567989   .2288731      6.851   0.000       1.119406    2.016572
------------------------------------------------------------------------------
```

7. Model:

log-frequency = CHD + BP + CHL + BP*CHL

```
. loglin pop  chd bp chl, fit(chd, bp, chl, bp chl)
Variable chd = A
Variable bp = B
Variable chl = C
Margins fit: chd, bp, chl, bp chl
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -108.30371
Iteration 1: Log Likelihood = -102.64697
Iteration 2: Log Likelihood = -102.54492


Poisson regression                                  Number of obs    =      32
Goodness-of-fit chi2(15)    =    54.512             Model chi2(16)   =1589.715
Prob > chi2                 =    0.0000             Prob > chi2      =  0.0000
Log Likelihood              =  -102.545             Pseudo R2        =  0.8857
```

```
------------------------------------------------------------------------------
     pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |   2.587845   .1075225     24.068   0.000       2.377105    2.798585
      B2 |  -.3017865   .1405951     -2.146   0.032       -.577348   -.0262251
      B3 |   .0650634   .1275828      0.510   0.610      -.1849944    .3151211
      B4 |  -.4750583   .1480435     -3.209   0.001      -.7652182   -.1848983
    BC22 |    .086675   .1945032      0.446   0.656      -.2945443    .4678943
    BC23 |   .1739533   .2499885      0.696   0.487      -.3160152    .6639218
    BC24 |   .1791841    .318915      0.562   0.574      -.4458778    .8042461
    BC32 |   .5082823   .1699628      2.991   0.003       .1751613    .8414032
    BC33 |   .3269788   .2231387      1.465   0.143      -.1103651    .7643227
    BC34 |   .5686604   .2741297      2.074   0.038       .0313761    1.105945
    BC42 |   .3643067   .1974599      1.845   0.065      -.0227075     .751321
    BC43 |   .6060867   .2438457      2.486   0.013       .1281578    1.084015
    BC44 |   1.001151   .2882806      3.473   0.001       .4361318    1.566171
      C2 |   .0411582   .1283272      0.321   0.748      -.2103586    .2926749
      C3 |  -.8671007    .168533     -5.145   0.000      -1.197419   -.5367821
      C4 |  -1.521027    .216483     -7.026   0.000      -1.945326   -1.096728
   _cons |   2.118789   .1356619     15.618   0.000       1.852896    2.384681
------------------------------------------------------------------------------
```

8. Model:

log-frequency = CHD + BP + CHL

```
. loglin pop  chd bp chl, fit(chd, bp, chl)
Variable chd = A
Variable bp = B
Variable chl = C
Margins fit: chd, bp, chl
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -121.24414
Iteration 1: Log Likelihood = -114.89258
Iteration 2: Log Likelihood = -114.76953
Iteration 3: Log Likelihood = -114.76904

Poisson regression                                    Number of obs    =      32
Goodness-of-fit chi2(24)      =    78.960             Model chi2(7)    =1565.267
Prob > chi2                   =    0.0000             Prob > chi2      =  0.0000
Log Likelihood                =  -114.769             Pseudo R2        =  0.8721

------------------------------------------------------------------------------
     pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |   2.587845   .1075225     24.068   0.000       2.377105    2.798585
      B2 |  -.2239276   .0840022     -2.666   0.008      -.3885688   -.0592864
      B3 |   .3875415   .0725428      5.342   0.000       .2453602    .5297227
      B4 |  -.1091995   .0814328     -1.341   0.180      -.2688048    .0504059
      C2 |    .307701   .0652134      4.718   0.000       .1798851    .4355168
      C3 |  -.5951669   .0830387     -7.167   0.000      -.7579197    -.432414
      C4 |   -1.05544   .0974332    -10.832   0.000      -1.246406   -.8644745
   _cons |   1.923189   .1217978     15.790   0.000        1.68447    2.161909
------------------------------------------------------------------------------
```

Summary results:

Denoted CHD = 1, BP = 2, and CHL = 3.

Model	*Likelihood ratio Chi-square*	*Degree of freedom*	*p-value*
All pairwise association			
1. $u_{123 = 0}$	**4.775**	**9**	**0.8534**
Conditional independence			
2. $u_{23} = u_{123} = 0$	**24.401**	**18**	**0.142**
3. $u_{13} = u_{123} = 0$	**24.060**	**12**	**0.020**
4. $u_{12} = u_{123} = 0$	**30.404**	**12**	**0.002**
Partial independence			
5. $u_{12} = u_{13} = u_{123} = 0$	**54.512**	**15**	**<0.001**
6. $u_{12} = u_{23} = u_{123} = 0$	**54.854**	**21**	**<0.001**
7. $u_{13} = u_{23} = u_{123} = 0$	**48.508**	**21**	**<0.001**
Complete independence			
8. $u_{12} = u_{13} = u_{23} = u_{123} = 0$	**78.906**	**24**	**<0.001**

Followings are explanations of model selection. These are for illustration only, not for quoted in the research report.

Examine all models

So G^2 is non-significant so we do not need the three-way interaction.

How many two-way interaction do we need?

If we examine the fitted values for the parameters in relation to their standard errors (called standardized values) in the Stata output with "resid" option, we can determine which interaction terms can be discarded. Since only terms involving interaction between variables 1 and 3 and 2 and 3 are significantly different from zero we can omit all but these interactions. Thus model 2 looks promising.

Considering Model 1 - No thee-way interaction

$H_0 : u_{123} = 0$

Expected values have to be obtained iteratively. When this is done obtain $G^2 = 4.77$

Parameter	No.	This problem
u	1	1
$u_{1(i)}$	r-1	3
$u_{2(j)}$	c-1	3
$u_{3(k)}$	l-1	1
$u_{12(ij)}$	(r-1)(c-1)	9
$u_{13(ik)}$	(r-1)(l-1)	3
$u_{23(jk)}$	(c-1)(l-1)	3
Total		23

df = 4 x 4 x 2 - 23 = 32 - 23 = 9

P-value = 0.8534, thus we conclude that the model fits extremely well to the data.

Considering model 8: The independent model
Main effects.

H_0 : three variables are mutually independent or
$H_0 : u_{12} = u_{13} = u_{23} = u_{123} = 0$

Expected values are calculated as

$$\hat{e}_{ijk} = \frac{n_{i++}n_{+j+}n_{++k}}{n^2_{+++}}$$

$$\chi^2_p = 99.54 \quad G^2 = 78.96$$

$$\begin{aligned} df &= 4 \text{ x } 4 \text{ x } 2 - (4 - 1) - (4-1) - (2 - 1) - 1 \\ &= 32 - 3 - 3 - 1 - 1 \\ &= 24 \end{aligned}$$

Since χ^2_p and G^2 are highly significant we reject H_0 and conclude that the model does not provide and adequate fit.

Considering model 2: The conditional independence model

$$\ln e_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)}$$

This is the conditional independence model which implies no association between blood pressure (BP) and serum cholesterol level (CHL) for both CHD and no CHD patients. But BP and CHL each is associated with CHD.
G^2 = 24.4 with 18 *df*. gives a p-value of 0.142. So we conclude that this model provides an adequate fit to the data.

Selection of the final model

Comparing the two models that adequately fitted the data (Models 1 and 2) we have

$$\begin{aligned} G^2_2 - G^2_1 &= 24.4 - 4.77 = 19.63 \\ \text{with } 18 - 9 &= 9 \; df \end{aligned}$$

Use Stata to find a p-value

```
. disp chiprob(9, 19.63)
.02033827
```

Since this is significant (p-value = 0.02) we conclude that the addition of the parameter *u*$_{23}$ to model 2 causes a significant improvement in fit and consequently a model which includes two-way interactions between all pairs of variables is needed. Thus Model 1 is the best model for describing the data. We can examine the residuals of the model using the same command of Stata as that being used previously plus an option - "resid" as follows:

```
    . loglin pop  chd bp chl, fit(chd, bp, chl, chd bp, chd chl, bp
    chl resid
Variable chd = A
Variable bp = B
Variable chl = C
Margins fit: chd, bp, chl, chd bp, chd chl, bp chl
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0:   log likelihood = -172.80125
Iteration 1:   log likelihood = -78.652413
Iteration 2:   log likelihood = -77.677338
Iteration 3:   log likelihood = -77.676644
Iteration 4:   log likelihood = -77.676644

Poisson regression                                Number of obs   =         32
                                                  LR chi2(22)     =    1639.45
                                                  Prob > chi2     =     0.0000
Log likelihood = -77.676644                       Pseudo R2       =     0.9134

------------------------------------------------------------------------------
     pop |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |   3.495052   .3489747     10.015   0.000       2.811075     4.17903
    AB22 |   .0913573   .4512986      0.202   0.840      -.7931717    .9758862
    AB23 |  -.5623115   .3508082     -1.603   0.109      -1.249883    .1252599
    AB24 |  -1.342433   .3429665     -3.914   0.000      -2.014635   -.6702307
    AC22 |   .0384452    .303493      0.127   0.899      -.5563903    .6332806
    AC23 |  -.5872324   .3285031     -1.788   0.074      -1.231087    .0566217
    AC24 |  -1.203873   .3265999     -3.686   0.000      -1.843997   -.5637491
      B2 |  -.3905691     .46064     -0.848   0.397      -1.293407    .5122686
      B3 |   .6053894   .3613974      1.675   0.094      -.1029365    1.313715
      B4 |   .7873584   .3578944      2.200   0.028       .0858984    1.488819
    BC22 |   .0865808   .1945052      0.445   0.656      -.2946424    .4678041
    BC23 |   .1759085   .2501767      0.703   0.482      -.3144288    .6662459
    BC24 |   .1846617   .3200576      0.577   0.564      -.4426397     .811963
    BC32 |   .5090796   .1700817      2.993   0.003       .1757255    .8424336
    BC33 |   .3106373    .223539      1.390   0.165       -.127491    .7487656
    BC34 |   .5236173   .2756988      1.899   0.058      -.0167425    1.063977
    BC42 |    .367129   .1987235      1.847   0.065      -.0223619    .7566198
    BC43 |   .5494595   .2463114      2.231   0.026        .066698    1.032221
    BC44 |   .8502137   .2934776      2.897   0.004       .2750082    1.425419
      C2 |   .0038244   .3214342      0.012   0.991      -.6261751    .6338238
      C3 |  -.3031287   .3571926     -0.849   0.396      -1.003213    .3969559
      C4 |  -.3836105   .3751534     -1.023   0.307      -1.118898    .3516766
   _cons |   1.254175   .3508825      3.574   0.000       .5664583    1.941892
------------------------------------------------------------------------------
```

```
pop  chd   bp  chl   cellhat    resid   stdres
  2    1    1    1     3.505   -1.505   -0.804
  3    1    1    2     3.518   -0.518   -0.276
  3    1    1    3     2.588    0.412    0.256
  4    1    1    4     2.388    1.612    1.043
  3    1    2    1     2.372    0.628    0.408
  2    1    2    2     2.596   -0.596   -0.370
  1    1    2    3     2.088   -1.088   -0.753
  3    1    2    4     1.944    1.056    0.758
  8    1    3    1     6.421    1.579    0.623
 11    1    3    2    10.724    0.276    0.084
  6    1    3    3     6.469   -0.469   -0.185
  6    1    3    4     7.386   -1.386   -0.510
  7    1    4    1     7.702   -0.702   -0.253
 12    1    4    2    11.162    0.838    0.251
 11    1    4    3     9.854    1.146    0.365
 11    1    4    4    12.282   -1.282   -0.366
117    2    1    1   115.495    1.505    0.140
121    2    1    2   120.482    0.518    0.047
 47    2    1    3    47.412   -0.412   -0.060
 22    2    1    4    23.612   -1.612   -0.332
 85    2    2    1    85.628   -0.628   -0.068
 98    2    2    2    97.404    0.596    0.060
 43    2    2    3    41.912    1.088    0.168
 20    2    2    4    21.056   -1.056   -0.230
119    2    3    1   120.579   -1.579   -0.144
209    2    3    2   209.276   -0.276   -0.019
 68    2    3    3    67.531    0.469    0.057
 43    2    3    4    41.614    1.386    0.215
 67    2    4    1    66.298    0.702    0.086
 99    2    4    2    99.838   -0.838   -0.084
 46    2    4    3    47.146   -1.146   -0.167
 33    2    4    4    31.718    1.282    0.228
```

Summarize findings

There is a positive association between high blood pressure (level 4 of BP) and CHD and a positive association between high cholesterol (level 4 of CHL) and CHD. Low levels of each of these are 'protective' (i.e., negative coefficients).

The lack of a three-way interaction implies that:

a) interaction between CHD and BP is the same at all levels of serum CHL.

b) interaction between CHD and CHL is the same at all levels of BP.

7.6 Further readings

A good and comprehensive book is Agresti (1991); page 130-153. Another more practical book is given by Selvin (1995); page 293-364. All key concepts can be found in these books.

Selvin (1995) provided the simplest possible introduction to the log-linear model by applying the concept of log-linear model for 2-by-2 table (page 307-314), and for R-by-C table (page 314). The author demonstrated that the log-linear model applied to a 2-by-2 table produces the same estimate values, the same chi-square statistics, and the same results as that described in Chapter 2. Agresti (1991); page 133-134 had shown that when it is natural to regard one variable as a response and other as explanatory variables, certain log-linear models are equivalent to logistic regression model which had discussed in Chapter 6. Upton (1998) demonstrated that log-linear model can be used as a tool for exploratory data analysis. It serves as a useful guide for more complicated statistical modeling. The author also provided a comprehensive review of key concepts of this method that worth reading.

A closely related topic is capture-recapture model and Poisson regression. A readable introduction and practical example of this topic can be found in Selvin (1995); page 342-349, and page 455-488, respectively.

Chapter references

Agresti, A. (1990). ***Categorical data analysis*****. New York: John Wiley & Sons.**

CCEB (Centre for Clinical Epidemiology and Biostatistics).

The University of Newcastle, Australia. (1993) *STAT402:Analysis of categorical data - Log-linear model.* Newcastle: The University of Newcastle, NSW. Australia.

Everitt, B.S. The Analysis of Contingency Tables. Chapman and Hall, London, 1977

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. 2nd edition. New York: John Willey & Sons.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

Upton G.J.G. (1991) The exploratory analysis of survey data using log-linear models. *The Statistician*. 40; 169-82.

Exercise

Following data is from Everitt (1977); page 73. Your are assigned to find the 'best' log-linear model to describe the associations between adversity of school condition and home condition and deviant behavior in the classroom using the following data and summarize your findings.

		Adversity of school condition						Total
		Low		Medium		High		
Risk index		Not at risk	At risk	Not at risk	At risk	Not at risk	At risk	
Behavior	Not deviant	16	7	15	34	5	3	80
	Deviant	1	1	3	8	1	3	17
Total		17	8	18	42	6	6	97

Chapter 8

Special Topics for Categorical Data Analysis

Chapter Objectives

After completion this chapter, readers should be able to:

- **describe concepts underlying tests with continuity correction for 2-by-2 Table**
- **calculate exact p-value**
- **describe how odds ratio is an estimator of relative risk and when to be cautions**
- **describe basic concepts of analysis of categorical data from survey data**

8.1 Tests with continuity correction for 2-by-2 Table

In deriving the distribution of the chi-square statistic essentially we are employing a continuous probability distribution, namely the chi-square distribution as an approximation to the discrete probability distribution of frequencies (eg, the multinomial distribution). To improve the approximation, Yates (1934) suggested a correction. However, several authors questioned appropriateness of the continuity correction. Agresti (1990) provided a comprehensive summary on page 68 as well as Daniel (1991) on page 548-549, and StataCorp (1999) on page 406 of volume 1:A-G. All of them suggested not to use it, rather, use Fisher's exact test when in doubt of insufficient sample.

8.2 Exact methods

Exact methods are for small sample. A contingency table where the number of cells with expected value of less than 5 greater than 20% of the total number of cell is said to be small sample. In this case, asymptotic methods are not valid. Chi-square test is an asymptotic method, so not appropriate. For 2-by-2 contingency Table, Fisher's exact test is the appropriate one. For a table larger than 2-by-2 Table, the equivalent exact test is called Freeman-Halton Conditional Exact Test. Agresti (1990) provided a comprehensive review on page 59-67. Stata can calculate for these even for fairly large sample and for a table larger than 2-by-2 Table as shown below.

Example 8.1 For 2-by-2 Table

Altman (1991); page 254-256 illustrated calculation manually the Fisher's exact test using the data in the table below.

Table 8.1 Number of subjects by spectacle wearing status by juvenile delinquents status - data for example 8.1

Spectacle	Juvenile delinquents		
wearers	Yes	No	Total
Yes	1	5	6
No	8	2	10
Total	9	7	16

Here is an immediate form of "tabulate" command (see StataCorp., 1999; Volume 4: Su-Z, page 157-174).

```
. tabi 1 5 \ 8 2, exact

           |          col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |         1          5 |         6
         2 |         8          2 |        10
-----------+----------------------+----------
     Total |         9          7 |        16

           Fisher's exact =                 0.035
   1-sided Fisher's exact =                 0.024
```

Alternatively we can use either "csi" or "cci" command with "exact" option (see StataCorp., 1999; Volume 1: A-G, page 366-414).

Example 8.2 For 2-by-5 Table

A hypothetical data adapted form Example 4.1.

Table 8.2 Number of subjects by type of psychiatric disorder by blood group - data for example 8.2

Psychiatric disorder	Blood group				Total
	A	B	AB	O	
Schizophrenia	7	2	1	28	18
Neurosis	1	2	5	7	15
Depressed	4	3	1	1	9
Total	12	7	7	16	42

```
. tabi 7 2 1 8 \ 1 2 5 7 \ 4 3 1 1, chi2 exact

           |                    col
       row |         1          2          3          4 |     Total
-----------+--------------------------------------------+----------
         1 |         7          2          1          8 |        18
         2 |         1          2          5          7 |        15
         3 |         4          3          1          1 |         9
-----------+--------------------------------------------+----------
     Total |        12          7          7         16 |        42

          Pearson chi2(6) =  12.1167   Pr = 0.059
           Fisher's exact =                 0.043
```

Note that Pearson chi-square leads to non-significant result while exact test is significant. The exact test is preferred in this example due to small sample. The larger the sample size, the closer the p-values from asymptotic and exact tests.

However the above examples are the test statistics. As this book had advocated estimation-based approach throughout, some references for estimating confidence intervals recently published were provided as follows:

- Brenner and Quan (1990) : exact confidence limit for binomial proportions

- Agresti (1999) : confidence intervals for the odds ratio with small sample

- **Hirji (1994) : exact analysis for pair binary data**
- **Korn and Glaubard (1998) : exact confidence intervals for proportions from survey data**

Interesting arguments of exact methods for binomial proportions were given by Agresti and Coull (1998).

These references were also full of other references at the end of their papers. Stata provides some of these (see StataCorp., 1999; Volume 1: A-G, page 366-414). Again StataCorp (1999) provides useful references for the exact estimations as well.

8.3 Odds ratio (OR) as an estimator of relative risk (RR)

Simple mathematical proves that Odds ratio is an approximate relative risk was given clearly in Everitt (1977); page 31-33. Armitage and Berry (1994); page 509.

Davies, Crombie, and Tavakoli (1998) showed that if the odds ratio is interpreted as a relative risk it will always overstate any effect size: the odds ratio is smaller than the relative risk for odds ratios of less than one, and bigger than the relative risk for odds ratios of greater than one. The authors further demonstrated that the extent of overstatement increases as both the initial risk increases and the odds ratio departs from unity. However, serious divergence between the odds ratio and the relative risk occurs only with large effects on groups at high initial risk. Therefore qualitative judgments based on interpreting odds ratios as though they were relative risks are unlikely to be seriously in error. In studies which show reductions in risk (odds ratios of less than one), the odds ratio will never underestimate the relative risk by a greater

percentage than the level of initial risk. In studies which show increases in risk (odds ratios of greater than one), the odds ratio will be no more than twice the relative risk so long as the odds ratio times the initial risk is less than 100%.

Lee (1999) provided simple methods for checking for possible errors in reported odds ratios, relative risk and confidence intervals.

8.4 Analysis of categorical data from survey data

Survey data generally have some special characteristics different from other type of study design such as sampling probability (i.e., sampling weight), clustering, and stratification. These were arise from the design of data collection procedure. An excellent and precise description how these designs affect the analysis of data is given by StataCorp (1999); page 321-333 of Stata user's guide. Commands of Stata for these type of data can be found in StataCorp (1999); page 15-99 of Volume 4: Su-Z. Most of the commands can be used in the same ways as those had been illustrated in previous chapters but started with "svy" which stand for "survey". For example, "svytab" is an equivalent of "tabulate" ordinary command mostly used in the previous chapters.

Chapter references

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

Agresti, A. (1999). On logit confidence intervals for the odds ratios with small samples. *Biometrics*. 55: 597-602.

Agresti, A., and Coull, B.A. (1998). Approximate is better than "exact" intervals estimation of binomial proportions. *Am.*

Stat. Assoc. 52(2): 199-126

Altman D.G. (1991). Practical statistics for medical research. London: Chapman and Hall.

Armitage, P., and Berry, G. (1994). _Statistical methods in medical research_. Oxford: Blackwell Scientific Publications.

Brenner, D.J., and Quan, H. (1990). Exact confidence limits for binomial proportions - Pearson and Hartley revisited. _The Statistician_. 39: 391-397.

Daniel, W.W. (1991). _Biostatistics: A foundation for analysis in the health sciences_. 5th ed. New York: John Wiley & Sons.

Davies, H.T.O., Crombie, I.K., and Tavakoli, M. (1998). When can odds ratios mislead? _BMJ_ 316:989-991

Everitt, B.S. (1977). _The Analysis of Contingency Tables_.London: Chapman and Hall.

Hirji, K.F. (1994). Exact analysis for paired binary data. _Biometrics_. 50, 964-974.

Korn, E.L., and Graubard, B.I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. _Survey Methodology_. 24(2): 193-201

Lee, P.N. (1999). Simple methods for checking for possible errors in reported odds ratios, relative risk and confidence intervals. _Stat. Med_. 18: 1973-1981

StataCorp. (1999). _Stata statistical software: Release 6.0_. College Station. TX: Stata Corporation.

BIBLIOGRAPHY

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley & Sons.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.

Agresti, A. (1999). On logit confidence intervals for the odds ratios with small samples. *Biometrics*. 55: 597-602.

Agresti, A., and Coull, B.A. (1998). Approximate is better than "exact" intervals estimation of binomial proportions. *Am. Stat. Assoc*. 52(2): 199-126

Altman, D.G. (1991). *Practical statistics for medical research*. London: Chapman and Hall.

Armitage, P., and Berry, G. (1994). *Statistical methods in medical research*. Oxford: Blackwell Scientific Publications.

Bland, M. (1998). *An introduction of medical statistics*. Oxford: Oxford University Press.

Brenner, D.J., and Quan, H. (1990). Exact confidence limits for binomial proportions - Pearson and Hartley revisited. *The Statistician*. 39: 391-397.

Breslow, N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, 33 (l), 38-44.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88 (421), 9-25.

CCEB (Centre for Clinical Epidemiology and Biostatistics), The University of Newcastle, Australia. (1993). *STAT402:Analysis of categorical data - 2-by-2 Table.* Newcastle: The University of Newcastle, NSW.

Australia.

CCEB (Centre for Clinical Epidemiology and Biostatistics), The University of Newcastle, Australia. (1993). *STAT402:Analysis of categorical data - Logistic regression.* Newcastle: The University of Newcastle, NSW. Australia.

CCEB (Centre for Clinical Epidemiology and Biostatistics), The University of Newcastle, Australia. (1993). *STAT402:Analysis of categorical data - Log-linear model.* Newcastle: The University of Newcastle, NSW. Australia.

Conover, W. J. (1980). *Practical Nonparametric Statistics*. 2nd ed. New York: John Wiley & Sons.

Daniel, W.W. (1991). *Biostatistics: A foundation for analysis in the health sciences*. 5th ed. New York: John Wiley & Sons.

Daniel, W. W. (1995). *Biostatistics: A foundation for analysis in the health sciences*. 6th ed. New York: John Wiley & Sons.

Davies, H.T.O., Crombie, I.K., Tavakoli, M. (1998). When can odds ratios mislead? *BMJ*. 316:989-991

Dawson-Sounder, B., and Trapp, R.G. (1990). *Basic and clinical biostatistics*. Norwalk, Connecticut: Appleton & Lange.

Diggle, P.J., Liang, K-Y, and Zeger, S.L. (1994). *Analysis of longitudinal data*. New York: Oxford University Press.

Dunn, G., and Everitt, B. (1995). *Clinical biostatistics: An Introduction to evidence base medicine*. London: Edward Arnold.

Everitt. B.S. (1994). *Statistical methods for medical investigations*. 3rd edition. London: Edward Arnold.

Everitt, B. S. (1995). The analysis of repeated measures: A practical review with examples. *The Statistician*, 44(l), 113-135.

Everitt, B.S. (1977). *The Analysis of Contingency Tables*.London: Chapman and Hall.

Fienberg, S.E. (1980). *The analysis of cross-classified categorical data*. 2nd edition. Cambridge: The MIT Press.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. 2nd edition. New York: John Willey & Sons.

Friedman, L.M., Furberg, S.C.D., and DeMets, D.L. (1996). *Fundamental of clinical trials*. 3rd edition. St.Louis: Mosby.

Glantz SA. (1992). *Primer of biostatistics*. 3rd edition. New York: McGraw-Hill, Inc.

Goodman, L.A. (1965). On simultaneous confidence intervals for multinomial proportions. *Technometrics.* 7:247-254.

Guyatt, G., Jaeschke, R., Heddle, N., Cook, D., Shannon, H., and Walter S. (1995). Interpreting study results: confidence intervals. *Canadian Medical Association Journal.* 152:169-173.

Harper, R., and Reeves, B. (1999). Reporting of precision of estimates for diagnostic accuracy: a review. *BMJ*. 318: 1322-1323.

Harville, David A., & Mee, Robert W. (1984). A mixed-model procedure for analyzing ordered categorical data. *Biometrics*, 40:393-408.

Hirji, K.F. (1994). Exact analysis for paired binary data. *Biometrics*. 50:964-974.

Hosmer, D.W. Lemeshow, S. (1989). *Applied Logistic Regression*. New York: John Wiley & Sons.

Jaeschke, R., Guyatt, G., Shannon, H., Walter, S. Cook, D. Heddle, N. (1995). Assessing the effects of treatment: measures of association . *Canadian Medical Association Journal*. 152: 351-357

Kenward, M. G., Lesaffre, E., & Molenberghs, G. (1994). An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics,* 50:945-953.

Kleinbaum, D.G. (1994). *Logistic Regression: A self-learning text*. New York Springer-Verlag.

Kleinbaum, D.G. (1996). *Survival analysis: A self-learning text.* New York: Springer-Verlag.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. (1982). *Epidemiologic research: principles and qualitative methods*. London: Lifetime Learning Publications.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E, and Nizam, A. (1998). *Applied regression analysis and other multivariable methods*. Pacific Grove: Duxbury Press.

Knapp, RG, Miller, MC. (1992). *Clinical epidemiology and biostatistics.* Pensylvania: Harwal Publishing Company.

Korn, E.L., and Graubard, B.I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*. 24(2): 193-201

Laird, Nan M., & Ware, James H. (1982). Random effects models for longitudinal data. *Biometrics,* 38:963-974.

Landis, J.R. and Cock, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*.

33:159-174.

Lang, TA., Secic, M. (1997). ***How to report statistics in medicine: annotated guidelines for authors,*** **editors, and reviewers. Philadelphia: American College of Physician.**

Lee, P.N. (1999). Simple methods for checking for possible errors in reported odds ratios, relative risk and confidence intervals. ***Stat. Med*****. 18: 1973-1981**

Liang, K-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. ***Biometrika*** **73(l):13-22.**

Meinert, C.L. and Tonascia, S. (1986). ***Clinical trials, design, conduct, and analysis*****. New York: Oxford University Press.**

Rabe-Hesketh, S. and Everitt, B. (1998). ***A handbook of statistical analysis using Stata*****. London: Chapman & Hall/CRC.**

Richardson, J.T.E. (1994). The analysis of 2×1 and 2×2 contingency tables: an historical review. ***Statistical Methods in Medical Research*****. 3:107-133.**

Rosner, Bernard. (1995). ***Fundamentals of Biostatistics*****. 4th ed. Belmont, California: Duxbury Press.**

Rothman, K.J. and Greenland, S. (1998). ***Modern epidemiology*****. 2nd edition. Philadelphia: Lippincott-Raven Publishers.**

Schlesselman, J.J. (1982). ***Case-control studies: design conduct analysis*****. New York: Oxford University Press.**

Selvin, S. (1995). ***Practical biostatistical methods*****. Belmont: Duxbury Press.**

Senie, R.T., Rosen, P.P, Lesser, M.L., and Kinne, D.W. (1981). Breast self-examination and medical examination related to breast cancer stage. ***Am J Public Health*****. 71(6):583-90.**

Sribney, W. A. (1999). Comparison of different tests for trend.

Stata Corporation, http://www.stata.com/support/faqs/stat/ (connected at 12 December 1999).

Stasny, E.A. and Bauer, H.R. (1990). Symmetry and quasi-symmetry: an example in modeling pairs of sounds from children's early speech. *Stat. Med*. 9:1143-1155.

StataCorp. (1999). *Stata statistical software: Release 6.0*. College Station. TX: Stata Corporation.

Stiratelli, Robert, Laird, Nan, & Ware, James H. (1984). Random-effects models for serial observations with binary response. *Biometrics.* 40:961-971.

Stromberg, U. (1996). Collapsing ordered outcome categories: a note of concern. Am. J. Epidemiology 144(4): 421-424.

Upton G.J.G. (1991) The exploratory analysis of survey data using log-linear models. *The Statistician*. 40; 169-82.

Wonnacott, T.H., Wonnacott, R.J. (1990). *Introductory statistics*. 5th Edition. New York: John Willey & Sons.

Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika*, 75(4):621-629.

Zeger, S.L., and Karim, M.R. (1991). Generalized linear models with random effects: A Gibbs sampling approach. *J. Am. Stat. Assoc*. 86(413): 79-86.

Zeger, S.L., Liang, K-Y., and Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44:1049-1060.

APPENDIX

ANSWERS TO THE EXERCISES

Answers for the exercise in Chapter 2

Question 1.

i) It is a cross-sectional study.

The type of the study is a cross-sectional study since both gender and being on diet status were measured at the same time. What is known first is the grand total of 350 study subjects. Thus we need to note that the data in this table is the grand total only fixed.

Appropriate null hypothesis can be stated in two forms- general and specific forms. For the general form, we can state as follows:

Ho: There is no association between gender and being on diet status.
HA: There is an association between gender and being on diet status.

For the specific form, we can state as follows:

$$H_0 : \pi_{ij} = \pi_{i+}\ \pi_{+j} \quad \text{where } \pi_{i+} = \pi_{i1} + \pi_{i2} \quad ; i = 1,2.$$

$$\pi_{+j} = \pi_{1j} + \pi_{2j} \quad ; j = 1,2.$$

$$H_0 : \pi_{ij} \neq \pi_{i+}\ \pi_{+j}$$

This is the hypothesis of no association or independence.

ii) We can use the chi-square test

$$\chi^2 = \frac{\left(\left|14\times152-159\times25\right|\right)^2 350}{173\times177\times39\times311} = 3.21$$

Note that the smallest expected value is $\frac{39\times173}{350}$ = 19.3 so the chi-square test is appropriate.

We compare the value of with the chi-square distribution with 1 degree of freedom, the probability of observing a value as large or lager if H_0 is true is p-value > 0.05. Note that it is recommended that we should report the precise p-value instead of the "p-value > 0.05". One can use STATA to find the precise p-value using the "display" command. In this case we can do by executing the command as shown bellow and obtaining the p-value of 0.073.

```
. display chiprob(1, 3.21 )  →  command
. 0731895                    →  result
```

Thus the null hypothesis is not rejected. Therefore we have no sufficient evidence to conclude that there is an association between gender and diet.

iii) For a cross-sectional study we can calculate the relative risk or odds ratio as the measure of association (see Altman, 1991, page 266-269 for more details). The different between two proportions is not advisable here. Following is the example of calculation using the relative risk. Additionally, the attributable risk may also be useful in convincing policy makers.

We choose relative risk as a measure of association in this case assuming that SEX determine ON DIET STATUS or "column total fixed" table.

$$RR \quad = \quad \frac{\frac{14}{173}}{\frac{25}{177}} \quad = \quad \mathbf{0.57}$$

so boys are 0.57 times as likely to be on a diet as girls.

$$\textbf{Var}\ \ (ln\,\hat{RR}) \quad = \quad \left[\frac{159}{173 \times 14} + \frac{152}{177 \times 25}\right]$$

$$= \quad \mathbf{0.099999}$$

$$\textbf{95\%CI for ln(RR):}\ \ ln\left(0.57 \pm 1.96\sqrt{.099999}\right)$$

$$= \quad \textbf{ln(-1.18, 0.058)}$$

95% CI for RR= Exp(-1.18) to Exp(0.058)
= 0.307 to 1.059

iv) Among a total of 173 boys, 8.1% were on diet whilst among 177 girls, the corresponding rate was 14.1%. Boys as less likely to be on diets than girls (RR = 0.57; 95% CI: 0.31 to 1.06). However, this is not statistically significant (p-value = 0.073).

STATA commands:
The "csi" is an immediate command to estimate the relative risk which is the risk ratio in the output. It also provide the proportions to be reported as the descriptive components of the study results. The proportions are the column percents, assuming that SEX determine ON DIET STATUS or "column total fixed" table. If otherwise, i.e., one need to report the odds ratio, the "cci" command will be used. The proportions are

the row percents, assuming that ON DIET STATUS was known first then it was cross-classified by SEX or "row total fixed" table. Chi-square test and Fisher's exact test can also be estimated by the two commands. The output of both commands are as follows:

```
. csi 14 25 159 152

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        14          25  |         39
        Noncases |       159         152  |        311
-----------------+------------------------+----------
           Total |       173         177  |        350
                 |                        |
            Risk |   .0809249    .1412429 |   .1114286
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |        -.0603181       | -.1257701    .0051339
      Risk ratio |         .572948        |  .3082788    1.064846
 Prev. frac. ex. |         .427052        | -.0648459    .6917212
 Prev. frac. pop |        .2110857        |
                 +-----------------------------------------------
                             chi2(1) =     3.21  Pr>chi2 = 0.0730
```

By the above output, we quote 8.1% as the percentage of boy who were on diet and 14.1% as that for female. Then the RR (0.57) and 95%CI of RR (0.31 to 1.06). Finally, we quote the p-value (0.073).

```
. cci 14 25 159 152

                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        14          25  |         39       0.3590
        Controls |       159         152  |        311       0.5113
-----------------+------------------------+----------------------
           Total |       173         177  |        350       0.4943
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |        .5353459        |  .2710314    1.058345  (Cornfield)
 Prev. frac. ex. |        .4646541        | -.0583453    .7289686  (Cornfield)
 Prev. frac. pop |        .2375563        |
                 +-----------------------------------------------
                             chi2(1) =     3.21  Pr>chi2 = 0.0730
```

By the above output, we quote 35.9% as the percentage of boy among those who were on diet and 51.1% as that for female. Then the OR (0.53) and 95%CI of RR (0.27 to 1.06). Finally, we quote the p-value (0.073).

Note that both RR and OR are similar. It is the case when the event (ie. being on diet) is rare. In this case, it is 8% in boy and 14% in female.

Question 2.

The data can be summarized as a 2x2 table as follows.

Diet	Cause of death		Total
	CVD	Non CVD	
High Salt	5	2	7
Low Salt	30	23	53
Total	35	25	60

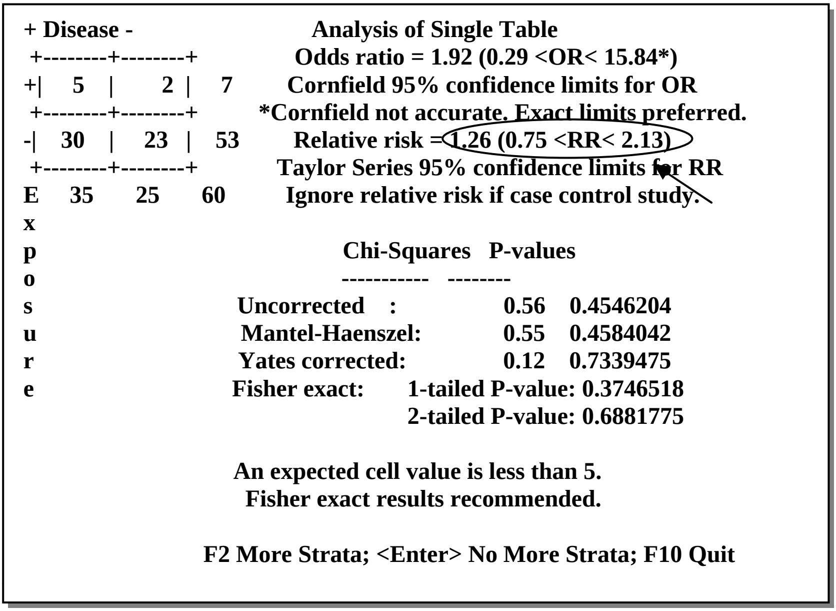
Firstly, the appropriate proportions should be determined. Based on the study design, the above table is "column total fixed". Thus the column percents are appropriate. Secondly, the measure of association should be estimated. Since this study cannot yield the incidence, therefore OR is appropriate. Thirdly, 95% CI of the OR should be calculated. Lastly, the hypothesis should be tested. The above statistics can be obtained by the same manner as that were done in Question 1. Details for estimating of 95% CI and the test hypothesis are provided as follows:

Estimating the 95% confidence interval :

The standard formula for estimating the CI was for large sample assuming normal distribution. In this study, the smallest expected value = = 2.92 which is too small to use the

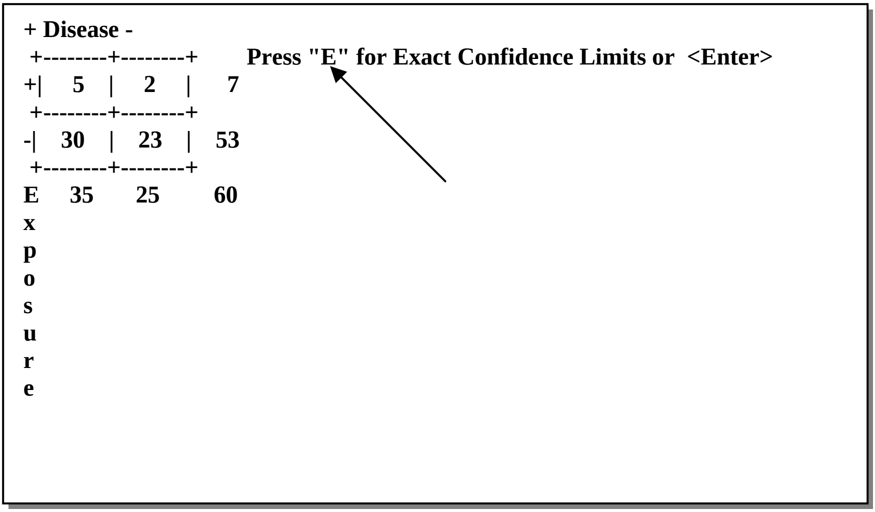
normal distribution. The exact CI is more appropriate. One method is proposed by Mehta, Patel, and Gray (1985). This yields the exact 95% confidence interval from 0.28 to 21.63 which can be easily calculated using "STATCALC" in "Epi Info" statistical package as shown bellow.

Step 1



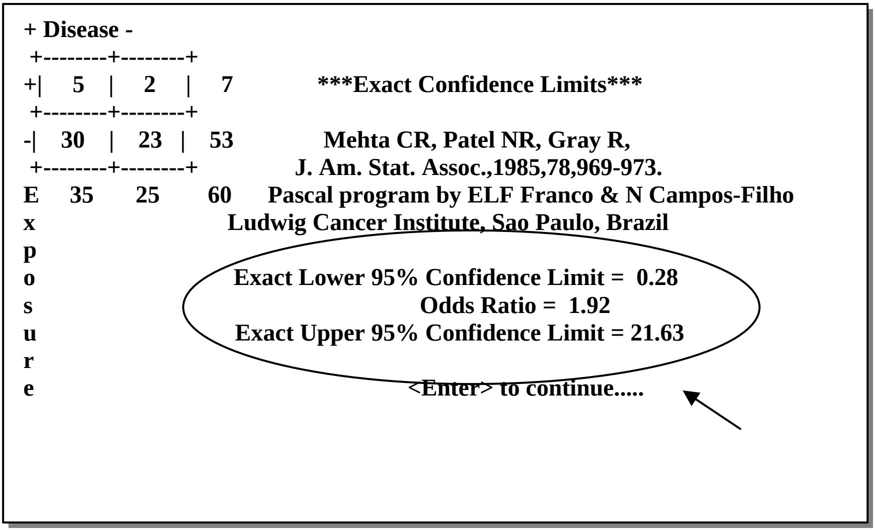
```
+ Disease -                      Analysis of Single Table
 +--------+--------+              Odds ratio = 1.92 (0.29 <OR< 15.84*)
+|   5   |     2 |    7         Cornfield 95% confidence limits for OR
 +--------+--------+          *Cornfield not accurate. Exact limits preferred.
-|  30   |   23  |   53         Relative risk = 1.26 (0.75 <RR< 2.13)
 +--------+--------+            Taylor Series 95% confidence limits for RR
E    35      25      60         Ignore relative risk if case control study.
x
p                                    Chi-Squares   P-values
o                                    -----------   --------
s                          Uncorrected    :             0.56    0.4546204
u                          Mantel-Haenszel:             0.55    0.4584042
r                          Yates corrected:             0.12    0.7339475
e                         Fisher exact:       1-tailed P-value: 0.3746518
                                              2-tailed P-value: 0.6881775

                          An expected cell value is less than 5.
                           Fisher exact results recommended.

                       F2 More Strata; <Enter> No More Strata; F10 Quit
```

Step 2



```
+ Disease -
 +--------+--------+      Press "E" for Exact Confidence Limits or  <Enter>
+|   5   |   2   |    7
 +--------+--------+
-|  30   |  23   |   53
 +--------+--------+
E   35     25      60
x
p
o
s
u
r
e
```

Step 3



```
+ Disease -
 +--------+--------+
+|   5   |   2   |   7          ***Exact Confidence Limits***
 +--------+--------+
-|  30   |  23  |   53           Mehta CR, Patel NR, Gray R,
 +--------+--------+          J. Am. Stat. Assoc.,1985,78,969-973.
E   35     25     60     Pascal program by ELF Franco & N Campos-Filho
x                     Ludwig Cancer Institute, Sao Paulo, Brazil
p
o                      Exact Lower 95% Confidence Limit =  0.28
s                                        Odds Ratio =  1.92
u                      Exact Upper 95% Confidence Limit = 21.63
r
e                                      <Enter> to continue.....
```

Testing the hypothesis :

Let π_1 be the proportion of men dying form CVD on a hi salt diet and π_2 be the corresponding proportion of men dying form other causes.

$$H_0 : \pi_1 = \pi_2$$

Note that the smallest expected value = $\frac{7 \times 25}{60}$ **= 2.92 which is too small to use the chi-square distribution. Thus we use Fishers' exact test**

Observed table	**p-value = 0.25**
One-tail	**p-value = 0.37**
Other-tail	**p-value = 0.31**
Two-tailed	**p-value = 0.69**

Thus the null hypothesis is not rejected and conclude that we have no sufficient evidence to concluded that the proportions of men aged 50-54 on a high salt diet dying from CVD and other causes are different.

STATA commands:

. cci 5 30 2 23, exact

```
                                                                     Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |         5          30  |         35      0.1429
        Controls |         2          23  |         25      0.0800
-----------------+------------------------+----------------------
           Total |         7          53  |         60      0.1167
```

```
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         1.916667       |  .3853557           .  (Cornfield)
 Attr. frac. ex. |         .4782609       | -1.595005           .  (Cornfield)
 Attr. frac. pop |          .068323       |
                 +-----------------------------------------------
                               1-sided Fisher's exact P = 0.3747
                               2-sided Fisher's exact P = 0.6882
```

Summarized findings:

Among a total of 35 CVD patients, 14.3% were high salt diet whereas there were 8.0% among 25 non-CVD patients. This case-control study failed to find a statistically significant relationship between high salt diet (p-value = 0.688) although it is suggested that those who had high salt diet are more likely to develop CVD than those who were not (OR = 1.91; 95% CI: 0.28 to 21.63).

Question 3.

i) This is a case-control study.

Let the proportion of cases using OC's be $_1$ and the corresponding proportion of controls be $_2$.

$$H_0 : \pi_1 = \pi_2$$

ii) The smallest expected value = 19.42.
So we can use the chi-square approximation.

$$\chi^2 = \frac{\left(\left|29\times1607-205\times2135\right|\right)^2 1976}{234\times1742\times164\times1812}$$

$$= \mathbf{5.84}$$

Comparing this with the chi-square distribution with 1 df., the probability of observing a value as large or lager if H_0 is true is $p < 0.05$. Thus we rejected H_0 and conclude that there is an association between oral contraceptive use and myocardial infarction since a statistically significantly larger proportion

of cases used OC's than controls. Note that it is recommended that we should report the precise p-value instead of the "p-value < 0.05". One can use STATA to find the precise p-value using the "display" command. In this case we can do by executing the command as shown bellow and obtaining the p-value of 0.157.

```
. display chiprob(1, 5.84 )   →   command
. 01566583                    →   result
```

iii) Since this is a case-control study we can calculate only an odds ratio.

$$OR = \frac{29 \times 1607}{205 \times 135} = 1.68$$

$$var(lnOR) = \frac{1}{29} + \frac{1}{205} + \frac{1}{135} + \frac{1}{1607} = 0.0474$$

95% CI for ln OR:

$$ln\left(1.68 \pm 1.96\sqrt{0.0474}\right) \quad = \quad ln(0.09, 0.95)$$

95% CI for OR: = 1.10 to 2.57

Thus cases had 1.7 times the odds of using OC's than controls.

iv) Among a total of 234 MI cases, 12.4% used OC whereas among 1742 controls there were 7.8%. This case-control study suggested a statistically significantly larger proportion of cases used OC's than controls (p-value = 0.016). The odds of a case using OC's were 1.7 times higher in cases than controls (95% CI: 1.10-2.57). If MI in women is

considered to be a rare disease then we could say that OC use increases the risk of MI 1.7 times.

STATA commands:

. cci 29 205 135 1607

```
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        29         205  |        234      0.1239
        Controls |       135        1607  |       1742      0.0775
-----------------+------------------------+----------------------
           Total |       164        1812  |       1976      0.0830
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         1.683939       |  1.102089    2.573581  (Cornfield)
 Attr. frac. ex. |         .4061541       |  .0926326    .6114364  (Cornfield)
 Attr. frac. pop |         .0503353       |
                 +-----------------------------------------------
                               chi2(1) =     5.84  Pr>chi2 = 0.0156
```

Question 4.

i) The null hypothesis :

H_0 : Proportion of cases exposed to factor E = proportion of controls exposed to factor E.

Note : This is always the underlying hypothesis being tested. Taking the study design into account and introducing some notation reduces the hypothesis to H_0 : $\pi_{12} = \pi_{21}$ and H_0 : $\pi = 0.5$ where π is the probability of the case being exposed and the control not exposed given that the pair is discordant.

ii) We use McNemar’s test to test the hypothesis. This test statistics has a chi-square distribution

$$\chi^2 = \frac{(20-5)^2}{20+5} = \frac{15^2}{25} = 9$$

Comparing this with the chi-square distribution with 1 df., the probability of observing a value as large or lager if H_0 is true is p-value < 0.005. Thus we reject Ho and conclude that a larger proportion of cases were exposed to factor E than controls.

In the situation where sample size is small, ie. $n_{12} + n_{21}$ is smaller than 20, we need to use the binomial exact probability test. However manually computation is tedious. We can easily obtain the exact probability using STATA. For this exercise, we can use an immediate form of the command “bitest” which is “bitesti” followed by n, x, and the probability of the event which is 0.5. The exact p-value is 0.004 as shown below.

```
. bitesti 25 20 0.5

        N   Observed k   Expected k   Assumed p   Observed p
------------------------------------------------------------
       25         20         12.5       0.50000      0.80000

  Pr(k >= 20)           = 0.002039  (one-sided test)
  Pr(k <= 20)           = 0.999545  (one-sided test)
  Pr(k <= 5 or k >= 20) = 0.004077  (two-sided test)
```

iii) $OR = \frac{20}{5} = 4$

The odds of case being exposed to factor E are 4 times the corresponding odds of a control being exposed. If one can assume that the disease being studied is rare then this can be

interpreted as the estimated relative risk being 4 so that people exposed to factor E are 4 times more likely to get the disease than those unexposed.

iv) 95% CI for

$$\mathbf{ln}\, OR = ln\left(4 \pm 1.96\sqrt{\frac{2}{20}+\frac{1}{5}}\right) = ln(0.406,\ 2.366)$$

95% CI for OR = 1.50 to 10.65

Note that the formula for the variance of lnOR used may not be appropriate due to the small cell frequency (ie. 5). An exact method can be used based on the exact CI for the binomial proportion representing the probability of falling into the n_{12} cell given the pair is discordant $(n_{12}/(n_{12}+n_{21}))$ **and then using the formula** $OR_L = \dfrac{\pi_L}{1-\pi_L}$ **and** $OR_u = \dfrac{\pi_u}{1-\pi_u}$ **to obtain a CI for OR.**

To calculate the exact 95% CI for x = 20 and n = 25 ,

from $\pi_L = [d - (d^2 - 4ae)^{1/2}]/2a$

$d = 25[2(20-1) + 3.84]$

$= 1046$

$a = 25[25+3.84]$

$= 721$

$e = [20-1]^2$

$= 361$

Thus $\pi_L = [1046-(1046^2 - (4x721x361))^{1/2}]/(2x721)$

$= 0.566$

from $\pi_U = [b+(b^2 - 4ac)^{1/2}]/2a$

where

$$b = 25[2(20+1)+3.84]$$
$$= 1146$$
$$a = 25[25+3.84]$$
$$= 721$$
$$c = [20+1]^2$$
$$= 441$$

Thus $\pi_U = [1146 + (1146^2 - (4x721x441))^{1/2}]/(2x721)$

$$= 0.936$$

So 95% CI for OR :

$$OR_L \frac{0.566}{1-0.566} = 1.30 \qquad OR_u = \frac{0.936}{1-0.936} = 14.62$$

That is we can be 95% certain that the true odds ratio lies between 1.30 and 14.62. This is slightly different from that obtained from STATA shown below which were due to rounding error.

STATA Commands

. mcci 15 20 5 60

```
                 | Controls               |
Cases            |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
         Exposed |        15          20  |         35
       Unexposed |         5          60  |         65
-----------------+------------------------+----------
           Total |        20          80  |        100

McNemar's chi2(1) =      9.00        Pr>chi2 = 0.0027
Exact McNemar significance probability       = 0.0041
```

```
Proportion with factor
        Cases            .35
        Controls          .2     [95% conf. interval]
                   ---------     --------------------
        difference       .15       .0465157   .2534843
        ratio           1.75       1.208304   2.534545
        rel. diff.     .1875        .077082    .297918

        odds ratio         4       1.456814   13.63903   (exact)
```

v) **A 1:1 matched case-control study conducted among 100 pairs of cases and controls. Cases who exposed to the factor were 35% whereas controls who exposed to the factor were 20%. There is a statistically significant relationship between expose to the factor and disease (p-value = 0.003). Cases were 4 times more likely to have been exposed to the factor than controls (95%CI : 1.46 to 13.64).**

Question 5.

Overall remarks: Outcome of this study is "Vaccination status" or V and an exposure of interest which is "receptive perception on vaccination of the mothers" or R. The remaining variables are regarded as controlled variables which include "sex of the children" (S) and "whether or not their parents living together" (P). We will use these notation throughout.

i) **Ignoring the effects of S and P :**

	V+	V-	Total
R+	158	242	400
R-	42	158	200
Total	200	400	600

Let π_1 be the proportion of mothers with receptive perception
who adopted vaccination
and π_2 be the corresponding proportion without receptive perception.

$$H_0 : \pi_1 = \pi_2$$

We calculate Pearson's Chi-square to test this hypothesis:

$$\chi^2 = 20.54$$

Comparing this with the chi-square distribution with 1 df., the probability of observing a value as large or lager if H_0 is true is $p < 0.05$. Thus we reject H_0 in flavor of the alternative hypothesis and conclude that there is a statistically significant difference in proportions of receiving vaccination in the receptive and non-receptive groups. In other words, there is a statistically significant association between V and R.

```
. disp chiprob(1, 20.54)
5.840e-06
```

This means 5.840×10^{-06} or 0.000005.84. However we report this as p-value < 0.001.

Since this is a prospective study the RR and OR could be estimated.

$$RR = \frac{158/400}{42/200} = \frac{158\times 200}{400\times 42} = \frac{79}{42} = 1.88$$

$$OR = \frac{158\times 158}{42\times 242} = 2.46$$

Note : Vaccination is not a rare outcome so the odds ratio is not a good approximation to the relative risk.

RR = 1.88 ⇨ People with receptive attitudes are 1.9 times likely to be vaccinated than those with unreceptive attitudes.

OR = 2.46 ⇨ the odds of a person with receptive attitudes being vaccinated are 2.5 time that for a person with non-receptive attitudes.

$$var(ln\,RR) \qquad = \frac{242}{400\times 158} + \frac{158}{200\times 42} = 0.0226$$

95% CI ln RR : ln 1.88 ± 1.96 ($\sqrt{.0226}$) = ln(0.336, 0.926)

95% CI RR : 1.40 to 2.52

$$var(lnOR) \qquad = \qquad \frac{1}{158} + \frac{1}{242} + \frac{1}{42} + \frac{1}{158} = 0.0406$$

95% CI ln OR : ln 2.46 ± 1.96($\sqrt{.0406}$) = ln (0.505, 1.295)

95% CI OR : 1.66 to 3.65

STATA Commands:

Create a data file

Since we will use the data for stratified analysis in the succeeding sections, we need to create a data file. This data file will also be used for further chapter in logistic regression. To do this, first we assign variable name and code. For simplicity of typing, let's assign as the following:

v	**=**	**Vaccination status** **(1=vaccinated, 0=not vaccinated)**
r	**=**	**Receptive perception** **(1=positive receptive perception,** **0=negative perceptive reception)**
p	**=**	**Parents living together** **(1=yes, 0=no)**
s	**=**	**Sex of the children** **(1=boy, 0=girl)**
freq	**=**	**Frequency**

Step 1: Enter the following data into the data editor of STATA

p	s	v	r	freq
1	1	1	1	68
1	1	1	0	17
1	1	0	1	172
1	1	0	0	43
1	0	1	1	8
1	0	1	0	12
1	0	0	1	52
1	0	0	0	78
0	1	1	1	1
0	1	1	0	4

p	s	v	r	freq
0	1	0	1	9
0	1	0	0	36
0	0	1	1	81
0	0	1	0	9
0	0	0	1	9
0	0	0	0	1

Step 2: Expand the data file *(The data file as the above format is a frequency form - one row contains several number of records. To get a usual one row per one record format, we need to expand the datafile.)*

```
. expand freq
(584 observations created)
```

Step 3: Analyze the data

```
. cs v r , or

                 | r                      |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |       158          42  |        200
        Noncases |       242         158  |        400
-----------------+------------------------+----------
           Total |       400         200  |        600
                 |                        |
            Risk |      .395         .21  |   .3333333
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |              .185      |   .1109627    .2590373
      Risk ratio |          1.880952      |    1.40057    2.526101
 Attr. frac. ex. |          .4683544      |    .286005    .6041331
 Attr. frac. pop |               .37      |
      Odds ratio |           2.45612      |   1.656968    3.639969  (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =    20.53  Pr>chi2 = 0.0000
```

ii) Ignoring P, the effect of R on V controlling for the effect of S is :

Boys

	V+	V-	Total
R+	69	181	250
R-	21	79	100
Total	90	79	350

$$OR = \frac{69 \times 79}{21 \times 181} = 1.43 \qquad OR = \frac{89 \times 79}{21 \times 66} = 5.49$$

Girls

	V+	V-	Total
R+	89	61	150
R-	21	79	100
Total	110	140	250

$$RR = \frac{69 \times 250}{21 \times 100} = 1.31 \qquad RR = \frac{89 \times 79}{21 \times 100} = 2.83$$

Compairing the relation risk estimates or odds ratios for boys and girls with the crude estimate we see that the estimates for girls are higher than for boys, the crude estimates (ie., OR=2.46, RR=1.88) lying between the gender-specific estimates.

This suggests interaction or effect modification - the effect of receptive perception on receiving vaccination depends on whether the person is boy or girl.

STATA Commands:

```
. cs v r if s == 0, or

                 | r                      |
                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |        89          21  |       110
        Noncases |        61          79  |       140
-----------------+------------------------+----------
           Total |       150         100  |       250
                 |                        |
            Risk |  .5933333         .21  |       .44
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .3833333       |  .2712962    .4953705
      Risk ratio |         2.825397       |  1.889054    4.225855
 Attr. frac. ex. |         .6460674       |  .4706345    .7633615
 Attr. frac. pop |         .5227273       |
      Odds ratio |         5.488681       |  3.079367    9.776585  (Cornfield)
                 +-----------------------------------------------
                             chi2(1) =    35.78  Pr>chi2 = 0.0000
```

```
. cs v r if s == 1, or

                 | r                      |
                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |        69          21  |        90
        Noncases |       181          79  |       260
-----------------+------------------------+----------
           Total |       250         100  |       350
                 |                        |
            Risk |      .276         .21  |  .2571429
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |             .066       | -.0311774    .1631774
      Risk ratio |         1.314286       |  .8550348    2.020206
 Attr. frac. ex. |         .2391304       |  -.169543    .5050011
 Attr. frac. pop |         .1833333       |
      Odds ratio |         1.434096       |   .826136    2.487454  (Cornfield)
                 +-----------------------------------------------
                             chi2(1) =     1.63  Pr>chi2 = 0.2019
```

STATA commands for stratified analysis

```
. cs v r, by(s)

               s |       RR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   2.825397     1.889054   4.225855          12.6
               1 |   1.314286     .8550348   2.020206            15
-----------------+-------------------------------------------------
           Crude |   1.880952      1.40057   2.526101
    M-H combined |   2.004141     1.501398   2.675227
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    6.496  Pr>chi2 = 0.0108
```

iii) Ignoring S, the effect of R on V controlling for the effect of P is :

Parents living together

	V+	V-	Total
R+	76	224	300
R-	29	121	150
Total	105	345	450

$$OR = \frac{76 \times 121}{29 \times 224} = 1.42$$

$$RR = \frac{76 \times 300}{29 \times 150} = 1.31$$

Parents NOT living together

	V+	V-	Total
R+	82	18	100
R-	13	37	50
Total	95	55	150

$$OR = \frac{82 \times 37}{13 \times 18} = 12.97$$

$$RR = \frac{82 \times 100}{13 \times 50} = 3.15$$

Again, there appears to be interaction with the effect of receptive perception of vaccination on receiving vaccination dependent on whether or not the parents of the children lived together - the effect is larger in those whose parents did not than did not.

STATA Commands:

```
. cs v r if p == 0, or
```

```
                 | r                        |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        82          13  |         95
        Noncases |        18          37  |         55
-----------------+------------------------+----------
           Total |       100          50  |        150
                 |                        |
            Risk |       .82         .26  |   .6333333
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |              .56       |   .4169898    .7030102
      Risk ratio |          3.153846      |   1.958292    5.079297
 Attr. frac. ex. |          .6829268      |   .4893508    .8031224
 Attr. frac. pop |          .5894737      |
      Odds ratio |          12.96581      |   5.791289    29.02214  (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =    45.01  Pr>chi2 = 0.0000
```

```
. cs v r if p == 1, or
```

```
                 | r                        |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        76          29  |        105
        Noncases |       224         121  |        345
-----------------+------------------------+----------
           Total |       300         150  |        450
                 |                        |
            Risk |  .2533333    .1933333  |   .2333333
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |              .06       |  -.0201005    .1401005
      Risk ratio |          1.310345      |   .8958644    1.916589
 Attr. frac. ex. |          .2368421      |  -.1162403    .4782396
 Attr. frac. pop |          .1714286      |
      Odds ratio |           1.41564      |   .8767834     2.28471  (Cornfield)
                 +-----------------------------------------------
                               chi2(1) =     2.01  Pr>chi2 = 0.1560
```

```
. cs v r, by(p)
```

```
               p |       RR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   3.153846      1.958292   5.079297      8.666667
               1 |   1.310345      .8958644   1.916589      19.33333
-----------------+-------------------------------------------------
           Crude |   1.880952       1.40057   2.526101
    M-H combined |   1.880952      1.406467   2.515511
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =    7.990  Pr>chi2 = 0.0047
```

iv) Stratifying both S and P simultaneously, the effects of R on V controlling for the effect of S and P are :

Boys & Parents living together

	V+	V-	Total
R+	68	72	240
R-	17	43	60
Total	85	215	300

$OR = 1.0$

$RR = 1.0$

Girls & Parents living together

	V+	V-	Total
R+	8	52	60
R-	12	78	90
Total	20	130	150

$OR = 1.0$

$RR = 1.0$

Boys & Parents NOT living together

	V+	V-	Total
R+	1	9	10
R-	4	36	40
Total	5	45	50

$OR = 1.0$

$RR = 1.0$

Girls & Parents NOT living together

	V+	V-	Total
R+	81	9	90
R-	9	1	10
Total	90	10	100

$OR = 1.0$

$RR = 1.0$

Within each S & P stratum there is no relationship between receptive perception of vaccination and receiving vaccination. The effect of receptive perception on vaccination of the mothers on receiving vaccination is similar across strata (ie., OR=1 and RR=1). The crude estimates (ie., OR=2.46, RR=1.88) lying outside the stratum-specific estimates. Thus

"sex of the children" and "whether or not their parents living together" jointly appear to be confounders.

STATA Commands:

The following commands are to generate a composite variable named "sp" containing combination of variable "s" and "p".

```
. gen sp = .
(600 missing values generated)

. replace sp = 1 if s == 0 & p == 0
(100 real changes made)

. replace sp = 2 if s == 1 & p == 0
(50 real changes made)

. replace sp = 3 if s == 0 & p == 1
(150 real changes made)

. replace sp = 4 if s == 1 & p == 1
(300 real changes made)

. tab sp
         sp |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        100       16.67       16.67
          2 |         50        8.33       25.00
          3 |        150       25.00       50.00
          4 |        300       50.00      100.00
------------+-----------------------------------
      Total |        600      100.00
```

The following 4 commands are to obtain estimates of the 4 strata.

```
. cs v r if sp == 1, or

                 | r                      |
                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+----------
           Cases |        81           9  |         90
        Noncases |         9           1  |         10
-----------------+------------------------+----------
           Total |        90          10  |        100
                 |                        |
            Risk |        .9          .9  |         .9
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |                0       | -.1959964    .1959964
      Risk ratio |                1       |  .8043074    1.243306
 Attr. frac. ex. |                0       | -.2433058    .1956926
 Attr. frac. pop |                0       |
      Odds ratio |                1       |         0    7.003225  (Cornfield)
                 +-----------------------------------------------
                             chi2(1) =     0.00  Pr>chi2 = 1.0000

. cs v r if sp == 2, or

                 | r                      |
                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |         1           4  |         5
        Noncases |         9          36  |        45
-----------------+------------------------+----------
           Total |        10          40  |        50
                 |                        |
            Risk |        .1          .1  |        .1
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |                0       | -.2078856    .2078856
      Risk ratio |                1       |  .1250732    7.995315
 Attr. frac. ex. |                0       | -6.995315    .8749268
 Attr. frac. pop |                0       |
      Odds ratio |                1       |         0    7.846757  (Cornfield)
                 +-----------------------------------------------
                             chi2(1) =     0.00  Pr>chi2 = 1.0000
. cs v r if sp == 3, or

                 | r                      |
                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |         8          12  |        20
        Noncases |        52          78  |       130
-----------------+------------------------+----------
           Total |        60          90  |       150
                 |                        |
            Risk |  .1333333    .1333333  |  .1333333
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |                0       | -.1110433    .1110433
      Risk ratio |                1       |  .4348194    2.299805
 Attr. frac. ex. |                0       | -1.299805    .5651806
 Attr. frac. pop |                0       |
      Odds ratio |                1       |  .3922252    2.556461  (Cornfield)
                 +-----------------------------------------------
                             chi2(1) =     0.00  Pr>chi2 = 1.0000

. cs v r if sp == 4, or

                 | r                      |
                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |        68          17  |        85
        Noncases |       172          43  |       215
-----------------+------------------------+----------
           Total |       240          60  |       300
                 |                        |
            Risk |  .2833333    .2833333  |  .2833333
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |-----------------------+----------------------
 Risk difference |                0      | -.1274779    .1274779
      Risk ratio |                1      |  .6376779     1.56819
 Attr. frac. ex. |                0      | -.5681899    .3623221
 Attr. frac. pop |                0      |
      Odds ratio |                1      |   .536736    1.860847  (Cornfield)
                 +----------------------------------------------
                             chi2(1) =     0.00  Pr>chi2 = 1.0000
```

The following 2 commands are to perform stratified analyses.

```
. cc v r, by(sp)

              sp |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               1 |          1            0    7.003225           .81
               2 |          1            0    7.846757           .72
               3 |          1     .3922252    2.556461          4.16
               4 |          1      .536736    1.860847      9.746667
-----------------+-------------------------------------------------
           Crude |    2.45612     1.656968    3.639969
    M-H combined |          1     .6072276    1.646829
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(3) =     0.00  Pr>chi2 = 1.0000

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.00
                                                Pr>chi2 =    1.0000
. cs v r, by(sp)

              sp |       RR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               1 |          1     .8043074    1.243306           8.1
               2 |          1     .1250732    7.995315            .8
               3 |          1     .4348194    2.299805           4.8
               4 |          1     .6376779     1.56819          13.6
-----------------+-------------------------------------------------
           Crude |   1.880952      1.40057    2.526101
    M-H combined |          1     .7542183    1.325876
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(3) =    0.000  Pr>chi2 = 1.0000
```

v) Individually “sex of the children” and “whether or not their parents living together” appear to be effect modifiers but in combination they confound the relationship between “receptive perception of vaccination” and “receiving of vaccination”.

vi) No, it is not a determinant of receiving vaccination as overall there is no association between perceptions and vaccination.

vii) Summary findings:

A cohort study involves 600 mothers who delivered their babies at a hospital. Among a total of 400 mothers who have positive receptive perception on vaccination, 39.5% of them have their children vaccinated. Whilst among a total of 200 mothers who have negative receptive perception on vaccination, 21.0% of them have their children vaccinated. After controlling for the effect of sex of the children and whether or not their parents living together, receiving vaccination of the children is not affected by what the receptive perception of the mothers is (adjusted RR = 1, 95%CI: 0.75 to 1.32). Detailed analysis suggested that individually "sex of the children" and "whether or not their parents living together" are significant effect modifiers (p-value = 0.011 and 0.005 respectively) but in combination they appear to confounder of the relationship between "receptive perception of vaccination" and "receiving of vaccination".

Note:

i) Question (a) is known as the crude analysis. Someone called a bivariate analysis. This serves as a good start for exploratory data analysis. That is, in most cases, it is not be valid by its own since most of health outcomes caused by several factors and they are sometime inter-correlated. However the crude bivariate analysis provides clues for further analysis which will be discussed in the chapter of "Logistic regression". Questions (b, c and d) are the stratified analysis. For question (b), the stratified variable is "sex of the children" whereas the variable "whether or not their parents living together" is the stratified variable for Question (c). In Question (d), a combination of the previous two variables forms its stratified variable.

ii) ***In the above analysis we calculate both OR and RR. In reality we need to do only one and RR is the most appropriate measure of association since this is a cohort study. However this is done to get a feel on how the OR is affected by the rate of the outcome. Moreover if we have many more controlled variables, we cannot use stratified analysis. In such case, we need to use "Logistic regression". Then the OR will be reported. We will discuss this again in the in the chapter of "Logistic regression".***

iii) ***Systematic approach for stratified analysis, one need to report the following 4 components before interpretation. That is, 1) the crude estimate which always has only one, 2) the stratum-specific estimates which vary depending on number of stratum, 3) the adjusted estimate which always has only one, and 4) a p-value of the test for homogeneity of the stratum-specific estimates. This can be easily accomplished by the last STATA command. Then we first consider the last component. If p-value is less than 0.05, it suggested that there is a statistically significant difference of the estimates across strata. Then we can conclude that there is a significant modification effect where the stratified variable is the effect modifier. In this case, the stratum-specific estimates and their confidence intervals are to be reported, discard the adjusted one. On the other hand, if p-value is greater than 0.05, we will conclude that there is no significant interaction effect. However, this test is lack of power. We recommend that the similarity of the stratum-specific estimates be based on judgement. If there seems to be clinically meaningful differences, we report the stratum-specific estimates and their confidence intervals. The adjusted estimates may be reported for discussion purpose. If they are more or less the same clinically, then we report the adjusted estimate ad its confidence interval. The role of the stratified variable can be determined by comparing the adjusted estimate with the crude one. If they are more or less the same, the effect the stratified variable on the relationship between the exposure of interest and***

the outcome is minimal or none. On the other hand, if they are different clinically, we will say that there is a confounding effect and the stratified variable is the confounder. There is no test statistics for this effect since it is not a chance bias but a systematic bias.

Answers for the exercise in Chapter 3

Question 1.

The data can be displayed as the 2 x C Table as follows:

	Diagnosis					
	S	AD	N	PD	SS	
D	105	12	18	47	0	182
$\overline{D}$	8	2	19	52	13	94
	113	14	37	99	13	276

i) H_0 : There is no association between diagnosis and prescription of drugs

or

H_0 : $\pi_{ij} = \pi_{i+}\pi_{+j} \qquad i = 1,2 \qquad j = 1,\ldots,5$

Where π_{ij} is the theoretical probability for cell (i,j) and π_{i+} ,π_{+j} are the row and column totals.

$$\chi^2 = 84.19$$

Comparing this with the chi-square distribution with 4 df., the probability of observing a value as large or lager if H_0 is true is p < 0.001. Thus we reject the null hypothesis and conclude that there is a strong association between diagnosis and prescription of drugs.

ii) Smallest expected values :

$$\frac{94 \times 13}{276} = 4.43$$

$$\frac{94 \times 14}{276} = 4.77$$

$$\frac{94 \times 37}{276} = 12.60$$

There are 2 out of 10 cells with expected value of less than 5.

There are two cells (20%) with expected values less than 5 but none less than 1. Thus the chi-square approximation to Pearson's Chi-square statistic should not be too bad.

STATA Commands :

```
. tabi 105 12 18 47 0 \ 8 2 19 52 13, col chi2

           |                          col
       row |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
         1 |       105         12         18         47          0 |       182
           |     92.92      85.71      48.65      47.47       0.00 |     65.94
-----------+-------------------------------------------------------+----------
         2 |         8          2         19         52         13 |        94
           |      7.08      14.29      51.35      52.53     100.00 |     34.06
-----------+-------------------------------------------------------+----------
     Total |       113         14         37         99         13 |       276
           |    100.00     100.00     100.00     100.00     100.00 |    100.00


Pearson chi2(4) =  84.1885   Pr = 0.000
```

Note that the above command requested the column percents, ie. assuming column total fixed, for simplicity of interpretation.

iii) This cross-sectional study involved 276 psychiatric patients. There were very high proportion of treatment include drugs for those who were diagnosed as schizophrenia (93%) and personality disorder (86%) whereas all those who were diagnosed as special symptoms, their treatments did not

include drug. There is a strong association between prescription of drugs and diagnosis ($\chi^2 = 84.19, 4df\,;\ p - value < 0.001$). Based on the proportions of treatment include drug across group of patients, it suggests that the diagnoses of schizophrenia, personality disorder and special symptoms contribute to the association in that the proportion of treatment include drug for schizophrenia and for personality disorder is higher but lower for special symptoms than that of the treatment did not include drug. The remaining groups of patients, the proportion are similar between the two groups of treatments.

Question 2.

The data can be displayed as the 2 x C Table as follows:

	Cigarettes per day						
	0	<5	5-14	15-24	25-49	50+	Total
Cases	7	55	489	475	293	38	1357
Controls	61	129	570	431	154	12	1357
	68	184	1059	906	447	50	2714

i) Ignoring the effect of ordinality of number of cigarettes smoked per day, we can test for association as follows:

H_0 : the proportion of cases falling into each smoking category
= the proportions of controls in each category or

$$\chi^2 = 137.72$$

(Note smallest expected value $= \dfrac{50\times 1357}{2714} = 25$**.)**

Comparing this with the chi-square distribution with 5 df., the probability of observing a value as large or lager if H_0 is true is $p < 0.001$. This leads us to reject Ho and conclude that the distribution of cases and controls across categories of smoking are different.

STATA Commands :

The first step we input the following data into data editor of STATA.

smk	case	freq
0	1	7
0	0	61
2.5	1	55
2.5	0	129
9.5	1	489
9.5	0	570
19.5	1	475
19.5	0	431
37	1	293
37	0	154
50	1	38
50	0	12

```
. expand freq
(2702 observations created)


. tab smk case, chi2

           |         case
       smk |         0          1 |     Total
-----------+----------------------+----------
         0 |        61          7 |        68 
       2.5 |       129         55 |       184 
       9.5 |       570        489 |      1059 
      19.5 |       431        475 |       906 
        37 |       154        293 |       447 
        50 |        12         38 |        50 
-----------+----------------------+----------
     Total |      1357       1357 |      2714
```

```
Pearson chi2(5) = 137.7193   Pr = 0.000
```

ii) Taken into account of the effect of ordinality of number of cigarettes smoked per day, we can test for association as follows:

H_0 : $f_1 = f_2 = f$ **where** $$f_i = \sum_{j=1}^{c} x_j \pi_{ij} \ ,$$

x_j = score assigned for category j
i = 1, 2 and j = 1, 2, 3, ..., 6

Taking the row total fixed (writing the table as a 2×6) and using the second method described in the module give the following:

Cases	**p_{1j}**	**.005**	**.041**	**.360**	**.350**	**.216**	**.028**	**1.995**
Controls	**p_{2j}**	**.045**	**.095**	**.420**	**.318**	**.113**	**.009**	**1.000**
Weight			**0**	**2.5**	**9.5**	**19.5**	**37**	**50**
	p_j	**.025**	**.068**	**.390**	**.334**	**.165**	**.018**	**1.000**

We notice that there are smaller proportions of cases than controls in the none and light smoking (<5, 5-14) categories and more cases in the heavier smoking categories.

Using weight equal to the midpoints of class intervals:
Mean scores cases :

$$f_1 = (0 \times 0.005) + (2.5 \times 0.041) + (9.5 \times 0.360) + (19.5 \times .350) + (37 \times 0.216) + (50 \times 0.028)$$

$$= 19.74$$

Mean score for controls :

$$f_2 = (0 \times 0.045) + (2.5 \times 0.095) + (9.5 \times 0.420) + (19.5 \times 0.318) + (37 \times 0.113) + (50 \times 0.009)$$

$$= 15.06$$

$$var(f_1) = \frac{1}{1357}\left\{531.54 - 19.74^2\right\} = 0.1045$$

$$var(f_2) = \frac{1}{1357}\left\{336.62 - 15.06^2\right\} = 0.0809$$

$$Q = \frac{(19.74 - 15.06)^2}{(.1045 + .0809)} = \mathbf{118.14}$$

Comparing this with the chi-square distribution with 1 df., the probability of observing a value as large or lager if H_0 is true is p < 0.001.Thus there is a highly statistically significant test for linear trend.

Now taking the column totals fixed and using method 1 with the same weights:

$$\sum_j n_{1j}x_j = \mathbf{(0 \times 7) + (2.5 \times 55) + (9.5 \times 489) + (19.5 \times 475)}$$

$$\mathbf{+ (37 \times 293) + (50 \times 38)}$$

$$\mathbf{= 26786.5}$$

$$\sum_j n_{+j}x_j = \mathbf{(0 \times 68) + (2.5 \times 184) + (9.5 \times 1059) + (19.5 \times 906)}$$

$$\mathbf{+ (37 \times 447) + (50 \times 50)}$$

$$\mathbf{= 47226.5}$$

$$\sum_j n_{+j}x_j^{\,2} = \mathbf{(0 \times 68) + (2.5^2 \times 184) + (9.5^2 \times 1059) + (19.5^2 \times 906)}$$

$$\mathbf{+ (37^2 \times 447) + (50^2 \times 50)}$$

$$\mathbf{= 1178174.25}$$

$$\chi_1^2 \qquad = \quad \frac{2714\left[2714\ (26786.5) - 1357\ (47226.5)\right]^2}{1357 \times 1357\ \left\{2714\ (1178174.25) - (47226.5)\ \right\}}$$

= 113.02

Again, this is highly statistically significant for linear trend.

Now we test for departure from linear trend.

137.72 – 113.02 = 24.70 on (5-1) = 4 df

STATA gives p-value < 0.001 as shown below.

```
. disp chiprob(4, 24.7)
.0000578
```

Thus while there is a highly statistically significant linear trend, this not explain all of the association between cigarette smoking and lung cancer.

We are also use method 2 and calculate Pearson's Chi-square rather than Neyman Chi-square.

f_1 = 19.74 and f_2 = 15.06 as before

Under $H_0 : f_1 = f_2 = f$

Where f = **(0×0.025) + (2.5×0.068) + (9.5×0.390) + (19.5×0.334) + (37×0.165) + (50×0.018)**

= 17.39

Calculate

$$var\,(f_1) \quad and \quad var\,(f_2) \;\; under \;\; H_0 : \; f_1 = f_2 = f$$

$$var\,(f_1) = \frac{\sum x_j^2 p_j - f^2}{n_{1+}}$$

$$= \frac{1}{1357}\left[433.51 - 17.39^2\right]$$

$$= \mathbf{0.0966}$$

Similarly $var\,(f_2) = 0.0966$

$$X_P^2 = \frac{(19.74 - 15.06)^2}{(0.966 + 0.966)}$$

$$= \mathbf{113.37}$$

Which is close to X_1^2 **obtained from method 1.**

Note : These should be equal except for rounding error in the calculation.

STATA Commands :

The following three commands involve test for trend. They yield similar results. This is to let us to have a feel about how the test for trend works. The first one performs a nonparametric test for trend across ordered groups. This test, developed by Cuzick (1985), is an extension of the Wilcoxon

rank-sum test and is a useful adjunct to the Kruskal-Wallis test. The formula for the test statistic is given by Cuzick (1985) and Altman (1991). The second and the third one are an ordinary two sample t-test and its non-parametric equivalent.

```
. nptrend smk, by(case)

      case     score       obs      sum of ranks
         0         0      1357     1637236.5
         1         1      1357     2047018.5

     z  = 10.59
  P>|z| =  0.00
```

```
. ttest smk, by(case)

Two-sample t test with equal variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |    1357    15.06264    .2846381    10.48535    14.50426    15.62102
       1 |    1357     19.7395    .3234075    11.91352    19.10507    20.37393
---------+--------------------------------------------------------------------
combined |    2714    17.40107    .2200029    11.46129    16.96968    17.83246
---------+--------------------------------------------------------------------
    diff |           -4.676861    .4308263               -5.521642    -3.83208
------------------------------------------------------------------------------
Degrees of freedom: 2712

                      Ho: mean(0) - mean(1) = diff = 0

     Ha: diff < 0               Ha: diff ~= 0              Ha: diff > 0
       t = -10.8556                t = -10.8556              t = -10.8556
   P < t =   0.0000          P > |t| =   0.0000          P > t =   1.0000
```

```
. ranksum smk, by(case)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

    case |      obs    rank sum    expected
---------+---------------------------------
       0 |     1357   1637236.5   1842127.5
       1 |     1357   2047018.5   1842127.5
---------+---------------------------------
combined |     2714     3684255     3684255

unadjusted variance   4.166e+08
adjustment for ties   -42251123
                     ----------
adjusted variance     3.744e+08

Ho: smk(case==0) = smk(case==1)
             z = -10.589
    Prob > |z| =   0.0000
```

The test above gives Z = 10.59 thus Z^2 = 112.15 which approximately equal to chi-square test for trend as obtained by manual calculation shown earlier (113.02 or 113.37). The more relevant command is shown below. The command "tabodds" is used with case-control and cross-sectional data. It tabulates the odds of failure against a categorical explanatory variable. It also performs an approximate chi-squared test of homogeneity of odds and a test for linear trend of the log odds against the numerical code used for the categories of the explanatory variable. Both of these tests are based on the score statistic and its variance.

```
. tabodds case smk
------------+------------------------------------------------------------
        smk |      cases     controls       odds      [95% Conf. Interval]
------------+------------------------------------------------------------
          0 |          7           61    0.11475        0.05249   0.25087
        2.5 |         55          129    0.42636        0.31095   0.58459
        9.5 |        489          570    0.85789        0.76027   0.96806
       19.5 |        475          431    1.10209        0.96737   1.25557
         37 |        293          154    1.90260        1.56540   2.31243
         50 |         38           12    3.16667        1.65478   6.05987
------------+------------------------------------------------------------
Test of homogeneity (equal odds): chi2(5)  =   137.67
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =   112.98
                                  Pr>chi2  =   0.0000
```

Thus we can also test for departure from linear trend using results from this command.

137.67 – 112.98 = 24.69 (This is similar to 24.70 which was obtained by manual calculation)

iii)	**None**	**vs**	**<5**	**OR = 3.72**
	None	**vs**	**5 –14**	**OR = 7.48**
	None	**vs**	**15 – 24**	**OR = 9.60**
	None	**vs**	**25 - 49**	**OR = 16.58**
	None	**vs**	**50+**	**OR = 27.59**

We see that the olds of a case being in the ‘smoker’ group increases with increasing levels of smoking as compared to the olds of the controls being in the ‘smoker’ group.

iv)	**None**	**vs**	**<5**	**OR = 3.72**
	<5	**vs**	**5-14**	**OR = 2.01**
	5-14	**vs**	**15-24**	**OR = 1.28**
	15-24	**vs**	**25-49**	**OR = 1.73**
	25-49	**vs**	**50+**	**OR = 1.66**

While there are increasing odds of a case being in the higher smoking group compared to controls, the largest increase is in going from “none” to “<5 cigarettes per day”.

v) Hospital controls were used. These may not be representative of the general population from which cases came. Also smoking is higher in hospital patients than the general population since smoking is related to number of diseases besides lung cancer. As we can see that 95% of controls were smokers which is very large.

Answers for the exercise in Chapter 4

i) We wish to test the null hypothesis

H_0 : there is on association between age and frequency of breast self-examination or

H_0 : $\pi_{ij} = \pi_{i+} \; \pi_{+j}$ where π_{ij} , π_{i+} and π_{+j} are the population proportion falling into cell (i,j), row i and column j (i = 1,...,3; j = 1,...,3) respectively.

We calculate Pearson's Chi-square statistic

$$\chi^2 = 25.09 \text{ on } 4 \text{ df.}$$

Comparing this with the chi-square distribution with 4 df., the probability of observing a value as large or lager if H_0 is true is $p < 0.001$. Thus we reject H_0 and conclude that there is a statistically significant association between age and frequency of breast self examination.

ii) The cell Chi-square statistics are :

		BSE	
Age	Monthly	Occasionally	Never
<45	8.79	0.1	6.18
45-59	0.15	0.04	0.03
60+	6.05	0.17	3.58

Note that chi-square = $(O - E)^2 / E$. Example of calculation of chi-square for the first cell: $8.79 = (91-66.78)^2 / 66.78$.

Comparing each of these with a chi-square distribution with 1 df, we have observed frequencies being statistically significantly different from that expected

under H_0 in the <45 year olds for “monthly” and “never” and for the 60+ year olds in “monthly” and “never”. However this analysis dose not tell us the direction of the difference. Calculating the expected values for these 4 cells indicates there are more than expected in “monthly” and fewer than expected for “never” in the young age group. The pattern is reversed in the oldest age group. Thus it appears that younger women tend to do breast self-examination more frequently than the older women. This pattern is also seen if we look at the “row percents” as follows:

```
. tabi 91 90 51 \ 150 200 155 \ 109 198 172, row

           |               col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        91         90         51 |       232
           |     39.22      38.79      21.98 |    100.00
-----------+---------------------------------+----------
         2 |       150        200        155 |       505
           |     29.70      39.60      30.69 |    100.00
-----------+---------------------------------+----------
         3 |       109        198        172 |       479
           |     22.76      41.34      35.91 |    100.00
-----------+---------------------------------+----------
     Total |       350        488        378 |      1216
           |     28.78      40.13      31.09 |    100.00
```

Note that the proportions in bold letters at the bottom are the one under the null hypothesis. Proportions in circles appear to be different from their corresponding null proportion.

Calculating “local” odds ratios for adjacent rows and columns

$$OR_{11} = \frac{91 \times 200}{150 \times 90} = 1.35$$

$$OR_{12} = \frac{90 \times 155}{200 \times 51} = 1.37$$

$$OR_{21} = \frac{150 \times 198}{200 \times 109} = 1.36$$

$$OR_{22} = \frac{200 \times 172}{198 \times 155} = 1.12$$

Thus in all cases, the younger age group is more likely to perform breast self-examination more frequently than the older age group.

Looking at the statistics from STATA. First we entry the data file as follows:

age	bse	freq
45	1	91
45	2	90
45	3	51
52	1	150
52	2	200
52	3	155
60	1	109
60	2	198
60	3	172

Then expand the data file to get the individual records of data.

```
. expand freq
(1207 observations created)
```

Now we are ready for the analysis. We can use the following two commands.

```
. tab age bse, all

           |               bse
       age |         1          2          3 |     Total
-----------+---------------------------------+----------
        45 |        91         90         51 |       232
        52 |       150        200        155 |       505
        60 |       109        198        172 |       479
-----------+---------------------------------+----------
     Total |       350        488        378 |      1216

          Pearson chi2(4) =  25.0860   Pr = 0.000
 likelihood-ratio chi2(4) =  25.1923   Pr = 0.000
               Cramer's V =   0.1016
                    gamma =   0.1897  ASE = 0.038
          Kendall's tau-b =   0.1234  ASE = 0.025
```

```
. spearman  age bse

 Number of obs =     1216
Spearman's rho =       0.1378

Test of Ho: age and bse independent
      Pr > |t| =       0.0000
```

```
        . correlat  age bse
(obs=1216)

         |      age      bse
---------+------------------
     age |   1.0000
     bse |   0.1394   1.0000
```

```
. nptrend  bse, by(age)

      age       score       obs       sum of ranks
       45          45       232         121878
       52          52       505       304487.5
       60          60       479       313570.5

      z   =   4.85
          P>|z| =  0.00
```

→ $\boxed{Z^2 = 4.85^2 = 23.523 \approx \chi^2}$

These following measures of association were taken into account of the ordinality of the age and BSE Variables. The scores 45, 52 and 60 were assigned to age and 1, 2, 3 to BSE (note open ended intervals for age make it difficult to know which scores to assign to categories <45 and 60+).

Gamma	=	0.19
Tau-b	=	0.12
Pearson correlation	=	0.14
Spearman correlation	=	0.14

Thus there was a “weak” linear correlation between age and BSE frequency.

Test for departure form linear trend : 25.086 - 23.523 = 1.563 on 3 df. resulting in p-value of 0.668 as shown below. Thus most of the association between age and BSE is explained by a linear association.

```
. disp chiprob(3,1.563)
.66780811
```

In summary, there is a significant association between age and frequency of BSE such that younger women tend to perform BSE more frequently than older women who are more likely to perform occasionally or never. However the magnitude of association is weak.

Answers for the exercise in Chapter 5

i) Observer’s marginal distribution

		Category		
		1	2	3
Anaesthetist	1	0.36	0.33	0.31
	2	0.40	0.47	0.13

Anaesthetist 1 use each category with equal frequency while Anesthetist 2 tends to use categories 1 and 2 more frequently than category 3. Anesthetist 2 only classifies patients as "unsuitable" 13% of time while Anesthetist 1 classifies one-third of patients as "unsuitable"

STATA Commands:

1. The command to obtain marginal proportions

```
. tabi 15 3 0 \ 1 12 8 \ 0 0 6, row col

           |                col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        15          3          0 |        18
           |     83.33      16.67       0.00 |    100.00
           |     93.75      20.00       0.00 |     40.00
-----------+---------------------------------+----------
         2 |         1         12          8 |        21
           |      4.76      57.14      38.10 |    100.00
           |      6.25      80.00      57.14 |     46.67
-----------+---------------------------------+----------
         3 |         0          0          6 |         6
           |      0.00       0.00     100.00 |    100.00
           |      0.00       0.00      42.86 |     13.33
-----------+---------------------------------+----------
     Total |        16         15         14 |        45
           |     35.56      33.33      31.11 |    100.00
           |    100.00     100.00     100.00 |    100.00
```

2. The command to obtain a data file

```
. tabi 15 3 0 \ 1 12 8 \ 0 0 6, replace

           |                col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        15          3          0 |        18
         2 |         1         12          8 |        21
         3 |         0          0          6 |         6
-----------+---------------------------------+----------
     Total |        16         15         14 |        45

          Pearson chi2(4) =  41.4432   Pr = 0.000
```

The output of the above command can be ignored. What of interest is the data file. The “replace” option automatically provides us the data file as follows:

row	col	pop
1	1	15
1	2	3
1	3	0
2	1	1
2	2	12
2	3	8
3	1	0
3	2	0
3	3	6

We can see this at the data editor or use “list command. The variable “row” is Anaesthetist 1 assessment, “col” is Anaesthetist 2 assessment, and “pop” is the number so assessed by both.

3. The command to test for symmetry and Marginal homogeneity

```
. symmetry row col [freq=pop], trend exact

----------+--------------------------
          |           col
      row |   1      2      3    Total
----------+--------------------------
        1 |  15      3      0     18
        2 |   1     12      8     21
        3 |   0      0      6      6
          |
    Total |  16     15     14     45
----------+--------------------------

                                        Chi-Squared    df      Prob>chi2
-------------------------------------+-------------------------------
Symmetry (asymptotic)                |      9.00      2         0.0111
Marginal homogeneity (Stuart-Maxwell) |      9.00      2         0.0111
-------------------------------------+-------------------------------
Linear trend in the (log) RR         |      8.33      1         0.0039
-------------------------------------+-------------------------------
Symmetry (exact significance probability)                       0.0049
```

ii)	**Nominal**	**Ordinal**	
		$W_{ij} = 1 - \frac{\lvert i-j \rvert}{c-1}$	$W_{ij} = 1 - \left(\frac{i-j}{c-1}\right)^2$
Kappa	**0.596**	**0.680**	**0.773**
Var$_0$(K)	**0.010385**	**0.011740**	**0.019737**
$Z = \frac{K}{\sqrt{var_0(K)}}$	**5.85**	**6.28**	**5.50**
P-value	**< 0.001**	**< 0.001**	**< 0.001**
95%CI Kappa*	**0.40 to 0.80**	**0.47 to 0.89**	**0.50 to 1.05**

***Note that 95%CI of Kappa =K** $\pm 1.96\sqrt{var_0(K)}$

In all cases p<0.001 so we reject Ho and conclude that the level of agreement achieved by the anaesthetists is statistically significantly better than that expected by chance.

STATA Commands:

The following three commands are to obtain results as summarized above.
The first command is for nominal data.

The second one is for ordinal with weight of $W_{ij} = 1 - \frac{\lvert i-j \rvert}{c-1}$ **.**

The last command is also for ordinal outcome using weight

$$W_{ij} = 1 - \left(\frac{i-j}{c-1} \right)^2 \quad .$$

```
. kap row col [freq=pop], tab

           |                 col
       row |         1          2          3 |     Total
-----------+---------------------------------+----------
         1 |        15          3          0 |        18
         2 |         1         12          8 |        21
         3 |         0          0          6 |         6
-----------+---------------------------------+----------
     Total |        16         15         14 |        45

             Expected
Agreement   Agreement     Kappa         Z         Pr>Z
-----------------------------------------------------
  73.33%      33.93%     0.5964       5.85      0.0000

. kap row col [freq=pop], wgt(w)

Ratings weighted by:
   1.0000   0.5000   0.0000
   0.5000   1.0000   0.5000
   0.0000   0.5000   1.0000

             Expected
Agreement   Agreement     Kappa         Z         Pr>Z
-----------------------------------------------------
  86.67%      58.37%     0.6797       6.27      0.0000
. kap row col [freq=pop], wgt(w2)

Ratings weighted by:
   1.0000   0.7500   0.0000
   0.7500   1.0000   0.7500
   0.0000   0.7500   1.0000

             Expected
Agreement   Agreement     Kappa         Z         Pr>Z
-----------------------------------------------------
  93.33%      70.59%     0.7733       5.50      0.0000
```

If the weight is to specify arbitrarily, we can define our own weight as follows:

```
. kapwgt mine 1 \ .8 1 \ 0 .8 1

. kap row col [freq=pop], wgt(mine)

Ratings weighted by:
   1.0000   0.8000   0.0000
   0.8000   1.0000   0.8000
   0.0000   0.8000   1.0000
```

```
                Expected
Agreement   Agreement     Kappa         Z          Pr>Z
-----------------------------------------------------------
  94.67%      73.04%     0.8022       5.22       0.0000
```

iii) While the anaesthetists tend to use the categories of the scale with different frequencies (Anesthetist 1 is more conservative in classifying more patient as unsuitable while Anaesthetist 2 tends to classify patients in the middle category), they can reach agreement better than that expected by chance. Thus the scale is reliable.

Answers for the exercise in Chapter 6

Part 1. Bivariate analysis : examine relationship between each variable and survival, one variable at a time.

1.1 Entering the data into Stata in the following format.

center	age	survive	inflam	appear	freq
1	1	1	1	1	9
1	1	1	1	2	7
1	1	1	2	1	4
1	1	1	2	2	3
1	1	2	1	1	26
1	1	2	1	2	68
1	1	2	2	1	25
1	1	2	2	2	9
1	2	1	1	1	9
1	2	1	1	2	9
1	2	1	2	1	11
1	2	1	2	2	2
1	2	2	1	1	20
1	2	2	1	2	46
1	2	2	2	1	18
1	2	2	2	2	5
1	3	1	1	1	2
1	3	1	1	2	3
1	3	1	2	1	1
1	3	1	2	2	0
1	3	2	1	1	1

center	age	survive	inflam	appear	freq
1	3	2	1	2	6
1	3	2	2	1	5
1	3	2	2	2	1
2	1	1	1	1	6
2	1	1	1	2	7
2	1	1	2	1	6
2	1	1	2	2	0
2	1	2	1	1	11
2	1	2	1	2	24
2	1	2	2	1	4
2	1	2	2	2	0
2	2	1	1	1	8
2	2	1	1	2	20
2	2	1	2	1	3
2	2	1	2	2	2
2	2	2	1	1	18
2	2	2	1	2	58
2	2	2	2	1	10
2	2	2	2	2	3
2	3	1	1	1	9
2	3	1	1	2	18
2	3	1	2	1	3
2	3	1	2	2	0
2	3	2	1	1	15
2	3	2	1	2	26
2	3	2	2	1	1
2	3	2	2	2	1
3	1	1	1	1	16
3	1	1	1	2	7
3	1	1	2	1	3
3	1	1	2	2	0
3	1	2	1	1	16
3	1	2	1	2	20
3	1	2	2	1	8
3	1	2	2	2	1
3	2	1	1	1	14
3	2	1	1	2	12
3	2	1	2	1	3
3	2	1	2	2	0
3	2	2	1	1	27
3	2	2	1	2	39
3	2	2	2	1	10
3	2	2	2	2	4
3	3	1	1	1	3
3	3	1	1	2	7
3	3	1	2	1	3

center	age	survive	inflam	appear	freq
3	3	1	2	2	0
3	3	2	1	1	12
3	3	2	1	2	11
3	3	2	2	1	4
3	3	2	2	2	1

1.2 Performing bivariate data analysis

Step 1.1 Recode the outcome to be 0-1 variable, where 1=survive and 0=dead and perform a univariable analysis to determine magnitude of the problem under investigation.

```
. recode  survive 1=0 2=1
(771 changes made)
```

```
. ci survive [freq=freq]

Variable |      Obs          Mean    Std. Err.       [95% Conf. Interval]
---------+-------------------------------------------------------------
 survive |      764      .7251309    .0161625        .6934026    .7568592
```

Proportion of three-year survival was 72.5% (95%CI: 69.3% to 75.7%).

Step 1.2 Crude analysis to determine the effect of CENTER on SURVIVE

```
. tab center survive [freq=freq], row chi2

           |        survive
    center |         0          1 |     Total
-----------+----------------------+----------
         1 |        60        230 |       290
           |     20.69      79.31 |    100.00
-----------+----------------------+----------
         2 |        82        171 |       253
           |     32.41      67.59 |    100.00
-----------+----------------------+----------
         3 |        68        153 |       221
           |     30.77      69.23 |    100.00
-----------+----------------------+----------
     Total |       210        554 |       764
           |     27.49      72.51 |    100.00

          Pearson chi2(2) =  10.9948   Pr = 0.004
```

```
. cci 171 82 230 60
```

```
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       171          82  |        253       0.6759
        Controls |       230          60  |        290       0.7931
-----------------+------------------------+----------------------
           Total |       401         142  |        543       0.7385
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         .5440085       |   .3696901    .8005537  (Cornfield)
 Prev. frac. ex. |         .4559915       |   .1994463    .6303099  (Cornfield)
 Prev. frac. pop |         .3616484       |
                 +-----------------------------------------------
                               chi2(1) =     9.61  Pr>chi2 = 0.0019
```

```
. cci 153 68 230 60
```

```
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       153          68  |        221       0.6923
        Controls |       230          60  |        290       0.7931
-----------------+------------------------+----------------------
           Total |       383         128  |        511       0.7495
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         .5869565       |    .392766    .8771556  (Cornfield)
 Prev. frac. ex. |         .4130435       |   .1228444     .607234  (Cornfield)
 Prev. frac. pop |         .3275862       |
                 +-----------------------------------------------
                               chi2(1) =     6.79  Pr>chi2 = 0.0092
```

Step 1.3 Crude analysis to determine the effect of AGE on SURVIVE

```
. tab age survive [freq=freq], row chi2
```

```
           |        survive
       age |         0          1 |     Total
-----------+----------------------+----------
         1 |        68        212 |       280
           |     24.29      75.71 |    100.00
-----------+----------------------+----------
         2 |        93        258 |       351
           |     26.50      73.50 |    100.00
-----------+----------------------+----------
         3 |        49         84 |       133
           |     36.84      63.16 |    100.00
-----------+----------------------+----------
     Total |       210        554 |       764
           |     27.49      72.51 |    100.00

          Pearson chi2(2) =   7.4526   Pr = 0.024
```

```
. cci 258 93 212 68
                                                         Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       258          93  |        351      0.7350
        Controls |       212          68  |        280      0.7571
-----------------+------------------------+----------------------
           Total |       470         161  |        631      0.7448
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         .8898357       |     .62034    1.276478  (Cornfield)
 Prev. frac. ex. |         .1101643       |  -.2764783      .37966  (Cornfield)
 Prev. frac. pop |         .0834101       |
                 +-----------------------------------------------
                            chi2(1) =     0.40  Pr>chi2 = 0.5269
```

```
. cci 84 49 212 68
                                                                     Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |        84          49  |        133      0.6316
        Controls |       212          68  |        280      0.7571
-----------------+------------------------+----------------------
           Total |       296         117  |        413      0.7167
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         .5498652       |   .3525543    .8574515  (Cornfield)
 Prev. frac. ex. |         .4501348       |   .1425485    .6474457  (Cornfield)
 Prev. frac. pop |         .3408163       |
                 +-----------------------------------------------
                            chi2(1) =     7.00  Pr>chi2 = 0.0081
```

Step 1.4 Crude analysis to determine the effect of APPEAR on SURVIVE

```
. recode   appear 1=0 2=1
(72 changes made)
```

```
. cc  survive  appear  [freq=freq]

                 | appear                 |             Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       323         231  |        554       0.5830
        Controls |        97         113  |        210       0.4619
-----------------+------------------------+----------------------
           Total |       420         344  |        764       0.5497
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         1.628911       |  1.183985    2.241029  (Cornfield)
 Attr. frac. ex. |         .3860928       |  .1553948    .5537763  (Cornfield)
 Attr. frac. pop |         .2251046       |
                 +-----------------------------------------------
                               chi2(1) =     9.03  Pr>chi2 = 0.0027
```

Step 1.5 Crude analysis to determine the effect of INFLAM on SURVIVE

```
. recode  inflam 1=1 2=0
(72 changes made)
```

```
. cc  survive   inflam [freq=freq]

                 | inflam                 |             Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+----------------------
           Cases |       444         110  |        554       0.8014
        Controls |       166          44  |        210       0.7905
-----------------+------------------------+----------------------
           Total |       610         154  |        764       0.7984
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |          1.06988       |  .7237485    1.581853  (Cornfield)
 Attr. frac. ex. |         .0653153       | -.3816954    .3678301  (Cornfield)
 Attr. frac. pop |         .0523466       |
                 +-----------------------------------------------
                               chi2(1) =     0.11  Pr>chi2 = 0.7358
```

From the above analysis, we can report the results as follows:

In an exploratory data analysis, effect of each variable on survival were assessed. In summary, there is a statistically significant association between centre, age, appearance and survival (Table 1). The odds of surviving are lower in Boston and Glamorgan as compared with Tokyo; are lower in women aged 70+ as compared with those <50; are higher in those with benign nuclear grade as compared with malignant appearance. There is not a statistically significant relationship between inflammation and survival.

Table 1. Relationship between each variable and survival.

Variable	OR	95% CI	p-value
Center			0.004
Tokyo	1		
Boston	0.54	0.37 to 0.80	
Gilanorgar	0.59	0.39 to 0.88	
Age			0.024
<50	1		
50-69	0.89	0.62 to 1.28	
70+	0.55	0.35 to 0.86	
Inflammation			0.736
Greater	1		
Minimal	1.07	0.72 to 1.58	
Appearance			0.003
Malignant	1		
Benign	1.63	1.18 to 2.24	

Part 2. Stratified analysis : investigate effect of each variable (esp. interaction effect or any sparse data that could affect modeling) on the relationship between other variable and survival

```
. cc  survive inflam  [freq=freq], by( center)

          center |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               1 |    1.42735      .7839431   2.601131      8.472414 (Cornfield)
               2 |   1.647059      .7891301   3.441417      5.106719 (Cornfield)
               3 |   .6809927      .3073596   1.513491      7.475113 (Cornfield)
-----------------+-------------------------------------------------
           Crude |    1.06988      .7237485   1.581853                (Cornfield)
    M-H combined |   1.215654      .8130749   1.817561
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(2) =     2.86  Pr>chi2 = 0.2393

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.91
                                                Pr>chi2 =    0.3409
```

```
. cc  survive inflam  [freq=freq], by(age)

             age |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               1 |   1.080196     .5696079   2.051455      8.728571 (Cornfield)
               2 |   1.213333     .6855999   2.149144      10.25641 (Cornfield)
               3 |   .9102564     .3461299   2.404947      4.105263 (Cornfield)
-----------------+-------------------------------------------------
           Crude |    1.06988     .7237485   1.581853                (Cornfield)
    M-H combined |    1.10912     .7473107   1.646099
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(2) =     0.25  Pr>chi2 = 0.8819

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.26
                                                Pr>chi2 =    0.6075
```

```
. cc  survive inflam  [freq=freq], by( appear)

          appear |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   .8362229     .5209398   1.342702      18.77907 (Cornfield)
               1 |   .9271111     .3973375   2.169064      5.357143 (Cornfield)
-----------------+-------------------------------------------------
           Crude |    1.06988     .7237485   1.581853                (Cornfield)
    M-H combined |    .856396     .5639342   1.300531
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.04  Pr>chi2 = 0.8385

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      0.53
                                                Pr>chi2 =    0.4667
```

```
. cc  survive  appear  [freq=freq], by( center)
          center |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               1 |   2.131579     1.199045   3.788206      7.862069 (Cornfield)
               2 |   1.413631     .8263451      2.419      10.96047 (Cornfield)
               3 |   1.594406     .8929255   2.845867      9.058824 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.628911     1.183985   2.241029                (Cornfield)
    M-H combined |   1.674815     1.209033   2.320041
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(2) =     1.07  Pr>chi2 = 0.5849

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      9.68
                                                Pr>chi2 =    0.0019
```

```
. cc  survive appear  [freq=freq], by( age)

             age |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               1 |   2.485185     1.414493   4.364453      7.714286 (Cornfield)
               2 |   1.605178     .9981418   2.581484      13.20513 (Cornfield)
               3 |   .9078947     .4483703   1.839268             8 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.628911     1.183985   2.241029                (Cornfield)
    M-H combined |   1.647031     1.194434   2.271127
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(2) =     4.73  Pr>chi2 = 0.0938

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      9.36
                                                Pr>chi2 =    0.0022
```

```
. cc  survive  appear  [freq=freq], by( inflam)

          inflam |       OR       [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
               0 |   1.554622     .6293684   3.819763      3.863636 (Cornfield)
               1 |   1.723592     1.198791   2.478281      21.54098 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   1.628911     1.183985   2.241029               (Cornfield)
    M-H combined |   1.697894     1.209561   2.383381
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.04  Pr>chi2 = 0.8384

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =      9.46
                                                Pr>chi2 =    0.0021
```

We have investigate effect of some variables, none have been found to be significant effect modifier. However we will consider putting the interaction term "APPEAR*AGER" into the model for further investigation since the p-value of 0.094 seems to be convincing (rule of thumb cutpoint is 0.2).

Part 3. Multivariable analysis : investigate effect of each variable on survival adjusted simultaneously for effect of other variables using logistic regression

Fitting logistic regression model

Step 3.1 Generate all possible two-ways interaction terms

```
. gen ce_ag =  center* age

. gen ce_in =  center*  inflam

. gen ce_ap =  center*  appear

. gen ag_in = age* inflam

. gen ag_ap = age* appear

. gen in_ap =  inflam* appear
```

Step 3.2 All main effect and two-ways interaction were fitted

```
. xi:  logistic   survive   i.center  i.age  inflam  appear  i.ce_ag
i.ce_in i.ce_ap i.ag_in i.ag_ap  in_ap  [freq=freq]
i.center             Icente_1-3   (naturally coded; Icente_1 omitted)
i.age                Iage_1-3     (naturally coded; Iage_1 omitted)
i.ce_ag              Ice_ag_1-9   (naturally coded; Ice_ag_1 omitted)
i.ce_in              Ice_in_0-3   (naturally coded; Ice_in_0 omitted)
i.ce_ap              Ice_ap_0-3   (naturally coded; Ice_ap_0 omitted)
i.ag_in              Iag_in_0-3   (naturally coded; Iag_in_0 omitted)
i.ag_ap              Iag_ap_0-3   (naturally coded; Iag_ap_0 omitted)

Note: Ice_ag_4 dropped due to collinearity.
Note: Ice_ag_9 dropped due to collinearity.
Note: Ice_in_3 dropped due to collinearity.
Note: Ice_ap_3 dropped due to collinearity.
Note: Iag_in_3 dropped due to collinearity.
Note: Iag_ap_3 dropped due to collinearity.

Logit estimates                                   Number of obs   =        764
                                                  LR chi2(18)     =      38.60
                                                  Prob > chi2     =     0.0032
Log likelihood = -429.96552                       Pseudo R2       =     0.0430
------------------------------------------------------------------------------
 survive | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Icente_2 |   .4706606    .211485     -1.677   0.094       .1950886    1.135491
Icente_3 |   1.164176    .582414      0.304   0.761       .4366964    3.103543
  Iage_2 |   .9691127   .4075248     -0.075   0.941       .4250429    2.209611
  Iage_3 |   .8014983   .4652564     -0.381   0.703       .2569179    2.500408
  inflam |   .5532933   .3656969     -0.895   0.371       .1514807     2.02094
  appear |   .8291563   .5826706     -0.267   0.790       .2091578    3.286993
Ice_ag_2 |   .6338761   .1508398     -1.916   0.055       .3976023    1.010555
Ice_ag_3 |   .5490657   .1798358     -1.830   0.067       .2889547    1.043323
Ice_ag_6 |   .8515569    .220565     -0.620   0.535       .5125553    1.414773
Ice_in_1 |   1.805347   1.001043      1.065   0.287       .6089476    5.352314
Ice_in_2 |   2.667878   1.603839      1.632   0.103       .8212054    8.667224
Ice_ap_1 |   1.146924    .528486      0.297   0.766       .4648504    2.829802
Ice_ap_2 |   .7773183   .3401168     -0.576   0.565       .3297268    1.832498
Iag_in_1 |   .9257537   .6369526     -0.112   0.911       .2403463    3.565771
Iag_in_2 |   .9731143   .6295218     -0.042   0.966       .2738477    3.457949
Iag_ap_1 |     2.2147   1.158892      1.520   0.129       .7941532    6.176258
Iag_ap_2 |   1.589314   .7658008      0.962   0.336       .6181068     4.08654
   in_ap |   1.324014   .7168045      0.518   0.604       .4582084    3.825801
------------------------------------------------------------------------------
```

Step 3.3 Interaction terms were eliminated one at a time according to lack of statistical significance. All interaction terms were dropped (Output not shown).

Step 3.4 Determining model with only the main effects

```
. xi: logistic  survive  i.center i.age inflam appear [freq=freq]

i.center             Icente_1-3   (naturally coded; Icente_1 omitted)
i.age                Iage_1-3     (naturally coded; Iage_1 omitted)

Logit estimates                                   Number of obs   =        764
                                                  LR chi2(6)      =      24.49
                                                  Prob > chi2     =     0.0004
Log likelihood = -437.01748                       Pseudo R2       =     0.0273

------------------------------------------------------------------------------
 survive | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Icente_2 |   .5730183   .1212359     -2.632   0.008       .3785089    .8674828
Icente_3 |   .6464286   .1381113     -2.042   0.041       .4252643    .9826122
  Iage_2 |    .952617    .181028     -0.255   0.798       .6563909    1.382529
  Iage_3 |   .6544147   .1574908     -1.762   0.078       .4083229    1.048823
  inflam |   .9823988   .2160256     -0.081   0.936       .6384281    1.511693
  appear |   1.686077   .2987845      2.948   0.003       1.191348    2.386251
------------------------------------------------------------------------------

. lrtest, saving(0)
```

First we try removing "INFLAM" due to the highest p-value of 0.936.

```
. xi: logistic  survive  i.center i.age appear [freq=freq]
i.center              Icente_1-3   (naturally coded; Icente_1 omitted)
i.age                 Iage_1-3     (naturally coded; Iage_1 omitted)

Logit estimates                                   Number of obs   =        764
                                                  LR chi2(5)      =      24.49
                                                  Prob > chi2     =     0.0002
Log likelihood = -437.02075                       Pseudo R2       =     0.0273

------------------------------------------------------------------------------
 survive | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Icente_2 |   .5715436   .1195508     -2.674   0.007       .3793168    .8611854
Icente_3 |   .6446424   .1359573     -2.082   0.037       .4263806     .974631
  Iage_2 |   .9529291   .1810479     -0.254   0.800       .6566591    1.382869
  Iage_3 |   .6542766   .1574399     -1.763   0.078       .4082583    1.048546
  appear |   1.677946   .2799015      3.103   0.002       1.210005    2.326851
------------------------------------------------------------------------------

. lrtest
Logistic:  likelihood-ratio test                      chi2(1)     =       0.01
                                                      Prob > chi2 =     0.9356
```

"INFLAM" has no effect on the model. Now the above model suggests "AGE" might be able to be removed.

```
. lrtest, saving(1)

. xi: logistic  survive  i.center appear [freq=freq]
i.center              Icente_1-3   (naturally coded; Icente_1 omitted)

Logit estimates                                   Number of obs   =        764
                                                  LR chi2(3)      =      20.96
                                                  Prob > chi2     =     0.0001
Log likelihood = -438.78221                       Pseudo R2       =     0.0233
```

```
------------------------------------------------------------------------------
 survive | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Icente_2 |   .5173948   .1033642     -3.298   0.001       .3497603    .7653738
Icente_3 |   .6098661   .1262996     -2.388   0.017       .4064019     .915194
  appear |   1.674864   .2783321      3.103   0.002       1.209275    2.319711
------------------------------------------------------------------------------

. lrtest
Logistic:  likelihood-ratio test                      chi2(3)     =       3.53
                                                      Prob > chi2 =     0.3170
```

"AGE" also has no effect on the model and can be removed. We can notice that removing "INFLAM" and "AGE", the coefficient of other variables in the model were not effect. Precision of the estimate (i.e., the range of 95% confidence intervals) were also more or less the same as the model that all variables are in. Additionally age is known to have some effect of survival. We then decide to choose the model with all main effect to describe factors affecting survival since it is more informative. That is, the effects of each study variable was already adjusted for effects of all other potential confounders.

Thus the final model which all variables are retained in the model is

```
. xi: logistic  survive  i.center i.age inflam appear [freq=freq]
i.center            Icente_1-3   (naturally coded; Icente_1 omitted)
i.age               Iage_1-3     (naturally coded; Iage_1 omitted)

Logit estimates                                   Number of obs   =        764
                                                  LR chi2(6)      =      24.49
                                                  Prob > chi2     =     0.0004
Log likelihood = -437.01748                       Pseudo R2       =     0.0273

------------------------------------------------------------------------------
 survive | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Icente_2 |   .5730183   .1212359     -2.632   0.008       .3785089    .8674828
Icente_3 |   .6464286   .1381113     -2.042   0.041       .4252643    .9826122
  Iage_2 |    .952617    .181028     -0.255   0.798       .6563909    1.382529
  Iage_3 |   .6544147   .1574908     -1.762   0.078       .4083229    1.048823
  inflam |   .9823988   .2160256     -0.081   0.936       .6384281    1.511693
  appear |   1.686077   .2987845      2.948   0.003       1.191348    2.386251
------------------------------------------------------------------------------
```

Summary steps for logistic regression model fitting

1. **All main affects and two-way interactions were fitted.**

2. **Interaction terms were eliminated one at a time according to lack of statistical significance.**

3. **All interaction terms were dropped**

4. **Settled for main effect model - could drop the terms for inflammation and age**

5. **Decide to choose the model with all main effect based on clinical judgement.**

6. **Estimate adjusted odds ratios from the logistic regression model**

Part 4. Reporting the results

A study on three-year survival of breast cancer patients according to two histologic criteria, age, and diagnostic center involved 764 patients. Among these, 554 patients still survived at three years. The proportion of three-year survival was 72.5% (95%CI: 69.3% to 75.7%).

Table 2 summarizes effects of selected variables on the survival. In an exploratory data analysis, effect of each variable on survival were assessed. In summary, after adjusting for the effect of all other variables in the Table, there is a statistically significant association between centre and survival (p-value = 0.003) and appearance and survival (p-value = 0.002). The odds of surviving are lower in Boston and Glamorgan as compared with Tokyo, and are higher in those with benign nuclear grade as compared with malignant appearance. However the magnitude of such differences were quite small (i.e., all ORs close to 1 and the largest possible is 2.6 - based on the lower limit of 95%CI of OR for Boston compared to Tokyo). There is no statistically significant

relationship between inflammation and survival (p-value = 0.936) nor age and survival (p-value = 0.317). The similarity between crude odds ratios and adjusted ones suggested that there was no confounding effect of all variables presented in the Table. Interaction effects were also not detected.

Table 2. Crude odds ratios and odds ratios adjusted for the effects of all other variables in the Table describing relationship between the variable and survival.

Variable	No.	Survive (%)	Crude OR	Adjusted OR	95% CI	p-value*
Center						0.003
Tokyo	290	79.3	1			
Boston	253	67.6	0.54	0.57	0.38 to 0.87	
Gilanorgar	221	69.2	0.59	0.64	0.43 to 0.98	
Age						0.317
<50	280	75.7	1			
50-69	351	73.5	0.89	0.95	0.66 to 1.38	
70+	133	63.2	0.55	0.65	0.41 to 1.05	
Inflammation						0.936
Greater	154	71.4	1			
Minimal	610	72.8	1.07	0.98	0.64 to 1.51	
Appearance						0.002
Malignant	344	67.2	1			
Benign	420	76.9	1.63	1.69	1.19 to 2.39	

*** p-value from likelihood ratio tests**

Answers for the exercise in Chapter 7

1. Log-linear model can be fitted as follows:

Firstly we enter the data to Stata using the following format.

beh	ris	adv	freq
1	1	1	16
1	2	1	7
1	1	2	15
1	2	2	34
1	1	3	5
1	2	3	3
2	1	1	1
2	2	1	1
2	1	2	3
2	2	2	8
2	1	3	1
2	2	3	3

This is a 2-by-2-by-3 Table. The analysis using log-linear modeling applied to these data yields the following results for all eight possible models:

Saturated Model:

log-frequency = BEH + RIS + ADV + BEH*RIS
+ BEH*ADV + RIS*ADV
+ BEH*RIS*ADV

1. Model:

log-frequency = BEH + RIS + ADV + BEH*RIS
+ BEH*ADV + RIS*ADV

```
. loglin freq  beh ris adv, fit( beh, ris, adv, beh ris, beh adv,
ris adv)

Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv, beh ris, beh adv, ris adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -21.374741
Iteration 1: Log Likelihood = -20.854752
Iteration 2: Log Likelihood = -20.845886

Poisson regression                                     Number of obs    =       12
Goodness-of-fit chi2(2)     =     0.943                Model chi2(9)    =   99.775
```

```
Prob > chi2                    =    0.6241            Prob > chi2      =  0.0000
Log Likelihood                 =   -20.846            Pseudo R2        =  0.7053

------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -2.664174    .786254     -3.388   0.001      -4.205204   -1.123145
    AB22 |    .589859   .6150592      0.959   0.338      -.6156349    1.795353
    AC22 |   .7346284   .8388175      0.876   0.381      -.9094237    2.378681
    AC23 |   1.661531   .9665848      1.719   0.086      -.2329406    3.556002
      B2 |  -.8047879   .4337849     -1.855   0.064      -1.654991    .0454149
    BC22 |   1.555037   .5159305      3.014   0.003       .5438317    2.566242
    BC23 |    .609435   .7387509      0.825   0.409      -.8384902     2.05736
      C2 |  -.0110953   .3469391     -0.032   0.974      -.6910836    .6688929
      C3 |  -1.286667   .5106199     -2.520   0.012      -2.287464   -.2858708
   _cons |   2.765876   .2478813     11.158   0.000       2.280037    3.251714
------------------------------------------------------------------------------
```

2. Model:
log-frequency = BEH + RIS + ADV + BEH*RIS + BEH*ADV

```
. loglin freq  beh ris adv, fit( beh, ris, adv, beh ris, beh adv)
Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv, beh ris, beh adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -27.003891
Iteration 1: Log Likelihood = -26.049576
Iteration 2: Log Likelihood = -26.03476
Iteration 3: Log Likelihood = -26.034744

Poisson regression                                    Number of obs    =      12
Goodness-of-fit chi2(4)        =    11.321            Model chi2(7)    =  89.397
Prob > chi2                    =    0.0232            Prob > chi2      =  0.0000
Log Likelihood                 =   -26.035            Pseudo R2        =  0.6319

------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -2.867615   .8366197     -3.428   0.001      -4.507359    -1.22787
    AB22 |   .6747983   .5777875      1.168   0.243      -.4576444    1.807241
    AC22 |   .9484221   .8091944      1.172   0.241      -.6375697    2.534414
    AC23 |     1.7492   .9583727      1.825   0.068      -.1291761    3.627576
      B2 |   .2006706   .2247333      0.893   0.372      -.2397986    .6411397
      C2 |    .756326   .2527576      2.992   0.003       .2609301    1.251722
      C3 |  -1.056053    .410461     -2.573   0.010      -1.860542   -.2515639
   _cons |   2.336987   .2423964      9.641   0.000       1.861898    2.812075
------------------------------------------------------------------------------
```

3. Model:
log-frequency = BEH + RIS + ADV + BEH*RIS + RIS*ADV

```
. loglin freq  beh ris adv, fit( beh, ris, adv, beh ris, ris adv)
Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv, beh ris, ris adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -23.313782
Iteration 1: Log Likelihood = -22.448441
Iteration 2: Log Likelihood = -22.433456

Poisson regression                                    Number of obs    =       12
Goodness-of-fit chi2(4)     =      4.118              Model chi2(7)    =   96.600
Prob > chi2                 =     0.3903              Prob > chi2      =   0.0000
Log Likelihood              =    -22.433              Pseudo R2        =   0.6828

------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -1.974081   .4772605     -4.136   0.000      -2.909495   -1.038668
    AB22 |   .6747981   .5777873      1.168   0.243      -.4576443     1.80724
      B2 |  -.8648807   .4382707     -1.973   0.048      -1.723875    -.005886
    BC22 |    1.60107   .5130191      3.121   0.002       .5955707    2.606569
    BC23 |   .7537717   .7191361      1.048   0.295      -.6557092    2.163253
      C2 |   .0571585   .3381997      0.169   0.866      -.6057008    .7200178
      C3 |  -1.041454   .4748581     -2.193   0.028      -1.972159   -.1107491
   _cons |    2.70316   .2494214     10.838   0.000       2.214303    3.192017
------------------------------------------------------------------------------
```

4. Model:
log-frequency = BEH + RIS + ADV + BEH*ADV + RIS*ADV

```
. loglin freq  beh ris adv, fit( beh, ris, adv, beh adv, ris adv)
Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv, beh adv, ris adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -21.969559
Iteration 1: Log Likelihood = -21.339813
Iteration 2: Log Likelihood = -21.326447
Iteration 3: Log Likelihood = -21.326431

Poisson regression                                    Number of obs    =       12
Goodness-of-fit chi2(3)     =      1.904              Model chi2(8)    =   98.814
Prob > chi2                 =     0.5926              Prob > chi2      =   0.0000
Log Likelihood              =    -21.326              Pseudo R2        =   0.6985

------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -2.442347   .7372098     -3.313   0.001      -3.887252   -.9974424
    AC22 |    .948422   .8091944      1.172   0.241      -.6375698    2.534414
    AC23 |     1.7492   .9583727      1.825   0.068      -.1291762    3.627576
      B2 |  -.7537717   .4287465     -1.758   0.079      -1.594099    .0865559
    BC22 |   1.601069   .5130191      3.121   0.002       .5955704    2.606568
    BC23 |   .7537717   .7191362      1.048   0.295      -.6557093    2.163253
      C2 |  -.0619841   .3487103     -0.178   0.859      -.7454436    .6214755
      C3 |  -1.363537    .520226     -2.621   0.009      -2.383162    -.343913
   _cons |   2.749832   .2496033     11.017   0.000       2.260618    3.239045
------------------------------------------------------------------------------
```

5. Model:
log-frequency = BEH + RIS + ADV + BEH*RIS

```
. loglin freq  beh ris adv, fit( beh, ris, adv, beh ris)
Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv, beh ris
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -29.197449
Iteration 1: Log Likelihood = -27.87973
Iteration 2: Log Likelihood = -27.862915

Poisson regression                                Number of obs    =       12
Goodness-of-fit chi2(6)      =     14.977         Model chi2(5)    =   85.741
Prob > chi2                  =     0.0204         Prob > chi2      =   0.0000
Log Likelihood               =    -27.863         Pseudo R2        =   0.6061

------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -1.974081   .4772606     -4.136   0.000      -2.909495   -1.038667
    AB22 |   .6747982   .5777874      1.168   0.243      -.4576443    1.807241
      B2 |   .2006706   .2247333      0.893   0.372      -.2397986    .6411397
      C2 |   .8754688   .2380476      3.678   0.000        .408904    1.342034
      C3 |  -.7339692   .3511884     -2.090   0.037      -1.422286   -.0456525
   _cons |   2.227684   .2397259      9.293   0.000        1.75783    2.697538
------------------------------------------------------------------------------
```

6. Model:
log-frequency = BEH + RIS + ADV + BEH*ADV

```
. loglin freq  beh ris adv, fit( beh, ris, adv, beh adv)
Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv, beh adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -27.803162
Iteration 1: Log Likelihood = -26.767899
Iteration 2: Log Likelihood = -26.755905
Iteration 3: Log Likelihood = -26.75589
Poisson regression                                Number of obs    =       12
Goodness-of-fit chi2(5)      =     12.763         Model chi2(6)    =   87.955
Prob > chi2                  =     0.0257         Prob > chi2      =   0.0000
Log Likelihood               =    -26.756         Pseudo R2        =   0.6217
------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -2.442347   .7372098     -3.313   0.001      -3.887252   -.9974425
    AC22 |    .948422   .8091943      1.172   0.241      -.6375698    2.534414
    AC23 |     1.7492   .9583727      1.825   0.068      -.1291761    3.627576
      B2 |   .3117796   .2055417      1.517   0.129      -.0910747    .7146339
      C2 |   .7563261   .2527576      2.992   0.003       .2609302    1.251722
      C3 |  -1.056053    .410461     -2.573   0.010      -1.860542   -.2515639
   _cons |   2.274355    .239915      9.480   0.000       1.804131     2.74458
------------------------------------------------------------------------------
```

7. Model:
log-frequency = BEH + RIS + ADV + RIS*ADV

```
. loglin freq  beh ris adv, fit( beh, ris, adv, ris adv)
Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv, ris adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -24.179001
Iteration 1: Log Likelihood = -23.173065
Iteration 2: Log Likelihood = -23.154617
Iteration 3: Log Likelihood = -23.154602

Poisson regression                                    Number of obs    =      12
Goodness-of-fit chi2(5)     =     5.560               Model chi2(6)    =  95.158
Prob > chi2                 =    0.3514               Prob > chi2      =  0.0000
Log Likelihood              =   -23.155               Pseudo R2        =  0.6726
------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -1.548813   .2670647     -5.799   0.000       -2.07225   -1.025376
      B2 |  -.7537718   .4287465     -1.758   0.079      -1.594099    .0865559
    BC22 |    1.60107   .5130191      3.121   0.002       .5955706    2.606569
    BC23 |   .7537717   .7191362      1.048   0.295      -.6557093    2.163253
      C2 |   .0571586   .3381998      0.169   0.866      -.6057008    .7200179
      C3 |  -1.041454   .4748581     -2.193   0.028      -1.972158    -.110749
   _cons |   2.640529   .2470106     10.690   0.000       2.156397    3.124661
------------------------------------------------------------------------------
```

8. Model:
log-frequency = BEH + RIS + ADV

```
. loglin freq  beh ris adv, fit( beh, ris, adv)
Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -30.132355
Iteration 1: Log Likelihood = -28.606613
Iteration 2: Log Likelihood = -28.584061

Poisson regression                                    Number of obs    =      12
Goodness-of-fit chi2(7)     =    16.419               Model chi2(4)    =  84.299
Prob > chi2                 =    0.0216               Prob > chi2      =  0.0000
Log Likelihood              =   -28.584               Pseudo R2        =  0.5959

------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -1.548813   .2670646     -5.799   0.000       -2.07225   -1.025376
      B2 |   .3117797   .2055417      1.517   0.129      -.0910746     .714634
      C2 |   .8754688   .2380476      3.678   0.000       .4089041    1.342034
      C3 |  -.7339691   .3511884     -2.090   0.037      -1.422286   -.0456525
   _cons |   2.165052   .2372165      9.127   0.000       1.700117    2.629988
------------------------------------------------------------------------------
```

Summary results:
Denoted BEH = 1, RIS = 2, and ADV = 3.

Model	*Likelihood ratio Chi-square*	*Degree of freedom*	*p-value*
All pairwise association			
1. u_{123}	**0.943**	**2**	**0.624**
Conditional independence			
2. $u_{12} = u_{123} = 0$	**11.321**	**4**	**0.023**
3. $u_{13} = u_{123} = 0$	**4.118**	**4**	**0.390**
4. $u_{23} = u_{123} = 0$	**1.904**	**3**	**0.593**
Partial independence			
5. $u_{12} = u_{13} = u_{123} = 0$	**14.977**	**6**	**0.020**
6. $u_{12} = u_{23} = u_{123} = 0$	**12.763**	**5**	**0.026**
7. $u_{12} = u_{23} = u_{123} = 0$	**5.560**	**5**	**0.351**
Complete independence			
8. $u_{12} = u_{13} = u_{23} = u_{123} = 0$	**16.419**	**7**	**0.0216**

The model that fitted well to the data and yet less complicated is Model 7. It can also be expressed as

log-frequency = BEH + RIS + ADV + RIS*ADV

Comparing the two models that adequately fitted the data (Models 7 and 4) we have

$$G_7^2 - G_4^2 \quad = \quad \mathbf{5.560 - 1.904 = 3.656}$$

$$\textbf{with} \quad \mathbf{5 - 3} \quad = \quad \mathbf{2}\ \boldsymbol{df}$$

Use Stata to find a p-value as follows:

```
. disp chiprob(2, 3.656)
.16073473
```

Thus adding the term BEH*ADV to Model 7 did not improve the fit. This term can really be removed.

We examine further by comparing Model 7 with Model 3

$G_7^2 - G_3^2$ = **5.560 - 4.118 = 1.442**

with 5 - 4 = 1 *df* ; p-value = 0.230

and Model 7 with Model 1

$G_7^2 - G_1^2$ = **5.560 - 0.943 = 4.617**

with 5 - 2 = 3 *df*; p-value = 0.202

Thus there was no need to add any other two-way interaction terms to Model 7 and it is the best model for describing the data. We can examine the residual from the output below.

```
. loglin freq  beh ris adv, fit( beh, ris, adv, ris adv) resid
Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv, ris adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -24.179001
Iteration 1: Log Likelihood = -23.173065
Iteration 2: Log Likelihood = -23.154617
Iteration 3: Log Likelihood = -23.154602

Poisson regression                                 Number of obs    =      12
Goodness-of-fit chi2(5)     =     5.560            Model chi2(6)    =  95.158
Prob > chi2                 =    0.3514            Prob > chi2      =  0.0000
Log Likelihood              =   -23.155            Pseudo R2        =  0.6726

------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -1.548813   .2670647     -5.799   0.000       -2.07225   -1.025376
      B2 |  -.7537718   .4287465     -1.758   0.079      -1.594099    .0865559
    BC22 |    1.60107   .5130191      3.121   0.002       .5955706    2.606569
    BC23 |   .7537717   .7191362      1.048   0.295      -.6557093    2.163253
      C2 |   .0571586   .3381998      0.169   0.866      -.6057008    .7200179
      C3 |  -1.041454   .4748581     -2.193   0.028      -1.972158    -.110749
   _cons |   2.640529   .2470106     10.690   0.000       2.156397    3.124661
------------------------------------------------------------------------------
```

```
freq  beh  ris  adv   cellhat    resid   stdres
 16    1    1    1    14.021    1.979    0.529
 15    1    1    2    14.845    0.155    0.040
  5    1    1    3     4.948    0.052    0.023
  7    1    2    1     6.598    0.402    0.157
 34    1    2    2    34.639   -0.639   -0.109
  3    1    2    3     4.948   -1.948   -0.876
  1    2    1    1     2.979   -1.979   -1.147
  3    2    1    2     3.155   -0.155   -0.087
  1    2    1    3     1.052   -0.052   -0.050
  1    2    2    1     1.402   -0.402   -0.340
  8    2    2    2     7.361    0.639    0.236
  3    2    2    3     1.052    1.948    1.900
```

Conclusions:

The model with only one interaction term, i.e., risk index and adversity of school condition, fit the data adequately (p-value = 0.351). This is a partial independence model. It is implied that there is an association between the two variables whilst the behavior is completely independent. Therefore, behavior can be omitted from the table. The two-way contingency table of risk index and adversity of school condition is sufficient to describe this data. That is, from the following table

		Adversity of school condition						
		Low		Medium		High		
Risk index		Not at risk	At risk	Not at risk	At risk	Not at risk	At risk	Total
Behavior	Not deviant	16	7	15	34	5	3	80
	Deviant	1	1	3	8	1	3	17
Total		17	8	18	42	6	6	97

then we can get a simpler table shown below.

		Adversity of school condition			Total
		Low	Medium	High	
Risk index	**Not at risk**	**17**	**18**	**6**	**41**
	At risk	**8**	**42**	**6**	**56**
Total		**25**	**60**	**12**	**97**

From this table, we can analyze the data using approaches for analysis of a 2-by-C Table presented in Chapter 3.

We might examine further whether or not behavior can be disregarded, i.e., examining for collapsibility. Notice that the coefficient of RIS*ADV in the models with and without BEH, in ovals respectively, are almost identical. This suggested that the measure of association between risk index and adversity of school condition was not affected by whether or not behavior was accounted for.

```
. loglin freq  beh ris adv, fit( beh, ris, adv, ris adv)
Variable beh = A
Variable ris = B
Variable adv = C
Margins fit: beh, ris, adv, ris adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -24.179001
Iteration 1: Log Likelihood = -23.173065
Iteration 2: Log Likelihood = -23.154617
Iteration 3: Log Likelihood = -23.154602

Poisson regression                                  Number of obs    =      12
Goodness-of-fit chi2(5)     =     5.560             Model chi2(6)    =  95.158
Prob > chi2                 =    0.3514             Prob > chi2      =  0.0000
Log Likelihood              =   -23.155             Pseudo R2        =  0.6726

------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -1.548813   .2670647     -5.799   0.000       -2.07225   -1.025376
      B2 |  -.7537718   .4287465     -1.758   0.079      -1.594099    .0865559
    BC22 |    1.60107   .5130191      3.121   0.002       .5955706    2.606569
    BC23 |   .7537717   .7191362      1.048   0.295      -.6557093    2.163253
      C2 |   .0571586   .3381998      0.169   0.866      -.6057008    .7200179
      C3 |  -1.041454   .4748581     -2.193   0.028      -1.972158    -.110749
   _cons |   2.640529   .2470106     10.690   0.000       2.156397    3.124661
------------------------------------------------------------------------------
```

```
. loglin freq  ris adv, fit(ris, adv, ris adv)
Variable ris = A
Variable adv = B
Margins fit: ris, adv, ris adv
Note: Regression-like constraints are assumed.  The first level of each
variable (and all iteractions with it) will be dropped from estimation.

Iteration 0: Log Likelihood = -50.343819
Iteration 1: Log Likelihood = -45.486633
Iteration 2: Log Likelihood = -45.369843
Iteration 3: Log Likelihood = -45.369675

Poisson regression                                    Number of obs    =       12
Goodness-of-fit chi2(6)     =    49.990               Model chi2(5)    =  50.728
Prob > chi2                 =    0.0000               Prob > chi2      =  0.0000
Log Likelihood              =   -45.370               Pseudo R2        =  0.3586

------------------------------------------------------------------------------
    freq |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      A2 |  -.7537717   .4287465     -1.758   0.079      -1.594099    .0865559
    AB22 |   1.601069   .5130191      3.121   0.002       .5955705    2.606568
    AB23 |   .7537718   .7191362      1.048   0.295      -.6557093    2.163253
      B2 |   .0571586   .3381998      0.169   0.866      -.6057008     .720018
      B3 |  -1.041454   .4748581     -2.193   0.028      -1.972159   -.1107491
   _cons |   2.140066   .2425356      8.824   0.000       1.664705    2.615427
------------------------------------------------------------------------------
```