STATISTICA, anno LXXVI, n. 2, 2016

CIRCULAR STATISTICAL APPROACH TO STUDY THE OCCURRENCE OF SEASONAL DISEASES

Kishore Kumar Das [1]
Department of Statistics, Gauhati University, Assam, India
Sahana Bhattacharjee
Department of Statistics, Gauhati University, Assam, India.

1. Introduction

Seasonal diseases can be referred to as the diseases which are caused due to the effects of the seasons or due to the allergy of the weather. Seasonal change in the incidence of infectious diseases is a regular phenomenon in temperate and tropical climate (Grassly and Fraser, 2006). The study area we have chosen is the Kamrup (rural) district of Assam. With the "Tropical Monsoon Rainforest Climate", Assam is temperate (summer max. at $95-100^oF$ or $35-38^oC$ and winter min. at $43-46^oF$ or $6-8^oC$) and experiences heavy rainfall and high humidity (Singh, 1993). In the Statistics (2011-12), it had been reported that in Assam, 92.9% of persons had symptoms of chronic illness and sought medical care whereas 87.8% of persons had symptoms of acute illness and took treatment for the same. It had been seen that seasonal diseases constituted most of these diseases. For instance, 1386 per $1,00,000$ persons suffered from Diarrhoea/Dysentery; 2732 per $1,00,000$ persons suffered from Hypertension; 356 per $1,00,000$ persons suffered from Tuberculosis and 680 per $1,00,000$ persons suffered from Asthma. Most of these diseases are seasonal in nature and the number of people suffering from these diseases has increased with time (Statistics, 2012-13).

Although seasonal variation encompasses cyclic change in disease occurrence, cyclic variation is often neglected in health services research (Christiansen *et al.*, 2012). The studies on seasonal diseases that have been undertaken for Assam have primarily reported the months (seasons) during which the prevalence of such diseases is the highest (Baruah *et al.*, 2007). No further results indicating the average or median time period and the probablistic pattern of occurrence of diseases have been reported so far. This paper primarily aims to incorporate these factors in studying the occurrence of seasonal diseases by taking into account their cyclical variation.

[1] Corresponding Author. E-mail: sahana.bhattacharjee@hotmail.com

The seasonal diseases tend to recur in some fixed month (season) every year and so, the period of the cycle is a year. This is a property possessed by circular random variable (r.v), which repeats itself after an interval 2π and is represented as a point in the circumference of a unit circle. It is usually reported in degree and converted into radians for computational convenience. The main advantage of modeling the occurrence of seasonal disease as a circular r.v lies in the fact it will not only give us an idea about the modal time period of occurrence, but also let us assess the mean and median time period. Furthermore, the circular descriptive statistical tools provide us with certain measures using which we can have an insight into the underlying distribution of the sample and detect the presence of seasonality in the data. These cannot be assessed with the aid of linear statistical tools.

The primary data collected for the purpose of the study will contain information on individuals who are suffering from non-seasonal diseases or from no disease at all. Since we are interested only in analyzing the occurrence of seasonal diseases, it becomes necessary on our part to consider only those individuals who are suffering from seasonal diseases. We, thus, end up with the Censored Sample, it being censored in the sense that only the observations of interest are included in the sample. The analysis pertaining to the censored circular sample require some special circular descriptive statistical techniques, which has not yet been devised in the literature.

Identification of the effect of seasons (months) in the occurrence of seasonal diseases will require some specialized tests. Apart from the Pearson χ^2 test for heterogeneity, another early work in detection of seasonality in epidemiological data using circular statistical tools can be attributed to Edwards (1961) who devised a test based on weights placed around a unit circle. But the Edwards method has the drawback of performing poorly in small samples. David and Newell (1965) proposed a test based on the maximum difference in the number of occurrences in all possible pairs of 6-month divisions of the year. None of the above authors have worked towards identifying the presence of seasonality in the censored circular sample in their respective study. Therefore, there remains a need to develop a test of detection of seasonality for the censored circular sample.

In the circular statistics literature, linear-circular regression has been proposed where the response variable is linear and continuous whereas the predictor is circular (see Mardia and Jupp, 2000, p.257). Here, we may be interested in modeling the occurrence of seasonal diseases w.r.t. the months (seasons) which are circular r.v's, in which case, the response variable will be dichotomous in nature. No such method dealing with binary response and circular predictor has been proposed in literature till date.

It can thus be seen that an efficient and detailed account on the occurrence of seasonal disease will require an analysis of a censored circular sample, detection of presence of seasonal effects in it and finally, the development of a binary-circular regression model. This will eventually help the health officials of Assam in understanding the proper underlying pattern of occurrence of seasonal diseases. This in turn will envisage a better health scenario, ensuring an overall improvement of the place. Keeping these things in mind, we have set our objectives as

- developing Circular Descriptive Statistics for the censored circular sample and drawing inference about the overall pattern of occurrence of seasonal diseases - both month-wise and season-wise;
- devising a new test of detection of seasonality for the censored circular sample for detecting month-wise and season-wise variation in the occurrence of seasonal diseases; and finally,
- proposing a regression model for analyzing binary response from a circular predictor and apply it to the dataset under consideration in the study.

2. Study area and Data

The data has been taken from the project entitled "Statistical Modeling in Circular Statistics: An Application to Health Science" sponsored by University Grants Commission (UGC), New Delhi, India. The study area has been chosen to be Kamrup (rural) district of Assam, India, where 3508 individuals were surveyed for attaining the concerned objective. The period of the study is from the year 2013 to 2014.

Out of a wide range of diseases that were reported to have occurred during the study period, we have neglected those with probability of occurrence almost nearing 0. We were, thus, left with the following diseases - *Hypertension, Diabetes, Asthma, Typhoid, Malaria, Poisoning, Heart problem, COPD, Diarhhoea, Dysentery, Pneumonia, Cancer, Jaundice, Skin disease, Gynecological problem, Tuberculosis, Mental illness, Sexually Transmitted Disease (STD), Urinary tract infection.* Of these, only the five diseases viz., Diabetes, Poisoning, Cancer, Gynecological problem and Mental illness are non-seasonal diseases. We have, therefore, considered the data only on the seasonal diseases. A total of 2700 cases of seasonal diseases has been reported.

In the present study, the classification of the seasons has been done as Winter (January-February), Pre-monsoon/Summer (March-May), Monsoon (June-September) and Post-monsoon (October-December). For studying the variation in occurrence month-wise, data are presented in form of a 12-series monthly totals, where it consists of the number of individuals affected with a seasonal disease in a given month of a year. Similarly, for studying the variation in occurrence season-wise, data are presented in form of a 4-series seasonal totals, where it comprises of the number of individuals affected with a seasonal disease during a particular season of a year. Since the lengths of some of the months are equal with each other, whereas some differ by only a day or two, it is appropriate only to adjust the number of cases (frequencies) so that they correspond to "months" of equal length, without significantly affecting the actual frequencies corresponding to every month. Thus, we group the data for this case by attributing the class interval $(0^o, 30^o)$ to the month of January, and so on (Mardia and Jupp, 2000).

But in case of season-wise analysis, the seasons differ by a significant length and so, adjusting the number of cases (frequencies) so that they correspond to "seasons" of equal length will highly distort the actual frequencies and eventually give misleading results. That is why we opt for grouping the cases in unequal

intervals, the width of the intervals being proportional to the length of the seasons. Thus, we group the data for this case by attributing the class interval $(0^o, 58^o)$ to the season "Winter", $(58^o, 149^o)$ to the season "Pre-monsoon", $(149^o, 269^o)$ to the season "Monsoon" and $(269^o, 360^o)$ to the season "Post-monsoon".

3. Circular Descriptive Statistics for censored circular sample

The sample data has been restricted to the one containing only the observations of interest and the remaining observations are not of interest. So, the sample data has been described as being dichotomous in nature. In the present study, circular descriptive statistics and tests have been developed considering only the observations of interest. We, therefore, term this sample as censored circular sample. To fulfil the objectives of the paper, we define the *Circular Descriptive Statistics for censored circular sample*, which are as follows:

Often, circular data that arises in terms of angles come in a grouped form. Analogous to the linear case, we make the assumption that all the observations in an interval are concentrated at the mid-point of that interval. Therefore, if the original n observations are grouped into k classes with the i^{th} class having a mid-point of α_i and frequency f_i, then a typical observation can be denoted by α_{ij}, $i = 1, 2, \ldots, k; j = 1, 2, \ldots, f_i; \left(\sum_{i=1}^{k} f_i = n\right)$. The polar to rectangular transformation for each observation transforms them to the following form:

$$(\cos\alpha_{ij}, \sin\alpha_{ij}) \qquad i = 1, 2, \ldots, k; j = 1, 2, \ldots, f_i \tag{1}$$

The angular observations α_{ij}'s, measured in radians, are treated as unit vectors and their resultant is defined as follows:

Definition 1. *The* ***Censored Sample Resultant Vector*** *of these n unit vectors is obtained by adding them over all the components and is given by:*

$$\left(\sum_{i=1}^{k}\sum_{j=1}^{f_i}(\cos\beta_{cij}), \sum_{i=1}^{k}\sum_{j=1}^{f_i}(\sin\beta_{sij})\right) = (C_c, S_c) \tag{2}$$

where

$$\beta_{cij} = \begin{cases} \begin{cases} \pi/2\lceil\alpha_{ij}\rceil, & \text{if } \lceil\alpha_{ij}\rceil \text{ is odd} \\ \pi/2\lfloor\alpha_{ij}\rfloor, & \text{if } \lfloor\alpha_{ij}\rfloor \text{ is odd} \end{cases} & \text{if the observation is not of interest} \\ \alpha_{ij} & \text{if the observation is of interest} \end{cases} \tag{3}$$

$\lceil x\rceil$ and $\lfloor x\rfloor$ denoting the ceiling and floor function of x respectively and

$$\beta_{sij} = I_s\alpha_{ij} \tag{4}$$

Here, β_{cij} and β_{sij} are measured in radians and I_s is the ***sine indicator variable*** *defined as*

$$I_s = \begin{cases} 0, & \text{if the observation is not of interest,} \\ 1, & \text{if the observation is of interest,} \end{cases} \tag{5}$$

We denote by R_c, the length of the censored resultant vector given by

$$R_c = \left(C_c^2 + S_c^2\right)^{\frac{1}{2}}$$

R_c lies in the interval $(0, n)$. A particular case of the above definition is the sample resultant vector (see Mardia and Jupp, 2000, p.15), when there is complete absence of observation of no interest.

It can be seen that the usual linear techniques fail to give a unique measure for the mean direction of a set of angular observations, because it will depend upon the choice of zero direction and the sense of rotation (clockwise or anti-clockwise) (Rao and SenGupta, 2001). Consequently, the direction of the resultant vector of the censored sample is considered as a measure of the mean direction of the sample and is defined as follows:

DEFINITION 2. *The* ***Censored Sample Circular Mean Direction****, denoted by $\bar{\alpha}_{c0}$, is defined as the quadrant specific inverse of the tangent of the ratio of S_c to C_c and is given by*

$$\bar{\alpha}_{c0} = \arctan^* \left(\frac{S_c}{C_c}\right)$$

where

$$\bar{\alpha}_{c0} = \begin{cases} \arctan\left(\frac{S_c}{C_c}\right), & \text{if } C_c > 0, S_c \geq 0, \\ \frac{\pi}{2}, & \text{if } C_c = 0, S_c > 0, \\ \arctan\left(\frac{S_c}{C_c}\right) + \pi, & \text{if } C_c < 0, \\ \arctan\left(\frac{S_c}{C_c}\right) + 2\pi, & \text{if } C_c < 0, S_c < 0, \\ \text{undefined}, & \text{if } C_c = 0, S_c = 0. \end{cases}$$

arctan is operated so as to provide us with an unique inverse of $\frac{S_c}{C_c}$ on $[0, 2\pi]$ and hence, a unique mean direction.
A particular case of the above definition is the sample circular mean direction (see Rao and SenGupta, 2001, p.13), when there is complete absence of observation of no interest.

DEFINITION 3. *For the purpose of robust estimation of the population measure of central tendency and to have an idea about the direction which divides the sample into two equal halves, we need to consider the measure for circular median of the sample. A* ***Censored Sample Circular Median Direction*** *of angular measurements is any angle ϕ such that*

- *half of the data points lie in the arc $[\phi, \phi + \pi)$, and*
- *the majority of the data points are nearer to ϕ than to $\phi + \pi$. Further, it can be seen that when the sample size n is odd, the sample median is one of the data points. When n is even, it is convenient to consider the midpoint of two appropriate adjacent data points analogous to the linear case.*

It is equivalent to finding ϕ by minimizing the function

$$d(\alpha) = \pi - \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i} I_m \mid \pi - \mid \alpha_{ij} - \alpha \mid\mid \tag{6}$$

where

$$I_m = \begin{cases} 0, & \text{if the observation is not of interest,} \\ 1, & \text{if the observation is of interest,} \end{cases} \tag{7}$$

and which is a function of α.

A particular case of the above definition is the sample circular median direction (Mardia and Jupp, 2000; Fisher, 1993), when there is complete absence of observation of no interest.

The quantity $d(\alpha)$ appearing in equation (6) gives a measure of the spread of the angles $\alpha_{ij}{}'s$ about the angle α (see Mardia and Jupp, 2000, p.19). Analogous to the linear case, where the median minimizes the mean deviation of a set of observations (representing dispersion in the dataset), the circular median direction of the censored sample also minimizes the sample measure of dispersion $d(\alpha)$. Based on this analogy, an equivalent measure of mean deviation for censored circular sample is as defined below:

Definition 4. *A measure of spread of censored angular data associated with the median direction ϕ is the* ***Censored Sample Circular Mean Deviation*** *given by*

$$d(\phi) = \pi - \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i} I_m \mid \pi - \mid \alpha_{ij} - \phi \mid\mid$$

which is nothing but the minimum value taken by $d(\alpha)$ and I_m is as defined in (7).

A particular case of the above definition is the sample circular median direction (see Fisher, 1993, p.36), when there is complete absence of observation of no interest.

Definition 5. *The* ***Censored Sample Mean Resultant Length*** *$\bar{R}_c$ is a measure of concentration of a censored angular dataset around its mean direction and is defined as*

$$\bar{R}_c = R_c/n = \left(\bar{C_c}^{\,2} + \bar{S_c}^{\,2}\right)^{\frac{1}{2}} \tag{8}$$

where

$$\bar{C}_c = \frac{C_c}{n}, \bar{S}_c = \frac{S_c}{n}$$

C_c and S_c being defined as in (2).

Evidently, $\bar{R}_c$ lies between 0 and 1. If the directions are tightly clustered, it would indicate a greater concentration around the mean direction and so, $\bar{R}_c$ will lie nearer to 1. On the contrary, if the angles are widely dispersed, it would mean

that there is no apparent concentration of the observations towards the mean direction. So, in this case, $\bar{R}_c$ will approximately be equal to 0. However, $\bar{R}_c \approxeq 0$ does not necessarily indicate that the observations of the censored sample are evenly dispersed about the unit circle. It simply indicates lack of concentration around the mean direction.

A particular case of the above definition is the sample mean resultant length (see Mardia and Jupp, 2000, pp.17-18), when there is complete absence of observation of no interest.

DEFINITION 6. *In the linear statistics literature, we have a measure of variance whose smaller value indicates lesser dispersion of the data and vice-versa. This calls for the establishment of an analogous measure of variance for censored circular sample, so as to facilitate comparison with data on the line. We have, thus, defined the measure of dispersion viz.* ***Censored Sample Circular Variance*** *as follows:*

$$V_c = 1 - \bar{R}_c$$

where $\bar{R}_c$ is as defined in (8).

V_c lies between 0 and 1. The interpretation of V_c is just the opposite of that of $\bar{R}_c$.

A particular case of the above definition is the sample circular variance (see Mardia and Jupp, 2000, p.18), when there is complete absence of observation of no interest.

Here, we are working with grouped data with the assumption that all the frequencies are concentrated at the mid-point of the class intervals. This assumption is fairly true for intervals below 45^o in length. In case the grouping is very coarse, say for those exceeding 45^o, the correction for grouping is needed which would, otherwise, exhibit misleading results. Following the calculations shown in Stuart and Ord (1987), it can be seen that the sample mean directions and the sample trigonometric moments do not need correction for grouping as the data are measured to the nearest 1^o or 5^o (Mardia and Jupp, 2000, p.23). In our study, the intervals into which we have classified the months are of equal width 30^o and so, they require no adjustment for grouping. But the groups corresponding to the seasons are of unequal widths with each of them exceeding 45^o. This would mean that an adjustment for the grouping corresponding to seasons is required.

We see that for season-wise grouping, h is not equal for all the intervals. So, we propose to consider h as the circular mean of all the widths (in radians) and then apply the following formula to adjust the resultant length of the censored circular sample pertaining to the season-wise analysis for the grouping error.

DEFINITION 7. *The* ***Adjusted Censored Sample Mean Resultant Length*** *is defined as:*

$$\bar{R}^*_{cp} = a\,(ph)\,\bar{R}_{cp}$$

where

$$a\left(ph\right)=\frac{\frac{ph}{2}}{\sin\frac{ph}{2}}$$

and in particular

$$a\left(h\right)=\frac{\frac{h}{2}}{\sin\frac{h}{2}}$$

$$\bar{R}_c^*=a\left(h\right)\bar{R}_c$$

Here, h is the mean length in radians of the class interval widths and $\bar{R}_{cp}$ is as defined in the subsequent section.

Consequently, for this sample, the variance, now termed as the **Adjusted Censored Sample Circular Variance**, becomes

$$V_c^*=1-\bar{R}_c^*$$

The interpretation of V_c^* remains the same as that of V_c. The length of the censored sample mean resultant vector corresponding to month-wise analysis, which doesn't require adjustment, is calculated using equation (8). In case the intervals are of equal width and the data does not contain any observation of no interest, the adjustment can be made using the formula given in (Mardia and Jupp, 2000, p.23).

The sample trigonometric moments are the sample analogs of the population trigonometric moments, which play a very vital role in the theory of circular distributions through determination of population mean and concentration measures. The censored sample trigonometric moments (about both zero and mean direction) are as defined below:

Definition 8. *The **Censored Sample p^{th} Order Trigonometric Moment about the Zero Direction** is calculated as*

$$m'_{cp}=\bar{C}_c\left(p\right)+i\bar{S}_c\left(p\right)$$

where

$$\bar{C}_c\left(p\right)=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\left(\cos p\beta_{cij}\right),\bar{S}_c\left(p\right)=\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\left(\sin p\beta_{sij}\right)$$

where β_{cij} and β_{sij} are as defined in (3) and (4) respectively. It can again be written as

$$m'_{cp}=\bar{R}_{cp}e^{(i\bar{\alpha}_{cp})}$$

where $\bar{R}_{cp}$ and $\bar{\alpha}_{cp}$ denote the censored sample mean resultant length and the censored sample circular mean direction of the transformed dataset $\{p\alpha_1,p\alpha_2,\ldots,p\alpha_n\}$.

In particular,

$$m'_{c1}=\bar{R}_ce^{(i\bar{\alpha}_{c0})}$$

A particular case of the above definition is the sample p^{th} order trigonometric moment about zero direction (see Mardia and Jupp, 2000, pp.20-21), when there is complete absence of observation of no interest.

For the censored sample pertaining to the season-wise analysis that has undergone adjustment for grouping, the sample moments about zero direction have to be adjusted. The corrected moments are as defined below:

DEFINITION 9. *The* ***Adjusted Censored Sample*** p^{th} ***Order Trigonometric Moment about the Zero Direction*** *is given by:*

$$m_{cp}^{*\prime} = \bar{C}_c^*(p) + i\bar{S}_c^*(p)$$

where

$$\bar{C}_c^*(p) = a(ph)\,\bar{C}_c(p), \bar{S}_c^*(p) = a(ph)\,\bar{S}_c(p)$$

Here, we see that the mean direction of the adjusted vector remains the same as

$$\begin{aligned} arctan^*\left(\frac{\bar{S}_c^*(p)}{\bar{C}_c^*(p)}\right) &= arctan^*\left(\frac{\bar{S}_c(p)}{\bar{C}_c(p)}\right) \\ &= \bar{\alpha}_{cp} \end{aligned}$$

Thus, we have,

$$m_{cp}^{*\prime} = \bar{R}_{cp}^* e^{(i\bar{\alpha}_{cp})}$$

In particular,

$$m_{c1}^{*\prime} = \bar{R}_c^* e^{(i\bar{\alpha}_{cp})}$$

DEFINITION 10. *The* ***Censored Sample*** p^{th} ***Order Trigonometric Moment about the Mean Direction*** *is calculated as*

$$m_{cp} = \bar{C}_c'(p) + i\bar{S}_c'(p)$$

where

$$\bar{C}_c'(p) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\left\{\cos p\left(\beta_{cij} - \bar{\alpha}_{c0}\right)\right\},$$

$$\bar{S}_c'(p) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\left\{\sin pI_s\left(\alpha_{ij} - \bar{\alpha}_{c0}\right)\right\}$$

wherein β_{cij} *and* I_s *are as defined in (3) and (5) respectively.*

In particular,

$$m_{c1} = \bar{R}_c$$

as

$$\sum_{i=1}^{k}\sum_{j=1}^{f_i}\sin I_s\left(\alpha_{ij} - \bar{\alpha}_{c0}\right) = 0. \tag{9}$$

and

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\cos\left(\beta_{cij} - \bar{\alpha}_{c0}\right) = \bar{R}_c. \tag{10}$$

(9) and (10) are analogous to the equations

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

and

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2$$

that exist in linear statistics literature. For the censored sample corresponding to the season-wise analysis, the particular case of the above definition holds for the unadjusted censored sample mean resultant length $\bar{R}_c$.

A particular case of the above definition is the sample p^{th} order trigonometric moment about mean direction (see Mardia and Jupp, 2000, p.21), when there is complete absence of observation of no interest.

Analogous to the linear measures of sample skewness and kurtosis respectively given by

$$\gamma_1 = \frac{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^3}{n}}{\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right\}^{\frac{3}{2}}}$$

and

$$\beta_2 = \frac{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^4}{n}}{\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right\}^{2}}$$

the skewness and kurtosis measures for censored circular sample are defined as follows:

Definition 11. *The **Censored Sample Circular Skewness** is defined as*

$$\hat{s} = \frac{\left\{\bar{R}_{c2} \sin\left(\bar{\alpha}_{c2} - 2\bar{\alpha}_{c0}\right)\right\}}{V_c^{\frac{3}{2}}} \tag{11}$$

For symmetric unimodal datasets (such as the ones from Von Mises or Wrapped normal distribution), $\hat{s}$ is nearly zero, indicating symmetry. Analogous to the linear case, a positive (negative) value of the measure indicates that the underlying distribution is positively (negatively) skewed.

For the censored sample pertaining to the season-wise analysis of occurrence of seasonal diseases, we replace $\bar{R}_{c2}$ and V_c in (11) by $\bar{R}^*_{c2}$ and V_c^* respectively. This gives us the **Adjusted Censored Sample Circular Skewness**, defined as:

$$\hat{s}^* = \frac{\left\{\bar{R}^*_{c2} \sin\left(\bar{\alpha}_{c2} - 2\bar{\alpha}_{c0}\right)\right\}}{V_c^{*\frac{3}{2}}}$$

The interpretation of $\hat{s}^*$ remains the same as that of $\hat{s}$.

A particular case of the above definition is the sample circular skewness measure (see Mardia and Jupp, 2000, p.22), when there is complete absence of observation of no interest.

DEFINITION 12. *The* ***Censored Sample Circular Kurtosis*** *is defined as:*

$$\hat{k} = \frac{\left\{\bar{R}_{c2}\cos\left(\bar{\alpha}_{c2} - 2\bar{\alpha}_{c0}\right) - \bar{R}_c{}^4\right\}}{V_c^2} \tag{12}$$

Similar to the skewness measure for the censored sample for season-wise analysis, the ***Adjusted Censored Sample Circular Kurtosis*** *is obtained by replacing* $\bar{R}_{c2}$, $\bar{R}_c$ *and* V_c *in (12) by* $\bar{R}^*_{c2}$, $\bar{R}^*_c$ *and* V^*_c *respectively and is given by*

$$\hat{k}^* = \frac{\left\{\bar{R}^*_{c2}\cos\left(\bar{\alpha}_{c2} - 2\bar{\alpha}_{c0}\right) - \bar{R}^{*4}_c\right\}}{V_c^{*2}}$$

The interpretation of $\hat{k}^*$ remains the same as that of $\hat{k}$.

For unimodal datasets with a peak that can be well approximated by a wrapped normal distribution, $\hat{k}$ is nearly 0, i.e., mesokurtic. Analogous to the linear case, a positive (negative) value of the measure indicates that the underlying distribution is sharp (flat) peaked. The numerators in the expressions (11) and (12) have arisen due to the interpretation of trigonometric moments of concentrated distributions on the circle.

A particular case of the above definition is the sample circular kurtosis measure (see Mardia and Jupp, 2000, p.22), when there is complete absence of observation of no interest.

3.1. Rayleigh Uniformity Test for Censored Circular Sample

One of the most vital test of hypothesis that arises in the circular statistics literature is the test of uniformity, i.e. whether the distribution on a circle is uniform or not. If the circular r.v. θ has a uniform distribution on the circle, then $E[(\cos\theta, \sin\theta)^T] = 0$ and so, it is intuitively reasonable to reject uniformity when its estimate viz. the vector sample mean $(\bar{C}, \bar{S})$ is far from 0 or equivalently, when $\bar{R}$ is large (see Mardia and Jupp, 2000, pp.94-95). Thus, a test statistic based on $\bar{R}$ would prove to be fruitful.

In case of censored circular sample, under the hypothesis of uniformity of the observations, it can be seen that the asymptotic distribution of $2n\bar{R}_c{}^2$ is χ^2_2. This constitutes the **Rayleigh Uniformity Test for Censored Circular Data** of which the general Rayleigh Uniformity Test is a particular case when there is complete absence of observation of no interest.

In case of the censored sample corresponding to season-wise analysis, the test statistic undergoes adjustment and the adjusted test statistic becomes $\dfrac{2n}{\{a(h)\}^2}\bar{R}^{*2}_c$ which follows χ^2_2 distribution asymptotically.

4. BINARY LOGISTIC REGRESSION FOR LINEAR RESPONSE AND CIRCULAR PREDICTOR

In the circular statistics literature, many situations may call for the analysis and prediction of a dichotomous linear outcome from a circular predictor. In such a

case, analogous to the linear case (see Peng *et al.*, 2002, pp.4-5) one may think of carrying out a *Binary Logistic Regression for Circular Predictor and Linear Response*. The proposed model is described as follows:

Let Y be the binary linear response variable representing the occurrence of event of interest and let

$$Y = \begin{cases} 1, & \text{if the event is of interest,} \\ 0, & \text{if the event is not of interest.} \end{cases} \tag{13}$$

Further, let π be the probability of happening of Y and (1-π) be the probability of non-happening of Y. Let α denote the circular predictor variable which is continuous in nature and measured in radians. Since α is a continuous circular r.v, its probability density function exists and has the following properties (see Rao and SenGupta, 2001, p.25):

$$\begin{aligned} &(i) f(\alpha) \geq 0; \\ &(ii) \int_0^{2\pi} f(\alpha)\, d\alpha = 1; \\ &(iii) f(\alpha) = f(\alpha + k.2\pi) \end{aligned}$$

for any integer k, i.e. f is a periodic function with period 2π.

Since a circular r.v. cannot be treated in a similar manner as a linear r.v, we need to transform α in such a way that the periodicity is taken into account. We are to look for a function $f(x)$ so that

$$\lim_{x \longrightarrow 360^o} f(x) = f(0)$$

We, thus, perform the sine and cosine transformation of α, and include the resulting components viz. $\sin\alpha$ and $\cos\alpha$ as covariates in the regression model. Then the Binary Logistic Regression for Circular Predictor and Linear Response consists in predicting the logit of Y from $\sin\alpha$ and $\cos\alpha$. Logit of Y is the natural logarithm (ln) of the odds of Y, where odds is the ratio of probability of happening of Y to that of probability of non-happening of Y. The mathematical form of the proposed model is given by:

$$\ln\left(\frac{\pi}{1-\pi}\right) = a + b\sin\alpha + c\cos\alpha \tag{14}$$

From (14) the probability of the occurrence of event of interest is predicted as follows:

$$\pi = Pr(Y = Outcome\ of\ interest|\alpha) = \frac{e^{a+b\sin\alpha+c\cos\alpha}}{1+e^{a+b\sin\alpha+c\cos\alpha}}$$

Analogous to the linear case, the coefficients a, b and c are estimated using the Maximum Likelihood Estimation method. The regression coefficients represent the change in the logit of the dependent variable with an unit change in the

TABLE 1
Month-wise occurrence of seasonal diseases in the Kamrup (rural) district of Assam, India.

Month	Angular range (in degrees)	Number of cases (frequency)
January	$0 - 30$	222
February	$30 - 60$	309
March	$60 - 90$	242
April	$90 - 120$	215
May	$120 - 150$	276
June	$150 - 180$	158
July	$180 - 210$	219
August	$210 - 240$	308
September	$240 - 270$	122
October	$270 - 300$	180
November	$300 - 330$	235
December	$330 - 360$	214

corresponding predictor. The statistical significance of the individual regression coefficients can be assessed using the Wald's Chi-square statistic and the goodness of fit of the logistic model can be assessed using three measures viz., the Hosmer-Lemeshow (H-L) test statistic, the Cox and Snell's R^2 index and Nagelkerke's R^2 index (Peng *et al.*, 2002). The interpretation of the two indices is same as that of the coefficient of determination in linear regression, but Cox and Snell's R^2 can never reach a maximum of value 1. However, Nagelkerke's R^2 which can be seen as an amendment of Cox and Snell's R^2, can reach a maximum value of 1. The H-L statistic tests the validity of the hypothesis that the predicted probabilities fit the actual probabilities well. A positive (negative) value of the estimated regression coefficient implies that there is a direct (inverse) relationship between the logit of Y and the predictor. The odds ratio are interpreted in the same manner as in the case of linear predictor.

5. Analysis

5.1. Data sets

The month-wise and season-wise adjusted frequencies of seasonal diseases in our study area have been displayed in Table (1) and Table (2) respectively.

5.2. Censored Circular Descriptive statistics for month-wise analysis

Here, the individual is of interest if he is suffering from (reported) any kind of seasonal disease during a month; otherwise, is not of interest. Thus, only those individuals will be included in the sample who are suffering from (reported) any kind of seasonal diseases in a month. This gives us the censored circular sample for month-wise analysis. The circular descriptive statistics of this censored sample are listed in Table (3).

TABLE 2
Season-wise occurrence of seasonal diseases in the Kamrup (rural) district of Assam, India

Season	Angular range (in degrees)	Number of cases (frequency)
Winter	$0 - 58$	544
Pre-monsoon	$58 - 149$	742
Monsoon	$149 - 269$	798
Post-monsoon	$269 - 360$	616

TABLE 3
Censored Circular Descriptive Statistics for data shown in Table (1)

Statistics	Values (in Radians)
Censored Sample Circular Mean Direction	1.27
Censored Sample Mean Resultant Length	0.06
Censored Sample Circular Variance	0.94
Censored Sample Circular Median Direction	1.31
Censored Sample Circular Mean Deviation	1.48
Censored Sample Circular Skewness	0.04
Censored Sample Circular Kurtosis	-0.04

5.3. Censored Circular Descriptive statistics for season-wise analysis

Same is for the season-wise analysis, wherein the individual is of interest if suffering from (reported) any kind of seasonal disease during a season; otherwise, it is not of interest. This provides us with the censored circular sample for season-wise analysis. Table (4) lists the circular descriptive statistics of this censored sample.

5.4. Censored Sample Trigonometric Moments for month-wise analysis

The trigonometric moments about the Zero and the Mean Direction for censored circular sample pertaining to the month-wise analysis has been displayed in the

TABLE 4
Censored Circular Descriptive Statistics for data shown in Table (2)

Statistics	Values (in Radians)
Censored Sample Circular Mean Direction	1.35
(Unadjusted) Censored Sample Mean Resultant Length	0.06
Adjusted Censored Sample Mean Resultant Length	0.07
Adjusted Censored Sample Circular Variance	0.93
Censored Sample Circular Median Direction	1.81
Censored Sample Circular Mean Deviation	1.40
Adjusted Censored Sample Circular Skewness	0.09
Adjusted Censored Sample Circular Kurtosis	-0.08

TABLE 5
Censored Sample Trigonometric Moments about the Zero Direction for data shown in Table (1)

Order of moments (p)	$\left(\bar{C}_c(p), \bar{S}_c(p)\right)$ (in Radians)
1	$(0.02, 0.06)$
2	$(0.02, 0.05)$
3	$(0.00, -0.04)$
4	$(-0.13, 0.03)$

TABLE 6
Censored Sample Trigonometric Moments about the Mean Direction for data shown in Table (1)

Order of moments (p)	$\left(\bar{C}'_c(p), \bar{S}'_c(p)\right)$ (in Radians)
1	$(0.06, 0.00)$
2	$(0.01, -0.05)$
3	$(0.02, 0.04)$
4	$(-0.08, -0.10)$

tables Table (5) and Table (6) respectively.
It can be seen from Tables (3), (5) and (6) that

$$\begin{aligned}\left(\bar{R}_c \cos\bar{\alpha}_{c0}, \bar{R}_c \sin\bar{\alpha}_{c0}\right) &= (0.06\cos(1.27), 0.06\sin(1.27)) \\ &= (0.02, 0.06) \\ &= m'_{c1}\end{aligned}$$

which proves the particular case under definition (8) and

$$\begin{aligned}\bar{R}_c &= 0.06 \\ &= 0.06 + i.(0.00) \\ &= m_{c1}\end{aligned}$$

This proves the particular case under definition (10) for the censored sample for month-wise analysis.

5.5. Censored Sample Trigonometric Moments for season-wise analysis

In the following section, the trigonometric moments about the Zero and the Mean Direction of the censored circular sample corresponding to the season-wise analysis have been depicted through the Tables (7) and (8) respectively.
From the Tables (4), (7) and (8), it can be observed that

$$\begin{aligned}\left(\bar{R}^*_c \cos\bar{\alpha}_{c0}, \bar{R}^*_c \sin\bar{\alpha}_{c0}\right) &= (0.07\cos(1.35), 0.07\sin(1.35)) \\ &= (0.02, 0.07) \\ &= {m^*_{c1}}'\end{aligned}$$

TABLE 7
Adjusted Censored Sample Trigonometric Moments about the Zero Direction for data shown in Table (2)

Order of moments (p)	$\left(\bar{C}_c^*(p), \bar{S}_c^*(p)\right)$ (in Radians)
1	$(0.02, 0.07)$
2	$(0.02, 0.08)$
3	$(0.01, -0.51)$
4	$(-0.32, 0758)$

TABLE 8
Censored Sample Trigonometric Moments about the Mean Direction for data shown in Table (2)

Order of moments (p)	$\left(\bar{C}_c'(p), \bar{S}_c'(p)\right)$ (in Radians)
1	$(0.06, 0.00)$
2	$(0.02, -0.07)$
3	$(0.35, 0.29)$
4	$(-0.71, 0.20)$

which proves the particular case under definition (9) and

$$\begin{aligned} \bar{R}_c &= 0.06 \\ &= 0.06 + i.(0.00) \\ &= m_{c1} \end{aligned}$$

This proves the particular case under definition (10) for the censored sample for season-wise analysis.

5.6. Rayleigh Uniformity Test for Censored Circular Data for month-wise analysis

If there were no variation with respect to months, the angular observations corresponding to Table (1) could be regarded as being drawn from the uniform distribution on the circle. Hence, we may frame our null hypothesis to be tested as:

H_0: The occurrence of seasonal diseases does not have any month-wise variation.

For the data shown in Table (1), the test statistic is

$$2n\bar{R_c}^2 = 22.24$$

and its tabulated value at 1% level of significance is 9.21. Also, the p-value of the test is 0.00.

Hence, our null hypothesis is strongly rejected and we conclude that *the occurrence of seasonal diseases has month-wise variation.*

5.7. Rayleigh Uniformity Test for Censored Circular Data for season-wise analysis

If there is no variation season-wise, the angular observations corresponding to Table (2) could be regarded being drawn from the uniform distribution on the circle. Hence, we may frame our null hypothesis as:

H_0: The occurrence of seasonal diseases does not have any season-wise variation.

and the alternative as

H_1: The occurrence of seasonal diseases has season-wise variation.

The value of the test statistic obtained on the basis of data shown in Table (2) is

$$\frac{2n}{\{a\,(h)\}^2}\bar{R}_c^{*^2} = 19.71.$$

The tabulated value at 1% level of significance is 9.21 and the p-value of the test is 0.00.

Hence, our null hypothesis is strongly rejected and we conclude that *the occurrence of seasonal diseases has season-wise variation.*

5.8. Binary Logistic regression for linear response and circular predictor for month-wise analysis

The individual is of interest if he is suffering from (reported) any kind of seasonal diseases during a month and otherwise is not of interest. So we define the linear response Y as

$$Y = \begin{cases} 1, & \text{if the individual suffers from seasonal disease,} \\ 0, & \text{if the individual does not suffer from seasonal disease.} \end{cases} \tag{15}$$

The continuous circular predictor is the month of a year which is represented in terms of angle as mentioned in section (2).
The Table (9) summarizes the results of the Binary Logistic regression for linear response and circular predictor for month-wise analysis of occurrence of seasonal diseases.

Interpretation: From the Table (9), it can be seen that the fitted model has been found to be:

$$\text{predicted } \ln\left(\frac{\pi}{1-\pi}\right) = 1.750 - 1.020\sin\alpha - 0.212\cos\alpha$$

The log of the odds of a person to suffer from seasonal disease during a month is significantly negatively related to the both the sine component ($p < 0.05$) and the cosine component ($p < 0.05$) of the months modeled as a circular r.v. It is

TABLE 9
Binary Logistic Regression analysis predicting the occurrence of seasonal diseases from months

Predictor	β	p	Odds Ratio
Constant	1.750	0.000	5.755
$\sin\alpha$	-1.020	0.000	0.361
$\cos\alpha$	-0.212	0.001	0.809
Goodness of fit test statistics	Value	p-value	
Cox and Snell's R^2	0.063		
Nagelkerke's R^2	0.100		
Hosmer and Lemeshow test statistic	14.5	0.06	

TABLE 10
Binary Logistic Regression analysis predicting the occurrence of seasonal diseases from seasons

Predictor	β	p	Odds Ratio
Constant	1.490	0.000	4.436
$\sin\alpha$	-0.521	0.000	0.594
$\cos\alpha$	-0.063	0.329	0.939
Goodness of fit test statistics	Value		
Cox and Snell's R^2	0.021		
Nagelkerke's R^2	0.033		
Hosmer and Lemeshow test statistic	15.3	0.05	

equivalent to saying that higher the value of the sine and the cosine components of α (holding the effect of the other predictor constant in either case), the less likely it is that a person would suffer from a seasonal disease in that month. Both the R^2 values and the insignificant H-L test statistic indicate that the model is a fairly good fit to the data.

5.9. Binary Logistic regression for linear response and circular predictor for season-wise analysis

Similarly, we categorize the individuals and model their season-wise occurrence of disease as response variable Y as in the category mentioned in equation (15).

The circular continuous predictor is the season of a year which is represented in terms of angle as mentioned in section (2).
Table (10) summarizes the results of the Binary Logistic regression for linear response and circular predictor for season-wise analysis of occurrence of seasonal diseases.
Interpretation: Table (10) shows the fitted model to be:

$$\text{predicted } \ln\left(\frac{\pi}{1-\pi}\right) = 1.490 - 0.521\sin\alpha - 0.063\cos\alpha$$

The log of the odds of a person to suffer from seasonal disease in a season is

significantly negatively related to the sine component ($p < 0.05$) of the months modeled as a circular r.v whereas the cosine component is not significantly related to it ($p > 0.05$). In other words, a high value of sine component of α (holding the effect of the other predictor constant in either case) means that a person to suffer from a seasonal disease in that season is less likely. Both the R^2 values and the non-significant value of the H-L statistic indicate that the model is a fairly good fit to the data.

6. Conclusion

In this section, we present the point-wise conclusion of the study on month-wise and season-wise occurrence of seasonal diseases respectively in the Kamrup (rural) district during the years 2013 and 2014:

6.1. Month-wise occurrence of seasonal diseases

- It is clear from Table (1) that the modal month of occurrence of seasonal diseases is February. Thus, the underlying distribution is unimodal in nature (Mardia and Jupp, 2000, pp.6,8)

- From the Table (3), it can be seen that the both the preferred month and median month of occurrence for seasonal diseases is March, as the censored sample mean direction and median direction is 1.27 radian, i.e., 72.77^o and 1.31 radian, i.e., 75.06^o respectively. A value 0.94 of the censored sample variance indicates that the data is highly dispersed. The data is marginally positively skewed as shown by the censored sample skewness measure. Finally, the underlying distribution is found to be platykurtic (marginal negative kurtosis) as the censored sample kurtosis has come to be -0.04. The above two measures jointly indicate that the underlying distribution is nominally well-approximated by the Wrapped Normal Distribution.

- The result of the Rayleigh Uniformity Test for censored sample confirms that there is a month-wise variation in the occurrence of seasonal diseases.

- The logistic regression analysis for censored sample pertaining to month-wise analysis reveals that the likelihood of occurrence of seasonal disease in a month increases from March to June and it decreases from September to December.

6.2. Season-wise occurrence of seasonal diseases

- From Table (2), it can be inferred that the modal season of occurrence of seasonal diseases is Monsoon. Thus, the underlying distribution is unimodal in nature.

- From the Table (4), it can be clearly observed that the both the preferred season and median season of occurrence for seasonal diseases is Pre-monsoon,

as the censored sample mean direction and median direction is 1.35 radian, i.e., 77.35^o and 1.81 radian, i.e., 103.71^o respectively. Since the censored sample variance is 0.93, the data can be said to be highly dispersed. The marginally positively skewness of the data is confirmed by the adjusted censored sample skewness value of 0.09. Finally, as the adjusted censored sample kurtosis has come to be -0.08, the underlying distribution can be inferred to be platykurtic (marginal negative kurtosis) . The above two measures jointly indicate that the underlying distribution is nominally well-approximated by the Wrapped Normal Distribution.

- The result of the Rayleigh Uniformity Test for censored data corresponding to the season-wise analysis shows that there is a season-wise variation of the occurrence of seasonal diseases.

- One can observe from the regression analysis of seasons that the likelihood of occurrence of seasonal diseases decreases from Winter to Pre-monsoon and increases from Pre-monsoon to Post-monsoon.

Further, the statistical nature of the underlying distribution of the dataset is found to be similar, both through the month-wise and season-wise analysis. This establishes the validity of the adjustment made to the censored sample for carrying out the season-wise analysis.

Acknowledgements

Kishore Kumar Das, Principal Investigator of the project entitled "Statistical Modeling in Circular Statistics: An Application to Health Science " acknowledges the financial support received from the UGC, India. The unpublished data has been taken from the same. Sahana Bhattacharjee has been granted a Junior Research Fellowship for pursuing full-time Doctoral (Ph.D) Program under the Innovation in Science Pursuit for Inspired Research (INSPIRE) programme sponsored by the Department of Science and Technology, New Delhi. The authors would like to express their gratitude towards the anonymous referees for their helpful comments and suggestions which has helped improve the clarity of the paper.

Appendix

A. Proofs

Theorem 13. *In the Rayleigh's test of Uniformity for Censored Circular Data, under the hypothesis of uniformity, the asymptotic distribution of the test statistic* $2n\bar{R}_c^2$ *is* χ_2^2*, where the symbols are as explained in section 3.*

Proof. By the virtue of Multivariate Central Limit Theorem, for large n, we have

$$\begin{pmatrix} \bar{C}_c \\ \bar{S}_c \end{pmatrix} \sim N\left[E\begin{pmatrix} \bar{C}_c \\ \bar{S}_c \end{pmatrix}, \begin{pmatrix} Var(\bar{C}_c) & Cov(\bar{C}_c, \bar{S}_c) \\ Cov(\bar{C}_c, \bar{S}_c) & Var(\bar{S}_c) \end{pmatrix}\right]$$

Let α'_{ij}'s denote the censored sample observations. Under the hypothesis of uniformity of α'_{ij}'s, i.e., under the assumption that the α'_{ij}'s have come from Circular Uniform distribution, the expectations, variances and covariances of $\bar{C}_c$ and $\bar{S}_c$ are as obtained below:

$$
\begin{aligned}
E(\bar{C}_c) &= E\left(\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}(\cos\beta_{cij})\right) \\
&= E\left(\frac{1}{n}\sum_{i}\sum_{j}\left(\cos\alpha'_{ij}\right)\right) \text{(summation being over only the } i \text{ and } j \\
&\qquad\qquad \text{corresponding to the observations in the censored sample)} \\
&= \int_0^{2\pi}\frac{1}{n}\sum_{i}\sum_{j}\left(\cos\alpha'_{ij}\right)\frac{1}{2\pi}d\alpha'_{ij} \\
&= \frac{1}{n}\sum_{i}\sum_{j}\frac{1}{2\pi}\int_0^{2\pi}\left(\cos\alpha'_{ij}\right)d\alpha'_{ij} \\
&= \frac{1}{n}\sum_{i}\sum_{j}\frac{1}{2\pi}\left(\sin 2\pi - \sin 0\right) \\
&= 0
\end{aligned}
$$

Similarly, it can be shown that

$$E(\bar{S}_c) = 0$$

$$
\begin{aligned}
Var(\bar{C}_c) &= Var\left(\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}(\cos\beta_{cij})\right) \\
&= Var\left(\frac{1}{n}\sum_{i}\sum_{j}\left(\cos\alpha'_{ij}\right)\right) \\
&= E\left\{\left(\frac{1}{n}\sum_{i}\sum_{j}\left(\cos\alpha'_{ij}\right)\right)\right\}^2 - \left[E\left(\frac{1}{n}\sum_{i}\sum_{j}\left(\cos\alpha'_{ij}\right)\right)\right]^2 \\
&= E\left\{\left(\frac{1}{n^2}\sum_{i}\sum_{j}\left(\cos^2\alpha'_{ij}\right)\right) + 2\left(\frac{1}{n^2}\sum_{i\neq i\prime, j\neq j\prime}\cos\alpha'_{ij}\cos\alpha'_{i\prime j\prime}\right)\right\} - 0 \\
&= \frac{1}{n^2}\sum_{i}\sum_{j}E\left(\cos^2\alpha'_{ij}\right) + \frac{2}{n^2}\sum_{i\neq i\prime, j\neq j\prime}E\left(\cos\alpha'_{ij}\cos\alpha'_{i\prime j\prime}\right) \\
&= \int_0^{2\pi}\frac{1}{n^2}\sum_{i}\sum_{j}\left(\cos^2\alpha'_{ij}\right)\frac{1}{2\pi}d\alpha'_{ij} + \frac{2}{n^2} \\
&\quad \sum_{i\neq i\prime, j\neq j\prime}E\left(\cos\alpha'_{ij}\right)E\left(\cos\alpha'_{i\prime j\prime}\right) \ (\because \alpha'_{ij}\text{'s are independent}) \\
&= \frac{1}{2n^2}\frac{1}{2\pi}\sum_{i}\sum_{j}\int_0^{2\pi}\left(1+\cos 2\alpha'_{ij}\right)d\alpha'_{ij} + 0 \\
&= \frac{1}{2n^2}\frac{1}{2\pi}\sum_{i}\sum_{j}\int_0^{2\pi}d\alpha'_{ij} + \frac{1}{2n^2}\frac{1}{2\pi}\sum_{i}\sum_{j}\int_0^{2\pi}\cos 2\alpha'_{ij}d\alpha'_{ij} \\
&= \frac{1}{2n^2}\frac{1}{2\pi}\sum_{i}\sum_{j}(2\pi-0) + \frac{1}{2n^2}\frac{1}{2\pi}\sum_{i}\sum_{j}\frac{1}{2}\left\{\sin 2\left(2\pi\right) - \sin 0\right\} \\
&= \frac{1}{2n^2}\frac{1}{2\pi}\ n\ 2\pi \\
&= \frac{1}{2n}
\end{aligned}
$$

Similarly, we have

$$Var(\bar{S}_c) = \frac{1}{2n}$$

Again,

$$\begin{aligned}
Cov(\bar{C}_c, \bar{S}_c) &= Cov\left\{\left(\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}(\cos\beta_{cij})\right), \left(\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}(\sin\beta_{sij})\right)\right\} \\
&= Cov\left\{\left(\frac{1}{n}\sum_{i}\sum_{j}(\cos\alpha'_{ij})\right), \left(\frac{1}{n}\sum_{i}\sum_{j}(\sin\alpha'_{ij})\right)\right\} \\
&= E\left\{\left(\frac{1}{n}\sum_{i}\sum_{j}(\cos\alpha'_{ij})\right).\left(\frac{1}{n}\sum_{i}\sum_{j}(\sin\alpha'_{ij})\right)\right\} \\
&- E\left(\frac{1}{n}\sum_{i}\sum_{j}(\cos\alpha'_{ij})\right)E\left(\frac{1}{n}\sum_{i}\sum_{j}(\sin\alpha'_{ij})\right) \\
&= \frac{1}{n^2}E\left\{\sum_{i}\sum_{j}(\cos\alpha'_{ij})\sum_{i}\sum_{j}(\sin\alpha'_{ij})\right\} - 0 \\
&= \frac{1}{n^2}E\left\{\sum_{i}\sum_{j}(\cos\alpha'_{ij})(\sin\alpha'_{ij})\right\} \\
&\qquad (\because \text{Other terms get cancelled as } \alpha'_{ij}\text{'s are independent}) \\
&= \frac{1}{2n^2}E\left\{\sum_{i}\sum_{j}\sin 2\alpha'_{ij}\right\} = \frac{1}{2n^2}\int_0^{2\pi}\frac{1}{2\pi}\sum_{i}\sum_{j}\sin 2\alpha'_{ij}d\alpha'_{ij} \\
&= \frac{1}{2n^2}\frac{1}{2\pi}\sum_{i}\sum_{j}\int_0^{2\pi}\sin 2\alpha'_{ij}d\alpha'_{ij} \\
&= \frac{1}{4n^2}\frac{1}{2\pi}\sum_{i}\sum_{j}[-\{\cos 2(2\pi) - \cos 0\}] = 0
\end{aligned}$$

Thus, we see that for large n,

$$\begin{pmatrix} \bar{C}_c \\ \bar{S}_c \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{1}{2n} & 0 \\ 0 & \frac{1}{2n} \end{pmatrix}\right]$$

$$\begin{aligned}
\therefore \left(\frac{\bar{C}_c - 0}{\frac{1}{\sqrt{2n}}}\right)^2 + \left(\frac{\bar{S}_c - 0}{\sqrt{2n}}\right)^2 &\sim \chi^2_2 \\
\Rightarrow 2n\left(\bar{C}_c^2 + \bar{S}_c^2\right) &\sim \chi^2_2 \\
\Rightarrow 2n\bar{R}_c^2 &\sim \chi^2_2
\end{aligned}$$

Hence the proof.

THEOREM 14. *For the censored sample pertaining to the season-wise analysis, which requires adjustment for grouping, the adjusted test statistic is* $\frac{2n}{\{a(h)\}^2}\bar{R}_c^{*^2}$ *which is asymptotically distributed as* χ_2^2.

PROOF. The Multivariate Central Limit Theorem states that for large n,

$$\begin{aligned}\begin{pmatrix}\bar{C}_c\\ \bar{S}_c\end{pmatrix} &\sim N\left[E\begin{pmatrix}\bar{C}_c\\ \bar{S}_c\end{pmatrix}, \begin{pmatrix}V(\bar{C}_c) & Cov(\bar{C}_c,\bar{S}_c)\\ Cov(\bar{C}_c,\bar{S}_c) & V(\bar{S}_c)\end{pmatrix}\right]\\ \Longrightarrow \begin{pmatrix}a(h)\bar{C}_c\\ a(h)\bar{S}_c\end{pmatrix} &\sim N\left[a(h)\,E\begin{pmatrix}\bar{C}_c\\ \bar{S}_c\end{pmatrix}, \{a(h)\}^2\begin{pmatrix}V(\bar{C}_c) & Cov(\bar{C}_c,\bar{S}_c)\\ Cov(\bar{C}_c,\bar{S}_c) & V(\bar{S}_c)\end{pmatrix}\right]\end{aligned}$$

It then follows from the relation $\bar{C}_c^* = a(h)\bar{C}_c$, $\bar{S}_c^* = a(h)\bar{S}_c$ and the previous proof that

$$\begin{pmatrix}\bar{C}_c^*\\ \bar{S}_c^*\end{pmatrix} \sim N\left[\begin{pmatrix}0\\ 0\end{pmatrix}, \begin{pmatrix}\frac{\{a(h)\}^2}{2n} & 0\\ 0 & \frac{\{a(h)\}^2}{2n}\end{pmatrix}\right]$$

$$\begin{aligned}\therefore \left(\frac{\bar{C}_c^*-0}{\frac{a(h)}{\sqrt{2n}}}\right)^2 + \left(\frac{\bar{S}_c^*-0}{\frac{a(h)}{\sqrt{2n}}}\right)^2 &\sim \chi_2^2\\ \Longrightarrow \frac{2n}{\{a(h)\}^2}\left(\bar{C}_c^{*^2}+\bar{S}_c^{*^2}\right) &\sim \chi_2^2\\ \Longrightarrow \frac{2n}{\{a(h)\}^2}\bar{R}_c^{*^2} &\sim \chi_2^2\end{aligned}$$

Hence the proof.

LEMMA 15.

$$\sum_{i=1}^{k}\sum_{j=1}^{f_i}\sin I_s(\alpha_{ij}-\bar{\alpha}_{c0}) = 0 \qquad (9)$$

PROOF.

$$\begin{aligned}\bar{S}_c &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\sin(\beta_{sij})\\ &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\sin(I_s\alpha_{ij})\\ &= \frac{1}{n}\sum_i\sum_j\sin(\alpha'_{ij})\\ \Longrightarrow n\bar{S}_c &= \sum_i\sum_j\sin(\alpha'_{ij})\end{aligned}$$

α'_{ij} being the observations of interest in the sample. Again,

$$\begin{aligned} \bar{C}_c &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\cos\left(\beta_{cij}\right) \\ &= \frac{1}{n}\sum_{i}\sum_{j}\cos\left(\alpha'_{ij}\right) \\ \implies n\bar{C}_c &= \sum_{i}\sum_{j}\cos\left(\alpha'_{ij}\right) \end{aligned}$$

Again, as $\bar{\alpha}_{c0}$ is the direction of the resultant vector $\bar{R}_c$ of the censored sample, we have, from the theory of Vector Algebra,

$$\bar{C}_c = \bar{R}_c\cos\bar{\alpha}_{c0} \text{ and } \bar{S}_c = \bar{R}_c\sin\bar{\alpha}_{c0}.$$

$\therefore$

$$\begin{aligned} \sum_{i=1}^{k}\sum_{j=1}^{f_i}\sin I_s\left(\alpha_{ij}-\bar{\alpha}_{c0}\right) &= \sum_{i}\sum_{j}\sin\left(\alpha'_{ij}-\bar{\alpha}_{c0}\right) \\ &= \sum_{i}\sum_{j}\sin\alpha'_{ij}\cos\bar{\alpha}_{c0} - \sum_{i}\sum_{j}\cos\alpha'_{ij}\sin\bar{\alpha}_{c0} \\ &= n\bar{S}_c\cos\bar{\alpha}_{c0} - n\bar{C}_c\sin\bar{\alpha}_{c0} \\ &= n\bar{R}_c\sin\bar{\alpha}_{c0}\cos\bar{\alpha}_{c0} - n\bar{R}_c\cos\bar{\alpha}_{c0}\sin\bar{\alpha}_{c0} \\ &= 0 \end{aligned}$$

Hence the proof.

LEMMA 16.

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\cos\left(\beta_{cij}-\bar{\alpha}_{c0}\right) = \bar{R}_c \tag{10}$$

PROOF. It follows from the previous proof that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{f_i}\cos\left(\beta_{cij}-\bar{\alpha}_{c0}\right) &= \frac{1}{n}\sum_{i}\sum_{j}\cos\left(\alpha'_{ij}-\bar{\alpha}_{c0}\right) \\ &= \frac{1}{n}\sum_{i}\sum_{j}\cos\alpha'_{ij}\cos\bar{\alpha}_{c0} + \frac{1}{n}\sum_{i}\sum_{j}\sin\alpha'_{ij}\sin\bar{\alpha}_{c0} \\ &= \bar{C}_c\cos\bar{\alpha}_{c0} + \bar{S}_c\sin\bar{\alpha}_{c0} \\ &= \bar{R}_c\cos^2\bar{\alpha}_{c0} + \bar{R}_c\sin^2\bar{\alpha}_{c0} \\ &= \bar{R}_c \end{aligned}$$

Hence the proof.

References

I. Baruah, N. G. Das, J. Kalita (2007). *Seasonal prevalence of malaria vectors in sonitpur district of assam, india.* Journal of Vector Borne Diseases, 44, pp. 149–153.

C. F. Christiansen, L. Pedersen, H. T. Sorensen, K. J. Rothman (2012). *Methods to assess seasonal effects in epidemiological studies of infectious diseases - exemplified by application to the occurrence of meningococcal disease.* Clinical Microbiology and Infection, 18, no. 10, pp. 963–969.

H. A. David, D. J. Newell (1965). *The identification of annual peak periods for a disease.* Biometrics, 21, pp. 645–650.

J. H. Edwards (1961). *The recognition and estimation of cyclic trends.* Ann Hum Genet, 25, pp. 83–86.

N. I. Fisher (1993). *Statistical Ananlysis of Circular Data.* Cambridge University Press.

N. C. Grassly, C. Fraser (2006). *Seasonal infectious disease epidemiology.* Proceedings B of The Royal Society, 273, p. 1600.

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

J. C. Peng, K. L. Lee, G. M. Ingersoll (2002). *An introduction to logistic regression analysis and reporting.* Journal of Educational Research, 96, pp. 3–14.

J. S. Rao, A. SenGupta (2001). *Topics in Circular Statistics.* World Scientific Publishing Co. Pte. Ltd., Singapore.

R. L. Singh (1993). *India, A Regional Geography.* National Geographical Society of India, India.

A. Stuart, J. K. Ord (1987). *Kendall's Advanced Theory of Statistics.* Vol. 1: Distribution Theory. Wiley, Griffin, London.

V. Statistics (2011-12). *AHS Fact Sheet Assam, Annual Health Survey Fact Sheet, Assam.* Vital Statistics Division, Office of the Registrar General and Census Commissioner, New Delhi, India.

V. Statistics (2012-13). *AHS Fact Sheet Assam, Annual Health Survey Fact Sheet, Assam.* Vital Statistics Division, Office of the Registrar General and Census Commissioner, New Delhi, India.

Summary

In this paper, we aim to analyse the occurrence of seasonal diseases in the Kamrup (rural) district of Assam during the years 2013 and 2014. This leads us to work with a censored circular sample containing only the observations of interest and thus, develop new circular descriptive statistics for analysis of Censored Circular sample, both month-wise and season-wise. Since the seasons differ by a significant length, we propose to group the cases in unequal intervals, the width of the intervals being proportional to the length of the seasons. The Rayleigh Uniformity Test has also been proposed for the censored sample, using which the presence of seasonal effect in both month-wise and season-wise occurrence is assessed. Finally, a logistic regression model for predicting binary response from circular predictor has been proposed to predict the occurrence of seasonal diseases from months and seasons. It is revealed that the occurrence of seasonal diseases is highest in the month of February or equivalently, during the Monsoon season. The distribution of occurrence of seasonal diseases both month-wise and season-wise is found to be marginally positively skewed and platykurtic, indicating that it can be moderately well-approximated by a wrapped normal distribution. Rayleigh Uniformity Test results for both month and season wise analysis suggest the presence of seasonal effect. The regression analysis shows that likelihood of occurrence of seasonal disease increases from March to June or Winter to Pre-monsoon and decreases from September to December or Pre-monsoon to Post-monsoon.

Keywords: Censoring; Rayleigh Uniformity Test; Binary Logistic Regression