The undersigned certify that we have read a thesis, entitled “On measuring agreement for categorical data” submitted to the Graduate School by Tang Pik Ha ( ) in partial fulfillment of the requirements for the degree of Master of Philosophy in Statistics. We recommend that it be accepted.

Prof. T. S. Lau
Supervisor

Prof. S. H. Cheung

Prof. W. Y. Poon

Prof. S. K. Tse
External Examiner
DECLARATION

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institution of learning.
I would like to express my gratitude to my supervisor, Prof. Tai-Shing Lau, for his guidance during the period of this research program. I would also like to thank for the entire staff of the Department of Statistics. Furthermore, I would like to take this opportunity to give thanks to my family, my classmates and my friends for their encouragement and support.
ABSTRACT

Cohen's kappa coefficient is commonly used to measure agreement between two observers for nominal data. This thesis is concerned with the measure of agreement with kappa for $m \times m$ contingency table in modelling approach. A model for agreement assuming marginal homogeneity for $2 \times 2$ table has been proposed by Bloch & Kraemer (1989), which yields an estimator of kappa. Agresti (1989) proposed an agreement model for $m \times m$ contingency table. In this model, the Maximum Likelihood (ML) estimator of kappa is derived. In this thesis, the small to moderate sample performance of the ML estimate of the kappa is investigated. The asymptotic variance of the ML estimate is also studied. A simulation study is conducted for $3 \times 3$ table. The results show that the asymptotic theory for the estimates holds with moderate samples.
摘要

在名義數據，Cohen 的卡巴系數通常被採用為量度兩位觀察者之間的一致性。這篇論文有關在 $m \times m$ 列聯表中，以模型為基礎的卡巴量度一致性。在 $2 \times 2$ 列聯表中，Bloch&Kraemer(1989)建議一個假設邊際同質性的一致模型，產生了卡巴的估計值。當伸延至 $m \times m$ 列聯表的時候，Agresti(1989)建議另一個一致模型，找出卡巴的最大似然估計。本論文會研究卡巴的最大似然估計值於小至中等樣本的特性，並且研究它的漸近極方差。因此，我們進行了一個 $3 \times 3$ 列聯表的模擬研究。結果顯示當中等樣本時，漸近理論成立。
# Contents

1 Introduction ........................................... 1  
   1.1 Agreement analysis ................................. 2  
   1.2 Outline of the thesis ............................... 6  

2 Review ................................................ 8  
   2.1 Chance-corrected measures ......................... 8  
   2.2 Statistical Modelling Approach .................... 15  

3 Model-based kappa .................................... 17  
   3.1 An agreement model with kappa as parameter ....... 17  
   3.2 Parameter Estimation ............................... 21  
   3.3 Asymptotic variance-covariance matrix .............. 25  
      3.3.1 Fisher Information ............................. 25  
      3.3.2 Computational detail ........................... 27  
   3.4 Illustrative Example ............................... 30  

4 Simulation Study .................................... 33  
   4.1 Design ............................................. 33
4.2 Results .............................................. 37
4.3 Discussion ........................................... 40

5 Conclusion ............................................ 42

Tables .................................................. 44

Figures .................................................. 49

Bibliography ......................................... 51
Chapter 1

Introduction

Data analysis is applied in various aspects, such as medical and psychological researches, etc. There are two main kinds of data: nonmetric (qualitative) and metric (quantitative). Nonmetric data are characteristics, attributes, or any categorical properties that can be used to identify a subject. These nonmetric data can be measured by nominal or ordinal scales (Hair et al., 1995). Categorical variable, whose levels do not have a natural ordering, is nominal (Agresti, 1990). Examples of nominal variables are

- religious affiliation (Catholic, Protestant, Jewish),
- voting behaviour (Democrat, Independent, Republican),
- marital status (single, married, widowed, divorced), and
- gender (female, male).

In working with these variables, the analyst may assign numbers to each category. However, these numbers only represent the corresponding categories
but do not indicate any amount of characteristics. The distinct levels of such variables differ in quality, not quantity (Agresti, 1984).

1.1 Agreement analysis

Many statistical research problems involve the measurement of agreement between two observers who classify responses among $m$ nominal categories rather than the association between them. Actually, the extent of agreement is more interesting than the extent of association. For example, clinicians may be asked to classify patients by their abnormality using a discrete rating system. So, one may concern how agree between the clinicians. Different observers can make different responses for a given subject. This is because several types of measurement error can occur, especially when the response variable has subjective rating scale. This is often the case for categorical responses. Discrepancies between ratings may be due to classification errors by the observers and the categories which do not having precise definition. Different observers may have different perceptions about the meaning of the categories. Although there is a common perception, measurement variability can happen. For example, an observer can show variability in repeated observations to a given subject. Issues arising in interobserver agreement analysis also apply to intra-observer agreement analysis (Agresti, 1992). The problem of observer agreement is particularly important in the fields of medical, psychological, educational and sociological applications.
Bishop et al. (1975) defined the distinction between 'agreement' and 'association' for nominal data as follows:

- When two responses are said to be agreed, they must fall into the identical category.

- While two responses are associated perfectly, we only require that we can predict the category of one response from the category of the other response.

A contingency table may be used to show high agreement and low association, or low agreement and high association. Thus, perfect association does not generally imply perfect agreement. Table 1.1 below illustrates a case in which two observers are strongly associated but no agreement takes place (Light, 1971).

Table 1.1 Hypothetical example of perfect association but zero agreement

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>c₁</th>
<th>c₂</th>
<th>c₃</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>c₁</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td>c₂</td>
<td>0</td>
<td>0</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>c₃</td>
<td>50</td>
<td>0</td>
<td>0</td>
<td>50</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50</td>
<td>50</td>
<td>50</td>
<td>150</td>
</tr>
</tbody>
</table>

From Table 1.1, there is strong association between the two observers in a predictive sense. Provided with one observer's responses, the other observer's responses can be predicted. For instance, Observer 1 classifies the subjects into c₁, but Observer 2 classifies them into c₂; when Observer 1 classifies the subjects into c₂,
Observer 2 classifies them into C3. It shows a perfect association between the two observers. Hence, we can predict the response of the second observer from the response of the first observer. However, based on the main diagonal that indicates agreement, one may conclude that this example shows no agreement.

To generalize our examination to a more general setting and further explain the terminology 'agreement' for nominal scale response, Table 1.2 displayed a $m \times m$ contingency table for measuring agreement that categorizes the responses of two observers to the same set of $n$ subjects (Bishop et al., 1975; Hubert, 1977).

### Table 1.2 Notation for two observer agreement

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$c_j$</th>
<th>$c_m$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1$</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
<td>$n_{1j}$</td>
<td>$n_{1m}$</td>
<td>$n_1$.</td>
</tr>
<tr>
<td>$c_i$</td>
<td>$n_{i1}$</td>
<td>$n_{i2}$</td>
<td>$n_{ij}$</td>
<td>$n_{im}$</td>
<td>$n_i$.</td>
</tr>
<tr>
<td>$c_m$</td>
<td>$n_{m1}$</td>
<td>$n_{m2}$</td>
<td>$n_{mj}$</td>
<td>$n_{mm}$</td>
<td>$n_m$.</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$n_1$</td>
<td>$n_2$</td>
<td>$n_j$</td>
<td>$n_m$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

Assumption for Table 1.2 can be summarized as follows: (1) Two observers independently classify each of $n$ subjects into one of the $m$ nominal (i.e., unordered) and mutually exclusive categories $c_1, c_2, ..., c_m$. (2) $n_{ij}$ is the number of subjects assigned in category $c_i$ by Observer 1 and in category $c_j$ by Observer 2. (3) $n_i$ is the total number of responses in row $i$; similarly, $n_j$ is the total number of
responses in column $j$. The total number of responses given by each observer is represented by $n$. (4) The number of rows is equal to the number of columns. (5) As described by Goodman and Kruskal (1954), the categories in the rows must appear in the same order as the categories in the columns. Hence, two rows cannot be permuted unless the corresponding columns are similarly permuted. There is a similar assumption given by Bloch and Kraemer (1989) that using an index of agreement is appropriate if the responses of the $m$ ($m \geq 2$) ratings on a subject are being assessed as potentially interchangeable.

Thus, a natural question of how agreement should be measured arises. Since the agreement between the two observers are defined by those entries along the main diagonal of the contingency table, the natural measure of agreement is

$$p_o = \frac{\sum_{i=1}^{m} n_{ii}}{n}$$

(Bishop et al., 1975). It could be considered as raw form of an agreement statistic. Perfect agreement occurs when $p_o=1$.

Berry and Mielke (1988) pointed out that a universal measure of agreement should, as a minimum, include seven basic attributes. One of the attributes is corrected for chance, which states that any agreement coefficient should reflect the amount of agreement in excess of what would be expected by chance. In addition, several authors (Brennan and Prediger, 1981; Cicchetti et al., 1985; Conger, 1985 and Krippendorff, 1970) have supported chance-corrected measures of agreement. As a consequence, my study just focuses on this attribute about the measure of agreement.
1.2 Outline of the thesis

This thesis is organised as follows. In Chapter 2, some measures of interobserver agreement for nominal scales, which incorporate a correction for chance agreement are reviewed. One of the pioneers to this problem is Scott (1955) who developed a statistic ($\pi_s$) for reporting the extent of interobserver agreement for nominal scale under the assumption of marginal homogeneity. Cohen (1960) introduced the kappa coefficient ($\kappa_C$) which is the proportion of agreement between two observers after chance agreement is removed. The formulae for calculating the Scott's coefficient and Cohen's kappa are to be presented in the chapter. Regardless of the appealing feature of kappa coefficient, which corrects chance, Bloch and Kraemer (1989) and Kraemer (1992) commented that Cohen's kappa is defined on an ad hoc approach. Therefore, Bloch and Kraemer (1989) proposed a model for agreement with marginal homogeneity to find the maximum likelihood estimator of kappa for 2 × 2 table.

To generalize the agreement model, recent development has been shifted to the dimension of $m \times m$ table. Agresti (1989) expanded the context of agreement model to more than two categories of responses. This study will investigate an agreement model for cell probabilities that contain kappa measure of agreement as a model parameter for $m \times m$ table.

In Chapter 3, a simple model for cell probabilities that contain the kappa measure of agreement as model parameter will be described for $m \times m$ table. The maximum likelihood estimators of the kappa (model-based kappa, $\kappa_M$) and mar-
ginal probabilities and their asymptotic variance-covariance matrices are derived.

In Chapter 4, the performance of model-based kappa is investigated through Monte Carlo studies under a situation where the number of categories is three. It is interesting to evaluate the small-sample estimate of model-based kappa in the agreement model. The results of the Monte Carlo studies are summarized and discussed. In Chapter 5, a brief conclusion is made.
Chapter 2

Review

In this chapter, a review of two traditional measures of agreement for nominal data in the two-way contingency table, the Scott's statistic ($\pi_S$) and Cohen's kappa ($\kappa_C$), is presented. A recommendation is given for using the agreement index. The work of Bloch and Kraemer's (1989), which proposed a model for agreement to obtain the MLE of kappa for $2 \times 2$ table, is then reviewed briefly.

2.1 Chance-corrected measures

Most early statistical researches dealing with observer's agreement focus on indexes for the measurement of agreement. The objective of these studies is to develop a statistic that indicated the degree of agreement between two observers.

The most commonly used measures of interobserver agreement concerning data of nominal scales incorporate with a correction for chance agreement. However, the definition of chance agreement is not the same for all coefficients. Two
chance-corrected coefficients are proposed by Scott's (1955) and Cohen's (1960), termed $\pi_S$ and $\kappa_C$ respectively. They are used as a descriptive statistic indicating degree of beyond-chance agreement between two observers. Before Cohen (1960) named the agreement index as kappa statistic without assuming identical marginals, Scott (1955) had developed a statistic similar to kappa involving the assumption of marginal homogeneity. For these two measures, independence between observers is assumed in deriving the proportion of agreement expected by chance.

There are five typical assumptions of the coefficient of agreement (Cohen, 1960, Sevensson, 1993, Scott, 1955). First, two observers independently classify each of $n$ subjects into one of $m$ nominal categories. Second, the categories of the nominal scale are mutually exclusive and exhaustive. Third, the subjects to be judged are independent. Fourth, the chance expected agreement is determined by the joint probabilities of the marginals. Fifth, the observers are equally competent to make judgment.

To calculate Scott and Cohen's coefficient, sample proportions $p_{ij}$ are used for each cell entry instead of frequency responses by the observers in classification table displayed on a $m \times m$ table (see Table 1.2), where $p_i = \sum_{j=1}^{m} p_{ij}$ and $p_j = \sum_{i=1}^{m} p_{ij}$.

The empirical expression for the chance-corrected coefficient can be written as

$$\frac{p_0 - p_e}{1 - p_e},$$
where $p_e$ is defined differently by Scott and Cohen.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>$p_o$</th>
<th>$p_e$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_S$(Scott, 1955)</td>
<td>$\sum_{i=1}^{m} p_{ii}$</td>
<td>$\sum_{i=1}^{m} \left( \frac{p_i + p_{i\cdot}}{2} \right)^2$</td>
</tr>
<tr>
<td>$\kappa_C$(Cohen, 1960)</td>
<td>$\sum_{i=1}^{m} p_{ii}$</td>
<td>$\sum_{i=1}^{m} p_i.p_{i\cdot}$</td>
</tr>
</tbody>
</table>

It can be denoted that the observed proportion of observations on which two observers agree as $p_o$, and the proportion of observations for which agreement is expected by chance as $p_e$. Thus, $p_o - p_e$ is the proportion of agreement beyond what is expected by chance. $1 - p_e$ is the maximum possible proportion of agreement beyond what is expected by chance. Then, the resulting coefficient by Scott or Cohen is the proportion of agreement between the two observers after chance agreement is removed. A zero value of coefficient indicates that the observed agreement is likely to be chance expected. Coefficient equals 1 means a total interobserver agreement and a negative value means observed agreement is less than chance agreement (Berry and Mielke, 2001; Fleiss, 1981).

The chance correction for $\kappa_C$ and $\pi_S$ is based on the observed marginal distributions for each observer. When the observers have the same observed marginals, then $\pi_S$ equals to $\kappa_C$.

It is worthwhile to note that, as an index of agreement, Scott's coefficient assumes marginal homogeneity while Cohen's one does not. Some researchers may be interested to know whether the marginal homogeneity is necessary. Note that when two observers are unbiased relative to each other, they have identical marginal distribution. Conceptually, perfect agreement between two observers implies
marginal homogeneity. To achieve perfect agreement, Agresti (1997) stated that all observations would need to fall into the main diagonal. As a result, both observers have the same number of observations in each category. Then, the row marginal percentages equal to the corresponding column marginal percentages.

According to Agresti (1992), the marginal distributions \( \{ \pi_i = \sum_j \pi_{ij}, i = 1, \ldots, m \} \) and \( \{ \pi_j = \sum_i \pi_{ij}, j = 1, \ldots, m \} \) describe how observers separately allocate subjects into the response categories, where \( \pi_{ij} \) is the cell probability in the two-way contingency table. Discrepancies between these marginal distributions are referred to the bias of one observer relative to another. Lack of bias means that \( \pi_i = \pi_j \) for all \( i \). Bias decreases as the marginal distributions become nearly equivalent. The term 'bias' here is defined as marginal heterogeneity for two observers. Therefore, the marginal homogeneity is a consequence of agreement between the observers.

Examination of the degree of the marginal distributions of the observers is necessary before using the index of agreement. Instead of ignoring marginal disagreement, Zwick (1988) suggested that researchers should study it to decide whether it reflects important observer differences or only random error. The marginal homogeneity can be tested using McNemar's test (1947) for 2x2 table. For 3x3 or 4x4 table, a test statistic by Fleiss and Everitt (1971) can be used. For higher dimension of square contingency table, Stuart's test (1955) should be employed.

As mentioned before, Cohen’s kappa does not assumed marginal homogeneity. The following examples show a larger value of Cohen’s kappa with worse mar-
ginal total agreement while a smaller value of kappa with better marginal total agreement.

In the first example, Tables 2.2 and 2.3, extracted from Feinstein and Cicchetti (1990), illustrate the value of $\kappa_C$ would be higher even if the observer's bias is more severe in $2 \times 2$ table.

Table 2.2  2×2 table with marginal homogeneity

<table>
<thead>
<tr>
<th>Observer A</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>45</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>No</td>
<td>25</td>
<td>15</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>30</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 2.3  2×2 table without marginal homogeneity

<table>
<thead>
<tr>
<th>Observer A</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observer B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>25</td>
<td>35</td>
<td>60</td>
</tr>
<tr>
<td>No</td>
<td>5</td>
<td>35</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
<td>70</td>
<td>100</td>
</tr>
</tbody>
</table>

Tables 2.2 and 2.3 have the same proportion of agreement $p_o=0.6$. Then, the marginal homogeneity of the table is tested with McNemar's statistic, $X^2 = \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$.

For Table 2.2, $X^2 = 2.5$ ($P$-value=0.1138) is smaller than critical value $X^2_{0.05}=3.84$ ($>0.05$). As a result, the null hypothesis of equal marginal probabilities is not rejected at $\alpha=0.05$. For Table 2.3, $X^2 = 22.5$ ($P$-value=0) so that the obtained
value is significant and the hypothesis of equal marginal probabilities is rejected. Although the marginal total of Table 2.2 would achieve better agreement than that of Table 2.3, the value of $\kappa_C$ is higher in Table 2.3 ($\kappa_C=0.26$) than in Table 2.2 ($\kappa_C=0.13$).

The second example is used to illustrate the value of $\kappa_C$ would be higher even if the observer's bias is more severe in 3×3 contingency table. Table 2.4, adopted from Agresti (1997) and Table 2.5, a hypothetical table, are used.

**Table 2.4 3×3 table with marginal homogeneity**

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>c₁</th>
<th>c₂</th>
<th>c₃</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>c₁</td>
<td>24</td>
<td>8</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>c₂</td>
<td>8</td>
<td>13</td>
<td>11</td>
<td>32</td>
</tr>
<tr>
<td>c₃</td>
<td>10</td>
<td>9</td>
<td>64</td>
<td>83</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>30</td>
<td>88</td>
<td>160</td>
</tr>
</tbody>
</table>

**Table 2.5 3×3 table without marginal homogeneity**

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>c₁</th>
<th>c₂</th>
<th>c₃</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>c₁</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>c₂</td>
<td>2</td>
<td>13</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>c₃</td>
<td>0</td>
<td>57</td>
<td>48</td>
<td>105</td>
</tr>
<tr>
<td>Total</td>
<td>42</td>
<td>70</td>
<td>48</td>
<td>160</td>
</tr>
</tbody>
</table>

Fleiss and Everitt (1971) modified a Stuart-Maxwell Statistic to test marginal
homogeneity in $3 \times 3$ table. The statistic becomes

$$X_3^2 = \frac{\bar{n}_{23}d_1^2 + \bar{n}_{13}d_2^2 + \bar{n}_{12}d_3^2}{2(\bar{n}_{12}\bar{n}_{23} + \bar{n}_{13}\bar{n}_{13} + \bar{n}_{13}\bar{n}_{23})}, \quad (2.1)$$

where $d_i = n_{ii} - \bar{n}_{ii}$ and $\bar{n}_{ij} = (n_{ij} + n_{ji})/2$. For Table 2.4, $X_3^2=0.5854$ ($P$-value=0.4442) is smaller than the critical value $X_{2,0.05}^2=5.99 (>0.05)$, the null hypothesis of equal marginal probabilities is not rejected at $\alpha=0.05$. For Table 2.5, $X_3^2=59$ ($P$-value=0), gives strong evidence of marginal heterogeneity. Similarly, Tables 2.4 and 2.5 above have the same proportion of agreement $p_o=0.63$. Although the marginal total of Table 2.4 would suggest better agreement than that of Table 2.5, the value of $\kappa_C$ is higher in Table 2.5 ($\kappa_C=0.47$) than in Table 2.4 ($\kappa_C=0.39$).

By these two examples, it seems unreasonable to obtain larger kappa value when greater observer's bias presents in Table 2.3 and Table 2.5. Therefore, we recommend that kappa should be used as a measure of agreement when marginal homogeneity is achieved.

Furthermore, Zwick (1988) proposed two stages in the assessment of observer agreement: first, the investigation of marginal homogeneity; second, Scott's $\pi_S$ as a measure of chance-corrected agreement if marginal homogeneity holds. The rationale of this approach is that, one should stop if the hypothesis of marginal homogeneity is rejected. The degree of disagreement between observers can be expressed as the discrepancies between their marginal distributions. However, if marginal differences are small, Scott's coefficient can be applied and the value of $\kappa_C$ will be closed to that of Scott's $\pi_S$ at anytime.
2.2 Statistical Modelling Approach

As stated before, Bloch & Kraemer (1989) commented that much of the work done on the statistic of kappa are developed in an intuitive, largely ad hoc, approach with kappa as a descriptive statistic. They proposed a model for agreement for $2 \times 2$ table under marginal homogeneity ($\pi_i = \pi_{ij} = \pi_j$, $i, j = 1, 2$) or when the observers are considered unbiased relative to each other. Hence, the model with marginal homogeneity can be written as Table 2.6:

Table 2.6: Model for agreement with marginal homogeneity in $2 \times 2$ table

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>$c_1$</th>
<th>$c_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_1$</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
</tr>
<tr>
<td>$c_2$</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
</tr>
<tr>
<td></td>
<td>$\pi_1$</td>
<td>$\pi_2$</td>
</tr>
</tbody>
</table>

where the cell probabilities $\{\pi_{ij}\}$ in Table 2.6 express as the function of marginal probabilities $\pi_i$ and kappa $\kappa$,

$$\pi_{11} = \pi_1^2 + \kappa \pi_1 (1 - \pi_1),$$

$$\pi_{22} = \pi_2^2 + \kappa \pi_2 (1 - \pi_2),$$

and

$$\pi_{12} = \pi_{21} = \pi_1 \pi_2 (1 - \kappa).$$

Bloch and Kraemer (1989) derived the maximum likelihood estimate for this kappa coefficient. Suppose that in a sample of $n$ subjects, the observed frequencies
of responses are shown in Table 1.2 with \( m=2 \). To obtain the MLE, multinomial distribution is used. The likelihood function is

\[
\frac{n!}{n_{11}!n_{12}!n_{21}!n_{22}!} \left[ \pi_1^2 + \kappa \pi_1 (1 - \pi_1) \right] ^{n_{11}} \left[ \pi_2 (1 - \kappa) \right] ^{n_{12} + n_{21}} \left[ \pi_2^2 + \kappa \pi_2 (1 - \pi_2) \right] ^{n_{22}}.
\]

Setting the log-likelihood function equal to zero, the maximum likelihood estimators (MLEs), \( \hat{\pi}_1 \) and \( \hat{\kappa}_I \) for \( \pi_1 \) and \( \kappa \) are given respectively as

\[
\hat{\pi}_1 = \frac{2n_{11} + n_{12} + n_{21}}{2n},
\]

\[
\hat{\kappa}_I = \frac{4(n_{11}n_{22} - n_{12}n_{21}) - (n_{12} - n_{21})^2}{(2n_{11} + n_{12} + n_{21})(2n_{22} + n_{12} + n_{21})} = \frac{p_0 - p_e}{1 - p_e},
\]

where

\[
p_0 = \frac{n_{11} + n_{22}}{n},
\]

\[
p_e = \left( \frac{n_1 + n_{1.}}{2n} \right)^2 + \left( \frac{n_2 + n_{2.}}{2n} \right)^2.
\]

So, \( \kappa_I \) has the form of chance corrected agreement. The estimator of \( \kappa_I \) equals to the coefficient proposed by Scott (1955) as a measure of agreement between two observers when they have the same marginal distribution. Meanwhile, a \( \sin^{-1} \) transformation was derived to stabilize the variance of \( \hat{\kappa}_I \) by Bloch & Kraemer (1989). With this transformation, calculations of confidence intervals are eased. In conclusion, Bloch and Kraemer (1989) recommended that kappa is used to measure agreement when the assumption of equal marginal distribution is desirable.
Chapter 3

Model-based kappa

In this chapter, an agreement model for $m \times m$ table is described. In this model, the maximum likelihood (ML) estimators of kappa and marginal probabilities are obtained in the modelling approach. And, the asymptotic variance-covariance matrix for the estimates is derived. Besides, an illustrative example is given.

3.1 An agreement model with kappa as parameter

Bloch and Kraemer (1989) have developed the $\kappa_I$ in case of $2 \times 2$ table. It is interesting to measure the agreement in multicategory level. In this section, a simple model for cell probabilities where kappa is the parameter in $m \times m$ table is introduced.

Agresti (1989) exhibited a simple quasi-symmetry model below in which kappa
includes all relevant information about the structure of agreement and disagree-
ment in the model-building approach. A simple model for cell probabilities in
which kappa is the parameter is then exhibited. Therefore, a combined approach
of model-building and summary descriptive measure is possible.

The 'quasi-symmetry model' has multiplicative form (see Agresti, 1990, P.355):

\[ \pi_{ij} = a_i b_j c_{ij} \quad \text{where} \quad c_{ij} = c_{ji} \quad \text{all} \quad i, j = 1, \ldots, m, \]

and all parameters are positive. An agreement model is derived from the quasi-
symmetry model by imposing constraints of marginal homogeneity \( \pi_i = \pi_i = \pi_i, \)
a_1 = \pi_i, b_j = \pi_j \quad \text{for} \quad i, j = 1, \ldots, m \quad \text{and constant} \quad c_{ij}, i \neq j \quad \text{and by substituting} \quad \pi_{ij} \quad \text{into the chance-corrected coefficient} \quad \kappa = \sum \frac{\pi_{ii} - \sum \pi_{i}. \pi_{.i}}{1 - \sum \pi_{i}. \pi_{.i}}. \]

Following Agresti (1989), the agreement model is to express the cell probabilities in terms of the common
\( \kappa_M \) (Thereafter, we call this common kappa as model-based kappa),

\[ \pi_{ii} = \pi_i^2 + \kappa_M \pi_i (1 - \pi_i), \quad (3.1) \]

\[ \pi_{ij} = \pi_i \pi_j (1 - \kappa_M) \quad \text{for} \quad i \neq j, \quad (3.2) \]

where \( \{\pi_{ij}\} \) are expressed with \( \kappa_M \) and the marginal probabilities \( \pi_i. \)

In this model, \( \{\pi_{ij}\} \) satisfies symmetry and quasi-independence (Agresti,
1992). For symmetry, \( \pi_{ij} = \pi_{ji} \) for all \( i \) and \( j \). For quasi-independence, the rat-
ing by one observer is statistically independent of the rating by another observer,
given that the row response differs from the column response. That is, \( c_{ij} = c \) for
all \( i \neq j \) (Agresti, 1990, P.355). To test the fit of a symmetry model, the Pearson
statistic \( X^2 = \sum \sum \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij} + \hat{n}_{ji}} \) can be employed, based on \( df = m(m - 1)/2. \)
On the other hand, we have several comments on the model. Firstly, $\kappa_M$ has the characteristics of chance-corrected agreement as discussed in Chapter 2 where $p_o$ is the percentage of agreement: $\sum \pi_{ii}$, and $p_e$ is the percentage of agreement on the basis of chance alone: $\sum \pi_i^2$.

Secondly, Agresti (1992) noted that several authors (Bloch and Kraemer, 1989; Fleiss, 1971; Kraemer, 1979) have suggested the use of a separate kappa for category $i$, which is

$$\kappa_i = \frac{\pi_{ii} - \pi_i^2}{\pi_i(1 - \pi_i)} \quad i = 1, \ldots, m. \quad (3.3)$$

It can be rewritten as

$$\pi_{ii} = \pi_i^2 + \kappa_i \pi_i(1 - \pi_i).$$

where $\kappa_i$ represents the degree to which agreement for category $i$ exceeds that expected under rating independence. The model-based kappa is a weighted average of $m$ separate kappa,

$$\kappa_M = \frac{\sum \kappa_i \pi_i(1 - \pi_i)}{\sum \pi_i(1 - \pi_i)},$$

having value of one (or zero) if and only if each $\kappa_i$ has value of one (or zero).

When all $\kappa_i$ are the same, $\kappa_i = \kappa_M$.

Thirdly, from equations (3.1) and (3.2), we have

$$P(X_2 = j | X_1 = j) = \pi_j + \kappa_M(1 - \pi_j),$$

$$P(X_2 = j | X_1 = i) = \pi_j(1 - \kappa_M) \quad i \neq j,$$

where the above are the conditional probabilities of a second independent rating.
given the first. Therefore,

\[ P(X_2 = j|X_1 = j) - P(X_2 = j|X_1 = i) = \kappa_M \quad \text{for all } i, j, \]

\( \kappa_M \) represents the correspondence of a second opinion to the first, and is a measure of reproducibility.

In addition, since each cell probability in equations (3.1) and (3.2) is bounded by zero and its corresponding marginal probabilities, we obtain the boundary conditions for the cell probabilities and \( \kappa_M \) as below

\[ 0 \leq \pi_{ii} \leq \pi_i, \quad (3.4) \]

\[ 0 \leq \pi_{ij} \leq \min(\pi_i, \pi_j), \quad (3.5) \]

and by equations (3.1) and (3.4), we have

\[ \pi_i^2 + \kappa_M \pi_i (1 - \pi_i) \geq 0, \quad (3.6) \]

\[ \kappa_M \geq -\frac{\pi_i}{1 - \pi_i} \quad \text{for all } i. \quad (3.7) \]

By equation (3.7), the minimum value of \( \kappa_M \) depends on the marginal proportions. At the same time, by equations (3.2) and (3.5), the maximum value of \( \kappa_M \) is one. Consequently, we have

\[ -\min_i \left[ \frac{\pi_i}{1 - \pi_i} \right] \leq \kappa_M \leq 1. \]

When the model holds, \( \kappa_M = 1 \) if there is complete agreement (i.e., \( \sum \pi_{ii} = 1 \)), while \( \kappa_M = 0 \) if there is statistical independence of ratings. On the other hand, \( \kappa_M \leq 0 \) if observed agreement is less than or equal to chance agreement.

In conclusion, although the model-based kappa describes agreement in a table in which observers are compelled to have the same marginal distributions for their
ratings, there is a case in which model-based kappa describes both the pattern and the strength of agreement. For further discussion of this model, one can refer to Agresti (1989) and Agresti (1992).

3.2 Parameter Estimation

For a sample of \( n \) subjects rated by the observers with frequency responses as in Table 1.2, we assume a multinomial distribution for the cell counts \( n_{ij} \). With a \( m \times m \) table, we can write the cell probabilities of the agreement model in previous section as follows:

<table>
<thead>
<tr>
<th>Observer 1</th>
<th>( c_1 )</th>
<th>( c_2 )</th>
<th>...</th>
<th>( c_m )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( c_1 )</td>
<td>( \pi_1^2 + \kappa M \pi_1 (1 - \pi_1) )</td>
<td>( \pi_1 \pi_2 (1 - \kappa M) )</td>
<td>...</td>
<td>( \pi_1 \pi_m (1 - \kappa M) )</td>
<td>( \pi_1 )</td>
</tr>
<tr>
<td>( c_2 )</td>
<td>( \pi_2 \pi_1 (1 - \kappa M) )</td>
<td>( \pi_2^2 + \kappa M \pi_2 (1 - \pi_2) )</td>
<td>...</td>
<td>( \pi_2 \pi_m (1 - \kappa M) )</td>
<td>( \pi_2 )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( c_m )</td>
<td>( \pi_m \pi_1 (1 - \kappa M) )</td>
<td>( \pi_m \pi_2 (1 - \kappa M) )</td>
<td>...</td>
<td>( \pi_m^2 + \kappa M \pi_m (1 - \pi_m) )</td>
<td>( \pi_m )</td>
</tr>
<tr>
<td>Total</td>
<td>( \pi_1 )</td>
<td>( \pi_2 )</td>
<td>...</td>
<td>( \pi_m )</td>
<td>1</td>
</tr>
</tbody>
</table>

Suppose a random sample of size \( n \) is observed, the likelihood function for the model is

\[
L = C \prod_{i=1}^{m} \prod_{j=1}^{m} \pi_{ij}^{n_{ij}},
\]  

(3.8)
where \( C \) is a known constant and \( n_{ij} \) is the observed frequency in the \((i, j)\)th cell, for \( i, j = 1, \ldots, m \) with constraints \( \sum_{i=1}^{m} \sum_{j=1}^{m} n_{ij} = n \) and \( \sum_{i=1}^{m} \sum_{j=1}^{m} n_{ij} = 1 \). From equations (3.1) and (3.2),

\[
L \propto \prod_{i=1}^{m} [\pi_i^2 + \kappa_M \pi_i (1 - \pi_i)]^{n_{ii}} \prod_{i \neq j} \pi_i \pi_j (1 - \kappa_M)^{n_{ij} + n_{ji}}.
\]

It is convenient to work with the log-likelihood function, the log-likelihood function of the model is:

\[
\ln L = \text{constant} + \sum_{i=1}^{m} (n_{ii} + n_{i} - n_{ii}) \ln(\pi_i) + \sum_{i=1}^{m} n_{ii} \ln[\pi_i + \kappa_M (1 - \pi_i)] \\
+ (n - \sum n_{ii}) \ln(1 - \kappa_M). 
\]  

(3.9)

Let \( \beta = (\pi_1, \pi_2, \ldots, \pi_{m-1}, \kappa_M)' \) be the vector of unknown parameters. Since there is a constraint that \( \sum \pi_i = 1 \) so that we only need to find the maximum likelihood estimate (MLE) of \( \beta \). The MLE of \( \beta \) is the vector \( \hat{\beta} = (\hat{\pi}_1, \hat{\pi}_2, \ldots, \hat{\pi}_{m-1}, \hat{\kappa}_M)' \) which maximizes the likelihood function (3.8). It is equivalent to maximize the log-likelihood function (3.9).

Since this is a constraint optimization problem, method of Lagrange multiplier can be employed in this case. It is equivalent to maximize

\[
L_1 = \sum_{i=1}^{m} (n_{ii} + n_{i} - n_{ii}) \ln(\pi_i) + \sum_{i=1}^{m} n_{ii} \ln[\pi_i + \kappa_M (1 - \pi_i)] \\
+ (n - \sum n_{ii}) \ln(1 - \kappa_M) + \lambda (1 - \sum \pi_i), 
\]

where \( \lambda \) is the Lagrange multiplier.

Differentiating equation (3.10) with respect to \( \pi_i \) yields

\[
\frac{\partial L_1}{\partial \pi_i} = \frac{n_{ii} + n_{i} - n_{ii}}{\pi_i} + \frac{n_{ii} (1 - \kappa_M)}{\pi_i + \kappa_M (1 - \pi_i)} - \lambda = \frac{n_{ii} + n_{i} - x_i}{\pi_i} - \lambda,
\]
where

\[
x_i = \frac{\kappa_M n_{ii}}{\pi_i + \kappa_M (1 - \pi_i)}.
\]  

(3.11)

Assuming \(x_i\) was known and setting \(\frac{dL_1}{d\pi_i}\) to zero, we have

\[
\pi_i = \frac{n_i + n_{i} - x_i}{\lambda},
\]

By summing over \(i\) gives

\[
\lambda = \sum_{i=1}^{m} (n_i + n_{i} - x_i),
\]

and then yields

\[
\hat{\pi}_i = \frac{n_i + n_{i} - x_i}{\sum_{i=1}^{m} (n_i + n_{i} - x_i)}.
\]

Then, differentiating the log-likelihood (3.10) with respect to \(\kappa_M\) yields

\[
\frac{\partial L_1}{\partial \kappa_M} = \sum_{i=1}^{m} \frac{n_{ii}(1 - \pi_i)}{\pi_i + \kappa_M (1 - \pi_i)} - \frac{n - \sum_{i=1}^{m} n_{ii}}{1 - \kappa_M}.
\]

Equating this derivative to zero and substituting with \(x_i\),

\[
\sum_{i=1}^{m} x_i (1 - \pi_i)(1 - \kappa_M) = \kappa_M (n - \sum_{i=1}^{m} n_{ii}).
\]

On the other hand, by equation (3.11)

\[
x_i \hat{\pi}_i + \kappa_M x_i (1 - \pi_i) = \kappa_M n_{ii}.
\]

Summing over \(i\) on both sides, it follows that

\[
\hat{\kappa}_M = \frac{\sum_{i=1}^{m} x_i}{n}.
\]
Finally, we get the maximum likelihood estimators (MLEs), \( \hat{\pi}_i \) and \( \hat{\kappa}_M \) as

\[
\hat{\pi}_i = \frac{n_i + n_i - x_i}{\sum_{i=1}^{m}(n_i + n_i - x_i)} \quad i = 1, \ldots, m,
\]

and

\[
\hat{\kappa}_M = \frac{\sum_{i=1}^{m} x_i}{n},
\]

respectively.

In order to obtain the ML estimates of the parameters, an iterative procedure is necessary. The procedure is written in the following steps.

**Step 1** Start with the initial value for the \( \hat{\pi}_i \) and \( \hat{\kappa}_M \):

The procedure starts with assigned values for \( \hat{\pi}_i \) and \( \hat{\kappa}_M \). Then, \( x_i \) can be considered as being known and new estimates form. As a rule of thumb, the Scott’s \( \pi_S \) is a convenient choice of the starting values:

\[
\hat{\pi}_i = \frac{n_i + n_i}{2n} \quad i = 1, \ldots, m,
\]

and

\[
\hat{\kappa}_M = \frac{p_0 - p_e}{1 - p_e} = \frac{\sum_{i=1}^{m} n_{ii} - n \sum_{i=1}^{m} \hat{\pi}_i^2}{n(1 - \sum_{i=1}^{m} \hat{\pi}_i^2)}.
\]

**Step 2** Iteration is proceeded by conducting each of the following three steps (3.12), (3.13), (3.14), (3.12),... in turn, each case is replaced by updated values previously until convergence occurs:

\[
x_i = \frac{\hat{\kappa}_M n_{ii}}{\hat{\pi}_i + \hat{\kappa}_M (1 - \hat{\pi}_i)} \quad i = 1, \ldots, m,
\]

\[
\hat{\pi}_i = \frac{n_i + n_i - x_i}{\sum_{i=1}^{m}(n_i + n_i - x_i)} \quad i = 1, \ldots, m,
\]

\[
\hat{\kappa}_M = \frac{\sum_{i=1}^{m} x_i}{n}.
\]
\( \hat{\kappa}_M \) estimates the same characteristic as Cohen's kappa when the model holds (Agresti, 1989). However, as estimators of \( \pi_{ij} \), cell probabilities \( \hat{\pi}_{ij} \) under a suitable model are better than the sample proportion \( p_{ij} \). Model-based \( \hat{\kappa}_M \) is better than Cohen's. It is well known that model-based estimation leads to an increase of asymptotic precision when the assumed model describes the data adequately (Altham, 1984; Basu et al., 1995; Bishop et al., 1975).

### 3.3 Asymptotic variance-covariance matrix

The asymptotic variance-covariance matrices of the MLEs are derived below by inverting the Fisher information matrix. First, the general idea of Fisher information will be given.

#### 3.3.1 Fisher Information

It is well known that the maximum likelihood estimator of \( \beta \) satisfies

\[
\hat{\beta} \overset{d}{\rightarrow} N[\beta, I^{-1}(\beta)],
\]

where \( \overset{d}{\rightarrow} \) denotes converge in distribution and \( I(\beta) \) is the Fisher information.

Therefore, large sample sizes allow statements to be made about the variances of MLE. Recall that the information matrix is the negative expected value of the matrix of second-order partial derivatives of the log-likelihood. For large samples, the inverse of the expected information matrix provides the asymptotic variance-covariance matrix of the MLE \( \hat{\beta} \). When several parameters are estimated, the expected information is in a matrix form, with the \( (i, j) \)th element being the
expected value of second-order partial derivative of the log-likelihood with respect to the $i$th and $j$th parameters. Suppose for $m$ parameters, $\beta$ is written as a vector $\beta = (\beta_1, \beta_2, ..., \beta_m)'$.

The expected information matrix is

$$I(\beta)_{m \times m} = 
\begin{bmatrix}
I_{1,1} & I_{1,2} & \cdots & I_{1,m} \\
I_{2,1} & I_{2,2} & \cdots & I_{2,m} \\
\vdots & \vdots & \ddots & \vdots \\
I_{m,1} & I_{m,2} & \cdots & I_{m,m}
\end{bmatrix}$$

where

$$I_{i,i} = -E \left[ \frac{\partial^2 \ln L(\beta)}{\partial \beta_i \partial \beta_i} \right] \quad i = 1, ..., m,$$

$$I_{i,j} = -E \left[ \frac{\partial^2 \ln L(\beta)}{\partial \beta_i \partial \beta_j} \right] \quad i \neq j.$$

Finally, the asymptotic variance-covariance matrix of the MLE of $\beta$ can be obtained by taking the inverse of the expected information matrix as follows:

$$\text{Cov}(\beta) = [I(\beta)]^{-1} = 
\begin{bmatrix}
\text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \cdots & \text{Cov}(\hat{\beta}_1, \hat{\beta}_m) \\
\text{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \text{Var}(\hat{\beta}_2) & \cdots & \text{Cov}(\hat{\beta}_2, \hat{\beta}_m) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(\hat{\beta}_m, \hat{\beta}_1) & \text{Cov}(\hat{\beta}_m, \hat{\beta}_2) & \cdots & \text{Var}(\hat{\beta}_m, \hat{\beta}_m)
\end{bmatrix}$$
3.3.2 Computational detail

In this section, we derive the asymptotic variance-covariance matrix for $\hat{\beta}=(\hat{\pi}_1, \hat{\pi}_2, ..., \hat{\pi}_{m-1}, \hat{\kappa}_M)'$ for the agreement model (3.1) and (3.2). The procedure to obtain the expected information matrix is as follows:

Step 1. The first derivatives of the log-likelihood function (3.9) substituted $\pi_m = 1 - \sum_{i=1}^{m-1} \pi_i$, with respect to $\pi_i$ and $\kappa_M$ for $i = 1, ..., m - 1$ yields

$$\frac{\partial \ln L}{\partial \pi_i} = \frac{n_i + n_i - n_{ii}}{\pi_i} - \frac{n_m + n_m - n_{mm}}{1 - \sum_{i=1}^{m-1} \pi_i} + \frac{n_{ii}(1 - \kappa_M)}{\pi_i + \kappa_M(1 - \pi_i)}$$

$$+ \frac{n_{mm}(\kappa_M - 1)}{1 + \sum_{i=1}^{m-1} \pi_i(\kappa_M - 1)},$$

$$\frac{\partial \ln L}{\partial \kappa_M} = \sum_{i=1}^{m-1} \frac{n_{ii}(1 - \pi_i)}{\pi_i + \kappa_M(1 - \pi_i)} + \frac{n_{mm} \sum_{i=1}^{m-1} \pi_i}{1 + \sum_{i=1}^{m-1} \pi_i(\kappa_M - 1)} - \frac{n - \sum_{i=1}^{m} n_{ii}}{1 - \kappa_M}.$$ 

Step 2. Take the second derivative of each above with respect to $\pi_i$ and $\kappa_M$ for $i = 1, ..., m - 1$

$$\frac{\partial^2 \ln L}{\partial \pi_i^2} = -\frac{n_i + n_i - n_{ii}}{\pi_i^2} - \frac{n_m + n_m - n_{mm}}{(1 - \sum_{i=1}^{m-1} \pi_i)^2} - \frac{n_{ii}(1 - \kappa_M)^2}{[\pi_i + \kappa_M(1 - \pi_i)]^2}$$

$$- \frac{n_{mm}(\kappa_M - 1)^2}{[1 + \sum_{i=1}^{m-1} \pi_i(\kappa_M - 1)]^2},$$

$$\frac{\partial^2 \ln L}{\partial \kappa_M^2} = -\sum_{i=1}^{m-1} \frac{n_{ii}(1 - \pi_i)^2}{[\pi_i + \kappa_M(1 - \pi_i)]^2} - \frac{n_{mm} \sum_{i=1}^{m-1} \pi_i^2}{[1 + \sum_{i=1}^{m-1} \pi_i(\kappa_M - 1)]^2} - \frac{n - \sum_{i=1}^{m} n_{ii}}{(1 - \kappa_M)^2},$$

$$\frac{\partial \ln L}{\partial \pi_i \partial \pi_j} = -\frac{n_m + n_m - n_{mm}}{(1 - \sum_{i=1}^{m-1} \pi_i)^2} - \frac{n_{mm}(\kappa_M - 1)^2}{[1 + \sum_{i=1}^{m-1} \pi_i(\kappa_M - 1)]^2} \quad i \neq j.$$
\[
\frac{\partial^2 \ln L}{\partial \tau_i \partial \kappa_M} = -\frac{n_{ii}}{[\tau_i + \kappa_M(1 - \tau_i)]^2} + \frac{n_{mm}}{[1 + \sum_{i=1}^{m-1} \tau_i(\kappa_M - 1)]^2}
\]

Step 3. Lastly, take the expectation of each of the above and multiply by -1 as the element of Fisher’s information matrix. The expected information matrix is evaluated by replacing the counts with their expected values, denoted by \( I_{i,j} \) for \( i, j = 1, ..., m \). The diagonal elements are given by

\[
I_{i,i} = \frac{n[2 - \tau_i - \kappa_M(1 - \tau_i)]}{\tau_i} + \frac{n[1 - \sum_{i=1}^{m-1} \tau_i(\kappa_M - 1)]}{1 - \sum_{i=1}^{m-1} \tau_i} + \frac{n\tau_i(1 - \kappa_M)^2}{\tau_i + \kappa_M(1 - \tau_i)} + \frac{n(1 - \sum_{i=1}^{m-1} \tau_i)(\kappa_M - 1)^2}{1 + \sum_{i=1}^{m-1} \tau_i(\kappa_M - 1)} \quad i = 1, ..., m - 1,
\]

\[
I_{m,m} = \sum_{i=1}^{m-1} \frac{n\tau_i(1 - \tau_i)^2}{\tau_i + \kappa_M(1 - \tau_i)} + \frac{n(1 - \sum_{i=1}^{m-1} \tau_i)(\sum_{i=1}^{m-1} \tau_i)^2}{1 + \sum_{i=1}^{m-1} \tau_i(\kappa_M - 1)} + \frac{n(1 - \sum_{i=1}^{m-1} \tau_i^2)}{1 - \kappa_M},
\]

and where the off-diagonal elements are

\[
I_{i,j} = \frac{n[1 - \sum_{i=1}^{m-1} \tau_i(\kappa_M - 1)]}{1 - \sum_{i=1}^{m-1} \tau_i} + \frac{n(1 - \sum_{i=1}^{m-1} \tau_i)(\kappa_M - 1)^2}{1 + \sum_{i=1}^{m-1} \tau_i(\kappa_M - 1)} \quad i \neq j = 1, ..., m - 1,
\]

\[
I_{i,m} = \frac{n\tau_i}{\tau_i + \kappa_M(1 - \tau_i)} - \frac{n(1 - \sum_{i=1}^{m-1} \tau_i)}{1 + \sum_{i=1}^{m-1} (\kappa_M - 1)} \quad i = 1, ..., m - 1.
\]

After taking the inverse of the Fisher information matrix and evaluated at the MLE of \( \beta \), the estimated asymptotic covariance matrix can then be obtained.

Finally, the kappa coefficient is used not only as a descriptive statistic, but also as a basis of statistical inference. Then, we can make statistical inference.
on these parameters such as hypothesis testing about the model-based kappa. Testing the null hypothesis $H_0: \kappa_M = \kappa_M^{(0)}$ can be accomplished by calculating the Wald Statistics

$$Z_0 = \frac{\hat{\kappa}_M - \kappa_M^{(0)}}{\text{ASE}(\hat{\kappa}_M)}$$

(3.15)

where $\text{ASE}(\hat{\kappa}_M)$ is the estimated asymptotic standard deviation of $\hat{\kappa}_M$.

In the next section, an illustrative example is used to demonstrate how the model-based approach provides better interpretation of the kappa rather than a single index of agreement.
3.4 Illustrative Example

In this section, a real data set is analyzed to illustrate the usage of model-based kappa. Gross (1971) has collected data based on two supervisors who were asked to rate independently the classroom style of 72 student teachers as authoritarian, democratic, or permissive. And it is analyzed by Agresti (1989) and Bishop et al. (1975). We use it to illustrate the application of the model-based kappa in the agreement model, with estimated expected frequencies given in Table 3.2.

<table>
<thead>
<tr>
<th>Rating by Supervisor 1</th>
<th>Authoritarian</th>
<th>Democratic</th>
<th>Permissive</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authoritarian</td>
<td>17(20.3)</td>
<td>4(4.6)</td>
<td>8(6.5)</td>
<td>29(31.4)</td>
</tr>
<tr>
<td>Democratic</td>
<td>5(4.6)</td>
<td>12(8.7)</td>
<td>0(3.5)</td>
<td>17(16.8)</td>
</tr>
<tr>
<td>Permissive</td>
<td>10(6.5)</td>
<td>3(3.5)</td>
<td>13(13.7)</td>
<td>26(23.7)</td>
</tr>
<tr>
<td>Total</td>
<td>32(31.4)</td>
<td>19(16.8)</td>
<td>21(23.7)</td>
<td>72</td>
</tr>
</tbody>
</table>

Parenthesized values denote estimated expected frequencies for agreement model (3.1) & (3.2)

The ML estimates and their estimated standard error are given by

<table>
<thead>
<tr>
<th>parameter</th>
<th>estimate</th>
<th>estimated SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi_1$</td>
<td>0.4368</td>
<td>0.0480</td>
</tr>
<tr>
<td>$\pi_2$</td>
<td>0.2337</td>
<td>0.0406</td>
</tr>
<tr>
<td>$\pi_3$</td>
<td>0.3295</td>
<td>0.0453</td>
</tr>
<tr>
<td>$\kappa_M$</td>
<td>0.37</td>
<td>0.0889</td>
</tr>
</tbody>
</table>
In this example, the usual single index from Cohen’s kappa is interpreted as the chance corrected agreement being 0.36. But, through this example, we will demonstrate how the model-based kappa from the agreement model can provide a more detailed interpretation.

First of all, we examine the marginal distribution of Table 3.2, the goodness of fit of the symmetry model and of the agreement model at 5% level of significance. Testing the null hypothesis of marginal homogeneity, the statistic by equation (2.1) $X^2 = 1.33 (P$-value$ = 0.2488)$ is smaller than the critical value $X^2_{2,0.05} = 5.99 (>0.05)$. The hypothesis of marginal homogeneity is not rejected. Then, testing the symmetry model, the Pearson statistic $= 0.67 (P$-value$ = 0.4131)$ is smaller than the critical value $X^2_{3,0.05} = 7.81 (>0.05)$. The symmetry model is not rejected. To test the agreement model, the Pearson Chi-squared statistic $= 7.7 (P$-value$ = 0.1736)$ is smaller than the critical value $X^2_{5,0.05} = 11.07 (>0.05)$, based on $df = m(m-1)-1$. We do not reject the agreement model. Thus, $\hat{\kappa}_M$ can be used to provide a description for the structure of the table.

The agreement between two supervisors ($\hat{\kappa}_M$) is 0.37 that represents quite weak agreement. Then, we use the model-based kappa in order to get more information about the structure of the table with equations (3.1) and (3.2). By the equation (3.1), the probability the supervisors agree that a teacher is in category $i$ exceeds the value corresponding to independency by around $0.37\hat{\pi}_i(1 - \hat{\pi}_i)$. Therefore, the corresponding probabilities that exceed are 0.091, 0.066 and 0.082 for category authoritarian, democratic and permissive respectively. In this case, any particular disagreement is around 63% as likely to occur as if the ratings are
statistically independent, i.e.,

\[ \frac{\hat{\pi}_{ij}}{\hat{\pi}_i \hat{\pi}_j} = 1 - \hat{\kappa}_M = 0.63 \quad \text{for all } i \neq j, \]

These are more informative interpretation for the kappa in terms of agreement and disagreement structure.

On the other hand, we can find the value of the separate kappa by equation (3.3) as \( \kappa_1=0.2321, \kappa_2=0.5556 \) and \( \kappa_3=0.3367 \). By these three indexes, the two supervisors agree most on the category democratic being higher than on the authoritarian and permissive in turn. Intuitively, the magnitude of agreement for each category is not identical. Then, one may use the ratio,

\[ \frac{\hat{\kappa}_M - \kappa_i}{\hat{A}\hat{S}\hat{E}(\hat{\kappa}_M)} \]

as an indicator of how far away between this separate kappa and the model-based kappa. We find that the ratios are 1.55, -2.088, 0.3745 for the corresponding separate kappa. The ratio is the greatest in the case of second separate kappa indicating the possible violation of the assumption of identical kappa though the chi-square test does not reject the hypothesis. One may conclude that the third separate kappa is near to the estimate of the model-based kappa. For the first and second separate kappa, they are slightly different from the estimate. As a result, the degree of agreement between two supervisors on each category are not the identical. In this example, the model-based kappa is just a weighted average of each separate kappa and is not identical to all of the separate kappa.

The properties of the estimator \( \hat{\kappa} \) will be investigated via Monte Carlo simulation study.
Chapter 4

Simulation Study

In this chapter, a simulation study is carried out to investigate the performance of the estimates of the model-based kappa developed in Chapter 3 and to have an insight into the appropriate sample size for asymptotic theory to hold. A FORTRAN program has been written and the procedure of the simulation study is presented.

4.1 Design

To investigate the small-sample properties of the estimator of model-based kappa, a simulation study is conducted for $3 \times 3$ table. Each table is generated with sample size $n$ fixed at 45, 90, and 180. For each sample size, model-based kappa $= 0.0, 0.2, 0.4, 0.6, 0.8$ for different designs of marginal probabilities $\pi_1, \pi_2$ and $\pi_3$ are constructed:

design 1: $(\pi_i = \pi_j: 1/3, 1/3, 1/3)$,
design 2: \((\pi_i = \pi_j: 0.2, 0.3, 0.5)\), and

design 3: \((\pi_i = \pi_j: 0.1, 0.2, 0.7)\).

These designs are used to represent a variety of different tables encountered in real life. The cell probability \(\pi_{ij}\) of each combination can be easily obtained under the agreement model in Chapter 3,

\[
\pi_{ii} = \pi_i^2 + k^M \pi_i (1 - \pi_i), \\
\pi_{ij} = \pi_i \pi_j (1 - k^M) \quad \text{for} \quad i \neq j.
\]

For every combination of sample size, design and model-based kappa, 1000 tables \((N=1000\) replications\) are drawn from a multinomial distribution with sample size \(n=45, 90\) and \(180\) and cell probabilities \(\pi_{ij}, i=1, 2, 3, j=1, 2, 3\). As a result, there are \(3 (\text{sample sizes}) \times 3 (\text{designs of } \pi_i) \times 5 (\text{designs of } k^M) \times 1000 (\text{tables}) = 45000\) trials in this study.

First, random sample of size \(n\) is generated from the multinomial distribution with 9 known cell probabilities using the IMSL random number generator RN-MTN in FORTRAN. Second, to compute the iterative solution of model-based kappa estimates in \(3\times3\) table, a FORTRAN program is written for this task. Third, for each simulation, the maximum likelihood estimates of the model-based kappa and its standard error are obtained. Let \(k^M\) denote the maximum likelihood estimate of the model-based kappa \((k^M)\). In estimating the kappa, the maximum number of iterations allowed is usually fixed at 40 and the tolerance level for convergence is usually 0.001. Finally, since the true value of the parameter under the model is known, five criteria are considered to evaluate the
performance of the ML estimates of kappa.

1. Simulated Bias of the estimate, $\text{bias}(\hat{\kappa}_M)$.

$\hat{\kappa}_M$ is the estimate of the $\kappa_M$ based on sample size $n$ in the $i$-th replication. The mean of the estimate ($\bar{\hat{\kappa}}_M$) is calculated by the average over the replications of the estimates $\hat{\kappa}_M$. The simulated bias is calculated by the difference between the mean and true agreement $\kappa_M$, and is defined as

$$\text{bias}(\hat{\kappa}_M) = \frac{\sum_{i=1}^{N} \hat{\kappa}_M - \kappa_M}{N}.$$  

If $\bar{\hat{\kappa}}_M$ is near to $\kappa_M$, the bias is small and the estimate is accurate.

2. Root Mean Square of estimate, $\text{RMS}_{\kappa_M}$.

The root mean square of the estimate of $\kappa_M$ is calculated as the square root of the average squared difference between the estimate and the $\kappa_M$

$$\text{RMS}_{\kappa_M} = \left[ \frac{\sum_{i=1}^{N} (\hat{\kappa}_M - \kappa_M)^2}{N} \right]^{1/2}.$$  

If $\text{RMS}_{\kappa_M}$ is small, the estimate is closed to the true $\kappa_M$ and the estimate is accurate.

3. Variance Ratio, VR

$\overline{\text{SE}}_{\kappa_M}$ is defined as the mean of the asymptotic standard error of $\hat{\kappa}_M$ over all replications, where SE is obtained from the inverse of Fisher information in Chapter 3. $\overline{\text{SE}}_i$ is the asymptotic Standard Error in the $i$th replication. And, SD is the sample standard deviation of the estimates of the model-based kappa. Since both SD and $\overline{\text{SE}}_{\kappa_M}$ are the estimates of the spread of
\( \hat{\kappa}_M \), their values should be closed. So, we calculate the variance ratio VR of \( \overline{SE}_{\kappa_M} \) and SD, where

\[
\overline{SE}_{\kappa_M} = \frac{\sum_{i=1}^{N} SE_i}{N}, \quad (4.3)
\]
\[
SD = \left[ \frac{\sum_{i=1}^{N} (\hat{\kappa}_{M_i} - \overline{\kappa}_M)^2}{N-1} \right]^{1/2}, \quad (4.4)
\]
\[
VR = \frac{\overline{SE}_{\kappa_M}}{SD}. \quad (4.5)
\]

If the ratio is closed to unity, the estimated standard error is accurate.

4. Simulated Probability of Type I error.

In testing the null hypothesis \( H_0 : \kappa_M = \kappa_M^{(o)} \), \( \kappa_M^{(o)} \) is a pre-specified value, the test statistic \( Z_i \) based on the \( i \)th replication is

\[
Z_i = \frac{\hat{\kappa}_{M_i} - \kappa_M}{SE_i}. \quad (4.6)
\]

If \( |Z_i| > Z_{\alpha/2} \), the null hypothesis is rejected at level of significance \( \alpha \), where \( Z_{\alpha/2} \) is the upper \( \alpha/2 \) percentage point of the standard normal distribution \( N[0, 1] \). We take the level of significance \( \alpha = 0.05 \).

Simulated probability of Type I error is the proportion of times that the null hypothesis is rejected in the total number of trials \( N=1000 \),

\[
\text{Simulated probability of Type I error} = \frac{\text{\# that the null hypothesis is rejected}}{N}
\]

If it is closed to \( \alpha \), it indicates that the type I error rate is controlled at the desired level in testing the null hypothesis about the value of \( \kappa_M \).
5. Significant Probability of the Kolmogorov-Smirnov one-sample test, SP

To test the null hypothesis that samples $Z_1, \ldots, Z_N \sim N[0, 1]$, the Kolmogorov-Smirnov one-sample test is used.

Define $Z_{(i)}$ is the i-th order statistic of $Z_1, \ldots, Z_N$ and $\Phi(.)$ is the cumulative distribution function of standard normal distribution. Then according to Gibbons (1971), the Kolmogorov-Smirnov one-sample test statistic is

$$D_N = \max\{\max_{1 \leq i \leq N} \left[\frac{1}{N} - \Phi(Z_{(i)})\right], \max_{1 \leq i \leq N} [\Phi(Z_{(i)}) - \frac{i - 1}{N}], 0\},$$

and the limiting distribution of $D_N$ for every $d \geq 0$ is

$$\lim_{m \to \infty} P(D_N \leq \frac{d}{\sqrt{N}}) = L(d),$$

where

$$L(d) = 1 - 2 \sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}.$$

If the significant probability $> probability of Type I error ($\alpha = 0.05$), the null hypothesis that sample $Z_1, \ldots, Z_N$ follows $N[0, 1]$ is not rejected at 5% level of significance.

4.2 Results

The results of the simulation study are summarized in Table 1a to 5c, Table 4.1 and Figure 1.1 to 1.5. From the simulation study, there are several findings. In particular, the effects of the sample size $n$ to the estimates will be discussed.
1. Accuracy

From Table 1a to 1c, the bias of the estimates are all small. In addition, when the sample size \( n \) increases, the bias decreases. Actually, if the sample is of at least 45, the bias of the ML estimate is negligible.

2. Effects of sample size

From Table 2a to 2c, all RMS\(_{\kappa_M}\) are small. RMS reduction may well be related to sample size as may be seen by comparing the results in Table 2a for samples of size 45 and 90. When the sample size \( n \) increases, most RMS\(_{\kappa_M}\)s decrease. In this case, small sample size (\( n=45 \)) is acceptable and moderate samples (\( n=90 \)) are good to obtain accurate estimates.

3. Effects of the magnitude of true agreement or the design of the marginal distribution

From Table 2a to 2c, for a fixed design and sample size, the RMS\(_{\kappa_M}\) with different true agreement are of similar values. On the other hand, for fixed true agreement and sample size, the RMS\(_{\kappa_M}\)s with different design are very closed. Hence, the effect of the magnitude of true agreement and design of marginal distribution is not important when comparing the effect of the sample size.

4. Accuracy of the estimated Standard error

In Table 3a to 3c, one can see that all Variance Ratios are between 0.95 and 1.08 and approximated to unity. It indicates that the estimated standard
error of the estimates is accurate.

5. Simulated probability of Type I error

According to Table 4a to 4c, when the sample size is small, most of the proportion is slightly greater than 0.05. This indicates that the null hypothesis is often wrongly rejected when testing the null hypothesis of the value of \( \kappa_M \). However, when the sample size is moderate or sufficiently large, the proportion is around 0.05. It indicates that the probability of type I error is at a desirable level.

6. Normality

From Table 5a to 5c, in testing the null hypothesis that samples \( Z_1, \ldots, Z_N \sim N[0, 1] \), the entries with rejection of normality are in bold. The results show that normality is rejected for small samples in some cases but the normality is not rejected for all cases with moderate or large samples.

For those entries with normality rejected, the normal Q-Q plot is employed to examine the shape of the distribution of samples \( Z_1, \ldots, Z_N \). A straight diagonal line indicates normal distributed sample. The plots are presented in Figure 1.1 to 1.5. Also, kurtosis and skewness are calculated as shown in Table 4.1. For kurtosis, the normal distribution has a zero value of the kurtosis statistic. A positive kurtosis indicates that the observations cluster more and have longer tails than those in the normal distribution. And, negative kurtosis indicates the observations cluster less and have shorter tails. For skewness, the normal distribution is symmetric and has a zero value.
of skewness statistic. A distribution with a significant positive skewness is skewed to left. A distribution with a significant negative skewness is skewed to right (Hair et al., 1995).

Table 4.1: The results of the skewness and kurtosis

<table>
<thead>
<tr>
<th>( \kappa_M )</th>
<th>sample size</th>
<th>design</th>
<th>skewness</th>
<th>kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>45</td>
<td>1</td>
<td>0.58</td>
<td>1.09</td>
</tr>
<tr>
<td>0.6</td>
<td>45</td>
<td>3</td>
<td>0.54</td>
<td>3.07</td>
</tr>
<tr>
<td>0.8</td>
<td>45</td>
<td>1</td>
<td>1.12</td>
<td>2.82</td>
</tr>
<tr>
<td>0.8</td>
<td>45</td>
<td>2</td>
<td>1.10</td>
<td>2.27</td>
</tr>
<tr>
<td>0.8</td>
<td>45</td>
<td>3</td>
<td>0.89</td>
<td>1.49</td>
</tr>
</tbody>
</table>

From Figure 1.1 to 1.5, the points do not show a straight line pattern in all figures. It indicates that the kurtosis and skewness of the sample \( Z_1, ..., Z_N \) are far away from zero. From Table 4.1, we can see that all the cases have positive kurtosis and skewness. It means that the distributions of samples have longer tails than the normal distribution and are slightly skewed to the left.

### 4.3 Discussion

Based on the results of the simulation study, the ML estimate of the model-based kappa performed well as both the bias and RMS of the estimates are small. It is seen that the model-based kappa estimate is accurate even the sample size
is small and would appear to be better with increasing sample size.

To achieve further statistical inference, the asymptotic standard error of the estimate is derived. It is compared with the empirical standard deviation in order to evaluate its accuracy. The results show that the asymptotic standard error is fairly accurate. Despite studying accuracy of the standard error, the normality of the estimate is also investigated. Since the MLE is asymptotically normal, it is interesting to study and suggest the reliable sample size for the asymptotic theory to hold. According to the results of the simulated probability of Type I error, all proportions are around 0.05 when sample size is moderate or sufficiently large. The Type I error rate is controlled at the desired level in testing the null hypothesis about the value of $\kappa_M$. At the same time, with the significant probability of the Kolmogorov-Smirnov test for testing the null hypothesis that samples $Z_1, ..., Z_N \sim N[0, 1]$, asymptotic normality seems to be hold when sample size is as small as $n=90$ for 3x3 table. Therefore, the sample size should be as small as moderate to ensure the ML estimate converges in distribution to normal.

From the results of the simulation study in the last section, the ML estimate of the model-based kappa is good in moderately large samples. In fact, in real situation, we do not know the true value of agreement that there is a problem of slightly bias or non-normality in using the estimator with small sample size. Thus, it is suggested to use the ML estimate of the model-based kappa with at least moderate sample size.
Chapter 5

Conclusion

In this thesis, traditional measures of agreement which are corrected for chance are reviewed and discussed. We have considered the properties of these measures and recommended that marginal homogeneity is assessed as the first step in the analysis of observer agreement. Also, because these measures are defined on ad hoc basis, a model proposed by Bloch and Kraemer (1989) is then used to find the estimator of kappa in $2 \times 2$ table.

With modelling approach, the agreement model discussed in Chapter 3 in which kappa as a parameter in the simple model for cell probabilities for $m \times m$ table, maximum likelihood estimators of the model-based kappa and the marginal probabilities are found. In order to have significance testing and construction of confidence intervals, the asymptotic variance-covariance matrices of the ML estimates are derived.

Although the computation of traditional kappa is generally simple, the algorithm of the iterative solution of the MLE converges quickly with computer. The
The greatest advantage of the model-based kappa is that it is estimated based on the likelihood function. Another advantage is that the interpretation of the kappa would be more informative while the traditional kappa could be interpreted as a single index only.

Finally, with our studies in the small to moderate sample performance of the model-based kappa in $3 \times 3$ contingency table through Monte Carlo studies, it is suggested that the ML estimates of the model-based kappa and its asymptotic standard error are acceptable in small samples and good for moderate sample generally.
## Tables for simulation study

### Table 1a: Simulated Bias of the Estimate for design 1

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>-0.0044</td>
<td>-0.0102</td>
<td>-0.0056</td>
<td>-0.0108</td>
<td>-0.0086</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>-0.0049</td>
<td>-0.0042</td>
<td>-0.0092</td>
<td>-0.0030</td>
<td>-0.0074</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>-0.0042</td>
<td>-0.0023</td>
<td>-0.0033</td>
<td>-0.0023</td>
<td>-0.0013</td>
</tr>
</tbody>
</table>

### Table 1b: Simulated Bias of the Estimate for design 2

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>-0.0149</td>
<td>-0.0114</td>
<td>-0.0129</td>
<td>-0.0091</td>
<td>-0.0034</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>-0.0023</td>
<td>-0.0047</td>
<td>-0.0069</td>
<td>-0.0067</td>
<td>-0.0025</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>-0.0041</td>
<td>-0.0022</td>
<td>-0.0013</td>
<td>-0.0041</td>
<td>-0.0031</td>
</tr>
</tbody>
</table>

### Table 1c: Simulated Bias of the Estimate for design 3

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>-0.0153</td>
<td>-0.0123</td>
<td>-0.0154</td>
<td>-0.0153</td>
<td>-0.0139</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>-0.0064</td>
<td>-0.0047</td>
<td>-0.0100</td>
<td>-0.0037</td>
<td>-0.0046</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.0029</td>
<td>-0.0029</td>
<td>-0.0036</td>
<td>-0.0027</td>
<td>-0.0012</td>
</tr>
</tbody>
</table>
Table 2a: The Root Mean Square of the Estimate for design 1

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.1162</td>
<td>0.1094</td>
<td>0.1124</td>
<td>0.1015</td>
<td>0.0785</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.0803</td>
<td>0.0788</td>
<td>0.0786</td>
<td>0.0864</td>
<td>0.0575</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.0568</td>
<td>0.0561</td>
<td>0.0566</td>
<td>0.0486</td>
<td>0.0391</td>
</tr>
</tbody>
</table>

Table 2b: The Root Mean Square of the Estimate for design 2

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.1075</td>
<td>0.1118</td>
<td>0.1161</td>
<td>0.0990</td>
<td>0.0802</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.0814</td>
<td>0.0822</td>
<td>0.0809</td>
<td>0.0725</td>
<td>0.0546</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.0564</td>
<td>0.0573</td>
<td>0.0574</td>
<td>0.0536</td>
<td>0.0407</td>
</tr>
</tbody>
</table>

Table 2c: The Root Mean Square of the Estimate for design 3

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.1237</td>
<td>0.1268</td>
<td>0.1386</td>
<td>0.1281</td>
<td>0.0966</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.0948</td>
<td>0.0943</td>
<td>0.0980</td>
<td>0.0879</td>
<td>0.0575</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.0648</td>
<td>0.0681</td>
<td>0.0700</td>
<td>0.0613</td>
<td>0.0611</td>
</tr>
</tbody>
</table>
Table 3a: Variance Ratio of $\frac{SE_{\kappa_M}}{SD}$ to SD for design 1

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.9535</td>
<td>1.0176</td>
<td>0.9738</td>
<td>0.9825</td>
<td>0.9774</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.9800</td>
<td>0.9900</td>
<td>0.9934</td>
<td>1.0007</td>
<td>0.9897</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.9829</td>
<td>0.9939</td>
<td>0.9691</td>
<td>1.0190</td>
<td>1.0111</td>
</tr>
</tbody>
</table>

Table 3b: Variance Ratio of $\frac{SE_{\kappa_M}}{SD}$ to SD for design 2

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>1.0729</td>
<td>1.0263</td>
<td>0.9883</td>
<td>1.0512</td>
<td>0.9796</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.9975</td>
<td>0.9888</td>
<td>1.0031</td>
<td>1.0138</td>
<td>1.0235</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>1.0225</td>
<td>1.0054</td>
<td>0.9943</td>
<td>0.9665</td>
<td>0.9788</td>
</tr>
</tbody>
</table>

Table 3c: Variance Ratio of $\frac{SE_{\kappa_M}}{SD}$ to SD for design 3

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>1.0322</td>
<td>1.0107</td>
<td>0.9650</td>
<td>0.9626</td>
<td>0.9917</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.9690</td>
<td>0.9741</td>
<td>0.9712</td>
<td>0.9826</td>
<td>0.9787</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>1.0138</td>
<td>0.9606</td>
<td>0.9597</td>
<td>0.9985</td>
<td>1.0001</td>
</tr>
</tbody>
</table>
Table 4a: Simulated Probability of Type I error for design 1

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.076</td>
<td>0.055</td>
<td>0.073</td>
<td>0.072</td>
<td>0.059</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.056</td>
<td>0.063</td>
<td>0.051</td>
<td>0.070</td>
<td>0.057</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.049</td>
<td>0.053</td>
<td>0.052</td>
<td>0.053</td>
<td>0.061</td>
</tr>
</tbody>
</table>

Table 4b: Simulated Probability of Type I error for design 2

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.036</td>
<td>0.050</td>
<td>0.066</td>
<td>0.041</td>
<td>0.086</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.058</td>
<td>0.066</td>
<td>0.048</td>
<td>0.052</td>
<td>0.052</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.044</td>
<td>0.056</td>
<td>0.053</td>
<td>0.055</td>
<td>0.063</td>
</tr>
</tbody>
</table>

Table 4c: Simulated Probability of Type I error for design 3

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.080</td>
<td>0.064</td>
<td>0.077</td>
<td>0.073</td>
<td>0.066</td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.077</td>
<td>0.080</td>
<td>0.069</td>
<td>0.071</td>
<td>0.065</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.061</td>
<td>0.065</td>
<td>0.059</td>
<td>0.057</td>
<td>0.060</td>
</tr>
</tbody>
</table>
Table 5a: Significant Probability SP for design 1

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.167</td>
<td>0.218</td>
<td>0.059</td>
<td><strong>0.000</strong></td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.168</td>
<td>0.606</td>
<td>0.345</td>
<td>0.245</td>
<td>0.183</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.211</td>
<td>0.759</td>
<td>0.763</td>
<td>0.138</td>
<td>0.079</td>
</tr>
</tbody>
</table>

Table 5b: Significant Probability SP for design 2

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.785</td>
<td>0.618</td>
<td>0.528</td>
<td>0.086</td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.781</td>
<td>0.238</td>
<td>0.926</td>
<td>0.516</td>
<td>0.084</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.548</td>
<td>0.534</td>
<td>0.295</td>
<td>0.267</td>
<td>0.574</td>
</tr>
</tbody>
</table>

Table 5c: Significant Probability SP for design 3

<table>
<thead>
<tr>
<th>$\kappa_M$</th>
<th>0.0</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 45$</td>
<td>0.664</td>
<td>0.767</td>
<td>0.074</td>
<td><strong>0.021</strong></td>
<td><strong>0.000</strong></td>
</tr>
<tr>
<td>$n = 90$</td>
<td>0.191</td>
<td>0.128</td>
<td>0.716</td>
<td>0.148</td>
<td>0.067</td>
</tr>
<tr>
<td>$n = 180$</td>
<td>0.655</td>
<td>0.078</td>
<td>0.991</td>
<td>0.540</td>
<td>0.150</td>
</tr>
</tbody>
</table>
Figures for simulation study

Figure 1.1 Normal Q-Q Plot of $Z_i$ in design 1, $\kappa_M = 0.6$ and $n=45$

Figure 1.2 Normal Q-Q Plot of $Z_i$ in design 3, $\kappa_M = 0.6$ and $n=45$
Figure 1.3 Normal Q-Q Plot of $Z_i$ in design 1, $\kappa_M = 0.8$ and $n=45$

Figure 1.4 Normal Q-Q Plot of $Z_i$ in design 2, $\kappa_M = 0.8$ and $n=45$

Figure 1.5 Normal Q-Q Plot of $Z_i$ in design 3, $\kappa_M = 0.8$ and $n=45$
Bibliography


