Circular Data: An Overview with Discussion of One-Sample Tests

By

Amanda Scott

Department of Mathematical Sciences
Montana State University
Bozeman, Montana
July 2002

A writing project submitted in partial
fulfillment of the requirements for the degree
Master of Science in Statistics

Section One: Introduction

Circular data is data that occurs around a circle, usually measured in degrees from 0° to 360° or in radians from 0 to 2π. It differs from traditional linear data in that it is closed and has no beginning or end along the real line. For instance, the angles 10° and 350° have an arithmetic mean of 180°, which is in the opposite direction. Intuitively, the mean should be 0°. This is a simple illustration of why new techniques have been developed to handle the unique nature of circular data.

In the following paragraphs, an overview of the different kinds of circular data, examples of their use, ways to display it graphically, common measures of location and dispersion, and common distributions fit to it will be given. Section Two will then deal with one-sample tests for uniformity and goodness-of-fit, while Section Three will conclude with a brief account of several other areas of analysis developed for circular data.

1.1: Types of Circular Data

There are two main types of circular data that occur frequently—vectorial and axial (Fisher, 1993, p. xvii). Vectorial data consists of directed line segments in which there is both an angle and direction associated with the point. The vanishing sights of homing pigeons or the directional preferences of migrating birds are examples of such. However, sometimes the direction is not of special importance, giving rise to axial data. Orientations of dragonflies with respect to the sun or orientations of fractures along a fault line represent axial data. With such data, both "ends of the axes" are recorded on the circle, yielding a bimodal distribution (two clusters of points) instead of a unimodal (one cluster) or multimodal (several clusters) distributions. This difference in data types poses no problem for analysis; simply double all of the axial data angles, reduce them modulo 360°, treat them as vectorial data in the analysis, and back transform them at the end when finished.

Circular data occurs in a wide variety of fields—biology, medicine, geology, meteorology, physics, and oceanography. Most studies of circular data are restricted to settings in which the period of the phenomenon at hand is already well defined. Estimating the period would be the subject of time series analysis, which is beyond the scope of this paper.

Some examples of situations in which circular data occur are given below. This list only scratches the surface of possible applications.

- *Studying the navigational systems of migrating birds*: Birds are placed in a planetarium with the constellations rotated, say 30°, clockwise from their normal positions. The number of directed movements made by the birds in a certain direction are recorded. Nonrandom movement is evidence that the birds are using the stars to navigate (Emlen, 1967).
- *Studying ozone concentrations*: Use wind direction to predict the ozone concentration in cities (Johnson & Wehrly, 1977).
- *Studying circadian or yearly rhythms*: Study body temperature fluctuations, sleep-wakefulness cycles, or hormone release throughout the day. Determining when the maximum and minimums occur can help scientists understand the human body better (Minors & Waterhouse, 1981).
- *Studying hospital emergency room entrance times*: Record times that people arrive at an emergency room, and use times to find busiest and slowest times of day. This information can help determine how many doctors and nurses to keep on staff throughout the day (Cox & Lewis, 1966, p. 254-55).

1.2: Summary Statistics

The first step in any study is to reduce the collected data into a manageable form using summary statistics. A visual representation of the data should always be obtained in order to assess the important characteristics of the sample. The simplest plot, taken from Fisher (1993, p. 16) is a scatter diagram or raw data plot in which points are plotted all around the circle (Figure 1.1).

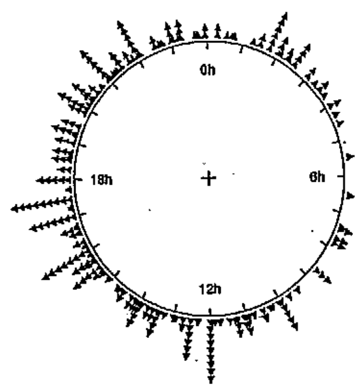

Fig. 1.1 A scatter diagram of arrival times over 24 hour period at an intensive care unit.

Another common plot is a histogram, which can be graphed either on the circle or along a line. A circle must be divided into arcs of equal length, taking the midpoint of each arc to be the center of each bar. Each rectangular bar then represents the relative frequency of points that fall into that arc (Figure 1.2a below from Fisher (1993, p. 19)). Since the periodic nature of the data may be lost in a linear plot, it is suggested that two complete cycles of the data be plotted on the same axis (Figure 1.2b below from Fisher (1993, p. 20)). Several variations of histograms are available, including the rose diagram introduced by Florence Nightingale in the 1850's (Cohen, 1984). Today, there are also several nonparametric density estimates in use as well, but the arbitrary choice of smoothing parameter can greatly affect the plots.

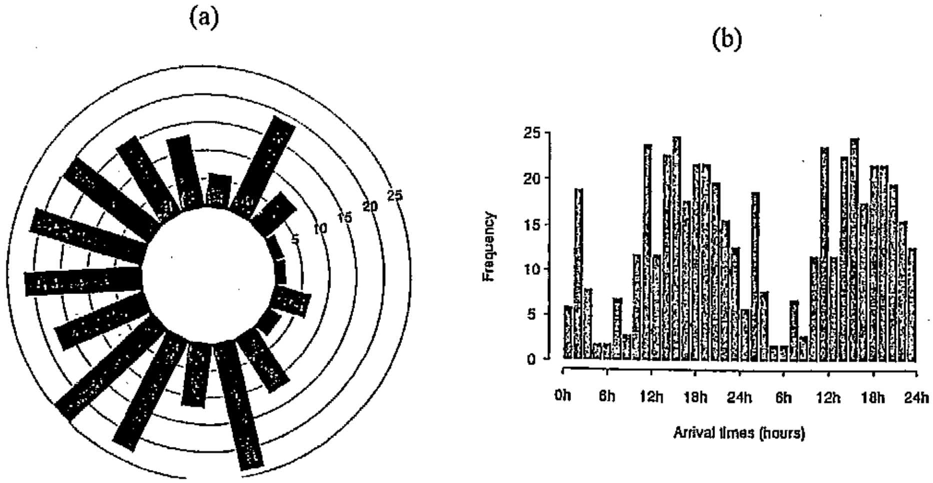

Fig. 1.2 (a) Circular histogram of arrival times at an emergency room over a 24-hour period. (b) Linear histogram of the same arrival times with two cycles shown.

The next step in summarizing the data is to find a measure of location. Obviously, the arithmetic mean of linear statistics is not applicable here. Two common measures of location are the mean vector and median direction. Sometimes one or the other will be easier to calculate, and both are only interpretable with unimodal samples. The mean vector **m** can be expressed in rectangular coordinates, $[\bar{x}, \bar{y}]$, or in polar coordinates, $[r, \bar{\phi}]$ (Batschelet, 1981, p. 231). These measures are obtained as follows (Batschelet, 1981, pp. 9-11):

Let $\phi_1, \phi_2, \ldots, \phi_n$ be a sample of vector angles on the unit circle.

Let $x_i = \cos\phi_i$, and $y_i = \sin\phi_i$.

Define $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}(\cos\phi_1 + \cos\phi_2 + \cdots + \cos\phi_n)$, and

$\bar{y} = \frac{1}{n}(y_1 + y_2 + \cdots + y_n) = \frac{1}{n}(\sin\phi_1 + \sin\phi_2 + \cdots + \sin\phi_n)$.

These are the rectangular coordinates of the mean vector **m**. Often it is useful to have the components of the mean vector in terms of the polar coordinates. The component r represents the mean vector length, ranging in value from zero to one, and $\bar{\phi}$ represents the angle of mean direction, ranging from 0° to 360°. To obtain r from $\bar{x}$ and $\bar{y}$, use the formula $r = (\bar{x}^2 + \bar{y}^2)^{1/2}$.

To calculate $\bar{\phi}$, use the following equations, conditioned on the values of $\bar{x}$ and $\bar{y}$:

$$\bar{\phi} = \begin{cases} \arctan\left(\frac{\bar{y}}{\bar{x}}\right); \bar{x} > 0 \\ \arctan\left(\frac{\bar{y}}{\bar{x}}\right) + 180°; \bar{x} < 0 \\ 90°; \bar{x} = 0, \bar{y} > 0 \\ 270°; \bar{x} = 0, \bar{y} < 0 \\ \mathit{undefined}; \bar{x} = \bar{y} = 0 \end{cases}$$

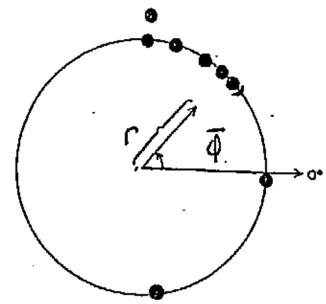

Fig. 1.3 The polar coordinates of the mean vector **m**.

See Figure 1.3, adapted from Mardia (1972, p. 23). Both the mean vector length and mean angle are important in hypothesis testing and confidence intervals. For bimodal axial data (Batschelet, 1981, p. 25), the angles should be doubled and the mean vector $\mathbf{m}_2$ is calculated using

$\bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n}\cos(2\phi_i)$ and $\bar{y}_2 = \frac{1}{n}\sum_{i=1}^{n}\sin(2\phi_i)$. Then, r_2 and $\bar{\phi}_2$ are calculated using the same formulas as were used for r and $\bar{\phi}$. The mean (undirected angle) $\bar{\phi}$ is obtained with either

$\bar{\phi} = \frac{\bar{\phi}_2}{2}$ or $\bar{\phi} = \frac{\bar{\phi}_2}{2} + 180°$.

For convenience, data can be collected in grouping intervals, every 5°, 15°, 30°, etc. This is called grouped data, and the mean vector length must be adjusted (Batschelet, 1981, p. 37-8), depending on the number of groups. For a large number of groups, the correction factor can be dropped. For less than twelve groups though, the corrected mean length is $r_c = cr$ where

$c = \frac{\lambda/2}{\sin(\lambda/2)}$, and λ is the class length in radians.

For a unimodal sample with n odd, the median direction (Fisher, 1993, p. 35) is a diameter with angle $\tilde{\phi}$ (measured on the side where sample points are concentrated) such that half of the data points lie on one side of the circle and the remaining half lie on the other side of the circle. In this case, the median is unique. If the sample size is even, the median direction is not unique, and the diameter passes through two sample points.

A measure of location is insufficient to describe the amount of clustering of the data points on the circle. Several possibilities exist. All depend in some fashion on r which is a natural measure of concentration because the more closely clustered points are, the closer the mean vector length is to one, and the more loosely clustered the points are, the smaller r is. The value 1-r is used instead so that small values will represent closely clustered points and large values will represent loosely clustered points, just as in linear statistics. Batschelet (1981, p. 34) uses $s^2 = 2(1-r)$ and $s = \sqrt{2(1-r)}$ as angular variance and angular standard deviation, respectively. He states that when deviations from the mean angle are small, $2(1-r)$ is approximately $\frac{1}{n}\sum_{i=1}^{n}(\phi_i - \bar{\phi})^2$. Mardia (1972, p. 24), on the other hand, discusses $s_o = \sqrt{-2\ln(1-r)}$ as standard deviation, which is also a good approximation of $\frac{1}{n}\sum_{i=1}^{n}(\phi_i - \bar{\phi})^2$. He uses a result from the wrapped normal distribution (see Section 1.3) that $1-r = e^{-\frac{1}{2}\sigma^2}$ to obtain the tranformation to 1-r. Mardia goes on to state that r and 1-r are more useful for theoretical applications. The differences between s and s_o are slight for small r. Fisher (1993, p.42) writes that when r is near zero, the two measures are approximately the same. The measure s_o ranges from zero to infinity (as r goes to zero), just as linear standard deviation does, but s ranges only from zero to $\sqrt{2}$. Finally, Fisher (1993, pp. 32-34) uses a sample angular dispersion measure $\hat{\delta}$ defined as $\frac{1-\hat{\rho}_2}{2r^2}$ where $\hat{\rho}_2 = \frac{1}{n}\sum_{i=1}^{n}\cos 2(\phi_i - \bar{\phi})$. He uses this measure in calculating confidence intervals about the true mean direction and in comparing several mean directions.

Example 1.1: Using ungrouped, vectorial data of 30 cross-bed azimuths of palaeocurrents from Fisher and Powell (1989), the mean vector m and the three measures of variance will be calculated. The data is unimodal in nature, so these measures will be interpretable. Table 1.1 lists the data, along with their corresponding cosine and sine values.

Table 1.1

ϕ_i (deg)	$\cos(\phi_i)$	$\sin(\phi_i)$	ϕ_i (deg)	$\cos(\phi_i)$	$\sin(\phi_i)$	ϕ_i (deg)	$\cos(\phi_i)$	$\sin(\phi_i)$
294	0.406737	-0.91355	229	-0.65606	-0.75471	290	0.34202	-0.93969
177	-0.99863	0.052336	239	-0.51504	-0.85717	245	-0.42262	-0.90631
257	-0.22495	-0.97437	277	0.121869	-0.99255	245	-0.42262	-0.90631
301	0.515038	-0.85717	250	-0.34202	-0.93969	214	-0.82904	-0.55919
257	-0.22495	-0.97437	287	0.292372	-0.9563	272	0.034899	-0.99939
267	-0.05234	-0.99863	281	0.190809	-0.98163	224	-0.71934	-0.69466
329	0.857167	-0.51504	166	-0.9703	0.241922	215	-0.81915	-0.57358
177	-0.99863	0.052336	229	-0.65606	-0.75471	242	-0.46947	-0.88295
241	-0.48481	-0.87462	254	-0.27564	-0.96126	186	-0.99452	-0.10453
315	0.707107	-0.70711	232	-0.61566	-0.78801	224	-0.71934	-0.69466

Therefore, $\bar{x} = \frac{1}{30}\sum_{i=1}^{30}\cos(\phi_i) = \frac{1}{30}(-8.9432) = -0.2981$, and

$\bar{y} = \frac{1}{30}\sum_{i=1}^{30}\sin(\phi_i) = \frac{1}{30}(-21.7155) = -0.7239$. The mean vector length r is

$r = \left((-0.2981)^2 + (-0.7239)^2\right)^{1/2} = (0.6128)^{1/2} = 0.7828$. Since $\bar{x} < 0$, the mean angle is

$\bar{\phi} = \arctan\left(\frac{-0.7239}{-0.2981}\right) + 180^\circ = 247.62^\circ$. So, the mean vector **m** is can be expressed

either as $[-0.2981, -0.7239]$ in rectangular coordinates or as $[0.7828, 247.62^\circ]$ in polar coordinates. The three measures of angular variance are computed as follows:

$s = \sqrt{2(1-0.7828)} = 0.6591$, $s_o = \sqrt{-2\ln(1-0.7828)} = 1.7475$, and

$\hat{\delta} = \frac{1-0.0417}{2(0.7828)^2} = 0.7819$ where $\hat{\rho}_2 = \frac{1}{30}\sum_{i=1}^{30}\cos 2(\phi_i - 247.62) = \frac{1}{30}(1.2519) = 0.0417$. □

When one variable is linear and the other circular, there are two correlation measures available. Which one to use depends on what type of relationship may exist between the two variables. If a cosine association exists between the two variables (a cosine curve performing a single cycle on a cylinder), then the multiple correlation between the linear variable X and the trigonometric components of the circular variable Θ, $(\sin\phi, \cos\phi)$, must be computed. The sample correlations r_{12}, r_{13}, and r_{23}, between X and sin ϕ, between X and cos ϕ, and between sin ϕ and cos ϕ, respectively, are calculated. To find the sample correlations, use the formula for any set of numbers $(u_1, v_1), \ldots, (u_n, v_n)$, $\dfrac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{\left[\sum_{i=1}^{n}(u_i-\bar{u})^2\sum_{i=1}^{n}(v_i-\bar{v})^2\right]^{\frac{1}{2}}}$ where $\bar{u}=\frac{1}{n}\sum_{j=1}^{n}u_j$ and $\bar{v}=\frac{1}{n}\sum_{j=1}^{n}v_j$. Then, the multiple correlation between X and the circular variable Θ is found using $R_n^2=\dfrac{\left(r_{12}^2+r_{13}^2-2r_{12}r_{13}r_{23}\right)}{\left(1-r_{23}^2\right)}$. This is called C-linear association. If no particular form of association is in mind beforehand, C-association can be calculated, which is much more involved. See Fisher (1993, p. 141-144) for more details. Testing for a significant relationship between the two variables can also be done.

The association between two circular variables can also be described. The more general T-monotone association can be used when no particular type of association is anticipated, and the more specific T-linear association is for strictly positive or negative relationships between the two variables. Computation of either correlation is very involved. Refer to Fisher (1993, pp. 146-154) for a more complete description of both.

3.5: Regression Models

To use circular data in simple and multiple linear regression settings, distinctions must be made between which role the circular variable plays and between when to use linear or nonlinear regression techniques. Three different regression models must be used for the three kinds of regression involving circular variables. These three cases, summarized in Table 3.2, are linear-circular regression, circular-linear regression, and circular-circular regression. In linear-circular regression, the circular variable is used to explain changes in the linear response variable, vice versa in circular-linear regression, and angular variables are both the explanatory and response variables in circular-circular regression. In each case, both linear and nonlinear forms of regression can be used.

Table 3.2

Regression Category	Response Variable	Explanatory Variable
Linear-circular	Linear	Circular
Circular-linear	Circular	Linear
Circular-circular	Circular	Circular

The most commonly used form of regression is the linear-circular case. The C-linear regression model used here can be viewed as a cosine curve performing a single cycle on a cylinder and can be fitted using ordinary least squares. It is written as

$$y_i = \beta_o + A\cos(wt - wt_o) + \varepsilon_i \tag{3.1}$$

where β_o is the mean level, A is the amplitude, w is angular frequency, and t_o is the peak phase.

The independent angular variable is t. The known period is T and is related to w by $w = \frac{2\pi}{T}$ or

$w = \frac{360^\circ}{T}$. The error terms may be independent and normally distributed, but they need not be. For example, repeated measurements on the same individual over time may be made and a curve fit to the data, but inference only applies to the individual that was measured and not to the population from which the individual was drawn. Minors and Waterhouse (1981) give many examples of studies done on circadian rhythms where this is usually the case.

Equation 3.1 does not look like a simple linear regression model until it is rewritten as

$$y_i = \beta_o + \beta_1 \cos(wt) + \beta_2 \sin(wt) + \varepsilon_i \qquad (3.2),$$

using the trigonometric identity $\cos(wt - \phi) = \cos(wt)\cos\phi + \sin(wt)\sin\phi$ (letting $\phi = wt_o$) where $\beta_1 = A\cos\phi$ and $\beta_2 = A\sin\phi$. Equation 3.2 is the simplest case of the more general trigonometric polynomial

$$y_i = \beta_o + A_1 \cos(wt - \phi_1) + A_2 \cos(2wt - \phi_2) + \ldots + A_k \cos(kwt - \phi_k) + \varepsilon_i \qquad (3.3),$$

which can provide a better fit of the data at the expense of estimating more parameters.

The nonlinear regression model of linear-circular association can no longer be thought of as a cosine curve on a cylinder. The only requirements that the curve meet are that it has one peak and one trough and has equal values at zero and 2π. One distortion of the cosine curve is when the peak and trough do not follow each other at an equally spaced interval. When this is the case, the correct model to use is

$$y_i = \beta_o + A\cos(\psi + v\cos\psi) + \varepsilon_i \qquad (3.4)$$

for $\psi = wt - \phi$. The additional parameter to estimate, v, is a skewness parameter ranging from -30° to 30°. A value of $v = 0^\circ$ reduces the model to the linear regression form in 3.1. A second possible distortion of the cosine pattern is when the peak is flatter or steeper than a cosine curve's peak is making the model

$$y_i = \beta_o + A\cos(\psi + v_1 \sin\psi) + \varepsilon_i \qquad (3.5)$$

appropriate. The parameter v_1 is a measure of peakedness; values close to –60° have flat peaks, and values close to 60° have steep ones. Solving for the parameters in these two nonlinear models requires using generalized least squares. See Figure 3.1 from Fisher (1993, p. 140) for both linear and nonlinear linear-circular regression model comparison.

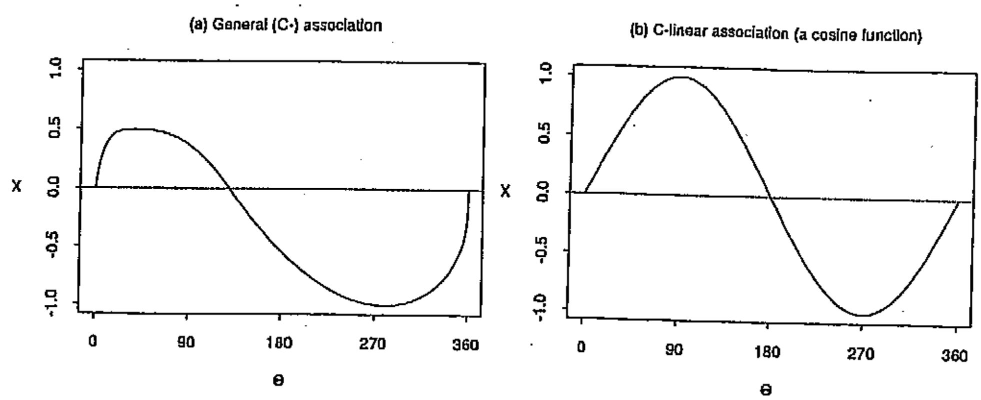

Fig. 3.1 (a) nonlinear linear-circular regression model—a periodic function from 0 to 2π. (b) linear linear-circular regression model—a cosine function from 0 to 2π.

Switching the roles of the angular and linear variables changes the nature of the model dramatically. In circular-linear regression, either the mean angle, dispersion, or both may depend on the explanatory variables. To model mean direction, use

$$\mu_i = \mu + g(\beta_1 x_1 + \beta_2 x_x + \ldots + \beta_k x_x) + \varepsilon_i \qquad (3.6)$$

where μ_i is the mean direction, and g is a link function mapping the real line to the circle. Maximum likelihood estimates of μ and $\boldsymbol{\beta}$ can be obtained using a system of iterative equations (Fisher, 1993, pp. 158-9). Since dispersion is not an easily definable characteristic of samples, modeling dispersion or both dispersion and mean angle requires the assumption of having drawn the sample from a von Mises population.

A model defining circular-circular regression is a problem that is not yet resolved.

3.6: Conclusion

This overview of multi-sample testing, computer procedures, confidence intervals, correlation analysis, and regression of circular data is intended to emphasize some interesting applications of circular data and some of the methods developed to support those applications. As more experience is gained in a wide range of experiments, the theory and methods for circular data will become more standardized and succinct, and as they gain publicity, fields that do not use them today may find applications for them in the future.

Literature Cited

Ajne, B. (1968). A simple test for uniformity of a circular distribution. *Biometrika*. **55**, 343-354.

Barndorff-Nielson, O. & Cox, D. R. (1980). Edgeworth and saddle-point approximations with statistical applications (with Discussion). *J. Royal Statist.* Soc. B **41**, 279-312.

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

Bingham, M. S. & Mardia, K. V. (1975). Maximum likelihood characterization of the von Mises distribution. *In* "Statistical Distributions in Scientific Work" (C. P. Patil *et al.*, eds), **3**, 387-398. Reidel, Dordrecht.

Bliss, C. I. (1970). *Statistics in Biology*, Vol. 2. New York: McGraw-Hill.

Cohen, I. B. (1984). Florence Nightingale. *Sci. Am.* **250**, 98-107.

Cox, D. R. & Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events*. London: Methuen & Co. Ltd.

Emlen, S. T. (1967). Migratory orientation in the indigo bunting (*Passerina cyanea*). Part I. Evidence for the use of celestial cues. *Auk* **84**, 309-342.

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

Fisher, N. I. & Powell, C. McA. (1989). Statistical analysis of two-dimensional palaeocurrent data: Methods and examples. *Aust. J. Earth Sci.* **36**, 91-107.

Greenwood, J. A. & Durand, D. (1955). The distribution of length and components of the sum of n random unit vectors. *Ann. Math. Stat.* **26**, 233-246.

Hodges, J. L., Jr. (1955). A bivariate sign test. *Ann. Math. Stat.* **26**, 523-527.

Johnson, R. A. & Wehrly, T. E. (1977). Measures and models for angular correlation and angular-linear regression. *J. R. Statist. Soc.* **B 39**, 222-9.

Kuiper, N. H. (1960). Tests concerning random points on a circle. *Konikl. Nederl. Akad. Wet. Proc.* Ser. A, **63**, 38-47.

Laubscher, N. F. & Rudolph, G. J. (1968). A distribution arising from random points on the circumference of a circle. *Nat. Res. Inst. Math. Sci. Report* **268**, Pretoria, South Africa, 1-15.

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

Minors, D. S. & Waterhouse, J. M. (1981). *Circadian Rhythms and the Human.* Bristol: John Wright & Sons, Ltd.

Rayleigh, Lord. (1880). On the resultant of a large number of vibrations of the same pitch and of arbitrary phase. *Phil. Mag.* **10**, 73-8.

Rao, J. S. (1969). Some contributions to the analysis of circular data. Thesis, Indian Statist Inst., Calcutta.

Rao, J. S. (1976). Some tests based on arc-lengths for the circle. Sankhya, Ser. B **33**, 1-10.

Watson, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika* **48**, 109-114.

TABLE 2.4

TEST	NULL HYPOTHESIS	ALTERNATIVE HYPOTHESIS	TEST STATISTIC	ADVANTAGES	DISADVANTAGES	NONREMOVABLE MODALITY
Rayleigh*ψ	uniform distribution	directedness or von Mises	r	r easy to calc. UMP for von Mises		unimodal
Hodges-Ajne's*ψ	uniform distribution	directedness	k=min # sample points on one side of diameter	easy to use	not as powerful as Rayleigh	unimodal
Rao Spacing*	uniform distribution	directedness	$U = \frac{1}{2}\sum_{i=1}^{n} \lvert T_i - \frac{360°}{n} \rvert$	easy to use powerful with multimodal data		multimodal
Range*ψ	uniform distribution	directedness	w=smallest arc containing all points	easy to use	only use for dist. with small angular deviation	unimodal
V-Test*ψ	uniform distribution	unimodal about θ_o or von Mises	$u = v\sqrt{2n}$ $v = r\cos(\bar{\phi} - \theta_o)$	more powerful than Rayleigh if direction centered around θ_o	not as powerful as Rayleigh if direction not centered around θ_o	unimodal
V Competitor*ψ	uniform distribution	unimodal about θ_o	k=# of points on side of circle opposite of θ_o	easy to use	not as powerful as V-Test	unimodal
Mean Direction*ψ	$\theta = \theta_o$	$\theta \neq \theta_o$	$\bar{\phi} \pm \sin^{-1}(z_{\alpha/2}\hat{\sigma})$ $\hat{\sigma}^2 = \frac{\hat{\delta}}{n}$	can test true mean direction explicitly	must use bootstrapping method for n less than 25	unimodal
Chi-squaredγ	data fits model	does not fit model	$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$	good approximation	only use with large sample size > 5k	multimodal
Kuiper's*	data fits model	does not fit model	$K = V_n\sqrt{n}$ $V_n = D^+ + D^-$	more powerful than chi-squared with bi & unimodal data		multimodal
Watson's U_n^2 *	data fits model	does not fit model	$U_n^2 = \sum_{i=1}^{n} v_i^2 - \sum_{i=1}^{n}\left(\frac{c_i v_i}{n}\right) + n\left[\frac{1}{3} - (\bar{v} - \frac{1}{2})^2\right]$	powerful for small samples		multimodal

* If data is grouped, r has been adjusted.
ψ If data is axial, angles have been doubled.
γ Expected frequency in each group is greater than or equal to 4.

Section Three: Additional Topics

Significant contributions to the field of circular data continue to be made regularly. It is an active field in which much work remains to be done until circular data acquires the breadth and depth of theory that linear data has. Some additional areas of application of circular statistics will be highlighted here—multi-sample tests, automated procedures, parameter estimation, correlation, and regression.

3.1: Multi-sample testing

In addition to testing a single sample of data, it is often useful to have procedures available for testing two or more samples of data. These procedures are designed to detect differences among samples, such as among a control treatment and different levels of a single factor. These tests rest on the following assumptions: each sample was a simple random sample from a population, samples are independent of each other, and grouping class sizes, if used, must be less than 5°. For the most part, it is also assumed that the data were drawn from a unimodal distribution and are most assuredly not uniformly distributed. Once satisfied, tests designed to be used with any of the following alternative hypotheses can be used: differences in mean angle exist; differences in angular deviation exist; differences in mean angle, angular deviation, or both exist; or differences in mean angle θ_1 or concentration parameter κ exist given a von Mises distribution. There are also omnibus tests that detect differences of any type. Finally, tests to detect differences in median direction or in overall distributional patterns exist. Both two-sample and multi-sample versions of tests are available.

One must keep in mind that just because significance is determined in any one test, the conclusion must still be drawn cautiously. A visual inspection of plots will aid the researcher in determining what it is about the samples' characteristics that cause significance to occur. For small numbers of samples, two useful plots for comparing them are scatter diagrams with different plotting symbols used for each sample and angular qq plots. Angular qq plots compare the general shape of two samples. If they are the same, the points will lie roughly along a 45° line through the origin (for construction details, see Fisher, 1993, pp. 111-2). For multiple samples that are all unimodal, side-by-side boxplots (constructed in the same way that boxplots for linear data are) can be used for comparison purposes.

These methods are best used in experimental designs with one factor at multiple levels. When additional factors are added with levels of each factor, the analysis becomes increasingly complex. Much work is being done in this area to determine what assumptions must hold, if any, in order to develop a general approach to analysis of variance of circular data.

3.2: Automated Procedures

The success of the tests and confidence intervals described hereafter is based to a great extent on the evolution of computer intensive procedures such as bootstrapping and randomization. When the sample size is small, between seven and twenty-five, or the sampling distribution of the statistic is not known, bootstrapping can provide reasonably accurate estimates of standard errors, and randomization (or permutation) can aid in completing hypothesis tests. Fisher (1993, pp. 200-18) gives some algorithms for programming bootstrapping and randomization methods. These procedures become especially useful in correlation analysis where computing the test statistics can be very involved.

3.3: Parameter Estimation and Confidence Intervals

The statistics for mean vector, von Mises parameters, and median direction have already been introduced in Sections 1.2 and 1.3. Here, the corresponding parameters for each of the statistics will be given in Table 3.1, and then methods for forming confidence intervals about each of them will be discussed.

Table 3.1

	Sample Statistic	Population Parameter
Mean Vector	m	μ_1
Mean Vector Length	r	ρ_1
Mean Angle/Direction	$\bar{\phi}$	θ_1
von Mises Concentration	$\hat{\kappa}$	κ
Median Direction	$\tilde{\phi}$	$\tilde{\mu}$

The confidence interval procedure for the mean direction θ_1 has already been described in Section 2.2 (iii). To summarize it once more, use a bootstrap procedure for samples less than 25 in size (Fisher, 1993, pp. 205-6). For samples with $n \geq 25$, calculate the formula

$\bar{\phi} \pm \sin^{-1}(z_{\alpha/2}\hat{\sigma})$ where $\hat{\sigma}^2 = \frac{\hat{\delta}}{n}$.

A confidence interval for the median direction $\tilde{\mu}$ can only be constructed if the data are relatively concentrated on one arc of the circle. To find the upper and lower endpoints of the confidence interval, $(\phi_{(L_m)}, \phi_{(U_m)})$, count off m observations to the left and right of the sample median $\tilde{\phi}$ (not including $\tilde{\phi}$ itself). The smaller of the observations that the count ends on is the lower endpoint, and the larger observation is the upper endpoint. The integer m depends on the sample size and the desired α-level. If $n < 16$, use Appendix A6 of Fisher (1993, p. 226) to find m to yield exact α–levels. If $n \geq 16$, use m = 1 + integer part of $\left(\frac{1}{2}\sqrt{n}z_{\alpha/2}\right)$ where $z_{\alpha/2}$ is the upper 100($\alpha/2$)% of a N(0, 1) distribution to obtain an approximate α–level (Fisher, 1993, pp. 72-3).

When the sample data fits a von Mises probability distribution, estimates of θ_1, κ, and ρ_1 can be used to form confidence intervals for both θ_1 and κ. The maximum likelihood estimate of θ_1 is $\bar{\phi}$, and the MLE of κ is the solution of $\frac{I_1(\hat{\kappa})}{I_o(\hat{\kappa})} = r$ for $\hat{\kappa}$. An approximate solution is $\hat{\kappa}_{ML} = \begin{Bmatrix} 2r + r^3 + {}^{5r^5}\!/_6 ; r < 0.53 \\ -0.4 + 1.39r + {}^{0.43}\!/_{(1-r)} ; 0.53 \le r < 0.85 \\ \frac{1}{r^3 - 4r^2 + 3r} ; r \ge 0.85 \end{Bmatrix}$. When the sample size is small, and r is less than 0.70, then the MLE of κ can be heavily biased. For $n \le 15$, use the estimate

$$\hat{\kappa} = \begin{Bmatrix} \max(\hat{\kappa}_{ML} - 2(n\hat{\kappa}_{ML})^{-1}, 0); \hat{\kappa}_{ML} < 2 \\ \frac{(n-1)^3 \hat{\kappa}_{ML}}{(n^3 + n)}; \hat{\kappa}_{ML} \ge 2 \end{Bmatrix}.$$

The MLE's can then be used to create confidence intervals for the mean angle and concentration parameter. The estimate of the standard error of $\bar{\phi}$ is $\hat{\sigma}_{VM} = \frac{1}{\sqrt{nr\hat{\kappa}}}$ where $z_{\alpha/2}$ is the upper 100($\alpha/2$) critical value from a N(0, 1) distribution. Then, the confidence interval for the mean angle θ_1 is $\bar{\phi} \pm \sin^{-1}(z_{\alpha/2}\hat{\sigma}_{VM})$ (Fisher, 1993, p. 89). When $\hat{\kappa} \ge 2$, a two-sided 100(1-α)% confidence interval for κ can be calculated using the values $a = \frac{(n - nr)}{\chi^2_{n-1}(1 - \alpha/2)}$ and $b = \frac{(n - nr)}{\chi^2_{n-1}(\alpha/2)}$ such that the upper and lower bounds are $\left(\frac{1 + \sqrt{1 + 3a}}{4a}, \frac{1 + \sqrt{1 + 3b}}{4b} \right)$ (Fisher, 1993, p. 91). Fisher (1993, pp. 88-91) also describes bootstrapping methods for each type of confidence interval.

3.4: Correlation Analysis

Correlation in linear statistics measures the strength of the linear association between two variables. New numerical measures of correlation have been developed for circular data. There are two settings of interest to consider: only one variable is circular while the other is linear or both variables are circular.

1.3: Circular Distributions

A brief summary of major families of distributions used for circular data is required here. A host of probability density functions (pdf) for circular data have been suggested. All are non-negative functions, either continuous or discrete, and have finite support from 0 to 2π. The most simple is the uniform distribution, in which any angle between 0° and 360° is equally likely. The pdf is $f(\phi)=\frac{1}{2\pi}$, $0 \le \phi \le 2\pi$. It is most commonly used as the null hypothesis model against which various alternative models are tested. The von Mises distribution plays the same role in circular statistics that the normal distribution does in linear statistics. It is a two parameter unimodal, symmetric distribution with pdf $f(\phi)=(2\pi I_o(\kappa))^{-1}\exp(\kappa\cos(\phi-\theta_1))$, $0 \le \phi \le 2\pi$, and $0 \le \kappa \le \infty$. In this pdf, $I_o(\kappa)$ is a modified Bessel function of order zero (see Fisher 1993, p. 48 or Batschelet 1981, p. 297 for a complete description), θ_1 is the population mean angle, and κ is a parameter of concentration. This is the only distribution for which the sample mean angle is the maximum likelihood estimate (MLE) of the population mean angle, and the MLE of ρ_1 (population mean vector length) is r (Bingham and Mardia, 1975). (Note, r can be converted into an estimate of κ by means of the equation $\frac{I_1(\hat{\kappa})}{I_o(\hat{\kappa})}=r$.)

Aside from these two most important distributions, various other linear distributions, such as the normal, cauchy, and poisson, have been modified and "wrapped" around the circle to obtain circular distributions. Distributions with multiple modes can be thought of as mixtures of several unimodal distributions. There are families of skewed distributions and of distributions with parameters that can alter the shape of their unimodal peaks (making them either flatter or steeper). A detailed presentation of some of these other distributions can be found in Mardia (1972, pp. 49-53), Fisher (1993, pp. 43-55), and Batschelet (1981, pp. 278-290).

This brief introduction to circular data, statistics, and distributions should suffice to make the presentation of subsequent material comprehensible.

Section Two: One-Sample Tests

2.1: Introduction

One-sample tests fall into one of two broad categories: tests for randomness and tests for goodness-of-fit. In testing for randomness, the question posed is whether or not there is a pattern in the sample data. For example in experiments, many times a researcher wants to know if the treatment induced causes the experimental units to prefer a particular direction over another. If not, then the observations will tend to be randomly scattered (or uniformly distributed) around the circle. If so, the points should be clustered together in some way. Once the null hypothesis of random distribution of the points has been rejected, the researcher may attempt to fit a probability model to the data. Goodness-of-fit tests determine whether the particular model chosen is appropriate or not. Examples of the commonly used tests of each type will be given.

A variety of creative tests have been developed over the years. Some are very easy to apply while others are computer intensive; some are well-used by scientists while others are not very popular; some are more powerful in rejecting certain types of alternatives than others; and some are better used with unimodal samples than with multimodal ones. All of the tests assume that the sample of vectors is independent and is a random sample from a population of interest. The tests also depend on one or both of the following assumptions: if the data are grouped, r has been adjusted (*), and if the data are axial, the angles have been doubled ($^{\Psi}$). The corresponding symbol will be placed next to the test name to indicate which of the assumptions is required for the test to be performed correctly. Also, note that some tests can be used with several different alternative hypotheses. Nonstandard tables of critical values or p-values for tests can be found in

Batschelet (1981) as well as straightforward examples of most of the tests. Mardia (1972) gives a theoretical treatment of most of the tests described here.

Before choosing one of the tests for randomness or goodness-of-fit from those described below, a probability plot of the data against the null hypothesis model can be useful. For the uniform model, order the angles from smallest to largest, and let

$x_1 = \frac{\phi_{(1)}}{360^\circ}, x_2 = \frac{\phi_{(2)}}{360^\circ}, \ldots, x_n = \frac{\phi_{(n)}}{360^\circ}$. Plot the pairs $\left(\frac{1}{n+1}, x_1\right)\left(\frac{2}{n+1}, x_2\right), \ldots, \left(\frac{n}{n+1}, x_n\right)$. For a von Mises null distribution, calculate the quantiles $q_1, q_2, \ldots, q_n$ for a von Mises distribution, and then calculate $z_i = \sin\frac{1}{2}(\phi_i - \bar{\phi})$ for every observation. Order the z_i from smallest to largest, and plot the points $\left(\sin(\frac{1}{2}q_1), z_{(1)}\right), \left(\sin(\frac{1}{2}q_2), z_{(2)}\right), \ldots, \left(\sin(\frac{1}{2}q_n), z_{(n)}\right)$. In either case, there is evidence that the correct model has been chosen if the points lie along a 45° line passing through the origin. Since there is an aribitrary starting point for circular data, Fisher (1993, p. 65) desribes a method for extending the ends of the plot by a percentage of the data. A uniform probability plot for the paleocurrent data of Example 1.1 is shown in Figure 2.1. The points do not fall along the 45° line through the origin. However, formal testing should still be done since a visual assessment of the correct model is sometimes not trustworthy.

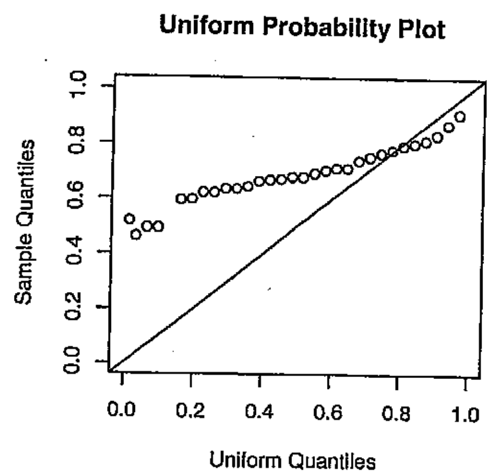

Fig. 2.1 Uniform probability plot for paleocurrent data.

2.2: Tests for Randomness

a) When there is an unspecified alternative direction, any of the following four tests can be used.

The null hypothesis (H_o) in each case is that the parent population is uniformly distributed.

i) Rayleigh Test *Ψ

Lord Rayleigh (1880) first studied the importance of using the mean vector length r in assessing uniformity of a sample of vectors. The concept is quite simple. If the points are randomly scattered about the circle, then r should be small. If the points are clustered in one area, then r will be close to one. When r is statistically significantly different from zero, the null hypothesis is rejected to conclude that there is some direction in the data. See Table H (Batschelet, 1981, p. 334). The Rayleigh test is very popular and can be used with the alternative of a von Mises distribution. It is the uniformly most powerful (UMP) test when testing against a von-Mises alternative (Mardia, 1972, p. 137).

Example 2.1: Using the data listed in Table 1.1 of 30 cross-bed azimuths of palaeocurrents, the Rayleigh test statistic is $r = 0.7828$. Table H indicates that for a sample size of 30 and an r of 0.7828, the p-value is less than 0.001. Therefore, there is statistically significant evidence that the palaeocurrents are not uniformly distributed. See Figure 2.2 taken from Fisher (1993, p. 61) below.

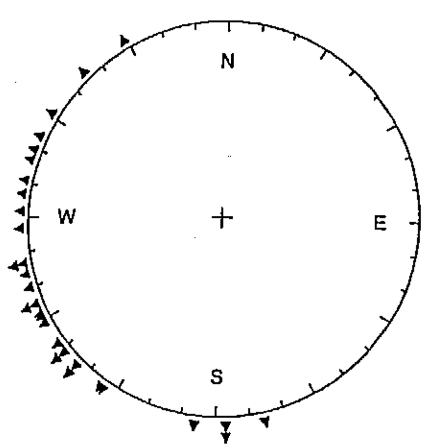

Fig. 2.2 Scatter diagram of 30 cross-bed azimuths of palaeocurrents.

ii) Hodges' and Ajne's Test *ψ

Ajne's test (Ajne, 1968) is a special case of the bivariate sign test by Hodges (1955). Its attraction is the ease of calculation of the test statistic. Rotate a diameter around the circle until a maximum and minimum number of points lie on either side of the line. Let k equal the minimum number of points on one side. If k is small relative to the sample size, then there is evidence that the population has some direction to it. Use Table J (Batschelet, 1981, p. 337) to assess significance. If the parent population is a von Mises distribution, then this test is not as powerful in detecting deviations from the null hypothesis as the Rayleigh test is.

Example 2.2: For the paleocurrent data, $k = 0$. For a sample size of thirty, the p-value is less than 0.001, evidence enough to reject the hypothesis of uniformity.

iii) Rao's Spacing Test *

Rao's Spacing test (Rao,1969, 1976) is unique in that it can detect deviations from uniformity in both unimodal and multimodal samples. If n points are randomly spaced, then the distance between neighboring points should be roughly $\frac{360^\circ}{n}$. If not, then very large or small distances between points might be expected. The test statistic $U = \frac{1}{2}\sum_{i=1}^{n} | T_i - \frac{360^\circ}{n} |$ is based on T_i, which is the distance between adjacent points. Large values of U provide evidence against the null hypothesis. Use Table L of Batschelet (1981, p. 339).

Example 2.3: For the paleocurrent data, first arrange the angles from smallest to largest, and find the arc length between consecutive angles to get each T_i. Subtract twelve from each one, take the absolute value, sum the deviations, and divide by two. Here, $U = 205$. From Table L, the p-value is less than 0.01, statistically significant evidence that the data is not uniformly distributed.

iv) The Range Test *Ψ

The range test (Laubscher & Rudolph, 1968) uses the smallest arc, w, capturing all of the data points as its test statistic. The premise is that the smaller the arc, the more concentrated the data, ruling out uniformity. Table M of Batschelet (1981, p. 340) provides critical values. This test only works well with distributions having small angular standard deviations.

Example 2.4: The smallest arc capturing all of the paleocurrent data points is $w = 329° - 166° = 163°$. The p-value from Table M is less than 0.005, again indicating a non-uniform distribution of the data.

Table 2.1 below summarizes the results of the tests done on the paleocurrent data in examples 2.1 through 2.4.

Table 2.1

TEST	STATISTIC	P-VALUE	DECISION
Rayleigh	$r = 0.7828$	< 0.001	Reject H_o
Hodges/Ajnes'	$k = 0$	< 0.001	Reject H_o
Rao's Spacing	$U = 205$	< 0.01	Reject H_o
Range Test	$w = 163°$	< 0.005	Reject H_o

b) In some studies, there is a specified alternative direction in mind beforehand. A homing location, topographic landmark, or magnetic orientation may have some hypothesized effect on the units. This particular direction is denoted by θ_o. The hypotheses for the next two tests are H_o: the parent population is uniformly distributed versus H_a: the parent population is clustered about θ_o.

i) The V Test *Ψ

This test, developed by Greenwood and Durand (1955), rests on the concept of the homeward component v defined as $v = r\cos(\bar{\phi} - \theta_o)$. This component is the projection of **m** onto the hypothesized direction θ_o and can range in value from −1 to 1. The test statistic used is $u = v\sqrt{2n}$. Rejecting the null hypothesis using Table I (Batschelet, 1981, p. 336) means that significant evidence exists that values are clustered around θ_o. When it is appropriate to specify an alternative direction, this test is more powerful than Rayleigh in detecting deviations from the null. This test and the next should only be used to test for uniformity. Testing if the mean direction differs from the homeward direction is the role of the test in part (iii).

Example 2.5: The directions that one hundred ants chose to travel when they were exposed to a black target are graphed in Figure 2.3 below. This data was taken from Fisher (1993, p. 243) and is reproduced in Table 2.2. If it is hypothesized that the ants will run toward the target, then $\theta_o = 180°$ is the homeward direction.

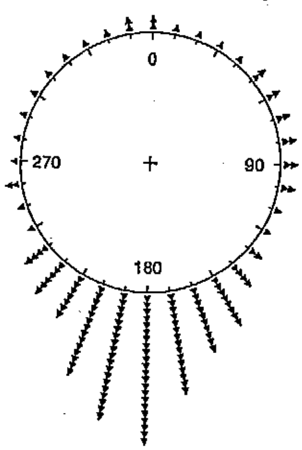

Fig. 2.3 Scatter diagram of 100 ant directions in presence of black target placed as shown.

Table 2.2

330	290	60	200	200	180	280	220	190	180
180	160	280	180	170	190	180	140	150	150
160	200	190	250	180	30	200	180	200	350
200	180	120	200	210	130	30	210	200	230
180	160	210	190	180	230	50	150	210	180
190	210	220	200	60	260	110	180	220	170
10	220	180	210	170	90	160	180	170	200
160	180	120	150	300	190	220	160	70	190
110	270	180	200	180	140	360	150	160	170
140	40	300	80	210	200	170	200	210	190

The summary values needed to compute the test statistic are $r = 0.6101$ and $\bar{\phi} = 183.14°$. Therefore, $v = 0.6101\cos(183.14 - 180) = 0.6092$, and $u = 0.6092\sqrt{200} = 8.6152$. From

Table I, the p-value is less than 0.0001, so there is significant evidence that the directions are clustered around 180°.

ii) Competitor of V Test *Ψ

This is a modification of the Hodges-Ajne test proposed by Batschelet (1981, pp. 64-6). The idea is to draw a diameter line perpendicular to the proposed direction, and let k equal the number of observations that fall on the side of the circle opposite that of θ_o. The test statistic is k, and the smaller it is, the more likely the sample was not drawn from a uniform distribution and that the points are clustered around θ_o. Table K (Batschelet, 1981, p. 338) lists p-values for a given n and k. This test is very easy to use.

Example 2.6: The number of observations that fall on the arc of the circle from 90° to 270° opposite the black target is k = 23. The p-value is less than 0.001, evidence that the points are not uniform and that they are clustered around 180°.

iii) Testing Mean Direction *Ψ

This test should be used (Fisher, 1993, p. 76) when the true mean direction is desired for a unimodal sample. The hypotheses are H$_o$: $\theta = \theta_o$ versus H$_a$: $\theta \neq \theta_o$. It is easiest to calculate a 100(1-α)% confidence interval and see if θ_o is included in the interval. For small samples of less than 25 observations, a bootstrapping method is required (Fisher, 1993, p. 75), and for samples with more than 25 observations, the confidence interval formula is $\bar{\phi} \pm \sin^{-1}(z_{\alpha/2}\hat{\sigma})$ where $\hat{\sigma}^2 = \frac{\hat{\delta}}{n}$ and $z_{\alpha/2}$ is the upper 100($\alpha/2$) critical value from a N(0,1) distribution. (Recall that $\hat{\delta} = \frac{1-\hat{\rho}_2}{2r^2}$ is a measure of sample dispersion.)

Example 2.7: Using the ant data of example 2.2, the values needed for a 95% confidence interval are $\bar{\phi} = 183.14°$, $r = 0.6101$, $\hat{\rho}_2 = 0.4452$, $\hat{\sigma} = \sqrt{\frac{0.7452}{100}} = 0.0863$, and $z_{.025} = 1.960$. The confidence interval is then $183.14° \pm \sin^{-1}(1.960 \cdot .0863)$, and the upper and lower bounds are (173.40°, 192.88°). Since the homeward direction 180° lies within these bounds, do not reject the null hypothesis that $\theta_o = 180°$.

Table 2.3 lists the results from the three tests done on the ant data.

Table 2.3

TEST	STATISTIC	P-VALUE	DECISION
V-Test	$u = 8.6152$	< 0.0001	Reject H_o
V-Competitor	$k = 23$	< 0.001	Reject H_o
Mean Direction	95% CI (173°, 193°)	n/a	Fail to Reject H_o

2.3: Goodness-of-Fit Tests

The hypotheses for the first three tests given are H_o: the parent population fits a particular model versus H_a: the parent population does not fit the proposed model where the model can be any probability model of interest. Note that goodness-of-fit tests can double as a test for randomness if the model of interest chosen is the uniform distribution, but they are usually used with more interesting distributions.

i) Chi-squared Test

This is the same chi-squared test that is used in linear statistics. The circle should be divided into arcs (not necessarily of equal length), and the frequency of sample points in each arc is counted (n_i). Then, the expected number of observations falling into each arc based on the chosen distribution is calculated, e_i. Let k equal the number of arcs. The test statistic is

$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - e_i)^2}{e_i}$ and has an approximate $\chi^2(k-1)$ distribution. Reject H_o for large values of χ^2. There is one special assumption here that cannot be violated: the expected frequency in each arc must be greater than or equal to four. Complying with this restriction implies that the sample size must be large, at least 5k, to be reasonably accurate, but when it is, it will work for both unimodal and multimodal samples.

Example 2.8: Applying the chi-squared test to strongly unimodal data, like the paleocurrent data, can be difficult, if not impossible. Even with only six arcs, three of the arcs have counts of zero in them, violating the assumption for this test. The sample size would have to be much bigger than thirty to apply this test to the paleocurrent data.

ii) Kuiper's Test *

In Kuiper's test (Kuiper, 1960), the empirical step cumulative distribution function (cdf) and theoretical cdf of a chosen model are graphed on top of one another. The test statistic uses the largest deviation of the empirical cdf above the theoretical cdf (D^+) and the largest deviation below the theoretical cdf (D^-). Define $V_n = D^+ + D^-$. The test statistic is $K = V_n\sqrt{n}$. Large deviations indicate a departure from the chosen model, so K large is evidence against the null hypothesis (see Table N in Batschelet, 1981, p. 341). This test is actually more powerful than the chi-squared test when unimodal or bimodal data is involved.

iii) Watson's U_n^2 Test *

Watson (1961) modified the method of using mean square deviation for circular data. Let F(ϕ) be the cdf of the theoretical distribution. Rearrange the data into ascending order $\phi_{(1)} \le \phi_{(2)} \le \ldots \le \phi_{(n)}$, and the test statistic is $U_n^2 = \sum_{i=1}^{n} v_i^2 - \sum_{i=1}^{n} \left(\frac{c_i v_i}{n}\right) + n\left[\frac{1}{3} - (\bar{v} - \frac{1}{2})^2\right]$ where

$v_i = F(\phi_{(i)})$ and $c_i = 2i - 1$. In theory, each of the observations is being subtracted from its expected value; the deviations are squared and averaged. A large mean square deviation from Table O (Batschelet, 1981, p. 342) would imply that the model does not fit the data well. This test can be used for mulitmodal samples and is particularly powerful for small sample sizes.

Example 2.9: From Examples 2.1-2.4, it is known that the palaeocurrent data is distributed significantly differently from the uniform distribution. Watson's test can be used to check if the data fits a von Mises distribution. Using Table E to get estimates for the von Mises cdf (Batschelet, 1981, p. 322-31) with a $\hat{\kappa}_{ML} \approx 2.7$ (See Section 3.3.), Watson's $U_n^2 = 1.09$. Table O then gives a p-value of less than 0.005. Therefore, there is significant evidence that the von Mises distribution fits the paleocurrent data very well.

2.4: Conclusion

The sheer number of tests can be overwhelming, but each test has its strengths, weaknesses, and purpose. By no means has every possible one-sample test been examined here. Modifications to these described here and tests simply for symmetry or for median direction are also available. Table 2.4 lists the salient facts pertaining to each test described here for quick reference use and side-by-side comparison.