ANKARA UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE

Ph.D. Thesis

APPLICATION OF CIRCULAR REGRESSION ANALYSIS ON BIOLOGICAL DATA

Desta Firdu MEKONNEN

DEPARTMENT OF ANIMAL SCIENCE

ANKARA
2017

THESIS APPROVAL

The thesis entitled "**APPLICATION OF CIRCULAR REGRESSION ANALYSIS ON BIOLOGICAL DATA**" carried out by Desta Firdu MEKOONEN satisfactory and recommended that it be accepted as dissertation for **DOCTOR OF PHILOSOPHY (Ph.D)** degree in the Department of Animal Science (Biometry and Genetics) in Ankara University Graduate School of Natural and Applied Science with the vote of the following jury on 21/04/2017.

Supervisor: Prof. Dr. Ensar BAŞPINAR
Ankara University, Department of Animal Science

Jury members:

Head : Prof. Dr. Ensar BAŞPINAR
Ankara University, Department of Animal Science

Member : Prof. Dr. Hülya BAYRAK
Gazi University, Faculty of Science, Department of Statistics

Member : Prof. Dr. Mustafa Muhip ÖZKAN
Ankara University, Department of Animal Science

Member : Prof. Dr. Handan ANKARALI
Düzce University, Faculty of Medicine, Department of Biostatistics

Member : Assoc. Prof. Dr. Serdal Kenan KÖSE
Ankara University, Faculty of Medicine, Department of Biostatistics

I approve the above decision.

Prof. Dr. Atila YETİŞEMİYEN
Graduate School Manager

ETHICS

I declare that all the information in this thesis prepared in accordance with the thesis rules of the Ankara University Graduate School of Natural and Applied Science. All information are correct and complete, that I have behaved in accordance with the scientific ethics in the course of the production of the information and cited all the resources I have used.

21.04.2017

Desta Firdu MEKONNEN

ABSTRACT

Ph.D Thesis

APPLICATION OF CIRCULAR REGRESSION ANALYSIS ON BIOLOGICAL DATA

Desta Firdu MEKONNEN

Ankara University

Graduate School of Natural and Applied Science

Department of Biometry and Genetics

Supervisor: Prof. Dr. Ensar BAŞPINAR

There are plenty of natural and artificial scenarios involve angular measurements. From our day-to-day activities to our metabolism, all have circularity in their nature, On the other hand, such scenarios are not studied using appropriate techniques until the emergence of circular statistics.

Circular statistics can be considered a new chapter of statistical analysis and modeling. Some statistical models consider circularity nature of data. On the contrary, some methods totally ignore circularity in observations even if it is clearly can be seen in the data in hand.

Even if there are some models that appreciate circularity in data they are very few. One of these areas is regression analysis of circular data. Even if regression analysis is a well-developed method when it comes to circular data it is still in its embryonic level. This is why the area is called the topic of current interest. There is no single universally accepted model (to our knowledge) in the area. Indeed, there are few circular regression models developed in the past 50 years for few data types. Applicability of these models in biological data is still in question.

In this thesis, we applied circular regression models on different biological data and examined strengths and weakness of these models based on acceptability, clarity, tractability and effectiveness when applied to those data.

Before jumping into circular regression analysis we have seen circular data, circular distributions, circular descriptive statistics and circular uniformity test since circular statistics is a new topic these chapters lay foundations to circular Regression methods.

May 2017, 110 pages

Key words: Circular data, Circular distributions, Circular Uniformity test, Circular Regression, Malaria, Crimean-Congo Hemorrhage Fever, Panic attack, Heart attack.

ACKNOWLEDGMENT

Firstly, I would like to express my gratitude to my advisor Prof. Dr. Ensar BAŞPINAR, a very understanding Professor for his everlasting helps. He was providing to me tirelessly with my study and any other matters I faced throughout my stay in Turkey. I am here because of his guidance throughout my courses and in writing of this thesis. I could not have imagined another advisor better understanding my situation more than he does.

Beside my advisor, I would like to thank my thesis committee Prof. Dr. Muhip ÖZKAN, and Assoc. Prof. Dr. Serdal Kenan KÖSE, their guidance brought me here others wise it would be unthinkable

I would also want to show my gratitude to my greatest professors Prof. Dr. Zahide KOCABAŞ and Prof. Dr. Orhan Kavuncu.

My special thanks to Mrs. Rabia ALBAYRAK DELİALİOĞLU, Mrs. Emel ÖZGÜMÜŞ DEMİR, Dr. Yasemin GEDİK, Ms. Özge ŞAHİN, without you people I can’t imagine I would be here.

My special Gratitude also goes to Ankara University Graduate school of Natural and Applied science, Ankara University Department of Animal Science, The Graduate School of Health Sciences Department of Biostatistics, Turkish Ministry of Education, TÜBİTAK, and YTB.

Desta Firdu MEKONNEN
Ankara, April 2017

TABLE OF CONTENTS

ABBRIVATIONS

AD	Anno Domino (years of the Lord)
ANOVA	Analysis of variance
BC	Before Christ
BCE	before the Common Era
c.d.f	circular distribution function
c.f	Characteristic function
c.r.v	circular random variable
CANOVA	circular Analysis of variance
CCD	continuous Circular distribution
CCHF	Crimean-Congo hemorrhage fever
circ.cor	circular correlation
cn	circular normal
Corr	correlation
cos	cosine
CP	characteristic properties
DCD	Discrete circular distribution
df	Degree of freedom
DfB	Degree of freedom between groups
DfT	Degree of freedom total
DfW	Degree of freedom within groups
Et al.	et alia (and others)
FP	Frequentist probability
Freq	Frequency
H_0	null hypothesis
H_1	Alternative Hypothesis
LSM	Least square Method
MCMC	Markov Chain Monte Carlo
MCMCM	Markov Chain Monte Carlo methods
MS	Mean of Squares

MSB	Mean of squares between groups
MSW	Mean of squares within groups
P. falciparum	*Plasmodium falciparum*
P. malariae	*Plasmodium malariae*,
P. ovale	*Plasmodium ovale*
P. vivax	*Plasmodium vivax*
pdf	probability density function
PPMC	Pearson Product Moment Correlation
R	Regression towards the Mean
Rad	Radian
Rufin	random uniform
Rvm	random Von Mises
sin	sine
SS	Sum of Squares
SSB	Sum of squares between groups
SST	Sum of squares total
SSW	Sum of squares within groups
TÜBİTAK	Türkiye Bilimsel ve Teknolojik Araştırma Kurumu
VM	Von Mises
WHO	World Health organization
YTB	Yurtdışı Türkler ve Akrabalar Topluluğu

LIST OF FIGURES

LIST OF CHARTS

1. INTRODUCTION

1.1 Background

Daniel J. Denis in his "Applied Univariate, Bivariate and Multivariate statistics" book Published in 2015 says "*learning new statistical techniques without consulting the earliest of Historical sources on those techniques a rather shallow and hollow experience*". It is better to start with the historical development of the subject matter "Statistics and Regression" and deductively narrow the topic to the main objective of the present work.

Varalakshmi et al. (2004) define statistics as "*Statistics is concerned with scientific methods for collecting, organizing, summarizing, presenting and analyzing data as well as deriving* ***valid concussions and making reasonable decisions*** *on the basis of this analysis*". Everything in statistics is buried deep in the concept of "*valid concussions and reasonable decisions*". To make valid conclusion and reasonable decisions we need sufficient knowledge and appropriate assumptions. John Tukey believes statistics is a science not a branch of mathematics but uses mathematical models and assumptions (Brillinger 2002). Besides Tukey, Stephenson (2000) argues statistics is neither a pure science nor a branch of mathematics. He clams statistics is a meta-science. Discovery of probability and its advancement as a game of chance plays grate role for structuration of statistics as a fields of inquiry (Lightner 1991).

The development and advancement of statistics in general and regression analysis in particular is believed to be very gradual and in different parts of the world independently. Fienberg (1992) criticizes even if understanding history of scientific fields such as statistics is as important as the science itself, writing some aspects of it is daunting task. Knowing the origin of statistics is very crucial. In the same writing Fienberg claims it helps statisticians to understand some of origins of their works and sense of what, statistically, discovery is all about.

Before becoming in modern sense and usage statistics was one of the oldest field of scientific enquiries. It has been used since the beginning of civilization.

The birth of modern statistics is credited to two British Statisticians (demographers) John Graunt (1620-1674) and Sir William Petty (1620-1687) where they developed early Human statistical and census methods that initiates the modern demography and census methods. John Graunt introduced the probability of survival to each year along with the life table. Although the original scope of statistics was limited to very few areas of interests and gradually increased the scope, areas of interests, and applications through time to too many areas of science and now statistics is mandatory to almost every aspect of human activities.

According to Shafer (www.probabilityandfinance.com, Anonymous 2017a) even if there were a number of attempts by Daniel Bernoulli (1700-1782), the most fruitful probabilistic work was publication of series of papers from 1771-1787 by Pierre Simon Laplace (1749-1827) that introduced of the concept of inverse probability. Later on in 19^{th} centuries, Laplace's papers collected and printed as book. The axiomatic approach of probability was introduced by Andre Kolmogorov, the concept which still in use for definition of probability in scientific problems.

Statistics is not remain as empirical science. It becomes also philosophical where there are two school of thoughts, classical and the Bayesian statistical school of thoughts. As one can see from many statistical texts and reference the development of statistics follows the classical (sometimes referred as traditional) statistical school of thoughts.

The same is true for correlation and regression analysis too (about their advancement). Even if correlation and regression analysis is frequently related to Carl Pearson due to the fact that he developed rigorous Mathematics of Pearson Product moment correlation (PPMC), Stanton (2001) claims it is the imagination of Sir Francis Galton that brought up the modern concept of correlation and regression. Stanton (2001) says "*Galton's Fascination with genetics and hereditary provided the initial inspiration that led to regression and the PPMC*". It was the problem of genetic hereditary enquiry that led

Galton to investigate further more on whether or not the characteristics of one generation is manifested on their offspring. Galton (1886) used a concept "regression towards mediocrity" to mean Regression towards the mean (RTM). He compared the height of children to their parents and found that they are closer to the mean than their parents are. According to Galton's theory, extremely short parents produce taller offspring and extremely tall parents produce shorter offspring than they do, in the long-range height tends to go towards the mean.

On the other hand, scientists give credits to the legendary mathematician Carl Friedrich Gauss or to his compatriot Adrien- Marie Legendre long before Galton or Pearson. Even there is a theory about allegedly the biggest scientific disputes in the early time of renaissance between these titans of scientific discovers on the ownership of the regression method.

It is believed to be gauss used the method regression probably in the name "trivial" but it was Legendre who published first on the topic. Legendre was also the believed to be the discoverer of the most famous statistical method, "least square Method (LSM)" in modern perspective even if the method was used long before him and "Polynomial regression" which may be the basis for circular regression methods. Later in 1809 Guess published a memoir that mentioned he used and has to be considered as the discoverer of LSM because he used the methods as early as 1795 to estimate the orbit of asteroids (Abdi, 2010).

Karl Pearson (the First professor of statistics in Britain) was a friend and colleague of John Galton and he wrote a bibliography of Galton after John Galton's death in 1911. In his four Bibliographies of Galton, he described the discovery of regression slope by the former. Previously Pearson believed it was Auguste Barvais (1811-1863), a French theorist who was interested in accuracy of astrological measurements, was responsible for the discovery of correlation and regression (Barnes 1998). Then after Galton work Pearson developed correlation and regression theories to almost to the way used today. He developed multiple correlation methods; different assumptions and requirements of correlation and regression analyses (Norton 1978).

Aldrich (2005) claims that, R.A. Fisher (1809-1962) in 1020s created modern regression analysis based on the two previously described theory of error theory of Gauss and correlation concept of Karl Pearson. According to him (Aldrich 2005) fisher's preconception of regression is marked by the two innovations of the conditionality of the two innovations of the conditional normal distribution of y's with specification linearity of x's and for inferential purpose x's can be considered as fixed but unknown. Latter on he, added to his regression analysis model goodness-of-fit" of regression Paper "the goodness-of-fit of regression formulae" 1922 using theory of previously examined theory of χ^2 and contingency table.

Fisher is also responsible for multivariate regression and some other nonlinear regressions. However, majority of statistical analyses we have seen so far is based on the **magnitude measure of variables of interest**. What if we need "**direction**" rather than the "**magnitude**" as we can see plenty of natural and artificial scenarios around us where the measure if interest is actually direction.

Measure of direction is probably the oldest of all human activities since the date human being wanted more than what his environment provides his needs. The History of directional analysis is closely related to earliest human migration and navigation systems. Sailors in the Mediterranean Sea used to a navigation system like wind direction, position of the sun and stars, the sea current direction and …so on.

Ancients Greeks used celestial navigation systems. In south china and Indian Ocean, navigators used a constant monsoon winds to judge their direction of movements (Linton 2013). According to Linton (2013), the Arab world contributed for development of directional judgement using a magnetic device called "Kamal", a device used to measure the altitude and longitude of celestial bodies including moon. Advancement in modern navigation system was initiated by commercial trade systems of Portuguese, Spaniards and Turks in the age of exploration in 1500's. In modern time, British commissioners were responsible for advancement of compass system.

On the other hand, a mathematical measurement of degree were believed to be originated in Egypt around fifteen hundred BC degree measurement system were taken from the sun's shadow (Wallis 2005). According to Wallis (2005), the history of the first navigation system that resembles the current compass were the "principle of astrolabe". Islamic Spain introduced the system to Europe. The first and second world war contributed for the advancement of compass system for military purpose.

Clock measurement is one of the two principal statistics of circular analysis. Time is very different from other physical measurements like Temperature, length and Mass in many ways (Jespersen and Fitz-Randolph 1999). We can see and feel distance; we can weigh any physical entity; we can feel Temperature, but we cannot see or feel time measurements, we relate time with events. That makes it difficult to have a clear sense of measurement on time.

It is a known Fact movement of sun, moon and stars marks awareness of time measurements and interval. According to Jespersen and Fitz-Randolph (1999), there are countless other cycles and rhythms going on around us. Biologists, Botanists and other life scientist study but not yet not fully understood "built-in" clocks that regulates basic life processes like earth beats, breathing, ministerial cycles in primates…etc.

Jespersen and Fitz-Randolph (1999) also clarified about Geographical process that happened million years ago and they speak about a term "geologic time".

Time in its cyclic nature was used to provide an adequate ways to get with environmental phenomena. The coming of and changing of seasons, the dark and daylight of the day; size and direction of the moon, flow of rivers; used to basics for current calendar system.

Topics and explanations in the above few paragraphs shows circular measurements of compass and time were at least as old as other scientific enquires like Biology and statistics in linear sense. On the other hand; since the emergence of statistics as field of

science in majority of studies, estimations, predictions and models on the subject remained linear. Gaile and Burt (1980) argues **traditional statistics** are linear in sense that measurements are distributed on a number line. The term "*Traditional statistics*" sometimes is confusing as it is used to comparison "the linear" statistics from any form of others methods. For example, the term Traditional statistics is used to compare a frequentist approaches with Bayesian methods. Duckworth and Stephenson (2012) published a paper in a topic "*Beyond traditional statistical methods*". In their paper, they presented the need for new and better statistical methods like dynamic graphic, nonlinear estimations, resampling and simulation based inference systems; they call it "modern method".

In this thesis by applying circular regression analysis to different biological data, we planned to examine the applicability of circular methods in general and circular regression methods in particular. When we say applicability we mean by both statistical tractability and logically mindfulness since there are plenty of models that are statistically tractable but logically meaningless.

1.2 Significance of Circular Statistics

The majority of statistical analysis so far was based on the assumption of linearity in the relationships of variables of interest. On the other hand, there are plenty of natural and artificial scenarios that have circularity (directionality) in their nature. Modeling these scenarios as if they are linear could lead to erroneous outputs.

Circular statistics can be available in any aspects of science. Srinivas and Rao (2016) published a paper on importance of circular data on **sports science**. They recommend coaches and sport science specialists has to focus on circular aspects of the science as well. Jeff's paper (2008) prepared for American political science associations annual meeting used a second Iraqi war and Gun violence in large American cities data to demonstrate how circular statistics in vital to **political and crime study**. Jeff and Hangartner (2010) publish another paper how crime is distributed over 24 hours in America and they applied circular regression method to find out the regression of crime

over time of the day. Guterman et al. (2009) paper also used circular statistics for **psychophysical** research.

The majority of circular statistical analysis used data from **Geography**: wind direction, altitude, and latitude because availability and modeling of such data are relatively easier than data say, in biology and day-to-day activities where the data structure is complex for circular analysis.

Circular statistics is even applicable in our **simple day today activities**. For example, Corcoran et al. (2008) used circular statistical methods to analyze "*journey to work*" data in southeast Queensland region in Australia. Their finding in the study area shows that there is strong spatial patterns of flows for the study area when used circular statistics.

Circular statistics is related to geometry since it uses trigonometric expressions and derivations. It is better to see the development of circular statistics from a geometry point of view as well. Due to this fact, circular statistics needs more sophisticated mathematical knowledge than the traditional statistical methods.

There is also Bayesian analysis for circular data recently developed for circular ANOVA and even for regression analysis involving circular variables using wrapped distribution as a basis for obtaining circular data from linear counterparts (Ravindran and Ghosh 2011).

1.3 Problem of Statement

Even if there are plenty of natural and artificial scenarios that can/should be analyzed using circular statistical methods, the method is a recent idea. It has been done linear analyses or some sorts of time series analyses for data that are actually have to be analyzed using circular methods. For example, many seasonal scenarios are more suitable to be analyzed using circular methods rather than time series analysis.

As it was mentioned in the introduction chapter, circular statistical analysis methods can be considered in its "infant level" when compared to other methods of estimation and prediction.

Indeed, there are some methods and models in the topic. Nonetheless, the majority of them lack robustness.

Greenwood and Duran (1953) presented burning argument on the topic "*Although numerous Theorems have been developed, the question of how to recognize a cycle remain open*" back in 1953. Even if that much time gap passed between Greenwood's and Duran's question and today there is not internationally accepted theorems and models on the topic yet. Moreover, **these methods are not yet effectively applicable to biological data**.

1.4 Objectives of the Thesis

In the thesis, we study and evaluate different statistical and biological problems that are related to circular statistical analyses methods since circular statistical analysis methods are very different in many ways from the usual statistical methods we know. We planned to address very important objectives about the topic mentioned above and our goals can be summarized as follows

Objective 1. The principal objective of this thesis is to perform circular regression analysis on different biological data and evaluate applicability, tractability, and meaningfulness of these models both mathematically and biologically.

Objective 2. Circular analysis methods are very new subject so jumping directly into regression analysis might create some confusion on the matter. Due to this fact, before doing regression analysis we setted an objective to give some revelation aspects of circular data by addressing questions like

I. What is meant by circular data?

II. What makes circular data unique?

Objective 3. To evaluate some of statistical values like “circular probability distributions, circular descriptive statistics, circular uniformity test and whether or not our data sets follow a specific distribution.

Objective 4. The final principal goal of the thesis is to compare losses and gains when used circular analysis methods mathematically, cost effectiveness, and with some related issues over some other “traditional” methods.

Objective 5. We also have our emphasis on the limitations of circular methods and models

Objective 6. Beside application and comparison of methods and models, we also set a goal to give ideas about future works on the topic.

2. MATERIALS AND METHODS

2.1 Materials

In this chapter, some of the methods applied to accomplish the general goals of this thesis are presented. Specific methodologies are presented in respective chapters since we used different methods and data in different chapters.

The biggest challenge faced to accomplish objectives of the paper was finding appropriate data for circular analysis in general and circular regression analysis in particular. The majority of data that are found online and even in some laboratories are not suitable for circular analysis. Moreover, since circular data are longitudinal data type finding such data consume a great deal of time. Due to this fact, we are forced to keen on secondary and simulated data; and modify them in a way they can be used for circular analysis. Different data sources and materials are used in the paper and they are presented as follows.

2.1.1 India

Among materials we used one is an open access malaria data has been used for "*association of climatic variability, vector population and malaria disease in the district of Visakhapatnam*", India: a Modeling and prediction analysis" (Stuckey et al. 2014) in which data collection methods have been described in detail. Monthly malaria data were collected from the district for the 2005-2011 period. From the data, we chose the *Plasmodium Falciparum* case the 28th indexed study area in the paper as shown in Figure 2.1 of the shaded area.

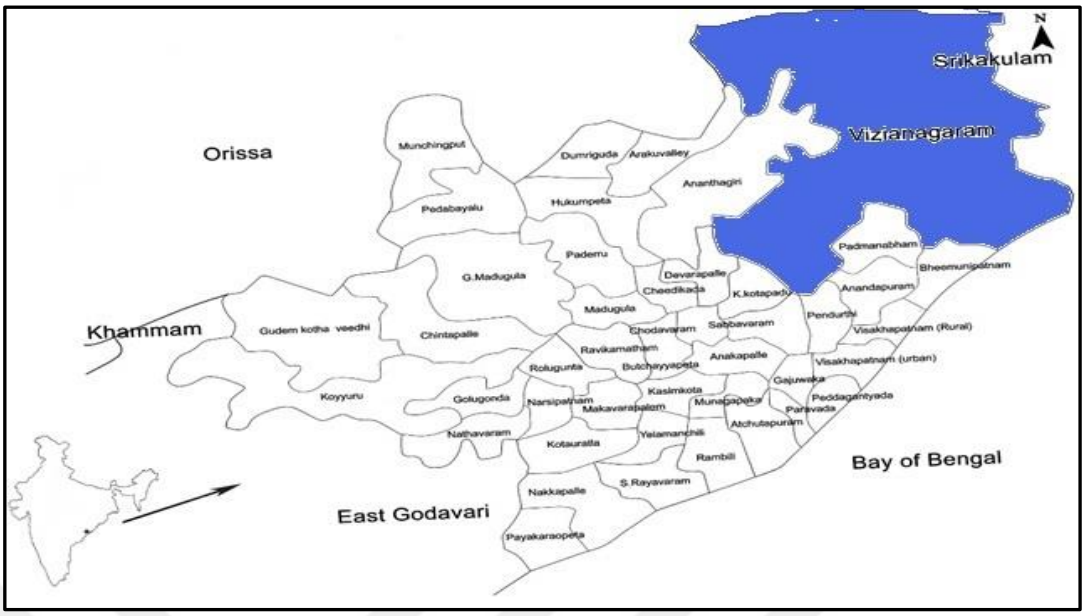


Figure 2.1 Map of the study area Visakhapatnam district

Chart 2.1 *Plasmodium Falciparum* data in Visakhapatnam (U)

			2005	2006	2007	2008	2009	2010	2011
S.No	Mandal Name	Month	Pf	Pf	Pf	Pf	Pf	Pf	Pf
28	Visakhapatnam (U)	January	58	52	77	57	71	66	67
		February	44	64	79	62	84	60	100
		March	53	72	78	83	127	120	166
		April	72	73	111	125	125	163	196
		May	112	135	118	145	206	170	280
		June	154	165	164	202	150	204	318
		July	149	233	176	360	217	416	430
		August	231	178	231	284	408	441	431
		September	204	182	219	295	448	225	388
		October	113	154	167	162	206	184	191
		November	86	137	97	128	123	133	138
		December	65	61	55	70	69	74	102

The data in chart 2.1 is used for calculation of circular descriptive statistics and linear-circular regression.

2.1.2 Turkey

The second material used in the paper is occurrence of CCHF (Crimean-Congo hemorrhage Fever) in Turkey.

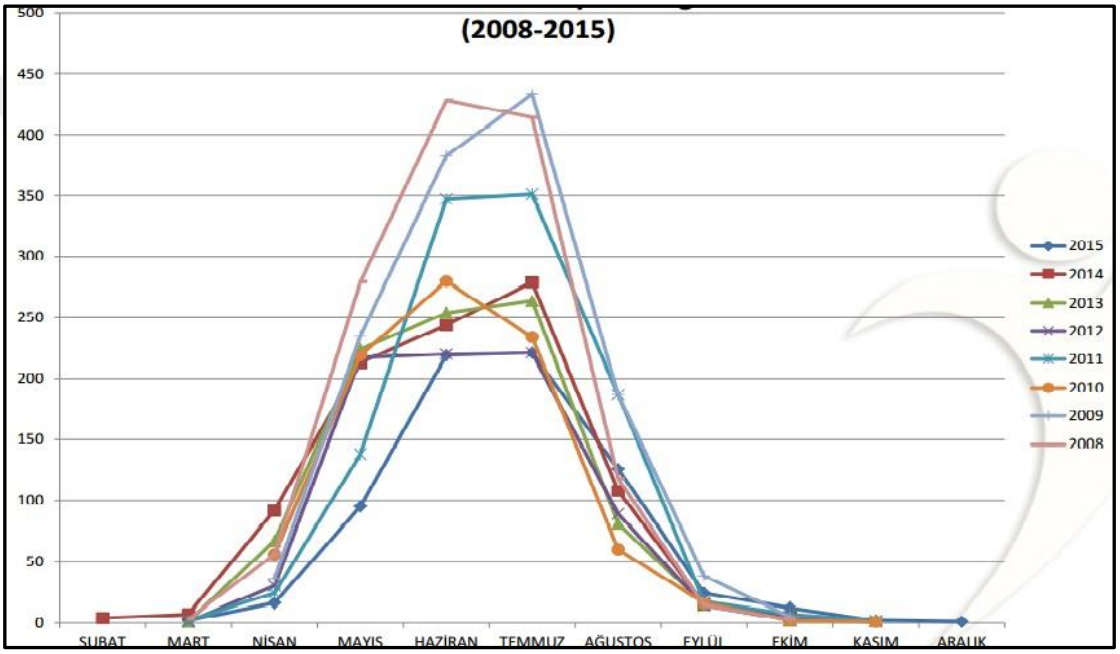


Figure 2.2 Monthly data plots of CCHF from the year 2008-2015

Data plot in figure 2.2 is used for analysis of circular-linear regression.

2.1.3 Taiwan

Chart 2.2 Maximum panic attack data in a week in Taiwan

Days	Maximum attack
Monday	10
Tuesday	8
Wednesday	8
Thursday	9
Friday	9
Saturday	12
Sunday	12

The data in chart 2.2 is used for circular uniformity test.

2.1.4 The other material we used is simulated data

Chart 2.3 head and tail of birthdate and heart attack date simulated data

Index	Birthdate	Heart attack date
1	0.278187	1.232262
2	5.26886	-1.88462
3	0.384266	0.076561
4	2.168721	0.157084
	…	…
	…	…
	…	…
1332	0.762404728	0.233955145
1333	2.435280699	0.013507469
1334	1.169476352	0.078194345
1335	2.033624633	0.125431734
1332	0.762404728	0.233955145
1333	2.435280699	0.013507469

The heart attack simulated data in chart 2.3 is used for circular-circular regression analysis.

For plotting, analysis, calculation and abstaining CCHF data we used R, excel, *GetData Graphics Digitizer 2.26* software.

** **Remark**: three dots in rows and columns of any table indicates in this paper data lines that are not specified because they are too long.

2.2 Methods

Circular analysis in general and circular regression methods, in particular, is a new area of statistical methods. Knowing introductory aspects of circular statistics has a huge advantage in circular regression analysis. Due to this fact, it will be better and sometimes mandatory to introduce concepts of statistical data structures and some introductory methods before doing circular regression.

2.2.1 Circular data

Even if there were numerous attempts to use circular analysis methods in different fields of applied sciences, Fisher (1992) gives credits to G.S. Watson and E.J. Williams for pioneering of circular statistical analysis as it is used today. These two statisticians (1956) introduced methods to infer about mean direction and dispersion of single sample from Von Mises distribution. They also applied comparison of two samples and more samples for circular data, which is analogs to ANOVA of the linear data analysis.

According to Fisher (1992), Mardia's paper (1972) covers many aspects of estimation and inferential statistics of circular statistics stochastically. His paper covers summarization and goodness-of-fit test as well as parametric and non-parametric approaches.

Many statisticians including Rao (2001), Fisher (1995), Downs and Mardia (2002) argues that within two decades after Mardia's Book (1972) some advanced forms of circular statistical analysis such as circular correlation, Regression, time series analysis of circular data, large sample methods and bootstraps, non-parametric density estimations and special smothering are developed.

The circular analysis is very different in many ways from customary analysis used in the majority of predictions and estimation methods.

The first unique feature of the circular analysis is the data type. Circular data arises mainly from two types of measurements, which are **compass** and **clock in circular sense** Rao (2001). The direction of migrations of birds Ozarowska (2013), wind direction, latitude and longitude of the globe, the direction of epidemic diseases are few examples of directions. There are plenty of examples and research problems for a time in a circular sense. Data types like an emergency in the hospital within 24 hours, the number of accidental death in the USA within a year Herone (2016), the occurrence of any epidemic disease in cyclical time Legrand (2007) are very few of them.

The natural measurement of circular data is degrees or radians. Therefore, circular time measurements have to be converted into degrees or radians by the following conversion equation bellow of circular analysis purpose.

$$\theta = \frac{2\pi * x}{y} \tag{2.1}$$

Where;

θ is time in radian,

X is time to convert in to radian,

Y is time measurement considered full cycle of time.

Normally Radian is considered a better measurement than degrees because degree measurements are just division of into 360° arbitrarily equal partitions. On the other

hand, radian measurements are related to the radius of a circle. Radian measurement gives the ratio of the arc length to the radius of a circle, which is mathematical; fulfill the notion of differentiating trigonometric calculations.

$$\frac{d}{dx}sinx = cosx \; and \; \frac{d}{dx}cosx = -sinx$$

On the other hand, if one uses degree measurements it includes additional factors and complicate the calculation

$$\frac{d}{dx}\sin(x) = \left(\frac{\pi}{180}\right)\cos x \text{ and } \frac{d}{dx}\cos(x) = -\left(\frac{\pi}{180}\right)\sin x$$

Since radian is the ratio of two lengths in same unit radian measurements are considered "*unites*". This notion gives an easier way to calculate some linear displacement from circular motion and helps to get results with sensible measurement units. For instance, in figure 2.3, arc length $\widehat{0P} = R\theta$, differentiating both side of this equation with respect to time will result a linear velocity ν and angular velocity ω with an equation ν=ωR. If ω is measured in radian/second (r/s) and R in centimeter (cm), the resulting linear velocity will have cm/s units, which is natural measurement of velocity. If we were using degree measurement instead of radian measurements, the resulting unit would be degree/second, which is not natural measurement of velocity. The question raises here is "if radian measurements are the "right" measures of angles why degree measures are so popular?" Kupkova (2005) argues this is due to simplicity of whole number system in degree measurements.

The other unique feature of circular data is there is no natural zero point and sense of direction. 45° (1/4π rad) to the east might be zero for mathematician or 60° (2/3π rad) to the north may be zero points for ecologist. Due to this fact, rank-based statistics have to be done with ultimate caution. In circular data collection, the only requirement is all measurements should have the same zero point. This means if one decide February as a starting zero point all measurements have to be referenced from February. In this sense, April will have π/6 radian value.

Circular measurement can be done into two ways. As shown in the figure 2.3, measurements can be Arc length of $\widehat{0p}$ on the circumference of the circle or an angle those points make from the center of the circle. In circular data analysis majority of the objectives are to assess directions so knowing magnitudes of any data has less importance. Due to this fact (Mardia and Jeff, 1972), usually circular data are represented in unit circle that means a circle with unit radius.

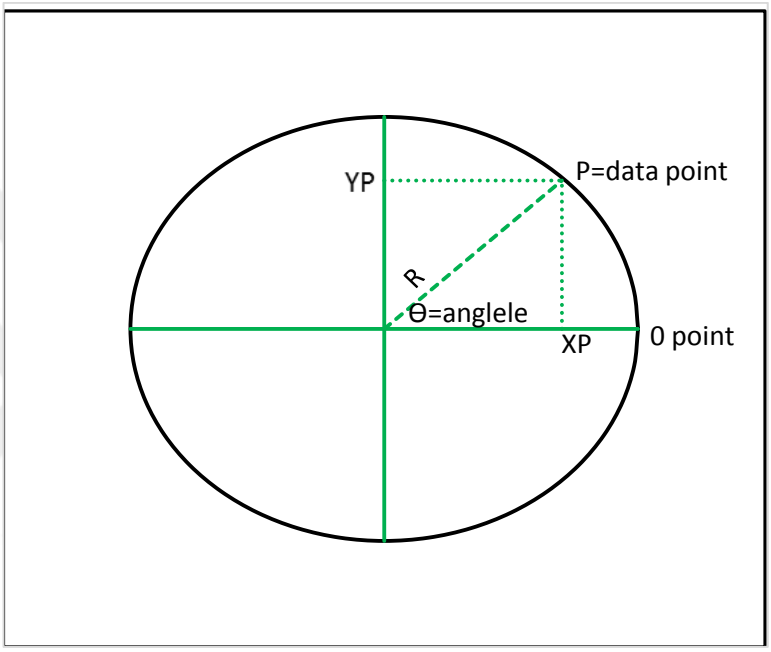


Figure 2.3 Data point P on the unit circle

2.2.2 Circular distributions

In linear statistical analysis and inference, identifying the distribution of the data is a key to any statistical analysis and decision-making. It helps to choose the right statistical method (Duckworth and Stephenson 2012). The same is true for circular statistics too. Prior to any circular analysis knowing the distribution of the data in hand is very important. Usually, it is mandatory too, since we use Maximum Likelihood Estimation (MLE) or Ordinary Least Square Estimation (OLS) methods for inference or predictions. Moreover, knowing the distribution of circular variable helps to assess if there is multimodality. When there is multimodality in the distribution statistical

methods we use will change, otherwise, we may arrive at completely different result than we supposed to have Rao (2001).

Circular distribution is a probability distribution whose total probability is concentrated on the circumference of a unit circle (Mardia and Jeff, 1972). In this basis, any measurement that has cyclic nature can have circular distribution.

In circular logic, each data point on the circumference of a circle represents direction. Therefore, this is how to assign probabilities for different data points. The range of circular random variable (c.r.v) measured in radian may represent to be $[0,2\pi)$ or $[-\pi, \pi)$ similar wise in degree measure $[0,360°)$ or $[-180°, 180°)$. In both ways, circular random variables are bound to full circle.

Let *X* be a random variable on the circumference of a unit circle $X^2+Y^2=1$;
The distribution function of F of X is given by the equation
$F(\theta) = \Pr(0 \le X \le \theta)$Where $0 < \theta \le 2\pi$ and measuring Θ with its components (cosΘ, sinΘ) in an anticlockwise (can be clockwise as well) direction from positive X-axis (starting point is mainly arbitrary). Mardia (1972) argues it is unnatural to restrict the domain of the distribution to $0\text{-}2\pi$. Therefore, we can extend the domain by defining the distribution over any multiple cycles of a circle as follow

$$F(\theta + 2\pi) - F(\theta) = 1 \qquad (2.2)$$

Where

$$\infty < \theta < \infty$$

2.2.2.1 Types of circular distribution

A. Discrete circular distribution: discrete circular distribution is the circular distribution type assigning probability masses only to a countable number of directions. A discrete circular distribution sometimes referred as circular Lattice distribution.

Mardia (1972) explained as

$$\Pr(\theta = v + 2\pi r/m) = pr(\theta) \text{ for r=1,2,…,m-1}$$

and

$$pr(\theta) \geq 0 \text{ and } \Sigma pr(\theta)=1$$

When Point v+2πr/m are spaced equally on the circumference of a circle (unit circle), this distribution may be seen as an m-sided polygon. When points on the circumference go to infinity the polygon goes to a perfect circle. If all data points have same weight then the probability of any point equals 1/m.

For v=0, the completive distribution function of discrete circular distribution is given by

$$\phi_p = \sum_{r=0}^{m-1} pre^{\frac{2\pi r}{m}} \tag{2.3}$$

For $p = 0(\text{mod}\, m)$ $\phi_p = 1$

Where the distribution is uniform

$\phi_p = 1$ if $p = 0(\text{mod}\, m) = 0$

A. Continuous Circular Distribution

Circular Continuous distribution (ccd) is a distribution assigning probability density (pdf) for absolutely continuous direction.

Continuous distribution has the following axiomatic properties

$(i).f(\theta) \geq 0;$

$(ii).\int_0^{2\pi} f(\theta)d\theta = 1;$

$(iii).f(\theta) = f(\theta + k.2\pi)$

The first property implies - the probability of any direction θ is equal to or greater than zero. The second property implies integral of all probabilities is one. The third and most the unique property of circular distribution implies random variables θ_1, θ_2, θ_3..,θ_n have same probability after making full cycle (2π) with k=1,2,3,… a number of rotations.

2.2.2.2 The characteristic function

Mardia (1972) advises to refer characteristics function of a circular variable for a better understanding of their characteristics. In this explanation and formulation, let us borrow his ideas;

Let Z be a random variable on the unit circle. We shall identify Z with a random variable θ where $0<\theta\leq 2\pi$

We can express Z with its characteristic function as follows

$$Z = e^{i\theta} \tag{2.4}$$

Mardia also warns notationally we shall not distinguish between a random variable and its values.

As it can be done in linear form the function, characteristic function can be found using integrals as follows

$$\phi(t) = E\left(e^{i\theta}\right) = \int_0^{2\pi} e^{it\theta} dF(\theta) \tag{2.5}$$

Where F(θ) is c.d.f (circular distribution function)

Using expression 2.5

$$\int_0^{2\pi} e^{it(\theta+2\pi)} dF(\theta) = \int_0^{2\pi} e^{it\theta} dF(\theta) \tag{2.6}$$

Expression 2.6 implies that

$$e^{2it\theta} = 1 \tag{2.7}$$

Whenever there is ϕ with |ϕ (t)|≠0 and according to Mardia (1972) this suggest that the function ϕ(t) should only defined for integer values of t's. Moreover, the theory of Fourier functions of periodic function states that it is sufficient to take t as integer.

Based on the above explanation c.f ϕ_p of θ is express as follows

$$\phi(t) = \int_0^{2\pi} e^{ip\theta} dF(\theta) \tag{2.8}$$

for p=…,-2,-1,0,1,2,…

2.2.2.3 Ways of obtaining circular distributions

Similar to the linear case there are plenty of circular distributions. Many important circular models can be generated using specified circular probability distributions on a real line or on a plane (Rao 2001). Knowing ways of generating any distribution is as important as the distribution parameters itself because it helps to understand the underlying logic behind the distribution generated.

In this chapter three of the four basic methods of generating circular distribution is presented.

I. By wrapping a linear distribution around the unit circle.
II. Through characterizing properties such as maximum entropy.
III. By transforming bivariate linear random variable to just its directional components, the so-called offset distribution.

2.2.2.4 Wrapped distributions

Theoretically, any distribution on a real line can be wrapped on a circle and produce circular distribution. Basically, wrapping linear variable on to a circumference of a unit circle is the most common way to get circular distributions.

Given X_1, X_2, $X_3 \ldots X_n$ random variable with distribution function F(x) on a line, random variable Xw of the wrapped distribution is given by

$$X_w = X \,(\mathrm{mod}\, 2\pi) \tag{2.9}$$

The distribution has a probability distribution function as follows

$$f_W(\theta) = \sum_{-\infty}^{\infty} f(\theta + 2k\pi) \tag{2.10}$$

Where

$\theta \in [0, 2\pi)$

Wrapped distribution has many distribution families. The most famous wrapped distributions are Wrapped normal, Wrapped Cauchy, General wrapped stable. These distributions are derived from similar distribution in a linear line. Exponential wrapped distributions by wrapping classical exponential on a circumference of a unit circle.

Rao's (2001) wrapped exponential distribution has probability density function (pdf)

$$f(\theta) = \frac{\lambda e^{-\lambda\theta}}{1 - e^{-2\pi\lambda}} \tag{2.11}$$

Where $\theta \in [0,2\pi)$

The probability density function (pdf) of wrapped exponential distribution is strictly decreasing on the interval $[0,2\pi)$ if $\lambda>0$ and strictly increasing on the interval $[0,2\pi)$ if $\lambda<0$.

Days pass by new families of wrapped distribution emerges. For instance, Joshi and Jose (2016) presented wrapped Lindley distribution, a very close distribution to exponential distribution.

For better understanding, the logic of wrapping pdf of Lindley linear and its wrapped form is presented as follows as presented by (Joshi and Jose 2016) and (Ghitany et al. 2008).

pdf of Linear Lindley distribution

$$f(x) = \frac{\lambda^2}{\lambda+1}(1+x)e^{-\theta x} \tag{2.12}$$

for $x > 0$ *and* $\lambda > 0$

pdf of wrapped Lindley Distribution

Taking a random variable $\theta = x(\mathrm{mod}\, 2\pi)$ such that $\theta \in [0,2\pi)$ from a linear Lindley distribution on a real line generate wrapped variable

$$g(\theta) = \sum_{m=0}^{\infty} f_x(\theta + 2m\pi)$$

$$\Rightarrow \left(\frac{\lambda^2}{1+\lambda}\right)\sum_{m=0}^{\infty} e^{-\lambda(\theta+2m\pi)}(1+\theta+2m\pi)$$

$$\Rightarrow \left(\frac{\lambda^2}{1+\lambda}\right)e^{-\lambda\theta}\left[(1+\theta)\sum_{m=0}^{\infty} e^{-2\pi\lambda m} + 2\pi\sum_{m=0}^{\infty} m e^{-2\pi\lambda m}\right]$$

$$\Rightarrow \left(\frac{\lambda^2}{1+\lambda}\right)e^{-\lambda\theta}\left[\frac{1+\theta}{1-e^{-2\pi\lambda}} + \frac{2\pi e^{-2\pi\lambda}}{\left(1-e^{-2\pi\lambda}\right)^2}\right] \tag{2.13}$$

for $\theta \in [0,2\pi) and \lambda > 0$

The above derivation of Lindley wrapped circular distribution from Lindley linear distribution can be used as a basis for other wrapped distribution from their linear counterparts.

2.2.2.5 Using characteristics properties to get circular distribution

Construction of circular distribution creates plenty of problems even for well-trained statisticians because methods for linear data or distribution cannot be utilized due to the disparity of the topologies on the line and the circle and the periodicity requirement has to be mate for a circular distribution to be meaningful (Sengupta 2004). According to Sengupta (2004) using characterization methods of distributions on a line can produce a new univariate, bivariate and more distributions on a circle, cylinder or on a torus.

For simplicity let's take univariate characterization based on Shannon's Entropy.

Let *X* be a continuous random variable with pdf =f(x;n), then Shannon's information is defined as follows

$$H(f) = -\int_{-\infty}^{\infty} f(x;n) \log f(x;n) dx \tag{2.14}$$

Sengupta uses majorization and Schur-convex function for information measure on non-negative, let

$$\begin{cases} f(x;n) > 0 & for\ x \in (a,b) \\ otherwise & 0 \end{cases} \tag{2.15}$$

Let's consider the F-class parametric density function $f(x;n)$ that satisfies constraints

$$\int_a^b T_j(x) f(x;n) dx = \tau_j \tag{2.16}$$

for $j = 1,2,...,k$

For given a function $T_1(x), T_2(x), ..., T_k(x)$ on (a, b) whose definite integral exists and constraint

$\tau_1, \tau_2, ..., \tau_k$, then variational methods for this function and its constraints can be used to find the class of densities that maximize H(f) over F. the maximum entropy over the class F can be attained by exponential family of distributions with the density of form as follows

$$f(x;n) = C.e^{\left[\sum n_i \tau_i(x_i)\right]} \tag{2.17}$$

Where C is normalizing constant to be determined by invoking constraints in equation 5.14 and 5.15 where is exists $n_1, n_2, ..., n_k$ that satisfies 5.16. For the alternative proof of characterization methods of circular probability Sengupta (2004) recommend theorem 13.2.1 of (Kegan et al. 1973).

2.2.2.6 Circular uniform distribution

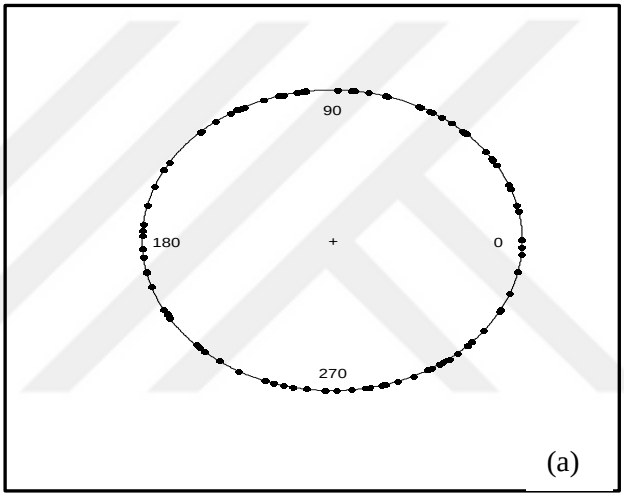
One of the most important distributions of circular analysis is circular uniform distribution. When data values distributed uniformly on the circumference of the circle there will not be preferred direction (circular mean), circular variance, and comparison of means (ANOVA). Since all directions are equally likely, the distribution is also called isotopic or random distribution. The circular uniform distribution shows a constant density. This distribution is independent of measured angles. This distribution has maximum entropy when there is not a question of mean direction since the distribution doesn't have mean direction.

$$f(\theta) = \frac{1}{2\pi} \tag{2.18}$$

Where $0 \leq \theta < 2\pi$.

The cumulative distribution function of Uniform distribution is given by the equation bellow where $F(0) = 0$ and $F(2\pi) = 1$

$$F(\theta) = \frac{\theta}{2\pi} \tag{2.19}$$



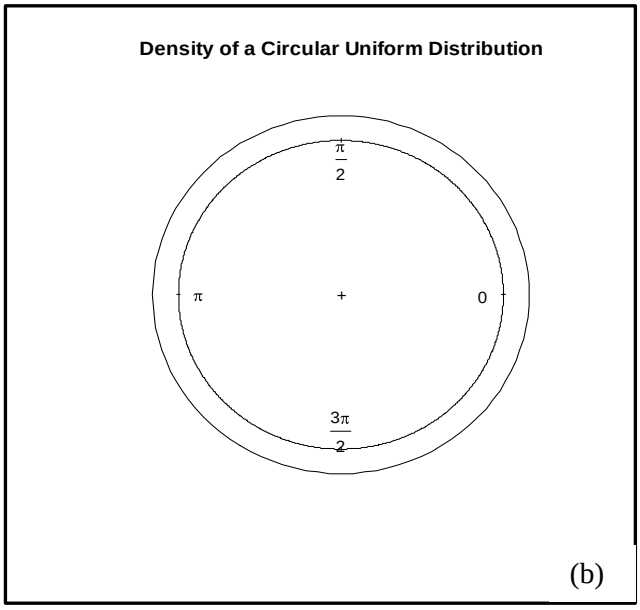


Figure 2.4.a. data point plot of circular data, Figure b. Density of a circular uniform distribution

There is an old saying “a picture is a worth a thousand word”. One of the integral parts of inferential statistics is a data visualization. Knowing the distribution of circular variables helps to assess if there is multimodality. If there is multimodality in distribution the statistical method we use will change, otherwise, we may arrive at completely different result than we supposed to have” (Mekonnen 2017). When it comes to circular data visualization more important than linear data types. Based on the idea above, circular distribution of uniform simulated data using r-software packages is presented in figure 2.4a.b Shows circular uniform data points on the circumference of a circle (preferably unit circle). Figure 2.4a shows circular data points on the circumference of a circle whereas figure 2.4b shows circular density distribution where there is no peaks and troughs to indicate these data plots follow a uniform distribution.

2.2.2.7 Circular normal distribution

Normal distribution is the basis of the assumption of plenty of statistical decision-making processes especially when one has to use parametric analysis. This is the distribution first described by Carl Friedrich Gauss two century ago (Roussas 2003). Gauss proved that Normal distribution can be derived from a logic what is now called Maximum Likelihood with a single assumption of the most probable value. Since then the distribution became the basis of numerous statistical analysis and modeling assumptions. In addition, it becomes a starting argument for developments of other distributions. Its popularity is not from nothing, this distribution gains its popularity from assumptions that it approximates too many natural phenomena with the notion of the Central limit theorem (Ghasemi and Zahediasl 2012). It is two parametric namely, mean and standard deviation, distribution with pdf as bellow

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)}{2\sigma^2}} \tag{2.20}$$

Where μ is a location parameter and σ is a scale parameter. It is not surprising that Richard Von-Mises in 1918 used this distribution as a basis to develop his continuous

probability distribution on the circle. His distribution is also called circular normal distribution because it shares many desirable properties circular distribution have.

Angular random variable θ or a linear variable X, (wrapped on the circumference of a circle) said to have Von-Mises or circular normal distribution if it has the density function. Although names the Von-Mises distribution VM(μ,κ) and Circular normal CN(μ,κ) distribution are used interchangeability, in this paper we prefer to use Circular normal unless we are forced to use Von-Mises distribution VM(μ,κ) when for simulation purpose.

$$f(\theta,\mu,\kappa)=\frac{1}{2\pi I_0\kappa}e^{k\cos(\theta-\mu)} \tag{2.21}$$

Where

$0 \le \theta < 2\pi$, $0 \le \mu < 2\pi$,

$K \ge 0$ implies a concentration parameter

$I_0(k)$ is modified Bessel function of the first kind and order zero.

$$I_0(\kappa)=\frac{1}{2\pi}\int_0^{2\pi}\exp(\kappa\cos\theta)d\theta=\sum_{r=0}^{\infty}\left(\frac{\kappa}{2}\right)^{2r}\left(\frac{1}{r!}\right)^2 \tag{2.22}$$

The above Bessel function is based on the Bessel equation of differential form as bellow

$$x^2y''+xy'+(x^2-p^2)y=0 \tag{2.23}$$

As it is known from differential function solving procedures the solution of the differential equation is not constant (numeric) but it is a function (may be linear). Plugging the above equation in the equation bellow will give us Bessel function

$$y(x)=x^n\sum_{k=0}^{\infty}b_kx^k \tag{2.24}$$

The process of finding Bessel function and using it in circular distribution is handy for especially non-mathematises. Fortunately, there are plenty of package programs to ease the problem. Some properties of the circular-normal distribution is explained by Gubel et al. (2012) they can be summarized as bellow.

Properties of circular normal distribution

I. **Symmetry**: circular distribution is symmetrical distribution about mean direction μ and (μ+π).

II. **Mode**: - since the cosine function has a maximum value at zero, the circular normal distribution is maximum at θ=μ.

$$f(\mu)=\frac{e^{\kappa}}{2\pi I_0(\kappa)} \tag{2.25}$$

III. **Antimode**: - antimode of a circular normal distribution is at μ±π, since cosine of π =-1 which is the minimum value.

$$f(\mu\pm\pi)=\frac{-e^{\kappa}}{2\pi I_0(\kappa)} \tag{2.26}$$

IV. The role of k

When one see the ratio of mode versus antimode the result equation is

$$\frac{f(\mu)}{f(\mu+\pi)}=\frac{\dfrac{e^{\kappa}}{2\pi I_0(\kappa)}}{\dfrac{-e^{\kappa}}{2\pi I_0(\kappa)}}=e^{2\kappa} \tag{2.27}$$

V. Trigonometric Moment of CN distribution

The above equation shows the effect of the ratio of the mode and antimode is only depend on the concentration parameter k. the higher the ratio shows the wider the gap between mode and antimode, the wider the gap results in the Peaker the distribution. This is due to the higher concentration of the data towards the mean direction. The concentration parameter k plays the role of variance hence standard deviation in linear normal distribution.

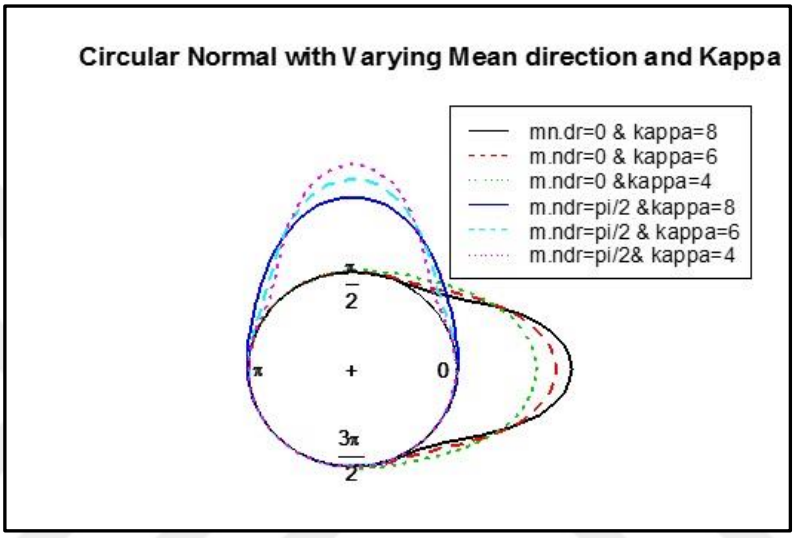


Figure 2.5 Circular normal distribution with varying mean direction and Kappa

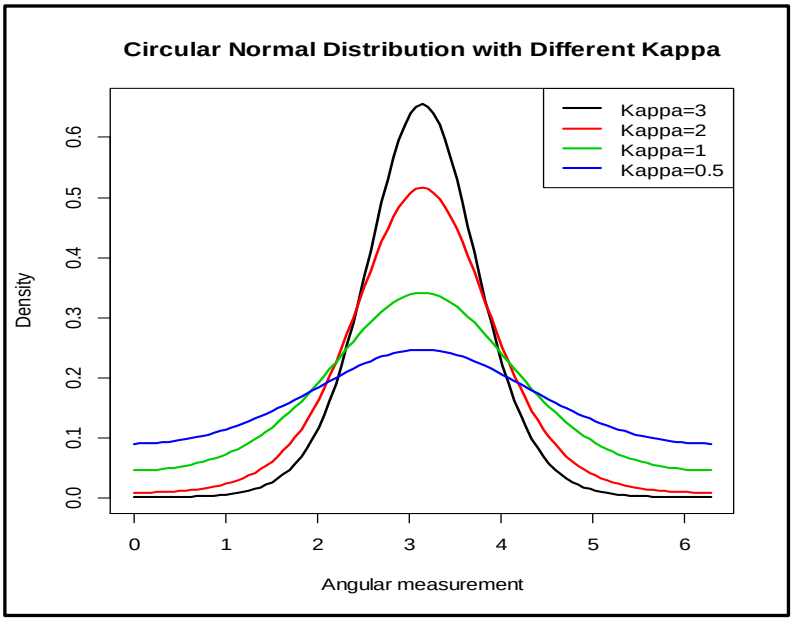


Figure 2.6 Circular normal distributions with varying concentration parameters

In the figure 2.5, circular variables are represented circularly with varying mean direction and circular concentration parameter (kappa). As it is shown in the figure when kappa decreases for same mean direction the plot, become flatter and flatter. This implies that the data show more variation since concertation parameter is the opposite of dispersion parameter, which is equivalent to variance in linear data. Circular data can be represented in linear plots as well as shown in figure 2.6.

In linear case, as bivariate is extension of distribution of variable in line into a plane of two dimensions, bivariate circular distribution is extension of univariate circular distribution in a complex plane into toroidal sphere (Guterman 2009).

The probability density function of two circular random variables (linear variables wrapped into circle) θ_1 and θ_2.

Bivariate and multivariate distributions are mostly expressed in Von Mises or normal distribution cases. This is due to the complexity of circular distributions. The importance of bivariate normal distributions in molecular science pinpointed by Mardia et al. (2008).

$$f(\theta_1,\theta_2)=\frac{\exp\{\kappa_1\cos(\theta_1-\mu_1)+\kappa_2\cos(\theta_2-\mu_2)+\lambda_{12}\sin(\theta_1-\mu_1)\sin(\theta_2-\mu_2)\}}{T(\kappa_1,\kappa_2\lambda_{12})} \quad (2.28)$$

For $-\pi \leq \theta_1$ and $\pi \leq \theta_2$

Where K1, $K_2 \geq 0$, $-\infty < \lambda_{12} < \infty$, $-\pi \leq \mu_1, \mu_2 < \pi$

The normalizing constant T (.) is given by

$$T(\kappa_1\kappa_2\lambda_{12})=4\pi\sum_{m=0}^{\infty}\binom{2m}{m}\left(\frac{\lambda}{2}\right)^{2m}\kappa_1^{-m}I_m(\kappa_1)\kappa_2^{-m}I_m(\kappa_2) \quad (2.29)$$

2.2.2.8 Axial Data

When variables are distributed on the circumference of a full circle, their probability is bound to (0-2π). On the other hand, there are some data sets spread over one-half of the circle and some other on the other half of the same circle. When these data do not have an identifiable orientation on a circle, it is difficult to consider this data as circular data sets. This kind of data sets is called axial data. Axial data sets are plenty in nature Arnold and Sengupta (2004) presented an example of "*propagations of crack or on fault in mining walls*". Using ordinary circular methods such as Von Mises mean direction methods to analyze such data would give a faulty result. Due to this fact, analysts recommend a method called angle doubling. The angle which is just multiplying the measured angles by two; treat them as a circular variable, and seek a circular distribution to approximate. The detail of axial data is beyond this paper and we recommend the book "Advances *in directional and linear statistics*" edited by Martin and Sengupta (2011) and contributed by varies experts in the field of statistics, physics, and mathematics for a festschrift of prominent statistician Rao Jammalamadaka.

2.2.3 Circular descriptive statistics

Circular descriptive statistics are different from a linear descriptive statistics due to the fact that in circular descriptive statistics one has to address dimensionality of the data in hand. As it was mentioned in the introductory chapter the data type in circular statistics are mainly compass measurements and time in a circular sense. When one works with time measurements data points have to be converted into degrees or radians, more sensibly into radians using expression 2.30.

$$adjustment = 2\pi \frac{y}{x} \tag{2.30}$$

Where y is time to be converted and x is a full cycle of a time.

2.2.3.1 Circular mean direction

Circular mean direction is calculated using sine and cosine components of the circular data. When grouped data involved as the case in malaria study of this paper individual sine/cosine components are multiplied by each frequencies. Malaria case are treated as grouped data and considered to be happened at the middle of each month. Expressions below used to calculate descriptive statistics of grouped circular data.

$$\overline{C}_n = \frac{1}{n}\sum_{i=1}^{k} f_i(\cos\theta_i) \tag{2.31}$$

$$\overline{S}_n = \frac{1}{n}\sum_{i=1}^{k} f_i(\sin\theta_i) \tag{2.32}$$

$$\overline{R} = \sqrt{\overline{C}^2 + \overline{S}^2} \tag{2.33}$$

Where $n = \sum f_i$

$$\cos(\overline{\theta}) = \frac{\overline{C}_n}{\overline{R}} \tag{2.34}$$

$$\sin(\overline{\theta}) = \frac{\overline{S}_n}{\overline{R}} \tag{2.35}$$

$$\overline{\theta} = \arctan\left(\frac{\sin(\overline{\theta})}{\cos(\overline{\theta})}\right) \tag{2.36}$$

Circular mean direction is quadrant dependent, the actual mean direction should be specified based on what quadrant sine and cosine components fall.

$$\begin{cases} if\ C_n > 0\ and\ \ S_n \geq 0; \boldsymbol{\mu_{dir}} = \mu_{cal} \\ \quad if\ C_n = 0\ \ and\ S_n > 0\ ; \boldsymbol{\mu_{dir}} = \dfrac{\pi}{2} \\ \quad if\ C_n < 0\ ; \boldsymbol{\mu_{dir}} = \mu_{cal} + \pi \\ if\ C_n \geq 0\ \ and\ \ S_n < 0\ ; \boldsymbol{\mu_{dir}} = \mu_{cal} + 2\pi \\ if\ C_n = 0\ and\ \ \ S_n = 0\ ; \boldsymbol{\mu_{dir}} = undefined \end{cases}$$

2.3.2 Circular dispersion measures

$\overline{R}$ is called the mean resultant length and it is not only used to specify mean direction of circular variable but also it can postulate the length of circular data vectors. It varies from zero to one. $N - \overline{R}$ is another related measure of variation in circular data. It can be as big as a number of observations N where all observations are line in a single line. In which case variance become zero. On the other hand $\overline{R}$ can be as small as zero where all observations are uniformly distributed on the circumference of a circle. $\overline{R}$ is a measure of dispersion which is analogues to variance in linear analysis. Circular variance is approximated using $\overline{R}$ by equation below

$$Var = 2(1 - \overline{R}) \tag{2.37}$$

The suffix 2 in expression 2.37 is to be able to get the small variation between circular variables otherwise, the variation would be unseen especially for data concentrated around the mean direction. Fisher (1995) presented circular standard deviation to be

$$\sigma = \sqrt{(-2\log \overline{R})} \tag{2.38}$$

For the data concentrated around the mean direction, circular standard deviation is calculated as in the expressions 2.31-2.36

$$\sigma = \sqrt{(2Var)} \tag{2.39}$$

Fisher (1995) also used

$$\hat{\sigma} = \frac{1-\hat{\rho}_2}{2\bar{R}^2} \tag{2.40}$$

As a measure of circular standard deviation where

$$\hat{\rho}_2 = \frac{1}{n}\sum_{i=1}^{n}\cos 2(\theta_i - \bar{\theta}) \tag{2.41}$$

to calculate confidence intervals.

$$\hat{\mu}(\theta) = \bar{\theta} \pm \sin^{-1}(Z_{\alpha/2}, \hat{\sigma}) \tag{2.42}$$

2.4 Circular Uniformity Test

In any circular analysis, testing whether circular data is distributed uniformly or not is the basic aspect. It helps to know randomness or "Isotopy" of the data in hand. As it is pinpointed in the introduction chapter, if a data is distributed uniformly on a circle it does not have preferred mean direction. Circular uniform distribution is a distribution with a maximum entropy (Rao 2001). In any circular estimation or inference, the first step should be knowing whether a data is distributed uniformly or not. Follmann and Proschan (1999) conducted circular uniformity test for correlated angular measurements. Borgan et al. (2002) made a data driven circular uniformity test.

Unlike linear data, uniformity test in circular data is very difficult property to verify by eyes even with more sophisticated plots. Due to this fact, to verify circular uniformity there is a need for statistical testing. Fortunately, there are few statistical tests to reject or accept the null hypotheses claims the data in hand is uniformly distributed or it follows some other distributions. Circular uniformity test has very practical application whether some phenomena happened uniformly in some specific period or not. For example, whether emergency administration of heart-related patients is uniform over 24 hour of the day; terrorist attacks follow some months of a year.

There are different tests used for uniformity of circular data. Some of the methods used to test whether circular variable follows uniform distribution or some other distribution. For instance. Rao (2001) tested circular uniformity against general wrapped stable distribution. There are some tests used to test simply whether circular variables follow circular distributions or not (Ajne 1968).

As it was pinpointed in the previous sections, the Statistical analysis in circular data is in its formative stages. A many of analysis methods and models in the topic needs extensive verification. Moreover, comparison of these analysis methods and models is crucial to get the best one out of them. Due to this fact, we presented and compared some tests used to test the uniformity of circular data based on applicability and tractability of these tests on biological data.

2.4.1 Rayleigh test

Rayleigh's test is based on the idea of rejecting the null hypothesis calming uniformity when sample mean direction is significantly far from zero and it is score test of uniformity within Von Mises distribution Matwiri et al. (2013).

Rayleigh uniformity test is a unimodal departure from uniformity (Pewsey et al. 2013). This test takes into account the two conditions about the mean directions whether it is specified or not.

When the mean direction is specified, the alternative Hypothesis has a different condition than when it is not specified.

When the mean direction is specified hypothesis testing of uniformity given as bellow.

$\boldsymbol{H_0}$: Angles are distributed uniformly on the circumference of a circle.

$\boldsymbol{H_1}$: Angles are distributed unimodally around a specified mean direction.

When the alternative hypothesis says "…unimodal around a specified mean" it means there is preferred direction but only with one mode. In reality, a circular distribution that deviates from uniformity is not only unimodal. The situation where circular distribution will be discussed later.

The test statistics in the condition where mean direction is specified is based on the measurement of concentration parameter $\overline{R}$ and given as follows

$$\overline{R}_0 = \overline{R}(\cos\overline{\theta} - \mu_0) \tag{2.43}$$

Calculation of the mean direction, resultant vector and quadrant specification is presented in equations 2.31-2.36.

The P-value also found as follows

Let $Z_0 = \overline{R}_0\sqrt{2N}$ and $P_z = \Pr(Z < Z_0)$

Then the p-value under the null hypothesis is calculated as follows

$$P = 1 - P_Z + \exp\left(\frac{-Z/2}{\sqrt{2\pi}}\left[\frac{\left(Z_0 - Z_0^3\right)}{16n} + \frac{15Z_0 + 305Z_0^3 - 125Z_0^5 + 9Z_0^7}{4608n^2}\right]\right) \tag{2.44}$$

Under the second condition where the mean direction is not specified the hypothesis testing of uniformity is as follows

Let $Z = N\overline{R}$ where $Z \sim (0,1)$

$$P = \begin{cases} \exp(-Z_0)\left[1 + \frac{\left(2Z - Z^2\right)}{4n} - \frac{\left(24Z + 32Z^2 + 76Z^3 - 9Z^4\right)}{288n^2}\right]; n < 50 \\ \exp(-Z);\ for.n \geq 50 \end{cases} \tag{2.45}$$

2.4.2 The χ^2 test

The chi-squared test can be also used to test the uniformity of circular variables uniformly distributed or not especially for the data points fairly large enough (>30). If our data is grouped data like in the case of malaria in months the groupings must not be less than four.

To calculate the expected values the formula bellow is used

$$\hat{f}_i = \frac{observations}{groups} \tag{2.46}$$

Then the Chi-squared test is calculated as follow

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(f_i - \hat{f}_i\right)}{\hat{f}_i} \tag{2.47}$$

2.4.3 Rao's spacing test

Rao's spacing uniformity test is considered as one of the most powerful tests when compared to other uniformity tests such as Rayleigh-test and Kuiper's v test on many aspects of uniformity tests for circular data (Russel and Levitin 1995, Rao 1972) Berens (2009) published a MATLAB toolbox for Rao's spacing test of uniformity test. The logic behind Rao's spacing test is if underlying distribution is uniform. Successive data points (observations) should be approximately evenly spaced about 360°/N apart over specified direction.

The test statistics for Rao's spacing test is U-statistics.

$$U = \frac{1}{2}\sum_{i=1}^{n} | T_i - \lambda | \tag{2.48}$$

And $\begin{cases} T_i = f_{i+1} - f_i \ \ for\ 1 \le i \le n-1 \\ T_n = (360° - f_n) + f_1 \ for\ i = n \end{cases}$

Since the sum of the positive deviation must be equal to the sum of negative deviation. The absolute value will be eliminated hence the test will have simpler expression form as follows.

$$U = \sum_{i=1}^{n} (T_i - \lambda) \tag{2.49}$$

2.5 Correlation and Regression Involving Circular Measurements

2.5.1 Circular correlation

Correlation statistics involving circular measurements are of two type. The first type is when one of the claimed correlated variables is circular and the other is linear. The second type of circular correlation is where both variables are of a circular type. Sometimes measuring circular correlation (association) is explained as circular regression (may be wrongly). For example Kemper et al. (2012) used the term association when they actually did circular regression analysis.

2.5.1.1 Circular linear (circular-linear) correlation

The Mardia's concept of the linear-circular correlation coefficient is presented in this chapter as follows and the name is given in Honor of K.V.Mardia (born 1935).

A measure of association between two random variables X and Θ; where X is defined (-∞, +∞) and Θ is defined (0,2π), the population correlation coefficient is given by expression 2.50 below.

$$\rho_{X\theta}^2 = \frac{4(V_c^2 E_{Xs}^2 + V_S^2 E_{XC}^2 - \beta_2 E_{XC} E_{Xs})}{\sigma^2 (4V_C^2 V_S^2 - \beta_2^2)} \tag{2.50}$$

Where

$$E(X)=0;\quad E(\sin\theta)=0;\qquad E\left(X\sin\theta\right)=E_{XS};\ \ E\left(X\cos\theta\right)=E_{XC}$$

$$Var\left(\cos\theta\right)=V_C;\quad Var\left(\sin\theta\right)=V_S;\qquad Var(X)=\sigma;\ \beta_2=E(\sin 2\theta)$$

The above expression can be used when one of the variables is linear and the other is circular. Unlike regression analysis involving circular data, it does not distinguish which variable is circular and which one is linear.

For a sample correlation coefficients Expression (2.51) is commonly used

$$r^2=\frac{r_{XC}^2+r_{XS}^2-2r_{XC}r_{XS}r_{CS}}{1-r_{CS}^2} \tag{2. 51}$$

Where

$$r_{XC}=corr\left(linear,\cos\left(\theta\right)\right);\quad r_{XC}=corr\left(linear,\sin\left(\theta\right)\right);\quad r_{XC}=corr\left(\sin(\theta),\cos\left(\theta\right)\right)$$

2.5.1.2 Circular-circular correlation

As it was mentioned in the methodology chapter, there are few correlation measures presented by different statisticians (Rao 2001, Pauen and Ivanova 2013, Kempter et al. 2012) for data points involving circular data but these measures are a bit sophisticated and need complex mathematical trigonometric knowledge in advanced level. Out of them, the one easily understandable is a model used by NCSS (Statistical Software) referring Rao and SenGupta (2001).

$$r_c=\frac{\sum_{k=1}^{n}\sin\left(\theta_i-\bar{\theta}\right)\sin\left(\phi_i-\bar{\phi}\right)}{\sqrt{\sum_{k=1}^{n}\sin^2\left(\theta_i-\bar{\theta}\right)\sum_{k=1}^{n}\sin^2\left(\phi_i-\bar{\phi}\right)}} \tag{2.52}$$

This method is used to calculate correlation between heart attack date and birthdate data.

Significance of this correlation can be calculated using the expression as follows

$$Z_r = r_C \sqrt{\frac{n\lambda_{2,0}\lambda_{0,2}}{\lambda_{2,2}}} \tag{2.53}$$

Where

$$\lambda_{ij} = \frac{1}{n}\sum \sin^i\left(\theta_i - \bar{\theta}\right)\sin j\left(\phi_i - \bar{\phi}\right)$$

2.5.2 Regression analysis involving circular data

Regression analysis that involves circular data is quite different from other forms of regression analyses in many aspects of the regression methods and processes. There are three basic regression types in circular data. Based on what parts of the data is measured in a circular. These three regression types are summarized in chart 2.4.

Chart 2.4 Regression types involving circular data

Regression-type	Response variable	Explanatory variable
Linear-Circular (LC)	Linear	Circular
Circular-Linear (CL)	Circular	Linear
Circular-Circular(CC)	Circular	Circular

Linear-circular regression: Regression analysis when the explanatory variable is (are) measured in linear and the response variable is circular form the regression type in this case normally referred as **Linear-Circular** regression. There is misunderstanding generally when one of the regression components is circular and the other is Linear. For

example, in some references, there is a term **Circular-Linear** where explanatory variable is (are) linear and the response variable is circular, and in some other references, there are terms **Linear-circular** when the explanatory variable(s) is (are) linear and the response variable is circular (Rao 2001). In some other texts, it simply referred as **Circular regression** without specifying the regression components. In our perception, it would give better understanding if the regression nomenclature were based on the response variable because the naming of the explanatory variable (s) will be easier even if they are more than one. In next, few sub chapters each regression type is examined using appropriate data sets.

2.5.2.1 Linear-circular regression

Regression analysis where the response variable is linear (magnitude) and the explanatory variable is circular is called **Linear-circular regression analysis**. In this analysis type, observations enter into the model as a trigonometric polynomial. Step by step analysis process is presented in this subchapter using malaria data in India that with a little modification. There are few attempts for prediction of linear response variable (magnitude) from circular explanatory variables (direction or time in circular sense) (Kamper et al. 2012). Hussin (2006) did a hypothesis testing for linear-circular regression analysis.

These models lack clarity in application and interpretation. Therefore, we used a malaria data explained in the material section of this paper. The good thing about that data is it can be used as frequency data and bivariate data.

Malaria case frequencies are taken as a dependent variable and time of the year measured in radian as explanatory variable.

As it is proven in section 2.5.1.1 after being confident Malaria case and months of the year in radian have circular-linear association we proceed into sinusoidal regression

analysis using least square method and Fourier analysis (Attinger 1996). Our equation of regression looks like bellow.

$$y_j = A_0 + A_1\cos(\omega\theta - \varphi) + A_2\cos(2\omega\theta - \varphi) + ... + A_p\cos(p\omega\theta - \varphi) + B_1\sin(\omega\theta - \varphi) + B_2\sin(2\omega\theta - \varphi) + ... + B_p\sin(p\omega\theta - \varphi) + \varepsilon_j \quad (2.54)$$

Where

θ is the independent angle and P is degree(order) of polynomial

A_0 is a mean level.

ω is angular frequency.

$A_1, A_2, ... A_p$ and $B_1, B_2, ... B_p$ are coefficients to be predicted from the model.

φ is the peak of angular time where the frequency is highest. It is called acrophase.

The above equation is the Fourier series equation

We used ordinary least square method

$$\begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & \cos(\omega\theta_1 - \varphi_1) & \cos(2\omega\theta_1 - \varphi_2) & ... & \cos(p\omega\theta_1 - \varphi_p) & \sin(p\omega\theta_1 - \varphi_1) & \sin(p\omega\theta_1 - \varphi_2) & ... & \sin(p\omega\theta_1 - \varphi_p) \\ 1 & \cos(\omega\theta_1 - \varphi_1) & \cos(2\omega\theta_1 - \varphi_2) & ... & \cos(p\omega\theta_1 - \varphi_p) & \sin(p\omega\theta_{21} - \varphi_1) & \cos(p\omega\theta_2 - \varphi_2) & ... & \sin(p\omega\theta_2 - \varphi_p) \\ . & . & . & ... & . & . & . & ... & . \\ . & . & . & ... & . & . & . & .... & . \\ 1 & \cos(\omega\theta_1 - \varphi_1) & \cos(2\omega\theta_1 - \varphi_2) & ... & \cos(p\omega\theta_1 - \varphi_p) & \sin(p\omega\theta_1 - \varphi_1) & \sin(p\omega\theta_1 - \varphi_2) & ... & \sin(p\omega\theta_1 - \varphi_p) \end{bmatrix} * \begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ \\ A_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

One can reach to the intended result using the above regression expression.

The biggest challenge in circular regression analysis is a determination of degrees of polynomials (Rao 2001). One way to tackle this challenge is to use an iterative method. When used software based circular regression analysis running the regression with one more degree of polynomial and see if regression coefficient is significant at that degree of polynomial. If the regression at this level is also significant run the regression until it is not significant anymore. We use excel regression analysis method for Fourier regression equation above.

2.5.2.2 Circular-linear regression

Circular-linear regression is a regression type whereby direction is predicted from magnitudes.

The regression equation is given in expression 2.55.

$$\mu(\theta) = \mu_0 + \sum \beta_i x_j \,(\mathrm{mod}\, 2\pi) \tag{2.55}$$

With a link function $g(x) = \tan^{-1}(\mathrm{sgn}(x)\,|x|^{\lambda})$, the parameter λ can be estimated from the data with analogous to estimation of Box-cox transformation as described by Fisher and lee (1992). This method is demonstrated on the CCHF data in Turkey.

As Rao (2001) specified the conditional distribution of the dependent angular measurement θ given the linear independent explanatory variable X is given by CN(μ+2πF(x),k). This method allows a direct estimation of μ and the concentration parameter k by the method of MLE.

2.5.2.3 Circular-circular regression

Circular-Circular regression is a type of regression whereby both the dependent variable and explanatory variable are arisen from circular data. These data having joint Probability density function $f(\theta.\phi).0 < \theta.\phi \leq 2\pi$. Out of regression analysis involving circular data the most complicated one is circular-circular regression. Even if there are few attempts to find a model for this kind of regression they still lacks robustness. For instance Kato's (2008) regression model as a Mobius transformation needs more understanding of dimensionality and physics logic.

For this regression type simulated circular data is used based on heart related studies in Turkey (Onat et al. 1993, Onat 2001, Şurdumavcı et al. 1991).

To predict ϕ (attackdate) from a given θ (birthdate) consider a conditional expectation of $(e^{i\phi}|\theta)$ vector.

$$E\left(e^{-i\phi} \mid \theta\right) = \rho(\theta)e^{i\mu(\theta)} = g_1(\theta) + ig_2(\theta) \tag{2.56}$$

$\mu(\theta)$ represents the conditional mean direction of Φ for a given θ and $0 \leq \rho(\theta) \leq 2\pi$ is the conditional concentration parameter towards this direction (Rao.2001).

Since $g_1(\theta)$ and $g_2(\theta)$ are circular with a period 2π they can be approximated with their Fourier series as follows with an appropriate degree of trigonometric polynomials (order of polynomial)

$$\begin{aligned} g_1 &= \sum_{k=0}^{m} \left(\alpha_k \cos k\theta + \beta_k \sin k\theta\right) \\ g_2 &= \sum_{k=0}^{m} \left(\delta_k \cos k\theta + \gamma_k \sin k\theta\right) \end{aligned} \tag{2.57}$$

The general linear model form from expression 2.57,

$$\begin{aligned} \cos\phi &= \sum_{k=0}^{m} \left(\alpha_k \cos k\theta + \beta_k \sin k\theta\right) + \varepsilon \\ \sin\phi &= \sum_{k=0}^{m} \left(\delta_k \cos k\theta + \gamma_k \sin k\theta\right) + \varepsilon \end{aligned} \tag{2.58}$$

Since the explanatory variable θ is taken from the uniform distribution, roots of Ordinary Fourier series find the least square estimate of regression coefficient as follows,

$$\hat{\alpha}_0 = \frac{1}{n}\sum_{i=1}^{n} \cos\phi_i$$

$$\hat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} \cos\phi_i \cos\, j\theta_i$$

$$\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n}\cos\phi_i \sin\, j\theta_i$$

$$\hat{\delta}_0 = \frac{1}{n}\sum_{i=1}^{n}\sin\phi_i$$

$$\hat{\delta}_j = \frac{1}{n}\sum_{i=1}^{n}\sin\phi_i \cos\, j\theta_i$$

$$\hat{\gamma}_j = \frac{1}{n}\sum_{i=1}^{n}\sin\phi_i \sin\, j\theta_i$$

The Dependent components and the design matrix structure of the Circular-Circular regression analysis is look likes bellow as specified in Rao (2001).

$$\begin{aligned} Y_{1i} &= \cos\phi_i ; i = 1,2,...,n \\ Y_{2i} &= \sin\phi_i ; i = 1,2,...,n \end{aligned} \tag{2.59}$$

$$\begin{aligned} Y^{(1)} &= (Y_{11}, Y_{12},..., Y_{1n})' \\ Y^{(2)} &= (Y_{21}, Y_{22},..., Y_{2n}) \end{aligned} \tag{2.60}$$

Error terms

$$\begin{aligned} \varepsilon^{(1)} &= (\varepsilon_{11}, \varepsilon_{12},..., \varepsilon_{1n})' \\ \varepsilon^{(2)} &= (\varepsilon_{21}, \varepsilon_{22},..., \varepsilon_{2n})' \end{aligned} \tag{2.61}$$

The design matrix

$$\underset{nx(2m+1)}{X} = \begin{bmatrix} 1 & \cos\theta_1 & \cos 2\theta_1 & ... & \cos m\theta_1 & \sin\theta_1 & \sin 2\theta_1 & ... & \sin m\theta_1 \\ 1 & \cos\theta_2 & \cos 2\theta_2 & ... & \sin\theta_2 & \sin\theta_2 & \sin 2\theta_2 & ... & \sin m\theta_2 \\ . & . & . & ... & . & . & . & ... & . \\ . & . & . & ... & . & . & . & ... & . \\ . & . & . & ... & . & . & . & ... & . \\ 1 & \cos\theta_n & \cos 2\theta_n & ... & \cos m\theta_n & \sin\theta_n & \sin 2\theta_n & .... & \sin m\theta_n \end{bmatrix} \tag{2.62}$$

Finally the parameters to be estimated from the model

$$\underset{\sim}{\lambda}^{(1)} = (\alpha_0, \alpha_1, ..., \alpha_m, \beta_1, \beta_2, ..._1, \beta_m)'$$
$$\underset{\sim}{\lambda}^{(2)} = (\gamma_0, \gamma_1, ..., \gamma_m, \delta_1, \delta_2, ..._1, \delta_m) \quad (2.63)$$

Together as regression equation form.

$$\begin{bmatrix} Y^{*(1)} \\ Y^{*(2)} \end{bmatrix} = \begin{bmatrix} \lambda^{(1)} \\ \lambda^{(2)} \end{bmatrix} * \underset{nx(2m+1)}{X} + \begin{bmatrix} \varepsilon^{(1)} \\ \varepsilon^{(2)} \end{bmatrix} \quad (2.64)$$

As it was mentioned in the Linear-circular regression chapter, the story of determination of degrees of polynomials is also difficult. All trigonometric-based analysis shares the same difficulty on the issue of determination of degree of polynomial where the regression level is "best" significant. If computer programs used for the regression analysis purpose, running the regression with one more degree of polynomial and see the effect of the added polynomial level just as the same manner with forward parameter inclusion methods in other forms of regressions.

The problem arises when one wants to preside with step by step determination of the degree of polynomial. Rao (2001) used a method called augmenting the design matrix X on the residual sum of squares. The logic behind augmenting these matrices is used (m+1) degree of polynomial until the reduction of error sum of square is no significant anymore. The other method Rao (2001) recommend is to assess the proportional reduction in the mean square error where $x_1.x_2.x_3.\ldots\ldots\ldots.x_p$ is used to predict Y instead of $x_1.x_2.x_3.\ldots\ldots\ldots.x_k$ where p>k. This logic is used in linear Predictors where a partial correlation ratio is taken in account.

Using the expression bellow one can approximately estimate the proportional reduction in regression sum of square.

$$\rho_i^{*2} = \frac{\rho *_{i0}^2 (m+1) - \rho_{i0}^{*2}(m)}{1 - \rho_{i0}^{*2}(m)} \quad (2.65)$$

Where $\rho^2(m+1)$ is a squared multiple correlation coefficient where $(m+1)^{th}$ degrees of polynomial is included and $\rho^2(m)$ is a squared multiple correlation coefficient where $(m+1)^{th}$ degree of polynomial is not included. If there is "significant" reduction in the regression sum of square where $(m+1)^{th}$ degree of polynomial is included, the regression should have that order of polynomial as well.

Setting up the explanatory variables for different order of polynomial for the sake of simplicity use Euler formula with Binomial theorem as follows

$$\begin{aligned}
&\sin(n\theta) = \frac{e^{in\theta} - e^{-in\theta}}{2i} \\
&= \frac{(\cos\theta + i\sin\theta)^n - (\cos\theta - i\sin\theta)^{n-k}}{2i} \\
&= \sum_{k=0}^{n}\binom{n}{k}\frac{\cos^k\theta(i\sin\theta)^{n-k} - \cos^k\theta(-i\sin\theta)^{n-k}}{2i} \\
&\sum_{k=0}^{n}\binom{n}{k}\cos^k\theta\sin^{n-k}\theta\frac{i^{n-k} - (-i)^{n-k}}{2i} \\
&\sum_{k=0}^{n}\binom{n}{k}\cos^k\theta\sin^{n-k}\theta\sin\left[\frac{1}{2}(n-k)\pi\right]
\end{aligned} \tag{2.66}$$

Based on this calculation first few multiple of sin functions of any angle.

$$\begin{aligned}
&\sin(2\theta) = 2\sin\theta\cos\theta \\
&\sin(3\theta) = 3\cos^2\theta\sin\theta - \sin^3\theta \\
&\sin(4\theta) = 4\cos^3\theta\sin\theta - 4\cos\theta\sin^3\theta \\
&\sin(5\theta) = 5\cos^4\theta\sin\theta - 10\cos^2\theta\sin^3\theta + \sin^5\theta
\end{aligned} \tag{2.67}$$

Similarly the cosine components of varying order of polynomials

$$
\begin{aligned}
\cos(n\theta) &= \frac{e^{in\theta} - e^{-in\theta}}{2} \\
&= \frac{e^{(i\theta)n} - e^{(-i\theta)n}}{2} \\
&= \frac{(\cos\theta + i\sin\theta)^n - (\cos\theta - i\sin\theta)^{n-k}}{2} \\
&= \sum_{k=0}^{n}\binom{n}{k}\frac{\cos^k\theta(i\sin\theta)^{n-k} + \cos^k\theta(-i\sin\theta)^{n-k}}{2} \\
&\sum_{k=0}^{n}\binom{n}{k}\cos^k\theta\sin^{n-k}\theta\frac{i^{n-k} + (-i)^{n-k}}{2} \\
&\sum_{k=0}^{n}\binom{n}{k}\cos^k\theta\sin^{n-k}\theta\cos\left[\frac{1}{2}(n-k)\pi\right]
\end{aligned}
\tag{2.68}
$$

Based on the above calculations first 5 order of polynomial multiple of cosine functions are as follows in

$$
\begin{aligned}
\cos(2\theta) &= \cos^2\theta + \sin^2\theta \\
\cos(3\theta) &= \cos^3\theta - 3\cos\theta\sin^2\theta \\
\cos(4\theta) &= \cos^4\theta - 6\cos^2\theta + 4\sin 4\theta \\
\cos(5\theta) &= \cos^5\theta\sin\theta - 10\cos^3\theta\sin^2\theta + 5\cos\theta\sin 4\theta
\end{aligned}
\tag{2.69}
$$

The calculation of partial correlation coefficients is needed for these 5 order of polynomials.

For the second order of polynomial

$$
\rho_2^{*2} = \frac{\rho^{*2}_{i0}(1,2) - \rho_{i0}^{*2}(1)}{1 - \rho_{i0}^{*2}(1)} \tag{2.70}
$$

For Third order of polynomial

$$
\rho_3^{*2} = \frac{\rho^{*2}_{i0}(1,2,3) - \rho_{i0}^{*2}(1,2)}{1 - \rho_{i0}^{*2}(1,2)} \tag{2.71}
$$

For fourth order of polynomial

$$\rho_4^{*2} = \frac{\rho *_{i0}^2 (1,2,3,4) - \rho_{i0}^{*2}(1,2,3)}{1 - \rho_{i0}^{*2}(1,2,3)} \tag{2.72}$$

For Fifth order of polynomial

$$\rho_5^{*2} = \frac{\rho *_{i0}^2 (1,2,3,4,5) - \rho_{i0}^{*2}(1,2,3,4)}{1 - \rho_{i0}^{*2}(1,2,3,4)} \tag{2.73}$$

3. RESULT AND DISCUSSION

3.1 Data plots

Malaria data presented in chart 2.1 of section 2.1.1 is used for circular plot, calculating descriptive statistics and linear-circular regression analysis. Before continuing to calculating descriptive statistics of circular data, the logic behind malaria data and why it is important to be seen in great attention is presented in this chapter in detailed.

Vector-borne disease are infectious disease transmitted by mostly bloodsucking insects (vectors). These vectors ingest disease-producing microorganisms during blood feed from infected host (human or human) and they transmit these microorganisms in to new host (human or animal) during next blood feed (http:// apps.who.Anonymous 2014a). There has been numerous researches conducted on the issue for ages. It has been one of the biggest challenges of human history and we have been implementing our ultima knowledge to tackle the problem, some time we win some times the problem beat us by far. According to WHO report (apps.who.int/, Anonymous 2014b) Vector-borne disease accounted for 17% of the estimated global burden of all infectious diseases. In our days, there are numerous international and national organizations including WHO releasing big funds on the researches to the issue. There are Daily, weekly, monthly and yearly reports of vector born disease in WHO websites.

Vector- borne diseases are mostly tend to be influenced by geographical and climatic Dynamics of our planet earth. Andrew et al. (2000) pinpointed that inter-annual and inter-decadal climatic variation have a direct influence on the epidemiology of vector borne diseases.

Among vector-borne disease and infections, there is no other disease that challenged our existence as malaria. Malaria is caused by the protozoan parasite Plasmodium. Human malaria is caused by four different species of Plasmodium: *P. falciparum, P. malariae, P. ovale and P. vivax* (WHO factsheet 2014). There are currently over 100 countries and

territories where there is a risk of malaria transmission, and more than 125 million international travelers visit these every year.

As one of vector-borne disease, malaria is also influenced by geographical and climatic dynamics. The map above is clearly shows that tropical areas are more affected than those cooler environments on the globe. It is not only geographical features that influence the incidence of malaria but also climatic and weather factors play big (if not the biggest) role for the incidence and transmission of malaria (Grassly and Fraser 2006, Stuckey et al. 2014, Cairn et al. 2015).

Visualization of the data says too many things about the data type and distribution forms. It also gives some important clue about what kind of analysis methods to follow and what types of results to expect. Due to this fact, it is customary to use a graphical representation of the data in hand.

In this section, the relationship of malaria data in Visakhapatnam district over time of the year plotted and examined if there is a tangible relationship between time of the years(months) and malaria distribution.

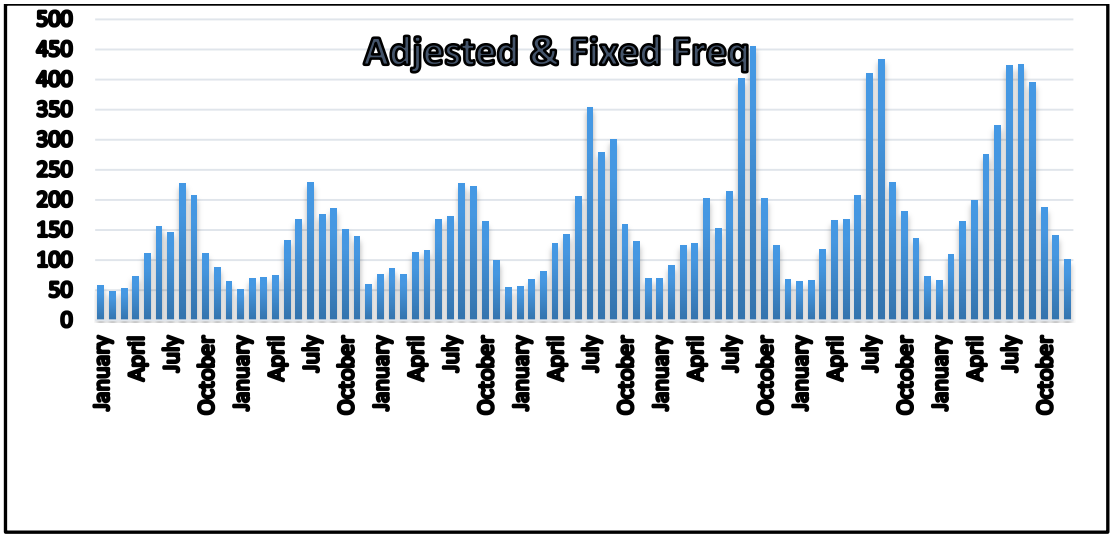


Figure 3.1 Time series plot of Adjusted and fixed frequencies of Plasmodiu Falciparum

Figure 3.1 indicates monthly malaria cases of the study area in 7 years from 2005-2011. In the plot, it is clear that malaria cases show some trend that means malaria is not uniformly distributed throughout the year in each year. There is also malaria case increment from year 2005 to 2011 roughly but it is not focus in this paper. Figure 3.1 also helps to specify whether there is multimodality in the data. As one can see from the figure, there is only one peak and one trough in a cycle (from January to next January) roughly. This indication intern helps the data are suitable for circular analysis because multimodality affects methods to use in circular analysis.

In circular statistical analyses, increment in years does not have importance because circularity is taken in months. This is to mean, data points in 2006 or in 2007 considered to be as the same frequency as long as they are happened in the same month, say March. To brief this idea malaria cases in April of 2008 and malaria cases in April of 2009 are logically considered as in one group even though there are 11 months' time gaps between them. Mathematically, cases that happened in same months are pooled together for calculation and analysis in circular methods. This is one of the main points that distinguishes circular analysis from time series statistics.

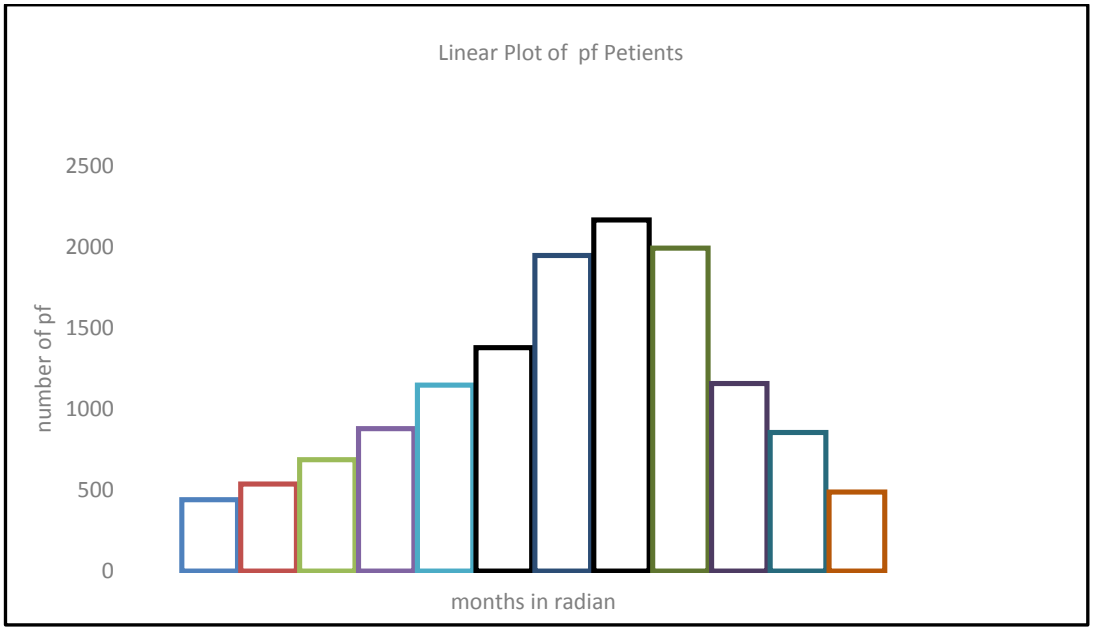


Figure 3.2 Time series plot of Adjusted and fixed frequencies (pooled)

Figure 3.2 displays pooled plots of the malaria in months that is the linear representation of circular data. In the figure, it looks the distribution is a bit skewed to the left. If the data were analyzed using linear methods skewness of it should be taken in account. In circular figures, the skewness of any distribution is based on the choice of starting point of any measurement. If February ($1/6\pi$ rad) were the choice of starting point, the January distribution would shift to the end and the distribution would look" more" normal. One of the advantages of circular data management is the choice of starting point, which is wholly arbitrary, so one can have a circular normal distribution opposite to when he manage linear data.

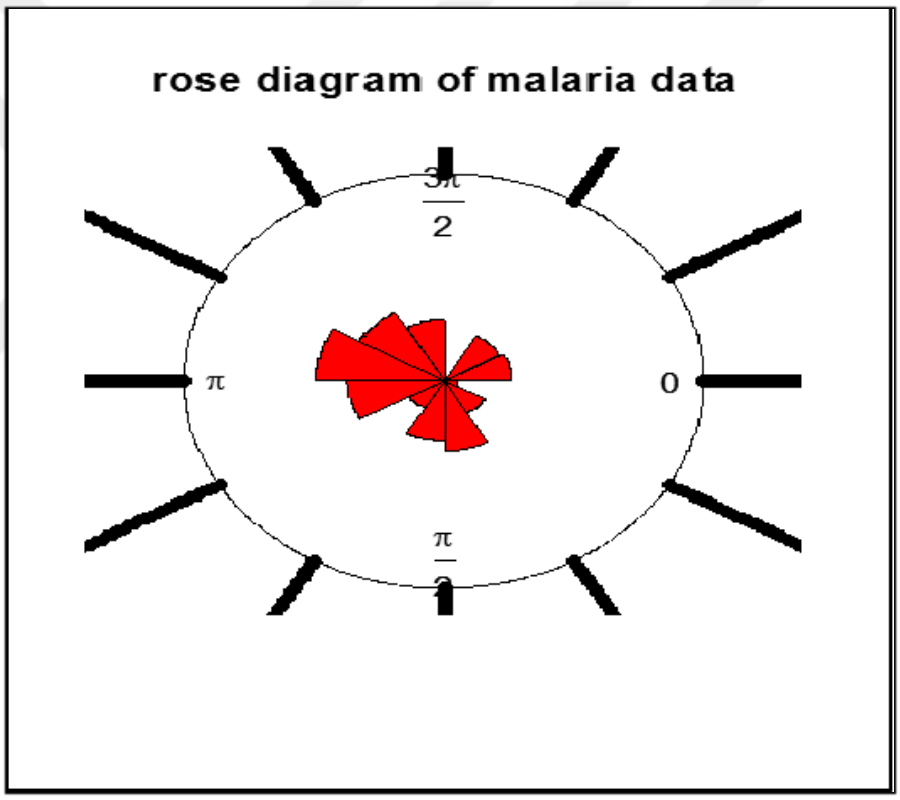


Figure 3.3 Circular plot (rose diagram) of data

There are few data plots for circular representation of circular data. Basically, there are three data plots for circular representation of circular data. One is a circular **dot plot** by which data points are plotted as a dot on the circumference of a circle as shown in figure 2.4a. The other is **circular histogram** by which frequencies are plotted on the circumference of a circle as a histogram. Out of these data visualizations of circular

data, the **rose diagram** is the most used and sophisticated one. Rose diagram is mostly used for the wind rose data because wind direction data is used frequently in weather forecasting and aviation industries. Rose diagram when applied to wind data is called "rose wind" to represent wind speed and directions daily, monthly, yearly, or seasonally. Circular data other than wind direction and speed can be successfully plotted with rose diagram.

In figure 3.3 circular plot of malaria occurrence in the study area in months is plotted as a rose diagram. The plot has 12 partitions (bins) to indicate data is plotted in 12 groupings or frequencies. Basically, any grouping (frequencies) are possible. In this data plots, different frequencies have different bin size indicating there are differences in the occurrence of malaria monthly. The smallest malaria occurrence can be seen in 10^{th} bin starting from January and counting clockwise direction. If there were no difference in malaria cases in months, bins would have similar size and show smooth sub circular shaded area within a circle. In the above rose diagram, there is also dot plots on the circumference of the circle. Dot plots in this data lack clarity due to extremely large sample is plotted.

3.2 Circular descriptive statistics

Since the *Plasmodium Falciparum,* data is a frequency data type (grouped data), calculating descriptive statistics should take into account this condition. The first step is to convert months into radian and adjust into the way one day represents 1 degree (0.017453 rad). Since months in a year do not have equal days, one has to adjust each month in the way to contribute equal frequencies. To do so, 365 days changed in a year to 360°.

First, 30 is divided by a number of days in each month we have the rate of each month's contribution to 360°.

$$adjustment = \frac{30}{x} \tag{3.1}$$

Where x in days of each month

Adjustment result is multiplied by the frequency of each month. But results found in this way does not have equal frequencies to the unadjusted frequencies due to round off error. To retain the total frequency, adjusted frequencies are summed and divide by the sum of frequencies before adjustment. Then, the result is multiplied by the adjusted frequencies. The resulting frequencies are shown in chart 3.1.

Chart 3.1 Frequency adjustment

Index	Months	Days	Adjustment	Month	Degree	**Radian**	**Pf**
1	January	31	0.967741935	1	1	**0.017453293**	**57**
2	February	28	1.071428571	2	30	**0.523598776**	**44**
3	March	31	0.967741935	3	60	**1.047197551**	**53**
4	April	30	1.000000000	4	90	**1.570796327**	**72**
5	May	31	0.967741935	5	120	**2.094395102**	**112**
6	June	30	1.000000000	6	150	**2.617993878**	**154**
7	July	31	0.967741935	7	180	**3.141592654**	**149**
8	August	31	0.967741935	8	210	**3.665191429**	**231**
9	September	30	1.000000000	9	240	**4.188790205**	**204**
10	October	31	0.967741935	10	270	**4.71238898**	**113**
	…	…	…	…	…	**…**	
	…	…	…	….	…	**….**	
79	July	31	0.967741935	6	180	**3.141593**	**430**
80	August	31	0.967741935	7	210	**3.665191**	**431**
81	September	30	1.000000000	8	240	**4.188790**	**388**
82	October	31	0.967741935	9	270	**4.712389**	**191**
83	November	30	1.00000000	10	300	**5.235988**	**138**
84	December	31	0.967741935	11	330	**5.759587**	**102**

From the chart 3.1, the seventh column is used which is radian measure of days, and ninth column that is adjusted and fixed frequencies for calculation of circular descriptive statistics.

3.2.1 Circular mean direction

Malaria case are treated as grouped data and considered to be happened in the middle of each month. Expressions 2.31-2.36 are used to calculate the mean direction of *plasmodium falciparum* data presented in chart 2.1 in the materials and methods chapter and the data frequency adjustment is presented in chart 3.1. For the sake of simplicity trigonometric components of each data, level is presented in chart 3.2.

Chart 3.2 Trigonometric components decomposition of time series data

Index	Radian (θ)	freq(f_i)	Cos(θ)	Sin(θ)	f_iCos(θ)	$f_i$sin(θ)
1	0.017453	57	0.9998	0.0175	56.9913	0.9948
2	0.523599	48	0.8660	0.5000	41.5196	23.9713
3	1.047198	52	0.5000	0.8660	26.0802	45.1723
4	1.570796	73	0.0000	1.0000	0.0000	73.2215
5	2.094395	110	-0.5000	0.8660	-55.1130	95.4585
6	2.617994	157	-0.8660	0.5000	-135.6306	78.3063
7	3.141593	147	-1.0000	0.0000	-146.6399	0.0000
	…	…	…	…	…	…
	…	…	…	…	…	…
	…		…	…	…	…
80	3.665191	424	-0.8660	-0.5000	-367.3446	-212.0865
81	4.18879	395	-0.5000	-0.8660	-197.2913	-341.7186
82	4.712389	188	0.0000	-1.0000	0.0000	-187.9746
83	5.235988	140	0.5000	-0.8660	70.1706	-121.5391
84	5.759587	100	0.8660	-0.5000	86.9354	-50.1922

$$n = \sum f_i = 13689$$

$$\overline{C}_n = \frac{1}{n}\sum_{i=1}^{k} f_i(\cos\theta_i) = \frac{1}{13689}\left(57(0.9998) + 48(0.8660) + ... + 100(0.8660)\right) = \text{-}0.328$$

$$\overline{C}_n = \frac{1}{n}\sum_{i=1}^{k} f_i(\sin\theta_i) = \frac{1}{13689}\left(57(0.0175) + 48(0.500) + ... + 100(-0.5000)\right) = \text{-}0.111$$

$$\overline{R} = \sqrt{\overline{C}^2 + \overline{S}^2} = \sqrt{(-0.328)^2 + (-0.111)^2} = 0.3465$$

Trigonometric mean directions with respect to mean resultant vector R is needed. To do so respective mean directions with mean resultant vector are divided as follows,

$$\cos(\overline{\theta}) = \frac{\overline{C}_n}{\overline{R}} = -\frac{0.328}{0.3465} = -0.94726$$

$$\sin(\overline{\theta}) = \frac{\overline{S}_n}{\overline{R}} - \frac{0.111}{0.3465} = -0.32046$$

$$\overline{\theta} = \arctan\left(\frac{\sin(\overline{\theta})}{\cos(\overline{\theta})}\right) = a\tan\left(\frac{-0.94726}{0.32046}\right) = 0.326216$$

Using the quadrant specificity property of mean direction presented in section 2.3.1, since summation of both sine and cosine are negative we add π on the calculated mean value so that the resulting mean direction is **0.326216085+3.141593=3.467809**.

The mean direction calculated in this way, considering a number of malaria cases as frequencies and taking as if every case happened in the middle of each month, has a very slight difference with when one calculate mean direction precisely from each data individually. However, the true mean resultant length appears to be smaller. Therefore, it needs correction using an equal spacing of each data point. For monthly circular data sets with equal spacing and adjusting of months to have 30 days.

Equation 3.2 gives sensible corrections expression.

$$\bar{R}_C = C\bar{R}_r \tag{3.2}$$

Where $\bar{R}_C$ is a corrected mean resultant length,

$C = \frac{\pi/M}{\sin(\pi/M)}$ is a correction factor and M is number of months in a year,

$\bar{R}_r$ is raw resultant vector before correction.

For the malaria data, where raw mean resultant length is calculated as **0.346520512** before correction. The corrected mean resultant length is **0.350511**, which is always greater than uncorrected resultant length because of the correction factor in this case **1.011515** radians in magnitude. Pewsey. et al. (2013. pp 31) presented effects of correction factors on resultant mean lengths.

The quadrant-specific mean direction is found to be 3.467809 radian (198.6908°) that Correspond to at the end of July.

3.2.2 Circular dispersion measures

The mean resultant length 0.3465 (0.351 after adjustment) is also used to specify the dispersion of *Plasmodium falciparum* in the study area. If all malaria cases were occurred in one month the resultant vector would have been 1. Similarly, if malaria case were occurred uniformly throughout the year the mean resultant vector would have been zero. For this data though, the calculated mean resultant vector is neither one nor zero. This indicates there is seasonality in the occurrence of *Plasmodium falciparum* malaria in the study area.

$N - \bar{R}$ (13688.572) is another related measure. If this data were uniformly distributed $N - \bar{R}$ would be equal to the sample size N. on the other hand if the data were

distributed uniformly $N - \overline{R}$ would be zero. This method works better for relatively small sample size.

The circular Variance of *plasmodium falciparum* malaria data is calculated as follows

$$Var = 2(1-0.351) = 1.298$$

As it was explained in the methodology section circular standard deviation is calculated using slightly different method than linear standard deviation whereby it is just square root of variance. Circular standard deviation is square root of negative logarithm of mean resultant vector. This expression sometimes referred as Fisher's standard deviation.

$$\sigma = \sqrt{\left(-2\log \overline{R}\right)} = 0.6747$$

As one can see from mean resultant vector the data is not concentrated around the mean direction. Therefore, calculation of standard deviation is better when used expression 2.38 rather than 2.39. When one uses the Fisher's expression of the circular standard deviation would be **0.959449**, which is less than the Rao (2001) standard deviation.

Angular mean direction would satisfy the notion of $\sum_{i=1}^{n} \sin(\theta_i - \overline{\theta})$, but since this data is grouped data type it need not satisfy the notion.

3.3 Circular Uniformity Test

As it is already clarified in section 2.4 of this paper, circular uniformity test is one of very important subject in circular data analysis because circular uniform distribution is central to circular analyses as normal distribution is central to linear analyses. There are few circular uniformity tests for circular data and three of them are presented here with their importance, weakness and strengths when used to test circular uniformity of different biological data.

3.3.1 Rayleigh test of *Plasmodium vivax* malaria case in Natavaram

One of the objectives of malaria transmission is studying whether there is seasonality or uniformly distributed around the year. The data (*Plasmodium vivax*) malaria case in Natavaram malaria data of Visakhapatnam district, we tested whether or not a Natavaram, 16^{th} index.

Chart 3.3 *Plasmodium vivax* malaria case in Natavaram

Years							
Months	2005	2006	2007	2008	2009	2010	2011
January	6	2	2	4	4	6	4
February	6	8	6	5	3	0	5
March	12	3	1	4	2	0	12
April	18	6	2	21	2	1	13
May	8	8	4	3	5	6	9
June	12	9	5	6	14	4	9
July	32	13	10	23	6	8	22
August	23	5	11	14	13	11	12
September	21	6	5	17	7	7	15
October	4	12	4	5	7	7	10
November	4	2	2	6	1	8	9
December	7	2	2	2	2	5	5
Total	153	76	54	110	66	63	125

Prior to any analysis plotting the data and see if it gives some clue about whether or not the data is uniformly distributed is advisable. Therefore, the rose diagram plot of *Plasmodium Vivax* is presented in figure 3.1 bellow.

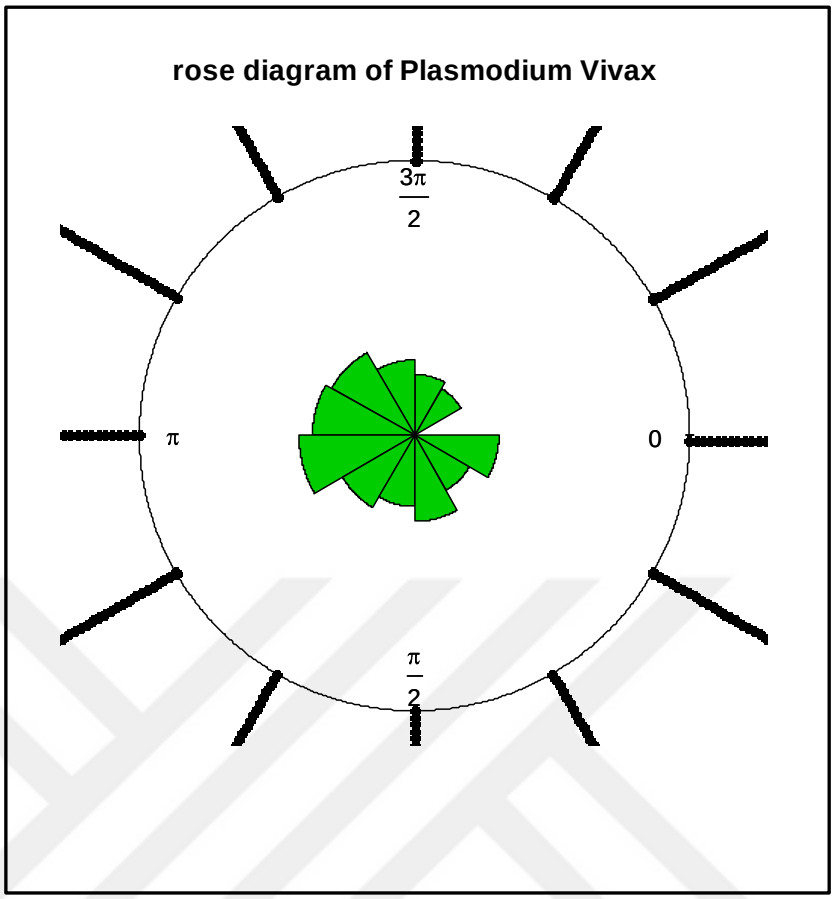


Figure 3.4 Rose diagram of *Plasmodium Vivax*

From the rose diagram in figure 3.4, even if it indicates variation it is difficult to guess whether *Plasmodium Vivax* data distributed uniformly or not around the year. Seeing circular plots one cannot identify whether data is distributed uniformly or not.

In the malaria data of Natavaram. 16th indexed study site, which is malaria case, since the objective is whether malaria case is distributed uniformly throughout the year or not and data points are 647 which is greater than 50, the p-value is found suing equation 2.45.

The two hypotheses are constructed as follows

$\boldsymbol{H_0}$: there is no mean direction

(Remember; uniform circular distribution does not have mean direction)

H_1: there is mean direction

Where n is sample size (angles) and $\bar{R}$ is mean resultant vector.

Chart 3.4 *Plasmodium vivax* malaria frequencies, sine, and cosine

rad(θ)	freq (f)	cos (θ)	sin (θ)	f*cos (θ)	f*sin(θ)
0.017453	28	0.999848	0.017452	27.99574	0.488667
0.523599	33	0.866025	0.500000	28.57884	16.50000
1.047198	34	0.500000	0.866025	17.00000	29.44486
1.570796	63	$6.13x10^{-17}$	1.000000	$3.86x10^{-15}$	63.0000
2.094395	43	-0.500000	0.866025	-21.5000	37.23909
2.617994	59	-0.86603	0.500000	-51.0955	29.50000
3.141593	114	-1.00000	$1.23x10^{-16}$	-114.000	$1.4x10^{-14}$
3.665191	89	-0.86603	-0.50000	-77.0763	-44.5000
4.18879	78	-0.50000	-0.86603	-39.0000	-67.5500
4.712389	49	$-1.8x10^{-16}$	-1.0000	$-9x10^{-15}$	-49.0000
5.235988	32	0.50000	-0.86603	16.0000	-27.7128
5.759587	25	0.866025	-0.50000	21.65064	-12.5000
sum	647			-191.447	-25.0902

From chart 3.4 Rayleigh test is calculated as follows

N= 647

$$\bar{C}_n = \frac{-191.447}{647} = -0.2959$$

$$\bar{S}_n = \frac{-25.0902}{647} = -0.03878$$

$$\bar{R}_n = \sqrt{(-0.2959)^2 + (-25.0902)^2} = 0.298429$$

$$Z = N\bar{R}_n^2 = 647(0.298429)^2 = 57.62179$$

$$P = \exp(-Z) = 9.44x10^{-26}$$

$Z_{critical}$ (647. 0.05) =2.9957

The P-value of Z using the second condition where the mean direction is not specified in equation 10.2 is $9.44x10^{-26}$. Since p-vale is less than 0.05 the null hypothesis that claims there is no preferred mean direction is rejected to prove *Plasmodium vivax* malaria case in Natavaram is not uniformly distributed around the year. In another word, it is to mean that *Plasmodium vivax* malaria case in Natavaram shows some seasonality.

The critical values of Rayleigh test are presented in a table form. The standard and extended tables of Rayleigh test table are presented in the appendix VII-IX at the end of this paper.

There are two basics assumptions in Rayleigh test

I. The data in hand have to be Unimodal data. On the other hand applying Rayleigh test for bimodal and multimodal data will give an erroneous result. This issue is considered to be the drawback of Rayleigh test.
II. The data are not diametrically bidirectional (not axial). Rayleigh test is not also applicable for axial data.

Rayleigh test does not differentiate the calculation of uniformity test for grouped data individually.

Pewsey et al. (2013) argue that Rayleigh test is less importance when data on the circle have multimode or more complex structure. In this situation, they recommend to use a series of uniformity tests so-called **Omnibus test**. These circular uniformity tests constitutes Hudges-Ajne's An, Kuiper's V_n Watson's U^2 and Rao's spacing tests. Rayleigh test is one of the oldest circular uniformity test. For instance, (Moore 1980) published the modification of Rayleigh test.

3.3.2 Circular χ^2 Test

The chi-squared test can be also used to test uniformity of circular variables uniformly distributed or not especially for the data points fairly large enough (>30). Our data is grouped data like in the case of malaria in months the groupings must not be less than four.

To calculate the expected values the formula bellow is used

$$\hat{f}_i = \frac{observations}{groups} = \frac{647}{12} = 53.91667$$

Then the Chi-squared test is calculated as follow

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(f_i - \hat{f}_i\right)}{\hat{f}_i} = \left(\frac{28-53.91667}{53.91667}\right) + \left(\frac{33-53.91667}{53.91667}\right) + \ldots + \left(\frac{25-53.91667}{53.91667}\right) = 11$$

For *Plasmodium vivax* malaria case in Natavaram presented in chart 3.3 the chi-squared test result is found to be

$$\chi^2_{critical}(11,0.05) = 19.675$$

Based on the chi-square test, the null hypothesis, which claims the distribution of *Plasmodium vivax* malaria case in Natavaram study site, is distributed uniformly around the year is accepted. Chi-square test demonstrates different result than Rayleigh test.

3.3.3 Rao's Spacing Test

To demonstrate circular analysis can be applied to different biological scenarios we chose another scenario with a different data set. Let's see of Panic attack is distributed uniformly around the week. There is some evidences panic attack varies with days of the week, seasons of the year even hours of the day. Such data have to be analyzed using circular methods taking in account circularity of the time they occur. Maximum

panic attack data collected and analyzed for panic attack ANOVA by Kao et al. (2014) from Taiwan is used to examine whether panic attack is randomly distributed throughout the week or not. Understanding driving factors of many cases including panic attack helps for precaution and/or easily manage that case.

Chart 3.5 maximum panic attack data in a week in Taiwan

Days	Maximum attack
Monday	10
Tuesday	8
Wednesday	8
Thursday	9
Friday	9
Saturday	12
Sunday	12

For starting pint, 0°=Monday and if we have one more assumption panic attack is randomly distributed within a given day. We allocated panic attack data in each day randomly within 24 hours of the day and record the resulting attack occurrence time. Convert the result into degrees and sort the data from the smallest to the highest ignoring the days when the attack occurs.

$$\lambda = \frac{360}{7 * 24} = 2.142857$$

Chart 3.6 data arrangement and calculation results in Panic attack data in Taiwan

No	Day	Attack time /day	Attack time /week	λ	Degrees	Ti	Ti-λ	|Ti-λ|
1	Monday	6	6	2.142	12.85714	2.1429	0.0000	0.000
2	Monday	7	7	2.14	15	2.1429	0.0000	0.000
3	Monday	8	8	2.14	17.14286	0.0000	-2.1429	2.143
4	Monday	8	8	2.142	17.14286	6.4286	4.2857	4.286
5	Monday	11	11	2.142	23.57143	2.1429	0.0000	0.000
	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…
65	Sunday	16	136	2.142	291.4286	2.1429	0.0000	0.000
66	Sunday	17	137	2.142	293.5714	4.2857	2.1429	2.143
67	Sunday	19	139	2.142	297.8571	6.4286	4.2857	4.286
68	Sunday	22	142	2.142	304.2857	68.571	66.428	66.429

Calculation of Rao's spacing test U is as follows

$$U = \frac{1}{2} \left| \left[(0.000 - 2.142857) + (0.000 - 2.142857) + \ldots + (66.42857 - 2.142857 \right] \right| = 169.2857$$

The p-value of U(68) of a 167.2857° is less than 0.001. Therefore. The null hypothesis claiming Panic attack in Taiwan is uniformly distributed throughout the week is rejected. Due to this fact, one can conclude that in the study area the maximum Panic attack shows some variation within a week.

3.4 Correlation and Regression Involving Circular Data.

3.4.1 Measure of Circular Correlations

3.4.1.1 Circular-Linear (Circular-Linear) Correlation

Using the Mardia's concept of the linear-circular correlation coefficient, the circular-linear correlation between malaria (*plasmodium falciparum*) and time of the year are analyzed to assess the association between malaria and the time of the data presented in chart 3.1. The sample circular-linear correlation coefficient between malaria and time of the year is calculated as follows using the expression 2.51.

First, linear correlation coefficients between malaria case, cosine and sine of time in radian measures' result is presented in chart 3.7.

Chart 3.7 individual correlation of Linear-sine. Linear-cosine and sine-cosine

	Linear	**Cos(θ)**	**Sin(θ)**
Linear	1	-0.7218	-0.26451
Cos(θ)	0.520992047	1	0.000984
Sin(θ)	0.069967872	$9.68492x10^{-07}$	1

To save space in chart 3.7, in the upper diagonal individual correlation coefficients are presented and in the lower diagonal their respective squared are given for the sake of easiness.

$$r^2 = \frac{r_{XC}^2 + r_{XS}^2 - 2r_{XC}r_{XS}r_{CS}}{1 - r_{CS}^2} = \frac{0.25099 + 0.069968 - 2(-0.7218 * -026451 * 0.000984)}{1 - 9.6849x10^{-07}}$$

$$= 0.59058$$

Based on the above calculation, r^2 found to be **0.59058** and r is just the square root of r^2 0.7685; it can be concluded that there is fairly enough association between these two variables. If one suspect the resulting value is not enough to conclude there is a circular-

linear association, F-test can be applied used in (Rao, 2001) to reject the null hypothesis claiming there is no circular-linear association. The problem with using this expression to calculate circular association is always the calculated result will be positive. In reality, relationships can be positive where both/all correlated variables go to the same ways and negative where variables have opposite relationships. In this calculation system, correlation coefficients are [0,1].

$$\frac{(n-3)r^2}{1-r^2} \sim F_{2,(n-3)} = \frac{(84-3)(0.59058)^2}{1-(0.59058)^2} = 116.8431 \sim 3.1065$$

$116.8431 \sim 3.1065$; Since F-calculated (116.84) is greater than F-table (2, 81, 0.05) =3.1065 the null hypothesis that claims "there is no correlation" is rejected. As a result, the calculated correlation between these two variables (malaria case and time of the year) is significant.

3.4.1.2 Circular-circular correlation

As it was mentioned in the methodology chapter, there are few correlation coefficients presented by different statisticians to calculate circular-circular correlations (Rao 2001, Pauen and Ivanova 2013, Kempter et al. 2012) but these methods are a bit sophisticated and need complex mathematical trigonometric knowledge in advanced level. The one easily understandable is a model used by NCSS (Statistical Software).

The heart attack data presented in chart 2.3 is used to demonstrate circular-circular correlation coefficients. Data points, their respective sine, and cosine components are presented in chart 3.8.

Chart 3.8 Heart attack date, birthdate and their respective sine and cosine components

Index	atack date(θ)	Birthdate (θ)	**sinθ**	**cosθ**	**sinϕ**	**cosϕ**	$\sin(\theta-\bar{\theta})$	$\sin(\phi-\bar{\phi})$	$\sin^2(\theta-\bar{\theta})$	$\sin^2(\phi-\bar{\phi})$	$\sin(\theta-\bar{\theta})*\sin(\phi-\bar{\phi})$
	1.232262	0.278187	0.9432	0.3321	0.2746	0.9616	0.8657	0.8756	0.7494	0.7667	0.7580
2	1.88462	5.26886	0.9512	-0.3087	-0.8491	0.5282	0.9918	-0.2239	0.9837	0.0501	-0.2220
3	0.076561	0.384266	0.0765	0.9971	0.3749	0.9271	-0.1090	0.9219	0.0119	0.8498	-0.1005
	…	…	…	…	…	…		…	…	…	…
	…	…	…	…	…	…		…	…	…	…
	…	…	…	…	…	…			…	…	…
1332	0.069301	0.37079	0.0692	0.9976	0.3624	0.9320	-0.1162	0.9165	0.0135	0.8401	-0.1065
1333	0.121208	3.917622	0.1209	0.9927	-0.7005	-0.7137	-0.0645	-1.0000	0.0042	1.0000	0.0645
1334	0.051763	2.533864	0.0517	0.99866	0.5710	-0.82095	-0.13363	-0.17985	0.017857	0.032346	0.024033
1335	0.04816	5.247077	0.0481	0.99884	-0.8604	0.509572	-0.13719	-0.24503	0.018824	0.060038	0.033617

Using expressions 2.31-2.36 the summary mean direction (or circular mean time in radian) for both birthdate and attack data is presented in chart 3.9.

Chart 3.9 summary mean direction and mean resultant vectors for both attack data and birthdate

	Attack date	Birthdate
N (sample size)	1335	1335
Sum of sines	231.333	-5.76816
Sum of cosines	1230.744	5.731769
Mean of sines	0.173283	-0.00432
Mean of cosines	0.921906	0.004293
Mean resultant vector (R)	0.93805	0.006091
Sine mean direction	0.184727	-0.70934
Cosine mean direction	0.98279	0.704866
Mean(direction)	0.185794	-0.78856
adjusted	0.185794	-0.78856

Using square rule of sine variables and C-C correlation expression presented in expression 6.63 circular- circular correlation coefficient of heart attack date and birthdate in radian is calculated as follows

$$\sin^2(\theta_i - \bar{\theta}) = \left(\sin\left(\theta_i - \bar{\theta}\right)\right)^2$$

$$r_c = \frac{\sum_{k=1}^{n} \sin\left(\theta_i - \bar{\theta}\right)\sin\left(\phi_i - \bar{\phi}\right)}{\sqrt{\sum_{k=1}^{n} \sin^2\left(\theta_i - \bar{\theta}\right)\sum_{k=1}^{n} \sin^2\left(\phi_i - \bar{\phi}\right)}} =$$

$$\frac{\sin(1.23-0.186)\sin(0.28-(-0.789)) + \sin(1.88-0.186)\sin(5.27-(-0.789))...}{\sqrt{\sin(1.23-0.186)^2\ \sin(0.28-(-0.789))^2 + \sin(1.88-0.186)^2\ \sin(5.27-(-0.789))^2...}} = -0.14221$$

The significance test is presented in the circular-circular regression subchapter.

3.4.2 Regression Analysis Involving Circular Data

Regression analysis that involves circular data is quite different from other forms of regression analysis in many aspects of the regression process. There are three basic regression analyses in circular data. Based on what parts of the data is measured in circular, regression analysis is summarized in chart 3.10.

Chart 3.10 Regression types involving circular data

Regression-type	Response variable	Explanatory variable
Linear-Circular (LC)	Linear	Circular
Circular-Linear (CL)	Circular	Linear
Circular-Circular(CC)	Circular	Circular

In next subchapters each regression type is examined using appropriate data sets.

3.4.2.1 Linear-circular regression

Regression analysis where the response variable is linear (magnitude) and the explanatory variable is circular (direction) is called Linear-Circular regression analysis. In this analysis, type observations enter into the model as a trigonometric polynomial. Step by step analysis process is presented in this subchapter using malaria data in India what was used to calculate circular descriptive statistics in previous subchapters with a little modification.

The modification is; malaria case frequencies are taken as a dependent variable and time of the year measured in radian as explanatory variable.

As it is proven in section 3.5.1.1 after being confident Malaria case and months of the year in radian have significant linear-circular correlation sinusoidal regression analysis using least square method and Fourier analysis is done. As it explained in section 2.4.1.1 the regression model is given as follows.

$$y_j = A_0 + A_1 \cos(\omega\theta - \varphi) + A_2 \cos(2\omega\theta - \varphi) + \ldots + A_p \cos(p\omega\theta - \varphi) + B_1 \sin(\omega\theta - \varphi) + B_2 \sin(2\omega\theta - \varphi) + \ldots + B_p \sin(p\omega\theta - \varphi) + \varepsilon_j$$

Where

θ is the independent angle and P is degree(order) of polynomial

A_0 is a mean level.

ω is angular frequency.

$A_{1,}A_{2,\ldots}A_p$ and $B_{1,}B_{2,\ldots}B_p$ are coefficients to be predicted from the model.

φ is peak of angular time where the frequency is highest. It is called acrophase.

One can reach to the intended result using the above regression expression. The biggest challenge in circular regression analysis is determination of degrees of polynomials (Rao, 2001). One way to tackle this challenge is to use iterative method. Running the regression with one more degree of polynomial and see if regression coefficient is significant. If the degree of polynomial at this level is also significant run the regression until the regression at this level is not significant anymore. Excel regression analysis method for Fourier regression expression is used.

Chart 3.11 Correlation of circular regression analysis at the 1st order of polynomial

Regression statistics	
Multiple R	0.765908098
R square	0.586615215
adjusted R square	0.56568434
Standard deviation	67.81767387
observations	84

From the above chart it is concluded that the dependent variable malaria case and components of the independent variable that is time of the year in radian measurement have **0.7659** correlation that shows a strong relationship between these variable. It also determined that R-square value 0.5866 is fair enough for our model. Even if R-square does not tell the whole story, there is a clue that this model explained about 58.66% source of variation. One can see that from adjusted R-square values as well.

Chart 3.12 ANOVA of regression analysis of the first order of polynomial

ANOVA					
	df	***SS***	***MS***	***F***	***Significance F***
Regression	4	515598.6	128899.6	28.02631	$1.7x10^{-14}$
Difference	79	363339.7	4599.237		
Total	83	878938.3			

From chart 3.12, the model for this (1^{st}) choice of order of polynomial is significant. When in one order of polynomial the coefficient of one of the component is significant, if one suspect that the coefficient of the next order of polynomial might be significant as well. So, the regression was run again with one more order of polynomial and found that the regression is significant at that level too (second order of polynomial). The regression for the third order of polynomial and found significant as well.

When the regression is computed for the fourth order of polynomial it is found both cos(ω4θ-acrophase) and sin(ω4θ-acrophase), not significant as indicated in row 9^{th} and 10^{th} of chart 3.13.

Chart 3.13 Regression coefficients chart of fourth order of polynomial

		coefficients	Standard dev.	t Stat	P-values	lower %95	Upper %95
1	*intercept*	148.152	55.98611	2.646227	0.009913	36.62189	259.682
2	*cos(ωθ- φ)*	-25.5752	39.52759	-0.64702	0.519593	-104.318	53.16778
3	*sin(ωθ- φ)*	5.818694	10.04018	0.579541	0.563961	-14.1824	25.81976
4	*cos(2ωθ- φ)*	124.9673	19.68664	6.347822	1.51E-08	85.74949	164.1851
5	*sin(2ωθ- φ)*	110.5624	46.19998	2.393126	0.019208	18.52731	202.5975
6	*cos(3ωθ- φ)*	33.37631	12.18262	2.739667	0.007681	9.107296	57.64533
7	*sin(3ωθ- φ)*	-41.1102	14.12306	-2.91085	0.004743	-69.2447	-12.9756
8	***cos(4ωθ- φ)***	**2.405278**	**12.54185**	**0.19178**	**0.848433**	**-22.5794**	**27.38993**
9	***sin(3ωθ- φ)***	**4.791897**	**11.54729**	**0.41498**	**0.679341**	**-18.2115**	**27.79527**

From raw nine and ten of chart 3.13, one can see that since p-values are not the significant inclusion of fourth order polynomial is not important. Due to this fact regression model with third order of polynomial is the **best model**.

Chart 3.14 Regression chart of third order polynomial

Regression statistics	
Multiple R	0.814898
R square	0.664059
adjusted R square	0.628225
Standard deviation	62.74514
observations	84

From chart 3.14, one can see that multiple correlation R is 0.815 that show there is a strong association between the dependent variable Malaria case in moths and components of the independent variable circular time in radian. R-square value 0.664 is **fair enough** for circular regression with multiple components. If R-squared is not tell the whole story adjusted R-squared 0.628 can be used. These whole scenarios show the

model at three order of polynomial is the "best model". The overall significance of the model is shown in the chart below and $5.37x10^{-15}$ value reveals it is significant.

Chart 3.15 Overall model ANOVA of regression of third order of polynomial

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	583666.9	72958.36	18.53169	$5.37x10^{-15}$
Difference	75	295271.4	3936.952		
Total	83	878938.3			

Chart 3.16 Regression coefficients chart of third order of polynomial

	Coefficient	**Standard dev.**	**t Stat**	**P-values**	**lower %95**	**Upper %95**
Intercept	145.1842	55.19465	2.630403	0.010296	35.27758	255.0908
cos(ωθ- φ)	-28.2717	38.76169	**-0.72937**	0.467986	-105.456	48.91265
sin(ωθ- φ)	5.124309	9.827532	0.521424	0.603568	-14.4448	24.69343
cos(2ωθ- φ)	124.2058	19.43201	6.391812	$1.15x10^{-08}$	85.51169	162.8999
sin(2ωθ- φ)	114.1844	45.17945	2.527353	0.013542	24.22063	204.1482
cos(3ωθ- φ)	34.93872	11.76197	2.970482	**0.003964**	11.51765	58.35979
sin(3ωθ- φ)	-41.1772	13.9749	-2.94651	**0.00425**	-69.0048	-13.3496

The best model is presented in chart 3.16 with respective coefficients. In the chart, the p-value of one of the components is not significant for Even if they are not significant at this level since the higher order polynomials are significant they have to be included in the model just like multiple linear regression where interaction case is significant then individual cases should be included. The final model is given in the expression bellow.

$$y = 145.1842 - 28.2717\cos(\omega\theta - \varphi) + 124.2058(2\omega\theta - \varphi) + 34.93872(3\omega\theta - \varphi) + 5.124309\sin(\omega\theta - \varphi) + 114.1844\sin(2\omega\theta - \varphi) - 14.1772(3\omega\theta - \varphi)$$

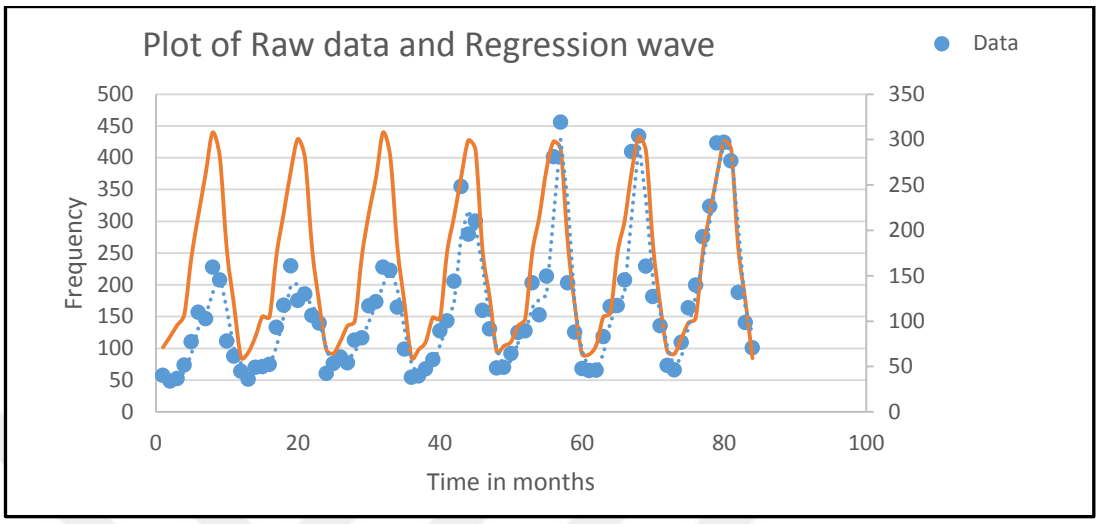


Figure 3.5 Plot of raw data frequency and predicted values

The main objective of linear-circular regression is to successfully predict a smooth sinusoidal curve from uneven peaks and troughs of raw data plots.

In figure 3.5, one can see that the regression wave has a similar trend with the mean of raw frequency wave. The regression wave looks like having high peak due to the high frequency of row data around month 57^{th}, 58^{th}, 80^{th}, 81^{th}. Considering these values as outliers is depending on non-statistical factors like whether there is an epidemic of malaria in these months or some other environmental and social factors like sudden population size change.

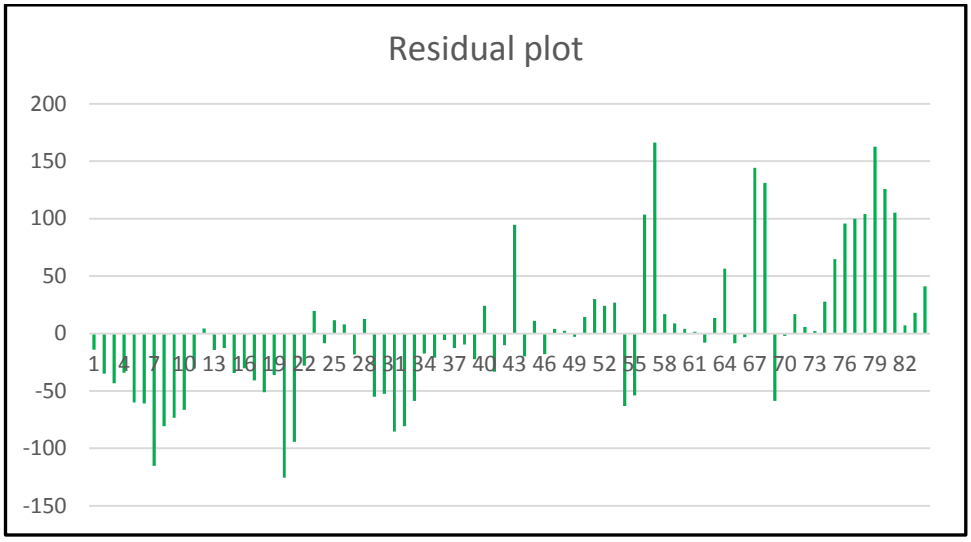


Figure 3.6 Residual plots

In a linear regression (including some nonlinear regression) normality of residual plot is important but in circular regression residuals should not distributed normally as it is shown in figure 3.6. In the figure, it is clearly indicated that there is some pattern in the residual plot. Studying this plot and come up with some sensible distribution (if there is) a matter of future work.

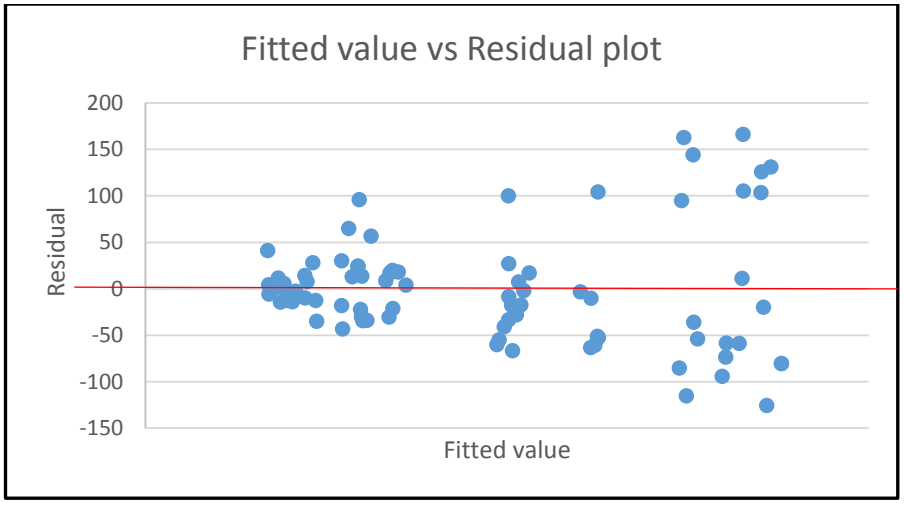


Figure 3.7 fitted Vs residual plots.

In circular regression analysis, the error term is not needed to be normally distributed.

Fitted value verses residual plot in figure 3.7 indicates some relationship as it is expected but the relationship is nonlinear if we were fitted line it would show heteroscedasticity. In circular case, heteroscedasticity is already expected.

3.4.2.2 Circular-linear regression

Linear circular regression is a regression type whereby direction (time in circular sense) is predicted from magnitude. As Fisher and Lee (1992) stated this regression type is one of the neglected regression types.

The method in this subchapter is demonstrated on the data that was presented by "Doç. Dr. Fazilet DUYGU (2016) for Turkish public health organization titled "Dünya ve *Türkiye'de Kırım-Kongo kanamalı ateşi epidemiyolojisi*" to mean "*Crimean-Congo hemorrhagic fever epidemiology in the world and Turkey*" in which data collection methods have been described in detail in their presentation. Monthly CCHF data were collected from 2008-2015 period data plots are presented in figure 3.8.

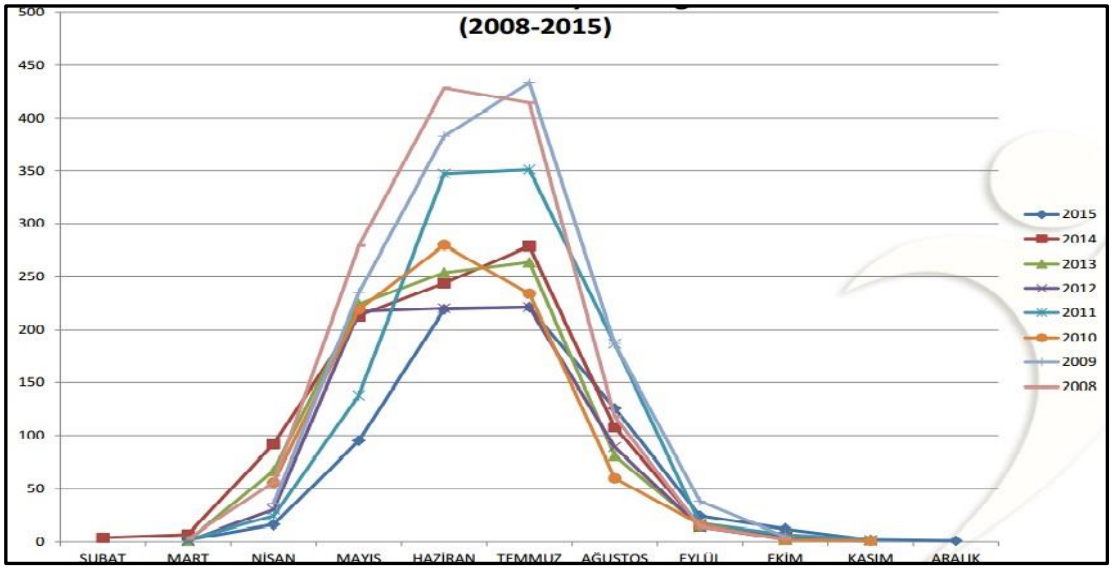


Figure 3.8 monthly data plots of CCHF from the year 2008-2015

As one can see from the plot data points are not given. Therefore there must be a way to obtain data points from the plot. To obtain data points from the data plot *GetData Graphics Digitizer 2.26* is used. Obtaining data from plots is simple;

- load the plot into the software,
- set the minimum X and Y axes of the plot,
- Stream very carefully on the data point their values wanted to be obtained.

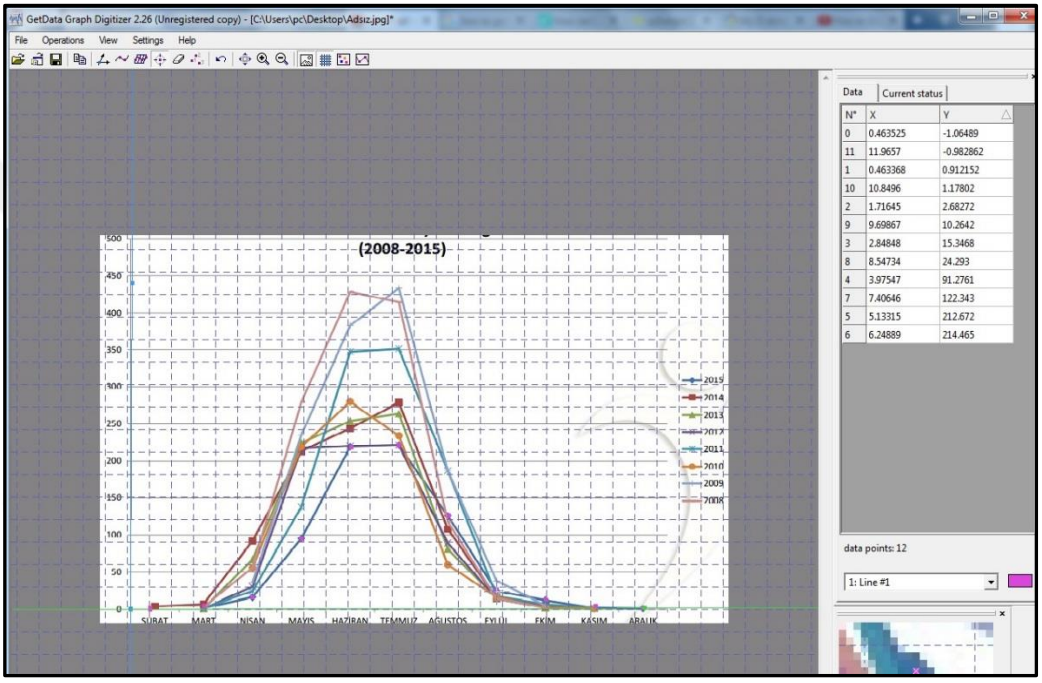


Figure 3.9 loading and obtaining data points from data plots

In figure 3.9, there are three fields in the page. The main and the biggest field the page is the area where data plot is loaded; in the top right corner there is the data view field and in the right bottom field there is a small area where exactly we point to on the data plot field. After finishing these process from file menu we export the data values in any appropriate format.

Chart 3.17 Radian measures of months and monthly CCHF data points (frequencies) from

Months	Days	Adjustment	Degree	Radian	08	09	10	11	12	13	14	15
January	31	0.96774	1	0.01745	0	0	0	0	0	0	2	1
February	28	1.07143	30	0.5236	1	1	1	1	2	2	7	3
March	31	0.96774	60	1.0472	2	2	2	3	7	1	89	6
April	30	1	90	1.5708	54	29	51	21	28	61	205	15
May	31	0.96774	120	2.0944	273	234	214	135	205	212	237	91
June	30	1	150	2.61799	416	372	272	337	215	246	271	212
July	31	0.96774	180	3.14159	403	421	227	342	216	254	104	214
August	31	0.96774	210	3.66519	110	180	57	183	85	75	13	120
September	30	1	240	4.18879	13	36	11	23	18	11	3	21
October	31	0.96774	270	4.71239	0	3	0	8	1	1	0	0
November	30	1	300	5.23599	0	0	0	0	0	0	0	0
December	31	0.96774	330	5.75959	0	0	0	0	0	0	0	0

As it was done in Malaria data previously, the first step is to convert months into radian and adjust to the way, one day represents 1 degree (0.017453 rad). Since months in a year do not have equal days, each month has to be adjusted in the way to contribute equal frequencies. To do so, we have changed 365 days in a year to 360°.

First, 30 is divided by a number of days in each month to have the rate of each month's contribution to 360°.

$Adjustment = \frac{30}{x}$;Where x is days of each month

The adjustment result is multiplied by the frequency of each month. To retain the total frequency adjusted frequencies are added and divided by the sum of frequencies before adjustment. Then, the result is multiplied by adjusted frequencies. The resulting frequencies is shown in chart 3.17.

Chart 3.18 frequency of CCHF in Turkey

index	Months	Days	Days Adjustment	Degree	**Radian**	Freq	Adjusted	**Adjusted and fixed**
1	January	31	0.96774	1	**0.01745**	0	0	**0**
2	February	28	1.07143	30	**0.52360**	1	1	**1**
3	March	31	0.96774	60	**1.04719**	2	2	**2**
4	April	30	1.00000	90	**1.57079**	54	54	**55**
	…	…	…	…	**…**	…	…	**…**
	…	…	…	…	**…**	…	…	**…**
	…	…	…	…	**…**	…	…	**…**
92	August	31	0.96774	210	**3.66519**	120	116	**119**
93	September	30	1.00000	240	**4.18879**	21	21	**22**
94	October	31	0.96774	270	**4.71238**	0	0	**0**
95	November	30	1.00000	300	**5.23598**	0	0	**0**
96	December	31	0.96774	330	**5.75958**	0	0	**0**

From chart 3.18 radian measure of time as circular dependent variable and adjusted & fixed frequencies as linear explanatory variable are used. The main objective here is to estimate the mean CCHF occurrence time from the frequency.

The methods uses consecutive iterative methods when apply to the data and find the log-likelihood (Anonymous 2017c).

Using R- package the result of the regression is presented as follows

lm.circular(y=y, x=x, init=c(5,1), type='c-l', verbose=TRUE)

Iteration 1 : Log-Likelihood = 18.11643

Iteration 2 : Log-Likelihood = 17.51982

Iteration 3 : Log-Likelihood = 19.75503

Iteration 4 : Log-Likelihood = 36.90548

Iteration 5 : Log-Likelihood = 38.85477

Iteration 6 : Log-Likelihood = 40.72392

Iteration 7 : Log-Likelihood = 42.35866

Iteration 8 : Log-Likelihood = 43.38337

Iteration 9 : Log-Likelihood = 43.91427

1

Iteration 19 : Log-Likelihood = 44.32773

Iteration 20 : Log-Likelihood = 44.32774

……

Iteration 37 : Log-Likelihood = 44.32774

Iteration 38 : Log-Likelihood = 44.32774

Iteration 39 : Log-Likelihood = 44.32774

Iteration 40 : Log-Likelihood = 44.32774

Iteration 41 : Log-Likelihood = 44.32774

Iteration 42 : Log-Likelihood = 44.32774

Iteration 43 : Log-Likelihood = 44.32774

Iteration 44 : Log-Likelihood = 44.32774

Iteration 45 : Log-Likelihood = 44.32774

Iteration 46 : Log-Likelihood = 44.32774

Iteration 47 : Log-Likelihood = 44.32774

Call:

lm.circular.cl(y = ..1, x = ..2, init = ..3, verbose = TRUE)

As Fisher and lee (1992) presented there are three models for angular response variable. R-uses the mixed model out of these three models.

p-values are approximated using normal distribution

Using the R-two step (Mixed) regression model iteratively the resulting coefficients are presented and interpreted as follows.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
[1,]	**-0.36607**	0.09178	3.989	3.32e-05 ***
[2,]	**0.16397**	0.09195	1.783	0.0373 *

Log-Likelihood: 44.33

Summary: (mu in radians)

mu: -0.3439 (0.09993) kappa: 1.674 (0.2201)

$\hat{\mu} = -0.36607 radian$ and $\hat{\beta} = 0.16397 radian$

Our models containing the above coefficients generally accepted to be best model of Circular-linear regression prediction methods.

$$\hat{\mu} = \mu_0 + \hat{b}_i x_i = -0.36607 + 0.16397 x_i$$

3.4.2.3 Circular-circular regression

To continue with Circular-Circular regression in this subchapter where both the response variable and explanatory variable are circular. The model used is centered model. The uncentered counterpart of the model presented by Rivest (1997) where directions are measured (-μ, μ) rather than origin (0, 0).

To demonstrate the model of circular-Circular regression simulated Heart attack data is used.

Before doing the analysis it is good to explain some aspects of Heart related disease and how it is better for analysis of circular regression.

Heart disease is the leading cause of death in the United States (David and Robert 2016). According to Onat et al. (1993), the prevalence of coronary Heart disease in

Turkey is 5.8%. Even if life factors such as age, Smoking status, Work or life-related stress, genetic…etc. influences the occurrence of Heart-related diseases especially sudden heart attack recently proven that the case is also related to time of the day. Muller et al. (1989) claim that the onset of myocardial infraction and sudden cardiac-related death is frequently triggered by daily activities based on the prior decade of data from their paper. In their paper, they argued that myocardial infraction is more likely to happen in mornings than late in the day. Because of the close relationship of myocardial infraction and sudden cardiac death, there is high probability of sudden cardiac death in mornings. They also supported their arguments with data onset of cardiovascular-disasters, myocardial infraction, sudden cardiac death and stroke is related to a disorder of transient myocardial ischemia.

The case does not only depend on the time of the day. According to an article published in The New York-times march 19. 1993 edition, sighting the work of Dr. Alan of the Robert wood Johnson medical school, Based on the Birthday finding 118.955 heart attack patients the found a very admirable correlation between Birthday and heart attack time in a year (Anonymous 2017b). Dr. Alan clams that, there are emotional relationships of the disorder and peoples' behavior around their birthdays including smoking and drinking that believed to trigger heart-related problems. Since then. Different cardiologists and statisticians become curious about the relationship between birthdays and heart attack days. Moreover, the majority of online medical and news outlets (to our knowledge) reported about the issue widely. According to the publication, there is also another claim about heart attack occurs more commonly in cold days that the warmer days regardless of time of the year. To some people, though, this kind of claims are considered as just conjure.

On the other Hand, Poltavskiy et al. (2016) studied the relationship between birth month and 5 major cardiovascular diseases. Their conclusion invites scientists for a better look into this dimension of study about heart-related cases.

For this paper to demonstrate the circular-circular relationship of birth date and heart attack date data is simulated using r with assumption birthdate in Ankara to be uniformly distributed around the year. This fact gives information that the independent variable which is birthdate to be "circular uniform distribution". Bivariate data is simulated with minimal but significant correlation with components of (sine and cosine) birthdates for heart attack-date.

Unlike other chronic heart-related problems that case can be found a long time after the real disease occurrence time due to delay to report to the hospital. A heart attack is very acute so hospital admission date can be considered the attack date. For specification of sample size Ankara population and Heart attack, the prevalence in the city is taken as a reference.

The prevalence of Heart stroke in Central Anatolia area that includes Ankara is estimated to be 7.8% (Ünal et al. 2013). 7.8% of approximately 9.030.837 population in central Anatolia in 2011 (Ünal et al. 2013) about 99.339 people reported stroke. According to the calculation burden of Ankara will be 44491. When on decide to take 3% of the Ankara Stroke patients the sample size become 1335.

1335 bivariate sample was simulated using R-software package in the following syntax with 0.3, 0.4 and 0.5 optimum correlation of individual components of the birthdate variables and attack date variables with the multiple population correlation coefficient 0.67 and P-value 0 (highly significant correlation).

```
> birthdate <- runif(1335.min=0. max=6.283185307)
> Attackdate <- atan2(0.3*cos(birthdate)
+              + 0.4*sin(birthdate).
+              0.5*(sin(birthdate))+ rvm(1335, 5.024972, 0.101588))
> circ.cor(birthdate. Attackdate. test=T)
```

r	*test.stat*	*p.value*
-0.1422142	-5.007078	5.526265×10^{-07}
r	test.stat	p.value

Since to present all data points is taking too much spaces only the head and tail of simulated data is presented in chart 3.19.

Chart 3.19 head and tail of simulated data

Index	Birthdate	Heart attack date
1	0.278187	1.232262
2	5.26886	-1.88462
3	0.384266	0.076561
4	2.168721	0.157084
	…	…
	…	…
	…	…
1332	0.762404728	0.233955145
1333	2.435280699	0.013507469
1334	1.169476352	0.078194345
1335	2.033624633	0.125431734
1332	0.762404728	0.233955145
1333	2.435280699	0.013507469

In principle, circular data can be negative where the sense of direction is opposite to the intended (expected) directions. In this data, since the data is time measurement, negative time in not acceptable. Using the absolute value command negative data points are changed into positive.

At this level, there is bivariate circular data representing birthdate and heart attack date measured in radian (0.2π).

The data having joint Probability density function $f(\theta.\phi).0 < \theta.\phi \leq 2\pi$. To predict ϕ (attackdate) from a given θ (birthdate) consider a conditional expectation of $(e^{i\phi}|\theta)$ vector.

$$E\left(e^{-i\phi} \mid \theta\right) = \rho(\theta)e^{i\mu(\theta)} = g_1(\theta) + ig_2(\theta) \tag{3.3}$$

$\mu(\theta)$ represents the conditional mean direction of Φ for a given θ and $0 \leq \rho(\theta) \leq 2\pi$ is the conditional concentration parameter towards this direction (Rao.2001).

Since $g_1(\theta)$ and $g_2(\theta)$ are circular with a period 2π they can be approximated with their Fourier series as follows with an appropriate degree of trigonometric polynomials (order of polynomial)

The main objective in this paper is assessing application of already available models. It is believed there is no need to go step by step calculation of each regression coefficient. One who want to follow step by step methods can use "Topics in Circular statistics" of Rao (2001) and also can follow steps presented in section 2.5.2 of this paper. Emphasis is given to examine applicability of these models in Biological scenarios. Due to this fact, using "CircStats" and "circular" packages of r-software is quite satisfactory. Based on the least square estimation methods and "CircStats" and "circular" packages in r-software in the way bellow

> *install.packages("CircStats")*

> *library(CircStats)*

> *install.packages("circular")*

> *library(circular)*

Chart 3.20 Circular-circular regression coefficients of Heart attack data

	Order 1	Order 2
Intercept	0.49466899 ($\alpha 0$)	-0.12129390 ($\delta 0$)
cos.alpha1	-0.56119174 ($\alpha 1$)	0.23982541($\delta 1$)
cos.alpha2	0.07016889 ($\alpha 2$)	-0.04892861 ($\delta 2$)
sin.alpha1	-0.57186760 ($\beta 1$)	-0.56766988 ($\beta 2$)
sin.alpha2	0.27529379 ($\gamma 1$)	0.14979634 ($\gamma 2$)

From the above chart 3.21 coefficients of different orders of polynomial and sine & cosine

$\hat{\alpha}_0 = 0.49466899$

$\hat{\alpha}_1$ = -0.56119174

$\hat{\alpha}_2 = 0.07016889$

$\hat{\beta}_1 = -0.57186760$

$\hat{\beta}_2 = -0.56766988$

$\hat{\delta}_0 = -0.12129390$

$\hat{\delta}_1 = 0.07016889$

$\hat{\delta}_2 = -0.04892861$

$\hat{\gamma}_1$ = -0.04892861

$\hat{\gamma}_1 = 0.14979634$

cos(attack-date) = 0.49466899 - 0.56119174*cos(birthdate) + 0.23982541*cos2(birthdate) - 0.57186760*sin(birthdate) - 0.56766988*sin2(birthdate)

sin(attack-date) = -0.12129390 + 0.07016889*cos(birthdate) - 0.04892861*cos2(birthdate) + 0.27529379 *sin(birthdate) + 0.14979634*sin2(birthdate)

For any birthdate data measured in radian for heart attack risk groups, one can predict the time where heart attack can be occur.

Based on this calculation for a person who was born on **19/08/1977** and from some other preliminary evidence the person believed to be in the group of heart attack risk. The expected heart attack time for that person can be predicted as follows

Step 1. Convert the person's birthdate into radian without giving any consideration to the year when he/she was born. There are 365 days in a year but 360° in a circle to adjust this problem multiply each day by 0.986301 conversion factor, and convert the resulting degree measure into radian.

For this particular example the person's birthdate in radian measure is 3.976482 radian

Step 2. Using the regression expression found from the above calculation, predict expected attack-date

cos(attack-date) = 0.49466899 - 0.56119174*cos(3.976482) + 0.23982541*cos2(3.976482) -0.57186760*sin(3.976482) - 0.56766988*sin2(3.976482)
sin(attack-date) = -0.12129390 + 0.07016889*cos(3.976482) - 0.04892861*cos2(3.976482) + 0.27529379 *sin(3.976482) + 0.14979634*sin2(3.976482)

Step 3. Find the cosine and sine of birthdate measured in radian in this example the sine and cosine of 3.976482 are -0.67126 and -0.74122.

Step 4. Predict the expected heart attack date using sine & cosine components of the birthdate and regression coefficients.

cos(attack-date) = 1.55504

sin(attack-date)= -0.45064

Using the inverse of tangent function

$$\tan^{-1} = \frac{1.55504}{-0.45064} = -0.28979$$

From the result of the above calculation, a negative sign shows to the opposite direction from the specified direction when the data was collected. In this data, the clockwise direction was specified as a direction of movement from the beginning point of reference. In this paper, January first was selected as a starting point and goes to February. The result -0.28979 indicates that the expected heart attack date is 0.28979 radian to counter-clockwise direction starting from January First. Therefore, the result becomes **13th of December** is expected heat attack date.

The main objective of this analysis is not to successfully predict heart attack date. It is to show how circular regression can be done using data collected in time-dependent variables.

In the R-Circular-Circular regression packages, there is a message says "*Higher order terms are not significant at the 0.05 level*". This gives the third and more order of polynomial is not important to our model and our best model is a model with a second order of polynomial.

4. CONCLUSION AND RECOMMENDATION

4.1 Conclusion

This thesis is aimed to accomplish application of circular regression analysis on different biological data. However, since circular statistics is new aspect of analysis we had to start our methods from very basics of circular data. We proceeded to circular distribution and descriptive statistics of circular data, deep in the chapters of the paper we analyzed circular Uniformity test and circular ANOVA before we enter into the main objective of this paper which is regression analysis involving circular data.

Circular data analysis is very different in many aspects from the usual data analysis methods used in many fields of estimation, inference, and predictions. The first unique feature of the circular analysis is a data type. Circular data arises from mainly two measurements; direction and clock. Circular data analysis is relatively new fields of statistics where its advancement is only 50 years old.

There are different challenges in the analysis of circular data since the topic is relatively new. These challenges raised from lack of adequate reference to the lack of tractable models to lack of appropriate data sources. Because circular data need more, time to collect due to the fact that circular data is longitudinal. Sometimes, it might need 10 and more years to collect readily analyzable data.

Data sets for this paper was extracted from secondary data sources and we used simulated data sets for the majority of our analyses due to lack of readily available primary data sources. We used malaria data in India, Panic attack data in Taiwan, Heart-related data in Turkey.

We used circular statistical methods and models to address crucial question “how to use circular models in biological data” These show that circular data analysis can be applicable in a variety of biological fields from different parts of the world.

Conclusive discussions on findings were given in respective chapters along with models and methods.

We dedicated this chapter for the overview of summary and to give some ideas to recommendations and future works.

In the introductory chapters we gave very detailed ideas about statistical analyses and developments of statistics from its formative stages and then we connected those overview to circular statistics as well.

Following Introductory chapters, we pinpointed objectives of this thesis, general materials, and methods. In each chapter we explained methods that are important in respective chapters.

Even if we faced diverse challenges in lack of data, shortage of references, deficiency of appropriate models we managed to use what in our hand and contributed to the relatively better understanding of circular statistics in different fields of study.

4.2 Recommendation and Future Work

For the work on the domain of circular statistical analysis in general and circular regression analysis in particular, improvements can be made in the area of data collection, and data managements.

On the data collection, when data are collected for circular analysis there should be a consideration of the cyclical nature of time or clock.

As distribution and data plotting are the backbone of statistical analysis there have to be better and easily understandable distribution parameters and plotting mechanisms.

There should be better and agreeable circular variance and standard deviation calculation methods.

As we have mentioned in chapters of the paper there should be circular post hock methods.

In circular regression methods, there are plenty of gaps between the model we have now and the “best” circular regression model would be.

REFERENCES

Abdi, H. 2010. Partial least square regression and projection on latent structure regression (PLS regression. DOI: 10.1002/wics.51.

Ajne, B. 1968. A simple test for circular uniform distributions. *Biomatika*. **55**(2), 343-354.

Aldrich, J. 2005. Fisher and Regression. *Statistical science*. **20**(4), 401-417.Linton, M.A. 2013. History of navigation. PDF generated. Thu, 04 Jul 2013 06:00:53 UTC.

Andrew, K., Steve, L., Ulisses, C. and Jonathan, P.2000. Climate change and vector-borne disease. *Bulletin of World Health Organization*. 2000, **78**(9)

Anonymous 2016 An Iterative guide to Fourier Transform. https://betterexplained.com/articles/an-interactive-guide-to-the-fourier-transform/ (26.10.2016, 17:58).

Anonymous 2017a. http://www.probabilityandfinance.com/articles/41.pdf. 12.12.2016.

Anonymous. 2014a. WHO (2014).A global brief on vector-borne disease.

http://apps.who.int/iris/bitstream/10665/111008/1/WHO_DCO_WHD_2014.1_eng.pdf (11.04.2017)

Anonymous. 2014b. WHO (2014) A global brief on vector-borne disease.

http://apps.who.int/iris/bitstream/10665/111008/1/WHO_DCO_WHD_2014.1_eng.pdf. (11.04.2017)

Anonymous.2017b.http://www.nytimes.com/1993/03/19/us/chances-of-heart-attack-are-greatest-on-birthday.html. 13.14.2017.

Anonymous.2017c.http://www.glennshafer.com/assets/downloads/articles/article50.pdf. G. Shefer, 09.01.2017(15:45)

Arnold, B.C. and Sengupta, A. 2004. Probability distributions and statistical inference for axial data. *Department of statistics, university of California.*

Attinger, O. 1996. Use of Fourier series for analysis of biological system. *Biophysical journal*. **6**(3), 291-304.

Barnes, T.J. 1998. A history of regression: Actors, networks, Machines and numbers. *Environment and Planning*. **30**(1998), 203-223.

Berens, P. 2009. A MATLAB toolbox for circular statistics. Journal of statistical software. **31**(10)

Borgan, N., Bogdan, K. and Futschik, A.2002. A data driven smooth test for circular uniformity. *The institute of statistical mathematics*. **54**(1), 29-44.

Brillinger, D.R. 2002. John W. Tukey. His life and proffetional contributions. *The annals of statists*. **30**(6), 1535-1575.

Cairn,M.,Walker,P.,Okell, L.,Griffin, J., Garske,T., Asante, K., Owusu-Agyei, S., Diallo, D., Dicko, A., Cisse, B., Greenwood, B., Chandramohan, D, Ghani, A. and Milligan, P. 2015. Seasonality in malaria transmission: implication for

case-management with long-acting artemisinin combination Therapy in Sub-Saharan Africa. *Malaria Journal*. DOI: 10.1186/s12936-015-0839-4.

Chandra, R., Neelapu, N. and Sidagam, N. 2015. Association of climatic variability, vector population and malaria in district of Visakhapatnam, India, A modelling and prediction analysis. *PLOS one*. **10**(6).

Corcoran, J., Chhetri, P. and Stimson, R. 2008. Using circular statistics to explore the geography of Journey to work. *Regional science association international*.*DOI:10.1111/j.1435-5957.2008.00164.x.*

Denis, D.J. 2015. Applied Univariate, Bivariate, and Multivariate statistics. Wiley and Sons, Incorporated, John, USA.

Downs, T.D. and Mardia, K.V. 2002. Circular regression. *Biomatika*. **89**(3), 683-697.

Duckworth, W.M. and Stephenson, W.R. 2012. Beyond traditional statistics. *The American statistician*. **56**(2), 230-233.

Duygu, F. 2016 Dünya ve Türkiye'de Kırım-Kongo kanamalı ateşi epidemiyolojisi. *Türkiye halk sağlık kurumu*.

Fienberg, S.E. 1992. A brief history of statistics in three and one-half chapter: essay review. *Statistical science*. **7**(2), 208-225.

Fisher, N. I. 1995. Statistical analysis of circular data. *Cambridge University Press*. NY, USA.

Fisher, N.I. and Lee, A.J. 1992. Regression model for Angular response. *Biomatrics*, (**48**),665-677

Follmann, D.A. and Proschan, M.A. 1999. A simple permutation-type method for testing circular uniformity with correlated angular measurments. *Biomatika*. **55**(3), 782-791.

Gaile, G. and Burt, J. 1980. Directional statistics. Norwich, England: Geo abstracts

Galton, F. 1886. Regression towards Mediocrity in hereditary stature. *Anthropological miscellanea.* 246-263.

Gattho, R. and Jammalamadaka, S.R. 2007. The generalized Von Mises distribution. *ELSEVIER*. **4**(3), 341-353.

Ghasemi, A. and Zahediasl, S. 2012. Normality test for statistical analysis: A guide for non-statisticians. *International journal of Endocrinology and Metabolism*. **10**(2), 486-489.

Ghitany,M.E., Atieh, B. and Natarajah, S. 2008. Lindley distribution and its application. *Mathematics and Computer in simulation (ELSEVIER)*. **78**(4), 493-506.

Gill, J. and Hangartner, D. 2010. Circular data in political science and how to handle it. *Oxford journals*. **18**(3), 316-336.

Grassly, N.C. and Fraser, C. 2006. Seasonal infectious diseases epidemiology. *Proc.Biol Sci.* **273**(1600), 2541-2550.

Gubel, E.J., Greenwood, J.A. and Duran, D. 2012. The circular normal distribution: Theory and table. *Journal of American statistical association*. **48**(261), 131-152.

Gubel, E.J., Greenwood, J.A. and Duran, D. 2012. The circular normal distribution: Theory and table. *Journal of American statistical association*. **48**(261), 131-152.

Guterman, P.S., Allison, R.S. and McCague, H. 2009. the application of circular statistics to Psychophysical research. *Proceedings of Fechner day*. **25**(2009).

Herone, M. 2016. Deaths: Leading causes for 2013. *National vital statistics reports*. **65**(2).

Hussin, A.G. 2006. Hypothesis testing of parameters for ordinary linear circular regression. *Pakistan journal of statistics and operation research*.**2** (2).

Jammalamadaka, S.R. 2001. Topics in circular statistics. Singapore, *World scientific publishing Co.Pte.Ltd.*

Jeff, G. 2008. Circular data in political science and how to handle it. *American political science annual meeting.* Boston, MA, August 29.

Jespersen, J. and Fitz-Randolph, J. 1999. From Sundials to Atomic Clocks: understanding time and frequency. US department of commerce. *Monograph 155*.

Joshi,S. and Jose,K.K. (2016) Wrapped Lindley distributions. *Journal of Communications in statistics: theory and methods*.

Kao, LT., Xirasagar, S., Chung, KH., Lin, Ch., Liu, SP. And Chung, SD. 2014. Weekly and Holiday related patterns of Panic attacks in Panic disorder: Population-Based study. Plos One. 9(7).

Kato, S. 2008. A circular-circular regression model. *Statistica sinica*. **18**(2), 663-645.

Kemper, R., Leibold, C., Buzsaki, G. and Schmidt, R. 2012. Quantifying circular-linear associations: Hippocampal phase precession. *Journal of Neuroscience Methods*.**207** (1), 113-24.

Kupkova, E. 2005. Radians vs Degrees. *Acta didactica universitstis comenianae mathematics*. Issue(5).

Legrand, J. (2007)Understanding the dynamics of Ebola epidemic. *Epidemoil Infect*.1 **34**(4), 610-621.

Lightner, J.E. 1991. A brief look at the History of probability and statistics. *The Mathematics teacher*. **84(**8), 623-630.

Mardia, K.V. and Jepp, P.E. 1972. Directional statistics. London, *Academic press Inc*.

Mardia, K.V., Hughes, G., Taylor, C. and Singh, H. (2008) Multivariate von Misses distribution with application to Bioinformatics. *Statistical society of Canada*. **36**(1), 99-109.

Martin, T.W. and SenGupta, A. (editors) 2011. Advances in directional and linear statistics. Physica-Verlag. London. Varies contributors.

Mekonnen, D. 2017. Application of circular ANOVA on Biological data: case study on Crimean-Congo Hemorrhage Fever.International annal of medicine. https://doi.org/10.24087/IAM.2017.1.3.67

Moore, B. 1980. A modification of Rayleigh test for vector data. *Boimatika*. **67**(1), 175-180.

Muller, JE, Tofler, GH. And Stone, PH. 1989. Circadian variation and triggers of onset of acute cardiovascular diseases. **79**(4), 733-743. doi.org/10.1161/01.CIR.79.4.733.

Mutwiri, R. M., Mwimbi, H. and Slotow, R. 2014. Approaches for testing uniformity hypothesis in angular of mega-herbivores. *International journal of science and research*.**5**(3).

Mutwiri,R.M., Mwimbi, H. and Slotow, R. 2013. Approaches for testing uniformity hypothesis in angular of mega-herbivores. *International journal of science and research*. **5**(3)

Norton, B.J. 1978. Karl Pearson and statistics: the social origin of scientific Innovation. *JSTOR* **8**(1), 3-34.

Onat, A. 2001. Risk factors and cardiovascular diseases in Turkey. *Journal of Atherosclerosis research*. **156**(1), 1-10.

Onat, A., Senocak, M., Şurdum-Avci, G. and Ornek, E.1993 Prevalence of coronary heart disease in Turkish Adults. *International journal of Cardiology*. **39**(1), 23-31.

ONAT, A., Şurdumavci, G., Şenocak,M.,Örnek,E.,Gözükara,Y., Karaaslan, Y., Özişik, U., İşler, M, Tabak, F. And Özcan, R. 1991. Survey on Prevalence of Cardiac Disease and its Risk Factors in Adults in Turkey: 3. Prevalence of Heart Diseases. *Turk kardiyolog derneği ARŞ*. **19**(1), 26-33.

Ozarowska, A. 2013. A new approach to evaluate multimodal orientation behavior of migratory passerine birds recorded in circular orientation cages. *Journal of experimental Biology*.**216**, 4038-4048.

Pauen, K. and Ivanova, G. 2013. Multiple Circular-Circular correlation coefficients for the quantification of phase synchronization process the brain. Biomedtech (berl). 58(2). 141-155

Pewsey, A., Neuhausen, M. and Ruxton, G. D. 2013. Circular statistics in R. UK, *Oxford University press.*

Pewsey, A., Neuhausen, M. and Ruxton, G. D. 2013. Circular statistics in R. UK, *Oxford University press.*

Poltavskiy, E. 2016. Birth month and cardiovascular disease risk association: Is meaningfulness in the eye of beholder. doi:10.5210/ojphi.v8i2.6643. *Online journal of public Health informatics*.

Ravindran, P. and Ghosh, S.K. 2011. Bayesian analysis of circular data using wrapped distributions. *Journal of statistical theory and practice*. **5**(4), 547-561.

Rivest, L.P. 1997. A decentered predictor for circular-circular regression. *Biomatika*. **84**(3), 717-726.

Robert, E. R. and David, P.R. 2016. Text book of family medicine. Ninth ed. ELSEVIER, 1110, Philadelphia, USA.

Roussas, G.G. 2003. Introduction to probability and statistical inference. *ELSEVIER SCIENCE (USA)*.

Russel, G.S. and Levtin, D.J. 1995. An expanded table of probability values for Rao's spacing test. *Communications in Statistics – simulation and computation*. **24**(4), 879-888.

Sengupta, A. 2004. On the constructions of probability distributions for directional data. *Bul.cal.Math.socienty*. **96**(2), 139-154.

Srinivas, M. and Rao, S. 2016. Importance of circular data in sport science. *ResearchGate*. **1**(2), 37-44.

Stanton, J. M. 2001. Galton, Pearson, and the peas: A brief History of linear regression for statistics instructors. Journal of statistics education. **9**(3).

Stephenson, D.B. and Doblas-Reyes, F.J. 2000. Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus*, **52**(3), 300-322.

Stuckey, E.M., Smith, T. and Chitnis, N. 2014. Seasonally dependent relationships between indicators of malaria transmission and disease provided by mathematical model simulations. *PLOS, Computational Biology*. 10(1371).

ÜNAL, B., ERGÖR, G., DİNÇ HORASAN, G. and SÖZMEN, K. 2013. Chronic disease and risk factors survey in Turkey. Republic of Turkey Ministry of Health, Public Health agency of Turkey.

Varalakshmi, T.V., Sundaram, G.G., Ezhilarasi, S., Indrani, B. and Suseela, N. 2004. Statistics. *Tamilnadu Textbook Corporation*, Chennai, India.

Wallis, D. 2005. History of Angle measurement. London.

APPENDIXES

Appendix I Circular plot of varying mean and concentration parameters

Appendix II Syntax of circular normal distribution with different concentration parameter kappa

Appendix III Density of a circular uniform distribution

Appendix IV Natavaram vivax rose diagram plot

Appendix V Generate bivariate circular-circular data

Appendix VI Critical *z* Values for the Rayleigh's Test

Appendix VII Standard Rao's spacing test critical values table

Appendix VIII Extended Rao's critical values table

Appendix IX Extended Rao's critical values table

Appendix I Circular plot of varying mean and concentration parameters

```
ff <- function(x) sqrt(x)/20
curve.circular(ff)
curve.circular(ff. to=6*pi. join=FALSE. nosort=TRUE. n=1001. modulo="asis".
shrink=1.2)
plot.function.circular(function(x) dvonmises(x. circular(0). 8). xlim=c(-2. 3.5).lwd=2.col=1.lty=1.
main="Circular Normal with Varying Mean direction and Kappa")
par(new=T)
plot.function.circular(function(x) dvonmises(x. circular(0). 6). xlim=c(-2. 3.5).lwd=2.col=2.lty=2)
par(new=T)
plot.function.circular(function(x) dvonmises(x. circular(0). 4). xlim=c(-2. 3.5).lwd=2.col=3.lty=3)
par(new=T)
plot.function.circular(function(x)dvonmises(x.circular(pi/2).8).xlim=c(-2.  3.5
).lwd=2.col=4..lty=1)
par(new=T)
plot.function.circular(function(x)dvonmises(x.circular(pi/2).6).xlim=c(-2.
3.5).lwd=2.col=5..lty=2)
par(new=T)
plot.function.circular(function(x)dvonmises(x.circular(pi/2).4).xlim=c(-2.
3.5).lwd=2.col=6..lty=3)
legend("topright". legend=c("mn.dr=0 & kappa=8"."mn.dr=0 & kappa=6"."mn.dr=0
&kappa=4"."mn.dr=pi/2 &kappa=8".
"mn.dr=pi/2 & kappa=6"."mn.dr=pi/2& kappa=4").
    col=c(1.2.3.4.5.6). lty=c(1.2.3.1.2.3))
```

Appendix II Syntax of circular normal distribution with different concentration parameter kappa

plot(function(x) dvonmises(circular(x). mu=circular(pi). kappa=3).

*type="l". lwd=2. col=1. main="Circular Normal Distribution with Different Kappa". xlab="Angular measurement".ylab="Density". from=0. to=2*pi)*

plot(function(x) dvonmises(circular(x). mu=circular(pi). kappa=2).

*type="l". lwd=2. col=2. main="Circular Normal Distribution with Different Kappa". xlab="angular measurement".ylab="Density". from=0. to=2*pi.add=TRUE)*

plot(function(x) dvonmises(circular(x). mu=circular(pi). kappa=1).

*type="l". lwd=2. col=3. main="Circular Normal Distribution with Different Kappa". xlab="angular measurement".ylab="Density". from=0. to=2*pi.add=TRUE)*

plot(function(x) dvonmises(circular(x). mu=circular(pi). kappa=0.5).

*type="l". lwd=2. col=4. main="Circular Normal Distribution with Different Dappa". xlab="angular measurement".ylab="Density". from=0. to=2*pi.add=TRUE)*

legend("topright". legend=c("Kappa=3". "Kappa=2". "Kappa=1". "Kappa=0.5"). col=1:4. lwd=2)

index 2 rose diagram of malaria data

dataC <- circular(petientday.type= "angles".units= "radians".template= "none".

+ modulo= "asis".zero= 0.rotation= "clock")

> x <- dataC

> y <- rose.diag(x. bins=18) # Points fall out of bounds.

> points(x. plot.info=y. stack=TRUE)

> y <- rose.diag(x. bins=12. prop=1.5. shrink=0.000002)

> points(x. plot.info=y. stack=TRUE)

> plot(x.main="rose diagram of malaria data")

> rose.diag(x. bins=12. add=TRUE. col=2)

> points(x. plot.info=y. stack=TRUE)

Appendix III Density of a circular uniform distribution

```
> data1 <- rcircularuniform(100. control.circular=list(units="degrees"))
> plot(data1)
> curve.circular(dcircularuniform. join=TRUE. xlim=c(-1.2. 1.2).
+ ylim=c(-1.2. 1.2). main="Density of a Circular Uniform Distribution")
```

Appendix IV Natavaram vivax rose diagram plot

vivaxC <- circular(dataC.type= "angles".units= "radians".template= "none".modulo= "asis".zero= 0.rotation= "clock")

> x <- vivaxC

> y <- rose.diag(x. bins=12. prop=1.5. shrink=0.000002)

> points(x. plot.info=y. stack=TRUE)

> plot(x.main="rose diagram of Plasmodium Vivax")

> rose.diag(x. bins=12. add=TRUE. col=3)

> points(x. plot.info=y. stack=TRUE)

Appendix V Generate bivariate circular-circular data

birthday <- rvm(10000, 5.024972 0.101588)

*attackday <- atan2(0.15*cos(data)*

*+ 0.25*sin(data).*

*0.35*sin(data))+ rvm(10000.5.024972.0.101588)*

Export generated data to excel (CSV)

Create (select) a directory where to save your generated data

write.csv (birthdays, file="exported_data.csv")

filename in r=birthdays. attackdays. file name in CSV=exported data

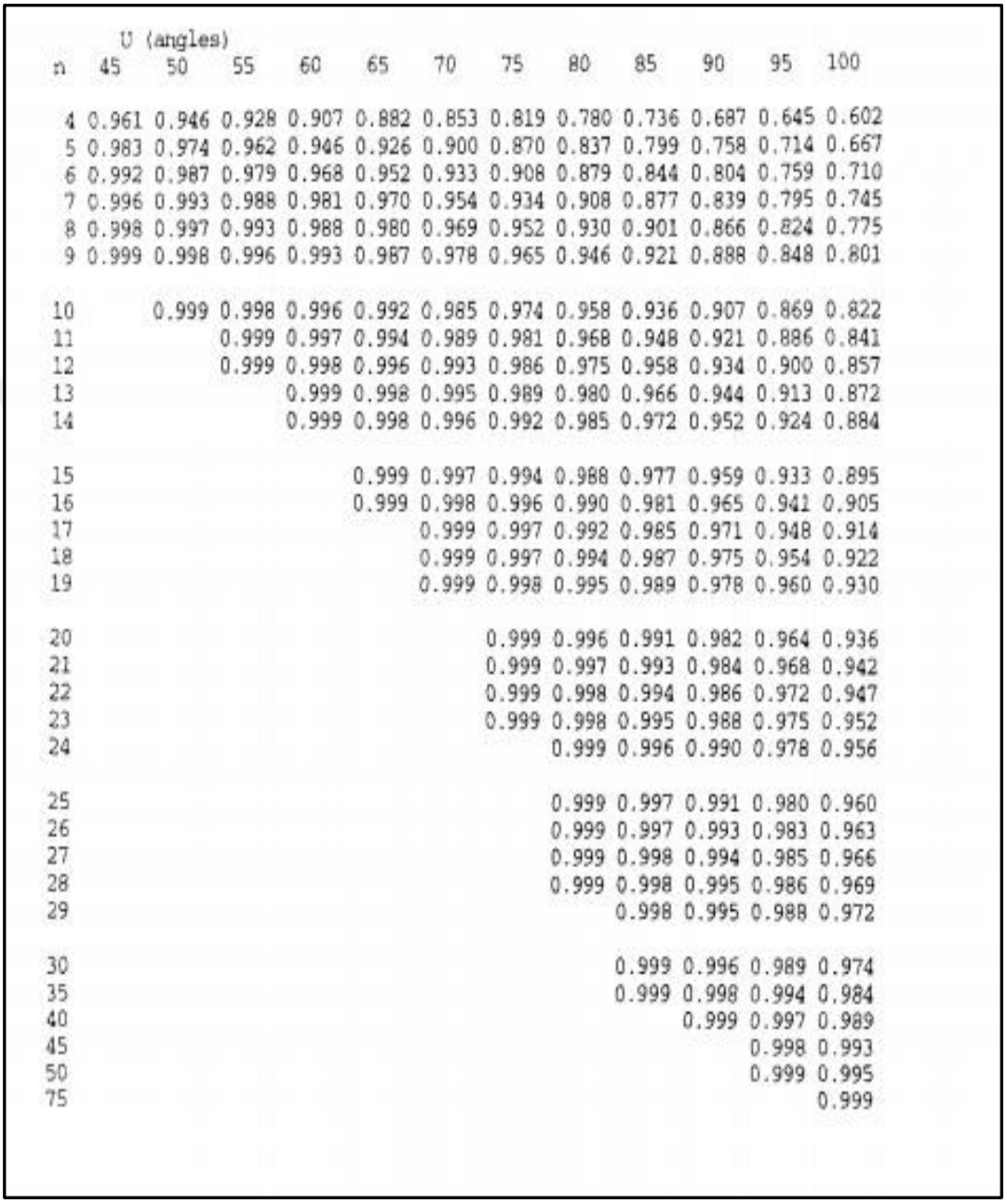
Appendix VI Critical *z* Values for the Rayleigh's Test

Taken from Zar. 1981 Table B.32

n	α: 0.50	0.20	0.10	0.05	0.02	0.01	0.005	0.002	0.001
6	0.734	1.639	2.274	2.865	3.576	4.058	4.491	4.985	5.297
7	0.727	1.634	2.278	2.885	3.627	4.143	4.617	5.181	5.556
8	0.723	1.631	2.281	2.899	3.665	4.205	4.710	5.322	5.743
9	0.719	1.628	2.283	2.910	3.694	4.252	4.780	5.430	5.885
10	0.717	1.626	2.285	2.919	3.716	4.289	4.835	5.514	5.996
11	0.715	1.625	2.287	2.926	3.735	4.319	4.879	5.582	6.085
12	0.713	1.623	2.288	2.932	3.750	4.344	4.916	5.638	6.158
13	0.711	1.622	2.289	2.937	3.763	4.365	4.947	5.685	6.219
14	0.710	1.621	2.290	2.941	3.774	4.383	4.973	5.725	6.271
15	0.709	1.620	2.291	2.945	3.784	4.398	4.996	5.759	6.316
16	0.708	1.620	2.292	2.948	3.792	4.412	5.015	5.789	6.354
17	0.707	1.619	2.292	2.951	3.799	4.423	5.033	5.815	6.388
18	0.706	1.619	2.293	2.954	3.806	4.434	5.048	5.838	6.418
19	0.705	1.618	2.293	2.956	3.811	4.443	5.061	5.858	6.445
20	0.705	1.618	2.294	2.958	3.816	4.451	5.074	5.877	6.469
21	0.704	1.617	2.294	2.960	3.821	4.459	5.085	5.893	6.491
22	0.704	1.617	2.295	2.961	3.825	4.466	5.095	5.908	6.510
23	0.703	1.616	2.295	2.963	3.829	4.472	5.104	5.922	6.528
24	0.703	1.616	2.295	2.964	3.833	4.478	5.112	5.935	6.544
25	0.702	1.616	2.296	2.966	3.836	4.483	5.120	5.946	6.559
26	0.702	1.616	2.296	2.967	3.839	4.488	5.127	5.957	6.573
27	0.702	1.615	2.296	2.968	3.842	4.492	5.133	5.966	6.586
28	0.701	1.615	2.296	2.969	3.844	4.496	5.139	5.975	6.598
29	0.701	1.615	2.297	2.970	3.847	4.500	5.145	5.984	6.609
30	0.701	1.615	2.297	2.971	3.849	4.504	5.150	5.992	6.619
32	0.700	1.614	2.297	2.972	3.853	4.510	5.159	6.006	6.637
34	0.700	1.614	2.297	2.974	3.856	4.516	5.168	6.018	6.654
36	0.700	1.614	2.298	2.975	3.859	4.521	5.175	6.030	6.668
38	0.699	1.614	2.298	2.976	3.862	4.525	5.182	6.039	6.681
40	0.699	1.613	2.298	2.977	3.865	4.529	5.188	6.048	6.692
42	0.699	1.613	2.298	2.978	3.867	4.533	5.193	6.056	6.703
44	0.698	1.613	2.299	2.979	3.869	4.536	5.198	6.064	6.712
46	0.698	1.613	2.299	2.979	3.871	4.539	5.202	6.070	6.721
48	0.698	1.613	2.299	2.980	3.873	4.542	5.206	6.076	6.729
50	0.698	1.613	2.299	2.981	3.874	4.545	5.210	6.082	6.736
55	0.697	1.612	2.299	2.982	3.878	4.550	5.218	6.094	6.752
60	0.697	1.612	2.300	2.983	3.881	4.555	5.225	6.104	6.765
65	0.697	1.612	2.300	2.984	3.883	4.559	5.231	6.113	6.776
70	0.696	1.612	2.300	2.985	3.885	4.562	5.235	6.120	6.786
75	0.696	1.612	2.300	2.986	3.887	4.565	5.240	6.127	6.794
80	0.696	1.611	2.300	2.986	3.889	4.567	5.243	6.132	6.801
90	0.696	1.611	2.301	2.987	3.891	4.572	5.249	6.141	6.813
100	0.695	1.611	2.301	2.988	3.893	4.575	5.254	6.149	6.822
120	0.695	1.611	2.301	2.990	3.896	4.580	5.262	6.160	6.837
140	0.695	1.611	2.301	2.990	3.899	4.584	5.267	6.168	6.847
160	0.695	1.610	2.301	2.991	3.900	4.586	5.271	6.174	6.855
180	0.694	1.610	2.302	2.992	3.902	4.588	5.274	6.178	6.861
200	0.694	1.610	2.302	2.992	3.903	4.590	5.276	6.182	6.865
300	0.694	1.610	2.302	2.993	3.906	4.595	5.284	6.193	6.879
500	0.694	1.610	2.302	2.994	3.908	4.599	5.290	6.201	6.891
∞	0.6931	1.6094	2.3026	2.9957	3.9120	4.6052	5.2983	6.2146	6.9078

(http://webspace.ship.edu/pgmarr/Geo441/Tables/Rayleighs%20z%20Table.pdf)

Appendix VII Standard Rao's spacing test critical values table



	U (angles)											
n	45	50	55	60	65	70	75	80	85	90	95	100
4	0.961	0.946	0.928	0.907	0.882	0.853	0.819	0.780	0.736	0.687	0.645	0.602
5	0.983	0.974	0.962	0.946	0.926	0.900	0.870	0.837	0.799	0.758	0.714	0.667
6	0.992	0.987	0.979	0.968	0.952	0.933	0.908	0.879	0.844	0.804	0.759	0.710
7	0.996	0.993	0.988	0.981	0.970	0.954	0.934	0.908	0.877	0.839	0.795	0.745
8	0.998	0.997	0.993	0.988	0.980	0.969	0.952	0.930	0.901	0.866	0.824	0.775
9	0.999	0.998	0.996	0.993	0.987	0.978	0.965	0.946	0.921	0.888	0.848	0.801
10		0.999	0.998	0.996	0.992	0.985	0.974	0.958	0.936	0.907	0.869	0.822
11			0.999	0.997	0.994	0.989	0.981	0.968	0.948	0.921	0.886	0.841
12			0.999	0.998	0.996	0.993	0.986	0.975	0.958	0.934	0.900	0.857
13				0.999	0.998	0.995	0.989	0.980	0.966	0.944	0.913	0.872
14				0.999	0.998	0.996	0.992	0.985	0.972	0.952	0.924	0.884
15					0.999	0.997	0.994	0.988	0.977	0.959	0.933	0.895
16					0.999	0.998	0.996	0.990	0.981	0.965	0.941	0.905
17						0.999	0.997	0.992	0.985	0.971	0.948	0.914
18						0.999	0.997	0.994	0.987	0.975	0.954	0.922
19						0.999	0.998	0.995	0.989	0.978	0.960	0.930
20							0.999	0.996	0.991	0.982	0.964	0.936
21							0.999	0.997	0.993	0.984	0.968	0.942
22							0.999	0.998	0.994	0.986	0.972	0.947
23							0.999	0.998	0.995	0.988	0.975	0.952
24								0.999	0.996	0.990	0.978	0.956
25								0.999	0.997	0.991	0.980	0.960
26								0.999	0.997	0.993	0.983	0.963
27								0.999	0.998	0.994	0.985	0.966
28								0.999	0.998	0.995	0.986	0.969
29									0.998	0.995	0.988	0.972
30									0.999	0.996	0.989	0.974
35									0.999	0.998	0.994	0.984
40										0.999	0.997	0.989
45											0.998	0.993
50											0.999	0.995
75												0.999

Standard Rao's spacing test critical values table

Appendix VIII Extended Rao's critical values table

	U											
n	105	110	115	120	125	130	135	140	145	150	155	160
4	0.559	0.515	0.471	0.428	0.386	0.345	0.305	0.267	0.230	0.197	0.166	0.138
5	0.617	0.566	0.514	0.461	0.410	0.360	0.313	0.270	0.233	0.199	0.169	0.141
6	0.657	0.601	0.544	0.488	0.434	0.381	0.330	0.283	0.239	0.199	0.163	0.132
7	0.691	0.634	0.574	0.514	0.453	0.394	0.338	0.286	0.239	0.196	0.160	0.128
8	0.721	0.661	0.598	0.533	0.469	0.406	0.346	0.290	0.239	0.193	0.154	0.120
9	0.746	0.685	0.619	0.551	0.483	0.415	0.351	0.291	0.237	0.189	0.148	0.114
10	0.768	0.706	0.638	0.567	0.495	0.423	0.355	0.292	0.235	0.185	0.143	0.108
11	0.787	0.724	0.655	0.582	0.506	0.431	0.359	0.292	0.232	0.181	0.137	0.101
12	0.804	0.741	0.671	0.595	0.516	0.437	0.362	0.292	0.230	0.176	0.132	0.096
13	0.819	0.757	0.685	0.607	0.525	0.443	0.364	0.291	0.227	0.172	0.126	0.090
14	0.833	0.771	0.698	0.618	0.533	0.448	0.366	0.290	0.224	0.167	0.121	0.085
15	0.846	0.784	0.710	0.628	0.541	0.453	0.368	0.289	0.221	0.163	0.116	0.080
16	0.857	0.796	0.722	0.638	0.548	0.457	0.369	0.288	0.218	0.158	0.111	0.075
17	0.868	0.807	0.732	0.647	0.555	0.461	0.370	0.287	0.214	0.154	0.107	0.071
18	0.877	0.817	0.742	0.656	0.562	0.465	0.371	0.286	0.211	0.150	0.102	0.067
19	0.886	0.826	0.752	0.664	0.568	0.469	0.372	0.284	0.208	0.146	0.098	0.063
20	0.894	0.835	0.761	0.672	0.574	0.472	0.373	0.283	0.205	0.142	0.094	0.060
21	0.901	0.844	0.769	0.680	0.579	0.475	0.374	0.281	0.202	0.138	0.090	0.056
22	0.908	0.852	0.777	0.687	0.585	0.478	0.374	0.279	0.199	0.135	0.087	0.053
23	0.914	0.859	0.785	0.694	0.590	0.481	0.374	0.278	0.196	0.131	0.083	0.050
24	0.920	0.866	0.792	0.700	0.595	0.484	0.375	0.276	0.193	0.128	0.080	0.047
25	0.925	0.872	0.799	0.707	0.600	0.486	0.375	0.274	0.190	0.124	0.077	0.045
26	0.930	0.878	0.806	0.713	0.604	0.489	0.375	0.273	0.187	0.121	0.074	0.042
27	0.935	0.884	0.812	0.719	0.609	0.491	0.375	0.271	0.184	0.118	0.071	0.040
28	0.939	0.890	0.818	0.724	0.613	0.493	0.376	0.269	0.181	0.115	0.068	0.038
29	0.943	0.895	0.824	0.730	0.617	0.496	0.376	0.268	0.179	0.112	0.065	0.035
30	0.946	0.900	0.829	0.735	0.622	0.498	0.376	0.266	0.176	0.109	0.063	0.034
35	0.961	0.920	0.854	0.759	0.640	0.508	0.375	0.258	0.163	0.095	0.051	0.025
40	0.972	0.936	0.874	0.780	0.657	0.516	0.374	0.249	0.152	0.084	0.042	0.019
45	0.979	0.949	0.891	0.798	0.672	0.523	0.373	0.241	0.141	0.074	0.035	0.015
50	0.985	0.959	0.905	0.815	0.685	0.530	0.372	0.234	0.131	0.065	0.029	0.011
75	0.997	0.985	0.952	0.874	0.740	0.557	0.363	0.200	0.093	0.036	0.011	0.003
100	0.999	0.995	0.974	0.912	0.779	0.577	0.353	0.173	0.067	0.020	0.005	0.001
150		0.999	0.992	0.955	0.836	0.607	0.335	0.132	0.036	0.007	0.001	
200			0.998	0.976	0.875	0.631	0.318	0.102	0.020	0.002		
300				0.993	0.924	0.668	0.289	0.063	0.006			
400				0.998	0.953	0.697	0.265	0.039	0.002			
500				0.999	0.970	0.721	0.245	0.025	0.001			
600					0.980	0.742	0.227	0.016				
700					0.987	0.760	0.210	0.011				

Appendix IX Extended Rao's critical values table

n	U 165	170	175	180	185	190	195	200	205	210	215	220
4	0.113	0.092	0.075	0.063	0.053	0.044	0.036	0.029	0.024	0.019	0.014	0.011
5	0.116	0.094	0.076	0.060	0.046	0.035	0.026	0.020	0.015	0.011	0.008	0.006
6	0.106	0.084	0.066	0.052	0.040	0.030	0.022	0.016	0.011	0.008	0.005	0.004
7	0.100	0.077	0.059	0.044	0.032	0.024	0.017	0.012	0.008	0.006	0.004	0.002
8	0.093	0.070	0.052	0.038	0.027	0.019	0.013	0.009	0.006	0.004	0.002	0.002
9	0.086	0.063	0.046	0.032	0.023	0.015	0.010	0.007	0.004	0.003	0.002	0.001
10	0.079	0.057	0.040	0.028	0.019	0.012	0.008	0.005	0.003	0.002	0.001	0.001
11	0.073	0.052	0.035	0.024	0.016	0.010	0.006	0.004	0.002	0.001	0.001	
12	0.068	0.047	0.031	0.020	0.013	0.008	0.005	0.003	0.002	0.001		
13	0.063	0.042	0.027	0.017	0.011	0.006	0.004	0.002	0.001	0.001		
14	0.058	0.038	0.024	0.015	0.009	0.005	0.003	0.002	0.001			
15	0.053	0.034	0.021	0.013	0.007	0.004	0.002	0.001	0.001			
16	0.049	0.031	0.019	0.011	0.006	0.003	0.002	0.001				
17	0.046	0.028	0.017	0.009	0.005	0.003	0.001	0.001				
18	0.042	0.025	0.015	0.008	0.004	0.002	0.001					
19	0.039	0.023	0.013	0.007	0.004	0.002	0.001					
20	0.036	0.021	0.011	0.006	0.003	0.001	0.001					
21	0.033	0.019	0.010	0.005	0.002	0.001						
22	0.031	0.017	0.009	0.004	0.002	0.001						
23	0.028	0.015	0.008	0.004	0.002	0.001						
24	0.026	0.014	0.007	0.003	0.001	0.001						
25	0.024	0.013	0.006	0.003	0.001							
26	0.023	0.011	0.005	0.002	0.001							
27	0.021	0.010	0.005	0.002	0.001							
28	0.019	0.009	0.004	0.002	0.001							
29	0.018	0.009	0.004	0.002	0.001							
30	0.017	0.008	0.003	0.001								
35	0.011	0.005	0.002	0.001								
40	0.008	0.003	0.001									
45	0.005	0.002	0.001									
50	0.004	0.001										
75	0.001											

Curriculum Vitae

Name and Surname : Desta Firdu Mekonnen
Place of birth : Dabre Sina
Date of birth : 18 / 01 / 1983
Marital status : Single
Foreign language **:** English, Turkish

Education (Institute and year)

High school : Adama Preparatory High School (1999-2003)

Bachelor : Mekelle University, Faculty of Agriculture and Natural Resources, Department of Animal, Range and Wildlife Science (ARWS) (2003 – 2007)

Work experience and year

Academic Staff; Mizan-Tepi University, Faculty of Agriculture and Natural Resources, Department of Natural Resources (2008-2010)

Articles

Desta Firdu Mekonnen and Ensar Başpınar. Application of Circular Regression Analysis on Biological Data: Case Study on Malaria Cases in District of Visakhapatnam, India. International Journal of Current Research 2017, 9 (4); 48594-48600.

Desta Firdu Mekonnen and Ensar Başpınar. Application of Circular ANOVA on Biological Data: Case Study on Crimean-Congo Hemorrhage Fever Cases in Turkey. International Annals of Medicine 2017, 1(3).

Other courses

Ankara University TÖMER Turkish and Foreign Language Teaching, Research and Application Centre, "Turkish courses" April 2010-September 2010 (4 months)

ANKARA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DOKTORA TEZİ

DAİRESEL REGRESYON ANALİZİNİN BİYOLOJİK VERİLERDE UYGULANMASI

Desta Firdu MEKONNEN

ZOOTEKNİ ANABİLİM DALI

ANKARA
2017

ÖZET

Doktora Tezi

DAİRESEL REGRESYON ANALİZİNİN BİYOLOJİK VERİLERDE UYGULANMASI

Desta Firdu MEKONNEN

Ankara Üniversitesi
Fen Bilimleri Enstitüsü
Zootekni Anabilim Dalı

Tez Danışmanı: Prof. Dr. Ensar BAŞPINAR

Açısal ölçümler içeren çok sayıda doğal ve yapay senaryolar bulunmaktadır. Günlük faaliyetlerimizden metabolizmamıza kadar hepsinin doğasında dairesellik vardır. Öte yandan, dairesel istatistikler ortaya çıkana kadar uygun teknikler kullanılarak bu senaryolar çalışılmamıştır.)

Dairesel istatistikler, istatistiksel analiz ve modellemenin yeni bir alanı olarak görülebilir. Bazı istatistiksel modeller verilerin dairesellik özelliklerini dikkate alırken bunun aksine bazı yöntemler, eldeki veriler açıkça dairesellik gösterseler bile tamamen daireselliği göz ardı etmektedir.

Verilerin daireselliğini dikkate alan bazı modeller olmasına karşın bu modeller oldukça azdır. Bu alanlardan biri dairesel regresyon analizidir. Regresyon analizi, gelişmiş bir yöntem olmasına karşın dairesel verilere gelince hala geliştirilmesi gereken başlangıç seviyesindedir. Bu alana trend konu denilmesinin nedeni budur. Bu alanda (bilgimize göre) evrensel olarak kabul görmüş tek bir model yoktur. Nitekim, son 50 yılda birkaç veri türü için geliştirilmiş çok az dairesel regresyon modeli vardır.Bu modellerin biyolojik verilere uygulanabilirliği hala bir soru işaretindedir. (Bu modellerin biyolojik verilere uygulanabilirliği hakkında sorular vardır.)

Bu tezde, farklı biyolojik veriler üzerinde dairesel regresyon modelleri uygulanmış ve bu modellerin kabul edilebilirlik, açıklık, anlamlılık ve etkinlik bakımından güçlü ve zayıf yönlerini araştırılmıştır.

Dairesel regresyon analizi yeni bir konu olduğundan bu yöntem ile analizlere geçilmeden önce yöntemin temelini oluşturan dairesel veriler, dairesel dağılımlar, dairesel tanımlayıcı istatistikler ve dairesel üniformite testleri yapılarak incelenmiştir.

Farklı zorluklarla karşı karşıya kalmamıza rağmen elimizdeki verileri kullanarak, biyolojik çalışmaların farklı alanlarında dairesel regresyon analizinin nispeten daha iyi anlaşılmasına katkıda bulunduk.

Nisan 2017, 21 Sayfa

Anahtar kelimeler: Dairesel veriler, Dairesel dağılımlar, Dairesel uniformite test, Dairesel regresyon, Sıtma, Kırım-Kongo kanamalı ateşi, Panik atak, Kalp krizi.

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

ÇİZELGELER DİZİNİ

1. GİRİŞ

Açısal ölçümler içeren çok sayıda doğal ve yapay senaryolar bulunmaktadır. Günlük faaliyetlerimizden metabolizmamıza kadar hepsinin doğasında dairesellik vardır. Öte yandan, dairesel istatistikler ortaya çıkana kadar uygun teknikler kullanılarak bu senaryolar çalışılmamıştır.

Dairesel istatistikler, istatistiksel analiz ve modellemenin yeni bir alanı olarak görülebilir. Bazı istatistiksel modeller verilerin dairesellik özelliklerini dikkate alırken bunun aksine bazı yöntemler, eldeki veriler açıkça dairesellik gösterseler bile tamamen daireselliği göz ardı etmektedir.

Verilerin daireselliğini dikkate alan bazı modeller olmasına karşın bu modeller oldukça azdır. Bu alanlardan biri dairesel regresyon analizidir. Regresyon analizi, gelişmiş bir yöntem olmasına karşın dairesel verilere gelince hala geliştirilmesi gereken bir seviyededir. Bu alana trend konu denilmesinin nedeni de budur. Bu alanda elde edilen bilgilere göre hala evrensel olarak kabul görmüş tek bir model yoktur. Nitekim son 50 yılda birkaç veri türü için geliştirilmiş çok az dairesel regresyon modeli bulunmaktadır. Bu modellerin biyolojik verilere uygulanabilirliği ise hala bir soru işaretidir.

Bu tezde, farklı biyolojik veriler üzerinde dairesel regresyon modelleri uygulanmış ve bu modellerin kabul edilebilirlik, açıklık, anlamlılık ve etkinlik bakımından güçlü ve zayıf yönlerini araştırılmıştır.

Dairesel regresyon analizi yeni bir konu olduğundan bu yöntem ile analizlere geçilmeden önce yöntemin temelini oluşturan dairesel veriler, dairesel dağılımlar, dairesel tanımlayıcı istatistikler ve dairesel üniformite testleri yapılarak incelenmiştir.

Bu tezin hazırlanması sırasında farklı zorluklarla karşı karşıya kalınmasına rağmen eldeki mevcut veriler kullanılarak, biyolojik çalışmaların farklı alanlarında dairesel regresyon analizinin nispeten daha iyi anlaşılmasına katkılarda bulunulmuştur.

Dairesel istatistiksel yöntemler kullanılarak analiz edilebilmesi gereken doğal ve yapay senaryolar bol olsa bile bu senaryoları, zaman serisi analiz gibi, doğrusal analiz yöntemleri kullanılarak analizi yapılmıştır. Örneğin, birçok mevsimlik değişim senaryolar zaman serileri analizi yerine dairesel yöntemler kullanılarak analiz edilmeye uygundur. Ancak olayların mevsimsel olarak değişimi için henüz evrensel olarak kabul edilmiş dairesel analiz yöntemi yoktur.

Özet bölümünde belirtildiği gibi, döngüsel istatistiksel analiz yöntemleri, diğer yöntemlerine karşı "primitif düzeyinde" düşünülür. Konuyla ilgili bazı yöntemler ve modeller vardır, ancak bu yöntemler ve modeller arası dayanıklılık yoktur.

1.1 Tezin Amacı

Dairesel istatistiksel analiz yöntemleri bilinen diğer istatistiksel yöntemlerden birçok açıdan farklı olduğu için farklı istatistiksel ve biyolojik problemlerin dairesel istatistiksel analiz yöntemleri kullanılarak incelenmesi ve değerlendirilmesi bu tezin ana amacı olmuştur.

Bu tezde, dairesel istatistiksel analiz yöntemleri ile ilgili olan farklı istatistiksel ve biyolojik problemler incelenerek değerlendirilmiştir. Çünkü dairesel istatistiksel analiz yöntemleri bilinen diğer istatistiksel yöntemlerden birçok açıdan farklıdır.

Bu tezde;

- Farklı biyolojik verilerde dairesel regresyon analizi yaparak bu modellerin hem matematiksel hem de biyolojik uygulanabilirliği ve anlamlılığını değerlendirmek,
- Dairesel regresyon analizine geçmeden önce yöntemin temelini oluşturan dairesel veriler, dairesel dağılımlar, dairesel tanımlayıcı istatistikler ve dairesel üniformite testlerinin yapılması,
- Dairesel regresyon yöntemini doğrusal yöntem ile karşılaştırıldığında zayıf ve güçlü taraflarının incelenmesi,

- Son olarak da, dairesel regresyon yöntem ve modellerinin sınırlamaları üzerinde durularak konuyla ilgili gelecekteki çalışmalara katkı sağlamak hedeflenmiştir.

2. MATERYAL VE METOT

2.1 Materyal

Bu tez çalışması yapılırken karşılaşılan en büyük problem dairesel regresyon analizi yapmak için gerekli olan uygun verilerin bulunamaması olmuştur. Ancak karşılaşılan bu problemi aşmak için farklı araştırmacılar tarafından yapılan çalışmalarda elde edilen ve farklı analiz yöntemleri kullanılarak incelenen veriler, dairesel analizler için uygun hale dönüştürülerek materyal olarak kullanılmıştır. Ayrıca dairesel analizler için uygun veriler "R simülasyon programı" ile elde edilerek bu tezde materyal olarak kullanılmıştır.

Bulunan verilerin çoğunun dairesel analiz için uygun olmaması sebebiyle analizlere geçilmeden önce veriler, dairesel analizler için uygun hale getirilerek değiştirilmiştir. Bu tezde kullanılan verilerin elde edilmesinde 4 farklı kaynaktan yararlanılmıştır. Bunlar;

1. Indiana'da sıtma hastalık verileri
2. Türkiye'de Kırım-Kongo kanamalı ateşi verileri
3. Tayvan'da panik-atak verileri
4. Ankara kalp kriz hastaları dayalı olarak kalp hastalık benzetim verileridir.

2.2 Metot

Dairesel regresyon analiz yöntemi oldukça yeni bir konu olduğundan analizlere geçmeden önce tanımlayıcı istatistikleri bilmek analizi kolaylaştırmada büyük bir avantaj sağlamaktadır. Bu nedenle, dairesel regresyon analizine geçmeden istatistiksel verilerin yapıları ve bazı tanıtıcı istatistikleri belirlenmiştir.

2.2.1 Dairesel veriler

Dairesel analizi diğer istatistik analiz yöntemlerinden ayıran ilk özelliği kullanılan veri türüdür. Dairesel veriler çoğunlukla yön ve dairesel zaman ölçüleri olmak üzere iki farklı ölçüm çeşidi kullanılarak elde edilmektedir. Kuşların göç yönü, rüzgâr yönü, salgın hastalıkların yönü, dünyanın enlem ve boylam derece değerleri **yön ölçümlerinden** birkaçına örnek olarak verilebilir. 24 saat içinde bir hastanede karşılaşılan acil durum vakaları, bir yıl içinde bir ülkede belirlenen kaza sonucu gerçekleşen ölüm sayısı, belirli bir zaman aralığında tekrarlanan salgın hastalıkların görülme oranları ise **dairesel zaman** ölçümlerine örnektir.

Dairesel verilerin doğal ölçüsü derece veya radyandır. Genel olarak radyan dereceden daha iyi bir ölçüm olarak kabul edilmektedir. Dolayısıyla dairesel zaman ölçümleri, aşağıdaki dönüşüm denklemi ile dereceye ya da radyana dönüştürülmesi gerekmektedir.

$$\theta = \frac{2\pi * x}{y}$$

θ: dairesel zaman ölçümleri

x: radyana dönüştürülen zaman ölçüleri

y: bir zaman ölçüsünün tam bir daireseli

Dairesel verilerin diğer bir özelliği ise doğal bir sıfır noktası ve yön tercihinin olmamasıdır. Doğudan 45° yönü bir matematikçi için sıfır kabul edilirken bir biyolog tarafından kuzeyden 60° sıfır kabul edilebilir. Bu yüzden, sıralama tabanlı istatistikler yapılırken son derece dikkatli olunmalıdır. Dairesel veriler elde edilirken dikkat edilmesi gereken tek şart tüm ölçümlerin aynı sıfır noktasından başlayarak aynı yöne gitmesidir.

2.2.2 Dairesel dağlımlar

Doğrusal istatistiksel analizlerde ve çıkarımlarda, verilerin dağılımını tanımlamak analiz ve karar vermek için önemli bir noktadır. Verilerin dağılımını bilmek doğru istatistiksel yöntemi seçmeye yardımcı olur. Verilerin dağılımının bilinmesinin doğru analiz yöntemini seçmedeki yardımı dairesel istatistiksel analizler için de geçerlidir. Herhangi bir dairesel analiz yapılmadan önce, verilerin dağılımını bilmek çok önemlidir. Dairesel değişkenlerin dağılımını bilmek multimodalite olup olmadığını değerlendirmeye yardımcı olur. Dairesel dağılımda multimodalite olduğu durumlarda uygulanacak istatistiksel yöntemler değişir.

Dairesel dağılım, toplam olasılığı birim çemberin çevresi üzerinde yoğunlaşan bir olasılık dağılımdır. Dairesel dağlımlar; kesikli dairesel dağılım ve sürekli dairesel dağılım olmak üzere ikiye ayrılır.

3. SONUÇ ve TARTIŞMA

Visakhapatnam şehrinde gözlemlenen *Plasmodium Falciparum* hastalığına ait yıllara ve aylara ait veriler çizelge 3. 1' de özetlenmiştir.

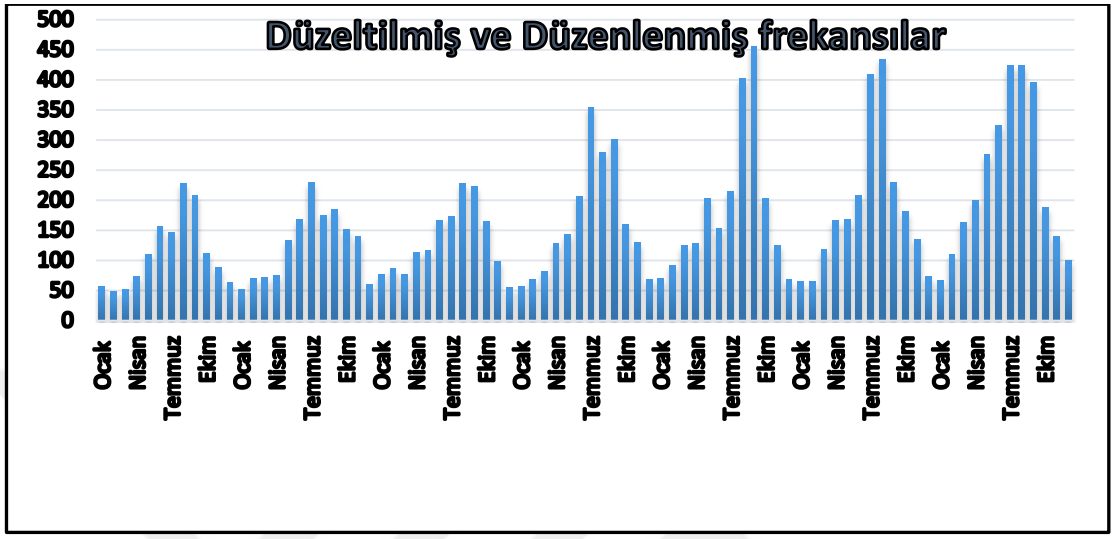
Çizelge 3.1 Plasmodium *Falciparum* verileri

Ay	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
Ocak	58	52	77	57	71	66	67
Şubat	44	64	79	62	84	60	100
Mart	53	72	78	83	127	120	166
Nisan	72	73	111	125	125	163	196
Mayıs	112	135	118	145	206	170	280
Haziran	154	165	164	202	150	204	318
Temmuz	149	233	176	360	217	416	430
Ağustos	231	178	231	284	408	441	431
Eylül	204	182	219	295	448	225	388
Ekim	113	154	167	162	206	184	191
Kasım	86	137	97	128	123	133	138
Aralık	65	61	55	70	69	74	102

Çizelge 3.2 Düzeltilmiş ve Düzenlenmiş veriler

No	**Aylar**	**Günler**	**Düzeltmeler**	**Derece**	**Radyan**	**Pf**
1	Ocak	31	0.967741935	1	**0.017453293**	**57**
2	Şubat	28	1.071428571	30	**0.523598776**	**44**
3	Mart	31	0.967741935	60	**1.047197551**	**53**
4	Nisan	30	1.000000000	90	**1.570796327**	**72**
5	Mayıs	31	0.967741935	120	**2.094395102**	**112**
6	Haziran	30	1.000000000	150	**2.617993878**	**154**
7	Temmuz	31	0.967741935	180	**3.141592654**	**149**
8	Ağustos	31	0.967741935	210	**3.665191429**	**231**
9	Eylül	30	1.000000000	240	**4.188790205**	**204**
10	Ekim	31	0.967741935	270	**4.71238898**	**113**
	...	…	…	…	**…**	
	…	…	…	…	**….**	
79	Temmuz	31	0.967741935	180	**3.141593**	**430**
80	Ağustos	31	0.967741935	210	**3.665191**	**431**
81	Eylül	30	1.000000000	240	**4.188790**	**388**
82	Ekim	31	0.967741935	270	**4.712389**	**191**
83	Kasım	30	1.00000000	300	**5.235988**	**138**
84	Aralık	31	0.967741935	330	**5.759587**	**102**

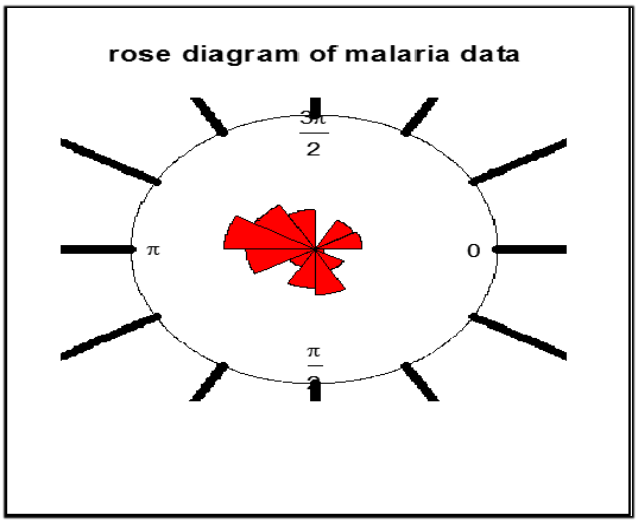
Çizelge 3.1'de özetlenen verilerin dairesel tanıtıcı istatistiklerini hesaplamadan önce grafiklerine görsel olarak bakmak önemlidir.



Şekil 3.1 7 yıllık sürede gözlenen aylık sıtma vakaları

Şekil 3.1'de 2005-2011 yılları arasındaki 7 yıllık sürede aylara göre tespit edilen sıtma vakalarını göstermektedir. Şekil incelendiğinde sıtma vakalarının uniform bir dağılım göstermediği ve yıl boyunca bir eğilim gösterdiği görülmektedir. Şekilde de görülebildiği gibi bir ayda tepe ve dip noktalarının varlığı bu verilerde multimodalite olmadığının bir göstergesidir. Bir veride multimodalite olmaması verilerin dairesel analiz için uygun olduğunu gösterir. Eğer verilerde multimodalite var ise o verilere dairesel analiz uygulanamaz.

Şekil 3.1 verilerin daireselliğini göstermediği için verilerin daireselliğini gösteren daha uygun şekil gerekmektedir. Ancak dairesel verilerin gösterimine uygun çok az sayıda veri grafiği bulunmaktadır. Genellikle, dairesel verilerin daireselliğini göstermek için kullanılan üç grafik türü bulunmaktadır. Bunlar "Dairesel nokta grafiği (circular dot plot)", "Dairesel histogram (circular histogram)" ve "Dairesel gül diyagramı (circular rose diagram)" olarak bilinir. Bu grafiklerden yaygın olarak kullanılanı ise dairesel gül diyagramıdır. Sıtma vakalarına ait dairesel gül diyagramı grafiği şekil 2'de gösterilmiştir.



Şekil 3.2 Sıtma vakalarına ait gül diyagramı

Şekil 3.2 aylara göre 12 frekansa sahip 12 bölüme (kutuya) ayrılmıştır. Bu veri grafiğinde, aylık sıtma vakaları arasındaki değişimi gösteren farklı alan büyüklükleri vardır. En az sıtma vakası, sıfır noktası Ocak ayı alınarak "saat yönünde" sayım yapıldığında 10. kutuda görülmektedir. Aylar boyunca sıtma vakalarında herhangi bir farklılık olmasaydı, kutular aynı boyutlarda olacak ve bir dairesel alan içerisinde hiçbir pürüzün olmadığı alt bir dairesel alan oluşmuş olacaktı.

3.1 Dairesel Tanımlayıcı İstatistikler

Dairesel tanımlayıcı istatistikler diğer tanımlayıcı istatistiklerden çok farklıdır. *Plasmodium Falciparum* gibi veriler frekans veri tipi (gruplanmış veriler) olduğundan tanımlayıcı istatistikler hesaplanırken bu durumun dikkate alınması gerekmektedir. Bu hesaplamada birinci adım olarak zaman verileri dereceye ya da radyana dönüştürülmesi gerekmektedir. Buna göre sıtma vakalarına ait verilerin radyana dönüştürülerek yapılan düzeltmeler Çizelge 3.2'de özetlenmiştir.

Sıtma vakalarına ait dairesel tanıtıcı istatistiklerin hesaplanması için çizelge 3.2'de yer alan altıncı ve yedinci sütunda yer alan frekanslar kullanılmıştır.

3.1.1 Dairesel ortalama yönü

Çizelge 3.2'de verilen sıtma vakalarının ortalama yönünün belirlenmesi için aşağıdaki ifadeler kullanılarak ortalama yön hesaplanmıştır.

$$n = \sum f_i = 13689$$

$$\overline{C}_n = \frac{1}{n}\sum_{i=1}^{k} f_i(\cos\theta_i) = \frac{1}{13689}\left(57(0.9998) + 48(0.8660) + \ldots + 100(0.8660)\right) = \text{-}0.328$$

$$\overline{C}_n = \frac{1}{n}\sum_{i=1}^{k} f_i(\sin\theta_i) = \frac{1}{13689}\left(57(0.0175) + 48(0.500) + \ldots + 100(-0.5000)\right) = \text{-}0.111$$

$$\overline{R} = \sqrt{\overline{C}^2 + \overline{S}^2} = \sqrt{(-0.328)^2 + (-0.111)^2} = 0.3465$$

$$\cos(\overline{\theta}) = \frac{\overline{C}_n}{\overline{R}} = -\frac{0.328}{0.3465} = -0.94726$$

$$\sin(\overline{\theta}) = \frac{\overline{S}_n}{\overline{R}} - \frac{0.111}{0.3465} = -0.32046$$

$$\overline{\theta} = \arctan\left(\frac{\sin(\overline{\theta})}{\cos(\overline{\theta})}\right) = a\tan\left(\frac{-0.94726}{0.32046}\right) = 0.326216$$

Sinüs ve kosinüsün toplamı negatif olduğu için ortalama yönün çeyrek özgüllük özelliğini kullanarak hesaplanan ortalama değer üzerine π eklenerek, ortalama yön **0,326216085 + 3,141593 = 3,467809** olarak hesaplanmıştır.

Plazmodium falciparum sıtma vakalarına ait verilerin dairesel varyansı aşağıdaki şekilde hesaplanmıştır.

$$Var = 2(1-\overline{R}) = 2(1-0.351) = 1.298$$

Dairesel standart sapma, $\overline{R}$'nin negatif logaritmasının kareköküdür. Bu ifade bazen Fisher'in standart sapması olarak bilinir.

$$\sigma = \sqrt{\left(-2\log \overline{R}\right)} = 0.6747$$

3.2 Dairesel Üniformite Testi

Dairesel üniformite testi, dairesel verilerin analizinde çok önemli bir konudur. Normal dağılım doğrusal analizlerin merkezi olduğu gibi, dairesel üniform dağılım da dairesel analizlerin merkezinde yer alır. Ancak dairesel veriler için çok az sayıda dairesel üniformite testi bulunmaktadır. Bu tezde dairesel üniformite testlerinden üç tanesini farklı biyolojik verilerde uygulanarak bu üniformite testlerinin zayıf ve kuvvetli olduğu yönlerin belirlenerek, bu üç test arasında karşılaştırma yapılmıştır. Yapılan karşılaştırma sonucuna göre bu üç üniformite testleri arasında en güvenilir olanının "Rao's spacing test" olduğu sonucuna varılmıştır.

3.3 Dairesel Verilere İlişkin Korelasyon ve Regresyon Analizleri

3.3.1 Dairesel korelasyon ölçüsü

Dairesel korelasyon yöntemleri kısmi korelasyon mantığı ve ifadelerini kullanır. Dairesel korelasyon ölçümünü hesaplamada, radyan zaman ölçüsünün sinüs ve kosinüs bileşenleri kullanılır.

3.3.1.1 Dairesel- doğrusal (dairesel-doğrusal) korelasyon

Sıtma hastalığı ile bir zaman dilimi arasında herhangi bir ilişki olup olmadığını belirlemek için Mardia'nın dairesel-lineer korelasyon katsayısı kavramını kullanarak, Çizelge 3.2'de verilen sıtma vakalarına ait verilerin frekansları ile radyan zaman ölçüsü arasındaki dairesel-doğrusal korelasyon analizi yapılmıştır.

İlk olarak, sıtma vakalarının frekansları ile radyan zaman ölçüsünün sinüsü ve kosinüsü arasındaki doğrusal korelasyon hesaplanmıştır. Daha sonra ise radyan zaman ölçüsünün sinüs ve kosünüs bileşenleri arasındaki doğrusal korelasyon katsayıları çizelge 3.3'de görülebileceği gibi hesaplanmıştır.

Çizelge 3.3 Frekans ile radyan zaman ölçüsünün bileşenlerine ait doğrusal korelasyon katsayıları

	Frekans	**Cos(θ)**	**Sin(θ)**
Frekans	1	-0.7218	-0.26451
Cos(θ)	0.520992047	1	0.000984
Sin(θ)	0.069967872	$9.68492x10^{-07}$	1

Çizelge 3.3'de bulunan katsayılar kullanılarak aşağıdaki doğrusal-dairesel korelasyon katsayı hesaplanmıştır.

$$r^2 = \frac{r_{XC}^2 + r_{XS}^2 - 2r_{XC}r_{XS}r_{CS}}{1 - r_{CS}^2} = \frac{0.25099 + 0.069968 - 2(-0.7218 * -026451 * 0.000984)}{1 - 9.6849x10^{-07}}$$
$$= 0.59058$$

Hesaplanan doğrusal-dairesel korelasyon katsayısı incelendiğinde gerçekten de zaman ve sıtma hastalığı arasında bir ilişki olduğunu ifade etmektedir.

3.3.1.2 Dairesel-dairesel korelasyon

Kalp krizi geçirme riski taşıyan hasta grubunda, kalp krizi geçirme zamanı ile doğum tarihi arasında herhangi bir ilişki olup olmadığını belirlemek için Çizelge 3, 4'de verilen kalp krizi geçirme zamanına ait radyan ölçüsü ile doğum tarihinin radyan ölçüsü arasındaki dairesel-dairesel korelasyon analizi yapılmıştır.

Çizelge 3.4 Kalp krizi ile doğum tarihlerine ilişkin simülasyon verilerin radyan ölçüsü

Sıra	Doğum tarihi	Kalp krizi geçirme tarihi
1	0.278187	1.232262
2	5.26886	-1.88462
3	0.384266	0.076561
4	2.168721	0.157084
	…	…
	…	…
	…	…
1332	0.762404728	0.233955145
1333	2.435280699	0.013507469
1334	1.169476352	0.078194345
1335	2.033624633	0.125431734
1332	0.762404728	0.233955145
1333	2.435280699	0.013507469

Aşağıda verilen ifade kullanılarak dairesel-dairesel korelasyon katsayısı –0.14221 olarak hesaplanmıştır.

$$r_c = \frac{\sum_{k=1}^{n} \sin\left(\theta_i - \bar{\theta}\right)\sin\left(\phi_i - \bar{\phi}\right)}{\sqrt{\sum_{k=1}^{n} \sin^2\left(\theta_i - \bar{\theta}\right)\sum_{k=1}^{n} \sin^2\left(\phi_i - \bar{\phi}\right)}} =$$

$$\frac{\sin(1.23-0.186)\sin(0.28-(-0.789))+\sin(1.88-0.186)\sin(5.27-(-0.789))...}{\sqrt{\sin(1.23-0.186)^2\sin(0.28-(-0.789))^2+\sin(1.88-0.186)^2\sin(5.27-(-0.789))^2...}} = -0.14221$$

3.3.2 Dairesel regresyon analizleri

Dairesel regresyon analizi, diğer regresyon analizlerinden gerek yöntem gerek sonuç gerekse de sonuçların yorumlanması gibi pek çok açıdan farklılık göstermektedir. Dairesel verilerin yapısına bağlı olarak regresyon analizinin yapılmasında kullanılan 3 çeşit regresyon analizi bulunmaktadır. Bu 3 çeşit regresyon analizi Çizelge 3.5'de özetlenmiştir.

Çizelge 3.5 Dairesel regresyon çeşitleri

Regresyon Çeşitleri	Bağımlı Değişken	Bağımsız Değişken
Doğrusal-Dairesel	Doğrusal	Dairesel
Dairesel-Doğrusal	Dairesel	Doğrusal
Dairesel-Dairesel	Dairesel	Dairesel

3.3.2.1 Doğrusal-dairesel regresyon analizi

Bağımsız değişkenin dairesel, bağımlı değişkenin ise doğrusal olduğu durumlarda regresyon analizi Doğrusal-Dairesel olarak isimlendirilir. Bu analizde veriler modele trigonometrik polinom olarak eklenirler. Sıtma hastalığına ait verilerde hastalık frekansı bağımlı değişken, düzeltilmiş zaman radyan ölçüsü ise bağımsız değişken olarak alınmıştır.

Daha önce hesaplanan korelesyon katsayına göre sıtma ile aylar arasında bir ilişki olduğunu bulduğumuz için en küçük kareler methodu ve Fourier dizi analizleri kullanılarak sinusoidal regresyon analizi yapılmıştır. Kullanılan regresyon denklemi aşağıdaki gibidir.

$$y_j = A_0 + A_1\cos(\omega\theta - \varphi) + A_2\cos(2\omega\theta - \varphi) + ... + A_p\cos(p\omega\theta - \varphi)$$
$$+ B_1\sin(\omega\theta - \varphi) + B_2\sin(2\omega\theta - \varphi) + ... + B_p\sin(p\omega\theta - \varphi) + \varepsilon_j$$

Dairesel regresyon analizi yapılırken karşılaşılan en büyük zorluk polinomların derecelerinin belirlenmesidir. Bu sorunun üstesinden gelmenin bir yolu iteratif yönteminin kullanılmasıdır. Bu yöntemde regresyonu derece derece artırılarak regresyon katsayısının anlamsız çıktığı dereceye kadar analiz devam ettirilir. Anlamsız çıktığı derecenin bir alt polinom derecesindeki regresyon denklemi en iyi regresyon modeli olarak kabul edilir. Aşağıda verilen Çizelge 3.6'da görülebileceği gibi 4. polinom derecesinde regresyon denklemi anlamsız bulunmuştur. Bu nedenle 3. polinom derecesindeki regresyon denklemi en iyi regresyon denklemi olarak alınmıştır.

Çizelge 3.6 Sıtma vakalarına ait ilk 4 polinom derecesinin P değerleri

Regresyon	1. derece	2. derece	3. derece	4. derece
P-cosine	0.083739976	$8.54725X10^{-09}$	0.00396409	0.848432743
P-sine	0.487609911	0.091924757	0.004250338	0.679340993

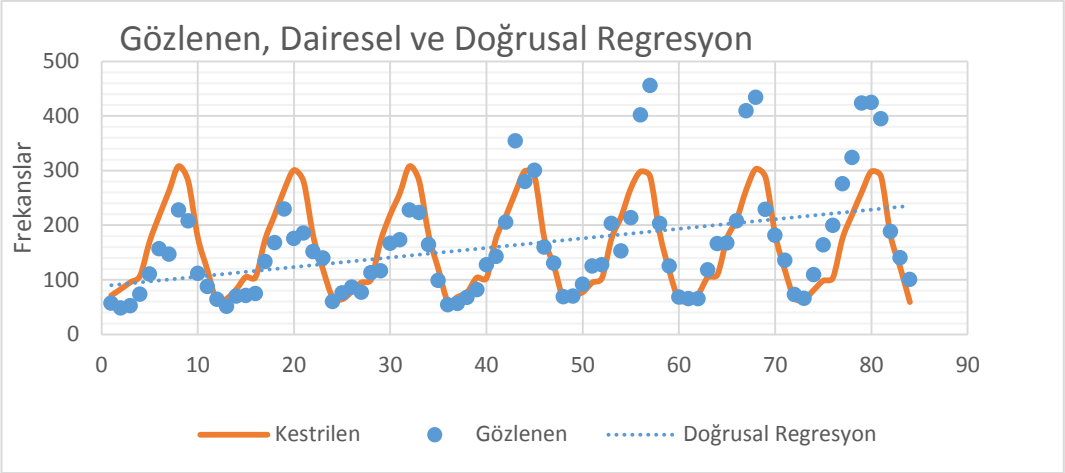
Çizelge 3.7 3. polinom derecesinin katsayıları

Katsayıların derecesi	Katsayılar
Kesişim	145,1841744
cos(ωθ- φ)	-28,27173067
sin(ωθ- φ)	5,124309228
cos(2ωθ- φ)	124,2057774
sin(2ωθ- φ)	114,1844083
cos(3ωθ- φ)	34,93872006
sin(3ωθ- φ)	-41,17719323

Çizelge 3.7 de verilen katsayılar kullanılarak aşağıdaki doğrusal-dairesel regresyon denklemi bulunmuştur.

$$y = 145.1842 - 28.2717\cos(\omega\theta - \varphi) + 124.2058\cos(2\omega\theta - \varphi) + 34.93872\cos(3\omega\theta - \varphi) + 5.124309\sin(\omega\theta - \varphi) + 114.1844\sin(2\omega\theta - \varphi) - 41.1772\sin(3\omega\theta - \varphi)$$

Gözlenen ve kestirilen frekanslar şekil 3.2'de verilmiştir. Şekil 3.2 incelendiğinde gözlenen verilerin frekansları ile regresyon dalgasının benzer bir eğilime sahip olduğu görülmektedir. Bu benzerlik dairesel regresyon analizi diğer regresyon analizlerinden daha iyi bir yöntem olduğunun göstergesidir.



Şekil 3.2 Gözlenen ve kestirilen değerler

3.3.2.2 Dairesel -dairesel regresyon analizi

Hem bağımsız hem de bağımlı değişkenin dairesel olduğu durumlarda regresyon analizi Dairesel-Dairesel olarak isimlendirilir. Ankara'da kalp krizi geçirme riski bulunan hastaların %3'ne ait veriler kullanılarak R paket programında 1335 hastaya ait doğum tarihleri ve kalp krizi geçirme tarihleri aşağıdaki gibi üretilmiştir.

birthdate <- runif(1335.min=0. max=6.283185307)
*> Attackdate <- atan2(0.3*cos(birthdate)*
*+ + 0.4*sin(birthdate).*
+ 0.5(sin(birthdate))+ rvm(1335, 5.024972, 0.101588))*
> circ.cor(birthdate. Attackdate. test=T)

r	*test.stat*	*p.value*
-0.1422142	-5.007078	5.526265×10^{-07}
r	test.stat	p.value

Üretilen verilerin R programında dairesel-dairesel korelasyon katsayısı (r=-0.1422142) anlamlı olduğu için aşağıdaki paket ile dairesel-dairesel regresyon analizi yapılmıştır.

install.packages("CircStats")

> library(CircStats)

> install.packages(“circular”)

> library(circular)

Yukarıdaki paket ile yapılan regresyon analizi sonucunda elde edilen katsayılar Çizelge 3.8’de verilmiştir.

Çizelge 3.8 Kalp krizi verilerine ait dairesel-dairesel regresyon katsayıları

	1. derece	2. derece
Kesişim	0.49466899 (α0)	-0.12129390 (δ0)
cos.alpha1	-0.56119174 (α1)	0.23982541(δ1)
cos.alpha2	0.07016889 (α2)	-0.04892861 (δ2)
sin.alpha1	-0.57186760 (β1)	-0.56766988 (β2)
sin.alpha2	0.27529379 (γ1)	0.14979634 (γ2)

Çizelge 3.8’de verilen katsayılar kullanılarak aşağıdaki dairesel-dairesel regresyon denklemi bulunmuştur.

cos(attack-date) = 0.49466899 -0.56119174*cos(birthdate) + 0.23982541*cos2(birthdate)- 0.57186760*sin(birthdate) - 0.56766988*sin2(birthdate)

sin(attack-date) = -0.12129390 + 0.07016889*cos(birthdate) - 0.04892861*cos2(birthdate) + 0.27529379 *sin(birthdate) + 0.14979634*sin2(birthdate)

4. TARTIŞMA VE SONUÇ

4.1 Tartışma

Bu tez çalışması ile dairesel regresyon analizinin farklı biyolojik veriler üzerinde uygulanması amaçlamıştı. Bununla birlikte, dairesel istatistikler analizin yeni bir konusu olduğu için, çalışmamıza dairesel verinin temellerinden başlamak zorundaydık. Dairesel verilerin dağılımları tespit edilmiş ve tanıtıcı istatistikleri hesaplanmıştır. Dairesel veri analizi, pek çok açıdan tahmin, çıkarım ve öngörülerin birçok alanında kullanılan olağan veri analizi yöntemlerinden çok farklıdır. Dairesel analizin ilk benzersiz özelliği bir veri türüdür. Dairesel veriler başlıca iki ölçümden kaynaklanır; Yön ve saat. Yuvarlak veri analizi, ilerlemesinin yalnızca 50 yaşında olduğu nispeten yeni istatistik alanlarıdır. Dairesel veri analizi, tahmin, çıkarım ve kestirim yapmak için yaygın olarak kullanılan diğer istatistik analiz yöntemlerinden oldukça farklıdır. Dairesel analizin diğer istatistik analiz yöntemlerinden ayıran ilk özelliği kullanılan veri türüdür Dairesel veriler çoğunlukla yön ve dairesel zaman ölçüleri olmak üzere iki farklı ölçüm çeşidi vardır. Dairesel analizler son 50 yılda gelişmeye başlamış yeni bir istatistik alandır. Dairesel verilerin analizinde farklı zorluklar vardır. Bu zorluklar, yeterli referans eksikliği, uygulanabilir model eksikliği ve uygun veri kaynaklarının olmamasıdır. Dairesel veriler zamana bağlı olduğundan, verilerin toplanması oldukça zaman almaktadır. Farklı zorluklarla karşı karşıya kalmamıza rağmen elimizdeki verileri kullanarak, biyolojik çalışmaların farklı alanlarında dairesel regresyon analizinin nispeten daha iyi anlaşılmasına katkıda bulunduk.

4.2 Öneriler

Genel olarak dairesel istatistiksel analiz alanındaki çalışmalar ve özellikle de dairesel regresyon analizinde, veri toplama ve veri yönetimi alanlarında geliştirmeler yapılabilir.

Veri toplama işleminde, dairesel analiz için veriler toplanırken zamanın daireselliğinin dikkate alınması gerekmektedir.

Dağılım ve veri grafikleri istatistiksel analizin omurgası olduğundan, daha kolay ve anlaşılabilir dağılım parametreleri ve veri grafikleri geliştirilmesi gerekmektedir.

Daha iyi ve anlaşılabilir bir dairesel varyans, standart sapma hesaplama ve çoklu karşılaştırma yöntemi olmalıdır.

Dairesel regresyon yöntemlerinde, şuanda elimizde bulunan ve gelecekte sahip olacağımız "en iyi" dairesel regresyon modeli arasında çok fazla boşluk bulunmaktadır. İleride yapılacak çalışmalar ile bu boşlukların doldurulması gerekmektedir.

ÖZGEÇMİŞ

Adı ve Soyadı : Desta Firdu Mekonnen
Doğum Yeri : Dabre Sina
Doğum Tarihi : 18 / 01 / 1983
Medeni Hali : Bekâr
Yabancı Dili **:** İngilizce, Türkçe

Eğitim Durumu (Kurum ve Yıl)

Lise **:** Adama Hazırlık Lisesi (1999-2003)

Lisans **:** Mekelle Üniversitesi, Ziraat ve Doğa Bilimleri Fakültesi, Zootekni, Mera ve Vahşi Yaşam Bilimleri Bölümü (2003 – 2007)

Çalıştığı Kurumlar ve Yıl

Öğretim Üyesi; Mizan-Tepi Üniversitesi, Ziraat ve Doğa Bilimleri Fakültesi, Doğa Bilimleri Bölümü (2008-2010)

Yayınları

Desta Firdu Mekonnen and Ensar Başpınar. Application of Circular Regression Analysis on Biological Data: Case Study on Malaria Cases in District of Visakhapatnam, India. International Journal of Current Research 2017, 9 (4); 48594-48600.

Desta Firdu Mekonnen and Ensar Başpınar. Application of Circular ANOVA on Biological Data: Case Study on Crimean-Congo Hemorrhage Fever Cases in Turkey. International Annals of Medicine 2017, 1(3).

Diğer Faaliyetler ve Kurslar

Ankara Üniversitesi TÖMER Türkçe ve Yabancı Dil Uygulama ve Araştırma Merkezi, "Türkçe Kursu" Nisan 2010-Eylül 2010 (4 ay)