CONFIDENCE INTERVALS (CI) FOR CONCENTRATION PARAMETER IN VON MISES DISTRIBUTION AND ANALYSIS OF MISSING VALUES FOR CIRCULAR DATA

SITI FATIMAH BINTI HASSAN

INSTITUTE OF GRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2015

CONFIDENCE INTERVALS (CI) FOR CONCENTRATION PARAMETER IN VON MISES DISTRIBUTION AND ANALYSIS OF MISSING VALUES FOR CIRCULAR DATA

SITI FATIMAH BINTI HASSAN

THESIS SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

INSTITUTE OF GRADUATE STUDIES
UNIVERSITY OF MALAYA
KUALA LUMPUR

2015

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: SITI FATIMAH BT HASSAN (I.C No: 841009135224)

Registration/Matric No: HHC 100006

Name of Degree: DOCTOR OF PHILOSOPHY

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

CONFIDENCE INTERVALS (CI) FOR CONCENTRATION PARAMETER IN VON MISES DISTRIBUTION AND ANALYSIS OF MISSING VALUES FOR CIRCULAR DATA

Field of Study: STATISTICS

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name: ASSOC. PROF. DR YONG ZULINA ZUBAIRI

Designation:

ABSTRACT

This study is on circular statistics that is also known as directional statistics. Directional statistics is a branch of statistics which deal with the data in angle form in which the method of analysis is different from linear data. For example, the distribution analogues to the normal distribution in linear data is known as circular normal distribution. This study comprises of four parts. The first part of the study focuses on the efficient approximation for the concentration parameter in von Mises distribution. Here, a new method of approximating the concentration parameter is proposed, and the performance of the proposed method is studied via simulation study.

The second part of the study is on the confidence intervals (CI) for the concentration parameter in von Mises distribution. Several methods in constructing the CI for the concentration parameter are proposed including CI based on circular population, CI based on the asymptotic distribution of $\hat{\kappa}$, CI based on the distribution of $\bar{\theta}$ and $\bar{R}$ and also CI based on bootstrap-*t* method. All proposed methods are validated via simulation study and the performance indicator such as an expected length and its coverage probability are evaluated.

The third part of the study is on the derivation of the circular distance for circular data. From this derivation, we construct the CI for the concentration parameter. Three different methods will be considered in proposing the new CI including mean, median and percentile. The simulation studies carried out to assess the performance of each proposed method.

The final part of this study is an analysis of missing values for circular variables. Missing values is a common problem that occurs in data collection. By ignoring the

existence of missing values, leads to the biasness and lack of efficiency of a statistics. In this study, three imputation methods are considered namely expectation-maximization (EM) algorithm and data augmentation (DA) algorithm. All proposed methods are compared to the conventional methods. The analyses are conducted by doing the simulation studies by varying the value of the concentration parameter. All the proposed methods from this study are illustrated using the real data consisting of data in angle form found in the literature.

ABSTRAK

Kajian ini adalah mengenai statistik membulat atau lebih dikenali dengan statistik berarah. Statistik berarah adalah suatu cabang bidang yang menggunakan data dalam ukuran sudut dan dengan itu kita memerlukan kaedah yang berlainan dalam menjalankan analisis data tersebut. Kajian ini terbahagi kepada empat bahagian utama. Dalam kajian ini, taburan von Mises akan digunakan sebagai taburan utama dalam melakukan perbincangan kajian. Taburan von Mises juga dikenali sebagai taburan normal membulat dan ia merupakan taburan yang menyerupai taburan normal seperti yang biasa digunakan dalam statistik linear. Bahagian pertama akan memberi focus kepada penganggaran untuk parameter menumpu dalan taburan von Mises. Dalam bahagian ini, kaedah penganggaran terbaru untuk parameter menumpu akan dicadangkan dan diikuti dengan kajian simulasi bagi menilai ketepatan prosedur yang telah dicadangkan.

Bahagian kedua akan membincangkan tentang selang keyakinan (SK) untuk parameter menumpu bagi taburan von Mises. Beberapa kaedah untuk menghasilkan selang keyakinan (SK) akan dicadangkan termasuk SK berdasarkan populasi membulat, SK berdasarkan taburan asimptotik $\hat{\kappa}$ matrix maklumat Fisher, SK berdasarkan taburan $\bar{\theta}$ dan $\bar{R}$ dan SK berdasarkan kaedah bootstrap-*t*. Semua kaedah yang telah dicadangkan akan disahkan melalui kajian simulasi dan penilaian bagi saiz selang dan kebarangkalian menumpu akan dinilai.

Bahagian ketiga kajian adalah untuk menghasilkn jarak membulat bagi data membulat. Berdasarkan penghasilan ini, didapati kita juga berjaya untuk menghasilkan SK bagi parameter menumpu.Tiga kaedah yang berbeza akan diperkenalkan untuk

menghasilkan SK termasuk min, median dan persentil. Kajian simulasi akan dilakukan bagi menilai ketepatan kaedah yang telah dicadangkan.

Bahagian terakhir dalam kajian ini adalah analisis data lenyap bagi pemboleh ubah membulat. Data lenyap merupakan suatu permasalahan biasa dalam pegumpulan data. Dengan hanya mengabaikan kewujudan data yang lenyap atau hilang akan membawa kepada bias dan menyebabkan ia menjadi kurang signifikan secara statistik. Dalam kajian ini, tiga kaedah imputasi akan dicadangkan termasuk 'expectation maximization' (EM) dan 'data augmentation' (DA). Semua kaedah yang dicadangkan akan dibandingkan dengan kaedah konvensional yang biasa digunakan. Untuk menentukan kaedah terbaik, analisis dibuat melalui kajian simulasi dengan mempelbagaikan nilai parameter menumpu. Akhir sekali, semua kaedah yang telah dicadangkan akan diilustrasi dengan menggunakan data sebenar dalam bentuk sudut yang diperolehi melalui kajian kesusasteraan.

ACKNOWLEDGEMENTS

Alhamdulillah, praise to Allah because I have successfully completed this research. I would like to express my sincere appreciation to my supervisors Prof. Madya Dr. Yong Zulina Zubairi and Prof. Dr. Abdul Ghapor Hussin for their good supervisions, continuous support and encouraging advices throughout the process in completing this research. Thank you very much for being very helpful and supportive.

A special thanks to my parents and husband for being very supportive and understanding throughout the process of completing this research. Without their love and support, it could be very tough journey for me to complete this thesis. Thank you for being there with me all the time.

I deeply thank all my colleague and my dearest friends who always being very supportive and helpful. Thank you for the motivation and advices throughout the completion of my study and thesis. We have shared the moment ups and downs together in the journey of completing this study. Without all the positive vibes from everyone of them, it might be very hard to endure some difficult times through this journey.

Lastly, I would like to thank all visiting lectures that gave comments and help me to improve my research. And, not forgotten, I also would like to thank the University of Malaya and Ministry of Education for providing the scholarship for my study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

The early studies of circular or directional data date back to the mid of the century in the field of astronomy where it was hypothesised that direction of stars were uniformly distributed (Watson, 1983). Books on methods of analysing circular data in the biological field were published include Batschelet (1981), Zar (1984), Upton and Fingleton (1989) and Cabrera *et al.* (1991).

In the last 20 years, vigorous development of statistical methods for analysing circular data can be observed in the general statistical methodology with wide application in various fields.

Circular data, however, are somewhat different from linear data due to the different topologies of the circle and the straight line. Angles are recorded in the range $(0°, 360°]$ degree or $[0,2\pi)$ radian, then the directions close to the opposite end points are near neighbour in a circular metric but maximally distant in linear metric. The statistical theories for line and circle are very different from one to another because the circle is a closed curve but the line is not. In real life, this kind of data can be easily found in the area of study such as:

i. **Meteorology**: wind and wave directions (Mardia, 1972; Bowers & Mould, 2000; Caires & Wyatt, 2003; Hussin *et al.*, 2004; Jammalamadaka & Lund,

2006; Gatto & Jamalamadaka, 2007; Hassan *et al.,* 2010a and Kamisan *et al.,* 2010)

ii. **Medical sciences**: the incidence of onsets of a particular disease at various times of the year (Mardia & Jupp 2000).

iii. **Biology**: bird orientation (Mardia, 1972), animal navigation (Batschelet, 1981)

iv. **Geology**: Azimuths of cross-beds in the upper Kamthi River (Sen Gupta & Rao, 1966)

v. **Geography**: the direction of the earthquake (Rivest, 1997)

vi. **Physics**: interference among oscillations with random phases (Beckmann, 1959)

vii. **Astronomy**: orbit plane that can be regarded as a point on the sphere (as discussed in Watson, 1970)

viii. **Criminology**: time pattern in crime incidence (Brunsdon & Corcoran, 2005)

The circular data cannot be treated as linear data due to its topology. Thus, the needs for statistical analysis as well as making statistical interpretation of this data are really indispensable. For example, as shown in Figure 1.1 for the measurements of wind direction data, the calculated arithmetic mean for 1° and 359° using conventional linear techniques is 180°. On the other hand, by using circular statistics, the mean direction should be 0°. The calculation of the mean direction can be done using the following formula that is totally different from the linear concept.

$$\text{Circular Mean, } \bar{\hat{\mu}} = \begin{cases} \tan^{-1}\left(\dfrac{S}{C}\right) & S > 0, C > 0, \\ \tan^{-1}\left(\dfrac{S}{C}\right) + \pi & C < 0, \\ \tan^{-1}\left(\dfrac{S}{C}\right) + 2\pi & S < 0, C > 0, \end{cases}$$

where $C = \sum \cos\left(x_j\right)$ and $S = \sum \sin\left(x_j\right)$.

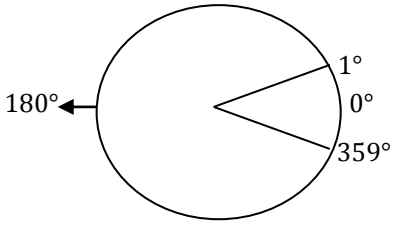

Figure 1.1: Arithmetic mean pointing the wrong way

Here, misinterpretation has occurred by someone making their interpretation without having any idea about the circular data. Making this such interpretation will lead to unbiased and misconception in our analysis. As a consequence, a lot of concepts applied in linear statistics are not quite developed for circular statistics.

The von Mises distribution is often used as the basis for parametric statistics inference and will be used in this study. The von Mises distribution, which is also known as the circular normal distribution, is an analogy to the normal distribution in linear data. This study focuses on the estimation of the concentration parameter in von Mises distribution. In this study, we develop an efficient method to approximate the concentration parameter. Later on, we continue with the approximation of the confidence intervals for the concentration parameter. The study of confidence intervals (CI) for parameter in various distributions has gained a lot of attention recently. As for circular data, few studies were done including by Stephen (1969), Fisher (1993), Khanabsakdi (1995 – 1996), Mardia and Jupp (2000) and Jammalamadaka (2001).

In this study, our focus is to find the CI for the concentration parameter in von Mises distribution. A few methods are developed to achieve this objective. The study begins with the approximation method based on circular population, and this is followed by CI based on the asymptotic distribution of $\hat{\kappa}$, CI based on the distribution of $\bar{\theta}$ and $\bar{R}$ and also CI based on bootstrap-*t* method. All these methods will be used to construct an

efficient CI for the concentration parameter in von Mises distribution. All proposed methods are validated via simulation study and its expected length as well as the coverage probability will be assessed.

Finally, the study also addresses the analysis of missing values for circular data. Missing values is a common problem in data analysis. Deleting or ignoring all missing values, may lead to a lack of statistical power. A few imputation methods have been developed for linear data case, but for circular case, method for handling missing values are still limited. Thus, in this study, two imputation methods are proposed to handle the problem of missing values in circular variables. These two methods are the expectation-maximization (EM) and the data augmentation (DA) algorithm. The analyses are carried out on data that follow von Mises distribution, and the performances of the proposed methods are compared with the conventional method which is the mean imputation method. The biasness are calculated to assess the performance of the proposed method and to identify the most feasible method. Finally, all the proposed methods will be illustrated using real data sets.

1.2 Problem Statement

- The von Mises distribution has two parameters namely the concentration parameter and the circular mean. In estimating the parameter, maximum likelihood estimation (MLE) is often used. For the concentration parameter, the solution of the MLE, however, is analytically intractable because of the presence of modified Bessel functions $I_0(\kappa)$, $I_1(\kappa)$, ... (Mardia, 1972; Batschelet, 1981 and Fisher, 1993). From the literature, the estimations of concentration parameter are given either for small and large concentration parameter only. Therefore, a new and efficient approximation of concentration parameter which applicable for both small and large concentration parameter is needed.
- Most study apply only on simple analysis which is descriptive statistics. The study on inferential statistics, for example, the confidence intervals that can be used in hypothesis testing are relatively few. In circular data, confidence intervals are mostly developed for parameter mean direction only. Therefore, it is necessary to obtain methods for constructing the confidence intervals for concentration parameter.
- Most researchers approach the problem in missing values by deleting or just ignoring them, this may lead to a lack of statistical power. Furthermore, the work on missing values for circular variables are relatively few. Therefore, it is deemed necessary to have methods of addressing the missing value problem for circular data.

1.3 Limitation of the Study

In this study, the simulation study was carried out using varies sample size range from 10 up to 500. However, we publish the simulation results up to 100 only. This is because for large sample size, the results always converge beyond the sample size 100. We did not consider the sample size which is more than 500 because of we have limitation in terms of computational time and limited availability of high performance computer to analyse such large data.

1.4 Objective

The main objective of the study is to propose an efficient confidence intervals for the concentration parameter for the von Mises distribution. The followings are the specific objectives of the study:

i. To develop an efficient method of approximating the concentration parameter in von Mises distribution.
ii. To propose new methods of constructing the confidence intervals for the concentration parameter in von Mises distribution.
iii. To propose a new statistic based on circular distance.
iv. To construct the confidence intervals using a new statistic based on circular distance.
v. To formulate a feasible method of imputing missing values for circular variables.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter presents the literature review that has led to this study. In Section 2.2, a brief explanation of circular statistics and other studies on circular data are discussed as well as its characteristics. Types of circular distributions are discussed in Section 2.3. The studies on confidence intervals for parameter in circular distribution and related studies are discussed in Section 2.4. In Section 2.5, a review of the methods that are used in handling the missing values for linear data, as well as circular data, are given. The details of the source of data and software used in the study are discussed in Section 2.6 under methodology.

2.2 Circular Statistics

Circular data can be defined as the data distributed on the circumference of the circle. This type of data arises in many fields such as earth sciences, meteorology, biology, physics, psychology, image analysis, medicine and astronomy (Mardia & Jupp, 2000). Many examples of circular data were given in the previous chapter. Further reading on circular data can be found in Mardia (1972), Batschelet (1981), Fisher (1993), Mardia and Jupp (2000) and Jammalamadaka and SenGupta (2001).

2.2.1 Numerical Statistics

Let θ_i where $i = 1, \ldots, n$ be observations from a random circular sample of size *n*. Thus, the descriptive statistics for circular data are described as follows.

i. Mean Direction

Each observation of θ_i is considered as a unit vector and the corresponding values of $\cos\theta_i$ and $\sin\theta_i$ are calculated. The resultant length which denoted by *R* is then given by

$$R = \sqrt{C^2 + S^2}\,, \tag{2.1}$$

where $C = \sum_{i=1}^{n} \cos\theta_i$ and $S = \sum_{i=1}^{n} \sin\theta_i$. Thus, the mean direction, denoted by $\bar{\theta}$, is given by

$$\bar{\theta} = \begin{cases} \tan^{-1}\left(\dfrac{S}{C}\right) & \text{if } C \geq 0 \\ \tan^{-1}\left(\dfrac{S}{C}\right) + \pi & \text{if } C < 0 \end{cases}. \tag{2.2}$$

ii. Median Direction

Mardia and Jupp (2000) defined the median as any point ϕ where half of the data lie in the arc $[\phi, \phi + \pi]$, and the other half of points are nearer to ϕ than $\phi + \pi$. On the other hand, Fisher (1993) defined the median direction as the

observation ϕ that minimise the summation of circular distances to all observations,

$$d(\phi) = \pi - \sum_{i=1}^{n} \left| \pi - \left| \theta_i - \phi \right| \right|. \tag{2.3}$$

iii. Mean Resultant Length

Mean resultant length denoted by $\bar{R}$ is defined as the length of the centre of the vector $z = C + iS$. $\bar{R}$ is useful for unimodal data to measure on how concentrated the data is towards the centre. Mean resultant length is given by

$$\bar{R} = \frac{R}{n} \text{ where } 0 \le \bar{R} \le 1. \tag{2.4}$$

The data is said to have small dispersion and more concentrated towards the centre if the value of $\bar{R}$ is close to 1.

iv. Sample Circular Variance

Sample circular variance, denoted by *V*, is given by

$$V = 1 - \bar{R}, \text{ where } 0 \le V \le 1. \tag{2.5}$$

The smaller the circular variance the more concentrated the samples. However, this measure is rarely used in circular statistics in comparison to the measure of the concentration parameter.

v. Sample Circular Standard Deviation

Sample circular standard deviation, denoted by *v*, is given by

$$v = \sqrt{-2\log(1-V)} \text{ where } V = \text{sample circular variance.} \tag{2.6}$$

vi. Concentration Parameter

Concentration parameter, denoted by κ, is the standard measure of dispersion for circular data. Best and Fisher (1981) defined the estimate for the concentration parameter obtained by the maximum likelihood method and it is given as follow

$$\kappa = \begin{cases} 2\bar{R} + \bar{R}^3 + \dfrac{5\bar{R}^5}{6}, & \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + \dfrac{0.43}{1-\bar{R}}, & 0.53 \le \bar{R} < 0.85, \\ \dfrac{1}{\bar{R}^3 - 4\bar{R}^2 + 3\bar{R}}, & \bar{R} \ge 0.85 \end{cases} \tag{2.7}$$

where $\bar{R}$ is mean resultant length. The value of κ lies in the range of $[0,\infty)$. The large value of κ indicates that the observations are highly concentrated in the direction of the mean direction $\bar{\theta}$. If κ is close to 0, it shows that the observations are uniformly distributed around the unit circle.

Unlike the linear data, circular data cannot be analysed directly using the default and built-in function available in many software. Hence, several researchers have developed statistical packages that can be used in that available software. They have written special programs in few commercial software that later on can easily be used by

another researcher. Cox (2001) has developed the programming in Stata that can be used in analysing circular data. Using the package by Cox (2001), the data are assumed recorded in degrees from North. The package consists of four different categories namely utilities, summary statistics and significance tests, univariate graphics and bivariate relationships. Hence, using circular statistics package by Cox (2001) in Stata, the researcher can obtain the descriptive statistics, graphical representation as well as finding the correlation between variables.

Jones (2006) used MATLAB to analysed the directional data. For this purpose, he developed the programs namely Vector_Stats. Specifically, this programs can cater only for two-dimensional directional data such as directions of cross beds. Vector_Stats in MATLAB can be used to calculate the descriptive statistics and generate plots for directional data. The program also provides analysis for single-sample inference on distribution and parameters such as the test of uniformity.

Later on, Berens (2009) developed a MATLAB toolbox for circular statistics namely CircStat. This package includes the descriptive statistics, inferential statistics and measure of association. Apart from that, this toolbox also provides an analysis for data which the underlying distribution is von Mises distribution since this is the most common distribution for circular data. By having this package, the statistical analysis of circular data can be done using MATLAB easily. As for illustration purpose, Berens shows an application on how descriptive statistics can be calculated using neuroscience data.

Lund and Agostinelli were the first who have written programs that is called CircStats package that can be used in analysing directional data in 2007 and later the latest version in 2012 (Lund & Agostinelli, 2012). This package is compatible to use in

R and the S-Plus language. It offers a wide range of circular statistical analysis including the descriptive statistics, graphical representations and inferential statistic. The programs written in this package are based on the description in Jammalamadaka and SenGupta (2001).

Apart of developing the package that compatible in analysing the circular data, there are few other software that offer basic statistical analysis for this type of data. Here, we reviewed the studies that have been done using certain softwares to analyse data. Hussin *et al.* (2006) carried out the circular data analysis using AXIS software (Handeson & Seaby, 2002). AXIS software is an exploratory software that is designed specifically for the directional data. The study focused on how the analysis can be done using the software itself. The analysis included basic summary statistics, circular plot, testing for uniformity or randomness and the comparison between samples. As for illustration purpose, they analysed two different Malaysian wind data set namely Southwest and Northeast.

ORIANA (Kovach Computing Services, 2009) is one of the commercial softwares that offers a basic analysis of circular data. This kind of software is user-friendly as it does not need ones to do the programming in order to carry out the analysis. ORIANA can be used to display the basic summary statistics and it can be very useful for the circular graphical representation as it offers a number of circular plots such as rose diagram, circular histogram, raw data plot and arrow data plot. Testing for uniformity and comparisons between samples also can be done using ORIANA. Hassan *et al.* (2009) has carried out the analysis of Malaysian wind data using ORIANA and discussed the features that are available to be used for circular statistical analysis.

From the studies that were described previously, it showed that this study has gained prior attention from researchers in various fields. This is due to the importance and the wide application of the circular data in many fields such as astronomy, geology, zoology, neuroscience and medical research.

2.2.2 Graphical representation

The graphical representations are often used, to summarise the data and to explore the circular samples. It also useful for the purpose of detecting outliers in circular sample. Below are types of graphical representation that are available for circular samples:

i. Q-Q Plot

Q-Q plot allows us to compare the distribution of two samples. For circular data, the Q-Q plot can be obtained using ORIANA software and using the CircStats package in R or S-Plus.

ii. Spoke Plot

Spoke plot is one of the graphical representation that is specifically useful for circular data. Zubairi *et al.* (2008) used the MATLAB software to develop this plot. This plot is useful for getting a general pattern of the linear relationship between two circular variables as well as calculating the linear correlation. It consists of inner, θ_i and outer, φ_i rings where $0^\circ \le \theta_i, \varphi_i < 360^\circ, i = 1, 2, \ldots$ in

which lines are used to connect the pairs of points (θ_i, φ_i). As an illustration of this plot, three different Malaysian wind data sets were used, and the correlation, as well as its linear relationship, were calculated in their study.

iii. Circular Boxplot

Boxplot is a common plot in real line data set and has been widely used in exploratory data analysis. Boxplot is useful to identify the existence of outliers in the sample. This type of boxplot, however, is not suitable for circular data due to a different topology of the circular data itself. Therefore, Abuzaid *et al*. (2012) has developed the circular version of boxplot namely as circular boxplot. In order to develop the circular boxplot, median direction, quartile of circular variables and circular interquartile range and fences are calculated.

2.3 Circular Distribution

A circular distribution is a probability distribution that the total probability is concentrated on the circumference of a unit circle. Each point located on the circumference represents a direction. The circular variables, μ is measured in radian and in the range of $[0, 2\pi)$ or $[-\pi, \pi)$.

2.3.1 Uniform Distribution

The uniform distribution which denoted as U_c is a basic distribution on the circle. In this distribution, all directions are equally likely. The probability density function (pdf) is given by:

$$f(\theta) = \frac{1}{2\pi}, \text{ where } (0 \le \theta < 2\pi). \tag{2.8}$$

While the distribution function is given by

$$F(\theta) = \frac{\theta}{2\pi}, \text{ where } (0 \le \theta < 2\pi). \tag{2.9}$$

For circular uniform distribution, the mean direction, μ is undefined and having mean resultant equal to 0. This distribution plays an important role in circular analysis because they represent the state of 'no mean direction' (Jammalamadaka & SenGupta, 2001).

2.3.2 Von Mises Distribution

In modelling the circular data, the widely used distribution on the circle is the von Mises distribution. This distribution is also known as the circular normal distribution, and it is an analogue to the normal distribution on the real line. The von Mises distribution is denoted by $VM(\mu, \kappa)$ where μ is the mean direction while κ is

the concentration parameter. The probability density function of von Mises distribution (Mardia & Jupp, 2000) is given by

$$f(\theta) = \frac{1}{2\pi I_0(\kappa)} \exp\left(\kappa \cos(\theta - \mu)\right),\ 0 \le \theta < 2\pi,\ 0 < \kappa < \infty, \tag{2.10}$$

where I_0 denotes the modified Bessel function of the first kind and order 0 and can be defined as:

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \exp(\kappa \cos\theta)\, d\theta. \tag{2.11}$$

Von Mises distribution is the most common distribution considered for unimodal samples for circular data. Some of its density properties are:

i. it is symmetrical about the mean direction μ

ii. it has a mode at μ

iii. it has anti mode at $(\mu \pm \pi)$.

The limiting forms for this distribution as given by Fisher (1993),

i. as $\kappa \to 0$, the distribution will converge to the uniform distribution, U_c

ii. as $\kappa \to \infty$, the distribution tends to the point distribution concentrated in the direction of μ.

2.3.3 Wrapped Normal Distribution

The wrapped normal distribution can be obtained by wrapping a normal distribution around a unit circle. This distribution is a symmetric unimodal two parameter distribution. The wrapped normal distribution is denoted by $WN(\mu,\rho)$ where μ is the mean direction while ρ is the mean resultant length. The probability density function is given by:

$$f(\theta)=\frac{1}{2\pi}\left(1+2\sum_{p=1}^{\infty}\rho^{p^2}\cos(\theta-\mu)\right),\ 0\le\theta<2\pi,\ 0\le\rho\le1. \tag{2.12}$$

If $f(\theta)$ is obtained by wrapping a normal distribution with variance σ^2, then

$$\rho=e^{-\frac{1}{2}\sigma^2},\quad \text{or}\quad \sigma^2=-2\log\rho. \tag{2.13}$$

The limiting forms for this distribution as given by Fisher (1993),

i. as $\rho\to0$, the distribution will converge to the uniform distribution, U_c

ii. as $\rho\to1$, the distribution tends to the point distribution concentrated in the direction of μ.

2.3.4 Wrapped Cauchy Distribution

The wrapped Cauchy distribution can be obtained by wrapping the Cauchy distribution on a real line with a density

$$f(\theta)=\frac{1}{\pi}\left(\frac{\sigma}{\sigma^2+(\theta-\mu)^2}\right),\ -\infty<\theta<\infty \tag{2.14}$$

around the circle. This distribution is a symmetric unimodal two parameter distribution. The wrapped Cauchy distribution is denoted by $WC(\mu,\rho)$ where μ is the mean direction while ρ is the mean resultant length. The probability density function is given by

$$f(\theta)=\frac{1}{2\pi}\left(\frac{1-\rho^2}{1+\rho^2-2\rho\cos(\theta-\mu)}\right),\ 0\le\theta<2\pi,\ 0\le\rho\le1. \tag{2.15}$$

The limiting forms for this distribution as given by Fisher (1993),

i. as $\rho\to0$, the distribution will converge to the uniform distribution, U_c

ii. as $\rho\to1$, the distribution tends to the point distribution concentrated in the direction of μ.

2.4 Confidence Intervals

Confidence intervals can be defined as an interval estimate of the point estimate or the parameter itself. As in Efron and Tibshirani (1993), knowing the interval estimate with its point estimate can say what the best guess is for θ, and how far in error that guess might be. In the perspective of linear statistics, this area has gained prior attention from many researchers. Many new and integrated approaches were developed to obtain an efficient approximation for confidence intervals based on different methods such as

confidence interval based on hypothesis testing, confidence interval based on bootstrap method which include percentile bootstrap, bootstrap-*t* and iterated bootstrap.

Expected length and coverage are usually used to assess the superiority of the proposed methods to approximate the confidence intervals. Expected length is defined as the class size of the estimated intervals. The coverage probability is the proportion of times that the estimated intervals cover the true parameter. Nominal coverage also known as target value is the confidence level that we consider when approximating the confidence intervals. Coverage error that is defined as the absolute difference between the nominal and actual coverage probability, is often used to assess the superiority of confidence intervals. The reliability of confidence interval is determined by its coverage (Letson & McCullogh, 1998). A good confidence interval should have a coverage probability that is close to a target value (nominal coverage) as well as having small coverage error. In the next section, the bootstrap method, a method that is widely used in constructing the confidence interval will be discussed.

2.4.1 Bootstrap Method

Bootstrap method was proposed by Efron (see Efron, 1979, 1987 and Efron & Tibshirani, 1993) and has gained so much attention and acceptance from researchers in various fields of study. The bootstrap method is procedures that create a number of sub-samples from a pre-observed data set by a simple random sampling with replacement. The sub-samples is then to be used in investigating the nature of the population without having any assumption about them. For the past few years, many studies have been developed on the bootstrap technique and confidence intervals in various research areas

(see Hall, 1986, 1987, 1988; Hall & Martin, 1988; DiCiccio & Efron, 1996; Letson & McCullogh, 1998; Polansky, 2000). This computer-based method is very useful in estimating the standard error and bias as well as approximating confidence intervals and other statistical measures (see Efron & Tibshirani, 1986).

How many bootstrap replications need to be considered in order to get a good estimation always becomes a question among researchers. Efron (1979) gives $B = n^n$ as the possible bootstrap replication. To estimate the standard error, 25 to 250 bootstrap replications usually considered. But, for another estimate such as confidence intervals, number of bootstrap replication will be increased. Bootstrap replications are dependent upon the value of X if $n \leq 100$ and the bootstrap replications is taken to be $B \leq 10000$ (Efron, 1979). As for circular distribution, Fisher (1993) takes number of bootstrap replication, $B = 200$ to approximate the confidence intervals for the mean direction.

In conclusion, Efron and Tibshirani (1993) gives rules of thumb in determining how many replications should be considered when doing the resampling:

i. Small number of bootstrap replication, for example, $B = 25$, is usually informative. $B = 50$ is considered as enough to give a good estimator of standard error.
ii. Very seldom $B > 200$ bootstrap replications needed in estimating a standard error. The number of replications generally in the range of 25 to 2000.

Efron and Tibshirani (1993) and Chernick (1999) give a comprehensive explanation on constructing the confidence interval based on several bootstrap methods. In this subsection, some reviews on confidence intervals based on various types of the

bootstrap method that motivate the study on confidence intervals in circular distribution are discussed.

Porter *et al.* (1997) and Polansky (2000) studied on bootstrap-*t* (see Efron, 1982) confidence intervals for small sample size. Porter *et al.* (1997) applied the non-parametric bootstrap *t* method to construct the confidence intervals for the mean parameter of an unknown distribution. The non parametric bootstrap is a distribution-free method where the original sample of size *n* is resampled *N* times with replacement. The study showed that the bootstrap-*t* is superior to the other method considered which is the Student's-*t.* The superiority of the method is assessed by the coverage probability of each method. The Student's-*t* gives a coverage probability that is less than the nominal as opposed to bootstrap-*t* which has a better coverage probability.

Letson and McCullogh (1998) discussed on different types of the bootstrap method in approximating confidence interval. Five different types of bootstrap techniques are considered which include single and double bootstrap. The performance of each method is assessed by its coverage error. Coverage error is defined as the absolute difference between the nominal coverage (target value) and the actual coverage (coverage probability). Shi's double bootstrap method is said to be superior because it gives good coverage in comparison to single bootstrapping methods.

Besides that, Polansky (2000) carried out the study on stabilizing the end points of bootstrap-*t* intervals, as well as its coverage error. This study was done for small sample size. The objective of his study is to improve the stability of the bootstrap-*t* method and preserve the coverage error. This is because the bootstrap-*t* is known to have a good coverage error. This method is said to be better as opposed to the

estimated-variance-stabilizing method by Tibshirani (1988) which only reduces the expected length but increases the coverage error.

In directional data, the bootstrap method in approximating the confidence intervals for parameter in circular distribution was developed by Ducharme (1985), Fisher and Hall (1992) and Fisher (1993). The details on confidence intervals for parameter in circular distribution will be discussed in the next subsection.

2.4.2 Confidence Intervals for Parameter in Circular distribution

In this subsection, the review on confidence intervals for parameter in circular distribution will be discussed. As explained in the previous subsection, the von Mises distribution is the most common distribution occurs in circular data, and it has two parameters namely the mean direction and concentration parameters.

Firstly, a review on confidence intervals for the mean direction will be discussed. Batschelet (1981) in his book has given the calculation of confidence intervals for mean angle with the samples were drawn from a von Mises distribution. For this purpose, calculation of mean vector length *r* and mean angle of the sample $\bar{\phi}$ are required. The angle of deviation δ will be determined based on *r* and the sample size *n* which can be referred from the chart given by Stephen (1962a & 1962b) and Brown and Mewaldt (1968). From their study, a $100(1-\alpha)\%$ confidence interval for the mean is $(\bar{\phi}-\delta, \bar{\phi}+\delta)$.

A Monte Carlo simulations study was carried out by Ducharme *et al.* (1985) to compare the performances of the bootstrap method with the other methods. A total of six different methods and non-parametric confidence cones for the mean direction based on the bootstrap method have been considered. As for performance indicator, coverage probabilities were assessed to compare the superiority of the method. Samples were drawn from different distributions, and they are considering bootstrap replication which is $B = 200$.

Upton (1986) has given the approximation for confidence interval for the mean direction in von Mises distribution. Two proposed methods in approximating $100(1-\alpha)\%$ of CI were discussed and conclude that the new methods are preferred for smaller value of n and $\bar{R}$. The methods of approximation $100(1-\alpha)\%$ of CI for the mean direction are given as follows:

i. Likelihood-based, $\bar{R} \leq 0.9$

$$\left(\hat{\mu} \pm \arccos\left[\frac{\sqrt{\frac{2n\left(2R^2 - nZ_\alpha\right)}{4n - Z_\alpha}}}{R} \right] \right)$$

ii. Likelihood-based, $\bar{R} \geq 0.9$

$$\left(\hat{\mu} \pm \arccos\left[\frac{\sqrt{n^2 - \left(n^2 - R^2\right)\exp\left(\frac{Z_\alpha}{n}\right)}}{R} \right] \right)$$

where Z_α is the α point of a χ_1^2 distribution.

Fisher (1993) described the steps for finding the confidence intervals for the mean direction based on the bootstrap method. He suggested three different techniques to determine the final confidence intervals for the mean direction. The basic method (Technique 1) to obtain a $100(1-\alpha)\%$ confidence interval for the unknown mean direction as given by Fisher

i. calculate $\gamma_b = \hat{\mu}_b^* - \bar{\theta}$, $(-\pi \le \gamma_b < \pi)$, $b = 1,\ldots,B$

ii. sort into increasing order to obtain $\gamma_{(1)} \le \ldots \le \gamma_{(B)}$

iii. let l = integer part of $\left(\frac{1}{2}B\alpha + \frac{1}{2}\right)$ and $m = B-1$. Thus, the confidence intervals for μ will be given as:

$$\left(\bar{\theta} + \gamma_{(l+1)}, \bar{\theta} + \gamma_{(m)}\right).$$

Zar (1999) has given calculations for confidence intervals for the mean direction. He considered two cases as follow

i. for $R \le 0.9$ and $R > \sqrt{\frac{\chi_{\alpha,1}^2}{2N}}$,

$$d = \arccos\left[\frac{\sqrt{\frac{2N\left(2R_n^2 - N\chi_{\alpha,1}^2\right)}{4N - \chi_{\alpha,1}^2}}}{R_n}\right].$$

ii. for $R > 0.9$

$$d = \arccos\left[\frac{\sqrt{N - \left(N^2 - R_n^2\right)\exp\left(\frac{\chi^2_{\alpha,1}}{N}\right)}}{R_n}\right].$$

where $R_n = R \cdot N$. Considering both cases, the $100(1-\alpha)\%$ confidence interval for the mean direction is given by $(\bar{\alpha} - d, \bar{\alpha} + d)$.

On the other hand, Jammalamadaka and SenGupta (2001) discussed on the construction of confidence intervals for the mean direction based on circular 'standard error' of the MLE for $\hat{\mu}$ in von Mises distribution. This method is applicable for large samples where $\hat{\sigma}_{\hat{\mu}} = \frac{1}{\sqrt{n\bar{R}\hat{\kappa}}}$. Hence, a $100(1-\alpha)\%$ confidence interval for the mean direction is given by

$$\left[\hat{\mu} - \arcsin\left(\tau_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\mu}}\right), \hat{\mu} + \arcsin\left(\tau_{\frac{\alpha}{2}}\hat{\sigma}_{\hat{\mu}}\right)\right].$$

Otieno and Anderson-Cook (2006b) discussed on three different bootstrap methods (Fisher & Hall, 1989) to estimate the confidence intervals for the preferred direction for a single population. The preferred direction that were used are the mean direction, the median direction and the Hodges-Lehman estimate (Otieno & Anderson-Cook, 2006a). A comparison study was carried out using three different types of the bootstrap technique and the performance of the methods were assessed by their coverage probability and expected length.

As for the concentration parameter in von Mises distribution, an early study was carried out by Stephens (1969), where several approximation of confidence for small concentration parameter were proposed. Apart from that, Batschelet (1981) has discussed on the confidence interval for the concentration parameter based on the chart given in Stephen (1962a & 1962b) and Brown and Mewaldt (1968).

Confidence intervals for the concentration parameter based on percentile bootstrap method can be found in Fisher and Hall (1992) and Fisher (1993). Steps on percentile bootstrap will be discussed in Chapter 4, and this method is used in the comparison study with our proposed methods.

Khanabsakdi (1995) proposed confidence intervals for the concentration parameter based on chi-squared variable. A $100(1-\alpha)\%$ confidence intervals for circular variance is given by

$$\frac{ns^2}{\chi^2_{\frac{\alpha}{2},\upsilon}} < \sigma^2 < \frac{ns^2}{\chi^2_{1-\frac{\alpha}{2},\upsilon}}$$

The lower and upper limits of the population circular variance are then to be transformed to lower and upper limits of the concentration parameter. This transformation is obtained using the relation between length of the sample mean vector *r*, sample circular standard deviation *s* (Batschelet, 1981) and the concentration parameter, κ. Comparison with the Stephen's formula has been carried out, and it showed that the method is more efficient than the previous one. However, this method is said to be applicable only when the data is highly concentrated.

2.5 Missing Values

Missing values is one of the problems that always occur in data analysis. This problem must be taken seriously because by ignoring or deleting the missing values that exist in our data will affect the statistical power and may lead to the biasness (Little & Rubin, 2002; Tsikriktsis, 2005). Nowadays, the studies of missing data were extensively done by researchers from various fields such as medical (Enders, 2006), environmental (Norazian *et al.*, 2008) and psychology (Baraldi & Enders, 2010) due to the importance of obtaining the complete data analysis.

There are several classifications of missing values. These classifications influence the optimal strategy for working with missing values. Little and Rubin (2002) gave the classification of missing values as follows.

i. Missing completely at random (MCAR)

MCAR occurs when the probability of missing data on a variable X is unrelated to the other measured variables and the values of X itself. MCAR is said to be not really suitable in practice because of the strict assumption that requires the missingness to be unrelated to the study variables (Raghunathan, 2004)

ii. Missing at random (MAR)

MAR occurs when the missingness is related to the other measured variable in the analysis, but not to the underlying values of incomplete variable. MAR is described as systematic missingness where the tendency for missing data is

correlated with other study-related variables in the analysis. This type of missingness is said to be most likely happen in real practice as it requires less stringent assumption about the reason for missing data (Baraldi & Enders, 2010).

iii. Missing not at random (MNAR)

The data is said to be MNAR if the probability of missing data is systematically related to the hypothetical values that are missing. It also can be described as the data that are missing based on the would-be values of the missing observations.

In the next subsection, the reviews on methods of handling the missing data are discussed. Further reading on missing data analysis can be found in Schafer (1997), Allison (2002), Little and Rubin (2002) and Baraldi and Enders (2010).

2.5.1 Traditional Approaches in Handling the Missing Values Problems.

Many extensive studies have been done in handling the data set with missing values problems. Before begin with the analysis, one should understand the nature of the missingness that occurs in the data. This is important in order to choose the best technique that can be applied in analysing such data. In this section, some reviews on traditional approaches in handling missing values problems are discussed.

The most common traditional approaches are deletion and replacement procedures (Peugh & Enders, 2004). The easiest way to handle missing values is by

deleting those observations with missing values and consequently lead up to 'complete' analysis. This is the default method that is usually used in most statistical and computer analysis package including SPSS, S-Plus, R and SAS. Despite simple and easy, this approach will decrease the sample size of the data and at the same time can reduced the power of statistics (Little & Rubin, 2002).

There are two types of deletion techniques. The first technique is the listwise deletion that also known as the complete-case analysis or case-wise deletion. Many reviews on this technique from the past researchers in various fields including Kim and Curry (1977), Tsikriktsis (2005), Peugh and Enders (2004) and Baraldi and Enders (2010). This type of approach tend to be the choice among the researcher because of the simplicity of the method itself whereby it produce the complete data set and the statistical analysis can be done by using standard analysis techniques (Baraldi & Enders, 2010). As stated in Kim and Curry (1977), this method eliminates from further analysis all cases with any missing data. As a result, it gives a large effect in the data where randomly deleting 10% of the data from each variable in a matrix out of five variables can easily cause an elimination of 59% of cases from analysis. In addition, by using listwise deletion it gives a conservative estimate of the parameters and lead to conservative results. By reducing the sample size, it may decrease the statistical power. Hence, this will lead to lack of statistically significant (Tsikriktsis, 2005; Baraldi & Enders, 2010). By using this method, the data are assumed as MCAR, which missingnes is unrelated to all measured variable.

The second technique in deletion procedure is pairwise deletion that is also known as available-case analysis. Pairwise deletion is an alternative approach over the listwise deletion especially for linear models. According to Allison (2002), pairwise deletion is more efficient than listwise deletion because more data are considered in

producing the estimates. Similarly, Baraldi and Enders (2010) stated that this pairwise deletion is an improvement over listwise deletion whereby it minimizes the number of cases discarded during the analysis. Monte Carlo studies have shown that listwise deletion gives less accurate estimates of population parameters such as correlations. Pairwise deletion is consistently more accurate than listwise deletion though the differences can sometimes be small (Hippel, 2004; Acock, 2005; Tsikriktsis, 2005).

Other conventional or traditional methods that were always applied by researchers are simple replacement procedure or also known as single imputation method. There are few types of single imputation method including mean imputation, regression imputation, hot deck imputation and many more. The method of imputation uses the idea of fills in the missing values with some possible value. The most common single imputation method is mean imputation. Usually, this type of imputation is simply can be applied while doing the analysis using most of the available statistical softwares such as SPSS, R and S-Plus.

In mean imputation method (Winkler & McCarthy, 2005; Tsikriktsis, 2005; Saunders *et al.*, 2006, Baraldi & Enders, 2010; Hassan *et al.*, 2010b), all missing values will be replaced with the mean of all available observations. Norazian *et al.* (2008) applied two different types of mean imputation methods namely mean-before-after method and mean-before method. It was found that mean-before-after gives the best result for predicting missing values. The idea of using the mean substitution may be based on the fact that the mean is a reasonable guess of a value for a randomly selected observation from a normal distribution. However, with missing values that are not strictly random, the mean substitution may be a poor guess. According to Acock (2005), mean substitution is especially problematic when there are many missing values. For example, if 30% of the respondents do not answer the question, it means that there are

30% of missing values. If a mean sample is substituted for each of them, then 30% of the sample has zero variance on that data. In this method, any type of missingness (regardless of whether data are MCAR, MAR or MNAR) will lead to biasness of any parameter except the mean (Peugh & Enders, 2004).

The second approach in replacement procedure is known as hot-deck imputation (Tsikriktsis, 2005; Saunders *et al.*, 2005). By using this imputation method, we will replace a missing value with the actual score from a similar case in the data set. In this procedure, a correlation matrix is used to determine the most highly correlated variables. Hot deck imputation works very well in the large samples whereby a similar case is easily to identify. However, in order to apply this method, the programming must be written because it is not one of the built-in function in the most common statistical software.

Another method of replacement procedure is regression imputation. Other studies discussed on regression imputation include Hippel (2004), Tsikriktsis (2005), Winkler and McCarthy (2005), Saunders *et al.* (2006) and Baraldi and Enders (2010). Regression imputation used the idea of substitutes missing values with predicted scores from a regression equation. In comparison to the previous techniques discussed, regression imputation uses the most sources of information to predict the missing values and provide better estimation for missing values. The steps in this approach involve estimating the relationship between the variables, and then uses the regression coefficients to estimate the missing value. The underlying assumption of regression imputation is the existence of a linear relationship between the predictors and the missing variable. Despite the strategy of using information from the complete variables is good. However, this imputation method also produced biased parameter estimates (Baraldi & Enders, 2010). Apart from that, as in Winkler and McCarthy (2005) and

Saunders *et al.* (2006), the problem might occur during the regression imputation. This is because of the difficulty to work out with the equation and also the correlation between the variables may be weak or different relationship may exist. However, this method is easy to apply to linear data since it is already defaults in certain statistical software packages.

Both imputation techniques that are the mean imputation and regression imputation seems to be much better rather than the deletion process. However, they are still lead to biasness because they fail to account for the variability that is present in the hypothetical data values (Baraldi & Enders, 2010). Hence, the researcher makes an attempt to introduce the modern missing techniques in order to provide better estimates as well as reduced the biasness in estimating the parameter for missing value cases. In the next subsection, the modern missing values techniques will be reviewed.

2.5.2 Modern Techniques in Handling the Missing Values Problem.

Apart of traditional approaches, there are a few modern approaches, and some of them are integrated from the traditional approach. Multiple imputation and maximum likelihood can be considered as the modern techniques in handling the missing values problem. Baraldi and Enders (2010) gives a good introduction to modern approaches that can be used for missing data analysis.

Multiple imputation is done by creating several copies of data set, and each of them consists of different imputed values. Many studies were applied the multiple imputation techniques in handling the missing data previously including Acock (2005),

Schafer and Schenker (2000), Kofman and Sharpe (2000), Barzi and Woodward (2003), Junninen *et al.* (2004), Sartori *et al.* (2004), Enders (2006), Tsechansky and Provost (2007), Baraldi and Enders (2010), Johnson and Young (2011). Generally, the multiple imputation is done using three steps or phases as follows:

i. Imputation phase

In this stage, specified number of data sets are generated where each of them consists different estimate of the missing values. According to Graham *et al.* (2007), 20 data sets are a good rule of thumb to be followed.

ii. Analysis phase

In this phase, the complete data sets are obtained. Hence, statistical analysis will be carried out using the same techniques.

iii. Pooling phase

Pooling phase is where all parameter estimates yield from different data sets will be gathered. Pooled parameter estimates were calculated by taking the average over all estimates. Rubin (1987) has given the formula on pooling the parameter estimates and standard errors.

Baraldi and Enders (2010) illustrated the multiple imputation in their study and concluded that this modern technique show some improvement in comparison to traditional approaches. Further details on multiple imputation can be found in Rubin (1987), Schafer (1997), Sinharay *et al.* (2001), Allison (2002) and Little & Rubin (2002).

Expectation-Maximization (EM) is the algorithm that can be used in maximizing the likelihood of a variety of missing data models. EM algorithm was first introduced by Dempster *et al*. (1977). In simplest way, this algorithm can be defined as 'fill in' the missing data based on the initial estimate, re-estimate the parameter based on available data and then fill in again iteratively till the estimates converge. It can be done in single imputation or integrated to get the better estimate by performing the multiple imputation. EM using the maximum likelihood approach where a new data set was created in which all missing values are imputed with maximum likelihood values. There are two steps in EM algorithm:

i. Expectation (E-step)

The E-step of EM is replacing the missing values observations, X_{mis} which require the estimation of $\theta^{(t)}$ to obtain complete data, when X_{obs} is given.

ii. Maximization (M-step)

In this step, $\theta^{(t+1)}$ is re-estimated by maximum likelihood based on X_{obs} and $\theta^{(t)}$ obtained in Step (1).

Steps (i) – (ii) will be repeated iteratively until $\theta^{(t)}$ and $\theta^{(t+1)}$ satisfied the convergence criteria and converge to a local maximum of the likelihood function.

By using EM single imputation, it tends to underestimate the standard error and thus lead to inaccurate estimation (Schafer, 1997). Otherwise, multiple estimation

allows pooling of the parameter estimates to obtain improved parameter estimates where it gives a different solution for each imputation. The steps in doing multiple imputation using EM algorithm can be found in Little and Rubin (2002), Sartori *et al.* (2004) and Acock (2005).

According to Barzi and Woodward (2004), from different imputation they used in their study, Expectation Maximization (EM) method is the most appealing because it allows any type and number of variables as well as gives the most reliable variance of estimate. EM is known to yield estimates with theoretical properties that the other imputation methods do not provide when the missing at random assumption is satisfied. EM requires specifying a joint probability distribution for the variable to be imputed and the predictor variables, and it provides maximum likelihood estimates in the presence of missing data. However, as the percentage of missing values over than 60%, there is no imputation method can be used to get a better estimation.

Junninen *et al.* (2004) compared of several techniques that can be used in handling missing values problems. Different methods were described in the study namely the regression based imputation, self-organizing map, multiple imputation and a hybrid model. The performance indicator was calculated to evaluate the accuracy of each method. Based on the results, he concluded that the method improved by a hybrid approach and multiple imputation method can be the chosen as the best methods.

Another modern technique that can be considered is data augmentation (DA) algorithm. DA algorithm was first proposed by Taner and Wong (1987). There are two steps in this method namely I-step and P-step. Briefly, the steps are described below:

i. Imputation (I-step)

Given a current guess of a parameter as $\theta^{(t)}$, draw independent *q* values of X_{mis}

$$X_{mis}^{(t+1)} = \left(x_{mis1}^{(t+1)}, x_{mis2}^{(t+1)}, \ldots, x_{misq}^{(t+1)} \right)$$

generated from the conditional predictive distribution of X_{mis}

$$X_{mis}^{(t+1)} \sim P\left(X_{mis} \mid X_{obs}, \theta^{(t)} \right).$$

ii. Posterior (P-step):

Draw new values of θ

$$\theta^{(t+1)} = \left(\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{n_0}^{(t+1)} \right)$$

which is calculated from the conditional distribution of X_{obs} and $X_{mis}^{(t+1)}$

$$\theta^{(t+1)} \sim P(\theta \mid X_{obs}, X_{mis}^{(t+1)}).$$

Steps (i) – (ii) will be repeated from the initial value $\theta^{(0)}$ for a value of *t*

$$\left\{ \left(\theta^{(t)}, X_{mis}^{(t)} \right) : t = 1, 2, \ldots \right\}$$

DA algorithm also can be applied in order to carry out the multiple imputation. I-step will be carried out in *Imputation Phase* to generate few different imputed values. Baraldi and Enders (2010) have carried out the multiple imputation using DA algorithm, and it showed that the estimates are better in comparison with the traditional method. As

for linear data analysis, statistical analysis can be done via selected software that offers this type of analysis such as S-Plus and R.

In circular or directional statistic field, there is still limited software that available for analysing such data. Hence, the programming needs to be written in order to do the statistical analysis for missing values with this kind of data. In this study, our objective is to propose the methods that can be used in handling the missing values problems for circular data. Three different techniques which include the conventional, as well as the modern techniques, will be discussed, and the performance of each method will be evaluated.

2.6 Methodology

This study can be divided into four parts. The first part begins with the development of a new estimation for the concentration parameters in the von Mises distribution. It is continued with proposals of a new approximation for confidence intervals (CI) for the concentration parameter. In the second part of the study, formulation of the four different proposed methods is given. Simulation studies are carried out to evaluate the performance of each proposed method and later on some illustrations of the proposed method is given by applying the method on real data sets. This is followed by the third part, in which a new measurement of circular distance is derived. This statistics is then used in the approximation for chi-squared distribution. Based on circular distance, a new CI is derived, and simulation study is performed to measure the performance of the three proposed methods. The final part of this study focuses on how to handle the missing values problem that occurs in univariate circular

data. Three imputation methods are considered in the study, and the performance are assessed via simulation study.

2.6.1 Source of Data

As illustrations of the proposed methods, a real data set obtained from Meteorological Service Department is used. Also, secondary data found in the literature are also used. Following are the descriptions of the data used. A list of the data sets can be found in the Appendix A and B respectively.

i. Kuala Terengganu (Wind direction data)

Daily wind direction data (in radian) recorded at maximum wind speed (in m/s) were considered. The data were collected at an altitude of 2.8 m to 40.1 m in the year 2004. A total of 50 measurements were recorded for the annual northeast monsoon in Kuala Terengganu station. All measurements were obtained from the Malaysian Meteorology Service Department.

ii. Humberside Coast (Wind and wave direction data)

In addition to the local data, the study utilises the data set collected along the Holderness Coastline, which is the Humberside coast of the North Sea, United Kingdom in October 1994. There were 85 measurements recorded over a period 22.7 days. For this purpose, four different data were recorded:

a. wind and wave direction measured by HF radar.

b. wind and wave direction measured by anchored wave buoy.

2.6.2 Software

In this study, several mathematical and statistical softwares were used. S-Plus is the main software used in this study. S-plus was first developed in the mid-1970s. Since then, it has undergone many changes. S-Plus language is designed to be flexible, and it is an interactive software for data analysis. In this study, programming are developed in S-Plus for analysis purposes and in carrying out the simulation studies were carried out to assess the accuracy of the proposed methods.

Another software that is used in this study is ORIANA software, one of the softwares designed specifically for circular data. ORIANA was first introduced on 31st December 2003 and the latest version of ORIANA 4.0 is updated on 16th May 2012. It can perform the basic statistics such as circular mean, median, mean vector length, concentration parameter, circular variance, standard deviation and also the confidence intervals for the mean. In addition, ORIANA is able to display several types of graphical representations such as rose diagram, linear histogram, raw data plot and many more. Distribution plots for comparing the data to certain distribution, scatter plot for preliminary data analysis and Q-Q plot in order to compare the distributions of two samples are also available. In this study, ORIANA is used to plot the circular data graphically.

The study also utilises Minitab to obtain a plot of linear graphical representations. Minitab is a statistical package developed by researchers in 1972 and widely use for statistical data analysis. In this study, Minitab version 16 is used.

2.6.3 Flow Chart of Research Design of the Study

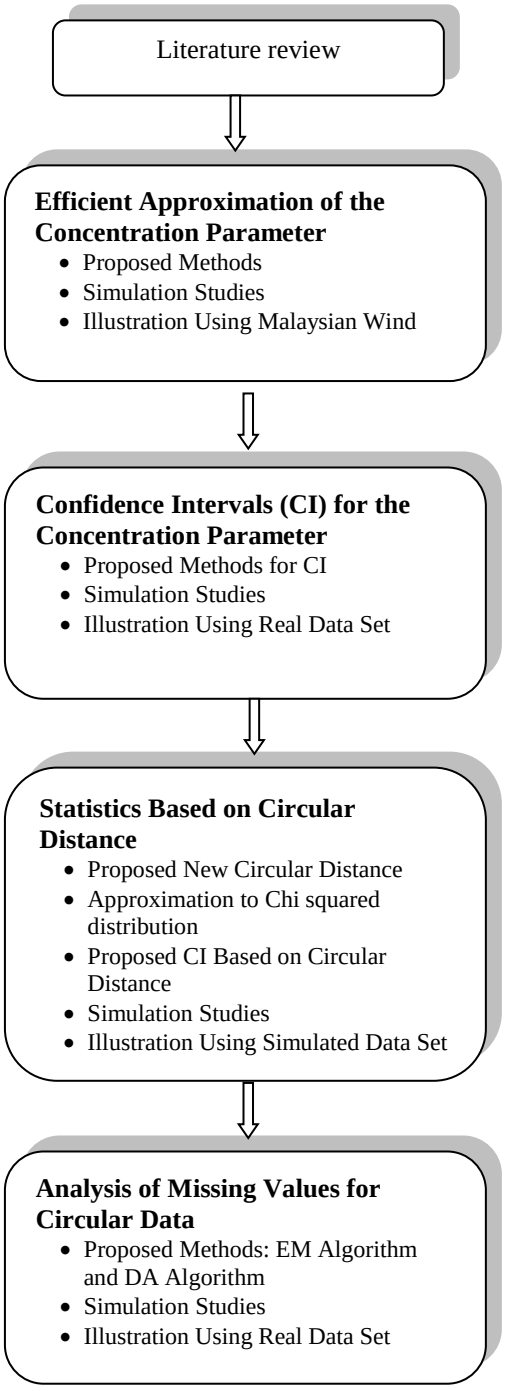

Figure 2.1: Flow chart of research design of the study

The study begins with the development of a new efficient approximation for the concentration parameter in von Mises distribution. New approximation methods based on modified Bessel function are proposed. The simulation studies are carried out to assess the performance of the proposed method with other methods. In the later part of the section, the new proposed method is then illustrated using real data set found in the literature.

Confidence intervals (CI) for concentration parameter in von Mises distribution are considered in the following section. Four different methods in approximating CI are proposed. To assess the accuracy of the proposed methods, a simulation study using S-Plus is carried out. Expected length and coverage probability are used to assess the performance of each method. Again, a simulation study is carried out in order to check on the performance and stability of each method. As for graphical presentation, two different softwares namely ORIANA and Minitab are used. ORIANA is used in plotting any graphical representation related to circular data while Minitab is used in plotting linear graphical representations.

In the next section, an approximation for circular distance is proposed. Based on the circular distance itself, a new CI is derived. Simulation study is carried out to identify the best percentile to get the most efficient CI for concentration parameter.

The last part of this study focuses on the analysis of missing values for circular data. This type of analysis has been well developed for the linear data, but there is somewhat limited study of circular data due to the complexity of the circular data itself. The analysis can be complicated by the fact of the characteristic of circular data itself. Thus, specific tools must be used to handle the analysis of the data. The analysis focuses on the appropriate procedure to deal with missing values. In this study, a few imputation methods for circular data are considered. The first method is known as the circular Expectation-Maximization (EM) algorithm and the second method is data augmentation

(DA) algorithm. Both methods are compared with the current method that is the circular mean. Simulation studies will be carried out to assess the performance of each method and the findings from each study are discussed. Finally, all proposed methods are illustrated using the wind and wave direction data.

CHAPTER 3

IMPROVED EFFICIENT APPROXIMATION OF CONCENTRATION PARAMETER FOR VON MISES DISTRIBUTION

3.1 Introduction

This chapter discusses a few improved approximation methods of the concentration parameter for von Mises distribution. In Section 3.2, a brief introduction of the parameter estimation for von Mises distribution is given. Details of the proposed approximation methods are given in Section 3.3. To assess the accuracy of the proposed approximation methods, simulation studies are carried out in Section 3.4. In Section 3.5, the proposed method is illustrated using the wind direction data. Finally, discussion and conclusion of the whole chapter are given in Section 3.6.

3.2 Background

For directional data, the distribution that is often used to describe its physical properties is the von Mises distribution and is named after the Austrian mathematician Richard Edler von Mises (1883-1953). As a continuous probability distribution, the von Mises is analogous to the normal distribution for linear data and has some similar characteristics with the normal distribution. Thus, the von Mises is also known as the

circular normal distribution. The von Mises distribution has two parameters namely the concentration parameter and the circular mean. In estimating the parameter, maximum likelihood estimation (MLE) is often used. For the concentration parameter, the solution of the MLE, however, is analytically intractable because of the presence of modified Bessel functions $I_0(\kappa)$, $I_1(\kappa)$, ... (Mardia, 1972; Batschelet, 1981 and Fisher, 1993). Thus, some approximations are applied instead.

A circular random variable θ follows the von Mises distribution, denoted by $VM(\mu_0, \kappa)$, with probability density function given by

$$g(\theta; \mu_0, \kappa) = \{2\pi I_0(\kappa)\}^{-1} \exp\{\kappa \cos(\theta - \mu_0\}, \quad (3.1)$$

where μ_0 $(0 \le \mu_0 < 2\pi)$ is the mean direction and κ is known as the concentration parameter. Also, $I_0(\kappa)$ denotes the modified Bessel function of the first kind and order zero of κ. The Bessel functions are solutions of a second-order differential equation known as the Bessel's differential equation and the probability density can also be expressed as a series of Bessel functions (Abramowitz & Stegun, 1974).

Some of the recent works on the von Mises distribution include a restricted maximum likelihood estimators (MLE) based on the assumption of large concentration parameters and when it is known apriori that the concentration parameters are subjected to a simple order restriction (Dobson, 1978). Best and Fisher (1981) provided an iterative algorithm using fixed points to obtain the MLE for κ in the von Mises-Fisher distribution and recently, Gatto (2008) extended the generalised von Mises in which Matlab was used to handle the computational aspects of the parameter estimation using MLE and trigonometric method of moments.

In this chapter, an improved efficient approximation of κ obtained from the MLE is proposed. Unlike other estimations that have been shown to be only applicable to either large or small κ, the proposed approximation is found to be suitable for all values of κ. The improved approximation is obtained by solving the piecewise polynomial functions involving the ratio of modified Bessel functions.

3.2.1 Parameter Estimation of the Von Mises Distribution

As mentioned earlier, this distribution has two parameters namely the circular mean and concentration parameter. Suppose $\theta_1,\ldots,\theta_n$ is a random sample from $VM\left(\mu_0,\kappa\right)$, the MLE of the mean direction, $\bar{\theta}$ is given by

$$\bar{\theta}=\begin{cases}\tan^{-1}\left(S/C\right), & S>0,\ C>0\\ \tan^{-1}\left(S/C\right)+\pi, & C<0\\ \tan^{-1}\left(S/C\right)+2\pi, & S<0, C>0\end{cases}, \tag{3.2}$$

where $C=\sum\cos\theta_i$ and $S=\sum\sin\theta_i$.

The MLE for κ, denoted by $\hat{\kappa}$ is given by the solution of

$$A\left(\hat{\kappa}\right)=\bar{R}=\left(\bar{C}^2+\bar{S}^2\right)^{\frac{1}{2}}, \tag{3.3}$$

where $\overline{R}$ is the mean resultant length and $A(\kappa)=\frac{I_1(\kappa)}{I_0(\kappa)}$, where I_0 and I_1 are the modified Bessel function of the first kind of order zero and one respectively. Further, the variance of $\hat{\kappa}$ is given by

$$\text{var}(\hat{\kappa})=\left\{n\left[1-\left(\frac{A(\hat{\kappa})}{\hat{\kappa}}\right)-A^2(\hat{\kappa})\right]\right\}^{-1}. \tag{3.4}$$

The parameter estimate $\hat{\kappa}=A^{-1}(\overline{R})$, however, cannot be simply evaluated. This is due to the presence of the modified Bessel functions in the formulation. Instead, an approximation of A^{-1} is used. The approximation can be obtained using iterative procedures in which the early version includes a tabulation of certain values of A^{-1} as described in Amos (1974).

From there on, several approximation of A^{-1} have been proposed in the literature. Among the approximation methods are Amos (1974), Mardia and Zemroch (1974), Dobson (1978), Best and Fisher (1981) and Hussin and Mohamed (2008). Some can be quite complicated in its derivation using sophisticated computer programme and algorithms, while some are simple and easy to derive. In this study, our objective is to propose an improve approximation for concentration parameter as we consider both small and large values of κ. In the next section, discussion will be on the current method that will be used in the comparison study.

3.2.2 Approximation for the Von Mises Concentration Parameter

As mentioned in the previous section, several approximations for $A^{-1}(x)$ for all x in $(0,1)$ can be found in the literature. In an early study, Amos (1974) proved

$$\frac{x}{\frac{1}{2}+\left(x^2+\frac{9}{4}\right)^{\frac{1}{2}}} < A(x) < \frac{x}{\frac{1}{2}+\left(x^2+\frac{1}{4}\right)^{\frac{1}{2}}}, \text{ for } x \geq 0, \tag{3.5}$$

and hence $A^{-1}(x)$ is approximately given by

$$f(x) = \frac{x}{1-x^2}\left[\frac{1}{2}+\left\{1.46\left(1-x^2\right)+\frac{1}{4}\right\}^{\frac{1}{2}}\right]. \tag{3.6}$$

Later on, Mardia and Zemroch (1975) provided a computer algorithm for calculating $A^{-1}(x)$ together with the tables which was obtained iteratively. Meanwhile, by using the power series for the Bessel function $I_0(x)$ and $I_1(x)$, Dobson (1974) gave the approximation of $A^{-1}(x)$ as follows

$$f(x) = \begin{cases} 2x + x^3 + \frac{5x^5}{6}, & x < 0.65 \\ \frac{9-8x+3x^2}{8(1-x)}, & x \geq 0.65 \end{cases}, \tag{3.7}$$

and has shown that the approximation gives less maximum relative error compared to Amos's approximation. Further, an improved approximation for $A^{-1}(x)$ was given by Best and Fisher (1981) which is

$$f(x)=\begin{cases}2x+x^3+\dfrac{5x^5}{6}, & x<0.53\\ -0.4+1.39x+\dfrac{0.43}{1-x}, & 0.53\le x<0.85,\\ \dfrac{1}{x^3-4x^2+3x}, & x\ge 0.85,\end{cases} \tag{3.8}$$

in which tabulated values are given in Fisher (1993).

In the following section, an improvement of the approximation by identifying a threshold for value of $A(\kappa)$ will be described in which the formulation as given by Fisher (1993) can be applied.

3.3 Proposed Method for Concentration Parameter

In this section, a new method of approximating the concentration parameter is proposed. This new method is developed based on modified Bessel Functions. Later, this new method will be validated via simulation studies with tabulated values of the concentration parameter and sample sizes.

By definition, $A(\kappa)=\dfrac{I_1(\kappa)}{I_0(\kappa)}=t$ and from the power series for the Bessel function $I_0(\kappa)$ and $I_1(\kappa)$, it is found that for small κ (Jammalamadaka & SenGupta, 2001),

$$A_s(\kappa)\approx\frac{\kappa}{2}\left(1-\frac{1}{8}\kappa^2+\frac{1}{48}\kappa^4-\ldots\right), \tag{3.9}$$

while for large κ,

$$A_l(\kappa) \approx 1 - \frac{1}{2\kappa} - \frac{1}{8\kappa^2} - \frac{1}{8\kappa^3} - \ldots . \tag{3.10}$$

In order to find κ such that $A_s(\kappa)$ and $A_l(\kappa)$ are close to each other, it is necessary that $\frac{A_s(\kappa)}{A_l(\kappa)} \approx 1$. In our case, we will consider the first term of $A_s(\kappa)$ and the first two terms of $A_l(\kappa)$ only for the purpose of simplicity of the calculation.

Thus,

$$\frac{A_s(\kappa)}{A_l(\kappa)} \approx \frac{\frac{\kappa}{2}}{1 - \frac{1}{2k}} \approx 1 \text{ or } \kappa^2 - 2\kappa + 1 \approx 0 . \tag{3.11}$$

Hence, $\kappa^2 - 2\kappa + \delta = 0$ or $\kappa = 1 \pm \sqrt{1-\delta}$ for small value of δ where $\delta \in [0,1]$

The above results indicate that the threshold value is in the interval [0, 2]. In order to find the threshold value, a simulation study is performed for various κ values that lies within the interval [0, 2]. The values of t_s and t_l where $A_s(\kappa) = t_s$ and $A_l(\kappa) = t_l$ are obtained where the difference between $A_s(\kappa)$ and $A_l(\kappa)$ is the smallest.

From Table 3.1, it can be seen that $\kappa_0 = 1.55$, where $A_l(\kappa_0) = 0.5918 = t_l$ and $A_s(\kappa_0) = 0.6355 = t_s$ give the smallest value of absolute difference of the computed values of t_l and t_s. By taking the average of t_l and t_s, we obtain a threshold value of approximately 0.6137.

Table 3.1: Numerical approximation of $A(\kappa)$.

κ	$A_l(\kappa)=t_l$	$A_s(\kappa)=t_s$	$\lvert A_l(\kappa)-A_s(\kappa)\rvert$
1.40	0.5335	0.5845	0.0510
1.45	0.5547	0.6012	0.0465
1.50	0.5741	0.6182	0.0441
1.55	**0.5918**	**0.6355**	**0.0436**
1.60	0.6082	0.6532	0.0451
1.65	0.6232	0.6716	0.0484
1.70	0.6372	0.6908	0.0537

Hence, we propose,

$$\hat{\kappa}=\begin{cases}A_s^{-1}(t), & t<0.6137\\ A_l^{-1}(t), & t\geq 0.6137\end{cases}, \tag{3.12}$$

where,

$$A_s(\kappa)=\frac{\kappa}{2}\left\{1-\frac{1}{8}\kappa^2+\frac{1}{48}\kappa^4+\ldots\right\}, \tag{3.9}$$

and

$$A_l(\kappa)=1-\frac{1}{2\kappa}-\frac{1}{8\kappa^2}-\frac{1}{8\kappa^3}+\ldots. \tag{3.10}$$

For $t<0.6137$,

$$\frac{\kappa}{2}\left\{1-\frac{1}{8}\kappa^2+\frac{1}{48}\kappa^4+\ldots\right\}=t \text{ or } \kappa^5-6\kappa^3+48\kappa-96t=0. \tag{3.13}$$

The solutions of the polynomial in (3.13) comprise of a real root and four complex roots. It can be obtained numerically from several mathematical packages. As an example, by using S-Plus, the polyroot function using the command line

```
polyroot(c(-96*t,48,0,-6,0,1))
```

would give the desired solution. This command line will give five roots that consist of four complex roots and one real root in which the real root is the estimated value of the concentration parameter.

For $t \geq 0.6137$, we obtain

$$1-\frac{1}{2\kappa}-\frac{1}{8\kappa^2}-\frac{1}{8\kappa^3}=t \text{ or } (8t-8)\kappa^3+4\kappa^2+\kappa+1=0 \qquad (3.14)$$

Similarly, the solution to the cubic polynomial in (3.14) can be obtained using SPlus with the command line

```
polyroot(c(1,1,4,(8*t-8)).
```

This command line will result in giving two complex roots and one real root which corresponds to the value of the concentration parameter.

3.4 Simulation Study

Computer programs were written using S-Plus to carry out the simulation study to assess the efficiency of the four different methods of approximating the concentration parameter as in (3.6 – 3.8) and the new proposed method as in (3.12). Circular samples

of length n = 30, 50 and 100 were generated from the von Mises distribution with mean 0 and κ = 0.5, 1.0, 1.5, 2.0, 4.0, 6.0, 8.0 and 10.0 respectively. Let s be the number of simulations and the following computations were obtained from the simulation study. Performance measures used in the study are given as follows.

i. Mean, $\bar{\hat{\kappa}} = \frac{1}{s}\sum \hat{\kappa}_j$,

ii. Absolute Relative Estimated Bias (AREB) = $\left(\frac{\left| \bar{\hat{\kappa}} - \kappa \right|}{\kappa} \right) \times 100\%$,

iii. Estimated Standard Errors (SE) = $\sqrt{\frac{1}{s-1}\sum \left(\hat{\kappa}_j - \bar{\hat{\kappa}} \right)^2}$,

iv. Estimated Root Mean Square Errors (RMSE) = $\sqrt{\frac{1}{s}\sum \left(\hat{\kappa}_j - \kappa \right)^2}$.

The simulation results with s = 5000 for various true values of concentration parameter and n = 30, 50 and 100 are shown in Tables 3.2, 3.3 and 3.4 respectively. The values of mean, absolute relative estimated bias (AREB), estimated standard error (SE) and estimated root mean square error (RMSE) were computed for all the Amos's (3.6), Dobson's (3.7), Best and Fisher's (3.8) and the new proposed method. When considering mean alone, Tables 3.2 to 3.4 show that the estimated mean obtained using the proposed method is close to the true mean for most of the given κ values as compared the other three methods.

Table 3.2: Simulation results for various value of parameter concentration, κ and $n = 30$

Performance Indicator	Concentration parameter, κ	$\hat{\kappa}^{\text{New}}$	$\hat{\kappa}^{\text{Amos}}$	$\hat{\kappa}^{\text{Best and Fisher}}$	$\hat{\kappa}^{\text{Dobson}}$
Mean	0.5	0.5950	0.5525	0.5956	0.5954
	1.0	1.0568	1.0357	1.0776	1.0736
	1.5	1.4620	1.6048	1.6036	1.5982
	2.0	2.1175	2.2066	2.1300	2.1261
	4.0	4.3814	4.8218	4.4030	4.3922
	6.0	6.5931	7.1735	6.6005	6.5977
	8.0	8.8172	9.4770	8.8211	8.8197
	10.0	11.0779	11.7895	11.0804	11.0795
Absolute Relative Estimate Bias (AREB)	0.5	0.1899	0.1051	0.1913	0.1908
	1.0	0.0568	0.0357	0.0776	0.0736
	1.5	0.0254	0.0698	0.0690	0.0655
	2.0	0.0587	0.1033	0.0650	0.0630
	4.0	0.0954	0.2055	0.1007	0.0981
	6.0	0.0988	0.1956	0.1001	0.0996
	8.0	0.1021	0.1846	0.1026	0.1025
	10.0	0.1078	0.1790	0.1080	0.1080
Estimate Standard Error (SE)	0.5	0.2568	0.2519	0.2589	0.2582
	1.0	0.2868	0.3429	0.3258	0.3219
	1.5	0.2740	0.4604	0.4099	0.4143
	2.0	0.4825	0.6081	0.5190	0.5110
	4.0	1.1833	1.3090	1.1692	1.1776
	6.0	1.7931	1.9080	1.7889	1.7906
	8.0	2.3841	2.4865	2.3820	2.3827
	10.0	3.0345	3.1253	3.0332	3.0337
Estimate Root Mean Square Error (ERMSE)	0.5	0.2728	0.2545	0.2750	0.2742
	1.0	0.2919	0.3434	0.3340	0.3294
	1.5	0.2500	0.4685	0.4186	0.4219
	2.0	0.4941	0.6379	0.5333	0.5244
	4.0	1.2239	1.5005	1.2191	1.2228
	6.0	1.8679	2.1852	1.8669	1.8674
	8.0	2.4948	2.8321	2.4944	2.4946
	10.0	3.1993	3.5474	3.1991	3.1992

To compare the performance of each method in the simulation studies, the value of the performance measure of AREB is used. From the simulation results in Tables 3.2, 3.3 and 3.4, generally it is observed that the measures of AREB for the proposed method are closer to zero for most of the values of κ as compared to the other estimates. However, it can be seen that for n = 30 and 50, and for very small values of κ, that is, for $\kappa \leq 1.0$, the approximations by Amos seem to show the smallest AREB value among the four approximation methods. Nevertheless, as size increases, specifically when n = 100 (see Table 3.4), the Amos method is only good for κ = 0.5. It can be inferred that for large values of κ, which is for κ > 1.0, the proposed method is consistently better than the other estimates with the smallest AREB when the sample size is $n \leq 50$.

As the sample size increases to 100, the proposed method seems to give the best estimate with the inclusive value of κ = 1. Thus, it can be deduced that for sample size is $n \leq 50$ and κ > 1.0, the proposed method is the best and as the sample size increase to 100 the proposed method is even better with a bigger range of values of κ, that is, $\kappa \geq 1.0$.

Table 3.3: Simulation results for various value of parameter concentration, κ and $n = 50$.

Performance Indicator	Concentration parameter, κ	$\hat{\kappa}^{\text{New}}$	$\hat{\kappa}^{\text{Amos}}$	$\hat{\kappa}^{\text{Best and Fisher}}$	$\hat{\kappa}^{\text{Dobson}}$
Mean	0.5	0.5525	0.5096	0.5526	0.5526
	1.0	1.0409	1.0050	1.0513	1.0478
	1.5	1.4520	1.5496	1.5577	1.5506
	2.0	2.0594	2.1413	2.0763	2.0773
	4.0	4.1747	4.6051	4.1959	4.1856
	6.0	6.3509	6.9265	6.3582	6.3555
	8.0	8.4384	9.0931	8.4424	8.4410
	10.0	10.5931	11.3004	10.5956	10.5947
Absolute Relative Estimate Bias	0.5	0.1050	0.0191	0.1051	0.1051
	1.0	0.0409	0.0050	0.0513	0.0478
	1.5	0.0320	0.0331	0.0385	0.0337
	2.0	0.0297	0.0707	0.0382	0.0387
	4.0	0.0437	0.1513	0.0490	0.0464
	6.0	0.0585	0.1544	0.0597	0.0593
	8.0	0.0548	0.1366	0.0553	0.0551
	10.0	0.0593	0.1300	0.0596	0.0595
Estimate Standard Error	0.5	0.2031	0.1960	0.2033	0.2032
	1.0	0.2282	0.2572	0.2459	0.2404
	1.5	0.2144	0.3356	0.3021	0.3114
	2.0	0.3493	0.4483	0.3817	0.3730
	4.0	0.8341	0.9501	0.8208	0.8288
	6.0	1.3271	1.4410	1.3235	1.3249
	8.0	1.7397	1.8442	1.7379	1.7386
	10.0	2.1748	2.2700	2.1737	2.1741
Estimate Root Mean Square Error	0.5	0.2089	0.1929	0.2091	0.2090
	1.0	0.2308	0.2544	0.2497	0.2439
	1.5	0.2030	0.3370	0.3042	0.3128
	2.0	0.3523	0.4666	0.3883	0.3801
	4.0	0.8227	1.0647	0.8170	0.8213
	6.0	1.3408	1.6389	1.3400	1.3404
	8.0	1.7614	2.0657	1.7610	1.7612
	10.0	2.2286	2.5457	2.2284	2.2285

Table 3.4: Simulation results for various value of parameter concentration, κ and $n = 100$

Performance Indicator	Concentration parameter, κ	$\hat{\kappa}^{\text{New}}$	$\hat{\kappa}^{\text{Amos}}$	$\hat{\kappa}^{\text{Best and Fisher}}$	$\hat{\kappa}^{\text{Dobson}}$
Mean	0.5	0.5279	0.4848	0.5279	0.5279
	1.0	1.0146	0.9684	1.0183	1.0171
	1.5	1.5055	1.5063	1.5216	1.5072
	2.0	2.0139	2.0893	2.0337	2.0443
	4.0	4.0653	4.4907	4.0855	4.0763
	6.0	6.1628	6.7345	6.1701	6.1674
	8.0	8.2637	8.9170	8.2677	8.2663
	10.0	10.3072	11.0122	10.3097	10.3089
Absolute Relative Estimate Bias	0.5	0.0559	0.0304	0.0558	0.0558
	1.0	0.0146	0.0316	0.0183	0.0171
	1.5	0.0036	0.0042	0.0144	0.0048
	2.0	0.0070	0.0446	0.0168	0.0221
	4.0	0.0163	0.1227	0.0214	0.0191
	6.0	0.0271	0.1224	0.0284	0.0279
	8.0	0.0330	0.1146	0.0335	0.0333
	10.0	0.0307	0.1012	0.0310	0.0309
Estimate Standard Error	0.5	0.1472	0.1423	0.1472	0.1472
	1.0	0.1639	0.1788	0.1698	0.1671
	1.5	0.2177	0.2263	0.2045	0.2140
	2.0	0.2343	0.3040	0.2590	0.2493
	4.0	0.5945	0.7062	0.5831	0.5898
	6.0	0.8990	1.0198	0.8960	0.8971
	8.0	1.2223	1.3372	1.2207	1.2212
	10.0	1.5120	1.6191	1.5111	1.5114
Estimate Root Mean Square Error	0.5	0.1482	0.1380	0.1482	0.1482
	1.0	0.1628	0.1766	0.1688	0.1661
	1.5	0.2176	0.2263	0.2052	0.2140
	2.0	0.2301	0.3118	0.2586	0.2509
	4.0	0.5584	0.7808	0.5529	0.5569
	6.0	0.8674	1.1563	0.8668	0.8670
	8.0	1.2040	1.5178	1.2038	1.2039
	10.0	1.5008	1.8085	1.5007	1.5007

Another measure of the performance, namely the measures of SE and RMSE are used. From Tables 3.2 to 3.4, we can see that the values of SE and RMSE for new proposed method are generally consistent for most of the tabulated κ values. Amos estimates give the smallest SE and RMSE for small value of κ, which κ is for $\kappa \leq$ 1.0, but become large as compared to other methods for κ > 1.0. Consistent with the earlier measure of AREB, it can be deduced that Amos estimate gives the best estimate for small κ (i.e. for $\kappa \leq 1.0$) but perform poorly for κ > 1.0. This suggests the superiority of the new proposed method as compared to the other two methods.

Using the measures of SE and RMSE, we note that the new proposed method gives almost similar value as compared to Best and Fisher's as well as Dobson's method. However, those measures did not elicit the superiority of the new proposed over the other two methods.

3.5 Illustrative Examples

As an illustration of the applicability of the proposed method, a bivariate data set was considered. The data was collected from along the Holderness Coastline, which is the Humberside Coast of the North Sea, United Kingdom in October 1994. A total of 85 measurements of wind direction using HF radar (x) and anchored buoy (y) were recorded over a period of 22.7 days. The data was fitted using the simple linear regression model proposed by Downs and Mardia (2002), and the model is given as below:

$$\hat{y}_i = 1.253 + 2\arctan\left\{0.906\tan\frac{1}{2}\left(x_i - 1.141\right)\right\}.$$

Our particular interest is on the estimation of the concentration parameter for the circular residuals, $\theta_i = \hat{y}_i - y_i$ based on the fitted model.

Table 3.5: Estimation of κ using the new proposed method

Data	$\hat{\kappa}^{\text{New}}$	$\hat{\kappa}^{\text{Amos}}$	$\hat{\kappa}^{\text{Best and Fisher}}$	$\hat{\kappa}^{\text{Dobson}}$
Concentration parameter	7.442	8.073	7.447	7.445

From Table 3.5, the estimated value of the concentration parameter for the residuals is high, and it can be proved using the circular plot as shown in Figure 3.1. Higher concentration parameter implies that the circular residuals are highly concentrated among each other as can be seen from Figure 3.1 where majority of the data are scattered around (-45^o, 45^o) with only a few observations fall outside the range. The results give almost the similar pattern as obtained in the simulation studies in Section 3.4.

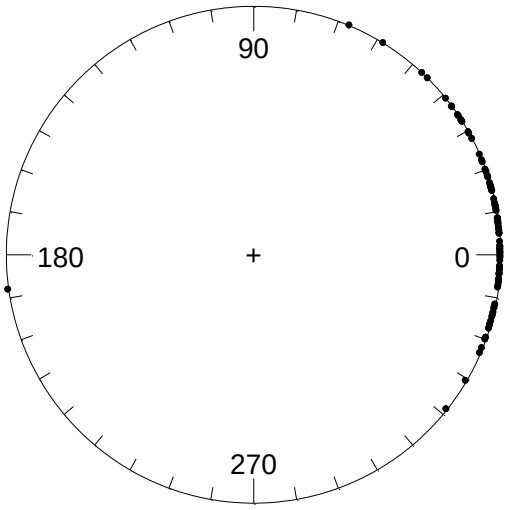

Figure 3.1: Circular plot for residuals

3.6 Discussion

This study is to propose an improved estimate of the concentration parameter, κ for the von Mises distribution that is applicable for both small and large values of κ. Based on the MLE, the estimate of κ is a piecewise function that involves a solution to a polynomial function that can be easily solved using S-Plus. To evaluate the performance of the proposed estimate, simulation studies were carried out to compare the three different approximation methods of concentration parameter κ, namely, the Dobson's method, Best & Fisher's method and Amos's method. Generally, it appears that, for both small and large values of κ, the proposed method shows a better performance than the Amos's, Dobson's and Best & Fisher's methods except for when $\kappa \leq 1$. The proposed method of approximation exhibits the least absolute relative bias for most of the κ values. The superiority of the proposed method is also observed with general consistent values of estimated SE and RMSE in comparison to the other methods considered. Unlike the Amos's method which is restrictive to small values of κ ($\kappa \leq 1$ for $n \leq 50$), the proposed method seems to be applicable to both small and large values of κ ($\kappa \geq 1$ for large sample size $n = 100$).

CHAPTER 4

CONFIDENCE INTERVALS FOR LARGE CONCENTRATION PARAMETER IN VON MISES DISTRIBUTION

4.1 Introduction

This chapter discusses on the approximation for confidence intervals (CI) of the concentration parameter for von Mises distribution. Section 4.2 commences with a brief introduction about the CI for parameter in circular distribution, in particular, the von Mises distribution. Details on the proposed approximation methods are given in Section 4.3. Four proposed method, as well as the current method by Fisher, are described. To assess the performance of the methods considered, simulation studies are carried out in Section 4.4. In Section 4.5, the proposed method is illustrated using the wind direction data. Finally, discussion and conclusion of the whole chapter are given in Section 4.6.

4.2 Background

This section is an extension of the previous study discussed in Chapter 3 where the new approximation for the concentration parameter was proposed. In this chapter, our particular interest is to find efficient confidence intervals (CI) for large concentration parameter, κ i.e $\kappa \geq 2$ (Mardia & Jupp, 2000) where the distribution becomes very concentrated around the angle μ with κ being the measure of the

concentration. As κ increases, the distribution of θ approaches a normal distribution with mean μ and variance $\frac{1}{\kappa}$. As κ approaches 0, the distribution tends to converge to a uniform distribution (Fisher, 1993; Mardia & Jupp, 2000).

Confidence intervals can be defined as an interval estimate of the point estimator or the parameter itself. As in Efron and Tibshirani (1993), knowing the interval estimate with its point estimate we can obtain the 'best guess' of the parameter and how far in error that guess might be. In the perspective of linear statistics, this area has gained great attention from many researchers. Many new and integrated approaches were developed to obtain an efficient approximation for CI based on different methods such as CI based on hypothesis testing and bootstrap, which include percentile bootstrap, bootstrap-*t* and iterated bootstrap. Few studies related to this topic can be found in Hall (1986, 1988), Porter *et al.* (1997), Polansky (2000), Sun and Wong (2007) and Asgharzadeh and Abdi (2011).

In circular statistics, some works were done in finding the CI for parameters in the unimodal distribution. Initially, Ducharme (1985) proposed the confidence cones for the mean directional vector by the bootstrap method for *F*-distribution on a *p*-dimensional sphere. Later on, Fisher and Hall (1989) came out with an alternative bootstrap algorithm to improve the method proposed by Ducharme (1985). In this work, they introduced a new approach that is based on pivotal statistics, and it is said to have a smaller coverage error as compared to non-pivotal statistics. For von Mises distribution, the CI for the mean direction based on the bootstrap method is discussed in Fisher (1993). Jammalamadaka and SenGupta (2001) discussed on the construction of CI for the mean direction based on circular 'standard error' of the MLE for the mean direction.

Details on confidence intervals for parameter in von Mises distribution were given in Section 2.4.2.

As for the concentration parameter in von Mises distribution, to our knowledge, works on the efficient CI are relatively few. An early study was carried out by Stephen (1969), where several approximations of CI for small concentration parameter are proposed. Steps for CI based on percentile bootstrap method can be found in Fisher (1993). Later on, Khanabsakdi (1995) proposed a new CI for the concentration parameter based on the Chi-square approximation and comparison with the previous method by Stephens showed that the new method is better than the previous one. Details on the four new proposed methods of approximating CI for the concentration parameter will be discussed in the next section. Also, the current method using the percentile bootstrap by Fisher (1993) is described.

4.3 Methods in Approximating Confidence Intervals (CI)

In this study, we have proposed four new methods of obtaining CI for the concentration parameter where $\kappa \geq 2$ due to its wide application in the real problem. The four methods considered are listed as below:

i. CI based on circular variance population which will be referred to as Method 1

ii. CI based on the asymptotic distribution of $\hat{\kappa}$ which will be referred to as Method 2

iii. CI based on the distribution of the mean direction, $\bar{\theta}$ and mean resultant length, $\bar{R}$ which will be referred to as Method 3

iv. CI based on bootstrap-*t* method which will be referred to as Method 4

In addition to the four methods, a current method based on percentile bootstrap (Fisher, 1993) will also be considered and used in the study. The performance of all the methods considered will be measured using the measurements of expected length and coverage probability.

4.3.1 Percentile bootstrap

Bootstrap method is one of the resampling techniques which has gained much attention in the past few years. Several bootstrap methods can be found in the literature; some of the widely used methods are percentile bootstrap, bootstrap-*t*, bias-corrected and accelerated bootstrap (BCA) and also calibration bootstrap. Efron and Tibshirani (1993) and Chernick (1999) gave a comprehensive review on constructing the CI based on several bootstrap methods including the percentile method. Other studies that discussed on the bootstrap method and CI can be found in Hall (1986), Porter *et al.* (1997) and Polansky (2000).

Percentile bootstrap is the simplest bootstrap method in approximating the confidence intervals. In circular statistics, Fisher (1993) described the percentile bootstrap method to approximate the CI for the concentration parameter. The following steps are carried out for simulation purpose:

Step 1: Resampling

Simulate *n* values $\theta_1^*,\ldots,\theta_n^*$ from $VM\left(\hat{\mu},\hat{\kappa}\right)$, where $\theta_i^*, i=1,2,\ldots,n$ and $0\le\theta<2\pi$.

Step 2: Bootstrap parameter estimate

The bootstrap parameter estimates for the bootstrap samples from Step 1 are obtained and labelled as $\hat{\kappa}_1^*$.

Step 3: Repetition

Steps 1 and 2 are repeated to obtain *B* bootstrap estimates $\hat{\kappa}_1^*,\ldots,\hat{\kappa}_B^*$ of the concentration parameter.

Step 4: Confidence intervals

i. To get a CI for κ, arrange the bootstrap estimates, $\hat{\kappa}_1^*,\ldots,\hat{\kappa}_B^*$ in increasing order:

$$\hat{\kappa}_{(1)}^*\le\ldots\le\hat{\kappa}_{(B)}^*.$$

ii. CI for κ is given as:

$$\left(\hat{\kappa}_{(l+1)}^*,\hat{\kappa}_{(m)}^*\right), \tag{4.1}$$

where l = integer part of $\left(\frac{1}{2}B\alpha+\frac{1}{2}\right)$ and $m = B - 1$.

4.3.2 New Proposed Methods for Confidence Intervals for Concentration Parameter

In this section, four different methods in approximating the confidence interval (CI) for large concentration parameter are constructed.

(i) Method 1: CI Based on Circular Variance Population

The CI of concentration parameter κ may be obtained by considering the wrapping of the normal distribution $N\left(\mu,\sigma^2\right)$ around the circle which gives the wrapped normal distribution given by $WN\left(\mu, A\left(\kappa\right)\right)$, where $A\left(\kappa\right)=\exp\left\{-\frac{\sigma^2}{2}\right\}$ or $\sigma^2=-2\ln\left(A\left(\kappa\right)\right)$ and the sample circular standard deviation, v is given by $v=\left\{-2\ln\left(1-V\right)\right\}^{\frac{1}{2}}$ (Fisher, 1993). However, $V=1-\bar{R}$, hence the sample circular standard deviation can be written as

$$\begin{aligned} v &= \left\{-2\ln\left(1-\left(1-\bar{R}\right)\right)\right\}^{\frac{1}{2}} \\ &= \left\{-2\ln\left(\bar{R}\right)\right\}^{\frac{1}{2}}. \end{aligned} \tag{4.2}$$

By using the standard result, the $100\left(1-\alpha\right)\%$ CI for the variance, σ^2 is given by

$$\frac{(n-1)v^2}{\chi^2_{n-1,\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)v^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}. \tag{4.3}$$

Rewriting (4.2) and using $\sigma^2 = -2\ln\left(A(\kappa)\right)$, equation (4.3) can be written as

$$\frac{(n-1)v^2}{\chi^2_{n-1,\frac{\alpha}{2}}} < -2\ln\left(A(\kappa)\right) < \frac{(n-1)v^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}. \tag{4.4}$$

Alternatively, we may write

$$\exp\left(-\frac{(n-1)v^2}{2\chi^2_{n-1,1-\frac{\alpha}{2}}}\right) < A(\kappa) < \exp\left(-\frac{(n-1)v^2}{2\chi^2_{n-1,\frac{\alpha}{2}}}\right). \tag{4.5}$$

$$A^{-1}(Y) < \kappa < A^{-1}(Z). \tag{4.6}$$

Thus, we may obtain the lower value, κ_L as well as the upper value, κ_U such that $\Pr\left(\kappa_L < \kappa < \kappa_U\right) = 1-\alpha$ where $\kappa_L = A^{-1}(Y)$ and $\kappa_U = A^{-1}(Z)$ respectively. The values of $A^{-1}(Y)$ and $A^{-1}(Z)$ in (4.6) may be estimated using the polyroot function in S-Plus as described in Hassan *et.al* (2012).

(ii) Method 2: CI based on the asymptotic distribution of $\hat{\kappa}$

Another procedure for finding the CI for κ is based on the normal distribution for the distribution of $\hat{\kappa}$ which is normally distributed with mean and variance given as below (Jamalamadaka & SenGupta, 2001),

$$\hat{\kappa} \sim N\left(\kappa, \frac{1}{n\left(1-\frac{\bar{R}}{\hat{\kappa}}-\bar{R}^2\right)}\right). \tag{4.7}$$

Hence, the 95% CI can be obtained by,

$$-B+\hat{\kappa}<\kappa<B+\hat{\kappa}, \tag{4.8}$$

where $B=\frac{1.96}{\left[n\left(1-\frac{\bar{R}}{\hat{\kappa}}-\bar{R}^2\right)\right]^{\frac{1}{2}}}$.

(iii) Method 3: CI Based on Distribution of Mean Direction, $\bar{\theta}$ and Mean Resultant Length, $\bar{R}$

We also propose CI of large κ based on the distribution of the mean direction, $\bar{\theta}$ and mean resultant length, $\bar{R}$. Let θ be a circular random variable from $VM\left(0,\kappa\right)$, then for large κ and following Hendricks *et al.* (1996),

$$2n\kappa A\left(\kappa\right)\left(1-\cos\bar{\theta}\right) \sim \chi_1^2 \text{ as } n\to\infty. \tag{4.9}$$

Alternatively, if we substitute with $\cos\bar{\theta}=\frac{\bar{C}}{\bar{R}}$, then we have

$$2n\frac{\kappa A\left(\kappa\right)}{\bar{R}}\left(\bar{R}-\bar{C}\right) \sim \chi_1^2 \text{ as } n\to\infty. \tag{4.10}$$

We note that,

$$2n\frac{\kappa A(\kappa)}{\bar{R}}\left(1-\bar{C}\right)=2n\frac{\kappa A(\kappa)}{\bar{R}}\left(1-\bar{R}\right)+2n\frac{\kappa A(\kappa)}{\bar{R}}\left(\bar{R}-\bar{C}\right). \tag{4.11}$$

Following the Cochran's theorem (see Stuart & Ord, 1991), we have

$2n\frac{\kappa A(\kappa)}{\bar{R}}\left(1-\bar{R}\right)\sim\chi^2_{n-1}$ and random variables $2n\frac{\kappa A(\kappa)}{\bar{R}}\left(1-\bar{R}\right)$ and $2n\frac{\kappa A(\kappa)}{\bar{R}}\left(\bar{R}-\bar{C}\right)$ are approximately independent for large κ. Further, from the decomposition in (4.11) we have,

$$2n\frac{\kappa A(\kappa)}{\bar{R}}\left(1-\bar{C}\right)\sim\chi^2_n. \tag{4.12}$$

In practice, the asymptotic result (4.12) is not adequate for moderately large values of κ (Mardia & Jupp, 2000). One way of improving the approximate (4.12) is to multiply $2n\frac{\kappa A(\kappa)}{\bar{R}}\left(1-\bar{C}\right)$ by a suitable constant so that its mean is approximately exactly the limiting value *n*. Following the idea of Stephens (1969) and Mardia and Jupp (2000), where γ is the average of γ_1 and γ_2 such that,

$$E\left[\frac{2n\gamma_1\left(1-\bar{C}\right)}{\bar{R}}\right]=n, \tag{4.13}$$

and

$$\gamma_2=\frac{2\kappa^2}{2\kappa+1}. \tag{4.14}$$

Hence, by solving for γ and averaging (4.13) and (4.14), we have,

$$2\gamma = \frac{1}{2}\left\{\frac{4\kappa^2}{2\kappa+1} + \frac{16n\kappa^2 + (8-8n)\kappa - 2n}{8n\kappa + 2n}\right\} = \frac{32n\kappa^3 + (4n+8)\kappa^2 + (4-6n)\kappa - n}{16n\kappa^2 + 12n\kappa + 2n}. \quad (4.15)$$

Thus, the decomposition (4.11) can be improved to,

$$\frac{2n\gamma}{\bar{R}}(1-\bar{C}) = \frac{2n\gamma}{\bar{R}}(1-\bar{R}) + \frac{2n\gamma}{\bar{R}}(\bar{R}-\bar{C}), \quad (4.16)$$

which gives,

$$\frac{2n\gamma}{\bar{R}}(1-\bar{R}) \sim \chi^2_{n-1}. \quad (4.17)$$

From (4.16) we have $\Pr\left(A^{-1} < 2\gamma < B^{-1}\right) = 1-\alpha$,

Thus,

$$\frac{1}{A} < \frac{32n\kappa^3 + (4n+8)\kappa^2 + (4-6n)\kappa - n}{16n\kappa^2 + 12n\kappa + 2n} < \frac{1}{B}. \quad (4.18)$$

Case 1

The lower limit for new confidence intervals is given as below

$$\frac{1}{A} < \frac{32n\kappa^3 + (4n+8)\kappa^2 + (4-6n)\kappa - n}{16n\kappa^2 + 12n\kappa + 2n}, \quad (4.19)$$

$$32nA\kappa^3 + (4nA + 8A - 16n)\kappa^2 + (4A - 6nA - 12n)\kappa - (nA + 2n) > 0. \quad (4.20)$$

Case 2

The upper limit for new confidence intervals is given as below

$$\frac{32n\kappa^3+(4n+8)\kappa^2+(4-6n)\kappa-n}{16n\kappa^2+12n\kappa+2n}<\frac{1}{B}, \tag{4.21}$$

$$32nB\kappa^3+(4nB+8B-16n)\kappa^2+(4B-6nB-12n)\kappa-(nB+2n)>0, \tag{4.22}$$

where $A=\frac{n(1-\bar{R})}{\bar{R}\chi^2_{(n-1),1-\frac{\alpha}{2}}}$ and $B=\frac{n(1-\bar{R})}{\bar{R}\chi^2_{(n-1),\frac{\alpha}{2}}}$.

The lower in (4.20) and upper limit in (4.22) can be obtained using the '*polyroot function*' in S-Plus in order to estimate $100(1-\alpha)\%$ of confidence interval for κ.

(iv) Method 4: CI Based on Bootstrap-*t* Method

From the literature, for large sample size, it has been shown that bootstrap-*t* gives a narrower expected length with smaller coverage error over bootstrap percentile and BCA method (Hall, 1988 and Porter *et al.* 1997). In this study, a new bootstrap-*t* method for constructing the confidence intervals for the concentration parameter is proposed. The simulation studies will be done using the steps described as follows:

Step 1: Resampling

Simulate *n* values $\theta_1^*,\ldots,\theta_n^*$ from $VM(\hat{\mu},\hat{\kappa})$.

Step 2: Bootstrap parameter estimate

i. The bootstrap estimate for the bootstrap sample from Step 1 is obtained and labelled as $\hat{\kappa}_1^*$.

ii. Calculate the standard error (SE) for the estimated bootstrap parameter and label as $\hat{S}_1^*$ where $\text{var}(\hat{\kappa}) = \frac{1}{n\left(1 - \frac{\bar{R}}{\hat{\kappa}} - \bar{R}^2\right)}$.

iii. Calculate the *t*-value given by

$t_1^* = \frac{\hat{\kappa}_1^* - \hat{\kappa}}{\hat{S}_1^*}$ where $\hat{S}_1^*$ is the estimate of the SE of $\hat{\kappa}^*$ based on the data in the first bootstrap sample.

Step 3: Repetition

Steps 1 and 2 are repeated to obtain *B* bootstrap *t*-values $t_1^*, \ldots, t_B^*$ of the concentration parameter.

Step 4: Confidence intervals

i. For $i = 1, \ldots, B$, $\bar{R}_i^* = \left(C_i^2 + S_i^2\right)^{\frac{1}{2}}$, $\hat{\kappa}_i^*$, $\hat{S}_i^*$ and t_i^* are calculated.

ii. To get a CI for κ, arrange the *t*-values, $t_1^*, \ldots, t_B^*$ in increasing order: $t_{(1)}^* \leq \ldots \leq t_{(B)}^*$.

iii. The $100(1-\alpha)\%$ CI for κ will be given as,

$$\left(\hat{\kappa} - t^*_{(1-\alpha)}S, \hat{\kappa} - t^*_{(\alpha)}S\right). \tag{4.23}$$

where $t^*_{(1-\alpha)}$ is $1-\alpha$ percentile of t^*_b values, $t^*_{(\alpha)}$ is α percentile of t^*_b values and *S* is estimated standard error for $\hat{\kappa}$.

4.4 Simulation Study

Simulation studies were carried out for three different sample sizes, *n* = 30, 50 and 100 respectively with various values of concentration parameter, namely κ = 2, 4, 6 and 8 for the confidence level, α = 0.05. Without loss of generality, the mean direction will be taken as 0 during the simulation study. Let *m* be the number of simulations, and the following computation were obtained. We define

i. Coverage Probability = $\frac{q}{m}$, where *q* = number of true value that falls in the CI and *m* = number of simulation.
ii. Expected Length = Upper limit – Lower limit.

Coverage probability can be defined as the proportion of a number that the CI contains the true value. In other words, the coverage probability is the actual probability that the interval contains the true concentration parameter for each method. The simulation studies were repeated for 5000 times and have been done at 95% of confidence level. Hence, the good indicator should give a coverage probability close to 0.95 which we refer to as nominal coverage probability or target value. Tables 4.1 and

4.2 showed the coverage probability and expected length respectively obtained from the simulation studies for different sample size and concentration parameter.

Table 4.1: Coverage probability for various value of κ for each sample size, n = 30, 50 and 100.

Sample size, n	Concentration parameter	Percentile Bootstrap By Fisher	Method 1	Method 2	Method 3	Method 4
30	2	0.902	0.842	0.961	0.741	0.935
	4	0.897	0.932	0.958	0.882	0.946
	6	0.895	0.940	0.959	0.917	0.945
	8	0.883	0.941	0.960	0.926	0.944
50	2	0.919	0.889	0.956	0.694	0.935
	4	0.920	0.926	0.952	0.885	0.946
	6	0.912	0.939	0.958	0.919	0.950
	8	0.908	0.941	0.951	0.922	0.943
100	2	0.930	0.914	0.971	0.594	0.939
	4	0.929	0.929	0.952	0.862	0.949
	6	0.926	0.933	0.951	0.909	0.947
	8	0.921	0.943	0.956	0.928	0.943

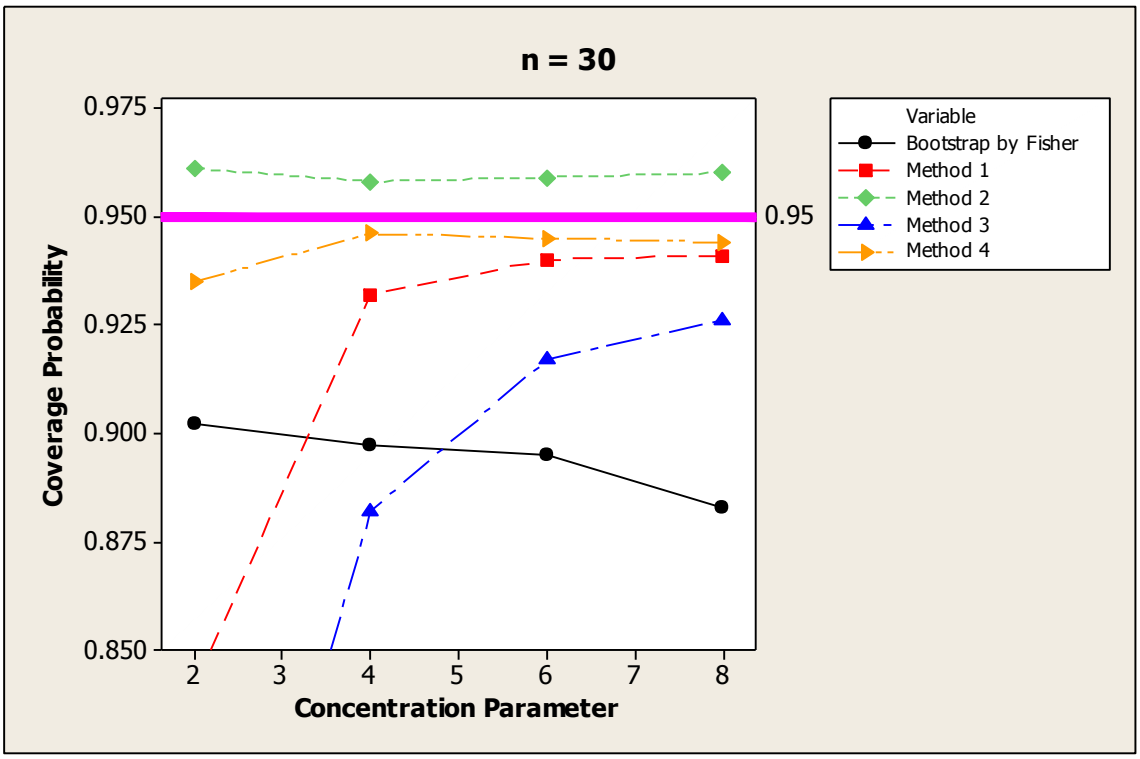

Figure 4.1: Coverage probability versus concentration parameter for *n* = 30

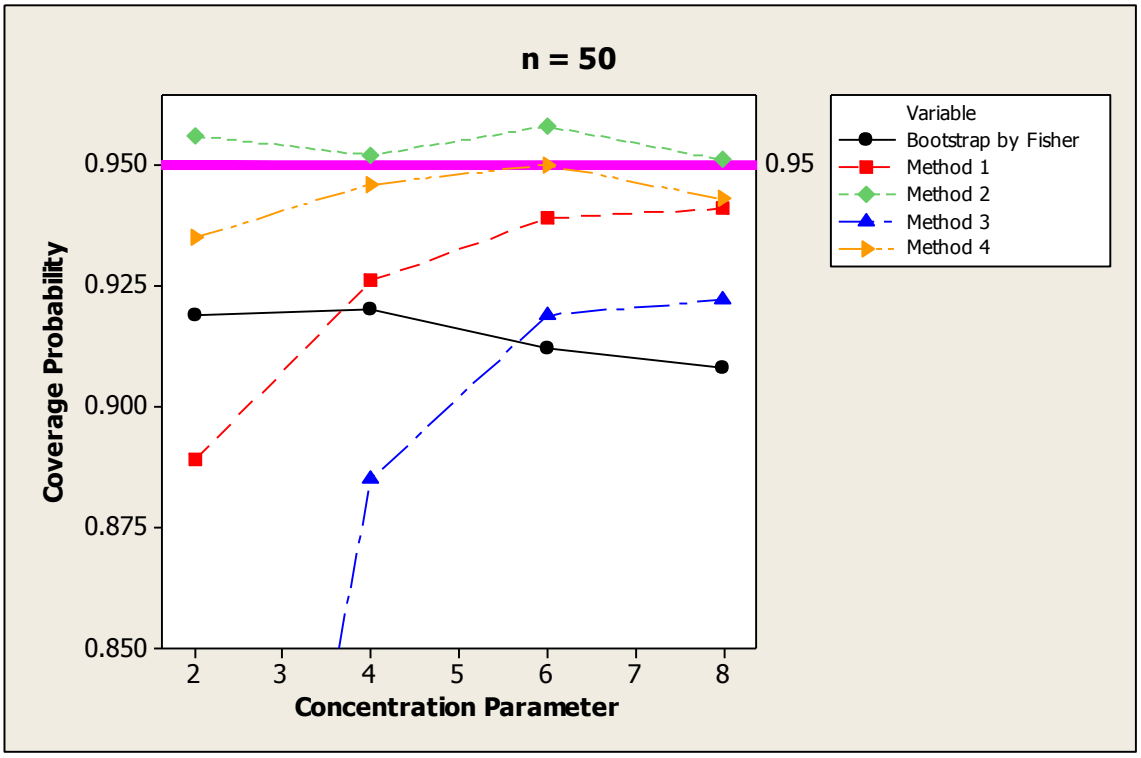

Figure 4.2: Coverage probability versus concentration parameter for *n* = 50

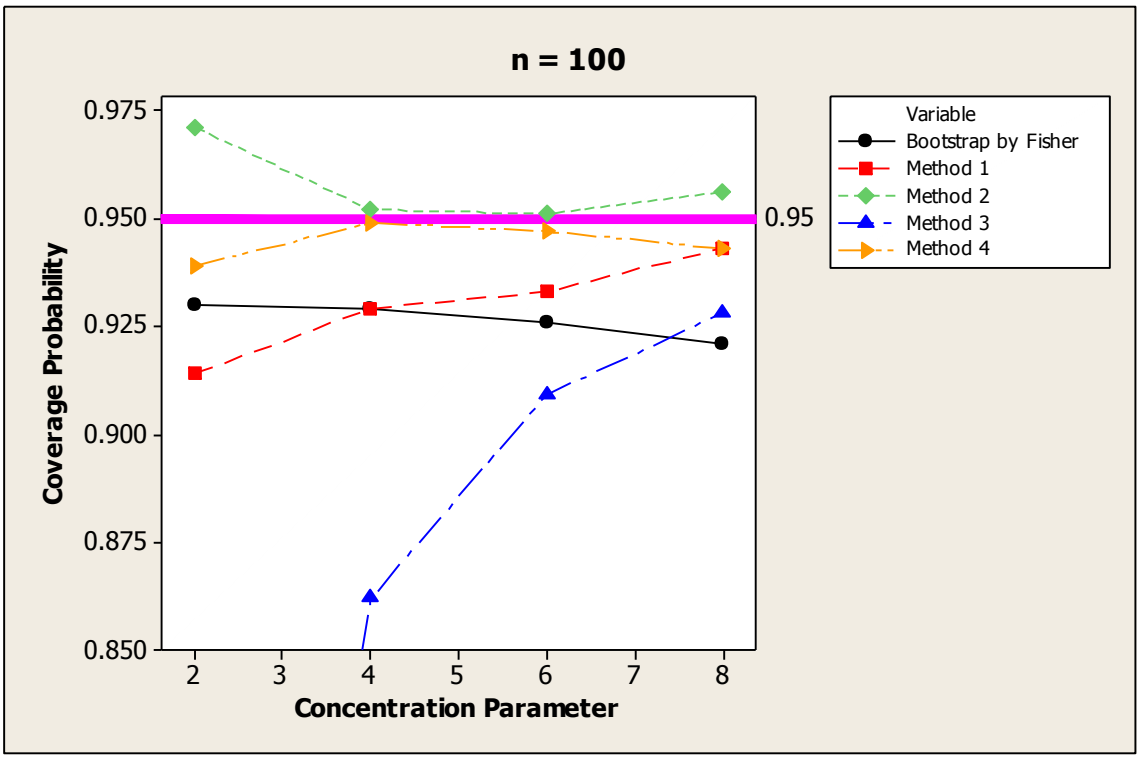

Figure 4.3: Coverage probability versus concentration parameter for *n* = 100

From the results obtained and displayed in Table 4.1, it can be seen that the coverage probability approaches to the target value as the value of concentration parameter increases for Method 1 to Method 4. Method 2 (CI based on the asymptotic distribution of $\hat{\kappa}$) gives consistently higher coverage probability than the target values for all different values of the concentration parameter. It can be seen that both Method 2 (CI based on the asymptotic distribution of $\hat{\kappa}$) and Method 4 (CI based on bootstrap-*t* method) have values close to the target value and Method 3 (CI based on distribution of mean direction and mean resultant length) gives lower coverage probability than the target value. The coverage probability by Fisher has the poorest performance with consistently having the lowest coverage probability. Therefore, by considering the coverage probability as the performance indicator, Method 4 (CI Based on Bootstrap-*t* Method) and Method 2 (CI based on asymptotic distribution of $\hat{\kappa}$) are the best as they give good coverage probability. This is followed by Method 1 and then Method 3.

Figures 4.1, 4.2 and 4.3 represent the coverage probability plot for each sample size. The thick pink line shows the target value (0.95) and can be labelled as a reference line for the other plots. From the results obtained and plots in Figures 4.1 to 4.3, looking at Method 1, it can be seen that the coverage probability become much closer to the target value as the value of concentration parameter increases. On the other hand, Method 2 gives consistently higher coverage probability than the target values for all different values of the concentration parameter. In which the plots are always above the reference line. Apart from that, Method 2 and Method 4 seems to give the values that are close to the target value in comparison to other values including the current method itself. We note that Method 3 gives quite poor coverage probability especially for $\kappa \leq 4$.

Table 4.2: Expected length for various value of κ for each sample size, n = 30, 50 and 100.

Sample size, n	Concentration parameter	Percentile Bootstrap By Fisher	Method 1	Method 2	Method 3	Method 4
30	2	2.185	1.561	1.974	1.339	1.851
	4	4.752	3.851	4.168	3.531	4.028
	6	7.477	6.194	6.455	5.779	6.316
	8	10.186	8.555	8.774	8.050	8.593
50	2	1.529	1.143	1.471	0.996	1.404
	4	3.280	2.851	3.111	2.649	3.025
	6	5.146	4.572	4.797	4.322	4.720
	8	6.934	6.237	6.439	5.946	6.341
100	2	1.005	0.777	1.010	0.683	0.980
	4	2.125	1.941	2.132	1.821	2.082
	6	3.344	3.112	3.283	2.970	3.247
	8	4.523	4.272	4.430	4.113	4.378

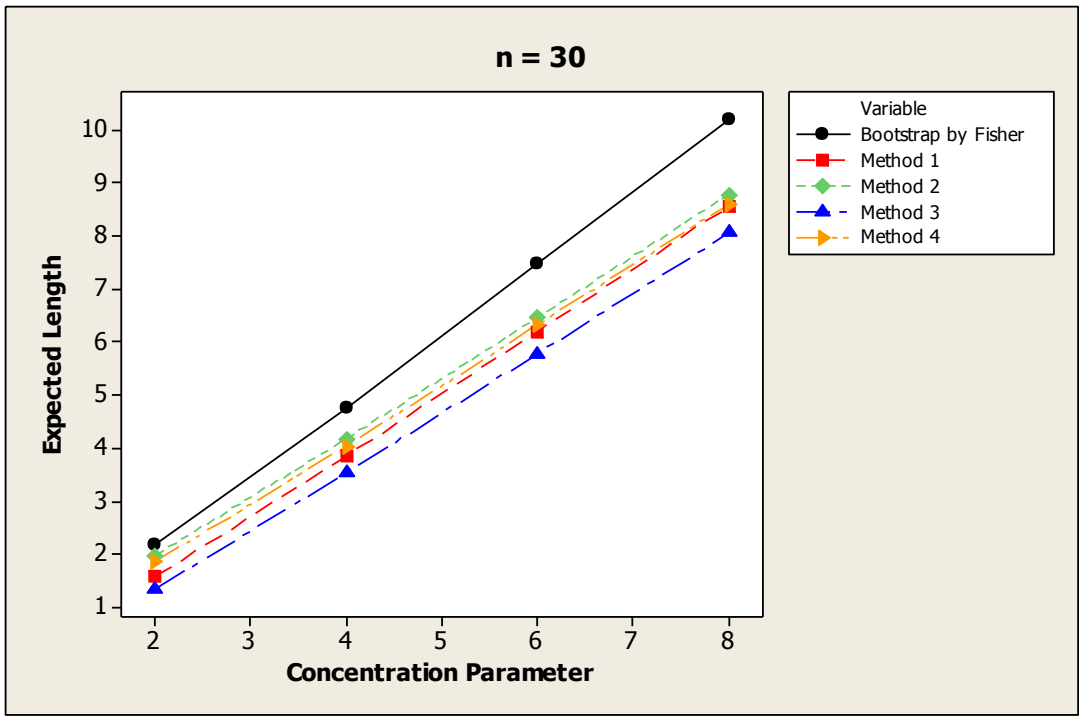

Figure 4.4: Expected length versus concentration parameter for *n* = 30

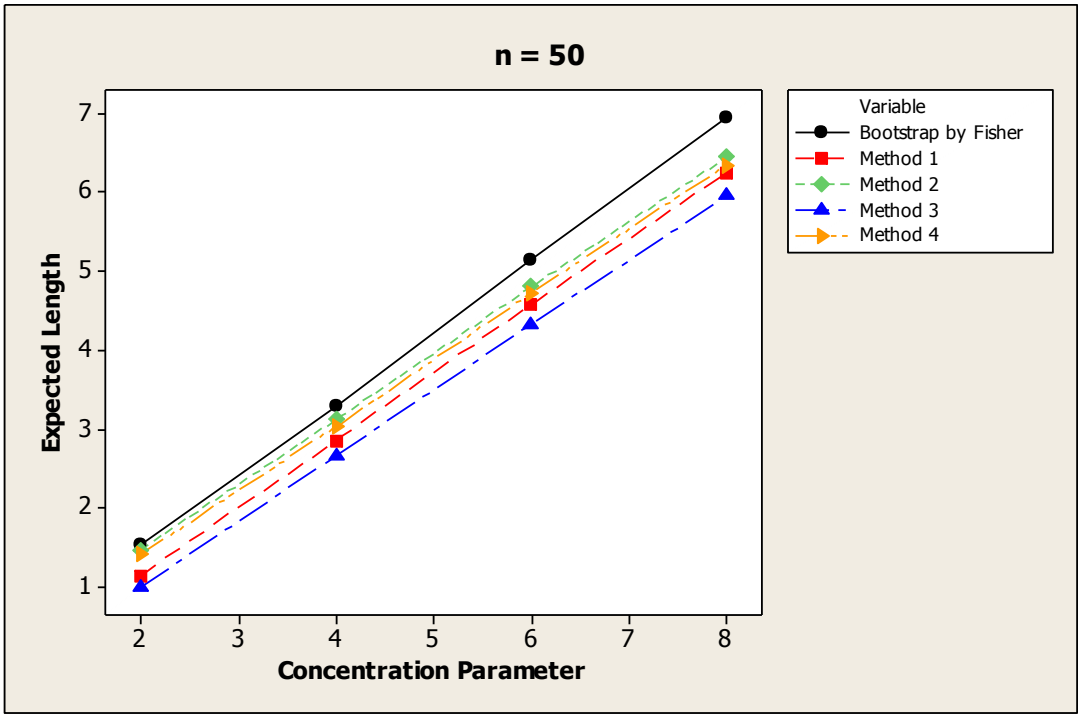

Figure 4.5: Expected length versus concentration parameter for *n* = 50

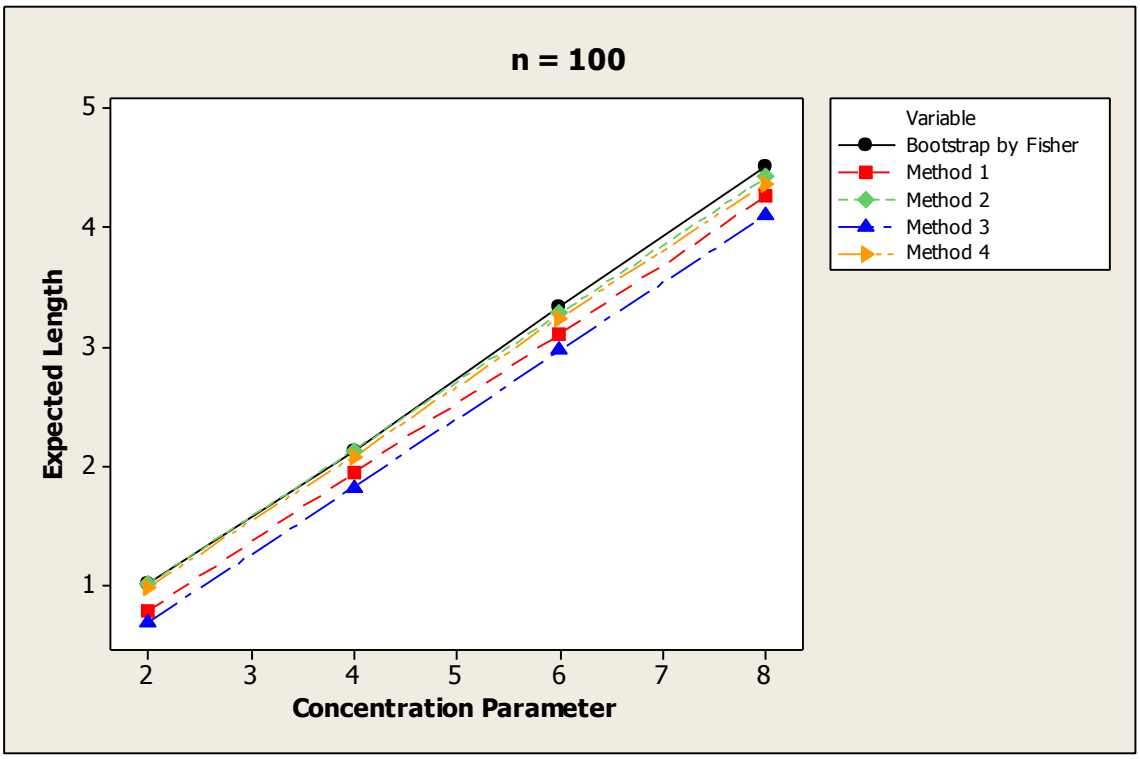

Figure 4.6: Expected length versus concentration parameter for n = 100

For further evaluation, we also consider the expected length for each method. Expected length can be defined as class size for each CI. Smaller values of expected length imply better approximation of CI as opposed to a wider length which represents a less efficient of the considered method. From Table 4.2, as the value of κ increases, it can be seen that the expected length for each method increases as well. It also shows that large concentration parameter results in larger expected length. It also noted that an increase of sample size results in a decrease of the expected length. In comparison of five methods, all the four proposed methods consistently give smaller length in comparison to the current method which is percentile bootstrap by Fisher (1993). Among the proposed method, Method 1 (CI based on circular variance population) and Method 4 (CI based on bootstrap-*t* method) give almost similar values of expected length. Method 2 (CI based on the asymptotic distribution of $\hat{\kappa}$) seems to give a large

expected length. From the results of the simulation study, it can be inferred that Method 3 is the superior method based on the performance of expected length.

For easy understanding, we can refer to Figures 4.4, 4.5 and 4.6 which represent the expected length for various values of the concentration parameter for sample size of 30, 50 and 100 respectively. It can be seen obviously that the current method (black line) lies above the rest of plots. As explained based on the results in Table 4.2, the blue line (Method 3) which represents the narrow length is always at the bottom of the other plots.

4.5 Illustrative Example

As an illustration of the proposed method, daily wind direction data (in radian) recorded at maximum wind speed (in m/s) was considered. Details of this data can be found in Section 2.6. Table 4.3 shows the concentration parameter and its upper and lower limit as well as their expected length for the five different methods including the current methods.

Table 4.3: Confidence intervals for wind direction data recorded at maximum wind speed at Kuala Terengganu

Data	Percentile Bootstrap By Fisher	Method 1	Method 2	Method 3	Method 4
Upper limit	3.380	3.027	2.749	2.838	2.678
Lower limit	7.342	6.026	6.003	5.630	5.891
Expected length	3.962	2.999	3.254	2.792	3.213

The concentration parameter calculated for the data is 4.931. Table 4.3 shows the upper and lower limit as well as their expected length for five different methods considered. It can be seen that the results obtained is consistent with the findings in the simulation studies. With reference to expected length, Method 3 (CI Based on Distribution of Mean Direction and Resultant Length) gives the smallest expected length in comparison to other methods. It also shows that the current method, namely the percentile bootstrap method gives the largest expected length of all the five methods considered. All these results seem to be similar with the findings from simulation studies.

4.6 Discussion

Several improved methods have been proposed for obtaining the CI of the concentration parameter for data with moderately large κ values in this study. All of the four methods proposed seems to perform relatively better than the existing method

by Fisher. Method 2 is superior in terms of simplicity in obtaining the CI, Method 4 is superior in terms of coverage probability and Method 3 is superior in terms of expected length.

Based on the performance using coverage probability, two methods namely Method 2 and Method 4 are both superior, Method 2 is appealing due to its simplicity in terms of calculating the CI. Yet, it is based on the asymptotic or limiting properties to the normal distribution. Method 4, however, preserves the original distribution, namely the von Mises distribution in obtaining the CI. Furthermore, based on the expected length, Method 4 performs much better than Method 2.

Alternatively, CI obtained by using the limiting property to the chi-square distribution namely Method 3 seems to do well in terms of the expected length. However, its performance seems to be somewhat average when coverage probability is concerned. Thus, this study provides several viable and improved methods of obtaining CI of the concentration parameter for data with high κ values.

In conclusion, several methods of obtaining CI for the concentration parameter for data with large values of κ are proposed. The proposed methods are viable and are improvements of the existing method by Fisher.

CHAPTER 5

A NEW STATISTIC BASED ON CIRCULAR DISTANCE

5.1 Introduction

This chapter proposes a new statistic for the von Mises distribution based on the circular distance between two observations. Sections 5.2 and 5.3 describes the proposed statistic and approximations to Chi-Squared distribution is discussed. This is followed by Section 5.4 which presents new confidence intervals based on the statistic that have been proposed in the previous section. Three different methods to estimate confidence intervals are discussed in this section. The simulation studies are carried out in Section 5.5 to assess the performance of the proposed method. In Section 5.6, the proposed method is illustrated using the wind direction data. Finally, discussion and conclusion of the whole chapter are given in Section 5.7.

5.2 Approximation to Chi Squared Distribution

In this section, we proposed a new statistic for a sample from von Mises distribution with large concentration parameter κ which can be approximated by Chi-squared distribution. Supposed $\theta_{1,\ldots},\theta_n$ are *i.i.d* circular sample located on the

circumference of a unit circle. Rao (1969) defined the circular distance between θ_i and θ_j as

$$d_{ij} = 1 - \cos\left(\theta_i - \theta_j\right). \tag{5.1}$$

On the other hand, Jammalamadaka and Sen Gupta (2001) gave an alternative definition at circular distance between two points θ_i and θ_j such that

$$\theta_{ij} = \pi - \left|\pi - \left|\theta_i - \theta_j\right|\right| \tag{5.2}$$

to ensure that θ_{ij} will take the smallest angle between θ_i and θ_j. The results of a new statistic are given here.

Proposition 1

Suppose $\theta_1,\ldots,\theta_n$ be *i.i.d* observation from a von Mises distribution with mean direction, μ and concentration parameter κ. Then for $j = 1,\ldots,n$,

$$G_j = \kappa\left[n - C\cos\theta_j - S\sin\theta_j\right] \sim \chi^2_{n-1} \text{ as } \kappa \to \infty, \tag{5.3}$$

where $C = \sum_{i=1}^{n}\cos\theta_i$ and $S = \sum_{i=1}^{n}\sin\theta_i$.

Proof

Suppose $\theta_1,\ldots,\theta_n$ is a random variable from $VM(\mu,\kappa)$. For any observation θ_i and large κ, it is shown by Jammalamadaka and SenGupta (2001) that,

$$\sqrt{\kappa}(\theta_i-\mu)\xrightarrow{d}N(0,1) \text{ as } \kappa\to\infty. \tag{5.4}$$

Since θ_i and θ_j are independent observations, then

$$\sqrt{\frac{\kappa}{2}}(\theta_i-\theta_j)\xrightarrow{d}N(0,1). \tag{5.5}$$

From the properties of the standard normal distribution, this can be approximated to Chi-squared distribution and it is given by

$$\frac{\kappa}{2}(\theta_i-\theta_j)^2\xrightarrow{d}\chi_1^2. \tag{5.6}$$

For large value of the concentration parameter, the distribution of von Mises distribution is said to be more concentrated. This highly concentrated distribution will lead to shorter circular distance between two points. From the second Taylor series expression, we have

$$\cos\alpha\approx 1-\frac{\alpha^2}{2} \text{ or } \frac{\alpha^2}{2}\approx 1-\cos\alpha. \tag{5.7}$$

Substitute for $\alpha=\theta_i-\theta_j$, we have,

$$\frac{\left(\theta_i-\theta_j\right)^2}{2} \approx 1-\cos\left(\theta_i-\theta_j\right)$$
$$=1-\cos\theta_i\cos\theta_j-\sin\theta_i\sin\theta_j. \tag{5.8}$$

Hence, by substituting (5.8) in (5.6), we have

$$\kappa\left(1-\cos\theta_i\cos\theta_j-\sin\theta_i\sin\theta_j\right)\sim\chi_1^2. \tag{5.9}$$

Further due to independent of θ_i and θ_j, for $i\neq j$,

$$\sum_{i\neq j}\kappa\left(1-\cos\theta_i\cos\theta_j-\sin\theta_i\sin\theta_j\right)\sim\chi_{n-1}^2. \tag{5.10}$$

or

$$\kappa\left[(n-1)-\cos\theta_j\sum_{i\neq j}\cos\theta_i-\sin\theta_j\sum_{i\neq j}\sin\theta_i\right]\sim\chi_{n-1}^2. \tag{5.11}$$

However, let $C=\sum_{i=1}^{n}\cos\theta_i=\sum_{i\neq j}^{n}\cos\theta_i+\cos\theta_j$ and $S=\sum_{i=1}^{n}\sin\theta_i=\sum_{i\neq j}^{n}\sin\theta_i+\sin\theta_j$.

Thus,

$$\kappa\left[(n-1)-\cos\theta_j\sum_{i\neq j}\cos\theta_i-\sin\theta_j\sum_{i\neq j}\sin\theta_i\right]$$

$$=\kappa\left[(n-1)-\cos\theta_j\left\{C-\cos\theta_j\right\}-\sin\theta_j\left\{S-\sin\theta_j\right\}\right]$$

$$=\kappa\left[n-1-C\cos\theta_j+\cos^2\theta_j-S\sin\theta_j+\sin^2\theta_j\right].$$

$$= \kappa\left[n - C\cos\theta_j - S\sin\theta_j \right] \tag{5.12}$$

Hence,

$$G_j = \kappa\left[n - C\cos\theta_j - S\sin\theta_j\right] \sim \chi^2_{(n-1)}. \tag{5.13}$$

In the following section, results of the Monte Carlo simulation studies are given for various values of the concentration parameter κ and sample size.

5.3 Simulation of the Approximated Chi-Squared Distribution

The proposed statistic in (5.13) can be used to approximate the sample from von Mises to Chi-squared distribution for large concentration parameter, κ. For this purpose, the Kolmogorov-Smirnov test is used to identify the suitable samples size as well as the concentration parameter that can be approximated. In this case, the sample sizes that will be considered are 10, 20, 30, 50, 70 and 100 respectively with concentration parameters 2, 4, 6, 8 and 10 respectively. Table 5.1 gives the percentage of the transformed sample values that follow Chi-Squared distribution with df (n - 1) as given in (5.3).

Table 5.1: The percentage of samples correctly approximated by the Chi-Squared distribution with df $(n-1)$.

Concentration parameter, κ	Sample size, n					
	10	20	30	50	70	100
2	53.2	64.4	82.1	99.0	100.0	100.0
4	49.8	67.0	85.9	99.9	100.0	100.0
6	50.0	67.0	89.0	99.8	100.0	100.0
8	49.7	70.8	90.3	100.0	100.0	100.0
10	48.4	69.8	90.9	100.0	100.0	100.0

From Table 5.1, the following results can be observed:

i. For $n = 10$, the percentage is a decreasing function for all κ.

ii. For the range $20 \le n \le 50$, the percentage is an increasing function for all κ, while constant for $70 \le n \le 100$.

iii. For any κ considered in the table and $n \ge 50$, more than 99% of the sample can be approximated to Chi-Squared distribution with df $(n - 1)$.

Based on these simulation studies, we may conclude that for any sample of size greater than 30 and $\kappa \ge 2$ can be approximated by Chi-Squared distribution with $(n-1)$ degree of freedom. In the following section, we describe the derivation of confidence intervals based on the new proposed statistic.

5.4 Estimation of Confidence Intervals (CI) for Concentration Parameter, κ

In this section, we propose the estimation of the confidence intervals of the concentration parameter based on the new proposed statistic.

Recall that $G_j = \kappa\left[n - C\cos\theta_j - S\sin\theta_j\right] \sim \chi^2_{(n-1)}$ as in (5.13). Hence, $100(1-\alpha)\%$ confidence intervals of κ is given by

$$\frac{\chi^2_{\left(n-1,\frac{\alpha}{2}\right)}}{\left[n - C\cos\theta_j - S\sin\theta_j\right]} < \kappa < \frac{\chi^2_{\left(n-1,1-\frac{\alpha}{2}\right)}}{\left[n - C\cos\theta_j - S\sin\theta_j\right]}. \tag{5.14}$$

Alternatively,

$$\frac{\chi^2_{\left(n-1,\frac{\alpha}{2}\right)}}{A_j} < \kappa < \frac{\chi^2_{\left(n-1,1-\frac{\alpha}{2}\right)}}{A_j}, \tag{5.15}$$

where $A_j = \left[n - C\cos\theta_j - S\sin\theta_j\right]$.

From (5.15), we will have a set of lower limits, $\kappa_1^L,\ldots,\kappa_n^L$ and upper limits, $\kappa_1^U,\ldots,\kappa_n^U$ respectively. In the following subsection, we consider three methods of estimating the confidence intervals based on the proposed statistic.

In the next subsection, three different methods namely mean, median and percentile will be considered.

5.4.1 Method 1: Mean

In this method, the mean for all pairs of confidence intervals will be taken as the final confidence interval. It can be obtained as below:

Lower limit = $\frac{1}{n}\sum_{j=1}^{n}\kappa_j^L$.

Upper limit = $\frac{1}{n}\sum_{j=1}^{n}\kappa_j^U$.

Hence, the proposed $100(1-\alpha)\%$ confidence intervals is given by

$$CI_{(mean)} = \left(mean\left(\kappa_j^L\right), mean\left(\kappa_j^U\right)\right). \tag{5.16}$$

5.4.2 Method 2: Median

The second method of estimating the confidence intervals is by considering the median for each set of lower and upper limit respectively. Suppose,

Lower limit = $med\left(\kappa_j^L\right)$ and upper limit = $med\left(\kappa_j^U\right)$

Hence, the $(1-\alpha)100\%$ confidence intervals is given by

$$\mathrm{CI}_{(med)} = \left(med\left(\kappa_j^L\right), med\left(\kappa_j^U\right)\right). \tag{5.17}$$

5.4.3 Method 3: Percentile

The third proposed method of CI is based on percentile. In order to get the confidence intervals based on percentile for each set of lower and upper limit respectively, three steps must be followed. The simulation study will be carried out to identify the most potential percentile that can be used as the CI. This can be done by considering the coverage probability and expected length of the concentration parameter. In order to obtain the final CI, these following steps must be followed:

Step 1: All values of the concentration parameter in lower limit and upper limit sets are sorted in ascending order. It then will be divided into various percentages for further evaluation.

Step 2: From the results, the most potential cut of point of percentile that will produce $(1-\alpha)100\%$ of target values is noted. We note that for $\alpha = 0.05$ or the 95% target values lie between 30th to 50th percentile.

Step 3: Finally, each new percentage in Step 2 will be examined to assess how well they produce the target value of 0.95 or 95% of CI.

Details description on identifying the most potential percentage to approximate the CI will be discussed in the simulation study in the subsequent section.

5.5 Simulation Study

Simulation studies were carried out to measure the performance of the proposed method for estimating confidence intervals of κ. The performance indicators used are coverage probability and expected length. Let *m* be the number of simulation and the following were calculated.

i. Coverage Probability = $\frac{q}{m}$, where *q* = number of true value falls in the CI .

ii. Expected Length = Upper limit – Lower limit.

The simulation studies were repeated for 5000 times. The first part of the simulation study is on identifying the feasible percentile when CI is constructed using percentiles method. This is followed by another simulation study in which performance of all the proposed methods are assessed.

5.5.1 Confidence Intervals based on percentile

As mentioned earlier, simulation studies were performed to identify the feasible percentile that contains the best CI or the one that gives the best coverage probability using the steps as described in Section 5.4.3. For this simulation studies, different samples from von Mises distribution with *n* = 30, 50, 70 and 100 with concentration parameters, $\kappa = 2, 4, 6, 8 \text{ and } 10$ will be used. This choice of parameter is based on the

results obtained in Table 5.1. Without loss of generality, the mean direction will be taken as 0 in this simulation study.

Table 5.2: Coverage probability for each percentage value for CI based on percentile

Sample size , n	κ	Percentile									
		10%	20%	**30%**	**40%**	**50%**	60%	70%	80%	90%	100%
30	2	0.234	0.774	**0.966**	**0.962**	**0.920**	0.846	0.752	0.676	0.622	0.596
	4	0.234	0.712	**0.944**	**0.944**	**0.870**	0.798	0.676	0.556	0.498	0.458
	6	0.212	0.700	**0.912**	**0.932**	**0.840**	0.748	0.654	0.538	0.448	0.406
	8	0.220	0.708	**0.918**	**0.930**	**0.856**	0.758	0.630	0.512	0.420	0.382
	10	0.258	0.728	**0.928**	**0.924**	**0.856**	0.736	0.620	0.488	0.394	0.332
50	2	0.090	0.660	**0.952**	**0.956**	**0.862**	0.738	0.596	0.478	0.412	0.376
	4	0.098	0.660	**0.934**	**0.924**	**0.786**	0.596	0.398	0.260	0.198	0.164
	6	0.112	0.636	**0.934**	**0.920**	**0.756**	0.550	0.356	0.234	0.156	0.120
	8	0.096	0.614	**0.926**	**0.910**	**0.782**	0.566	0.334	0.208	0.138	0.108
	10	0.114	0.580	**0.924**	**0.904**	**0.762**	0.554	0.348	0.208	0.142	0.112
70	2	0.032	0.566	**0.956**	**0.940**	**0.814**	0.576	0.410	0.284	0.218	0.176
	4	0.038	0.514	**0.924**	**0.902**	**0.736**	0.440	0.232	0.122	0.068	0.056
	6	0.058	0.562	**0.918**	**0.886**	**0.648**	0.368	0.164	0.088	0.052	0.040
	8	0.066	0.566	**0.918**	**0.888**	**0.622**	0.314	0.138	0.058	0.026	0.022
	10	0.040	0.526	**0.918**	**0.902**	**0.656**	0.344	0.166	0.078	0.032	0.030
100	2	0.010	0.422	**0.950**	**0.936**	**0.728**	0.414	0.230	0.130	0.072	0.050
	4	0.014	0.430	**0.904**	**0.878**	**0.578**	0.252	0.088	0.038	0.028	0.026
	6	0.008	0.470	**0.918**	**0.842**	**0.506**	0.196	0.058	0.014	0.002	0.000
	8	0.012	0.480	**0.912**	**0.834**	**0.488**	0.196	0.054	0.008	0.000	0.000
	10	0.016	0.434	**0.928**	**0.836**	**0.472**	0.170	0.042	0.008	0.002	0.002

Table 5.2 and 5.3 showed the coverage probability and expected length respectively calculated from the simulation studies for various intervals considered with different sample size and concentration parameter. Coverage probability can be defined as the proportion of a number that the CI contains the true value. In other words, the coverage probability is the actual probability that the interval contains the true concentration parameter. The simulation studies have been done at 95% of confidence level. Thus, using the measure of performance of coverage probability, values of coverage probability close to 0.95 is indicative of a good performance and will be referred to as nominal coverage probability or good target value.

Table 5.3: Expected length for each percentage value for CI based on percentile

Sample size, *n*	κ	Percentile									
		10%	20%	**30%**	**40%**	**50%**	60%	70%	80%	90%	100%
30	2	1.124	1.555	**1.938**	**2.286**	**2.602**	2.895	3.148	3.341	3.476	3.548
	4	2.246	3.186	**4.007**	**4.810**	**5.588**	6.332	6.985	7.508	7.879	8.058
	6	3.289	4.659	**5.897**	**7.147**	**8.323**	9.498	10.513	11.355	11.927	12.210
	8	4.400	6.313	**7.995**	**9.718**	**11.358**	12.950	14.323	15.472	16.312	16.740
	10	5.442	7.767	**9.914**	**12.057**	**14.137**	15.990	17.772	19.212	20.244	20.794
50	2	0.873	1.221	**1.518**	**1.778**	**2.030**	2.246	2.429	2.576	2.671	2.717
	4	1.742	2.418	**3.071**	**3.680**	**4.262**	4.799	5.277	5.660	5.923	6.043
	6	2.584	3.639	**4.615**	**5.588**	**6.491**	7.345	8.119	8.737	9.175	9.376
	8	3.442	4.870	**6.174**	**7.434**	**8.692**	9.869	10.952	11.781	12.378	12.649
	10	4.305	6.068	**7.711**	**9.266**	**10.856**	12.386	13.777	14.854	15.612	15.962
70	2	0.741	1.029	**1.272**	**1.494**	**1.704**	1.886	2.040	2.161	2.239	2.272
	4	1.479	2.065	**2.591**	**3.099**	**3.597**	4.049	4.447	4.764	4.984	5.076
	6	2.174	3.048	**3.854**	**4.659**	**5.423**	6.141	6.769	7.280	7.625	7.774
	8	2.878	4.010	**5.070**	**6.113**	**7.137**	8.131	8.983	9.701	10.173	10.383
	10	3.646	5.119	**6.466**	**7.818**	**9.165**	10.419	11.516	12.418	13.034	13.299
100	2	0.621	0.848	**1.055**	**1.239**	**1.413**	1.565	1.690	1.786	1.848	1.874
	4	1.214	1.678	**2.116**	**2.535**	**2.940**	3.306	3.619	3.871	4.041	4.111
	6	1.819	2.553	**3.240**	**3.885**	**4.509**	5.108	5.640	6.059	6.329	6.437
	8	2.431	3.411	**4.313**	**5.200**	**6.073**	6.886	7.609	8.202	8.577	8.743
	10	3.020	4.249	**5.398**	**6.523**	**7.611**	8.630	9.552	10.312	10.803	11.008

For further evaluation, we also consider the expected length for each method. Expected length can be defined as class size for each CI. Smaller values of expected length imply better estimation of CI as opposed to wider length which represents the less efficient of the proposed method. The following tables show the simulation results to determine the most potential percentile for the CI.

From Table 5.2, it can be seen that the for any fixed n and κ, coverage probability increases steadily as the value of percentile increase and reaches almost the target value of 0.95 for 30th, 40th and 50th percentile respectively then further decreases from 60th percentile till 100th percentile onwards. It also can be observed that, as the sample size increase, the coverage probability also decreases. For different values of the concentration parameter and sample size, it can be seen that the coverage probability are stable for each percentage values. As explained in Section 5.4.3, our main purpose is to identify the most potential percentile which gives the closest value to our 95% target value. It can be seen clearly that 30th to 50th percentile give the values that are close to 0.95. Hence, these particular percentiles are chosen to be used in estimating the confidence intervals of the concentration parameter, κ.

Apart from assessing the coverage probability, the expected lengths also have been considered. From Table 5.3, for any fixed n and κ, the expected length is an increasing function of the percentile. For any fixed κ, the expected length decreases as sample size increases. For any fixed n, the expected length increases as the concentration parameter value increases. Here, we take note of the behavior of expected lengths for the range of 30th to 50th percentile of the concentration parameter, κ.

For further evaluation, the intervals are further subdivided into various small intervals that lead to new, and these new percentiles will be compared with another method namely mean and median as discussed in Sections 5.3.1 and 5.3.2 respectively.

5.5.2 Confidence Intervals of Concentration Parameter, κ based on Mean, Median and Percentile

The second part of the simulation study is to compare the performance of the three proposed methods of obtaining the CI. For the comparison study, new simulation studies were carried out for different sample size, namely n = 30, 50 and 100 with various value of concentration parameter, κ = 2.0, 4.0, 6.0, 8.0 and 10.0 respectively for the confidence level at α = 0.05. The samples were drawn from von Mises distribution and without loss of generality, the mean direction will be taken as 0 during the simulation study. The purpose of this simulation studies is to find the most efficient CI for the concentration parameter, κ. As for performance indicator, the coverage probability and the expected length will be used. The simulation results are given in the following tables.

Tables 5.4 and 5.5 show the coverage probability and the expected length respectively obtained from the simulation studies at $\alpha = 0.05$. For the display and analysis purpose, we only include the 30th, 34th and 38th percentile as these percentile values give better coverage probabilities which they are close to target value in comparison to other percentiles in the range of 30th to 50th percentile. Hence, the results from these percentiles will be used as the comparison with the mean and median method.

Table 5.4: Coverage probability for various value of κ for each sample size, *n* = 30, 50, 70 and 100 at $\alpha = 0.05$.

Sample size, *n*	Concentration parameter, κ	Mean	Median	Percentile		
				30%	34%	38%
30	2	0.945	0.913	0.954	0.968	0.969
	4	0.904	0.862	0.927	0.945	0.948
	6	0.884	0.841	0.919	0.939	0.941
	8	0.880	0.833	0.921	0.938	0.940
	10	0.876	0.831	0.911	0.932	0.936
50	2	0.919	0.856	0.954	0.970	0.965
	4	0.844	0.765	0.923	0.943	0.934
	6	0.819	0.739	0.925	0.943	0.927
	8	0.805	0.725	0.917	0.939	0.925
	10	0.799	0.726	0.915	0.936	0.917
70	2	0.896	0.804	0.950	0.970	0.958
	4	0.796	0.689	0.918	0.946	0.922
	6	0.747	0.632	0.918	0.936	0.904
	8	0.726	0.621	0.915	0.936	0.899
	10	0.723	0.618	0.914	0.934	0.895
100	2	0.852	0.706	0.943	0.971	0.954
	4	0.713	0.564	0.916	0.946	0.913
	6	0.646	0.499	0.923	0.936	0.890
	8	0.622	0.478	0.914	0.929	0.880
	10	0.606	0.476	0.911	0.931	0.877

Table 5.5: Expected length for various value of κ for each sample size, *n* = 30, 50, 70 and 100 at $\alpha = 0.05$.

Sample size, *n*	Concentration parameter, κ	Mean	Median	Percentile		
				30%	34%	38%
30	2	2.528	2.700	1.967	2.089	2.208
	4	5.405	5.708	4.004	4.277	4.550
	6	8.315	8.741	6.052	6.475	6.895
	8	11.177	11.747	8.089	8.657	9.234
	10	14.048	14.739	10.129	10.849	11.560
50	2	1.912	2.045	1.508	1.618	1.723
	4	4.080	4.313	3.064	3.308	3.552
	6	6.263	6.592	4.629	5.011	5.392
	8	8.429	8.859	6.192	6.707	7.224
	10	10.569	11.090	7.730	8.382	9.030
70	2	1.592	1.704	1.263	1.362	1.457
	4	3.391	3.588	2.562	2.781	2.999
	6	5.234	5.519	3.893	4.237	4.578
	8	7.041	7.405	5.197	5.664	6.127
	10	8.814	9.266	6.494	7.070	7.653
100	2	1.324	1.417	1.053	1.129	1.203
	4	2.811	2.975	2.135	2.306	2.473
	6	4.330	4.568	3.238	3.504	3.770
	8	5.833	6.138	4.325	4.685	5.046
	10	7.317	7.690	5.405	5.863	6.312

Table 5.4 shows the coverage probability for all three different methods namely mean, median and percentile for different sample size and concentration parameter. For any fixed κ, it can be seen that the coverage probability decreases as the sample size increases for all the three different methods considered. The coverage probability is also a decreasing function of the concentration parameter for any fixed value of sample size. It is noted that the median gives the poorest performance in which the coverage

probability is far from the target values in comparison to the other methods. The results indicate that percentile method is the best method because the coverage probabilities are consistently close to the target values. From three different percentiles considered in this study, it can be seen that for $\kappa = 2$, 30th percentile gives the coverage probability that is close to the target value. For $n = 30$, the coverage probability for 34th and 38th percentile give almost similar values and close to the target value. For all sample sizes and κ, it is noted that 34th percentile consistently gives the best coverage probability with values close to the target value in comparison to other percentiles as well as the mean and median. Hence, it can be said that, using the coverage probability as the performance indicator, percentile method is superior to the mean and median. The best percentile to be used is at 34th for all κ while for $\kappa = 2$, the percentile at 30th may be considered to be good as well.

Table 5.5 shows the expected length obtained from the simulation studies. It can be observed that the median gives the widest length in comparison to the other methods. For any fixed *n*, the expected length is an increasing function of the concentration parameter, and for any fixed κ, it is a decreasing function of the sample size. In addition, it can be seen that percentile method gives the narrowest expected length as compared to the other methods. Hence, it also can be concluded that percentile method is the superior method as compared to the other methods using expected length as the performance indicator.

Thus, using both measures of performance namely coverage probability and expected length, it can be concluded that 34th percentile gives the most efficient CI in comparison to all methods as it give good coverage probability as well narrow expected length.

5.6 Illustrative Example

As an illustration of the proposed method, two simulated data set will be considered. The data were generated at two different sample sizes which are n = 50 and 70 with mean direction 0 and the concentration parameter, $\kappa = 6$. The upper and lower limits, as well as its expected length, are recorded in the following table.

Table 5.6: Confidence intervals for simulated based on new statistic for circular distance

Sample size, n		CI based on Mean	CI based on Median	CI based on Percentile at 0.34
50	Upper limit	4.876	5.016	3.879
	Lower limit	10.850	11.164	8.632
	Expected length	5.975	6.147	4.753
70	Upper limit	5.059	5.342	4.286
	Lower limit	9.907	10.462	8.393
	Expected length	4.848	5.120	4.108

Table 5.6 shows the upper and lower limits and the expected length for both simulated data. The value of expected length obtained based on 34th percentile is the smallest among three methods. It can be seen that the CI based on median gives the widest length in comparison to other methods. These results support the findings from the simulation studies where CI based on percentile will give better estimate than other methods. Considering both data, it can be concluded that CI based on percentile gives a precise of the CI for the concentration parameter.

5.7 Discussion

A new statistics based on the circular distance for a sample from von Mises distribution was proposed in this chapter. Based on the statistic proposed, new approaches to approximate the CI for the concentration parameter are developed. These three methods are CI based on mean, median and percentile. For CI based on percentile, three steps must be followed before a final percentile that gives the most efficient CI is obtained. Based on the simulation study, it is observed that the range of percentile that gives values that are close to the target value, 0.95 is from 30th to 50th percentile. A second simulation study was further carried out to assess the performance of each proposed method. From the simulation studies, it can be seen that the CI based on percentile consistently gives good coverage probability as well as the smallest expected length. The superiority of the CI obtained using percentile is also illustrated using real data sets. Hence, it can be concluded that the method based on percentile is the best to approximate the CI for the concentration parameter based on circular distance.

In summary, the contribution of a new statistics developed in this study is illustrated by the construction of new CI. All the three proposed methods of obtaining CI provide alternate approaches and are appealing due to the simplicity of getting the CI. However, based on simulation studies, CI based on percentile is the most superior of the three proposed method. Nevertheless, the three methods of constructing CI provides an alternative approach and have great potential for improvement in further works.

CHAPTER 6

ANALYSIS OF MISSING VALUES FOR CIRCULAR VARIABLES

6.1 Introduction

This chapter discusses on the analysis of missing values for circular data. It commences with a brief background discussion on missing values in Section 6.2. In Section 6.3, imputation methods of missing values for circular data are presented. To assess the accuracy of the methods considered, simulation studies are carried out in Section 6.4. In Section 6.5, the proposed data imputation method is illustrated using the Malaysian wind direction data. Finally, discussion and conclusion of the whole chapter are given in Section 6.6.

6.2 Background

Missing values is a common problem in data analysis. This kind of problems has been addressed as many in various research fields. As described in the literature review (Chapter 2), the missing values can be classified as missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). In this study, all missing values are treated under MAR case because it has wide application in practical life and require less assumption. Furthermore, MCAR is not a reasonable

assumption for missing data in many real situations (Little & Rubin, 2002; Raghunathan, 2004)

Many integrated approaches have been developed in handling missing values which can be classified as the traditional and modern approaches. Traditional approaches include listwise deletion, pairwise deletion and simple replacement procedures. On the other hand, several modern approaches are applied where some of them are integrated from the traditional approach. Imputation is one of the modern approaches and it is a class of methods by which estimation of the missing value of its distribution is used to generate predictions from a given model (Tsechansky & Provost, 2007).

In most common cases, deletion is the simplest way to deal with missing values. By deletion, it will lead to a complete data set and the usual analysis can be done. However, this approach decreases the sample size of data and at the same time will reduce the power of statistics which in turn, results in biased estimates when the excluded group is a selective subsample from the study population (Barzi & Woodward, 2004). Therefore, new integrated methods are needed in order to overcome this problem. The most popular methods that normally used is the replacement procedure. Replacement procedure (Tsikriktis, 2005) includes mean substitution, hot-deck imputation and regression imputation. The simplest replacement method is by using mean substitution where all the missing values are replaced with the mean of available observations.

As discussed in the literature review previously, numerous methods have been developed to handle missing values for linear data. For circular data, however, we found that it is somewhat limited. This is might be due to the complexity and topology of the

circular data itself as well as limited statistical software available to analyse such data. Hence, in this study we focus on handling the occurrence of missing data in univariate data with von Mises distribution. Von Mises has been extensively studied, and many inference techniques have been developed. Thus, this model usually considered for circular data in most applied problem (Jamalamadaka & SenGupta, 2001).

In the following section, three widely used methods of data imputation for circular data are considered. They are the mean, EM algorithm and DA algorithm. Mean imputation is chosen over the traditional approach as comparison method because of its simplicity as well as the most reliable method to be applied in this distribution. Apart from the traditional approach, modern approach in particular EM and DA are considered as these methods are proven to be excellent methods for handling missing data in various situation (Allison, 2002).

Using the simulation study, the performance of the considered methods will be measured using several indicators namely circular distance for parameter mean direction and bias as well as estimated root mean square error (ERMSE) for concentration parameter.

6.3 Data Imputation of Missing Values for circular data

As mentioned earlier, the study of missing values is confined to univariate circular variables, namely variables from the von Mises distribution. Here, several imputation methods are described namely the by circular mean, Expectation-Maximization (EM) algorithm and data augmentation (DA) algorithm.

6.3.1 Circular Mean

Imputation by circular mean is a common method used when some observations are missing. In this method, all missing values are replaced with the circular mean calculated from the available data. The steps in carrying out the imputation procedure are described as follows:

1. Generate a random number from von Mises distribution $X \sim VM\left(n,0,\kappa\right)$ where *n* is number of sample size.
2. Distribute *q* missing values, X_{mis} in data set

$$\begin{matrix} x_{mis1} \\ \vdots \\ x_{misq} \\ \\ x_{q+1} \\ \vdots \\ x_{n} \end{matrix}$$

3. Calculating the initial parameter based on available non-missing data.

   Mean direction,

$$\hat{\mu}^{(0)} = \begin{cases} \tan^{-1}\left(\dfrac{S}{C}\right) & S>0, C>0 \\ \tan^{-1}\left(\dfrac{S}{C}\right)+\pi & C<0\,, \\ \tan^{-1}\left(\dfrac{S}{C}\right)+2\pi & S<0, C>0 \end{cases} \tag{6.1}$$

where $C = \sum_{i=q+1}^{n} \cos(x_i)$ and $C = \sum_{i=q+1}^{n} \cos(x_i)$.

4. The first cycle of complete data can be obtained by imputing the missing values with the initial circular mean obtained in Step (3) and the new parameter which is mean direction and concentration parameter is calculated.

$$\begin{matrix} \hat{\mu}^{(0)} \\ \vdots \\ \hat{\mu}^{(0)} \end{matrix}$$

$$\begin{matrix} x_{q+1} \\ \vdots \\ x_n \end{matrix}$$

6.3.2 EM algorithm

As an alternative to the conventional method, we also applied a common modern approach of imputing the missing values in von Mises distribution namely the EM algorithm. EM algorithm was first introduced by Dempster *et al.* (1977). In simplest way, this algorithm can be defined as 'fill in' the missing data based on the initial estimate, re-estimate the parameter based on available data and then fill in again iteratively till the estimates converge. A brief explanation and examples of EM algorithm in a linear case can be found in the literature review in Section 2.5.

There are two steps in the EM algorithm which can be called as Expectation or *E-step* and Maximization or *M-step* (Dempster *et al.*, 1977). The general steps used in EM in this study are described as follows:

Step 1: Expectation or E-step

The E-step of EM is replacing the missing values observations, X_{mis} which require the estimation of $\theta^{(t)}$ to obtain complete data, when X_{obs} is given.

Step 2: Maximization or M-step

In this step, $\theta^{(t+1)}$ is re-estimated by maximum likelihood based on X_{obs} and $\theta^{(t)}$ obtained in Step (1).

Steps (1) – (2) will be repeated iteratively until $\theta^{(t)}$ and $\theta^{(t+1)}$ satisfied our convergence criteria and converge to a local maximum of the likelihood function.

In this particular study, the EM algorithm was performed by using the following steps:

i. E-Step

In this step, the expectation value is calculated from the non-missing values. This value is then used to impute all missing values. For example, the initial mean will be calculated as in (6.1).

Hence, the first cycle of complete data set is obtained by imputing an initial circular mean calculated using (6.1)

$$\begin{matrix} \hat{\mu}^{(0)} \\ \vdots \\ \hat{\mu}^{(0)} \\ x_{q+1} \\ \vdots \\ x_n \end{matrix}$$

Thus, this E-step can be generalised as follow:

$$\begin{matrix} \hat{\mu}^{(j-1)} \\ \vdots \\ \hat{\mu}^{(j-1)} \\ x_{q+1} \\ \vdots \\ x_n \end{matrix}$$

$j = 1 ,..., q$

ii. M-Step

After doing an imputation, the new complete data set will be obtained. The estimation of the new parameter will be calculated, and the steps will be repeated iteratively until the convergence criteria satisfied to get the final estimate.

$$\hat{\mu}^{(j)} = \begin{cases} \tan^{-1}\left(\dfrac{S^{(j)}}{C^{(j)}}\right) & S^{(j)} > 0, C^{(j)} > 0 \\ \tan^{-1}\left(\dfrac{S^{(j)}}{C^{(j)}}\right) + \pi & C^{(j)} < 0 \\ \tan^{-1}\left(\dfrac{S^{(j)}}{C^{(j)}}\right) + 2\pi & S^{(j)} < 0, C^{(j)} > 0 \end{cases}, \tag{6.2}$$

where $C^{(j)} = \sum_{i=q+1}^{n} \cos(x_i) + q\cos\left(\hat{\mu}^{(j-1)}\right)$ and $S^{(j)} = \sum_{i=q+1}^{n} \sin(x_i) + q\sin\left(\hat{\mu}^{(j-1)}\right)$ and

$$\hat{\kappa}^{(j)} = \begin{cases} 2\bar{R}^{(j)} + \bar{R}^{(j)3} + \dfrac{5\bar{R}^{(j)5}}{6}, & \bar{R}^{(j)} < 0.53 \\ -0.4 + 1.39\bar{R}^{(j)} + \dfrac{0.43}{1-\bar{R}^{(j)}}, & 0.53 \le \bar{R}^{(j)} < 0.85, \\ \dfrac{1}{\bar{R}^{(j)3} - 4\bar{R}^{(j)2} + 3\bar{R}^{(j)}}, & \bar{R}^{(j)} \ge 0.85, \end{cases} \tag{6.3}$$

where $\bar{R}^{(j)} = \left(\bar{C}^{(j)2} + \bar{S}^{(j)2}\right)^{\frac{1}{2}}, \bar{C}^{(j)} = \dfrac{1}{n-q}\sum_{i=q+1}^{n} \cos(x_i) + \cos\left(\hat{\mu}^{(j-1)}\right)$

and $\bar{S}^{(j)} = \dfrac{1}{n-q}\sum_{i=q+1}^{n} \sin(x_i) + \sin\left(\hat{\mu}^{(j-1)}\right).$

6.3.3 Data Augmentation (DA) algorithm

Another method of imputing missing values that are considered in this study is (DA) algorithm. DA algorithm was first proposed by Taner and Wong (1987). There are two steps in this method namely I-step and P-step. Briefly, the steps are described below:

Step 1: Imputation or I-step

Given a current guess of a parameter as $\Theta^{(t)}$, draw independent *q* values of X_{mis}

$$X_{mis}^{(t+1)} = \left(x_{mis1}^{(t+1)}, x_{mis2}^{(t+1)}, \ldots, x_{misq}^{(t+1)} \right), \tag{6.4}$$

generated from the conditional predictive distribution of X_{mis}

$$X_{mis}^{(t+1)} \sim P\left(X_{mis} \mid X_{obs}, \theta^{(t)} \right). \tag{6.5}$$

Step 2: Posterior or P-step

Draw new values of θ

$$\theta^{(t+1)} = \left(\theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{n_0}^{(t+1)} \right), \tag{6.6}$$

which is calculated from the conditional distribution of X_{obs} and $X_{mis}^{(t+1)}$

$$\theta^{(t+1)} \sim P(\theta \mid X_{obs}, X_{mis}^{(t+1)}). \tag{6.7}$$

Steps (1) – (2) will be repeated from the initial value $\Theta^{(0)}$ for a value of *t*

$$\left\{ \left(\Theta^{(t)}, X_{mis}^{(t)} \right) : t = 1, 2, \ldots \right\}, \tag{6.8}$$

which have $P(\theta, X_{obs} \mid X_{mis})$ as its stationary distribution (Schafer, 2010). Steps (1) – (2) will be repeated until $\theta^{(t)}$ and $\theta^{(t+1)}$ converge to our pre-set convergence criteria.

DA algorithm is somewhat similar to EM algorithm where E-step of EM calculates the expected complete-data while I-step in DA simulates a random draw of the complete data (Schafer, 1997). Discussion on DA algorithm and its application in missing values data analysis can be found in Section 2.4.

In this study, these following steps are used.

i. I-step

In I-step, q number of missing values will be generated using the initial parameter. Those generated data is then to be used in replacing the missing values to get the complete data set. For example, the first new complete data set can be obtained as follows. These values are then to be used to impute the missing values in the data set.

$Y^{(1)} \sim VM\left(q, \hat{\mu}^{(0)}, \hat{\kappa}^{(0)}\right)$, q = number of missing values

$$\begin{matrix} y_1^{(1)} \\ \vdots \\ y_q^{(1)} \\ x_{q+1} \\ \vdots \\ x_n \end{matrix}$$

Thus, I-step in DA algorithm can be generalised as follows

$Y^{(j)} \sim VM\left(q, \hat{\mu}^{(j-1)}, \hat{\kappa}^{(j-1)}\right)$, q = number of missing values

$$
\begin{matrix}
y_1^{(j)} \\
\vdots \\
y_q^{(j)} \\
x_{q+1} \\
\vdots \\
x_n
\end{matrix}
$$

where $j = 1, \ldots, q$.

ii. P-step

In P-step, the estimation of new parameter estimates will be calculated based on the completed data set obtained from I-step. The new parameter estimate obtained from complete data set is then to be used again for re-generating new imputed values using I-step. These two steps will be repeated until the convergence criteria satisfied to get the final estimate. Our P-step can be generalised as follow:

$$
\hat{\mu}^{(j)} = \begin{cases}
\tan^{-1}\left(\dfrac{S^{(j)}}{C^{(j)}}\right) & S^{(j)} > 0, C^{(j)} > 0 \\
\tan^{-1}\left(\dfrac{S^{(j)}}{C^{(j)}}\right) + \pi & C^{(j)} < 0 \\
\tan^{-1}\left(\dfrac{S^{(j)}}{C^{(j)}}\right) + 2\pi & S^{(j)} < 0, C^{(j)} > 0
\end{cases}, \qquad (6.9)
$$

where $C^{(j)} = \sum_{i=q+1}^{n} \cos(x_i) + \sum_{i=1}^{q} \cos(y_i^{(j)})$ and $S^{(j)} = \sum_{i=q+1}^{n} \sin(x_i) + \sum_{i=1}^{q} \sin(y_i^{(j)})$ and

$$\hat{\kappa}^{(j)} = \begin{cases} 2\bar{R}^{(j)} + \bar{R}^{(j)3} + \dfrac{5\bar{R}^{(j)5}}{6}, & \bar{R}^{(j)} < 0.53 \\ -0.4 + 1.39\bar{R}^{(j)} + \dfrac{0.43}{1-\bar{R}^{(j)}}, & 0.53 \le \bar{R}^{(j)} < 0.85, \\ \dfrac{1}{\bar{R}^{(j)3} - 4\bar{R}^{(j)2} + 3\bar{R}^{(j)}}, & \bar{R}^{(j)} \ge 0.85, \end{cases} \tag{6.10}$$

where $\bar{R}^{(j)} = \left(\bar{C}^{(j)2} + \bar{S}^{(j)2}\right)^{\frac{1}{2}}$, $\bar{C}^{(j)} = \dfrac{1}{n-q}\sum_{i=q+1}^{n}\cos(x_i) + \dfrac{1}{q}\sum_{i=1}^{q}\cos(y_i^{(j)})$,

$\bar{S}^{(j)} = \dfrac{1}{n-q}\sum_{i=q+1}^{n}\sin(x_i) + \dfrac{1}{q}\sum_{i=1}^{q}\sin(y_i^{(j)})$.

6.4 Simulation Studies

Simulation studies were carried out in order to evaluate the performance for each proposed method. In this study, our particular interest is in investigating the estimation of parameter in von Mises distribution, μ and κ after doing the imputation methods mentioned earlier. For this purpose, programmes are written using S-Plus. Three different method considered will be referred to as Method 1 (circular mean), Method 2 (EM algorithm) and Method 3 (DA algorithm). For each sample, we randomly assign 5%, 10%, 15%, 20%, 30% and 40% of the missing values, respectively. The simulation studies are repeated for 5000 times, and the values of *X* have been drawn from $X \sim VM(0, \kappa)$ where κ = 2, 4, 6 and 8 with different sample sizes, *n* = 30, 50 and 100.

As for the performance measures, the circular mean and circular distance (*d*) were calculated for parameter μ since this parameter is in circular form. For

concentration parameter κ, the mean, estimate bias (EB), and estimate root mean square error (ERMSE) were calculated.

Circular mean for the mean direction is calculated using the following formulae:

$$\bar{\hat{\mu}} = \begin{cases} \tan^{-1}\left(\frac{S}{C}\right) & S > 0, C > 0 \\ \tan^{-1}\left(\frac{S}{C}\right) + \pi & C < 0 \\ \tan^{-1}\left(\frac{S}{C}\right) + 2\pi & S < 0, C > 0 \end{cases}, \tag{6.11}$$

where $C = \sum \cos\left(\hat{\mu}_j\right)$ and $S = \sum \sin\left(\hat{\mu}_j\right)$.

Circular distance is the smaller measurements of the two arclengths between the points along the circumference. The value of circular distance is in the range between $[0, \pi]$. A smaller value of circular distance obtained shows a better estimation.

$$\text{Circular Distance, } d = \pi - \left|\pi - \left|\bar{\hat{\mu}} - \mu\right|\right|. \tag{6.12}$$

For concentration parameter, the mean of κ is obtained from the simulation study and given by

$$\text{Mean, } \bar{\hat{\kappa}} = \frac{1}{\text{simu}} \sum \hat{\kappa}_j \text{ ,} \tag{6.13}$$

where simu = number of simulation.

The mean is then to be used in calculating the estimate bias (EB). The EB is the absolute difference between the estimated parameter that we obtained from the complete data set and the data after imputing the missing values. The EB can be defined as:

$$\text{Estimated Bias, } \text{EB} = \left| \bar{\hat{\kappa}} - \kappa \right| . \tag{6.14}$$

Estimated Root Mean Square Errors (ERMSE) is the one that frequently used measure of the difference between values predicted by a model and the values observed. In this study, the ERMSE is calculated by

$$\text{ERMSE} = \sqrt{\frac{1}{\text{simu}} \sum \left(\hat{\kappa}_j - \kappa \right)^2} . \tag{6.15}$$

Tables 6.1 – 6.3 show the simulation results of the mean direction obtained for three different sample sizes, 30, 50 and 100 respectively using all three methods considered. Method 1 refers to the imputation by the conventional method using circular mean, Method 2 is imputation by EM algorithm while Method 3 is by DA algorithm. The simulation studies were carried out using the steps as described in Section 6.3.

Table 6.1 (a): Simulation results for mean direction for sample size, *n* = 30

Performance Indicator	Concentration Parameter	Percentage of missing values	Method 1	Method 2	Method 3
Circular Mean	2	5%	0.016	6.281	6.281
		10%	0.027	0.001	0.001
		15%	0.037	0.004	0.001
		20%	0.051	0.004	0.004
		30%	0.071	0.006	0.007
		40%	0.083	0.000	0.001
	4	5%	0.014	6.282	6.283
		10%	0.023	0.002	0.002
		15%	0.028	6.282	6.282
		20%	0.041	0.001	0.000
		30%	0.058	0.000	6.283
		40%	0.074	0.002	0.003
	6	5%	0.043	0.001	0.002
		10%	0.022	0.001	0.001
		15%	0.029	0.002	0.002
		20%	0.041	0.002	0.002
		30%	0.056	0.001	0.001
		40%	0.069	6.282	6.282
	8	5%	0.014	0.000	0.000
		10%	0.021	0.001	0.000
		15%	0.027	0.001	0.001
		20%	0.038	6.282	6.282
		30%	0.055	6.283	6.283
		40%	0.070	0.001	0.001

Table 6.1 (b): Simulation results of circular distance for mean direction for sample size, n = 30

Performance Indicator	Concentration Parameter	Percentage of missing values	Method 1	Method 2	Method 3
Circular Distance	2	5%	0.016	0.002	0.003
		10%	0.027	0.001	0.001
		15%	0.037	0.004	0.001
		20%	0.051	0.004	0.004
		30%	0.071	0.006	0.007
		40%	0.083	0.000	0.001
	4	5%	0.014	0.001	0.000
		10%	0.023	0.002	0.002
		15%	0.028	0.001	0.002
		20%	0.041	0.001	0.000
		30%	0.058	0.000	0.000
		40%	0.074	0.002	0.003
	6	5%	0.043	0.001	0.002
		10%	0.022	0.001	0.001
		15%	0.029	0.002	0.002
		20%	0.041	0.002	0.002
		30%	0.056	0.001	0.001
		40%	0.069	0.001	0.001
	8	5%	0.014	0.000	0.000
		10%	0.021	0.001	0.000
		15%	0.027	0.001	0.001
		20%	0.038	0.001	0.001
		30%	0.055	0.000	0.000
		40%	0.070	0.001	0.001

From Table 6.1, using the measure of circular distance for mean direction, it is found that as percentage of missing observation increase, the circular mean gets bigger using Method 1 and deviates from the true value of 0. For Methods 2 and 3, the circular mean remain about the true value as percentage of missing values increase. Using the measure of circular distance, we note that for both Methods 2 and 3 the circular distance remains relatively small as percentage of missing values increase. The results are

consistent for all different values of concentration parameter for each percentage of missing values. Thus, based on the measure of circular distance, it can be said that Method 2 and Method 3 are both superior for sample size of 30.

Table 6.2 (a): Simulation results of circular mean for mean direction for sample size, n = 50

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Circular Mean	2	5%	0.012	0.001	0.001
		10%	0.025	6.281	6.281
		15%	0.040	0.001	6.283
		20%	0.047	6.282	6.282
		30%	0.068	0.001	6.283
		40%	0.083	0.001	6.283
	4	5%	0.009	6.283	6.283
		10%	0.020	6.281	6.282
		15%	0.034	0.001	0.001
		20%	0.042	0.001	0.002
		30%	0.058	0.000	0.000
		40%	0.073	0.001	6.283
	6	5%	0.008	6.283	6.283
		10%	0.020	6.282	6.282
		15%	0.032	0.000	0.000
		20%	0.039	6.283	6.283
		30%	0.055	6.283	0.000
		40%	0.071	0.000	6.283
	8	5%	0.008	0.000	0.000
		10%	0.019	6.282	6.282
		15%	0.031	6.283	6.283
		20%	0.038	0.000	0.001
		30%	0.055	0.000	6.283
		40%	0.068	6.282	6.281

Table 6.2 and 6.3 show the simulation results for sample size of 50 and 100 respectively. The results exhibit the same pattern as in $n = 30$, where in general, the

mean values are closer to the true parameter. New means obtained by imputation using Method 2 and Method 3 are closer to the true values in comparison to current method namely Method 1. Similarly, as seen earlier from Table 6.1, the results for circular distance are consistently small for both proposed methods. These results give the same pattern for all values of concentration parameter for each percentage of missing values.

Table 6.2 (b): Simulation results of circular distance for mean direction for sample size, n = 50

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Circular Distance	2	5%	0.012	0.001	0.001
		10%	0.025	0.002	0.002
		15%	0.040	0.001	0.000
		20%	0.047	0.001	0.001
		30%	0.068	0.001	0.000
		40%	0.083	0.001	0.000
	4	5%	0.009	0.000	0.000
		10%	0.020	0.002	0.001
		15%	0.034	0.001	0.001
		20%	0.042	0.001	0.002
		30%	0.058	0.000	0.000
		40%	0.073	0.001	0.000
	6	5%	0.008	0.001	0.001
		10%	0.020	0.001	0.001
		15%	0.032	0.000	0.000
		20%	0.039	0.000	0.000
		30%	0.055	0.001	0.000
		40%	0.071	0.000	0.000
	8	5%	0.008	0.000	0.000
		10%	0.019	0.001	0.001
		15%	0.031	0.000	0.000
		20%	0.038	0.000	0.001
		30%	0.055	0.000	0.000
		40%	0.068	0.001	0.002

Table 6.3 (a): Simulation results of circular mean for mean direction for sample size, n = 100

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Circular Mean	2	5%	0.013	6.283	6.283
		10%	0.027	0.001	0.001
		15%	0.038	0.000	0.001
		20%	0.048	6.283	6.283
		30%	0.064	6.280	6.280
		40%	0.082	6.283	6.283
	4	5%	0.012	0.001	0.001
		10%	0.022	6.283	6.283
		15%	0.031	6.282	6.283
		20%	0.041	0.000	0.001
		30%	0.057	6.283	0.000
		40%	0.072	6.282	6.282
	6	5%	0.011	0.001	0.001
		10%	0.020	6.282	6.282
		15%	0.030	6.283	6.283
		20%	0.041	0.002	0.002
		30%	0.056	0.001	0.001
		40%	0.070	0.000	0.001
	8	5%	0.010	6.283	0.000
		10%	0.021	0.001	0.001
		15%	0.030	6.283	6.283
		20%	0.038	0.000	0.000
		30%	0.055	0.000	0.000
		40%	0.069	0.001	0.001

Table 6.3 (b): Simulation results of circular distance for mean direction for sample size, n = 100

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Circular Distance	2	5%	0.013	0.001	0.000
		10%	0.027	0.001	0.001
		15%	0.038	0.000	0.001
		20%	0.048	0.000	0.000
		30%	0.064	0.004	0.003
		40%	0.082	0.000	0.000
	4	5%	0.012	0.001	0.001
		10%	0.022	0.000	0.000
		15%	0.031	0.001	0.000
		20%	0.041	0.000	0.001
		30%	0.057	0.000	0.000
		40%	0.072	0.001	0.002
	6	5%	0.011	0.001	0.001
		10%	0.020	0.001	0.001
		15%	0.030	0.000	0.000
		20%	0.041	0.002	0.002
		30%	0.056	0.001	0.001
		40%	0.070	0.000	0.001
	8	5%	0.010	0.000	0.000
		10%	0.021	0.001	0.001
		15%	0.030	0.000	0.000
		20%	0.038	0.000	0.000
		30%	0.055	0.000	0.000
		40%	0.069	0.001	0.001

Tables 6.4- 6.6 show the simulation results for estimations of the concentration parameter after imputation have been done. Table 6.4 shows the results for sample size of 30. New means, estimate bias (EB) and estimate root mean square error (ERMSE) were calculated in order to evaluate the performance of each proposed method. In general, as the percentage of missing values increases, the EB and ERMSE values also

increase. Also, when the value of the concentration parameter become larger, it results in the increment of the EB and ERMSE values.

Table 6.4 (a): Simulation results of mean for concentration parameter, κ for sample size, n = 30

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Mean	2	5%	2.278	2.294	2.165
		10%	2.363	2.388	2.196
		15%	2.429	2.461	2.194
		20%	2.558	2.607	2.196
		30%	2.815	2.894	2.246
		40%	3.108	3.231	2.296
	4	5%	4.651	4.712	4.450
		10%	4.794	4.889	4.486
		15%	4.893	5.020	4.468
		20%	5.200	5.399	4.549
		30%	5.717	6.044	4.631
		40%	6.311	6.768	4.762
	6	5%	6.686	6.985	5.784
		10%	7.154	7.359	6.733
		15%	7.397	7.676	6.815
		20%	7.742	8.174	6.849
		30%	8.517	9.207	7.020
		40%	9.291	10.312	7.182
	8	5%	9.219	9.458	8.909
		10%	9.472	9.835	8.996
		15%	9.766	10.258	9.096
		20%	10.249	11.014	9.181
		30%	11.182	12.410	9.429
		40%	12.102	13.757	9.555

As we compare the three methods, using the measure of EB and ERMSE, Method 3 that is the imputation using DA algorithm, is the most superior method. It is consistently give the smallest ERMSE and EB as the percentage of missing values increase. Both methods 1 and 2 do not perform really well. Thus, from Tables 6.4, for

the concentration parameter we can conclude that Method 3 is the best method out of all three methods considered.

Table 6.4 (b): Simulation results of EB for concentration parameter, κ for sample size, n = 30

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Estimate Bias (EB)	2	5%	0.278	0.294	0.165
		10%	0.363	0.388	0.196
		15%	0.429	0.461	0.194
		20%	0.558	0.607	0.196
		30%	0.815	0.894	0.246
		40%	1.108	1.231	0.296
	4	5%	0.651	0.712	0.450
		10%	0.794	0.889	0.486
		15%	0.893	1.020	0.468
		20%	1.200	1.399	0.549
		30%	1.717	2.044	0.631
		40%	2.311	2.768	0.762
	6	5%	0.686	0.985	0.216
		10%	1.154	1.359	0.733
		15%	1.397	1.676	0.815
		20%	1.742	2.174	0.849
		30%	2.517	3.207	1.020
		40%	3.291	4.312	1.182
	8	5%	1.219	1.458	0.909
		10%	1.472	1.835	0.996
		15%	1.766	2.258	1.096
		20%	2.249	3.014	1.181
		30%	3.182	4.410	1.429
		40%	4.102	5.757	1.555

Table 6.4 (c): Simulation results of ERMSE for concentration parameter, κ for sample size, n = 30

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Estimate Root Mean Square Error (ERMSE)	2	5%	0.629	0.645	0.584
		10%	0.698	0.721	0.618
		15%	0.754	0.788	0.636
		20%	0.871	0.923	0.662
		30%	1.139	1.233	0.764
		40%	1.462	1.608	0.857
	4	5%	1.415	1.475	1.315
		10%	1.579	1.682	1.417
		15%	1.646	1.790	1.419
		20%	1.922	2.145	1.519
		30%	2.454	2.839	1.678
		40%	3.150	3.702	1.953
	6	5%	2.245	2.517	2.171
		10%	2.230	2.435	2.038
		15%	2.545	2.839	2.254
		20%	2.884	3.374	2.406
		30%	3.651	4.432	2.642
		40%	4.487	5.735	3.017
	8	5%	2.824	3.065	2.739
		10%	3.006	3.380	2.835
		15%	3.282	3.801	2.983
		20%	3.726	4.562	3.196
		30%	4.648	6.075	3.606
		40%	5.740	7.648	3.960

Table 6.5 (a): Simulation results of mean for concentration parameter, κ for sample size, $n = 50$

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Mean	2	5%	2.163	2.170	2.096
		10%	2.259	2.278	2.084
		15%	2.394	2.423	2.108
		20%	2.492	2.531	2.130
		30%	2.699	2.757	2.124
		40%	2.970	3.056	2.166
	4	5%	4.358	4.388	4.238
		10%	4.569	4.647	4.268
		15%	4.805	4.928	4.282
		20%	4.976	5.131	4.305
		30%	5.445	5.700	4.354
		40%	5.964	6.319	4.414
	6	5%	6.569	6.638	6.403
		10%	6.857	7.032	6.419
		15%	7.184	7.467	6.450
		20%	7.474	7.839	6.530
		30%	8.088	8.660	6.568
		40%	8.876	9.681	6.693
	8	5%	8.703	8.824	8.501
		10%	9.102	9.413	8.567
		15%	9.490	9.990	8.603
		20%	9.807	10.443	8.664
		30%	10.653	11.650	8.813
		40%	11.555	12.962	8.922

Table 6.5 (b): Simulation results for concentration parameter, κ for sample size, $n = 50$

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Estimate Bias (EB)	2	5%	0.163	0.170	0.096
		10%	0.259	0.278	0.084
		15%	0.394	0.423	0.108
		20%	0.492	0.531	0.130
		30%	0.699	0.757	0.124
		40%	0.970	1.056	0.166
	4	5%	0.358	0.388	0.238
		10%	0.569	0.647	0.268
		15%	0.805	0.928	0.282
		20%	0.976	1.131	0.305
		30%	1.445	1.700	0.354
		40%	1.964	2.319	0.414
	6	5%	0.569	0.638	0.403
		10%	0.857	1.032	0.419
		15%	1.184	1.467	0.450
		20%	1.474	1.839	0.530
		30%	2.088	2.660	0.568
		40%	2.876	3.681	0.693
	8	5%	0.703	0.824	0.501
		10%	1.102	1.413	0.567
		15%	1.490	1.990	0.603
		20%	1.807	2.443	0.664
		30%	2.653	3.650	0.813
		40%	3.555	4.962	0.922

Table 6.5 (c): Simulation results of ERMSE for concentration parameter, κ for sample size, n = 50

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Estimate Root Mean Square Error (ERMSE)	2	5%	0.441	0.446	0.420
		10%	0.493	0.509	0.424
		15%	0.605	0.632	0.457
		20%	0.699	0.737	0.493
		30%	0.889	0.949	0.512
		40%	1.170	1.263	0.572
	4	5%	0.931	0.955	0.885
		10%	1.112	1.187	0.971
		15%	1.284	1.409	0.989
		20%	1.448	1.602	1.019
		30%	1.897	2.176	1.142
		40%	2.454	2.839	1.261
	6	5%	1.464	1.519	1.402
		10%	1.705	1.868	1.515
		15%	1.943	2.221	1.578
		20%	2.215	2.582	1.675
		30%	2.807	3.415	1.798
		40%	3.617	4.485	1.991
	8	5%	1.902	1.998	1.843
		10%	2.207	2.488	1.991
		15%	2.526	2.999	2.090
		20%	2.770	3.408	2.172
		30%	3.628	4.674	2.468
		40%	4.522	6.030	2.664

Table 6.6 (a): Simulation results of mean for concentration parameter, κ for sample size, n = 100

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Mean	2	5%	2.130	2.137	2.043
		10%	2.217	2.232	2.039
		15%	2.313	2.335	2.045
		20%	2.419	2.449	2.052
		30%	2.628	2.675	2.058
		40%	2.862	2.924	2.065
	4	5%	4.242	4.273	4.088
		10%	4.437	4.501	4.117
		15%	4.587	4.683	4.094
		20%	4.800	4.931	4.119
		30%	5.237	5.447	4.145
		40%	5.718	6.013	4.176
	6	5%	6.421	6.496	6.203
		10%	6.640	6.791	6.184
		15%	6.917	7.148	6.211
		20%	7.214	7.522	6.237
		30%	7.849	8.339	6.287
		40%	8.509	9.185	6.299
	8	5%	8.523	8.658	8.257
		10%	8.833	9.105	8.289
		15%	9.164	9.579	8.299
		20%	9.504	10.062	8.320
		30%	10.306	11.174	8.400
		40%	11.179	12.387	8.475

Table 6.6 (b): Simulation results of estimate bias for concentration parameter, κ for sample size, n = 100

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Estimate Bias (EB)	2	5%	0.130	0.137	0.043
		10%	0.217	0.232	0.039
		15%	0.313	0.335	0.045
		20%	0.419	0.449	0.052
		30%	0.628	0.675	0.058
		40%	0.862	0.924	0.065
	4	5%	0.242	0.273	0.088
		10%	0.437	0.501	0.117
		15%	0.587	0.683	0.094
		20%	0.800	0.931	0.119
		30%	1.237	1.447	0.145
		40%	1.718	2.013	0.176
	6	5%	0.421	0.496	0.203
		10%	0.640	0.791	0.184
		15%	0.917	1.148	0.211
		20%	1.214	1.522	0.237
		30%	1.849	2.339	0.287
		40%	2.509	3.185	0.299
	8	5%	0.523	0.658	0.257
		10%	0.833	1.105	0.289
		15%	1.164	1.579	0.299
		20%	1.504	2.062	0.320
		30%	2.306	3.174	0.400
		40%	3.179	4.387	0.475

Table 6.6 (c): Simulation results of ERMSE for concentration parameter, κ for sample size, n = 100

Performance Indicator	Concentration Parameter	Percentage	Method 1	Method 2	Method 3
Estimate Root Mean Square Error (ERMSE)	2	5%	0.297	0.302	0.271
		10%	0.357	0.369	0.286
		15%	0.436	0.456	0.303
		20%	0.530	0.559	0.316
		30%	0.726	0.773	0.342
		40%	0.957	1.020	0.360
	4	5%	0.630	0.651	0.575
		10%	0.765	0.818	0.616
		15%	0.890	0.976	0.629
		20%	1.072	1.198	0.672
		30%	1.497	1.709	0.737
		40%	1.965	2.268	0.786
	6	5%	0.988	1.041	0.914
		10%	1.142	1.269	0.951
		15%	1.375	1.586	1.009
		20%	1.624	1.920	1.054
		30%	2.217	2.712	1.136
		40%	2.880	3.575	1.232
	8	5%	1.330	1.424	1.255
		10%	1.538	1.767	1.326
		15%	1.779	2.156	1.361
		20%	2.074	2.602	1.417
		30%	2.822	3.689	1.554
		40%	3.664	4.912	1.685

Tables 6.5 and 6.6 show the simulation results for sample size of 50 and 100 respectively. Similar to the previous results of n = 30, the simulation results also exhibit the same pattern. In general, an increment in the percentage of missing values being imputed using the proposed method has led to a divergence of new mean as well as having larger value of EB and ERMSE for all three methods. It also noted that, the larger the concentration parameter, the higher are the EB and ERMSE values. This is

true for all the concentration parameter values. Small values of EB and ERMSE are observed for smaller percentage of missing values such as 5%, 10%, 15% and 20%. However, it is worthwhile to observe that the new means are comparatively far from the initial values if the percentage of missing values are too high especially when it reached 40% level. At this percentage level, it tends to produce quite high value of EB and ERMSE. Thus, it can be inferred that when the percentage of missing values reach more than 40%, the proposed method are no longer suitable to be implemented in the analysis.

In conclusion, from Table 6.1-6.3, it can be said, both proposed methods which are Method 2 and Method 3 perform better than Method 1. It can be seen that imputation methods by Method 2 and Method 3 give an estimated parameter which is close to its true value and has shorter circular distance. In contrast, imputation by circular mean gives poor performance as the difference of the new means with the true values gets larger with an increase in the percentage of missing values.

Thus, based on all simulation results displayed in Tables 6.1 to 6.6, a few conclusions can be drawn. As mentioned earlier, for parameter mean direction, both proposed methods which are Method 2 and Method 3 give the best performance based on the calculated values of circular distance. Both proposed methods seem to give very consistent values for all different values of the concentration parameter and sample sizes at a different level of percentage of missing data. Unlike Method 2 and Method 3, Method 1 gives a poor estimate by exhibiting comparatively higher value of circular distance for each sample size and concentration parameter. Hence, in perspective of the mean direction, it can be said that Method 2 and Method 3 can be used to impute the data with missing values if our objective is to estimate the parameter mean direction only.

However, if our objective is to estimate both parameters in von Mises distribution, we have to consider the results obtained in Tables 6.4 to 6.6. Considering all three methods of estimating the parameter after doing an imputation, it can be seen that Method 3 gives the best estimate. Method 3 gives consistently small values of EB and ERMSE while estimating the concentration parameter and also giving short circular distance in estimating the mean direction. Thus, from this simulation studies, it can be concluded that Method 3 that is DA algorithm is the best method to impute the missing values distributed with von Mises distribution.

6.5 Illustrative Example

As an illustration of the proposed method, a bivariate data set was considered. A sample size of 85 observations is considered. The data was fitted using the simple linear regression model proposed by Downs and Mardia (2002), and the model is given as below:

$$\hat{y}_i = 1.253 + 2\arctan\left\{0.906\tan\frac{1}{2}\left(x_i - 1.141\right)\right\}$$

The circular residuals for the fitted model are calculated by:

$$\theta_i = \hat{y}_i - y_i$$

In this section, our particular interest is in testing the superiority of the imputation methods in the circular residuals data. The new parameter estimation after imputation method is calculated using three methods considered.

Table 6.7: Parameter estimation based on imputation method

Percentage	Mean Direction			Concentration Parameter		
	Method 1	Method 2	Method 3	Method 1	Method 2	Method 3
5%	0.146	0.154	0.159	7.786	7.846	7.483
10%	0.139	0.155	0.152	7.882	8.001	6.877
15%	0.128	0.153	0.161	8.172	8.362	6.726
20%	0.121	0.154	0.187	8.366	8.617	6.731
30%	0.097	0.133	0.106	9.485	9.777	6.322
40%	0.091	0.146	0.168	17.775	19.395	14.050

Table 6.8: Circular distance and estimate bias calculated using imputation method

Percentage	Circular Distance (CD)			Estimated Bias (EB)		
	Method 1	Method 2	Method 3	Method 1	Method 2	Method 3
5%	0.007	0.000	0.005	0.344	0.404	0.041
10%	0.014	0.001	0.001	0.44	0.559	0.565
15%	0.025	0.001	0.007	0.73	0.92	0.716
20%	0.032	0.001	0.034	0.924	1.175	0.711
30%	0.056	0.020	0.048	2.043	2.335	1.12
40%	0.062	0.008	0.015	10.333	11.953	6.608

The initial mean direction for the residuals data is 0.153 while the concentration parameter is 7.442. Table 6.7 shows the new mean direction and concentration parameter estimated using the considered methods. The missingness was tested at six different percentages as what we have done in simulation studies. As for performance indicator, CD was calculated to measure the biasness for the mean direction while absolute EB is calculated for the concentration parameter. From Table 6.8, it can be seen that the value of the CD calculated for Method 2 and Method 3 are small up to 20% of missing values in comparison to Method 1. These results are similar to the ones

obtained in the simulation studies. The value of EB also seems to exhibit the same pattern as simulation results in Section 6.4. Method 1 and Method 3 give small EB as compared to Method 2. Thus, in conclusion, by considering both parameters, Method 3 is the best imputation method in handling missing values for circular data with von Mises distribution.

6.6 Discussion

Three different methods have been considered in handling the missing values for circular data. In this chapter, we focus on circular variables distributed with von Mises distribution. Method 1, which is imputation by circular mean, is the common method and has been widely used for handling missing data in linear data, as well as few studies in circular. Method 2 and Method 3 namely EM algorithm and DA algorithm respectively are the method that have been used in linear study and here investigate the applicability of both algorithms in circular data for the purpose of improving the method of handling missing values.

From simulation studies shown in Section 6.4, a few conclusions can be drawn. Based on circular distance, both Method 2 and Method 3 are superior in which both methods give very small value of circular distance which imply that the new estimates are close to the initial parameter mean direction. However, if we consider the estimate bias which related to concentration parameter, Method 3 gives the smallest bias. Thus, considering both parameters, it can be said that Method 3 is the best method as it gives small values of circular distance as well as estimate bias.

All the methods considered have been illustrated using real data set found in the literature. The illustration results also supported the results obtained from simulation studies where the superior method is Method 3 if we consider both the mean direction and concentration parameter.

CHAPTER 7

CONCLUSIONS

7.1 Summary

In this chapter, we will summarise all findings that we obtained from this study. Four topics related to von Mises distribution are discussed. In the first part of the study, it focuses on the efficient approximation for the concentration parameter in von Mises distribution. Our purpose is to propose an improved estimate of the concentration parameter, κ for the von Mises distribution which is applicable for both small and large values of κ In this study, our proposed method will be compared with three different methods namely, the Dobson's method, Best & Fisher's method and Amos's method. From the simulation studies, it can be observed that, for both small and large values of κ, the proposed method shows a better performance than the Amos's, Dobson's and Best & Fisher's methods except for when $\kappa \leq 1$. This can be seen from the least absolute relative bias for most of the κ values as well as smaller values of estimated SE and RMSE in comparison to the other methods considered. Unlike the Amos's method which is restrictive to small values of κ ($\kappa \leq 1$ for $n \leq 50$), the proposed method seems to be applicable to both small and large values of κ ($\kappa \geq 1$ for large sample size n = 100).

In the second part of this study, our focus is on constructing the confidence intervals (CI) for the concentration parameter, κ in von Mises distribution. Four improved methods in obtaining the CI of the concentration parameter for data with

moderately large κ values were proposed. The following are the methods that we considered

i. CI based on circular variance population which will be referred to as Method 1
ii. CI based on the asymptotic distribution of $\hat{\kappa}$ which will be referred to as Method 2
iii. CI based on the distribution of mean direction and mean resultant length which will be referred to as Method 3
iv. CI based on bootstrap-*t* method which will be referred to as Method 4

In addition to the four methods, a current method based on percentile bootstrap by Fisher is also considered. From the simulation study, it is noted that all of the four proposed methods seems to perform relatively better than the existing method by Fisher. Method 2 is superior in terms of simplicity in obtaining the CI, Method 4 is superior in terms of coverage probability and Method 3 is superior in terms of expected length. All proposed methods

In the third part of the study, the objective is to propose a new statistic based on circular distance in von Mises distribution. The proposed statistics that we obtained can be used in approximating a sample from von Mises distribution to Chi-squared distribution. Apart from that, the statistics based on circular distance is used in constructing new CI for the concentration parameter. In this study, three different methods are considered namely mean, median and percentile. From the simulation study, we noted that the range of percentile that gives values that are close to 0.95 is from 30th to 50th percentile. Another simulation study is performed to assess the performance of all proposed methods. From the simulation, it can be concluded that the

CI based on percentile consistently gives good coverage probability as well as the smallest expected length.

In the final part of this study, we consider several imputation methods when there are with missing values problem in the circular data set. In this study, the data are in circular form and distributed with von Mises distribution. Three different methods have been considered namely Method 1, which is imputation by circular mean, Method 2 and Method 3, which is EM algorithm and DA algorithm respectively. From the simulation studies, by assessing the performance indicator, a few conclusions can be drawn. Based on circular distance, both Method 2 and Method 3 are superior in which both methods give very small value which imply that the new estimates are close to the initial parameter mean direction. However, if we consider the estimate bias which related to concentration parameter, Method 3 gives the smallest bias. Thus, considering both parameters, it can be said that Method 3 is the best method as it gives small values of circular distance as well as estimate bias.

7.2 Contributions

This particular study has contributed and benefited to circular data analysis in the following ways:

i. We have proposed a new approach to approximate the concentration parameter in von Mises which applicable for both small and large values.
ii. We have developed four different methods of constructing the CI for large concentration parameter in von Mises distribution.

iii. We have proposed a new statistic based on circular distance in von Mises distribution

iv. We have shown that, the new statistic based on circular distance can be used to approximate a sample from von Mises to Chi-squared distribution

v. We have developed the CI for the concentration parameter using the statistic based on circular distance that we obtained

vi. We have identified two feasible methods in handling the missing problem in circular data distributed with von Mises distribution.

7.3 Further Research

Apart from the contributions that have been obtained from this study, there are various possibilities for further research in this related area. Some suggestions that might be considered for future studies are as follows:

i. consider other circular distribution in approximating the confidence intervals for the parameter

ii. consider outliers while approximating the confidence intervals

iii. develop the method of identifying outlier using a new statistic based on circular distance

iv. extend the circular distribution or circular model that can be considered in handling the missing data problems

v. extend to another imputation method in handling the missing values

vi. consider the robustness for each method proposed

REFERENCES

Abramowitz, M. and Stegun, I. G. (1965). *Handbook of Mathematical Functions.* New York: Dover Publications Inc.

Abuzaid, A. H., Mohamed, I. and Hussin, A. G. (2012). Boxplot for circular Variables. *Computational Statistics, 27*(3): 381-392.

Acock. (2005) Working with missing values. *Journal of Marriage and Family, 67*: 1012-1028.

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage Publications.

Amos, D. E. (1974). Computation of modified Bessel functions and their ratios. *Mathematics of Computation, 28*(125): 239-251.

Asgharzadeh, A. and Abdi, M. (2011). Exact confidence intervals and joint confidence regions for the parameters of the Gompertz. *Pakistan Journal of Statistics, 27*(1): 55-64.

Baraldi, A. N. and Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology, 48*: 5-37.

Barzi, F. and Woodward, M. (2004). Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology, 160*(1): 34-45.

Batschelet, E. (1981). *Circular Statistics in Biology.* London: Academic Press Inc.

Beckmann, P. (1959). The probability distribution of the vector sum of *n* unit vectors with arbitrary phase distributions. *Acta Technica, 4*: 323-335.

Berens, P. (2009). CircStat: A MATLAB toolbox for circular statistics. *Journal of Statistical Software, 31*(10): 1-21.

Best, D. J. and Fisher, N. I. (1981).The bias of the maximum likelihood estimators of the von Mises-Fisher concentration parameters, *Communications In Statistics—Simulation And Computation, B10*(5): 493-502.

Bowers, J. A., Morton, I. D. and Mould, G. I. (2000). Directional statistics of the wind and waves. *Applied Ocean Research, 22*: 13–30.

Brown, I. L. and Mewaldt, L. R. (1968). Behavior of sparrows of the genus *Zonotrichia*,in orientation cages during the lunar cycle. *Z. Tierpsychol*. *25*(6): 668-700.

Brunsdon, C. and Corcoran, J. (2005). "Using circular statistics to analyse time patterns in crime incidence." *Computers, Environment and Urban Systems, 30*: 300-319.

Caires, S. and Wyatt, L. R. (2003). A linear functional relationship model for circular data with an application to the assessment of ocean wave measurements. *Journal of Agricultural, Biological, and Environmental Statistics, 8*(2): 153-169.

Chernick, M. R. (1999). *Bootstrap Method: A practitioner's Guide*. Canada: John Wiley & Sons.

Cox, N. J. (2001). *Analysing Circular Data in Stata*. Nasug: Boston.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological), 39*(1): 1 - 38.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*(3): 189-212.

Dobson, A. J. (1978). Simple approximations for the von Mises concentration statistics. *Journal of the Royal Statistical Society (Applied Statistics), 27*(3)*:* 345-347.

Downs, T. D. and Mardia, K. V. (2002). Circular regression. *Biometrika, 89*(2): 683-697.

Durcharme, G. R., Jhun, M., Romano, J. and Truong, K. N. (1985). Bootstrap Confidence Cones for Directional Data. *Biometrika, 72*(3): 637-645.

Enders, C. K. (2006). A primer on the use of modern missing-data methods in psychosomatic medicine research. *Psychosomatic Medicine, 68*: 427-436.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1): 1–26.

Efron, B. (1987). Better bootstrap confidence intervals (with comments). *Journal of the American Statistical Association*, *82*(397): 171–200.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

Efron, B. and Tibshirani, R. (1986). Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science, 1*(1): 54-75.

Fisher, N. and Hall, P. (1989). Confidence regions for directional data. *Journal of the American Statistical Association*, *84*(408): 996 - 1002.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. Cambridge: Cambridge University Press.

Fisher, N. I. and Hall, P. (1992). *Bootstrap methods for directional data. The Art of statistical science*, edited by K.V. Mardia. Wiley: New York.

Gatto, R. (2008). Some computational aspects of the generalized von Mises distribution, *Statistics and Computing, 18*: 321–331.

Gatto, R. and Jammalamadaka, S. R. (2007). The generalized von Mises distribution. *Statistical Methodology, 4*: 341-353.

Graham, J. W., Olchowski, A. E. and Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*: 206−213.

Hall, P. (1986). On the bootstrap and confidence intervals. *The Annals of Statistics, 14*(4): 1431-1452.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *The Annals of Statistics, 16*(3): 927–953.

Hall, P. and Martin, M. A. (1988). On bootstrap resampling and iteration. *Biometrika, 75*(4): 661-671.

Handerson, P. A. and Seaby, R. M. H. (2002). *Axis Software, version 1.1,* Pisces conservation Ltd.

Hassan, S. F., Hussin, A. G. and Zubairi, Y. Z. (2009). Analysis of Malaysian wind direction data using ORIANA. *Modern Applied Science, 3*(3): 115-119.

Hassan, S. F., Hussin, A. G. and Zubairi, Y. Z. (2010a). Estimation of functional relationship model for circular variables and its application in measurement problems. *Chiang Mai Journal of Science, 37*(2): 195 - 205.

Hassan, S. F., Hussin, A. G. and Zubairi, Y. Z. (2012). Improved efficient approximation of concentration parameter and confidence interval for circular distribution. *Science Asia, 38*: 118-124.

Hassan, S. F., Zubairi, Y. Z., and Hussin, A. G. (2010b). Analysis of missing values in simultaneous linear functional model for circular variables. *Scientific Research and Essays, 5*(12): 1483-1491.

Hendriks, H., Landsman, Z. and Ruymgaart, F. (1996). Asymptotic behavior of sample mean direction for spheres. *Journal of Multivariate Analysis, 59*: 141-152.

Hippel, P. T. V. (2004). Biases in SPSS 12.0 misssing value analysis. *The American Statistician, 58*(2): 160-164.

Hussin, A. G., Fieller, N. R. and Stillman, E. C. (2004). Linear regression for circular variables with application to directional data. *Journal of Applied Science & Technology, 8*, (1 & 2): 1-6.

Hussin, A. G., Jalaluddin, J. F. and Mohamed, I. (2006). Analysis of Malaysian wind direction data using AXIS. *Journal of Applied Sciences Research, 2*(11): 1019-1021.

Jammalamadaka, S. R. and Lund, U. J. (2006). The effect of wind direction on ozone levels: a case study. *Environmental and Ecological Statistics, 13*: 287 - 298.

Jammalamadaka, S. R. and SenGupta, A. (2001). *Topics in Circular Statistics.* Singapore: World Scientific Publishing Co. Pte.Ltd.

Johnson, D. R. and Young, R. (2011). Toward best practices in analyzing datasets with missing data: Comparisons and recommendations. *Journal of Marriage and Family, 73*(5): 926-945.

Jones, T. A. (2006). MATLAB functions to analyze directional (azimuthal) data-I: Single sample inference. *Computer and Geoscience, 32*: 166-175.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment, 38*: 2895-2907.

Kamisan, N. A. B., Hussin, A. G. and Zubairi, Y. Z. (2010). Finding the best circular distribution for southwesterly monsoon wind direction in Malaysia. *Sains Malaysiana, 39*(3): 387-393.

Khanabsakdi, S. (1995 – 1996). Inferential statistics for concentration of directional data using the chi-square distribution. *The Philippine Statistician, 44 - 45*(1-8): 61 - 67.

Kim, J. O. and Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research, 6*: 215-241.

Kofman, P. and Sharpe, I. (2000). "Imputation Methods for Incomplete Dependent Variables in Finance," *Econometric Society World Congress 2000 Contributed Papers 0409, Econometric Society.*

Kovach Computing Services (2009). *Oriana for Windows.* Kovach Computing Services, Wales. Version 3, URL http://www.kovcomp.co.uk/oriana/.

Letson, D. and McCullough, B. D. (1998). Better confidence intervals: the double bootstrap with no pivot. *Agricultural & Applied Economics Association, 80*(3): 552-559.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data* (2nd Ed.). New York: Wiley.

Lund, U. and, Agostinelli, C. (2012). CircStats: Circular Statistics. R package version 0.2-4, URL http://cran.r-project.org/web/packages/CircStats/index.html

Mardia, K. V. (1972). *Statistics of Directional Data*. London: Academic Press Inc.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. England: John Wiley & Sons Ltd.

Mardia, K.V. and Zemroch, P. J. (1975). Algorithm AS81. Circular Statistics. *Applied Statistics*, *24*: 147–148.

Norazian, M. N., Shukri, Y. A., Azam, R. N. and Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia, 34*: 341-345.

Otieno, B. S. and Anderson-Cook, C. M. (2006a). Hodges-Lehmann estimator of preferred direction for circular data. *Journal of Statistics and Applications, 1*(2-4): 155-169.

Otieno, B. S., Anderson-Cook, C. M. (2006b). On bootstrap confidence interval estimation of preferred direction for circular data. Technical paper No. 805, Grand Valley State Univ, USA.

Peugh, J. L. and Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*: 525-556.

Polansky, A. M. (2000). Bootstrap-*t* confidence intervals for small samples. *The Canadian Journal of Statistics*, *28*(3): 501-516.

Porter, P. S., Rao, S. T., Ku, J. Y., Poirot, R. L. and Dakins, M. (1997). Small sample properties of nonparametric bootstrap-*t* confidence intervals. *Technical Paper: Journal of the Air & Waste Management Association, 47:* 1197-1203.

Ragunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health, 25*: 88-117.

Rao, J. S. (1969). *Some Contributions to the Analysis of Circular Data*. Ph.D Thesis, Indian Statistical Institute, Calcutta, India.

Rivest, L. –P. (1997). A decentred predictor for circular-circular regression. *Biometrika, 84*(3): 717-726.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Hoboken, N J: Wiley.

Sartori, N., Salvan, A. and Thomaseth, K. (2005). Multiple imputation of missing values in a cancer mortality analysis with estimated exposure dose. *Computational Statistics & Data Analysis, 49*: 937-953.

Saunders, J. A., Howel, N. M., Spitznagel, E., Dore, P., Proctor, E. K. and Pescarino, R. (2006). Imputing missing data: A comparison of methods for social work researchers. *ProQuest Education Journals, 30*(1): 9-31.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schafer, J. L. and Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association, 95*: 144–154.

SenGupta, S. and Rao, J. S. (1966). Statistical analysis of crossbedding azimuths from the Kamthi formation around Bheemaram, Pranhita-Godavari valley. *Sankhya Series B, 28*: 165-174.

Sinharay, S., Stern, H. S., and Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6*: 317-329.

Stephen, M. A. (1962a). Exact and approximate tests for direction I. *Biometrika, 49*(3 and 4): 467-477.

Stephen, M. A. (1962b). Exact and approximate tests for direction II. *Biometrika, 49*(3 and 4): 547-552.

Stuart, A. and Ord, J. K. (1991). *Kendall's Advance Theory of Statistics, Vol. 2* (5th Ed.). Edward Arnold: London.

Stephens, M. A. (1969). Test for the von Mises distribution. *Biometrika, 56*(1): 149-160.

Sun, Y. and Wong, A. C. M. (2007). Interval estimationfor the normal correlation coefficient. *Stattistics & Probability Letters, 77*: 1652-1661.

Tanner, M. A and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association, 82*(398): 528 - 540.

Tibshirani, R. J. (1988). Variance stabilization and the bootstrap. *Biometrika, 75*(3): 433–444.

Tsechansky, M. S and Provost, F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research, 8*: 1625-1657.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management, 24*: 53-62.

Upton, G. J. G. and Fingleton, B. (1989). *Spatial Data Analysis by Example. Volume 2. Categorical and Directional Data.* New York: John Wiley.

Upton, G. J.G. (1986). Approximate confidence intervals for the mean direction of a von Mises ditribution. *Biometrika, 73*(2): 525-527.

Watson, G. S. (1970). Orientation statistics in the Earth sciences. *Bull. Geol. Inst. Univ. Uppsala, 2*(9): 73-89.

Watson, G. S. (1983). *Statistics on Spheres*. New York: Wiley.

Winkler, A. and McCarthy, P. (2005). Maximising the value of missing data. *Journal of Targetting, Measurement and Analysis for Marketing*, *13*(2): 168-178.

Zar J. H. (1999). *Biostatistical Analysis.* 4th edition. Prentice Hill.

Zar, J. H. (1984). *Biostatistical analysis*. (2nd Ed.). Englewood Cliffs, NJ: Prentice-Hall Inc. (7).

Zubairi, Y. Z., Hussain, F. and Hussin, A. G. (2008). An alternative analysis of two circular variables via graphical representation: An application to the Malaysian wind data. *Computer and Information Science, 1*(4): 3-8.

LIST OF PUBLICATIONS

Academic Journal

Hassan, S. F., Hussin, A. G. and Zubairi, Y. Z. (2012). Improved efficient approximation of concentration parameter and confidence interval for circular distribution. *ScienceAsia, 38*: 118–124. ***(ISI-Cited Publication)***

Hassan, S. F., Hussin, A. G., Zubairi, Y. Z. and Satari, S. Z. (2014). Some Confidence Intervals for Large Concentration Parameter in von Mises Distribution. *Pakistan Journal of Statistics, 30*(2): 273-284. ***(ISI-Cited Publication)***

Satari, S. Z., Hussin, A. G., Zubairi, Y. Z. and Hassan, S. F. (2014). A New Functional Relationship Model for Circular Variables. *Pakistan Journal of Statistics, 30*(3): 397-410. ***(ISI-Cited Publication)***

Conference Proceeding

Efficient Approximation for the von Mises Concentration Parameter. Proceeding in International Conference on Mathematics and Sciences (ICOMSc). October 12-13, 2011 at Majapahit Hotel, Jalan Tunjungan 65 Surabaya, Indonesia.

LIST OF ORAL PRESENTATIONS

Improved Eficient of the Concentration Parameter for Von Mises Distribution. Posgraduate Seminar on 12 May 2011 at Pusat Asasi Sains, UM.

Efficient Approximation for the von Mises Concentration Parameter. International Conference on Mathematics and Sciences (ICOMSc). October 12-13, 2011 at Majapahit Hotel, Jalan Tunjungan 65 Surabaya, Indonesia.

Approximation of Confidence Intervals for the Concentration Parameter, κ in Von Mises Distribution. 1st ISM International Statistical Conference 2012 on 4 - 6 September 2012 at Persada Johor, Malaysia.

Bootstrap-*t* confidence intervals for concentration parameter in von Mises distribution. 12th Islamic Countries Conference on Statistical Sciences (ICCS-12). December 19-22, 2012 at Doha, Qatar.

Appendix A. Wind direction data recorded at maximum wind speed at Kuala Terengganu

Obs. Number	Wind direction (radian)
1	1.571
2	1.571
3	0.698
4	0.873
5	1.222
6	5.411
7	1.047
8	0.698
9	1.222
10	6.283
11	1.047
12	1.047
13	0.698
14	0.349
15	1.047
16	1.047
17	1.047
18	1.047
19	1.047
20	0.524
21	1.047
22	0.175
23	6.109
24	5.934
25	0.524

Obs. Number	Wind direction (radian)
26	1.396
27	0.873
28	6.283
29	1.571
30	1.047
31	1.047
32	6.109
33	5.934
34	0.698
35	0.349
36	1.047
37	0.698
38	0.873
39	0.873
40	0.873
41	1.396
42	0.698
43	1.047
44	0.873
45	0.873
46	1.047
47	0.873
48	0.873
49	1.047
50	0.873

Appendix B. Wind direction data recorded using HF radar and anchored buoy.

Obs. Number	HF Radar (radian)	Anchored Buoy (radian)
1	0.790	1.154
2	0.715	1.154
3	0.975	1.007
4	0.970	1.178
5	0.993	0.859
6	0.902	1.007
7	0.943	1.056
8	1.728	1.400
9	1.445	1.497
10	1.679	1.693
11	1.703	2.012
12	1.862	1.792
13	1.726	1.766
14	1.790	1.669
15	1.831	1.400
16	1.719	1.400
17	1.646	1.375
18	1.622	1.056
19	1.342	1.178
20	1.176	1.276
21	1.325	1.693
22	1.103	1.325
23	6.131	6.062
24	5.719	5.988
25	5.713	5.988
26	5.487	5.498
27	5.742	5.276
28	5.728	5.302
29	5.610	5.620
30	5.463	5.744
31	5.427	5.644
32	5.418	5.669
33	5.406	5.744
34	5.472	5.547
35	5.401	5.498
36	5.420	5.400
37	5.276	5.449
38	1.728	4.786
39	5.512	5.449
40	5.486	5.178
41	5.444	5.620
42	5.518	5.130

43	5.505	4.541
44	5.558	5.571
45	5.420	5.620
46	5.398	5.473
47	5.334	5.327
48	5.418	4.835
49	5.418	5.032
50	5.338	5.842
51	5.470	5.571
52	5.455	5.522
53	5.555	5.473
54	5.462	5.522
55	5.401	5.522
56	5.316	5.376
57	5.439	5.081
58	5.408	5.473
59	5.431	5.449
60	5.473	5.915
61	5.460	5.351
62	5.364	5.571
63	5.444	5.376
64	5.350	5.327
65	5.202	4.983
66	5.161	4.786
67	5.062	4.908
68	5.145	4.517
69	5.212	4.835
70	5.238	4.417
71	5.238	4.417
72	4.970	5.007
73	4.947	5.473
74	4.887	5.400
75	4.872	4.859
76	4.589	4.859
77	4.510	4.761
78	4.319	4.639
79	4.427	4.664
80	4.436	4.664
81	4.451	4.074
82	3.840	4.295
83	3.819	4.098
84	4.159	4.173
85	3.987	4.122

Appendix C. Programming Script: Simulation study for estimation of concentration parameter using different methods.

```
#SKEn50k0.5s5000=simu.kap.est(50,0,0.5,5000)
#kapT(0,2,20)
#kap.EST(20,0,2)

   simu.kap.est=function(n,mu,kp,simu){

      kappaEst=matrix(0,nrow=simu,ncol=4)
      for(i in 1:simu){
      kappaEst[i,]=kap.EST(n,mu,kp)$kapp

   }
      dimnames(kappaEst)=list(NULL,c("new
   technique","amos","fisher","dobson"))

      meanKappa=colMeans(kappaEst)
      EstBias_meanKappa-kp
      AREB_((abs(EstBias))/kp)
      ESE_((1/(simu-1))*(colSums((kappaEst-meanKappa)^2)))^(1/2)
      ERMSE_((1/simu)*(colSums((kappaEst-kp)^2)))^(1/2)
      output_rbind(meanKappa,EstBias,AREB,ESE,ERMSE)

   list(result=output,kappaEst=kappaEst)
   }
```

Appendix D. Programming Script: Estimation of concentration parameter using different methods.

```
#kap.EST(20,0,2)

   kap.EST=function(n,mu,kp){

   theta = rvm(n,mu,kp)

   C=(1/n) * sum(cos(theta))
   S=(1/n) * sum(sin(theta))
   r=(C^2 + S^2)^(1/2)

   if(kp<2){
   kp.s=polyroot(c(-96*r,48,0,-6,0,1))
   kap.n=kapT(kp.s)
   }

   else if(kp>=2){
   kp.s=polyroot(c(1,1,4,8*r-8))
   kap.n=kapT(kp.s)
   }
   kap.a=(r/(1 - r^2)) * ((1/2) + ((1.46 * (1 - r^2)) + (1/4))^(1/2))

   m1=CirMe(theta)
   K_sum(cos(theta-m1))
   w_K/n

   if(w<0.53){
      kap.f=2*w+w^3+(5/6)*w^5
   }
   else if(w>=0.53&&w<0.85){
      kap.f=(-0.4)+1.39*w+0.43/(1-w)
   }
   else if(w>=0.85){
      kap.f=1/(w^3-4*w^2+3*w)
   }

   if(w<0.65){
      kap.d=2*w+w^3+(5/6)*w^5
   }
```

```
   else if(w>=0.65){
      kap.d=(9-8*w+3*(w^2))/(8*(1-w))
   }

   a0 <- 1
   a1 <- 1
   a2 <- 4
   a3 <- 8 * r - 8
   p <- (3 * a3 * a1 - (a2^2))/(3 * (a3^2))
   q <- (2 * (a2^3) - 9 * a1 * a2 * a3 + 27 * a0 * (a3^2))/(27 *
   (a3^3))
   p1 <- (3 * (r - 1) - 2)/(24 * ((r - 1)^2))
   q1 <- (4 - 9 * (r - 1) + 54 * ((r - 1)^2))/(432 * ((r - 1)^3))
   D <- ((p/3)^3) + ((q/2)^2)
   kap.n1=( - (q/2) + D^(1/2))^(1/3) + ( - (q/2) - D^(1/2))^(1/3) -
   1/(6 * (r - 1))


   output1=cbind(kap.n,kap.a,kap.c,kap.f,kap.d,kap.n1)
   output2=cbind(kap.n,kap.a,kap.c,kap.d)

   list(all.kapp=output1,kapp=output2,kp.s=kp.s)

   }


#kapT(0,2,20)

   kapT=function(kp.s){
   n=length(kp.s)
   ab=Mod(kp.s)
   kpp=Re(kp.s)
   ac=cbind(ab,kpp)
   ad=ab-kpp
   for(i in 1:n){
      if(ad[i]==0) {rt=kpp[i]}
}
rt
}
```

Appendix E. Programming Script: Confidence Interval for concentration parameter.

```
Real.CI=function(theta,B,ky.1,ky.2,alp){

   n=length(theta)

   #estimation of parameter
   mu.1=CirMe(theta)
   kp.1=est.kappa(theta)

   S1=matrix(0,nrow=B,ncol=n)
   S3=matrix(0,nrow=B)
   S5=matrix(0,nrow=B)
   Cs=matrix(0,nrow=B)
   Ss=matrix(0,nrow=B)
   Rbars=matrix(0,nrow=B)
   SEr=matrix(0,nrow=B)
   tt=matrix(0,nrow=B)

      for(i1 in 1:B){
      S1[i1,]=rvm(n,mu.1,kp.1)
      S3[i1]=est.kappa(S1[i1,])

      Cs[i1]=(1/n)*sum(cos(S1[i1,]))
      Ss[i1]=(1/n)*sum(sin(S1[i1,]))
      Rbars[i1]=((Cs[i1])^2+(Ss[i1])^2)^(1/2)
      SEr[i1]=(1/n)*((1-Rbars[i1]/S3[i1]-(Rbars[i1])^2)^(-1/2))

      tt[i1]=(S3[i1]-kp.1)/SEr[i1]
   }

   S4=sort(S3)
   ac=as.integer((1/2)*B*alp+(1/2))
   am=B-ac
   a2=ac
   a3=(1/2)*B*alp+(1/2)

   kLow=S4[ac+1]
   kUpp=S4[am]
   L.bFisher=kUpp-kLow
```

```
# upper n lower limit based on another method

   cc=B*(alp/2)
   kLo=S4[cc]
   kUp=S4[B+1-cc]
   L.Fam=kUp-kLo

   tt0=sort(tt)
   t01=tt0[B*(1-alp/2)]
   t02=tt0[B*(alp/2)]

   C=(1/n)*sum(cos(theta))
   S=(1/n)*sum(sin(theta))
   Rbar=(C^2 + S^2)^(1/2)

   sE=(1/n)*((1-Rbar/kp.1-(Rbar)^2)^(-1/2))

   tLow=kp.1-t01*sE
   tUpp=kp.1-t02*sE
   L.bootT=tUpp-tLow

   v=(-2*log(Rbar))^(1/2)
   k1=((n-1)*(v^2))/ky.1
   k2=((n -1)*(v^2))/ky.2
   R1=exp(-k1/2)
   R2=exp(-k2/2)

   if(Rbar<0.6137){
   kp.s=polyroot(c(-96*Rbar,48,0,-6,0,1))
   kapp=kapT(kp.s)

   nCI1=kapT(polyroot(c(-96*R1,48,0,-6,0,1)))
   nCI2=kapT(polyroot(c(-96*R2,48,0,-6,0,1)))

   sE=sqrt(n*(1-(Rbar/kapp)-Rbar^2))
   n2.CI1=(-1.96/sE)+kapp
   n2.CI2=(1.96/sE)+kapp

   if(n2.CI1<0){n3.CI1=0 }
      else{n3.CI1=n2.CI1}
```

```
    }

    else if(Rbar>=0.6137){
    kp.s=polyroot(c(1,1,4,8*Rbar-8))
    kapp=kapT(kp.s)

    nCI1=kapT(polyroot(c(1,1,4,8*R1-8)))
    nCI2=kapT(polyroot(c(1,1,4,8*R2-8)))

    sE=sqrt(n*(1-(Rbar/kapp)-Rbar^2))
    n2.CI1=(-1.96/sE)+kapp
    n2.CI2=(1.96/sE)+kapp

    if(n2.CI1<0){n3.CI1=0 }
       else{n3.CI1=n2.CI1}
    }

    A=(n*(1-Rbar))/(Rbar*ky.1)
    B=(n*(1-Rbar))/(Rbar*ky.2)

    N3.CI1=kapT(polyroot(c(-(n*A+2*n),(4*A-6*n*A-12*n),(4*n*A+8*A-
    16*n),32*n*A)))
    N3.CI2=kapT(polyroot(c(-(n*B+2*n),(4*B-6*n*B-12*n),(4*n*B+8*A-
    16*n),32*n*B)))

    L.pop=nCI2-nCI1
    L.norm=n2.CI2-n3.CI1
    L.tbar=N3.CI2-N3.CI1


    Res=cbind(mu.1,kp.1,kapp,L.bFisher,L.Fam,L.bootT,L.stephen,L.pop,L.
    norm,L.tbar,
    kLow,kUpp,kLo,kUp,tLow,tUpp,sCI1,sCI2,nCI1,nCI2,n3.CI1,n2.CI2,N3.CI
    1,N3.CI2)

list(Result=Res)

}
```

Appendix F. Programming Script: CI based on a new statistic

```
simu.CI.sort=function(n,mu,kp,ch1,ch2,simu)
{
   covP = matrix(0, nrow = simu, ncol = 10)
   expL = matrix(0, nrow = simu, ncol = 10)

   #simulation
   for(i in 1:simu) {
      covP[i,]=CI.to.sort(n,mu,kp,ch1,ch2)$cp
      expL[i,]=CI.to.sort(n,mu,kp,ch1,ch2)$EL
         }

   dimnames(covP)=list(NULL,c("p.1","p.2","p.3","p.4","p.5","p.6","p.7
   ","p.8","p.9","p.10"))
   dimnames(expL)=list(NULL,c("p.1","p.2","p.3","p.4","p.5","p.6","p.7
   ","p.8","p.9","p.10"))

   #calculation of perf indicator
   CovP=colMeans(covP)
   ExpL=colMeans(expL)

   list(CovP=CovP,ExpL=ExpL,covP=covP,expL=expL)

}


## ---------------------- CI sort based on percentile ---------------


CI.to.sort=function(n,mu,kp,ch1,ch2){
   count=0
   #set.seed(40)
   theta=rvm(n,mu,kp)
   C=sum(cos(theta))
   S=sum(sin(theta))
   A3=matrix(0,nrow=n)

   for(j in 1:n){
```

```
   A3[j]=n-C*cos(theta[j])-S*sin(theta[j])
   kLow=ch1/A3
   kUpp=ch2/A3

}

data=cbind(kLow,kUpp)
ExpL=data[,2]-data[,1]
nData=cbind(data,ExpL)
dimnames(nData)=list(NULL,c("Low","Upp","ExpL"))

#sort each column
a1=sort(data[,1])
a2=sort(data[,2])
a3=cbind(a1,a2)

b1=sort.col(nData,"@ALL","Low",T)
b2=sort.col(nData,"@ALL","Upp",T)
b3=sort.col(nData,"@ALL","ExpL",T)

d=matrix(0,ncol=10)
cp=matrix(0,ncol=10)
EL=matrix(0,ncol=10)

for(i in 1:10){
d[i]=i*(n/10)

## counting
   if(kp<b3[d[i],2]&&kp>b3[d[i],1]) {
      cp[i]=1
   }
   else {
      cp[i]=0
      count=count + 1
   }
EL[i]=b3[d[i],3]
}

list(d=d,cp=cp,EL=EL,nData=nData,a3=a3,b3=b3)
#list(d=d,cp=cp,data=data,nData=nData,a3=a3,b1=b1,b2=b2,b3=b3)
}
```

Appendix G. Programming Script: Calculating the CI based on a new statistic (mean, median and percentile)

```
CI.sort.real=function(n,mu,kp,ch1,ch2,Qr){
   set.seed(40)
   count=0

   theta=rvm(n,mu,kp)
   C=sum(cos(theta))
   S=sum(sin(theta))
   n=length(theta)
   kp=est.kappa(theta)

   A3=matrix(0,nrow=n)

   for(i1 in 1:n){

   A3[i1]=n-C*cos(theta[i1])-S*sin(theta[i1])
   kLow=ch1/A3
   kUpp=ch2/A3
   Gj=A3*kp
}

ExpL=kUpp-kLow
nData=cbind(kLow,kUpp,ExpL)
dimnames(nData)=list(NULL,c("Low","Upp","ExpL"))

#sort each column
a1=sort(kLow)
a2=sort(kUpp)
a3=a2-a1
a4=cbind(a1,a2,a3)
c1=colMeans(a1)
c2=colMeans(a2)
ELmean=c2-c1
d1=colMedians(kLow)
d2=colMedians(kUpp)
ELmed=d2-d1

a5=a1[(0.3*n):(0.5*n)]
a6=a2[(0.3*n):(0.5*n)]
```

```
a7=a3[(0.3*n):(0.5*n)]
a8=cbind(a5,a6,a7)

n2=length(a5)
d=matrix(0,ncol=6)
cp=matrix(0,ncol=6)
EL=matrix(0,ncol=6)

for(i2 in 1:6){
d[i2]=round(1+(i2-1)*Qr)

if(kp<a6[d[i2]]&&kp>a5[d[i2]]) {
      cp[i2]=1
   }
   else {
      cp[i2]=0
      count=count + 1
   }

EL[i2]=a7[d[i2]]

res1=rbind(kp,c1,c2,ELmean,d1,d2,ELmed)
res2=rbind(cp,EL)

plot(Gj)
list(Gj=Gj,a5=a5,a4=a4,a8=a8,d=d,res1=res1,res2=res2)

}
```

Appendix H. Programming Script: Analysis of missing values for circular data

```
mV3.Real=function(data,mIni,kIni,Per,cycle){
   #set.seed(30)
   n=length(data)
   data=as.matrix(data)
   d2=ContiNAmulti(data,Per) #distribute NA
   d3=na.exclude(d2) #data after excluding all NAs

   # parameter after excluding all NAs
   mu.1=CirMe(d3)
   kp.1=est.kappa(d3)

   mu1=matrix(0,nrow=cycle)
   kp1=matrix(0,nrow=cycle)
   nDr1=matrix(0,nrow=n, ncol=cycle)

   mu1[1]=mIni
   kp1[1]=kIni

   #replace the NA in dataset
   nDr1[,1]=replace(d2,is.na(d2),mu1[1])

for(j1 in 2:cycle){

   mu1[j1]=CirMe(nDr1[,j1-1])
   kp1[j1]=est.kappa(nDr1[,j1-1])

   #replace the NA in dataset
   nDr1[,j1]=replace(d2,is.na(d2),mu1[j1])

   final=j1
if((abs(kp1[j1]-kp1[j1-1])&&(pi-(abs(pi-abs(mu1[j1]-mu1[j1-
   1])))))<=0.0001)
   break


}
```

```
   mu2=matrix(0,nrow=cycle)
   kp2=matrix(0,nrow=cycle)
   nDr2=matrix(0,nrow=n, ncol=cycle)

   mu2[1]=mIni
   kp2[1]=kIni

   #identify the number of NAs and their location
   g=length(which.na(d2[,1]))

   gr=matrix(0,nrow=g, ncol=cycle)

   #generate value for imputation
   gr[,1]=as.matrix(rvm(g,mu2[1],kp2[1]))

   #replace the NA in dataset
   nDr2[,1]=replace(d2,is.na(d2),gr[,1])

for(j2 in 2:cycle){

   mu2[j2]=CirMe(nDr2[,j2-1])
   kp2[j2]=est.kappa(nDr2[,j2-1])

   #generate value for imputation
   gr[,j2]=as.matrix(rvm(g,mu2[j2],kp2[j2]))

   #replace the NA in dataset
   nDr2[,j2]=replace(d2,is.na(d2),gr[,j2])

}
   d=cbind(data,d2)
   para1=cbind(mu1,kp1,mu2,kp2)
   #para2=cbind(mu2,kp2)

par.est=cbind(final,mu1[2],kp1[2],mu1[final],kp1[final],mu2[cycle],kp2
   [cycle])

list(par.est=par.est,para1=para1)

}
```