CHAPTER 2

CIRCULAR DATA

2.1 Introduction

In this chapter, we review the definition of circular data. Circular data can be visualized as being distributed on the circumference of a unit circle in the range of 0 to 2π radian. The data are commonly found in many scientific fields such as meteorology and biology where researchers are interested in studying direction of wind and direction of movement of animals, respectively. In this chapter, we give explanations and present some example of circular descriptive statistics. Briefly, we introduce several distributions of circular data and review several appropriate circular plots for circular data and review goodness of fit (GOF) tests for the circular distributions.

2.2 Descriptive Statistics of Univariate Circular Data

Descriptive statistics can be used to summarize and describe data. Several descriptive statistics are often used at one time to give a full picture of the data. Some commonly used descriptive statistics for circular data are described below (see Fisher (1993)):

i. The mean direction, $\overline{\theta}$

Let $\theta_1,...\theta_n$ be a sample of circular data. The mean direction is defined by the angle made by the resultant vector with the horizontal line. Specifically, we have the resultant length of the resultant vector, *R* given by

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

where $C = \sum_{i=1}^{n} \cos\theta_i$ and $S = \sum_{i=1}^{n} \sin\theta_i$. The mean direction is given by

$$\bar{\theta} = \begin{cases} \tan^{-1}(\mathrm{S/C}), & \text{if } \mathrm{S} \geq 0, \mathrm{C} > 0, \\ \frac{\pi}{2}, & \text{if } \mathrm{S} > 0, \mathrm{C} = 0, \\ \tan^{-1}(\mathrm{S/C}) + \pi, & \text{if } \mathrm{C} < 0, \\ \tan^{-1}(\mathrm{S/C}) + 2\pi & \text{if } \mathrm{S} < 0, \mathrm{C} \geq 0, \\ \text{undefined}, & \text{if } \mathrm{S} = 0, \mathrm{C} = 0. \end{cases}$$

ii. Mean resultant length, $\bar{R}$

Mean resultant length is useful for unimodal data to measure how concentrated the data is towards the centre. It is defined by $\bar{R} = \frac{R}{n}$ and lies in the range $[0,1]$. When $\bar{R}$ is close to 1, all directions in the data set are almost similar. The data is said to have small dispersion and is more concentrated towards the centre.

iii. The median direction

Mardia and Jupp (2000) defined the median as any point ϕ, where half of the data lie in the arc $[\phi, \phi + \pi)$and the other points are nearer to ϕ than to $\phi + \pi$. Basically, for any circular sample, Fisher (1993) defined the median direction as the observation ϕ which minimizes the summation of circular distances to all observations, $d(\phi) = \pi - \sum_{i=1}^{n} \left|\pi - |\theta_i - \phi|\right|$ for i = 1,..., n. Fisher's definition is used to obtain the circular median in the Oriana statistical software package.

iv. The sample circular variance

The sample circular variance is defined by the quantity $V = 1 - \bar{R}$, where $0 \le V \le 1$. The smaller values of circular variance refer to a more concentrated sample. The sample circular standard deviation is defined by the quantity

$$v = \sqrt{-2\log(1-V)}$$

$$= \sqrt{-2\log \bar{R}} \qquad , \quad 0 < v < \infty$$

v. The circular range

The circular range is the length of the smallest arc which contains all the observations. Consider $\theta_1, \theta_2, \ldots, \theta_n$ in the range $0 < \theta_i \le 2\pi$. Let $\theta_{(1)} \le \ldots \le \theta_{(n)}$ be the linear order statistics of $\theta_1, \theta_2, \ldots, \theta_n$. The arc lengths between adjacent observations are

$$T_i = \theta_{(i+1)} - \theta_{(i)}, \quad i = 1, \ldots, n-1; \quad T_n = 2\pi - \theta_{(n)} + \theta_{(1)}.$$

The circular range w is

$$w = 2\pi - \max\left(T_1, \ldots, T_n\right)$$

vi. The concentration parameter

The concentration parameter, denoted by κ, is a standard measure of dispersion for circular data. Best and Fisher (1981) gave the maximum likelihood estimates of the concentration paramater κ as follows

$$\hat{\kappa} = \begin{cases} 2\bar{R} + \bar{R}^3 + \frac{5}{6}\bar{R}^5, & \text{if} \quad \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + \frac{0.43}{(1-\bar{R})}, & \text{if} \quad 0.53 \le \bar{R} < 0.85 \\ \left(\bar{R}^3 - 4\bar{R}^2 + 3\bar{R}\right)^{-1}, & \text{if} \quad \bar{R} \ge 0.85 \end{cases}$$

where $\bar{R}$ is mean resultant length.

vii. The circular quantile

The first and third quantile directions Q_1 and Q_3 is any solution of

$$\int_{\phi - Q_1}^{\phi} f(\theta)d\theta = 0.25 \text{ and } \int_{\phi}^{\phi + Q_3} f(\theta)d\theta = 0.25$$

respectively and ϕ is a median direction. Q_1 can be considered as the median of the first half of the ordered data and Q_3 as the median of the second.

2.3 Circular Graphs

There are several plots available for circular data, such as rose diagram, circular histogram, arrow histogram, raw data and simple circular plots. These plots is able to give a general picture of the data set such as the distribution of the data, the circular mean and its 95% confidence interval as well as the possible occurrence of outliers. The plots can be obtained from the Oriana statistical software. As an illustration, some plots of wind direction data (Fisher's (1993)) are shown in Figures 2.1-2.4.

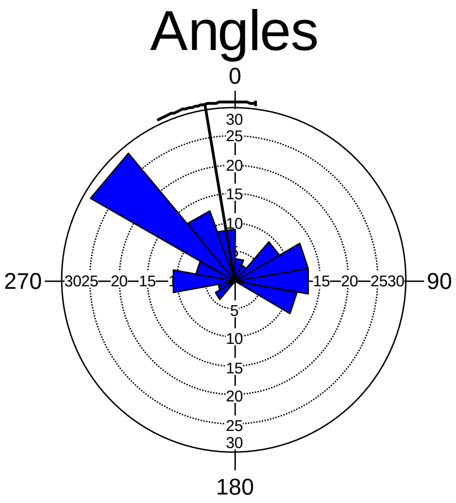

Figure 2.1: Rose histogram

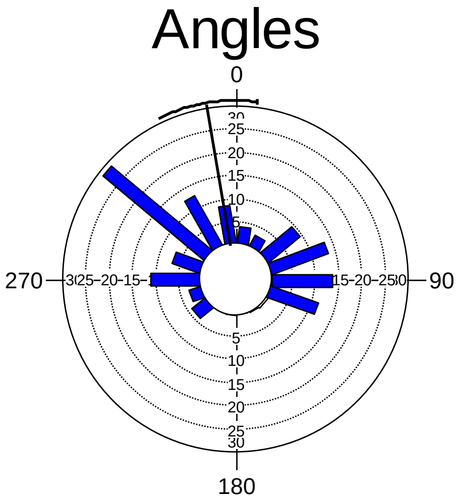

Figure 2.2: Circular histogram

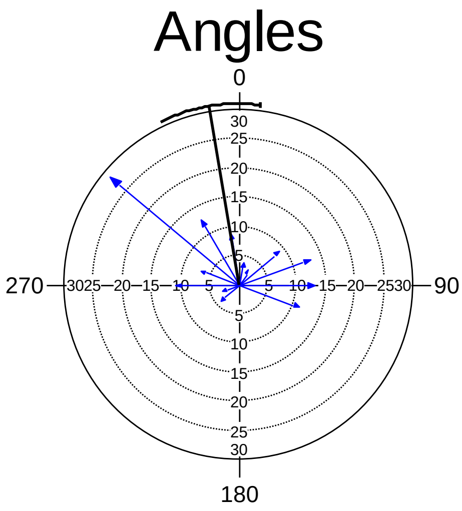

Figure 2.3: Arrow histogram

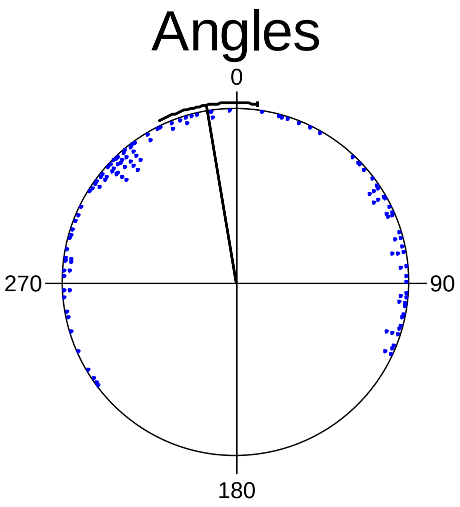

Figure 2.4: Raw data plot

Figure 2.1 shows the rose histogram which illustrates the frequencies of all observations in the given interval with the mean direction being 350° and the 95% confidence interval for the mean direction $[330°, 10°]$. The shape of rose histogram is actually based on the pattern of the rose flower. The circular histogram and arrow histogram given in Figure 2.2 and Figure 2.3 respectively are similar to the rose histogram but the frequency for circular histogram is represented by rectangular shape while that for arrow histogram is by arrows. On the other hand, the raw data plot in Figure 2.4 is the analogue of a simple dot plot illustrating the locations of all observations with mean direction and the 95% confidence interval for the mean direction included.

2.4 The Distributions of Circular Data

A circular distribution is a probability distribution whose total probability is concentrated on the circumference of a unit circle. Each point on the circumference represents a direction. The range of a circular random variable θ, measured in radians, may be taken to be $[0, 2\pi)$ or $[-\pi, \pi)$. Circular distributions are essentially of two types; they may be discrete, assigning probability masses only to a countable number of directions, or may be absolutely continuous with respect to the measure on the circumference of a unit circle.

Various distributions are available for circular data, for example, uniform distribution, wrapped Cauchy distribution, wrapped normal distribution, cardioid distribution, and others. Jammalamadaka and SenGupta (2001) reviewed the wrapped α stable distribution with the wrapped Cauchy and the wrapped normal distributions as the special cases. On the other hand, several bivariate circular distributions exist, such as the bivariate von Mises distribution, wrapped bivariate normal distribution and circular-linear distribution.

Several reviews have comprehensively discussed the circular distributions including Jammalamadaka and SenGupta (2001), Mardia (1972) and Fisher (1993). The von Mises distribution (also known as the circular normal distribution) is the most commonly used which is a continuous probability distribution on a circle. The von

Mises distribution may be thought of as a close approximation to the wrapped normal distribution, which is the circular analogue of the normal distribution.

2.4.1 The circular uniform distribution

The uniform distribution is a basic distribution on the circle. For this distribution, all directions of the data are equally likely; it is also known as random distribution. There is no certain concentration towards one or more preferred directions. The probability density function is given by

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

where $(0 \le \theta < 2\pi)$ and denoted by U_c. Furthermore, Mardia and Jupp (1972) stated that the circular uniform distribution has a unique property such that if θ_1 is uniformly distribution and θ_2 is chosen from any distribution, θ_1 and θ_2 are independently distributed then $\theta_1 + \theta_2$ is also distributed uniformly.

2.4.2 The von Mises distribution

The von Mises (*VM*) distribution was introduced by von Mises (1918) to study the deviations of measured atomic weight from integral values. The *VM* distribution has been extensively discussed where many inference techniques have been developed. The *VM* distribution is denoted by $VM(\mu, \kappa)$, where μ is the mean direction and κ is the concentration parameter. The probability distribution function for *VM* distribution is given by

$$f(\theta) = \left[2\pi I_0(\kappa)\right]^{-1} \exp\left[\kappa \cos(\theta - \mu)\right] \qquad 0 \le \theta < 2\pi, \quad 0 \le \kappa \le \infty$$

where

$$I_0(\kappa) = (2\pi)^{-1} \int_0^{2\pi} \exp[\kappa \cos(\vartheta - \mu)] d\vartheta$$

is the modified Bessel function of order zero; a series expansion and method for evaluating $I_0(\kappa)$ is given by Fisher (1993). The distribution function of VM is given by

$$F(\theta) = [2\pi I_0(\kappa)]^{-1} \int_0^{\theta} \exp[\kappa \cos(\alpha - \mu)] d\alpha \qquad , \alpha \in [0, 2\pi)$$

The parameter μ is the mean. As the parameter $\kappa \to 0$, the distribution converges to the uniform distribution U_c, while if $\kappa \to \infty$, the distribution tends to the point distribution concentrated in the direction μ.

As an illustration, we generate data from von Misses with different value of $\kappa = 3, 5, 10$ and fix the sample size n = 20 and the mean direction μ = 0. The data is given in Appendix 1 and the circular plots of the data are shown in Figures 2.5-2.7. All the three plots suggest that as κ increases, the generated data sets are more concentrated in the direction $\mu = 0$.

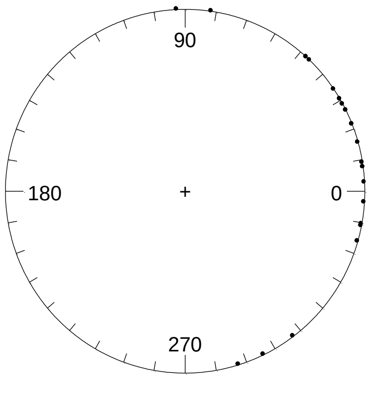

Figure 2.5: $VM(n = 20, \mu = 0, \kappa = 3)$

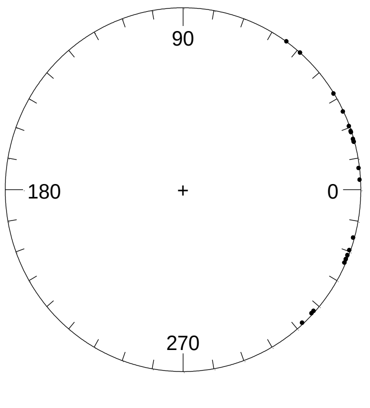

Figure 2.6: $VM(n = 20, \mu = 0, \kappa = 5)$

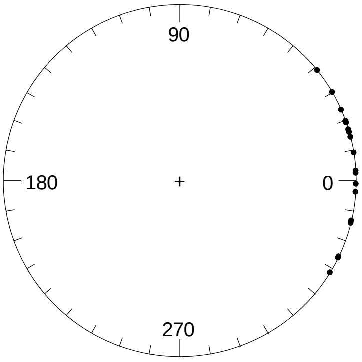

Figure 2.7: $VM(n = 20, \mu = 0, \kappa = 10)$

2.4.3 The general Wrapped Stable (*WS*) family distribution

Jammalamadaka and SenGupta (2001) discussed the general wrapped α-stable distribution which is constructed by using the characteristic function of the α-stable of the real line. The characteristic function as given by Lukacs (1970) is

$$\psi(t) = \begin{cases} \exp\left\{-\tau^{\alpha}|t|^{\alpha}\left[1 - i\beta\,\mathrm{sgn}(t)\tan\frac{\alpha\pi}{2}\right] + i\mu t\right\}, & \text{if } \alpha \in (0,1)\cup(1,2], \\ \exp\{-\tau|t| + i\mu t\}, & \text{if } \alpha = 1, \end{cases}$$

where $i = \sqrt{-1}$, $\tau \geq 0$, $|\beta| \leq 1$, $0 < \alpha \leq 2$, while μ is a real number. The density function of a wrapped α-stable random variable for $\theta \in [0,2\pi)$ is given by

$$f(\theta) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_{k=1}^{\infty}\exp\left\{-\tau^{\alpha}k^{\alpha}\right\}\cos\left\{k(\theta - \mu) - \tau^{\alpha}k^{\alpha}\beta\tan\frac{\alpha\pi}{2}\right\},$$

when $\alpha \in (0,1)\cup(1,2]$, with μ conveniently redefined as μ (mod 2π). Note that although there is generally no closed form expression for the density of an α-stable distribution on the real line, we are able to write such density for the wrapped case, at least as an infinite series.

The particular case corresponding to $\beta = 0$ gives us the symmetric wrapped stable (*SWS*) family of circular densities, which we will simply refer to as wrapped stable (*WS*), given by

$$f(\theta) = \frac{1}{2\pi} + \frac{1}{\pi}\sum_{k=1}^{\infty} \rho^{k^{\alpha}} \cos\{k(\theta - \mu)\},$$

where $\rho = \exp(-\tau^{\alpha})$. We shall denote such distributions as $WS(\alpha, \rho, \mu)$. The special case with $\alpha = 2$ and $\beta = 0$ gives us the wrapped normal density with $\rho = \exp(-\tau^2)$. When $\alpha = 1$and $\beta = 0$, it gives us the wrapped Cauchy density with $\rho = \exp(-\tau)$.

2.4.4 The wrapped Cauchy (*WC*) distribution

The wrapped Cauchy distribution is obtained by wrapping the Cauchy distribution on the real line around a unit circle. The wrapped Cauchy distribution is denoted by $WC(\mu, \rho)$, where μ is the mean direction and ρ is another measure of concentration parameter such that $\rho = A_1(\kappa)$, $A_1(\kappa) \equiv \frac{I_1(\kappa)}{I_0(\kappa)}$ is the ratio of two modified Bessel functions. The probability distribution function for the wrapped Cauchy distribution is given by

$$f(\theta) = \frac{1}{2\pi}\frac{1-\rho^2}{1+\rho^2 - 2\rho\cos(\theta - \mu)} \qquad 0 \le \theta < 2\pi, \quad 0 \le \rho \le 1$$

while the distribution function is given by

$$F(\theta) = \frac{1}{2\pi}\cos^{-1}\left(\frac{(1+\rho^2)\cos(\theta - \mu) - 2\rho}{1+\rho^2 - 2\rho\cos(\theta - \mu)}\right) \qquad 0 \le \theta < 2\pi,$$

As $\rho \to 0$, the distribution converges to a uniform distribution. On the other hand, as $\rho \to 1$, the distribution tends to a point distribution concentrated in the direction μ.

As an illustration, we generate data from the wrapped Cauchy distribution with different values of $\rho = 0.3, 0.7, 0.975$ and fix the sample size n = 20 and the mean direction μ = 0. The data sets are given in Appendix 2 and the plots of the data are in Figures 2.8-2.10. Generally, the generated data sets show similar behaviour as that of the von Mises distribution. When the measure of concentration parameter ρ gets close to 1, the distribution tends to a point distribution concentrated in the direction $\mu = 0$. However, since the wrapped Cauchy distribution is a heavy tail distribution, the generated data set has observations which are located further away from the rest, even for high value of ρ (see Figure 2.8).

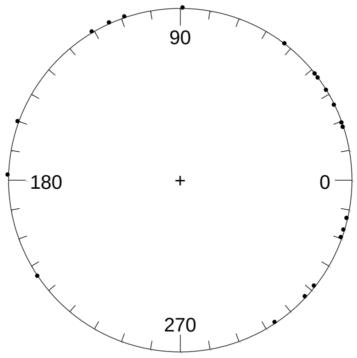

Figure 2.8: $WC(n = 20, \mu = 0, \rho = 0.3)$

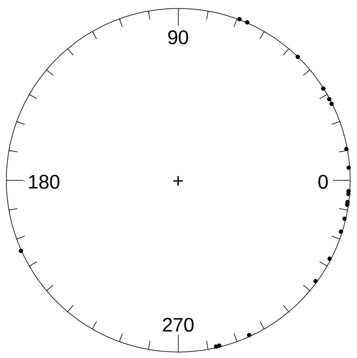

Figure 2.9: $WC(n = 20, \mu = 0, \rho = 0.7)$

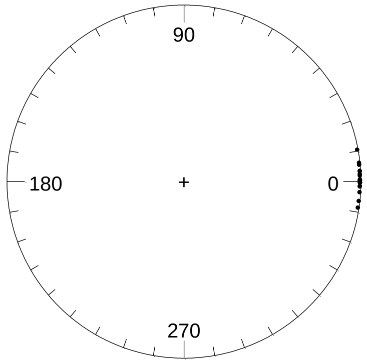

Figure 2.10: $WC(n = 20, \mu = 0, \rho = 0.975)$

2.4.5 The wrapped normal (*WN*) distribution

A wrapped normal distribution is obtained by wrapping a normal distribution around a unit circle. The normal distribution is denoted by $N(\mu_L, {\sigma_L}^2)$ where μ_L is the mean and ${\sigma_L}^2$ is the variance while the *WN* distribution is denoted by $WN(\mu, \rho)$, where μ is the mean direction and ρ is the measure of concentration parameter. Its probability distribution function is given by

$$f(\theta) = \frac{1}{\sigma\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left[\frac{-(\theta - \mu - 2k\pi)^2}{2\sigma^2}\right]$$

where σ^2 is the circular variance.

From Whittaker and Watson (1944), an alternative and more useful representation of this density is

$$f(\theta) = \frac{1}{2\pi}\left(1 + 2\sum_{k=1}^{\infty} \rho^{k^2} \cos k(\theta - \mu)\right), \quad 0 \le \theta < 2\pi \ , \qquad 0 \le \rho \le 1$$

The distribution is unimodal and symmetric about the value $\theta = \mu$. Unlike the von Mises distribution, the *WN* distribution possesses the additive property, that is, the convolution of two *WN* variables is also *WN*. Specifically, if $\theta_1 \sim WN(\mu_1, \rho_1)$, $\theta_2 \sim WN(\mu_2, \rho_2)$, and are independent, then $\theta_1 + \theta_2 \sim WN(\mu_1 + \mu_2, \rho_1\rho_2)$ (see Jammalamadaka and SenGupta (2001)).

For illustration, we generate data from the wrapped normal with different values of $\rho = 0.3, 0.7, 0.975$ and fix the sample size *n* = 20 and the mean direction $\mu = 0$. The data set are given in Appendix 3 and the circular plots of the data are in Figures 2.11-2.13. Unlike the wrapped Cauchy distribution, the behaviour of the wrapped normal distribution is closer to that of the von Mises distribution. As ρ increase from 0 to 1, the points are more concentrated in the direction $\mu = 0$. As the value of ρ gets smaller, the generated data set tends to be uniformly distributed.

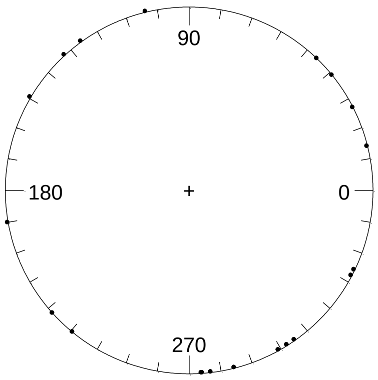

Figure 2.11: $WN(n = 20, \mu = 0, \rho = 0.3)$

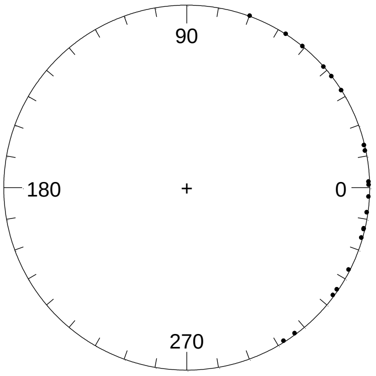

Figure 2.12: $WN(n = 20, \mu = 0, \rho = 0.7)$

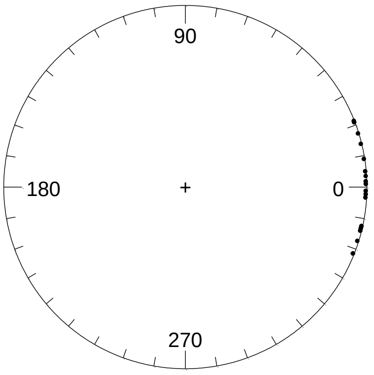

Figure 2.13: $WN(n = 20, \mu = 0, \rho = 0.975)$

2.4.6 Discussion

From the circular plots in Figures 2.5-2.7, we can see that as the concentration parameter κ increases, data from the von Mises distribution are more concentrated around the mean direction $\mu = 0$. Similarly for the wrapped Cauchy distribution and the wrapped normal distribution, as the measure of concentration parameter ρ increases, the data will be closer to the given mean direction $\mu = 0$. However, the wrapped Cauchy distribution has a long tail even for a data set with high concentration value ρ. Note that κ and ρ have a relationship such that $\kappa = A_1^{-1}(\rho)$, where the function $A_1(\rho) \equiv \frac{I_1(\rho)}{I_0(\rho)}$ is the ratio of two modified Bessel function. Appendix 4 gives the list of values of κ and the corresponding values of ρ.

2.5 Goodness of Fit Test

In the case of circular data, Watson (1961) proposed a test for goodness of fit U^2 of the von Mises distribution. However, the test can be extended to other circular distributions. Let $F_\kappa(\theta)$ be the distribution function of the von Mises distribution which is given by

$$z_i = F_\kappa(\theta_i) = \{2\pi I_0(\kappa)\}^{-1} \int_0^{\theta_i} e^{(\kappa \cos\theta)} d\theta .$$

The test statistic U^2 is given by

$$U^2 = \sum_{i=1}^n z_i^2 - \sum_{i=1}^n \left(\frac{c_i z_i}{n} \right) + n\left[\frac{1}{3} - (\bar{z} - \frac{1}{2})^2 \right] \qquad (2.1)$$

where $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$ and $c_i = 2i - 1$. The cut-off points were supplied by Stephens (1964). Then, we can plot a graph of the quantiles distribution of the data against the quantile of the von Mises distribution. If the quantiles are close to the straight line and the test statistic is smaller than the cut-off point, we can conclude that the data follow a von Mises distribution.

Brown (1994) obtained an alternative method which includes careful consideration of grouping effect. He first defined

$$Y_j = \sum_{i=1}^{j-1} (O_i - E_i) + \frac{1}{2}(O_j - E_j) \qquad (2.2)$$

where O_i is the ith observed value and E_i is the ith expected value. Then Brown's grouped version of U^2 is

$$U_d^2 = \frac{1}{n}\left\{\sum_{j=1}^{k} p_j Y_j^2 - \left(\sum_{j=1}^{k} p_j Y_j\right)^2\right\} + \frac{1}{6}\sum_{j=1}^{k} p_j^2\left(1-\frac{p_j}{2}\right) + \frac{1}{12n}\sum_{j=1}^{k} p_j\left(O_j - E_j\right)^2 \tag{2.3}$$

where $p_j = \dfrac{n_j}{N}$, n is a sample size and N is a population size. The statistic U_d^2 is invariant under cyclic permutations and order-reversing permutations of the cells. The null distribution of U_d^2 is close to that of U^2; if the value of Brown statistic is less than the cut-off point given in Table 6.5 of Mardia and Jupp (1972), we conclude that the fitted distribution is a good fit to the data.

In deciding whether a circular data set follows the von Mises (*VM*) distribution or the wrapped normal (*WN*) distribution, Kent (1976) highlighted the fact that both distributions are hardly distinguishable for $\kappa < 0.1$ or $\kappa > 10$. Kendall (1974) noted that for any analytical, computational and statistical purposes, the *WN* distribution is more convenient for use in some cases and the *VM* distribution in other cases. In this thesis, we refer to the suggestion of Collet and Lewis (1981) who concluded that a minimum sample size required in order to distinguish the two distributions is 200.

For circular regression model case, Lund (1999) assessed the goodness-of-fit of the least circular regression model by using the function $A(\hat{\kappa}) = \frac{1}{n}\sum_{i=1}^{n}\cos\left[y_i - \mu\left(\phi_i, X_i, \hat{\beta}_1, \hat{\beta}_2\right)\right]$ as an analogue of residuals sums of squares in linear regression model. Abuzaid (2009) improved the goodness-of-fit test given by

$$A^*(\hat{\kappa}) = \frac{1}{n}\sum_{i=1}^{n}\left[\cos^2\left(y_i - \hat{y}_i\right)\right] \tag{2.4}$$

where y the dependent angle, $\hat{y}$ the estimate dependent angle and $A^*(\hat{\kappa}) \in [0,1]$.

Therefore, the closer $A^*(\hat{\kappa})$ to 1 indicates a better goodness-of-fit of the model.

2.6 Practical Example

For illustration, we consider the Kuantan wind direction data measured in unit radian from the year 1999 to 2008. Table 2.1 give the yearly mean surface wind direction data obtained from the Malaysian Meteorological Department. Table 2.2 gives the values of circular descriptive statistics for the data. The mean direction μ is 84.65° and the concentration parameter for this data is 3.33. We can conclude that the data sets are concentrated in the east direction.

Table 2.1: Kuantan wind direction data

Year	Mean Surface Wind Direction (radian)
1999	0.28707
2000	1.46071
2001	0.87509
2002	1.64563
2003	1.56786
2004	1.33478
2005	1.80266
2006	2.15736
2007	1.73430
2008	1.67275

Table 2.2: Descriptive statistics

Variable	Angles
Mean Vector (µ)	84.65°
Length of Mean Vector (r)	0.88
Concentration	3.33
Circular Variance	0.12
Circular Standard Deviation	28.45°
Standard Error of Mean	10.57°

Figures 2.14 - 2.15 give the circular histogram and circular plot of the wind data respectively. It can be seen that the data are distributed with the mean direction close to 90° and is located in the middle of the circular histogram plot. However, there is one observation located a bit further from the rest. Further investigation is needed to understand this particular observation, which can be a candidate for outlier. Since the sample size of this data is small, we use result of Collet and Lewis (1981) such that we assume the data follow *WN* distribution.

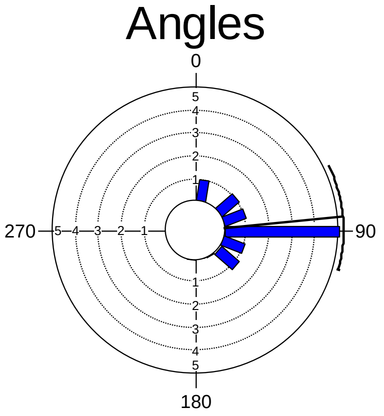

Figure 2.14: Circular Histogram

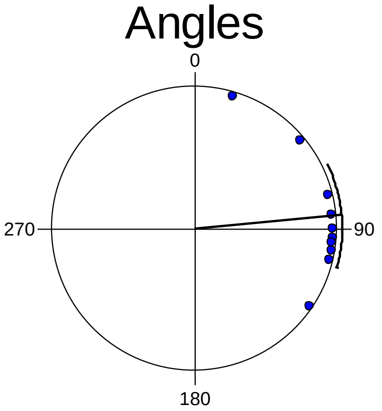

Figure 2.15: Circular plot of Kuantan wind data

2.7 Summary

In this chapter, we have discussed the difference of circular data from the linear data and argued on the need of special methods to analyse such data. We have reviewed circular descriptive statistics, circular graphs, and the goodness of fit test for circular data. However, the von Mises distribution and the wrapped normal distribution are indistinguishable when the sample size of the data set is small. In Section 2.6, we noted that the wrapped Cauchy distribution and the wrapped normal distribution are special cases of the wrapped stable family. In our study, we consider the wrapped normal distribution and compare the results obtained in the next chapter based on the *WN* distribution with the *VM* distribution.