Master's thesis in Statistics

Effects on Analysis Arising from Confidentialising Data Using Random Rounding

Xiangyin Chen (Monica)

University of Canterbury

Principal Supervisor: Dr. Jennifer Brown

Department of Mathematics and Statistics, University of Canterbury

Co-supervisor: Dr. Judi McWhirter

Department of Statistics, University of Waikato

Richard Penny

Statistics New Zealand

Abstract

Government statistical agencies collect data on individuals. These data can have personal information that will lead to individual identification. The information gathered is often released and used by other agencies. In order to preserve confidentiality (people's privacy) the data are treated in a way that prevents identification. In recent years there has been a rapid increase in the research in the area of confidentiality and statistical disclosure techniques (SDC). We focus on random rounding method, one of the SDC.

In this thesis we use rounded data which have been collected by Statistics NZ. We examine the effect of random rounding in contingency tables. We simulate data, based on rounded data, and actual data and use the general log-linear model and chi-square test for analysis.

Table of Contents

Chapter 1

Statistical Disclosure Control (SDC)

1.1. Introduction

In this chapter, we present an overview of the Statistical Disclosure Control (SDC) concepts, and the importance of data that are released to the public. We introduce several methods of releasing data: data can be released either as microdata or tabular data.

1.1.1. Importance of SDC

In recent years, the need to protect security and privacy information has arisen during data collection for statistical purposes, particularly if these data are to be disseminated. Statistical organizations need to ensure information collected from people or organizations is unable to be viewed by unauthorized people. SDC is a way to protect issues of privacy and individual information. Many countries have a legal framework of some form or another for dealing with protecting the privacy of the individuals (Willenborg and De Waal, 1996). As one example, the United Kingdom has a law named the Data Protection Act of 1984, which was established to protect the privacy of individuals. SDC also helps to protect the data obtained by a statistical office via a survey in case these data are lost or stolen. SDC also helps avoid problems of liability if a breach of confidence should arise. To take one example, imagine a student from Canterbury University gaining access to sensitive data from Statistics NZ. Statistics NZ trusts the University of Canterbury; they do not have the same level of confidence in the student. If the student then misuses the data - or gains access to data they should not - it is the University of Canterbury that will be held accountable. These

reasons explain why releasing data using SDC is a fundamental issue for a statistical office.

The privacy of the individual is very important. For example, suppose salary data have been collected in a small town. Assume these salary data are divided into three groups. If only a few people are categorized in the high salary group, it would be very easy to identify the people who earn a high salary and thus discover the amount of their salary.

Statistics Canada was the first to use a technique known as "random rounding" to deal with privacy issues. Random rounding was then used in the New Zealand Census of Population and Dwellings by the Department of Statistics in 1981 (Ryan and Penny, 1986). We will introduce this technique in Section 1.3.

1.1.2. Principles of Statistic Disclosure Control (SDC)

According to Willenborg and De Waal (1996), SDC incorporates two important concepts: (1) re-identification disclosure, and (2) prediction disclosure. Re-identification disclosure is when a non-authorized individual, known as an attacker, manages to deduce the value of a sensitive variable for the target individual after this individual has been re-identified. Prediction disclosure happens when the attacker is able to use the data, with some degree of confidence, to predict the value of a sensitive variable for some target individual. This problem of re-identification is a major concern.

SDC concerns safeguarding the confidentiality of information that has been collected from people or organizations. If unreleased data are published, this will lead to people's information being able to be indentified. For example, if the original data include the income levels, a user will know other peoples' exact salary, who earns this salary, what their job is, etc. People often do not want this

information revealed to other people. Therefore, we need to release data in a way that protects peoples' privacy. SDC is the way to release the data. If income data have been released in SDC, an attacker will find it hard to discover another persons' salary level, or who earns this salary. People are prevented from accessing actual data with SDC, but SDC is more about allowing access to data while still preserving respondent confidentiality. That means released data with SDC is more safety then allowing access to the actual data. We are concentrating on data releases with SDC rather than data access.

SDC can be used to protect against the identification of an individual or organization. Information on individuals, businesses, and other organizations are at risk of disclosure. Cuppen (2000) gives a definition of a sensitive cell as one where the contribution of an individual respondent contributing to that cell can be disclosed. The SDC method can be used to reduce the risk of disclosure, and to ensure that most of the information will not be stolen and will be protected for longer. The SDC method is built on disclosure risk - the data utility framework and the need to find the balance between managing disclosure risk while maximizing the amount of information in order to be able to prevent minimum information loss to users. When the data have been released will be loss some information, e.g. recoding method in Section 1.2 shows several categories are combined into one new single category.

Shlomo and Young (2006) defined disclosure risk as "identifying individuals in small cells in the data which then leads to attribute disclosure of other sensitive variables". This sort of risk usually happens when statistics are published at low aggregation level such as small geographies or small populations. Take the example of a small town, with only one doctor. If external users know the income group, age group, etc in this small town, and because doctors usually belong to a high salary group, this user will be identify this doctor's income, age, etc. The data breach will occur.

Three main approaches are used for SDC:

(1) legal provision, which has already been discussed;

(2) removing the sensitive information to protect privacy, e.g. the suppression method of SDC, which will be discussed in more detail in Section 1.2.1.

(3) fuzzing the information. i.e. adding noise to the original data via the rounding method, swapping the data, input perturbation etc. We introduce these methods in Section 1.2.

1.2. Data released in SDC

1.2.1. SDC for microdata

Microdata contain records which consist of information at the level of individual respondents such as a person, household or business. When the individual respondents are entered into a microdata set, e.g. "occupation" = "student", "sex" = "female", and "place of residence" = "Auckland", SDC techniques have to be used to protect each individual's privacy. The two main SDC techniques used when releasing microdata are local suppression and global recoding (Willenborg and De Waal, 1996).

• Local suppression: If a variable *X* in one or more records occurs in an unsafe combination, it is then replaced by a missing value. For example, take the combination variable “area” = ”Canterbury " and “enterprise” = ”ice cream firm”. Say Canterbury has only one ice cream firm. Therefore, if we were to replace the variable labels, it would uniquely identify the ice cream firm. In situations such as this, we would use the local suppression method to replace the data. The local suppression method will replace “enterprise” = ”ice cream firm” with “enterprise” = ”missing”, removing the possibility of the firm being identified.

• Global recoding: Several categories are combined into one new single category, one of which contains the variable *X.* Consider an example with the combination of variables “enterprise I” = ”ice cream firm”, “enterprise II” = ”butter making firm”, “enterprise III” = ”milk firm” and “area” = "Canterbury". The ice cream firm, butter making firm and milk firm all belong to the dairy products category. The global recoding method combines these three enterprises into one category: dairy production. The benefit of this is that the user will find it hard to identify this ice cream firm even if they know only one ice cream firm operates in Canterbury.

1.2.2. SDC for tabular data

1.2.2.1. Definitions

A table consists of a set of cells where each cell is characterized by a set of coordinates that consist of combinations of scores on different categorical variables (Willenborg and De Waal, 1996). A problem are often happens is that the variables are identifying variables. If a table aggregates individual responses, we must try to "disaggregate" the table values to protect the individual respondent's privacy and to prevent disclosure.

Tabular data can be classified into two classes. If the frequency or count data present the number of units of analysis in a cell, they are named count data. If the data present the amount each respondent contributes to its tabulation cells, they are called magnitude data (Cox, 1981; Federal committee on statistical methodology, 2005). With any table, the first step is to determine the sensitive cells. The tables of count (or frequency) data and tables of magnitude data use different ways to determine sensitive cells. This thesis describes how to determine the sensitive cells and which methods can be used concealing the sensitive cells.

1.2.2.2. SDC for tables of count data

We now discuss two classes of rules that limit disclosure of sensitive cell for tables of counts or frequencies as introduced in Cox (1981) and the Federal Committee on Statistical Methodology (2005).

The first class consists of table-specific rules, in particular the "*n*-threshold rule". We give an example using Table 1.1 where salary data have been collected from three different areas: area I, area II and area III. The data collector has divided the salary data into the three groups defined as salary group I, II and III. Table 1.1 stores the raw count data that represents the number of participants in each of the three salary groups based on different areas. If an agency requires 5 as the minimum number (threshold) of respondents for a published cell and a cell value is 3, then 2 (5 - 3) is the least amount of protection that must be added to the cell value. If an attacker knows that two people have a high salary in the small town, then the attacker will easily identify these two people, who are in "salary group III" with "area I". These two people's information will be identified. In a table of count data, if the number of respondents is less than some specified number (*n*-threshold rule) in one particular cell, that cell is called a sensitive cell.

Table 1.1 Number of people in each salary group based on different areas

	Salary group I	***Salary group II***	***Salary group III***	***Total***
Area I	23	15	**2**	40
Area II	18	**3**	27	48
Area III	21	22	30	73
Total	62	40	59	161

The second class of rules is special rules that are agency-specific and table-specific. These rules are designed to protect data considered sensitive by the agency. For more details, see Federal Committee on Statistical Methodology (2005, p14).

When a cell is identified as being sensitive, we must protect the sensitive cell as the next step. Several techniques can be used for concealing the sensitive information and thus reducing the disclosure risks. These are table redesign, suppression and rounding. We shall now give a brief overview of these techniques with some examples.

If a large numbers of sensitive cells have been found in certain categories, we can combine the variables, therefore reducing the detail of the table (Willenborg and De Waal, 1996). Note that combining columns or rows should result in larger cell counts. We will represent three methods using the example shown in Table 1.1.

•Table redesign: The objective of this method is to combine categories to reduce information displayed within Table 1.1. According to the table redesign method, Salary Group II and Salary Group III can be combined as one new variable to become Salary Group II & III, as illustrated in Table 1.2.

Table 1.2 Number of people in each salary group based on different areas (after table redesign)

	Salary group I	***Salary group II& III***
Area I	23	17
Area II	18	30
Area III	21	52

• Suppression: The technique known as primary suppression deletes the value of the sensitive cell and replaces it by a symbol. We use 5 as the threshold to apply to Table 1.1, and obtain the count cells (1,3), and (2,2) as the primary suppressions. The sensitive cell is suppressed and replaced by the symbol *X* (Willenborg and De Waal, 1996). Table 1.3 shows the raw data released using the primary suppression method.

Table 1.3 Number of people in each salary group based on different areas (after primary suppression)

	Salary group I	***Salary group II***	***Salary group III***	***Total***
Area I	23	15	*X*	40
Area II	18	*X*	27	48
Area III	21	22	30	73
Total	62	40	59	161

If the suppressed cell can be derived by subtractions from the marginal totals, suppression will be applied again which is named secondary suppression. Consider the following calculation: the cell count (2,2) can be written as cell (4,2) - cell (3,2) - cell (1,2) = 40 - 22 - 15 = 3. If given a huge table that selection of cells for secondary suppression is a complicated process. Table 1.4 presents the secondary suppressions, where, out of a total of 9 cells, only 5 cells are published. This leads to the biggest problem with suppression: when a sensitive cell has been suppressed in a table, it leads to a considerable amount of data loss. Note that suppression is not a common method to use for count data. This is because after primary suppression and secondary suppression, getting information from the table will become a complex task because information has been lost.

Table 1.4 Number of people in each salary group based on different areas (after secondary suppression)

	Salary group I	*Salary group II*	*Salary group III*	*Total*
Area I	23	*X*	*X*	40
Area II	18	*X*	*X*	48
Area III	21	22	30	73
Total	62	40	59	161

• Rounding: The suppression method leads to data loss. Thus the rounding method has been developed. Tabular data is rounded by the base value multiple of an integer. According to Willenborg and De Waal (1996), tabular data can be rounded by several methods. The first one is conventional rounding, where the original cells are rounded to the nearest multiple of a fixed rounding base. For example, using a rounding base of 5, if the original cells end in 3 or 4, these cells are rounded up and replaced by values ending in 5. If original cells ending in 1 or 2, these cells are rounded down and replaced by values ending in 0.

Conventional rounding is easy to use but it has its limitations. For example, if two cells have a value of 2, the marginal total is 4. When conventional rounding with rounding base 5 is applied to this table, each of these two cells will round to 0, but the marginal total will round to 5. Another method called random rounding has been developed. In Table 1.5, we used random rounding to base 5 for the original data of Table 1.1. Table 1.5 illustrates the problems with random rounding: the table after random rounding could lead to loss of confidence in the numbers.

Table 1.5 Number of people in each salary group based on different areas (after random rounding to base 5)

	Salary group I	*Salary group II*	*Salary group III*	*Total*
Area I	25	15	5	40
Area II	20	5	30	48
Area III	20	25	30	73
Total	62	40	59	161

Note: because of rounding, the numbers may not add up to 100%

The Federal Committee on Statistical Methodology (2005) introduced controlled rounding. This procedure has been developed to solve the additive problem. It is a special case of rounding, where the sum of the cells after rounding is equal to the appropriate published marginal totals. Table 1.6 displays the results of applying the controlled rounding procedure to Table 1.1.

Table 1.6 Number of people in each salary group based on different areas (after controlled rounding to base 5)

	Salary group I	*Salary group II*	*Salary group III*	*Total*
Area I	25	15	0	40
Area II	20	0	30	50
Area III	20	25	30	75
Total	65	40	60	165

Cox (1981) introduced a new disclosure control method, controlled tabular adjustment (CTA). This new approach is similar to controlled rounding. We need replace sensitive cells by safe values, a safe value being one that is a “sufficient

distance" away from the original cell value. Then we use linear programming to rebalance tabulations, i.e. we add a value of "sufficient distance" to the cell total. Note that this method is most valuable for releasing tables of magnitude data.

Adding random noise, input perturbation or data swapping is another disclosure protection method. Using these methods prior to releasing microdata ensures that any tables generated from the released microdata are fully protected. The random noise method of disguising sensitive variables, such as income, is to add or multiply by random numbers (Federal committee on statistical methodology, 2005).

Data swapping is a method that swaps the values of variables for records that match on a representative key (Federal committee on statistical methodology, 2005). Table 1.7 illustrates the data released after data swapping. In this example, the second area and the fourth area are swapped.

Table 1.7 Table released by data swapping

	Salary group
Area I	14
Area IV	22
Area III	25
Area II	21

1.2.2.3. SDC for Table of magnitude data

The first way suggested to determine the sensitive cells is to use the *(n,k)*-dominance rule. In recent years, the dominance rule, using two parameters *n* and *k* has been most commonly used. A cell is called sensitive if the sum of the largest *n* contributions account for more than *k%* of the total cell value (Willenborg and De Waal, 1996). The second way is using the *(p,q)* prior-posterior rule that also has two parameters *p* and *q* with *p>q*. It is assumed that each respondent can estimate the contribution of each other respondent to within less than *q%*. A cell is considered sensitive if it is possible for someone to estimate the contribution of an individual respondent to within less than *p%* (Willenborg and De Waal, 1996). For more information on this rule, the reader is referred to (Cox, 1987).

1.3. SDC with random rounding (controlled rounding)

1.3.1. SDC with random rounding

Random rounding is a technique that protects the confidential information with minimum loss of information. In random rounding, statistical agencies modify the original data by rounding up or rounding down by a base multiplier, e.g. Table1.8.

For rounding, base 3 or 5 are the most common choices. Statistics NZ chose base 3 for releasing the data from the 1981 Census (Ryan, 1981). Random rounding is more flexible to deal with than table redesign or suppression.

This report focuses on random rounding of two- and three-dimensional tables. According to the Bureau of Census, 90% of tables that have been disclosed are two-dimensional. Three-dimensional tables are the majority of higher dimensional tables. Controlled random rounding in three-dimensional tables is more difficult than in two-dimensional tables (Kelly et al, 1990).

1.3.2. Effect of random rounding

In random rounding, the original data are rounded by a multiple of a common base in order to protect the data. Table 1.8 demonstrates random rounding to base 3 and base 5.

Table 1.8 Random rounding to base 3 and base 5.

Original data	*Base 3*		*Base 5*	
	Rounded down (with probability)	***Rounded up (with probability)***	***Rounded down (with probability)***	***Rounded up (with probability)***
0	0 (null)	0 (null)	0 (null)	0 (null)
1	0 (2/3)	3 (1/3)	0 (4/5)	5 (1/5)
2	0 (1/3)	3 (2/3)	0 (3/5)	5 (2/5)
3	3 (null)	3 (null)	0 (2/5)	5 (3/5)
4	3 (2/3)	6 (1/3)	0 (1/5)	5 (4/5)
5	3 (1/3)	6 (2/3)	5 (null)	5 (null)
6	6 (null)	6 (null)	5 (4/5)	10 (1/5)
7	6 (2/3)	9 (1/3)	5 (3/5)	10 (2/5)
8	6 (1/3)	9 (2/3)	5 (2/5)	10 (3/5)

Note: base 3: 8 (original data) =6*1/3 + 9*2/3; base 5: 8 (original data) = 5*2/5+10*3/5**.**

The formula for finding the interval of possible data is 2 * (*base* - 1) +1. In other words, when rounding to base 3, the width of possible cells is 5. Put simply, using base 3 random rounding, a value that appears in a cell could be rounded from five original values (Ryan, 1981):

- rounded data - 2
- rounded data - 1
- rounded data
- rounded data + 1
- rounded data + 2

Using 5 as the rounding base, the figure published in a cell could be derived from any one of (2*(5-1)+1) nine parent (original) cells:

- rounded data - 4
- rounded data - 3
- rounded data - 2
- rounded data - 1
- rounded data
- rounded data + 4
- rounded data + 3
- rounded data + 2
- rounded data + 1

Note that this formula does not apply if the true value of cell is equal to zero. If the value of cell is equal to zero, the width of possible parent cells is equal to the

base number, e.g. if the cell value is 0, using base 3, possible parent cells are 0,1 and 2.

The discrepancy between the original data and its random rounded value is the value of "noise". The expected value of discrepancy is zero. For example, using data rounded to base 3 we assume that case 1 is x = 0 (mode 3), case 2 is x = 1 (mod 3), and case 3 is x = 2 (mod 3). For case 1: $d = 0$ with probability $p = 1$, the expected value discrepancy is 0. For case 2: $d = 1$ with probability $p = \frac{2}{3}$, and $d = -2$ with $p = \frac{1}{3}$, so the expected value discrepancy is $\frac{2}{3} \times (1) + \frac{1}{3} \times (-2) = 0$. For case 3: $d = 2$ with probability $p = \frac{1}{3}$, and $d = -1$ with $p = \frac{2}{3}$, so the expected value discrepancy is $\frac{1}{3} \times (2) + \frac{2}{3} \times (-1) = 0$ (Ryan and Penny, 1986).

According to Ryan (1981), the effect of random rounding can be obtained by the chi-square test. The examples below illustrates this process.

Cochran (1952) introduced the expression for computing the statistic x^2, namely :

$$x^2 = \sum \frac{(Observed \quad frequency - Expected \quad frequency)^2}{Expected \quad frequency} \tag{1.1}$$

where expected frequency is,

$$E_{ij} = \frac{n_{i.} n_{.j}}{N} \tag{1.2}$$

If the given table is a 2×2 table, e.g. Table 1.9, we can reduce the formula (1.1) to the following simplified form:

$$x^2 = \frac{N(ab - bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{1.3}$$

Table 1.9. General 2 x 2 contingency table

		A		
		Type I	***Type II***	***Total***
B	***Type I***	A	b	a + b
	Type II	C	d	c + d
	Total	a + c	b + d	N = a + b + c + d

Let us consider the case of the original data presented in Table 1.10.

Table 1.10 Original table in a 2x2 contingency table

5	7	12
11	12	23
16	19	35

Eight possible tables can be obtained by rounding to base 3. Next, let us calculate the chi-square test for each possible rounded table. The bias happened when the expected value of $\left(\hat{x}^2 - \overline{x}^2\right)$ is not equal to zero. Note $\hat{x}$ is calculate by Equation (1.1) for each possible rounded table. The $\overline{x}$ is the expected value of all possible rounded tables' $\hat{x}$.

This explanation shows us how to measure the bias when original table is known. Another way is to measure the effect of random rounding by using chi-square when the rounded table is given (Table 1.11).

Table 1.11 2x2 contingency table randomly rounded using base 3

3	6	9
9	12	21
12	18	30

Using base 3 rounding, Table1.11 could be derived from one of 625 (5^4) parent tables. Ryan (1981) shows the three steps to measure the effect of random rounding by chi-square:

(1) A cell's prior distribution of the parent table: base 3 has been assumed:

P (rounded data - 2) = 1/9, *P* (rounded data - 1) = 2/9, *P* (rounded data) = 3/9, *P* (rounded data + 1) = 2/9, *P* (rounded data + 2) = 1/9;

(2) Suppose these prior distributions are independent.

(3) Find the posterior distribution for each cell value, according to the random rounding method.

As discussed above, we present different types of data with different SDC methods. In this thesis, we focus on count data with random rounding. We want to show the effect of random rounding in two ways, one is to show the effect of random rounding using the chi-square test. The other one is to reflect the effect of random rounding on the log-linear model.

We consider two important issues with random rounding to show the bias effect of random rounding. Suppose we have 100 sample rounded tables. Next, we calculate the chi-square value of these 100 rounded tables, denoted as $x_1^2,...,x_{100}^2$. The expected value of $(x_1^2,...,x_{100}^2)$ is equal to the chi-square value of

the original table, according to expected value of discrepancy is zero (more detail in P20). This raises another important question regarding privacy and confidentiality. If Statistics NZ only published one random rounded table with the original table's chi-square value given to the n^{th} decimal place, is this safe to published while maintaining confidentiality? Chapter 3 will focus on this problem.

Another aspect of SDC relates to estimates in the log-linear model. Do estimated based on the rounded table change the hypothesis decision compared with estimates based on the actual table? Chapter 4 and 5 present this analysis for different contingency tables.

Chapter 2

Log-linear models

This chapter discusses the use of log-linear models for two-, three- and higher-dimensional contingency tables. We first give an overview of the generalized linear models (GLM) in which the response variable is from some member of the exponential family of distribution. We then introduce the concept of log-linear models and show ways to determine a minimal adequate model in R.

2.1. Generalized linear model (GLM)

2.1.1. The exponential family

The exponential family of distribution is a one-parameter distribution. A probability distribution $f(y)$ is said to be a member of the exponential family if it can be written as follows:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)} \tag{2.1}$$

where:

y : vector of measurements,

θ : parameter of the family,

a, b, s and t are known functions.

We can rewrite Equation 2.1 in the following form:

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \tag{2.2}$$

where:

$s(y) = \exp d(y)$,

and $t(\theta) = \exp c(\theta)$.

If $a(y) = y$, the distribution is called canonical (Dobson, 2001).

As an example, the binomial random variable $y \in \{0,1,2,3...,n\}$, with parameter π is shown below:

$$
\begin{aligned}
f(y;\pi) &= \binom{n}{y}\pi^{y}(1-\pi)^{n-y} \\
&= \exp\left(\log\left(\binom{n}{y}\Pi^{y}(1-\pi)^{n-y}\right)\right) \\
&= \exp\left(y\log(\pi) - y\log(1-\pi) + n\log(1-\pi) + \log\binom{n}{y}\right) \\
&= \exp\left(y\log(\frac{\pi}{1-\pi}) + n\log(1-\pi) + \log\binom{n}{y}\right)
\end{aligned}
$$

This can be rewritten as Equation 2.2 with $a(y) = y$, $b(\theta) = \log(\frac{\pi}{1-\pi})$, $c(\theta) = n\log(1-\pi)$, and $d(y) = \log(\binom{n}{y})$.

The Normal or Gaussian, Poisson, gamma, inverse Gaussian, geometric and negative binomial distributions can all be written in the form of Equation 2.2; therefore, they all belong to the exponential family.

2.1.2. Introduction to the generalized linear model (GLM)

Llinear regression has two kinds of variables in a contingency table, i.e. the explanatory variable and the response variable. Explanatory variables are also known as independent variables or predictors, while the response variables are called dependent variables. For example, we want to see how different factors affect the overall body health score of an individual. We look at three variables:

amount of exercise per week, smoking or not smoking, and age group. These are the explanatory variables. The response variable is then the body health score.

The GLM is defined in terms of a set of independent random variables $y_{1,...,}y_n$, where all independent random variables y_i have the same distribution, which belongs to the exponential family. GLM was first defined by Nelder and Wedderburn in 1972 (Dobson, 2001; McCullagh and Nelder, 1989).

Dobson (2001) suggests that we consider a smaller set of parameters $\beta_{1,...,}\beta_p$ (p<N) in GLM, because the parameter θ is usually not of direct interest. We allow the relationship between the response and explanatory variables of a GLM to be a function of the linear combination of explanatory variables, that is, for $i = 1,...,n$, we have Equation 2.3.

$$g(\mu_i) = x_i^T \beta = \eta \tag{2.3}$$

where:

g is the link function,

x_i is the explanatory variable, such as:

$$x = \begin{bmatrix} x_1^T \\ . \\ . \\ . \\ x_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & . & . & . & x_{1p} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ x_{N1} & . & . & . & x_{Np} \end{bmatrix}$$

and β is the parameters, such as:

$$\beta = \begin{bmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix}$$

The function $g(\cdot)$ is assumed to be is monotone and differentiable. This function is known as the link function. The distribution of the response variables depends on the form of the link function.

Essentially a GLM consists of three elements (Dobson, 2001):

● y_i, which has a distribution function f, which is a member of the exponential family; thus, $f(y_i;\theta_i) = \exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)]$ (note: each y_i has the canonical form);

● a linear predictor $\eta = x_i^T \beta$, where x_i is the explanatory variables and β is the parameters; and,

● a link function *g* such that $g(\mu_i) = x_i^T \beta$ where $\mu_i = E(y_i)$.

The GLM includes linear regression models, analysis of variance models, logit models, probit models, log-linear models and multinomial response models. A table of the link functions and their inverses, which is used for several exponential families of distributions for errors, is presented in Table 2.1 (McCullagh and Nelder, 1989). In this project, we focus on categorical data in the log-linear model, which will be discussed in Section 2.2.

Table 2.1 The link functions of common distributions.

Distribution	*Name*	*Link Function*	*Link inverse*
Normal	Identity	$\eta = \mu$	$\mu = \eta$
Exponential	**Inverse**	$\eta = \mu^{-1}$	$\mu = \eta^{-1}$
Gamma			
Inverse Gaussian	Inverse squared	$\eta = \mu^{-2}$	$\mu = \eta^{-1/2}$
Poisson	**Log**	$\eta = \ln(\mu)$	$\mu = \exp(\eta)$
Binomial	**Logit**	$\eta = \ln(\frac{\mu}{1-\mu})$	$\mu = \frac{\exp(\eta)}{1+\exp(\eta)} = \frac{1}{1+\exp(-\eta)}$
Multinomial			

2.1.3. Estimation

Two steps are used for model estimation:

(1) Define a measure of goodness of fit between the data and a corresponding set of fitted values generated by the model.

(2) Choose the parameter estimates as those that minimize the chosen goodness of fit criterion. We use the maximum likelihood method to obtain the estimates of the parameters given the data (McCullagh and Nelder, 1989).

To get the maximum likelihood estimate, we will need the likelihood function, which is the joint density. We then take the log of the likelihood function, called the log-likelihood function. The log-likelihood functions are the basis for deriving

estimators for parameters given an observed sample data. The shape of the log-likelihood function has the maximum point at the value p, which is defined as the Maximum Likelihood Estimate (MLE), which we will denote using the symbol $\hat{p}$.

The log-likelihood function for the exponential family of distribution is obtained by taking logs on both sides of Equation 2.2, which results in $l(\theta;y) = a(y)b(\theta) + c(\theta) + d(y)$. For example, consider the Poisson distribution $(x_1, x_2, ...x_n; \lambda)$. The Poisson mass function is:

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \tag{2.4}$$

The joint distribution is:

$$f(x_1,...,x_n;\lambda) = \prod_{i=1}^{n} \frac{\lambda^x e^{-\lambda}}{x_i!} = e^{-n\lambda} \frac{\lambda^{\sum_1^n x_i}}{\prod_{i=1}^{n} x_i} = L(\lambda) \tag{2.5}$$

where $L(\lambda)$ is the likelihood function.

We take logs of both sides to obtain the log-likelihood function:

$$\ln L(\lambda) = -n\lambda + \sum x_i \ln(\lambda) - \ln(\prod_{i=1}^{n} x_i) \tag{2.6}$$

The MLE of parameter the λ is obtained by maximizing λ. To find the maximum value of λ, we take the first derivation of the log-likelihood function and equate it to zero, i.e.:

$$\frac{d\ln(\lambda)}{dp} = -n + \sum_1^n X_i \frac{1}{\lambda} = 0 \tag{2.7}$$

The maximum likelihood estimator, $\hat{\lambda}$, is given by:

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} x_i}{n}$$

2.2. The log-linear model

2.2.1. Introduction to the log-linear model

The log-linear model is a linear model which describes the variations in the expectation of the logarithmic form of the response variable. We will focus on the analysis of categorical data in contingency tables by using log-linear models.

We know that with the joint Poisson distribution, the expected frequencies of the Poisson distribution are $E(Y_i) = \lambda_i$. When we have high-dimensional tables, the expected cell frequencies are given by the marginal probabilities multiplied by the fixed marginal total frequencies hence complicating the calculations. Thus we use the logarithm of GLM, which is a natural link function between $E(Y_i)$ and a linear combination of parameters, i.e.:

$$\eta_i = \log E(Y_i) = x_i^T \beta \qquad i = 1,..., N$$

This is the log-linear model (Dobson, 2001).

Conditions are needed to use the log-linear model. The response variable is a continuous variable which cannot be separated into a discrete contingency table (Jeansonne, 2002). If both the explanatory and response variables are categorical data, the log-linear model should be used. We focus on categorical data and log-linear models in this study, and any continuous data will be grouped into categorical data.

2.2.2. Contingency tables

2.2.2.1. Two-dimensional tables

A contingency table is a table of counts that usually shows relationship between two or more variables. In addition, most variables are categorical variables (Everitt, 1992).

First, we introduce two-dimensional tables in which a sample of *N* observations are classified into two categorical variables, one of which has *r* categories and the other one having *c* categories. If *r* and *c* both are equal to 2, we have the simplest form of contingency table, known as the 2 x 2 contingency tables. Otherwise, if either *r* or *c* is greater than 2, we then have an *r* x *c* contingency table (Everitt, 1992).

For example, in an investigation of the relationship between education and salary of sample size *n* = 50, the variable "education" could have two categories: no qualification and bachelor graduate; the variable "salary" has two categories: over or below $40,000 per annum. A table such as Table 2.2 is known as a 2 x 2 contingency table. A chi-square test can be used for a 2 x 2 contingency table. The null hypothesis states that knowing the level of the variable "salary" does not help you predict the level of the variable "education", i.e. the variables are independent. The alternative hypothesis is that knowing the level of the variable "salary" can help you predict the level of the variable "education".

The chi-square formula is given in chapter 1. Applying the chi-square formula to the data in Table 2.2 gives: $x^2 = \frac{150 \times (45 \times 55 - 25 \times 25)^2}{70 \times 80 \times 70 \times 80} = 16$ with 1 degree of freedom. We then use the chi-square table to find $p(x_1^2 > 16) = 0.0001$. Since the p-value (0.0001) is less than the significance level (0.05), we can reject the null

hypothesis. Thus, we conclude that there is a relationship between salary and education.

Table 2.2 2x2 contingency table.

Education	***Salary***		
	Below $40,000 p.a.	***Over $40,000p.a.***	***Total***
No qualification	45	25	70
Bachelor graduate	25	55	80
Total	70	80	150

Table 2.3 shows an *r* x *c* contingency table with *r* = *c* = 3. The variable "salary" is classified to three categories: “below $40,000”, “$40,000−$60,000”, and “over $60,000”. The variable "education" is classified to three categories: “no qualification”, ” bachelor graduate” and “post-graduate”.

Table 2.3 3x3 contingency table.

Education	***Salary***			
	Below $40,000 p.a.	***$40,000−$60,000 p.a.***	***Over $60,000 p.a.***	***Total***
No qualification	45	25	6	76
Bachelor graduate	25	55	16	96
Post graduate	5	30	55	90
Total	75	110	77	262

We can use the chi-square test for an *r* x *c* contingency table to indicate the existence of a relationship between two variables. The null hypothesis, H_0 , is that salary and education are independent. The alternative hypothesis, H_a, is

that salary and education are not independent. Firstly, we calculate the expected cell value of Table 2.3, according to the chi-square formula. Table 2.4 shows the expected frequency for the data of Table 2.3. Then, using the same method as we did for the 2 x 2 contingency table and by applying the chi-square formula, we obtain the chi-square value:

$x^2 = \frac{(45-21.76)^2}{21.76} + \frac{(25-31.91)^2}{31.91} + ... + \frac{(55-26.45)^2}{26.45} = 98.28$. Since the p-value (<0.01) is less than the significance level (0.05), we can reject the null hypothesis. Thus, we conclude that a relationship exists between salary and education (Everitt, 1992, p39). As the categorical order increases, the calculation becomes more complicated. Therefore we would like to apply the log-linear model to simplify calculations.

Table 2.4 Expected frequencies for the data of Table 2.3

Education	***Salary***			
	Below $40,000 p.a.	***$40,000−$60,000 p.a.***	***Over $60,000 p.a.***	***Total***
No qualification	21.76	31.91	22.34	76
Bachelor graduate	27.48	40.31	28.21	96
Post graduate	25.76	37.79	26.45	90
Total	75	110	77	262

The Christensen (1990) introduces the balanced analysis of variance (ANOVA) model:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

We can rewrite this expression as:

$$y_{ijk} = a + a_{1(i)} + a_{2(j)} + a_{12(ij)} + e_{ijk}.$$

Where:

i=1,…,I;

j=1,…J, and

k=1,…,K.

We assume that e_{ijk} is independent of $N(0,\delta^2)$.

We use the ANOVA model to test the relationship between the two factors in the log-linear model by testing whether their interaction term is zero. When the interaction term is not equal to zero, we write the two-dimension table as the full model $m_{ij} = a + a_{1(i)} + a_{2(j)} + a_{12(ij)}$ (2.8).

If the interaction term is equal to zero, then we have the sub-model of this full model:

$$m_{ij} = a + a_{1(i)} + a_{2(j)} \tag{2.9}$$

For instance, let us examine a contingency table that has the properties

$E(n_{ij}) = m_{ij}$ and $m_{ij} = n_{..}p_{ij}$. We are interested in the structure of m_{ij} to test the interaction term and whether it is equal to zero in the log-linear model.

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} + u_{12(ij)} \tag{2.10}$$

$$\log(m_{ij}) = u + u_{1(i)} + u_{2(j)} \tag{2.11}$$

We know $m_{ij} = n_{..}p_{ij}$ if the rows and columns of the table are independent, and $m_{ij} = n_{..}p_{ij} = n_{..}p_{i.}p_{j.}$ and (2.12) hold.

$$\log(m_{ij}) = \log(n_{..}p_{i.}p_{j.}) = \log(n_{..}) + \log(p_{i.}) + \log(p_{j.}) \tag{2.12}$$

The model shown in (2.12) is same as the model in (2.11). Therefore, the model in (2.12) holds and, the main effects are independent. Otherwise, the rows and

columns of the table have the interaction term, $m_{ij} = n_{..}p_{ij} \neq n_{..}p_{i.}p_{j.}$. We write m_{ij} in a log-linear model as shown in (2.10).

The Christensen (1990, p48) discussed several reasons for writing ANOVA type models for $\log(m_{ij})$ instead of for m_{ij}. "The first reason is that the large sample theory can be worked out. The second reason is that log-linear models arise in a natural fashion from the mathematics of Poisson sampling".

2.2.2.2. Three-dimensional and higher dimensional tables

Previously, we discussed two categorical variables in a contingency table. Now we will analysis three or more categorical variables in a contingency table. For example, Table 2.5 shows three variables being investigated. The first variable is "salary", the second variable is "education", and the third variable is "age", say “20−30”, or ”30−50”. The first two variables are as described for Table 2.2.

Table 2.5 Three-dimensional contingency table of salary level.

		Salary				
		Below $40,000 p.a.		***$40,000−$60,000 p.a.***		***Totals***
	Age	***20−30***	***30−50***	***20−30***	***30−50***	
Education	***No qualification***	35	25	15	30	105
	Bachelor graduate	20	5	25	45	95
	Totals	55	30	40	75	200

We can apply the chi-square test to check if variables are independent or not. However, the calculations are more complicated for multi-dimensional

contingency tables. Therefore we apply the log-linear model to test relationship between the factors.

Let p_{ij} represent the probability in the i^{th} row variable, and the j^{th} column variable.

We have seen previously the $E(n_{ij}) = m_{ij}$ where n_{ij} is the observed cell value in the i^{th} row and j^{th} column, and m_{ij} is the expected counts in the table. If the null hypothesis indicates mutual independence of the variables, the expected values may be obtained using the formula: $m_{ij} = n_{..} p_{ij}$ where $n_{..}$ is the sample size (Christensen, 1990).

We propose the null hypothesis, H_0, that salary, education and age are independent, and the corresponding alternative hypothesis may now be formulated as follows:

$$H0 : m_{ijk} = n_{i..} p_{i..} p_{.j.} p_{..k}$$

$$Ha : \text{correlation between variables.} \qquad (2.13)$$

For the counts of Table 2.5, we take the log of the equation (2.13), giving:

$$\log(m_{ijk}) = \log(n_{i..}) + \log(p_{i..}) + \log(p_{j.}) + \log(p_{..k}) \qquad (2.14)$$

The sub-models of model (2.14) are:

$$\log(m_{ijk}) = \log(n_{i..}) + \log(p_{i..}) + \log(p_{j.}) \qquad (2.15)$$

$$\log(m_{ijk}) = \log(n_{i..}) + \log(p_{i..}) \qquad (2.16)$$

We use the models (2.14), (2.15) and (2.16) to test whether the variables independent or not.

2.2.3. Model selection in R

We now want to test the relationship of the factors given a log-linear model. Christensen (1990) introduced a log-linear model for a three-dimensional table with the full model given as:

$$\log(m_{ijk}) = a + a_1(i) + a_2(j) + a_3(k) + a_{12}(ij) + a_{13}(ik) + a_{23}(jk) + a_{123}(ijk).$$

The corresponding ten sub-models are as shown below:

(i) $\log(m_{ijk}) = a + a_1(i)$

(ii) $\log(m_{ijk}) = a + a_1(i) + a_2(j)$

(iii) $\log(m_{ijk}) = a + a_1(i) + a_2(j) + a_3(k)$

(iv) $\log(m_{ijk}) = a + a_1(i) + a_2(j) + a_3(k) + a_{12}(ij)$

(v) $\log(m_{ijk}) = a + a_1(i) + a_2(j) + a_3(k) + a_{13}(ik)$

(vi) $\log(m_{ijk}) = a + a_1(i) + a_2(j) + a_3(k) + a_{23}(jk)$

(vii) $\log(m_{ijk}) = a + a_1(i) + a_2(j) + a_3(k) + a_{12}(ij) + a_{13}(ik)$

(viii) $\log(m_{ijk}) = a + a_1(i) + a_2(j) + a_3(k) + a_{13}(ik) + a_{23}(jk)$

(ix) $\log(m_{ijk}) = a + a_1(i) + a_2(j) + a_3(k) + a_{12}(ij) + a_{23}(jk)$

(x) $\log(m_{ijk}) = a + a_1(i) + a_2(j) + a_3(k) + a_{12}(ij) + a_{13}(ik) + a_{23}(jk)$

We compare the models (i) and (ii) to determined whether the variable a_2 is zero, according to the generalized ANOVA model. If the factor a_2 is not zero, we should prefer model (ii) over model (i). Following the same idea, we compare model (ii) and model (iii) to determine whether the factor a_3 is zero. If the factor a_3 is not zero, we should prefer the model (iii) over model (ii). It is also means that the three variables are not independent. We then go to next step to test

whether the interaction terms are zero. An interaction term is basically the product of two predictor variables of interest, such as a_{12} in our example.

We now compare models (iii) and (iv) to determine whether the a_{12} effect is zero, according to general ANOVA model. If the effect of a_{12} effect is not zero, we should prefer model (iv) over model (iii). Following the same idea, we compare models (iii) and model (v) to determine whether the effect of a_{13} is zero. Model (viii) holds if only if the effect of a_{123} can be dropped from the full model.

Three model selection methods are commonly used: the forward selection method, the backward selection method and the stepwise selection method (McCullagh and Nelder, 1989).

● Forward selection:

1. Compare models (iii) and (iv) to determined whether the effect of a_{12} is zero, according to the generalized ANOVA model. If the effect of a_{12} is not zero, we should prefer the model (iii) over model (iv). By the same idea, we also can start by comparing model (iii) comparison with model (v) to determine whether the effect of a_{13} is zero, or we could compare models (iii) and (vi) to determine whether the effect of a_{23} is zero.

2. We assume the effect of a_{12} is not zero. Now, we must compare models (iii) and model (v) to determine whether the effect of a_{13} is zero. If the effect of a_{13} is not zero, we must compare the models (vii) and (x) to determine whether the effect of a_{23} is zero.

● Backward selection:

1. Start from the full model (x) and compare it with models (vii) (viii) and (ix) to determine whether the effects of a_{23}, a_{12}, and/or a_{13} are equal to zero. If any effect is equal to zero, the full model will be reduced.

● Stepwise selection:

The stepwise procedure makes the decisions using the Akaike Information Criteria (AIC) in R. The AIC value is a measure of the goodness of fit of an estimated statistical model, the model have the smallest AIC value being the best. R provides a command "step" to choose the model with the best fit.

2.3. Summary

This chapter discusses the GLM to model non-normal data (Dobson, 2001), and the basic concepts of GLM. The contingency tables can be modelled using a GLM technique called the log-linear model. We present a method of testing relationships between each main effect. Two methods can be used to test a relationship between the effects: the first one uses the chi-square test ; the second one uses the log-linear model. When the categorical order increases, we would like to apply the log-linear model to simplify calculations.

Model selection is another main objective in this chapter. This chapter presents three model selection methods: the forward selection method, the backward selection method and the stepwise method. These three standard model selection techniques can be used to choose the best model and to identify the relationship among explanatory variables.

Chapter 3

Effect of random rounding and chi-square analysis

In this chapter, we analyze the difference between a rounded contingency table and the possible real tables it could be derived from. The rounded table and its possible real tables have been tested by the chi-square test.

3.1. Introduction

In the previous chapters, we introduced the idea that statistical organizations normally use random rounding to base 3 and 5 when publishing census data. This is to ensure some confidentiality for the data with random rounding. A user of a published table does not have the exact original cell counts. For example, if the published cell count is 6 then the user can only infer that the possible unrounded cell is one of 4, 5, 6, 7, or 8.

In this chapter, we illustrate the effect of random rounding on the analysis of a contingency table using the chi-square test. We start with a randomly rounded table and compared the chi-square tests of all possible unrounded tables (parent tables). Each rounded table has, N possible parent tables. We calculate the chi-square of the given rounded table, named as X_p^2 . We then calculate the individual parent tables' chi-squares, designated $X_n^2(n=1,...,N)$. Then we compare the randomly rounded and parent tables' chi-square values, and identify how many decimal places would need to be reported to be able to reconstruct the exact original table from the rounded table. This value is the threshold of maintaining confidentiality.

We report the rounded table's X_p^2 to the n^{th} decimal place, and report all parent tables' chi-squares to n^{th} decimal place. If at least two parent tables' chi-squares are equal to the rounded table's chi-square to the n^{th} decimal place (except for the parent table which is exactly same as the rounded table), the user has a high probability of inferring the original table. For example, if the rounded table's chi-square matches two parent tables' chi-square to four decimal places, then the user has a 50% chance of identifying the original table. If the rounded table's chi-square matches only one chi-square from the all possible parent tables (except for the parent table which is exactly same as the rounded table) at $(n+1)^{th}$ decimal places, the rounded table's X_p^2 cannot preserve confidentiality at the $(n+1)^{th}$ decimal place. We will call the $(n+1)^{th}$ decimal place DP. If the rounded table's published chi-square is reported with DP decimal places then a user will be able to reconstruct the original table.

In this chapter, we analyze the relationship between the DP and the variance of a rounded table when rounded using base 3 and base 5. We use a range of rounded tables and summarize the results in the Section 3.2.2.

3.2. Two-dimensional tables

3.2.1. r x c contingency tables

We would like to start with a 2X2 example. The two examples below illustrate a rounded table that is a two-way contingency table with 3 and 5 as the rounding bases. Each table rounded to base 3 has 625 possible original; each table rounded to base 5 has 6561 unrounded tables. However, the possible real unrounded tables are not all the tables that are possible, for example, Table 3.1 shows cell(1,1) is 9, the cell(2,1) is 12, and the marginal total cell(3,1) is 21. If

unrounded cells are 7 of cell(1,1), and 10 of cell(2,1), So the marginal total of cell(3,1) is 17 (10+7), It is outside the range of published marginal cell (21).This is because we should delete the possible tables which have marginal totals that are outside the range of the published rounded table's marginal totals. These examples show how to find DP and to delete unacceptable unrounded tables.

Table 3.1 is a 2X2 contingency table rounded to base 3 that has two variables grade and type. The variable grade has two categories, pass and fail. The variable type has been classified as type I or type II.

Table3.1. Exam grade

		Grade		
		Pass	***Fail***	***Total***
Type	***Type I***	9	6	15
	Type II	12	9	21
	Total	21	15	36

Firstly, we calculate the expected frequencies for each cell, using Formula (1.2). These are shown in Table 3.2.

Table 3.2. Expected frequencies for Table 3.1.

		Grade		
		Pass	***Fail***	***Total***
Type	***Type I***	8.75	6.25	15
	Type II	12.25	8.75	21
	Total	21	15	36

In Table 3.2 the first cell, named E_{11}, is obtained by:

$$E_{11} = \frac{(15 \times 21)}{36} = 8.75$$

The other cells follow a similar calculation.

Using Formula (1.1) we obtain the chi-square value:

$$\begin{aligned} X_p^2 &= \frac{(9-8.75)^2}{8.75} + \frac{(6-6.25)^2}{6.25} + ... + \frac{(9-8.75)^2}{8.75} \\ &= 0.0071 + 0.01 + ... + 0.0071 \\ &= 0.0294 \end{aligned}$$

Currently, we have the rounded parent table's chi-square (X_p^2). We will present two steps for obtaining the DP. First, we obtain the all possible real unrounded tables. In totals, 625 possible unrounded tables can be obtained from a two-way contingency table rounded using base 3. We then delete the unacceptable unrounded tables. After deletion we have 247 possible tables left.

The second step is to find the DP. We write the program in R to compute the chi-square for 247 unrounded tables $X_n^2 (n = 1,...,247)$, and then obtain the DP.

$$\text{2dp: = 0.03} \quad \begin{cases} table \longrightarrow No.103 \\ table \longrightarrow No.124\,(ignore) \end{cases}$$

$$\text{3dp:= 0.029} \quad \{table \longrightarrow No.124(ignore)$$

Using the example above, the rounded table has two unrounded tables that match its chi-square to two decimal places. Note that we always ignore one unrounded table's chi-square, because one possible unrounded table (No.124) which is the same as the rounded table will always exist. The user reconstruct the original table is $\begin{bmatrix} 8 & 6 \\ 12 & 8 \end{bmatrix}$(No.103) and $X^2 = 0.03$ in two decimal place. Therefore, the user can easily reconstruct the original table when the rounded

table's published chi-square is reported to 2 decimal places (i.e. the DP = 2). The rounded table is not safe to publish with the chi-square give to 2 decimal places.

We do a similar analysis for a two-way contingency table rounded using base 5. Firstly, we must compute the given table's chi-square, and then obtain all the possible unrounded tables. We obtain 587 out of 6561 possible tables. In the third step, we compute the chi-square for all the real unrounded tables and obtain the DP.

3.2.2. Data analysis and results

We randomly generated 100 two-way tables randomly rounded using base 3 as our possible rounded tables (see Section 1.3.2). We compute the DP of the 100 rounded tables and the variance, average and CV (variance/average) of each rounded table.

Figure 3.1 present three plots, showing average vs DP, variance vs DP, and CV vs DP for all tables. We can not see any trend between DP and average. The second plot shows a trend between DP and variance, suggesting that as variance increases, the DP becomes smaller. That means that as variance increases, the table becomes less confidential. The last plot shows a similar pattern to the second plot. Figure 3.2 clearly presents the relationship between DP and variance.

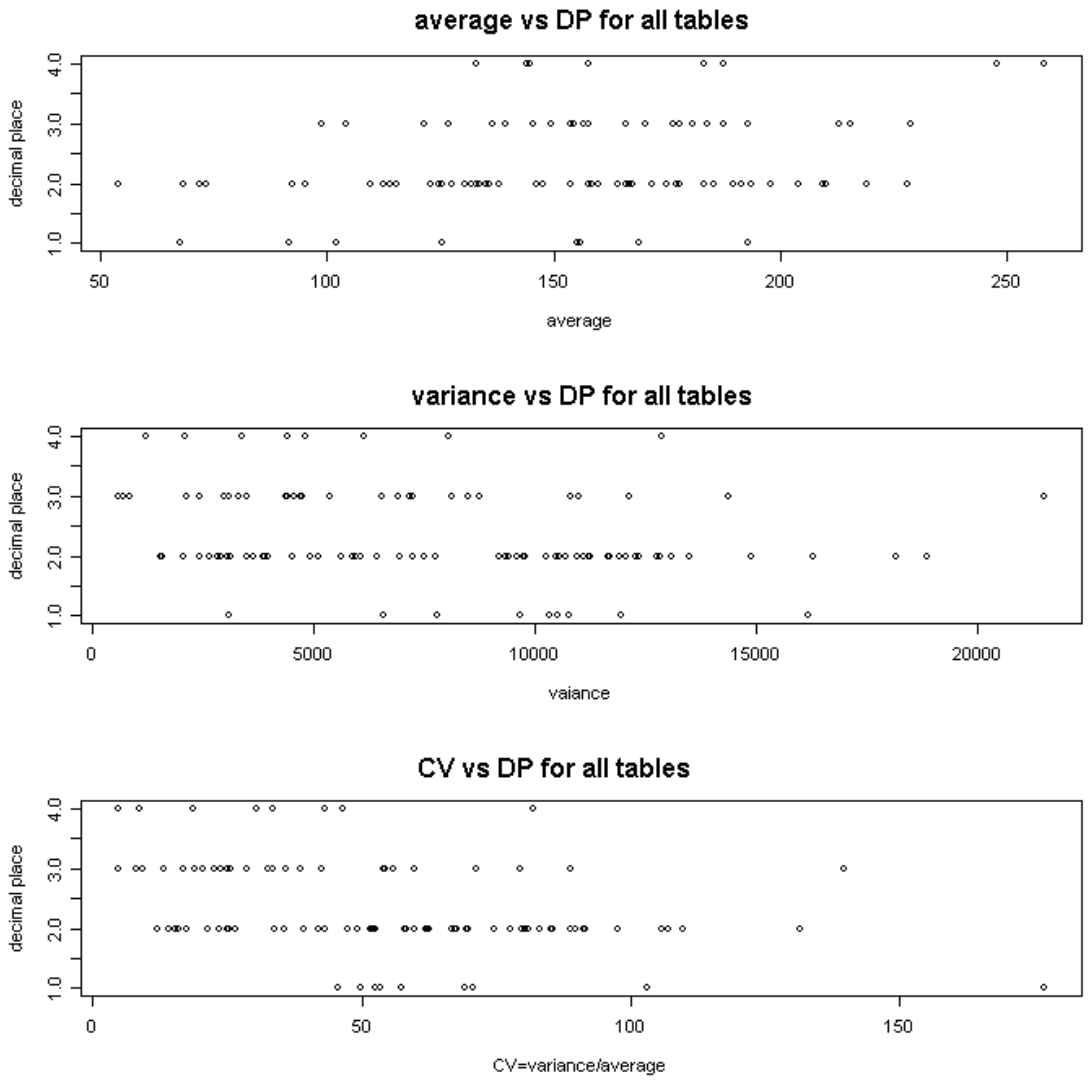

Figure 3.1. Three plots for two-way tables rounding using base 3 (RR3)

We randomly generated 400 two-way tables rounded using base 3 as our "published" tables (rounded tables). We want to see the relationship between the average or variance of table and decimal place of chi-squire. These were divided into four categories: high average with high variance (hahv), high average with low variance (halv), low average with high variance (lahv), and low average with low variance (lalv). We calculated the percentage of tables match in the rounded table to the n^{th} decimal place. For example, when the chi-square value is given

to 1dp, we find 245 unrounded tables for which the chi-square value is the same as that of the rounded table. The percentage in this case is 0.99 (245/247). When the chi-square value is given to 2dp, we get 150 unrounded tables where the chi-square value is the same as that of rounded table; here the percentage is 0.61 (150/247), etc. A bigger percentage means the published table is more confidential, because more possible unrounded tables have been found, and therefore, it is harder to identify which is the actual original table.

Figure 3.2 presents the percentage of the four groups of tables hahv, halv, lahv, and lalv. The high average plot has a higher percentage of matches than the low average plot based on same variance level (low or high). The low variance plot has a higher percentage of matches than the high variance plot based on same average level (low or high). The low variance plot has a higher percentage of matches than the high variance plot. The high percentage means it is harder to reconstruct the original table. Therefore, the rounded table with low variance is safer and more confidential than a rounded table with higher variance. The plot of the table with high average and low variance (halv) has the highest percentage of matches of the four categories, which suggests that a rounded table with low variance and high average gives more confidentiality than others.

The four plots in Figure 3.2 present the decreasing trend between percentage and chi-square value given to the n^{th} decimal place. More matched tables are found when the chi-square value is given to 1dp. Therefore, a rounded table where the chi-square is given to fewer decimal places will be more confidentiality.

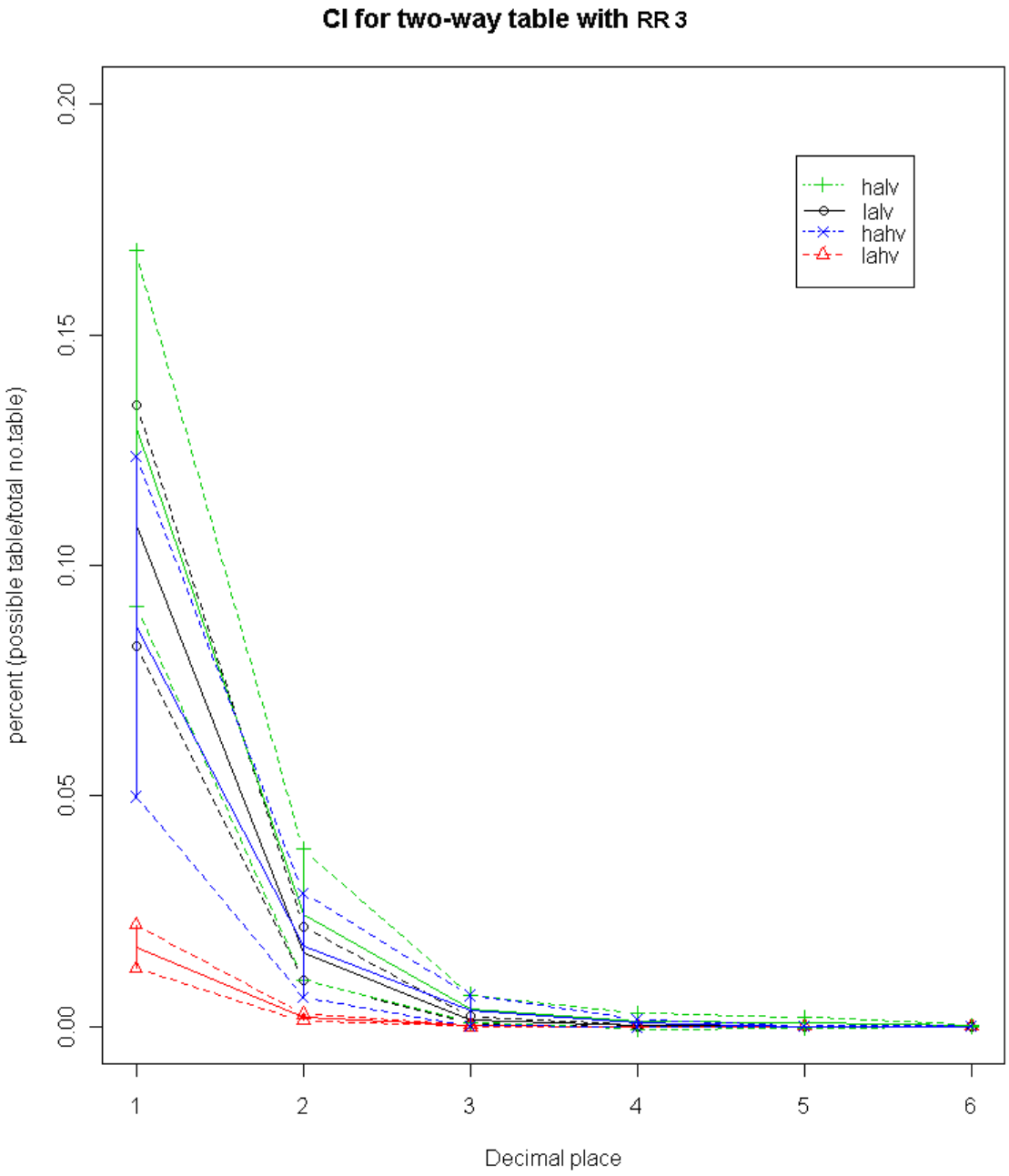

Figure 3.2. Percentage of matches vs. decimal place for four types of two-way contingency table randomly rounded using base 3

We repeat the study design for a two-way table rounded using base 5. Figure 3.3 shows the trends with the rounded table's variance and average are the same as

when base 3 is used. Next we compare the DP and variance plot between rounded tables using base 3 and 5.

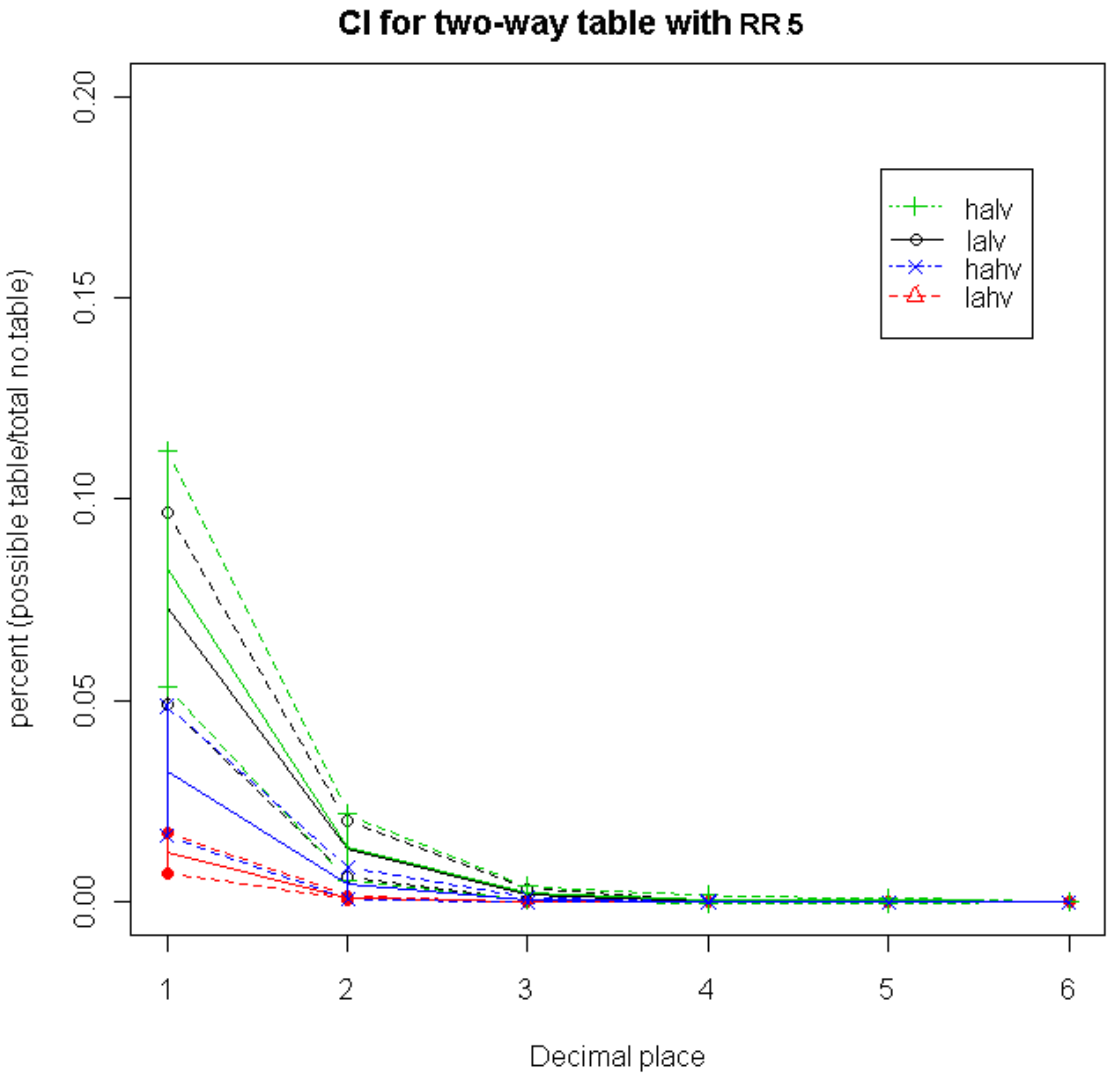

Figure 3.3 Percentage of matches vs. decimal place for four types of two-way contingency table randomly rounded using base 5

Figure 3.4 plots the variance and DP for tables randomly rounded using base 3 and 5 alongside each other, using different plot symbols to represent the two types of table. This graph reveals a decreasing trend. When the rounded table's published chi-square is given to DP decimal places, a user will be able to identify

the original table. The plot with base 5 appears to have a higher DP than the plot of base 3. When the DP is equal to 2, the plot points from the base 3 dataset appear more than the plot points from the base 5 dataset. When a table rounded using base 3 is published with the chi-square given to 2 decimal places, a user will more easily reconstruct the original table than she/he would with the table rounded to base 5 with the chi-square given to 2 decimal places.

As DP increases, we can see more plot points from the base 5 dataset appearing rather than the plot points of the base 3 dataset. Therefore, a table rounded using base 5 is more confidential than a table rounded using base 3, because in base 5 more decimal places need to be reported to be able to reconstruct the original table.

A smaller DP means that fewer decimal places need to be reported to be able to reconstruct the original table. Therefore, the user will easily be able to identify the original table. Decreasing the DP will reduce the confidentiality. This graph also presents the decreasing trend between the DP and the variance of a table. When variance increases, the DP decreases. Therefore, a table with low variance encourages confidentiality compared to a high variance table because more decimal places need to be reported to be able to reconstruct the original table.

Therefore, rounding a table using base 5 increases confidentiality compared with rounding using base 3 based on the chi-square having the same number of decimal places. A rounded table with low variance is safer and more confidential than a rounded table with high variance.

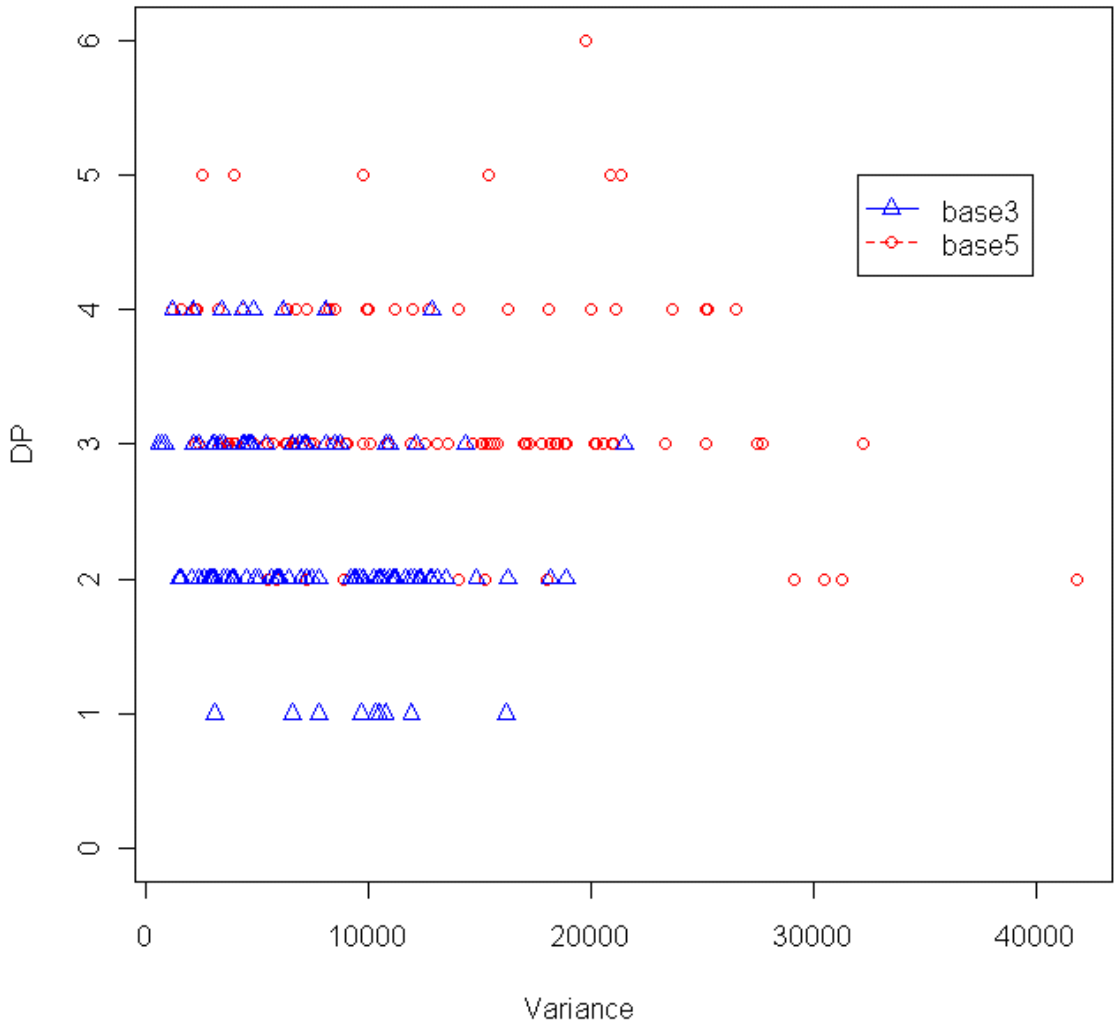

Figure 3.4. The variance and DP at base 3 and base 5

3.3. Multi-dimensional tables

3.3.1. r x c x n contingency tables

We now extend our analysis to a $2\times2\times2$ contingency table. We have used three-way $2\times2\times2$ contingency tables randomly rounded using base 3 and 5 as our "published" tables. We follow three steps similar to those used for two-way

tables. Firstly, we calculate the X_p^2 for the published tables. Secondly, all possible unrounded tables (parent tables) are constructed. A $2\times2\times2$ contingency table rounded using base 3 could be derived from 59691 possible parent tables, while a $2\times2\times2$ contingency table randomly rounded using base 5 could be derived form 274393 possible parent tables. Thirdly, we calculate the chi-square for all possible parent tables, and obtain DP.

Table 3.3. is known as a $2\times2\times2$ contingency table which has been randomly rounded using base 3, and which has three variables: "sex" , "grade" and "type" . The variable "grade" has two categories, "pass" and "fail". The variable "type" also has two categories: "type I" and "type II". "Sex" is, of course, divided into "male" and "female".

Table 3.3. A 2 x 2 x 2 contingency table correlating exam grade, sex and type

		Grade				
		Pass		***Fail***		***Total***
Sex		***Male***	***Female***	***Male***	***Female***	
Type	***Type I***	3	6	9	9	27
	Type II	9	12	3	6	30
	Total	12	18	12	15	57

Firstly, we calculate the expected frequencies for each cell. Here, we need to introduce a formula to calculate the expected value. The Everit (1992) introduced an expectancy formula as follows:

$$E_{ijk} = \frac{n_{i..} n_{.j.} n_{..k}}{N^2} \qquad (3.1)$$

where $n_{i..} = \sum_{j=1}^{c}\sum_{k=1}^{l} n_{ijk}$

$$n_{.j.} = \sum_{i=1}^{r}\sum_{k=1}^{l} n_{ijk}$$

$$n_{..k} = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ijk}$$

$$N = \sum_{i-1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l} n_{ijk} \qquad \text{for } i = 1,2,...,r,\ \ j = 1,2,...,c,\ \ \text{and } k = 1,2,...,l$$

For example, for the data shown in Table 3.4 we have:

$$n_{1..} = (3+6)+(9+9) = 27$$
$$n_{2..} = (9+12)+(3+6) = 30$$
$$n_{.1.} = (3+9)+(9+3) = 24$$
$$n_{.2.} = (6+12)+(9+6) = 33$$
$$n_{..1} = (3+9)+(6+12) = 30$$
$$n_{..2} = (9+3)+(9+6) = 27$$
$$N = 3+6+9+9+9+12+3+6 = 57$$

In Table 3.4 the first cell is named E111. We calculate the expected values using formula (3.1) and get the following result:

$$E_{111} = \frac{27\times 24\times 30}{57\times 57} = 5.9834$$

Table 3.4 shows the full set of expected frequencies.

Table 3.4. Expected frequencies for Table 3.3.

		Grade				
		Pass		Fail		Total
Sex		Male	Female	Male	Female	
Type	Type I	5.9834	8.2271	5.3850	7.4044	27
	Type II	6.6482	9.1413	5.9834	8.2271	30
	Total	12	18	12	15	57

We compute the chi-square (Everitt, 1992):

$$x^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{l}\frac{n_{ijk} - E_{ijk}}{E_{ijk}} \tag{3.2}$$

We compute the given example's chi-square value using formula (3.2) as follows:

$$x^2 = \frac{(3-5.983)^2}{5.983} + \frac{(9-6.6482)}{6.6482} + \ldots + \frac{(6-8.2271)^2}{8.2271} = 8.6774$$

Now we have the rounded table's $X_p^2(8.6774)$. We write the program in R to compute the chi-square for all possible parent tables. Then we can obtain DP.

$$\text{3dp: } = 8.677 \quad \begin{cases} table \longrightarrow No.11356 \\ table \longrightarrow No.29846 \ (ignore) \end{cases}$$

$$\text{4dp:= } 8.6774 \quad \{table \longrightarrow No.29846 \ (ignore)$$

In this example, the rounded table has two parent tables that match its chi-square to 3 decimal places. Note that we always ignore one unrounded table's chi-square, because there is always one possible unrounded table (No.29846) which is the same as the rounded table. Therefore, the user can easily reconstruct the original table when rounded tables published chi-square is reported to 3 decimal places, i.e. DP is 3. The rounded table can be safely published with the chi-square value given to less than 3 decimal places.

3.2.2. Data analysis and results

We repeat the study design with a three-way table. Figure 3.5 shows the trends with a three-way table rounded using base 3; variance and average are the same as for the two-way table. In Figure 3.5, the halv plot is very close to the lalv plot. This shows that the average is not a main reason for the percentage changing when tables have a low level of variance. The hahv plot clearly has a higher percentage of matches than the lahv plot, which shows that a high average leads to a percentage increase when a table has a high level of variance. The hahv and lahv plots in Figure 3.5 are well below the other plots, because the three-way table's sample size (three-way tables have more possible tables) is bigger than a two-way table's. The CI plot for three-way tables will be narrower than the CI plot for two-way tables. These four plots follow the same order as two-way tables: halv, lalv, hahv, and lahv (highest to lowest).

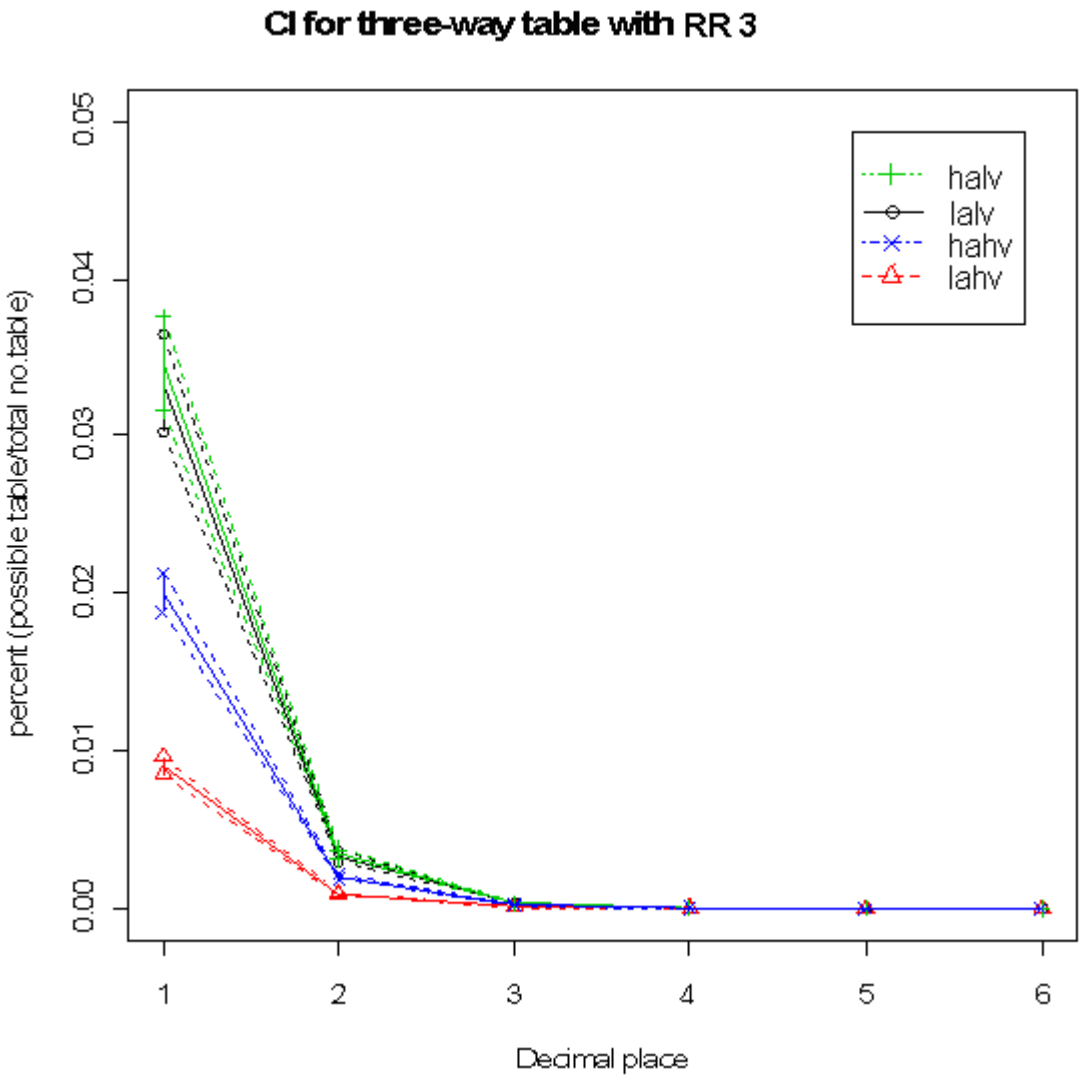

Figure 3.5 The variance and decimal place for a three-way table randomly rounded using base 3

Each three-way table randomly rounded using base 5 can produce a number of different possible actual tables. We can not produce all possible actual tables to analyse, because there are too many. We therefore produce a subset of all possible actual tables. Figure 3.6 presents the trends with the three-way table with rounding base 5 variance and average are the same as for the two-way table. These four plots follow the same order as two-way tables and three-way

tables with rounded using base 3, i.e. halv, lalv, hahv, and lahv (highest to lowest).

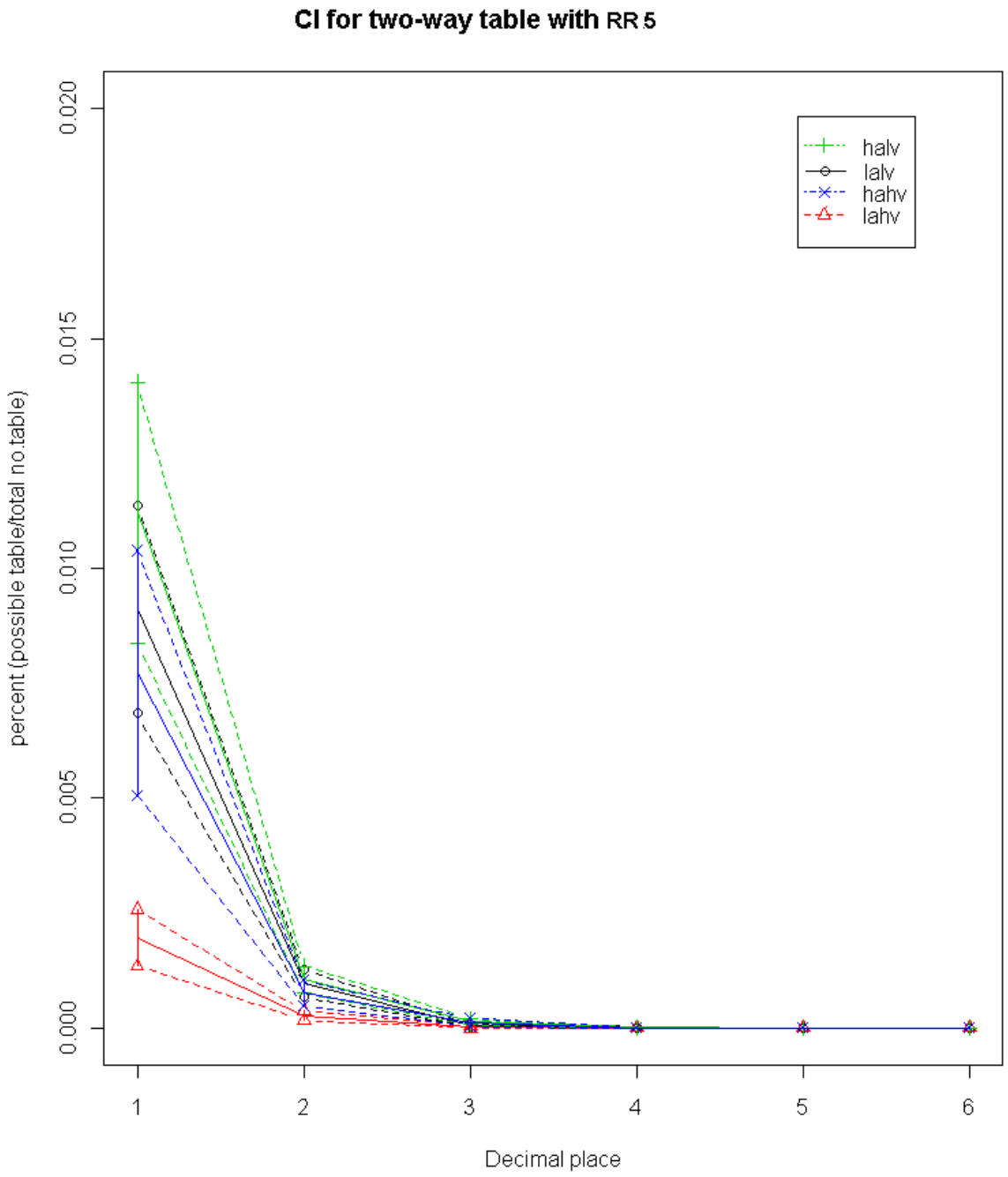

Figure 3.6 The variance and decimal place in three-way table with RR5

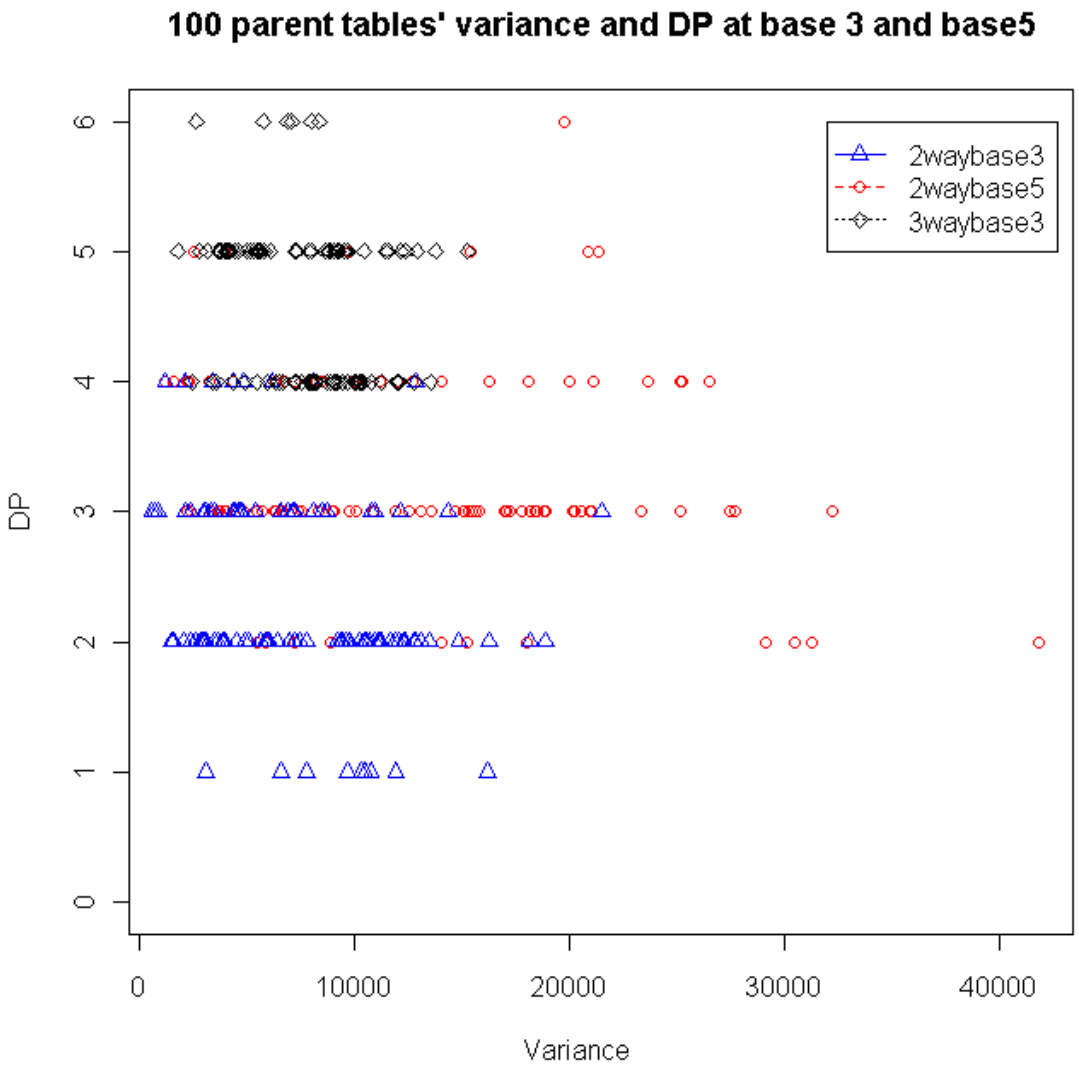

Figure 3.7. The variance and DP in two-, and three-way tables rounding using base 3 and base 5

Figure 3.7 plots the variance and DP in of two-way tables rounded using base 3 and 5 alongside those of a three-way contingency table rounded using base 3. Figure 3.7 shows a decreasing trend between a table's variance and the associated DP in the two-way tables rounded using base 3 and 5, and the three-way contingency table rounded using base 3. Data from the three datasets (each containing figures for 100 rounded tables) are represented by a different plot symbols. The plot showing the DP for the three-way contingency table appears to

be higher than the DP from the two-way contingency table. Therefore, the three-way contingency table has more confidential than the two-way contingency table, because in a three-way contingency table, more decimal places need to be reported to be able to reconstruct the original table. The three-way contingency table with rounded using base 3 is more confidential than the two-way contingency table rounded using base5.

This graph also presents the decreasing trend between a table's DP and its variance. As variance increases, the DP decreases. A table with high variance, will have less confidentiality compared with a table with low variance, because fewer decimal places need to be reported to be able to reconstruct the original parent table.

The most frequent DP appearing for the two-way contingency table rounded using base 3 is 2 or 3. The most frequent DP appearing for the two-way contingency with base 5 rouging is 3 or 4, while for the three-way table, the most frequent DP is 4 or 5.

3.4. Summary

This chapter suggests that a table with a high average is more confidential than another table with the same variance level. A table with low variance is more confidentiality than one with high variance, i.e. these four plots always follow the same order from highest to lowest: halv, lalv, hahv, and lahv.

The two-way contingency table with rounded using base 5 is more confidential than the two-way contingency table rounded using base3. The three-way contingency table is more confidential than the two-way contingency table,

because in three-way contingency tables more decimal places need to be reported to be able to reconstruct the original table.

Chapter 4

Effect of random rounding in log-linear models

This chapter gives the results of a study on census data analyzed with log-linear models. This chapter presents two related analyses on the effect of random rounding with log-linear models. Firstly the model using the random rounded table (model RR) is compared with the model using the actual data (model AC) to see whether or not they contain the same factors. Secondly, we compare the estimates from the RR and AC models and assess whether or not the hypothesis decisions change. We use these two methods to determine the effect of random rounding on model selection and interpretation.

4.1. Introduction

We obtained figures for age and sex for 1996, 2001 and 2006 censuses from the Statistic NZ table builder (http://wdmzpub01.stats.govt.nz/wds/TableViewer/tableView.aspx). We will discuss the effect of random rounding by comparing two different tables derived from the census data: the actual data and the rounded data. We assess two main performance measures to see if they are substantially affected by random rounding.

We present the effect of random rounding in three steps. In the first step, we use model selection to obtain a simplified model, and then check whether the simplified RR model and simplified AC model have the same factors or not. We then compare estimates in the RR model with those of the AC model to determine whether the hypothesis decision changes between models.

4.2. Data Analysis

4.2.1. Age by sex numerical table

We have four explanatory variables − year, age, sex, and region − in the given census table. The dependent variable is the count of people in each cell of the table. We denote the variables in this $3 \times 21 \times 2 \times 17$ table as *Y* for year, *A* for age, *S* for sex, and *R* for region. This published census table (which was rounded using base 3) can be produced from a number of possible actual tables. We can not produce all possible actual tables to analyze, because there are too many. As an alternative we produce a sample subset of all the possible actual tables. We used the upper limit tables (the worst tables), because rounded data in these are very different from the actual data. For example, the published cell count is 6 with rounding base three that the possible unrounded cell is 4, 5, 6, 7, or 8. So the unrounded cells 4 and 8 have the biggest discrepancy from the published cell. We called unrounded cell 4 and 8 are the worst possible parent cell to the published cell.

We have two ways of obtaining the upper limit tables. In the first method we choose the two worst possible parent tables. The worst dataset (WRa) was produced by adding 2 to the first row, subtracting 2 from the second row, and so on. The second worst dataset (WRb) was produced by adding 2 to the first column, subtracting 2 from the second column, and so on. Once we have obtained them, WRa and WRb will have rounded data that very different from the actual data.

In the second method, we randomly generate another 100 of the worst datasets (WR). The 100 worst datasets were produced by adding 2 or subtracting 2 randomly from each cell.

4.2.2. Model selection

In Chapter 2, we introduced our method of model selection. We want to obtain the minimally adequate model (the simplified model). We introduced four general models. The independent model is denote by (*Y, A, S, R*) with no interaction term. This first model has no relationship between variables. However, if an interaction term between each pair of variables appears, we symbolize this two-way interaction model by (*YA, YS, YR, AS, AR, SR*). If all interactions prove to be significant, then we will denoted the three-way interaction model by (*YAS, YSR, YAR, ASR*) and the four-way interaction model by (*YASR*) (Crawley, 2005).

General log-linear model with independent variable (*Y, A, S, R*):

$$\log \mu_{ijkl} = \lambda + \lambda_i^Y + \lambda_j^A + \lambda_k^S + \lambda_l^R \tag{4.1}$$

The second log-linear model has all four pairs of factors with conditional associations (*YA, YS,YR,AS,AR,SR*):

$$\log \mu_{ijkl} = \lambda + \lambda_i^Y + \lambda_j^A + \lambda_k^S + \lambda_l^R + \lambda_{ij}^{YA} + \lambda_{ik}^{YS} + \lambda_{il}^{YR} + \lambda_{jk}^{AS} + \lambda_{jl}^{AR} + \lambda_{kl}^{SR} \tag{4.2}$$

The third log-linear model has four triples of factors with conditional associations (*YAS, YSR, YAR, ASR*):

$$\log \mu_{ijkl} = \lambda + \lambda_i^Y + \lambda_j^A + \lambda_k^S + \lambda_l^R + \lambda_{ij}^{YA} + \lambda_{ik}^{YS} + \lambda_{il}^{YR} + \lambda_{jk}^{AS} + \lambda_{jl}^{AR} + \lambda_{kl}^{SR} + \lambda_{ijk}^{YAS} + \lambda_{ikl}^{YSR} + \lambda_{ijl}^{YAR} + \lambda_{jkl}^{ASR} \tag{4.3}$$

The fourth log-linear model is the most complex with all two-, three-, and four-way interactions (*YASR*):

$$\log \mu_{ijkl} = \lambda + \lambda_i^Y + \lambda_j^A + \lambda_k^S + \lambda_l^R + \lambda_{ij}^{YA} + \lambda_{ik}^{YS} + \lambda_{il}^{YR} + \lambda_{jk}^{AS} + \lambda_{jl}^{AR} + \lambda_{kl}^{SR} + \lambda_{ijk}^{YAS} + \lambda_{ikl}^{YSR} + \lambda_{ijl}^{YAR} + \lambda_{jkl}^{ASR} + \lambda_{ijkl}^{YASR} \tag{4.4}$$

We use a model selection process to find the minimally adequate model that describes the rounded table. We use the 'stepwise' procedure in R to determine

the model with best fit. We start with the maximum model, shown in Table 4.1. The output shows the steps that R goes through to determine the best fitted model. Note that the decisions are made using the AIC.

Table 4.1 Step output to determine the best model:

```
> step(glm4r)
Start:  AIC=23450.46
count ~ (year + age + sex + region)^4

                        Df  Deviance     AIC
- year:age:sex:region 640     758.3 22928.7
<none>                     3.818e-10 23450.5

Step:  AIC=22928.72
count ~ year + age + sex + region + year:age + year:sex + year:region +
    age:sex + age:region + sex:region + year:age:sex + year:age:region +
    year:sex:region + age:sex:region

                    Df Deviance     AIC
- year:sex:region   32    780.8 22887.2
<none>                    758.3 22928.7
- age:sex:region   320   2384.6 23915.0
- year:age:sex      40   2040.4 24130.9
- year:age:region  640   6497.6 27388.1

Step:  AIC=22887.24
count ~ year + age + sex + region + year:age + year:sex + year:region +
    age:sex + age:region + sex:region + year:age:sex + year:age:region +
    age:sex:region

                    Df Deviance     AIC
<none>                    780.8 22887.2
- age:sex:region   320   2403.7 23870.2
- year:age:sex      40   2063.7 24090.2
- year:age:region  640   6530.5 27357.0
```

Table 4.1 shows that the best fitted model (*YAS, YAR, ASR*) is:

$$\log \mu_{ijkl} = \lambda + \lambda_i^Y + \lambda_j^A + \lambda_k^S + \lambda_l^R + \lambda_{ij}^{YA} + \lambda_{ik}^{YS} + \lambda_{il}^{YR} + \lambda_{jk}^{AS} + \lambda_{jl}^{AR} + \lambda_{kl}^{SR} + \lambda_{ijk}^{YAS} + \lambda_{ijl}^{YAR} + \lambda_{jkl}^{ASR}.$$

Table 4.2 shows the summarized output of the (*YAS, YAR, ASR*) model. The null deviance is 1.8517×10^7, and the residual deviance is 7.078×10^2. We will use the (*YAS,YAR,ASR*) model as the best model for all further analysis.

Table 4.2 Output of model (*YAS, YAR, ASR*)

Table displays selected output only

```
> glm3r=glm(count~(year+age+sex+region)^3, family=poisson, data=recensus)
> glm5r<-update(glm3r,~.-year:sex:region)
> summary(glm5r)

Call:
glm(formula = count ~ year + age + sex + region + year:age +
    year:sex + year:region + age:sex + age:region + sex:region +
    year:age:sex + year:age:region + age:sex:region, family = poisson,
    data = recensus)

Deviance Residuals:
       Min          1Q      Median          3Q         Max
-2.1807901  -0.3660594  -0.0002461   0.3636733   2.1213848

Coefficients:
                                                  Estimate Std. Error   z value Pr(>|z|)
(Intercept)                                      3.224e+00  1.554e-01    20.743  < 2e-16 ***
year2001                                        -8.736e-02  1.784e-01    -0.490 0.624375
year2006                                        -1.882e-01  1.835e-01    -1.026 0.305044
age10-14 Years                                  -3.750e-02  2.265e-01    -0.166 0.868525
age100 Years and Over                           -2.014e+01  2.545e+03    -0.008 0.993684
age15-19 Years                                  -7.126e-01  2.623e-01    -2.717 0.006590 **
age20-24 Years                                   7.318e-02  2.249e-01     0.325 0.744932
age25-29 Years                                   1.865e-01  2.165e-01     0.862 0.388947
age30-34 Years                                   9.415e-02  2.147e-01     0.439 0.661004
age35-39 Years                                   2.700e-01  2.102e-01     1.285 0.198963
age40-44 Years                                   1.142e-01  2.186e-01     0.522 0.601559
age45-49 Years                                   9.047e-02  2.197e-01     0.412 0.680532
age5-9 Years                                    -6.615e-02  2.225e-01    -0.297 0.766266
age50-54 Years                                  -3.238e-01  2.375e-01    -1.363 0.172734
age55-59 Years                                  -5.794e-01  2.593e-01    -2.234 0.025465 *
age60-64 Years                                  -1.264e+00  3.164e-01    -3.994 6.50e-05 ***
age65-69 Years                                  -1.215e+00  3.310e-01    -3.670 0.000243 ***
age70-74 Years                                  -1.482e+00  3.686e-01    -4.021 5.79e-05 ***
age75-79 Years                                  -2.903e+00  6.437e-01    -4.510 6.50e-06 ***
age80-84 Years                                  -2.395e+00  6.183e-01    -3.874 0.000107 ***
age85-89 Years                                  -1.926e+01  1.839e+03    -0.010 0.991645
age90-94 Years                                  -2.052e+01  2.826e+03    -0.007 0.994206
age95-99 Years                                  -2.052e+01  2.826e+03    -0.007 0.994206
sexMale                                          4.858e-01  1.533e-01     3.168 0.001534 **
regionAuckland Region                            7.406e+00  1.555e-01    47.632  < 2e-16 ***
regionBay of Plenty Region                       5.877e+00  1.557e-01    37.754  < 2e-16 ***
regionCanterbury Region                          6.412e+00  1.556e-01    41.213  < 2e-16 ***
regionGisborne Region                            4.384e+00  1.565e-01    28.014  < 2e-16 ***
regionHawke's Bay Region                         5.395e+00  1.558e-01    34.623  < 2e-16 ***
regionManawatu-Wanganui Region                   5.852e+00  1.557e-01    37.587  < 2e-16 ***
regionMarlborough Region                         3.941e+00  1.571e-01    25.090  < 2e-16 ***
regionNelson Region                              3.953e+00  1.571e-01    25.169  < 2e-16 ***
regionNorthland Region                           5.419e+00  1.558e-01    34.781  < 2e-16 ***
regionOtago Region                               5.438e+00  1.558e-01    34.904  < 2e-16 ***
regionSouthland Region                           4.935e+00  1.561e-01    31.623  < 2e-16 ***
regionTaranaki Region                            5.065e+00  1.560e-01    32.474  < 2e-16 ***
regionTasman Region                              3.974e+00  1.570e-01    25.312  < 2e-16 ***
regionWaikato Region                             6.329e+00  1.556e-01    40.681  < 2e-16 ***
regionWellington Region                          6.411e+00  1.556e-01    41.209  < 2e-16 ***
regionWest Coast Region                          3.901e+00  1.572e-01    24.816  < 2e-16 ***
```

regionWest Coast Region	3.901e+00	1.572e-01	24.816	< 2e-16	***
year2001:age10-14 Years	-7.939e-02	2.650e-01	-0.300	0.764475	
year2006:age10-14 Years	-2.714e-01	2.811e-01	-0.965	0.334343	
year2001:age100 Years and Over	-4.968e-01	3.507e+03	-1.42e-04	0.999887	
year2006:age100 Years and Over	-4.067e-01	3.513e+03	-1.16e-04	0.999908	
year2001:age15-19 Years	3.083e-01	2.861e-01	1.078	0.281189	
year2006:age15-19 Years	1.011e-01	3.029e-01	0.334	0.738540	
year2001:age20-24 Years	-3.178e-01	2.795e-01	-1.137	0.255580	
year2006:age20-24 Years	-2.237e-01	2.828e-01	-0.791	0.429001	
year2001:age25-29 Years	-1.854e-01	2.535e-01	-0.731	0.464568	
year2006:age25-29 Years	-6.854e-01	2.845e-01	-2.410	0.015968	*
year2001:age30-34 Years	2.821e-01	2.446e-01	1.153	0.248738	
year2006:age30-34 Years	1.209e-01	2.560e-01	0.472	0.636800	
year2001:age35-39 Years	-3.432e-02	2.510e-01	-0.137	0.891250	
year2006:age35-39 Years	2.175e-01	2.504e-01	0.869	0.384873	
year2001:age40-44 Years	1.524e-01	2.570e-01	0.593	0.553304	
year2006:age40-44 Years	9.811e-02	2.662e-01	0.369	0.712417	
year2001:age45-49 Years	-1.669e-01	2.618e-01	-0.637	0.523867	
year2006:age45-49 Years	-1.240e-01	2.679e-01	-0.463	0.643619	
year2001:age5-9 Years	8.106e-02	2.522e-01	0.321	0.747866	
year2006:age5-9 Years	-2.093e-02	2.630e-01	-0.080	0.936573	
year2001:age50-54 Years	3.861e-01	2.660e-01	1.452	0.146533	
year2006:age50-54 Years	1.381e-01	2.824e-01	0.489	0.624794	
year2001:age55-59 Years	9.168e-02	3.041e-01	0.302	0.763013	
year2006:age55-59 Years	3.629e-01	2.994e-01	1.212	0.225415	
year2001:age60-64 Years	6.168e-01	3.475e-01	1.775	0.075867	.
year2006:age60-64 Years	6.096e-01	3.554e-01	1.715	0.086301	.
year2001:age65-69 Years	2.736e-01	3.925e-01	0.697	0.485850	
year2006:age65-69 Years	5.346e-01	3.847e-01	1.390	0.164630	
year2001:age70-74 Years	-6.587e-01	4.456e-01	-1.478	0.139357	
year2006:age70-74 Years	-2.783e-01	4.155e-01	-0.670	0.503036	
year2001:age75-79 Years	1.134e+00	6.902e-01	1.643	0.100390	
year2006:age75-79 Years	1.172e+00	6.918e-01	1.695	0.090123	.
year2001:age80-84 Years	8.366e-02	8.358e-01	0.100	0.920264	
year2006:age80-84 Years	8.463e-01	7.308e-01	1.158	0.246797	
year2001:age85-89 Years	1.722e+01	1.839e+03	0.009	0.992529	
year2006:age85-89 Years	1.475e-01	2.628e+03	5.62e-05	0.999955	
year2001:age90-94 Years	9.046e-02	3.468e+03	2.61e-05	0.999979	
year2006:age90-94 Years	1.599e-01	3.468e+03	4.61e-05	0.999963	
year2001:age95-99 Years	9.430e-02	3.468e+03	2.72e-05	0.999978	
year2006:age95-99 Years	1.564e-01	3.468e+03	4.51e-05	0.999964	
year2001:sexMale	-1.287e-02	5.396e-03	-2.386	0.017043	*
year2006:sexMale	-2.016e-02	5.378e-03	-3.749	0.000178	***
year2001:regionAuckland Region	1.277e-01	1.784e-01	0.716	0.474260	
year2006:regionAuckland Region	2.935e-01	1.836e-01	1.599	0.109846	
year2001:regionBay of Plenty Region	7.500e-02	1.787e-01	0.420	0.674663	
year2006:regionBay of Plenty Region	1.785e-01	1.838e-01	0.971	0.331494	
year2001:regionCanterbury Region	7.511e-02	1.786e-01	0.421	0.673994	
year2006:regionCanterbury Region	2.324e-01	1.837e-01	1.265	0.205847	
year2001:regionGisborne Region	1.881e-03	1.798e-01	0.010	0.991651	
year2006:regionGisborne Region	5.295e-02	1.849e-01	0.286	0.774581	
year2001:regionHawke's Bay Region	3.169e-02	1.789e-01	0.177	0.859390	
year2006:regionHawke's Bay Region	1.044e-01	1.840e-01	0.567	0.570617	

```
age95-99 Years:sexMale:regionWaikato Region                -7.128e-01  2.832e+03 -2.52e-04 0.999799
age10-14 Years:sexMale:regionWellington Region              1.812e-01  2.294e-01     0.790 0.429406
age100 Years and Over:sexMale:regionWellington Region      -7.286e-01  2.904e+03 -2.51e-04 0.999800
age15-19 Years:sexMale:regionWellington Region             -2.197e-01  2.501e-01    -0.878 0.379811
age20-24 Years:sexMale:regionWellington Region              3.603e-01  2.352e-01     1.532 0.125590
age25-29 Years:sexMale:regionWellington Region              4.327e-02  2.233e-01     0.194 0.846351
age30-34 Years:sexMale:regionWellington Region              4.019e-02  2.089e-01     0.192 0.847434
age35-39 Years:sexMale:regionWellington Region              2.970e-01  2.093e-01     1.419 0.155840
age40-44 Years:sexMale:regionWellington Region              3.814e-01  2.177e-01     1.752 0.079787 .
age45-49 Years:sexMale:regionWellington Region              1.718e-01  2.227e-01     0.771 0.440431
age5-9 Years:sexMale:regionWellington Region               -4.173e-02  2.185e-01    -0.191 0.848563
age50-54 Years:sexMale:regionWellington Region              3.807e-02  2.275e-01     0.167 0.867104
age55-59 Years:sexMale:regionWellington Region              1.321e-01  2.500e-01     0.528 0.597301
age60-64 Years:sexMale:regionWellington Region             -6.753e-03  2.814e-01    -0.024 0.980854
age65-69 Years:sexMale:regionWellington Region              3.634e-01  3.126e-01     1.162 0.245059
age70-74 Years:sexMale:regionWellington Region             -5.248e-01  3.796e-01    -1.383 0.166744
age75-79 Years:sexMale:regionWellington Region             -1.248e-01  4.673e-01    -0.267 0.789423
age80-84 Years:sexMale:regionWellington Region              1.037e+00  6.845e-01     1.515 0.129713
age85-89 Years:sexMale:regionWellington Region              1.577e+01  1.111e+03     0.014 0.988677
age90-94 Years:sexMale:regionWellington Region             -6.239e-01  2.832e+03 -2.20e-04 0.999824
age95-99 Years:sexMale:regionWellington Region             -8.861e-01  2.832e+03 -3.13e-04 0.999750
age10-14 Years:sexMale:regionWest Coast Region              1.310e-01  2.317e-01     0.565 0.571812
age100 Years and Over:sexMale:regionWest Coast Region      -8.733e-01  2.904e+03 -3.01e-04 0.999760
age15-19 Years:sexMale:regionWest Coast Region             -1.166e-01  2.526e-01    -0.462 0.644303
age20-24 Years:sexMale:regionWest Coast Region              4.419e-01  2.382e-01     1.855 0.063553 .
age25-29 Years:sexMale:regionWest Coast Region              1.253e-02  2.261e-01     0.055 0.955819
age30-34 Years:sexMale:regionWest Coast Region              1.554e-02  2.116e-01     0.073 0.941437
age35-39 Years:sexMale:regionWest Coast Region              3.290e-01  2.119e-01     1.553 0.120498
age40-44 Years:sexMale:regionWest Coast Region              4.763e-01  2.202e-01     2.163 0.030521 *
age45-49 Years:sexMale:regionWest Coast Region              2.675e-01  2.251e-01     1.188 0.234729
age5-9 Years:sexMale:regionWest Coast Region               -8.161e-02  2.210e-01    -0.369 0.711909
age50-54 Years:sexMale:regionWest Coast Region              1.474e-01  2.300e-01     0.641 0.521609
age55-59 Years:sexMale:regionWest Coast Region              1.873e-01  2.525e-01     0.742 0.458245
age60-64 Years:sexMale:regionWest Coast Region              7.414e-02  2.839e-01     0.261 0.793957
age65-69 Years:sexMale:regionWest Coast Region              4.783e-01  3.150e-01     1.518 0.128928
age70-74 Years:sexMale:regionWest Coast Region             -4.028e-01  3.818e-01    -1.055 0.291444
age75-79 Years:sexMale:regionWest Coast Region             -2.697e-02  4.695e-01    -0.057 0.954194
age80-84 Years:sexMale:regionWest Coast Region              1.130e+00  6.866e-01     1.646 0.099692 .
age85-89 Years:sexMale:regionWest Coast Region              1.577e+01  1.111e+03     0.014 0.988679
age90-94 Years:sexMale:regionWest Coast Region             -4.090e-01  2.832e+03 -1.44e-04 0.999885
age95-99 Years:sexMale:regionWest Coast Region             -8.623e-01  2.832e+03 -3.05e-04 0.999757
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.8519e+07  on 2141  degrees of freedom
Residual deviance: 7.8078e+02  on  672  degrees of freedom
AIC: 22887
```

We repeat the modelling process and use the 'stepwise' procedure of model selection for the worst two possible parent datasets. A comparison of the rounded table model with the models from the worst two datasets (WRa and WRb) shows that the same effects are seen in all three minimally adequate models. Table 4.3 shows the AIC, residual deviance and degrees of freedom (df) of the three different models and datasets. We see the AIC and residual deviances are very similar, with same degrees of freedom appearing for all three datasets.

In the next section, we will look at whether estimates produced by the three models lead to a change in hypothesis decision.

Table 4.3 Comparison of the factors in all three cases: WRa and WRb and RR

		Best Model
		(YAS, YAR, ASR)
RD	***RR***	780.78
	WRa	764.06
	WRb	769.11
DF	***RR***	672
	WRa	672
	WRb	672
AIC	***RR***	22887
	WRa	22994
	WRb	23018
P> chi-square		<0.01

Note:

RD: residual deviance

DF: degrees of freedom of residual deviance

RR: given rounded table

WRa: worst possible parent data for the given rounded table

WRb: second worst possible parent data for the given rounded table

4.2.3. Effect of random rounding

Currently, we have three datasets: the rounded table, the worst possible parent table (WRa), and the second worst possible parent table (WRb). This section compares the estimates from the three datasets which produced changes in the hypothesis decisions. We are interested in knowing whether estimates using the worst datasets can change hypothesis decisions compared to the RR model.

The best fitting model (*YAS, YAR, ASR*) produced 1470 estimates for each dataset. Table 4.4 shows the number of significant and non-significant estimates in all three datasets. The three datasets yielded different numbers of estimates that were significant at the 5% level.

Table 4.4 Number of estimates to be significant or not among the three datasets.

		(YAS, YAR, ASR)
Rounded table (RRT)	***No. of significant***	147
	No. of not significant	1323
	Total	1470
1st worse case (WRa)	***No. of significant***	228
	No. of not significant	1242
	Total	1470
2nd worse case (WRb)	***No. of significant***	159
	No. of not significant	1311
	Total	1470

Further investigation showed that the number of estimates that changed the hypothesis decision is in fact a lot higher than suggested in this table. Table 4.5 shows the number of estimates that produced a change in hypothesis decision between the rounded table and two worst datasets. The ideal case would be no change in the hypothesis decisions occurred after random rounding. If a high frequency of change is seen, then the error from random rounding will be of considerable concern, because it may lead to incorrect interpretations.

Table 4.5 Estimates in WRa and WRb that change the hypothesis decision

	WRa		*WRb*		*Total*
	No. of estimates producing change in the h0 decision	***No. of estimates producing no change in the h0 decision***	***No. of estimates producing change in the h0 decision***	***No. of estimates producing no change in the h0 decision***	
(YAS, YAR, ASR)	91	1379	62	1408	1470

Table 4.5 shows 6.1% (91/1470) of the estimates cause a change in the hypothesis decision between the rounded table (RR) and WRa, and 4.2% (62/1470) of the estimates change the hypothesis decision between the rounded table and WRb.

Table 4.6 compares WRa and WRb against the rounded table, where each dataset produced 1470 estimates. The resulting estimates are compared for change in significance. The number of estimates that are significant in WRa and WRb, but not in RR is always greater than the number of estimates that are significant in RR, but not in WRa and/or WRb. In other words, this simulation describes more estimates move from non-significant to significant compared to the other way round, when the worst possible parent tables have been constructed from the randomly rounded table. This addition of significant estimates is a concern for random rounding because of the effect of this on interpretation.

Table 4.6 Number of estimates changing the hypothesis decision (2)

		WRa		
		p>0.05	p<0.05	Total
RR	p>0.05	1237	86	1323
	p<0.05	5	142	147
	Total	1242	228	1470

		WRb		
		p>0.05	p<0.05	Total
RR	p>0.05	1286	37	1323
	p<0.05	25	122	147
	Total	1311	159	1470

This analysis has used the worst two datasets. The pattern used to generate these two datasets was to add 2 to the first row or column, subtracting from the second, and so on. Next, we generate another 100 worst cases that have been produced by adding or subtracting 2 randomly (WR1 to WR100). Note that the worst cases have been produced by adding or subtracting 2, because with base 3 rounding, the rounded cell is from an interval that has a width of 5.The WR cells are either the upper limit of the interval (RR + 2) or the lower limit of the interval(RR - 2). After constructing the 100 worst datasets in this way, we use the same study design as for the WRa and WRb datasets.

We calculate the number of significant and non-significant estimates for each of the WR datasets. The 100 WR models each contain 1470 estimates. For each, we calculate the percentage of estimates that are significant out of the total number of estimates (1470) for each WR dataset, named *per_wr*. Figure 4.7 shows 100 *per_wr* for the 100 WR datasets. The straight line is the percentage of the number of estimates that are significant in the RR model (147/1470). The *per_wr* points are consistently above that straight line. This means the WR datasets always have more significant estimates than the rounded table. This error could make a difference for the user who uses a randomly rounded table

instead of original parent table. For example, if estimates change the hypothesis decision (e.g. estimates are non-significant in the RR model but significant in the possible parent table) then some estimates could be removed from the model constructed from the RR table. If the real unrounded data were used, these estimates would have been included in the model.

Next, we will compare the datasets WR1-WR100 to the RR dataset and see how many estimates are significant in the WR datasets compared to the RR dataset.

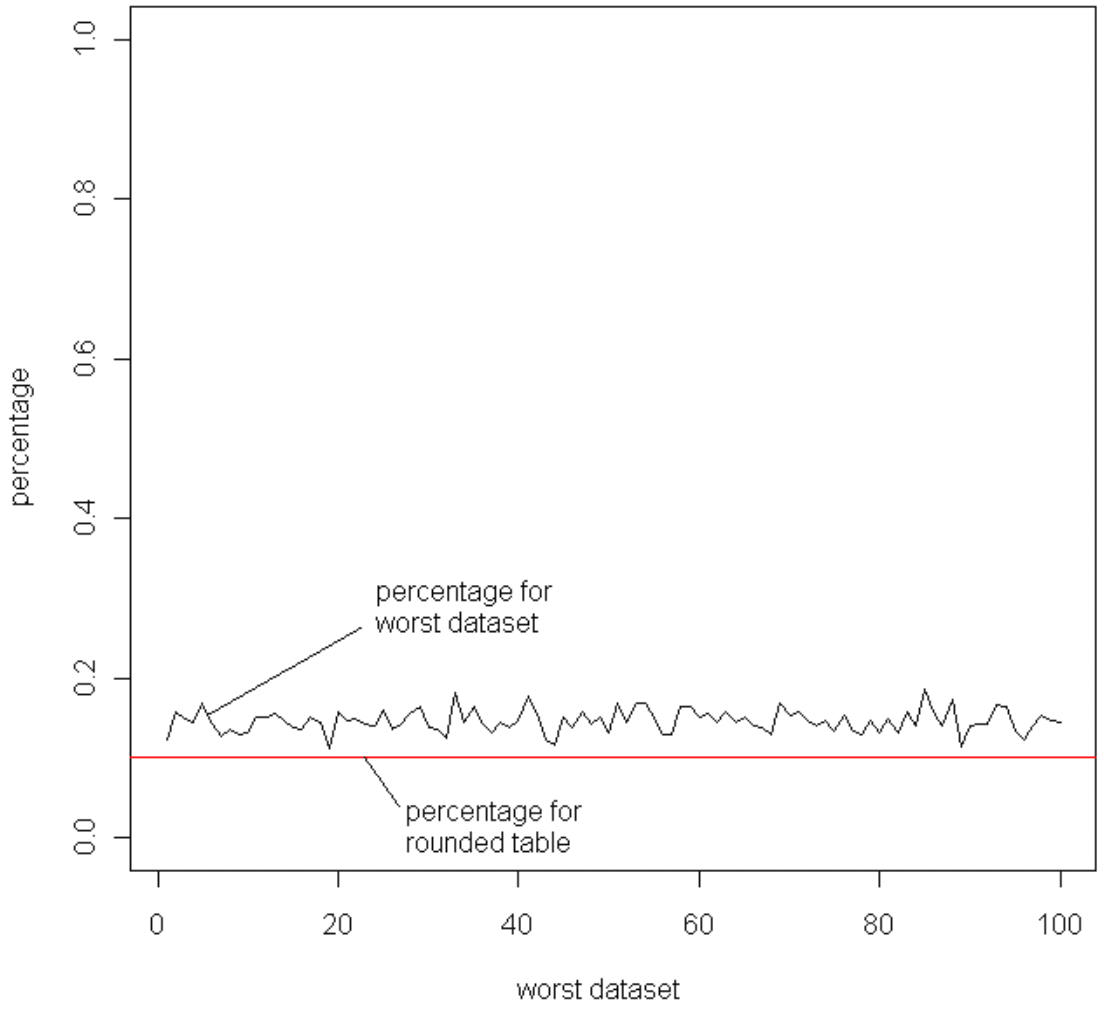

Figure 4.7 Percentage of estimates that are significant within WR datasets.

Table 4.8 shows the number of estimates that changed the hypothesis decisions in each WR dataset compared to the RR dataset. The first column shows the number of the WR dataset (e.g. WR1). The second column shows the number of estimates that are significant in that WR dataset, but not in the RR dataset. The third column shows the number of estimates that are not significant in the WR dataset, but are significant in the RR dataset. The fourth column shows the number of estimates that are significant in both the WR dataset and the RR dataset. The fifth column shows the number of estimates that are not significant in neither the WR dataset and nor the RR dataset.

Table 4.8 Number of estimates changing the hypothesis decision.

	Changed h0 decisions		***Unchanged h0 decisions***			***Changed h0 decisions***		***Unchanged h0 decisions***	
WR	***WR (p≤0.05) & RR (p>0.05)***	***WR (p>0.05) & RR (p≤0.05)***	***WR (p≤0.05) & RR (p≤0.05)***	***WR (p>0.05) & RR (p>0.05)***	***WR***	***WR (p≤0.05) & RR (p>0.05)***	***WR (p>0.05) & RR (p≤0.05)***	***WR (p≤0.05) & RR (p≤0.05)***	***WR (p>0.05) & RR (p>0.05)***
1	65	30	117	1258	***51***	121	20	127	1202
2	101	16	131	1222	***52***	76	11	136	1247
3	88	17	130	1235	***53***	128	26	121	1195
4	81	15	132	1242	***54***	123	23	124	1200
5	124	23	124	1199	***55***	94	17	130	1229
6	82	18	129	1241	***56***	66	23	124	1257
7	54	13	134	1269	***57***	66	22	125	1257
8	75	23	124	1248	***58***	118	24	123	1205
9	74	31	116	1249	***59***	117	22	125	1206
10	78	33	114	1245	***60***	112	37	110	1211
11	96	22	125	1227	***61***	108	27	120	1215
12	95	20	127	1228	***62***	90	24	123	1233
13	102	19	128	1221	***63***	98	13	134	1225
14	91	21	126	1232	***64***	84	18	129	1239
15	98	43	104	1225	***65***	89	15	132	1234
16	74	21	126	1249	***66***	81	23	124	1242
17	92	16	131	1231	***67***	92	36	111	1231
18	85	19	128	1238	***68***	64	20	127	1259
19	51	32	115	1272	***69***	105	5	142	1218
20	109	23	124	1214	***70***	93	13	134	1230
21	95	27	120	1228	***71***	98	14	133	1225
22	101	30	117	1222	***72***	83	15	132	1240
23	85	23	124	1238	***73***	78	20	127	1245

24	74	13	134	1249	***74***	84	15	132	1239
25	107	18	129	1216	***75***	82	33	114	1241
26	75	23	124	1248	***76***	93	14	133	1230
27	71	8	139	1252	***77***	65	15	132	1258
28	98	17	130	1225	***78***	57	13	134	1266
29	110	16	131	1213	***79***	88	18	129	1235
30	79	24	123	1244	***80***	84	39	108	1239
31	73	21	126	1250	***81***	88	16	131	1235
32	65	29	118	1258	***82***	73	28	119	1250
33	150	29	118	1173	***83***	96	12	135	1227
34	81	15	132	1242	***84***	94	36	111	1229
35	102	7	140	1221	***85***	140	13	134	1183
36	78	16	131	1245	***86***	115	30	117	1208
37	66	18	129	1257	***87***	76	15	132	1247
38	79	14	133	1244	***88***	115	9	138	1208
39	69	13	134	1254	***89***	46	26	121	1277
40	81	12	135	1242	***90***	81	21	126	1242
41	126	11	136	1197	***91***	97	33	114	1226
42	99	17	130	1224	***92***	81	18	129	1242
43	59	24	123	1264	***93***	110	11	136	1213
44	73	49	98	1250	***94***	117	23	124	1206
45	83	9	138	1240	***95***	75	26	121	1248
46	73	18	129	1250	***96***	69	35	112	1254
47	93	8	139	1230	***97***	81	22	125	1242
48	94	31	116	1229	***98***	95	16	131	1228
49	110	33	114	1213	***99***	97	27	120	1226
50	79	34	113	1244	***100***	80	15	132	1243

Table 4.8 shows that the number of estimates that are significant in the WR dataset, but not in the RR dataset is always greater than the number of estimates that are significant in the RR dataset, but not in the WR datasets. This is the same result that was obtained by comparing the RR dataset compared to the WRa and WRb datasets.

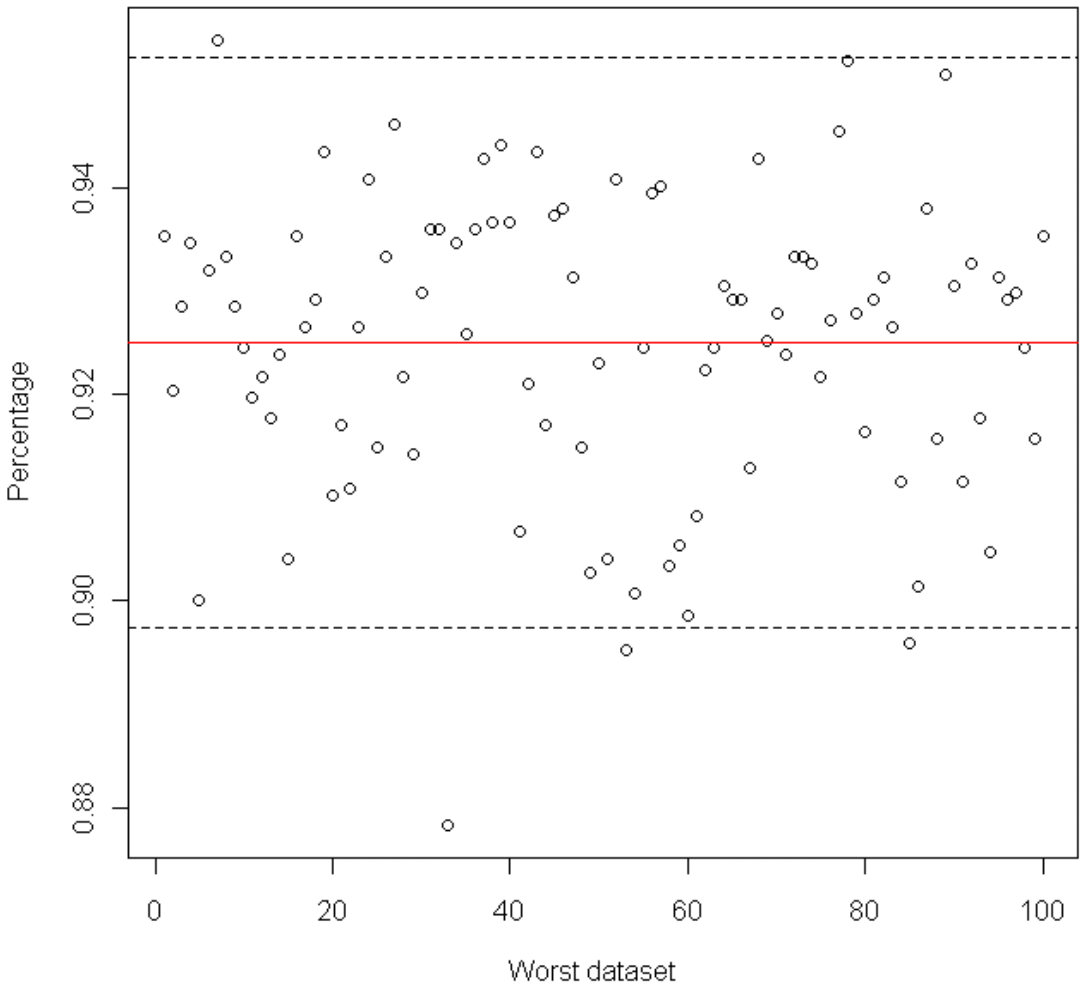

Figure 4.9 Percentage of estimate that do not change the hypothesis decision.

Figure 4.9 shows the percentage of the estimates (out of the total number of estimates, e.g. (117+1258)/1470) that do not change the hypothesis decision between each WR dataset and the RR dataset. The percentage confidence interval at the 5% level is [0.8974, 0.9527]. Therefore, almost over 90% estimates in each WR dataset do not change the hypothesis decision, as compared to the RR dataset.

4.3. Summary

We have presented two analyses to see the effect of random rounding with the log-linear model in a four-way contingency table. The rounded table and the worst possible parent tables have the same effects in all three minimally adequate models. The AIC and residual deviance are very similar, with the same degrees of freedoms being found in the rounded and the WR datasets. Random error is not a concern at this stage. In the second analysis, we looked at whether estimates in the WR datasets changed the hypothesis decision compared to the rounded dataset. In this stage, we found the number of estimates that change the hypothesis in each WR datasets compared to the RR dataset. The WR datasets always have more estimates that are significant than the rounded. This error could make a difference for the user who uses the RR table instead of the original table. This error is trivial in the four-way table because over 90% estimates in each WR dataset do not change the hypothesis decisions compared to the RR dataset.

We will repeat a similar study design with two- and three-way contingency tables in Chapter 5.

Chapter 5

Simulation for two- and three-way contingency tables

Chapter 4 presented the effect of random rounding for four-way contingency tables. We repeat a similar study design with two- and three-way contingency tables in chapter 5. This chapter gives the results of this study for two types of census data analysed with log-linear models. These two types of census data include a three-way contingency table and a two-way contingency table, which were obtained from the Statistic NZ table builder. The three-way contingency table correlates sex, region and qualification level.

(http://wdmzpub01.stats.govt.nz/wds/TableViewer/tableView.aspx). The two-way contingency table correlates income source and income level

(http://wdmzpub01.stats.govt.nz/wds/TableViewer/tableView.aspx).

5.1. Three-way contingency table

5.1.1. Tables

We have three explanatory variables: qualification, sex and region in the given census table. We denote these variables in this $13 \times 2 \times 17$ table as *Q* for qualification, *S* for sex, and *R* for region. This published census table is rounded using base 3 and is denoted as RR'. This table can be produced from a huge number of possible actual (parent) tables. These worst two datasets were generated using by the pattern described in chapter 4; these are called WRa' and WRb'. We also generated the 100 worst datasets randomly, denoted WR1'−WR100'. We present the effect of random rounding in this numerical table.

5.1.2. Model selection

We use the backward and ‘stepwise’ procedures in R to determine the best fitting model of RR'. The best fitting model is:

$$\log \mu_{ijk} = \lambda + \lambda_i^Q + \lambda_j^S + \lambda_k^R + \lambda_{ij}^{QS} + \lambda_{ik}^{QR} + \lambda_{jk}^{SR} + \lambda_{ijk}^{QSR}$$

This model will be called (*QSR*). We repeat the modelling process and use the 'stepwise' procure in model selection to find best fitting models for WRa’ and WRb’. A comparison of the RR' model with the WRa' model and the WRb' model, shows that the same effects are seen in all three minimally adequate models.

Table 5.1 shows the AIC and residual deviance (RD) are very similar with the same degrees of freedoms (DF) appearing in all three datasets in the best fitting model. In the next section, we will look at whether the number of estimates that produce a change in hypothesis decisions in the models.

Table 5.1 Comparison of the factors in three cases: WRa', WRb' and RR'.

		Best Fitting Model
		(QSR)
RD	***RR'***	4.815x10^-10
	WRa'	-7.241x10^-12
	WRb'	-3.376x10^-11
DF	***RR'***	0
	WRa'	0
	WRb'	0
AIC	***RR'***	5015.3
	WRa'	5022.1
	WRb'	5028.4
P> chi-square		<0.01

5.1.3. Effect of random rounding

Currently, we have 103 datasets: the rounded table, the worst two possible parent tables generated by rounding according to a pattern (WRa' and WRb') and the 100 worst possible parent tables generated by random rounding (WR1'−WR100'). This section compares the number of estimates that produce a change in hypothesis decision in each of the 103 datasets. We are interested in knowing whether estimates in the worst datasets change hypothesis decisions, compared to the RR'.

We calculate the number of estimates that are significant, or not, for each of the worst datasets. The 102 worst (WR') models each contained 442 estimates. For each, we calculate the percentage of estimates that are significant out of the total number of estimates (442) for each WR' dataset, named *per_wr'*. Figure 5.1 shows the *per_wr'* for each of the 102 WR' datasets. The straight line is the number of estimates that are significant from the rounded table (67/442). The *per_wr'* points are consistently above that straight line. This means that the WR' datasets always have more estimates that are significant than the rounded tables. This result is same as was seen with the four-way contingency tables in Chapter 4.

Next, we will compare each of the WR' datasets with the RR' datasets and see how many estimates do not produce a change in hypothesis decision in the WR' datasets, compared with the RR' dataset.

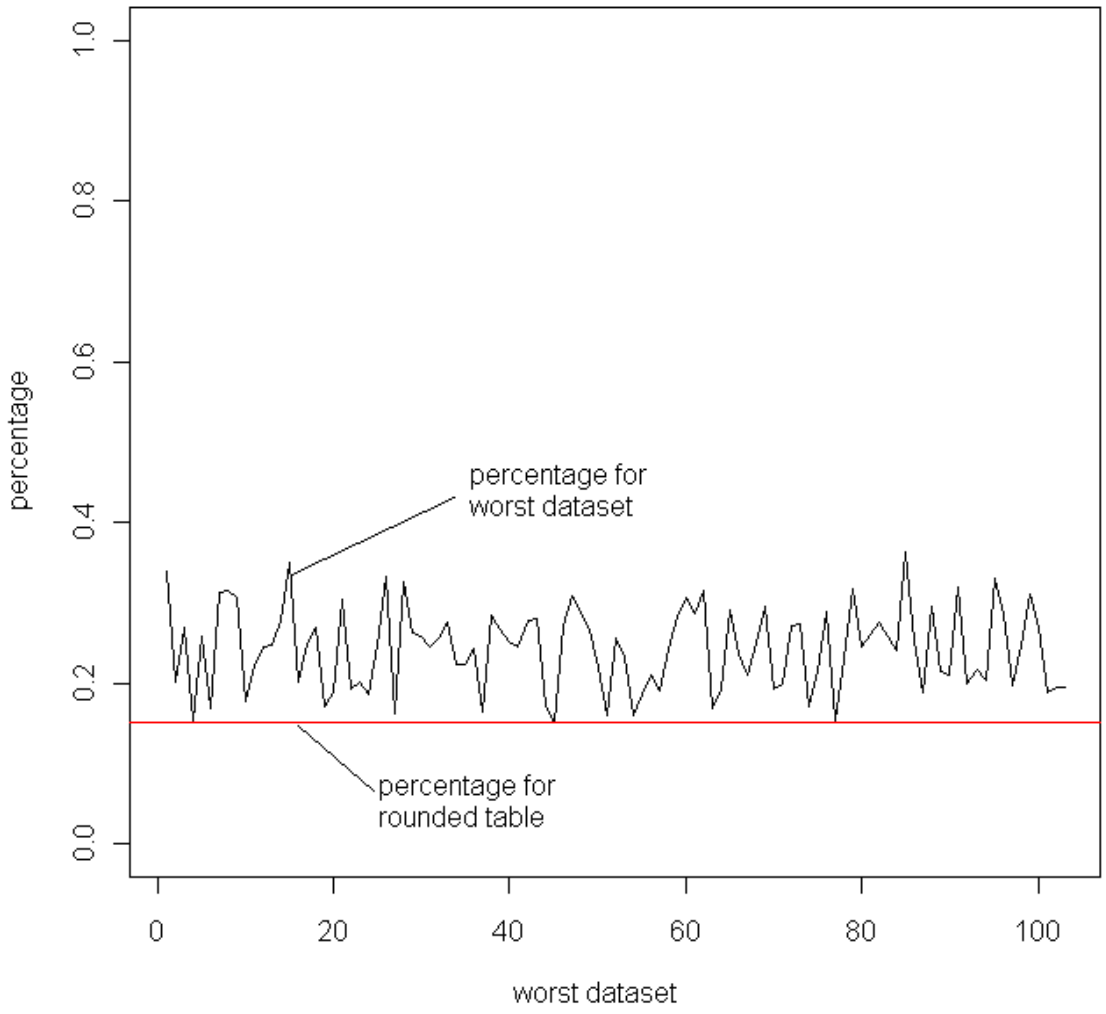

Figure 5.1 Percentage of significant estimates in each WR' dataset.

Table 5.2 shows the number of estimates that changed the hypothesis decision in each WR' dataset compared to the RR' dataset. The first column shows the number of the WR' dataset. The second column shows the number of estimates that were significant in the WR' dataset, but not in the RR' dataset. The third column shows the number of estimates that were not significant in the WR' dataset, but were significant in the RR' dataset. The fourth column shows the number of estimates that are significant in the both the WR' dataset and the RR' dataset. The fifth column shows the number of estimates that were not significant in neither the WR' dataset and nor the RR' dataset.

Table 5.2 Number of estimates producing a change hypothesis decision.

	Changed h0 decisions		Unchanged h0 decisions			Changed h0 decisions		Unchanged h0 decisions	
WR'	WR' (p≤0.05) & RR' (p>0.05)	WR' (p>0.05) & RR' (p≤0.05)	WR' (p≤0.05) & RR' (p≤0.05)	WR' (p>0.05) & RR' (p>0.05)	WR'	WR' (p≤0.05) & RR' (p>0.05)	WR' (p>0.05) & RR' (p≤0.05)	WR' (p≤0.05) & RR' (p≤0.05)	WR' (p>0.05) & RR' (p>0.05)
1	84	1	66	291	52	50	4	63	325
2	30	8	59	345	53	43	7	60	332
3	67	15	52	308	54	20	16	51	355
4	18	17	50	357	55	29	14	53	346
5	54	7	60	321	56	34	8	59	341
6	26	18	49	349	57	25	8	59	350
7	80	9	58	295	58	47	7	60	328
8	77	5	62	298	59	63	5	62	312
9	74	6	61	301	60	70	2	65	305
10	27	16	51	348	61	64	4	63	311
11	40	9	58	335	62	72	0	67	303
12	55	14	53	320	63	21	13	54	354
13	46	4	63	329	64	32	14	53	343
14	59	3	64	316	65	73	11	56	302
15	88	0	67	287	66	42	4	63	333
16	31	9	58	344	67	41	15	52	334
17	55	12	55	320	68	49	6	61	326
18	68	16	51	307	69	68	4	63	307
19	26	17	50	349	70	31	13	54	344
20	33	17	50	342	71	33	12	55	342
21	67	0	67	308	72	60	7	60	315
22	27	9	58	348	73	68	14	53	307
23	33	11	56	342	74	26	17	50	349
24	27	12	55	348	75	41	13	54	334
25	58	16	51	317	76	64	3	64	311
26	80	0	67	295	77	21	20	47	354
27	21	16	51	354	78	45	10	57	330
28	79	2	65	296	79	76	3	64	299
29	67	18	49	308	80	58	17	50	317
30	53	6	61	322	81	55	7	60	320
31	49	8	59	326	82	58	3	64	317
32	60	14	53	315	83	50	3	64	325
33	58	3	64	317	84	53	14	53	322
34	52	20	47	323	85	94	0	67	281
35	37	5	62	338	86	63	18	49	312
36	46	6	61	329	87	25	9	58	350
37	22	16	51	353	88	68	4	63	307
38	63	4	63	312	89	41	13	54	334

39	64	13	54	311	90	44	18	49	331
40	51	7	60	324	91	75	1	66	300
41	48	7	60	327	92	31	10	57	344
42	67	11	56	308	93	44	15	52	331
43	71	14	53	304	94	34	11	56	341
44	29	19	48	346	95	80	1	66	295
45	18	18	49	357	96	73	13	54	302
46	59	6	61	316	97	27	7	60	348
47	78	9	58	297	98	48	6	61	327
48	65	4	63	310	99	70	0	67	305
49	65	14	53	310	100	60	8	59	315
50	39	9	58	336	WRa'	23	7	60	352
51	21	17	50	354	WRb'	27	8	59	348

Table 5.2 shows that the number of estimates that are significant in the WR' datasets, but not significant in the RR' dataset is always greater than the number of estimates that are significant in the RR' dataset but not in the WR' datasets. This result is same as the result obtained for the four-way contingency table.

Figure 5.2 shows the percentage of the estimates (out of the total number of estimates, e.g. (66+291)/442) that do not change the hypothesis decision between each WR' dataset and the RR' table. The percentage confidence interval at the 5% level is [0.79, 0.94]. Therefore, around 87% estimates do not change the hypothesis decisions in each of the WR' datasets, compared with the RR' dataset. This result is very close to the result which was obtained for the four-way contingency table (Chapter 4).

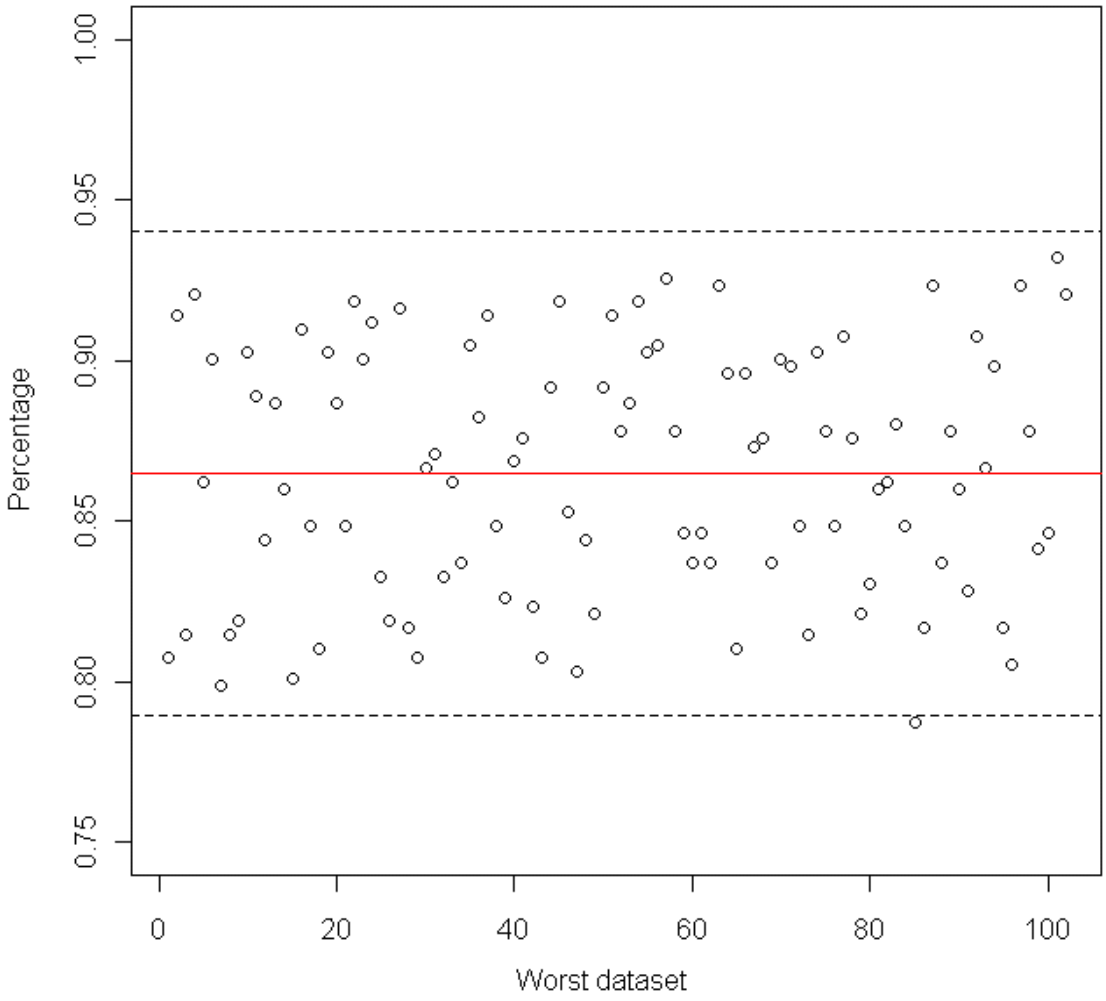

Figure 5.2 Percentage of estimates that do not change the hypothesis decision.

5.2. Two-way contingency tables

5.2.1. Tables

We have two explanatory variables − source and income − in the given census table. We denote these variables in this 15 x 14 table as *S* for source, and *I* for income. This published census table rounded using base 3, denoted as RR'', can

be produced from a huge number of possible actual (parent) tables. We generate the worst two datasets using the pattern described in chapter 4 these were denoted as WRa'' and WRb''. We also generated the worst 100 datasets randomly, denoted as WR1''−WR100''. We present the effect of random rounding in this numerical table.

5.2.2. Model selection

We used the backward and 'stepwise' procedure in R to determine the best fitting model of RR''. The best fitted model is $\log \mu_{ij} = \lambda + \lambda_i^S + \lambda_j^I + \lambda_{ij}^{SI}$, denoted as (*SI*). Note that most estimates are highly significant in this two-way contingency table. We repeat the modelling process and use the 'stepwise' procure in model selection to find the best fitting models for WRa'' and WRb''. A comparison of the RR'' model with the WRa'' model and the WRb'' model shows that the same effects are seen in all three minimally adequate models.

Table 5.3 shows the AIC, residual deviance and degrees of freedom of the three models and datasets. We see that AIC and residual deviance are very similar with the same degrees of freedom being seen in all three datasets in the best fitting model. In the next section, we will look at whether different estimates change the hypothesis decisions for all three datasets.

Table 5.3 Comparison of the factors in three cases: WRa'' and WRb'' and RR''.

		Best Fitted Model
		(SI)
RD	***RR''***	4.6802x10^-11
	WRa''	-5.0825x10^-11
	WRb''	-8.5050x10^-12
DF	***RR''***	0
	WRa''	0
	WRb''	0
AIC	***RR''***	2357.4
	WRa''	2357.1
	WRb''	2397.5
P> chi-square		<0.01

5.2.3. Effect of random rounding

Currently, we have 103 datasets: the rounded table, the worst two possible parent tables generated using a pattern (WRa'' and WRb'') and the worst 100 possible parent tables generated by random rounding (WR1''−WR100''). This section compares the number of estimates that change the in hypothesis decisions in all 103 datasets. We are interested in knowing whether estimates in the worst (WR'') datasets change the hypothesis decisions, compared to the RR'' dataset.

We calculate the number of estimates that are significant, or not, for each WR'' dataset. The 102 WR'' models each contained 210 estimates. For each, we calculate the percentage of estimates that are significant out of the total number of estimates (210) for each WR'' dataset, named *per_wr''*. Figure 5.3 plots the *per_wr''* of the 102 WR'' datasets. The straight line is the percentage of the

number of significant estimates from the rounded table (180/210). The points of *per_wr''* are mostly above that straight line. This means the WR'' datasets tend to have more estimates that are significant than the rounded table. This result is similar to the results for estimates produced by the four-way contingency table in Chapter 4, and for the three-way tables in this chapter. What differs is that, for occasionally per_wr'' was lower than the straight line. For this two-way table, for some WR'' datasets, the percentage of estimates that are significant increases when a random rounded table is constructed. For the three- and four- way tables, this never occurred and the percentage of estimates that were significant always decreased with the model from the random produced by the randomly rounded table.

Next, we will compare each WR'' dataset with the RR'' dataset and see how many estimates do not change the hypothesis decision in the WR'' datasets, as compared with the RR'' dataset.

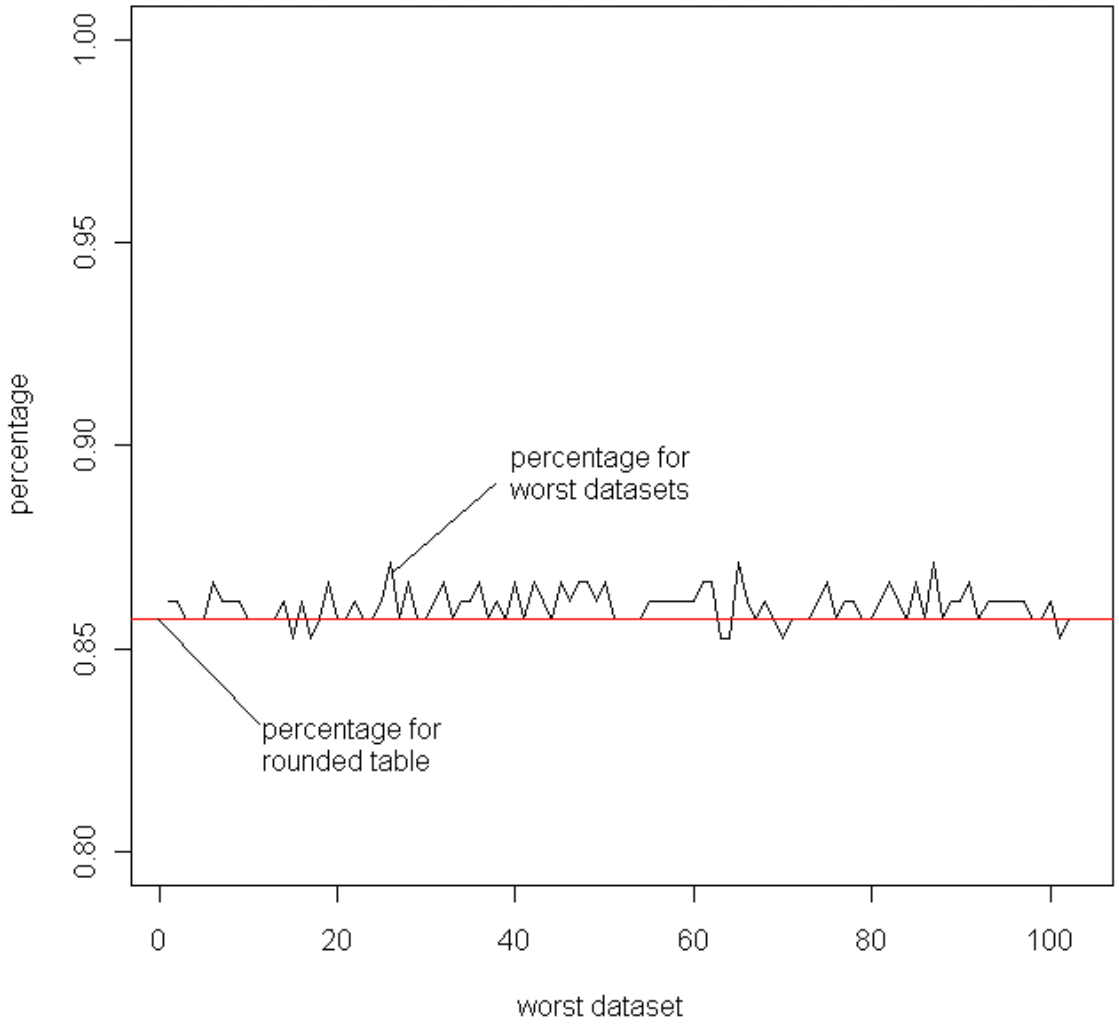

Figure 5.3 Percentage of estimates that are significant for each WR'' dataset.

Table 5.4 shows the number of estimates that changed the hypothesis decision in each WR'' dataset compared to the RR'' dataset. The first column shows the number of the WR'' dataset. The second column shows the number of estimates that are significant in WR'' dataset but not in the RR'' dataset. The third column shows the number of estimates are not significant in WR'' dataset but are significant in the RR'' dataset. The fourth column shows the number of estimates that are significant in both the WR'' dataset and the RR'' dataset. The fifth column shows the number of estimates are not significant in neither the WR'' dataset nor the RR'' dataset.

Table 5.4 Number of estimates that change the hypothesis decisions.

	Changed h0 decisions		Unchanged h0 decisions			Changed h0 decisions		Unchanged h0 decisions	
WR''	WR'' (p≤0.05) & RR'' (p>0.05)	WR'' (p>0.05) & RR'' (p≤0.05)	WR'' (p≤0.05) & RR'' (p≤0.05)	WR'' (p>0.05) & RR'' (p>0.05)	WR''	WR'' (p≤0.05) & RR'' (p>0.05)	WR'' (p>0.05) & RR'' (p≤0.05)	WR'' (p≤0.05) & RR'' (p≤0.05)	WR'' (p>0.05) & RR'' (p>0.05)
1	29	28	152	1	52	28	28	152	2
2	28	27	153	2	53	28	28	152	2
3	28	28	152	2	54	28	28	152	2
4	28	28	152	2	55	28	27	153	2
5	27	27	153	3	56	28	27	153	2
6	28	26	154	2	57	28	27	153	2
7	27	26	154	3	58	28	27	153	2
8	28	27	153	2	59	29	28	152	1
9	28	27	153	2	60	27	26	154	3
10	27	27	153	3	61	28	26	154	2
11	28	28	152	2	62	28	26	154	2
12	28	28	152	2	63	27	28	152	3
13	28	28	152	2	64	27	28	152	3
14	28	27	153	2	65	29	26	154	1
15	27	28	152	3	66	28	27	153	2
16	28	27	153	2	67	27	27	153	3
17	27	28	152	3	68	28	27	153	2
18	27	27	153	3	69	28	28	152	2
19	28	26	154	2	70	27	28	152	3
20	28	28	152	2	71	27	27	153	3
21	27	27	153	3	72	28	28	152	2
22	28	27	153	2	73	28	28	152	2
23	28	28	152	2	74	28	27	153	2
24	28	28	152	2	75	28	26	154	2
25	28	27	153	2	76	27	27	153	3
26	29	26	154	1	77	28	27	153	2
27	27	27	153	3	78	28	27	153	2
28	29	27	153	1	79	27	27	153	3
29	28	28	152	2	80	28	28	152	2
30	28	28	152	2	81	27	26	154	3
31	28	27	153	2	82	29	27	153	1
32	29	27	153	1	83	29	28	152	1
33	28	28	152	2	84	27	27	153	3
34	28	27	153	2	85	29	27	153	1
35	28	27	153	2	86	28	28	152	2
36	28	26	154	2	87	29	26	154	1
37	28	28	152	2	88	27	27	153	3
38	28	27	153	2	89	27	26	154	3
39	28	28	152	2	90	28	27	153	2

40	28	26	154	2	91	29	27	153	1
41	28	28	152	2	92	28	28	152	2
42	28	26	154	2	93	27	26	154	3
43	28	27	153	2	94	29	28	152	1
44	28	28	152	2	95	28	27	153	2
45	28	26	154	2	96	28	27	153	2
46	28	27	153	2	97	27	26	154	3
47	28	26	154	2	98	28	28	152	2
48	28	26	154	2	99	28	28	152	2
49	28	27	153	2	100	29	28	152	1
50	28	26	154	2	WRa''	28	27	152	3
51	28	28	152	2	WRb''	28	28	152	2

Table 5.4 shows that when the estimates are compared for significance, the number of estimates that are significant in WR'' but not in the RR'' dataset is slightly greater than the number of estimates are significant in the RR'' dataset but not in the WR'' datasets.

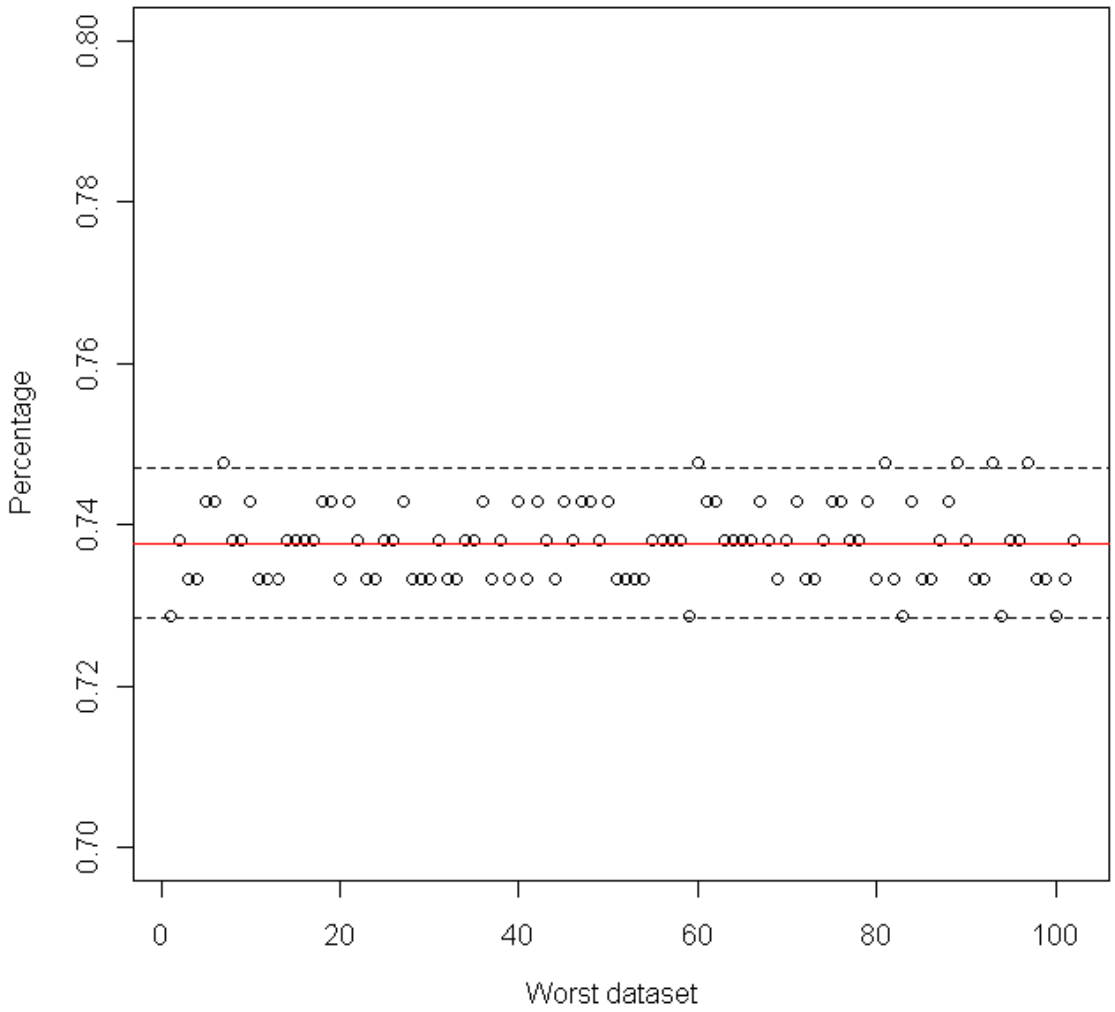

Figure 5.4 Percentage of estimate that do not change the hypothesis decision.

Figure 5.4 shows the percentage of estimates (out of the total number of estimates, e.g. (152+1)/210) that do not change the hypothesis decisions between each WR'' dataset and RR'' table. The percentage confidence interval at the 5% level is [0.73, 0.75]. Therefore, around 74% estimates do not change the hypothesis decision in each WR'' dataset, compared with the RR'' dataset. This result is very close to the result which was obtained for the four-way contingency table (Chapter 4).

5.3. Summary

The two-, three- and four-way contingency tables show similar results produced by the effect of random rounding. All minimally adequate models constructed from the random rounded table and from a range of possible original actual (parent) datasets had the same effects. However, what is a concern is that the model from the rounded dataset almost always had fewer estimates that are significant, compared with the actual datasets. With the three-, and four-way tables, all actual datasets, had more significant estimates than the rounded tables.

6. Reference

Christensen, R. (1990). *Log-linear models*. New York: Springer-Verlag.

Cox, L. H. (1981). Linear sensitivity measures in statistical disclosure control. *Journal of Statistical Planning and Inference, 5*(2), 153-164.

Cox, L. H. (1987). A Constructive Procedure for Unbiased Controlled Rounding. *Journal of the American Statistical Association, 82*, 520-524.

Crawley, M. (2005). *Statistics: an introduction using R*: Wiley.

Cuppen, M. (2000). Source Data Perturbation in Statistical Disclosure Control. *Statistics Netherlands*, 9-26.

Dobson, A. (2001). *An introduction to generalized linear models*: CRC Pr I Llc.

Everitt, B. (1992). *The analysis of contingency tables*: Chapman & Hall/CRC.

Federal committee on statistical methodology (2005). *Statistical policy working paper 22*. Washington.

Jeansonne, A. (2002). Loglinear Models Retrieved 6th, November,2009, from http://online.sfsu.edu/~efc/classes/biol710/loglinear/Log%20Linear%20Models.pdf

Kelly, J., Golden, B., Assad, A., & Baker, E. (1990). Controlled rounding of tabular data. *Operations Research*, 760-772.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models*: Chapman & Hall/CRC.

Ryan, M., & Penny, R. (1986). A problem associated with random rounding *The Newzealand Statistician, 21*, 7.

Ryan, M. P. (1981). Random rouding and chi-square analysis. *The Newzealand Statistician, 16*, 9.

Cochran, WG (1952). The Chi-squire Test of Goodness of Fit, Ann.Math.Statistics, 23.

Shlomo, N., & Young, C. (2006). Statistical Disclosure Control Methods Through a Risk-Utility Framework *Privacy in Statistical Databases* (Vol. 4302): Springer Berlin / Heidelberg.

Willenborg, L., & De Waal, T. (1996). Statistical Disclosure Control in Practice. Vol. 111. *Lecture Notes in Statistics, Springer-Verlag, New York*.

Appendix - Important R code

1. R code for 2x2 contingency table with rounding base 3:

a) R function "rounbase3" obtain the possible unrounded tables for 2x2 contingency table with rounding base 3.

```
> rounbase3<-function(paratable)
+ {
+ h<-array(0,c(2,2,249))
+ A<-array(0,c(2,2))
+ V1<-array(c((paratable[1]-2):(paratable[1]+2)))
+ V2<-array(c((paratable[2]-2):(paratable[2]+2)))
+ V3<-array(c((paratable[3]-2):(paratable[3]+2)))
+ V4<-array(c((paratable[4]-2):(paratable[4]+2)))
+ n1o<-sum(paratable[1]+paratable[3])
+ n2o<-sum(paratable[2]+paratable[4])
+ no1<-sum(paratable[1]+paratable[2])
+ no2<-sum(paratable[3]+paratable[4])
+ o=1
+ for(l in 1:5)
+ {
+    for(k in 1:5)
+    {
+       for(i in 1:5)
+       {
+            for(j in 1:5)
+            {
+                 V1o<-V1[i]+V3[j]
+                 V2o<-V2[l]+V4[k]
+                 Vo1<-V1[i]+V2[l]
+                 Vo2<-V3[j]+V4[k]
+
+                if((V1o<=(n1o+2)) & (V1o>=(n1o-2))& (V2o<=(n2o+2))& (V2o>=(n2o-2))& (Vo1<=(no1+2))&
+                          (Vo1>=(no1-2))& (Vo2<=(no2+2))& (Vo2>=(no2-2)))
+
+                   {
+
+                            A[1]<-V1[i]
+                            A[3]<-V3[j]
+                            A[2]<-V2[l]
+                            A[4]<-V4[k]
+                            o=1+
+                            o
+                            #print(o)
+                            #print(A)
+                            h[1:2,1:2,o]=A
+                    }
+             }
+       }
+
+    }
+ }
+ newh<-array(0,c(2,2,248))
+ newh[1:2,1:2,1:247]<-h[1:2,1:2,2:248]
+ #h
+ }
```

b) R function "chisq" to calculate the chi-square value of 2x2 contingency table.

```
> chisq<-function(table)
+ {
+   library(pgirmess)
+   sum1<-sum(valchisq(table),na.rm = FALSE)#same to table calculate
+   sum1
+ }
`
```

Note: R code chisq.test(table, correct=F) can get the same result as above chisq function. Crawley (2005) introduces (correct=F) means switch the correction off, the result will be same to the value we calculated by hand.

c) R function " chi_2way_b3" to calculate the chi-square value of 247 possible unrounded table together.

```
> chi_2way_b3<-function(baby_base3)
+ {
+ chi<-array(0,247)
+
+ for (p in 1:247)
+ {
+   chi[p]<-chisq(baby_base3[1:2,1:2,p])
+ }
+ chi
+ }
```

d) R function" percent_base3" to compute percentage that is no. of matched tables out of total no. of possible tables in n^{th} decimal place.

```
> percent_base3<-function(parm_ll)
+ {
+
+ percent<-array(0,c(1,6,100))
+ for(i in 1:100)
+ {
+ bb=rounbase3(parm_ll[1:2,1:2,i])
+ parent_chi=chisq(parm_ll[1:2,1:2,i])
+ bb_chi=chi_2way_b3(bb)
+
+ percen=array(0,6)
+ for(h in 1:6)
+ {
+ p_chi<-round(parent_chi,digits=h)
+ b_chi<-round(bb_chi,digits=h)
+ or<-array(0,247)
+ r=1
+ for(w in 1:247)
+ {
+ if(w!=124)#it alway have 124th baby =itself, so we don't look at this one.
+ {
+  if(b_chi[w]==p_chi) ##one part table have 247baby. w is the nth baby of this part
+    { #print("############")
+      #print("th parmets have the match baby on poistion")
+      #print(w)
+       r=1+
+       r
+      or[r]=w
+
+     }
+
+ }
+ }
+ or[or==0]=NA
+ or1=or[complete.cases(or)]#exclude the NA value
+ l=length(or1)
+ percen[h]=l/247
+ if(percen[h]==0){break}
+ #print(h)
+ }
+ percent[1:1,1:6,i]<-percen
+ }
+ percent
+
+ }
```

2. R code for 2x2 contingency table with rounding base 5:

a) R function "rounbase5" obtain the possible unrounded tables for 2x2 contingency table with rounding base 5.

```
> rounbase5<-function(paratable)
+ {
+ h<-array(0,c(2,2,2503))
+ A<-array(0,c(2,2))
+ V1<-array(c((paratable[1]-4):(paratable[1]+4)))
+ V2<-array(c((paratable[2]-4):(paratable[2]+4)))
+ V3<-array(c((paratable[3]-4):(paratable[3]+4)))
+ V4<-array(c((paratable[4]-4):(paratable[4]+4)))
+ n1o<-sum(paratable[1]+paratable[3])
+ n2o<-sum(paratable[2]+paratable[4])
+ no1<-sum(paratable[1]+paratable[2])
+ no2<-sum(paratable[3]+paratable[4])
+ o=1
+ for(l in 1:9)
+ {
+    for(k in 1:9)
+    {
+       for(i in 1:9)
+       {
+            for(j in 1:9)
+            {
+                 V1o<-V1[i]+V3[j]
+                 V2o<-V2[l]+V4[k]
+                 Vo1<-V1[i]+V2[l]
+                 Vo2<-V3[j]+V4[k]
+
+                if((V1o<=(n1o+4)) & (V1o>=(n1o-4))& (V2o<=(n2o+4))& (V2o>=(n2o-4))& (Vo1<=(no1+4))&
+                      (Vo1>=(no1-4))& (Vo2<=(no2+4))& (Vo2>=(no2-4)))
+
+                   {
+
+                           A[1]<-V1[i]
+                           A[3]<-V3[j]
+                           A[2]<-V2[l]
+                           A[4]<-V4[k]
+                           o=1+
+                           o
+                           #print(o)
+                           #print(A)
+                           h[1:2,1:2,o]=A
+                    }
+
+
+              }
+       }
+
+    }
+ }
+ newh<-array(0,c(2,2,2501))
+ newh[1:2,1:2,1:2501]<-h[1:2,1:2,2:2502]
+ #h
+ }
```

b) R function " chi_2way_b5" to calculate the chi-square value of 2501 possible unrounded table together.

```
> chi_2way_b5<-function(baby_base5)
+ {
+ chi1<-array(0,2501)
+
+ for (p in 1:2501)
+ {
+   chi1[p]<-chisq(baby_base5[1:2,1:2,p])
+ }
+ chi1
+ }
```

c) R function" percent_base3" to compute percentage that is no. of matched tables out of total no. of possible table in n^{th} decimal place.

```
> percent_base5<-function(parm_ll)
+ {
+
+ percent<-array(0,c(1,6,100))
+ for(i in 1:100)
+ {
+ bb=rounbase5(parm_ll[1:2,1:2,i])
+ parent_chi=chisq(parm_ll[1:2,1:2,i])
+ bb_chi=chi_2way_b5(bb)
+
+ percen=array(0,6)
+ for(h in 1:6)
+ {
+ p_chi<-round(parent_chi,digits=h)
+ b_chi<-round(bb_chi,digits=h)
+ or<-array(0,2501)
+ r=1
+ for(w in 1:2501)
+ {
+ if(w!=1251)#it alway have 124th baby =itself, so we don't look at this one.
+ {
+  if(b_chi[w]==p_chi) ##one part table have 247baby. w is the nth baby of this part
+    { #print("############")
+      #print("th parmets have the match baby on poistion")
+      #print(w)
+       r=1+
+       r
+      or[r]=w
+
+     }
+
+ }
+ }
+ or[or==0]=NA
+ or1=or[complete.cases(or)]#exclude the NA value
+ l=length(or1)
+ percen[h]=l/2501
+ if(percen[h]==0){break}
+ #print(h)
+ }
+ percent[1:1,1:6,i]<-percen
+ }
+ percent
+
+ }
```

3. R code to calculate the DP.

Below codes obtain the DP for 2x2 contingency table with rounding base 3. And the two-way table with rounding base 5, three-way table with rounding base 3 and 5, are following the similar idea**.**

```
> c1<-array(0,100) #null array to store 100 parents're X^2
> chi1<-array(0,c(1,247,100))#null array to store 100parents're babies' re X^2
> pp1_2way<-array(0,c(2,2,247))#null array tempory to store nth parent's 247 babies.
> dee3<-array(0,c(1,247,100)) #null array to store 100parents're babies' re X^2 at nth dp
> deee3<-array(0,247)#null array tempory to store nth parent's 247 babies're X^2 at nth dp.
> numP_dp_2wayb3<-array(0,100)
> r=1
>
> for(l in 1:100)
+ {
+ c1[l]<-chisq(parm100[1:2,1:2,l])#calculate parmt X^2
+ pp1_2way<-rounbase3(parm100[1:2,1:2,l])#one table have 247babys,
+ #pp1_2way store 100partmets 's babies.(each parmet have 247 babies)
+ chi1[1:1,1:247,l]<-chi_2way_b3(pp1_2way) #100parment's babies' X^2.
+
+ #######parent's X^2 in 4dp##############
+
+ c<-round(c1[l],digits=0)#partent round
+
+ ####babies(parent round at base 3) in 4dp###
+
+ dee3[1:1,1:247,l]<-round(chi1[1:1,1:247,l],digits=1)
+ deee3<-dee3[1:1,1:247,l]  # deee3 store nth parmet's 247baby's X^2
+
+ #######compare parent and itself's babies at nth dp, to find same X^2#######
+
+ for(w in 1:247)
+ {
+ if(w!=124)#it alway have 124th baby =itself, so we don't look at this one.
+ {
+  if(deee3[w]==c) ##one part table have 625baby. w is the nth baby of this part
+    { print("############")
+      print(l)
+      print("th parmets have the match baby on poistion")
+      print(w)
+       r=1+
+       r
+      numP_dp_2wayb3[r]=l
+
+        break
+     }
+
+ }
+ })
> numP_dp_2wayb3
```

4. R code to calculate the number of estimates are not significant.

Below codes obtain the number of estimates are not significant for original rounded age by sex table. Other tables obtain the number of estimates that are not significant are following similar idea.

```
> pva_round<-summary(glm5r)$coef[,4]# showed p-value
> write.csv(pva_round,file="p_round.csv")
> ##############
> p_round<-read.csv('p_round.csv')
> names(p round)
> attach(p_round)
> p_round_name=p_round$X
> pvalue_round=p_round$x
> pvalue_round[2]
> for(i in 1:1470)
+ {
+ if(pvalue_round[i]<=0.05){pvalue_round[i]=0}
+ else{pvalue_round[i]=1}
+ }
> pvalue_round
> sum(pvalue_round)
```

5. R code to calculate the number of estimates that have a change in hypothesis decision between the rounded table and worst tables.

Below codes obtain the number of estimates that have a change in the hypothesis decision for age by sex table. Other tables obtain the number of estimates that have a change in hypothesis decision, are following similar idea.

```
> per_deci=function(pvalue_worse)
+ {
+ h0=array(0,1470)  #  stroe the h0 desicision, #change the desicision and WR sig, but rounded not
+ h1=array(0,1470)  ##change the desicision and WR not sig, but rounded are
+ h2=array(0,1470)  ## haven't change the desicision,and both significant
+ h3=array(0,1470)  ## haven't change the desicision,and both non-significant
+
+ for(i in 1:1470)
+ {
+ if(pvalue_round[i]!= pvalue_worse[i] & pvalue_worse[i]==0 & pvalue_round[i]!= 0)
+ #WR sig, but rounded not
+ {h0[i]=1}#change the desicision and WR sig, but rounded not
+
+ if(pvalue_round[i]!= pvalue_worse[i] & pvalue_worse[i]!=0 & pvalue_round[i]== 0)
+ #WR not sig, but rounded are
+ {h1[i]=1}#change the desicision and WR not sig, but rounded are
+
+ if(pvalue_round[i]== 0 & pvalue_worse[i]==0){h2[i]=1}
+ # haven't change the desicision,and both significant
+
+ if(pvalue_round[i]== 1 & pvalue_worse[i]==1){h3[i]=1}
+ # haven't change the desicision,and both non-significant
+ }
+
+ WRsig_roudednot=sum(h0)
+ WRnot_roudedsig=sum(h1)
+ WR_RR_unchange_sig=sum(h2)
+ WR_RR_unchange_notsig=sum(h3)
+ WR_RR_change=sum(h0)+sum(h1)
+ WR_RR_unchange=sum(h2)+sum(h3)
+ print('No.of estimates are significant in WR but are non-significant in RR:')
+ print(WRsig_roudednot)
+ print('No.of estimates are non-significant in WR but are significant in RR:')
+ print(WRnot_roudedsig)
+ print('No.of estimates are changed the dicision:')
+ print(WR_RR_change)
+ print('No.of estimates are not changed the dicision and both significant in WR and RR:')
+ print(WR_RR_unchange_sig)
+ print('No.of estimates are not changed the dicision and both non-significant in WR and RR:')
+ print(WR_RR_unchange_notsig)
+ print('No.of estimates are not changed the dicision:')
+ print(WR_RR_unchange)
+ }
```