The Philippine Statistician, 1995-1996
Vol. 44-45, Nos. 1-8, pp. 61-67

Inferential Statistics for Concentration of Directional Data Using the Chi-Square Distribution

Surin Khanabsakdi[1]

ABSTRACT

For any population distribution we often consider 2 important characteristics: central tendency and dispersion. In directional or circular data, the central tendency is measured by the mean direction and the dispersion by the concentration value. This paper is concerned with estimation of and hypothesis testing on the measure of concentration.

KEY WORDS: Angular data, Bessel function, Circular data, Concentration, Directional data, Von Mises distribution.

1. INTRODUCTION

Directional data is also often called "cyclic data" or "circular data" or "angular data". Examples of such data are those on time within each day, on months in each year, and on the direction of homeward flying of a bird. In circular statistics, the observed value is the angle on a unit circle. On the linear scale, we frequently consider its characteristics, especially central tendency and dispersion, and/or test hypothesis concerning these parameters. The same is true for directional data. There are several test statistics available for testing hypothesis of central tendency in samples from circular population but only a few on concentration. This paper is concerned with the estimation of and hypothesis testing on the parameter of dispersion or concentration.

In any circular population, we are often concerned with the mean direction or length of the population mean vector which indicates the central tendency and the concentration parameter which indicates the dispersion of population data. There are relationships among these parameters. For example, the mean direction is an angle, say θ, of the population mean vector. The length, say ρ, of the population mean vector may be large or small. If the directions of the unit vectors are very different, the length is small, which implies greater data dispersion. On the other hand, if the directions of the unit vectors are not quite different, the length is large, which implies little dispersion. Since the data are angles of the unit vector, the mean direction, θ, varies from 0 to 2π and the length, ρ, varies from 0 to 1.

Batschelet (1981) transformed ρ to population circular standard deviation, $\sigma = [2(1-\rho)]^{1/2}180/\pi$ which varies from 0 to $180\sqrt{2}/\pi$. In other words, "concentration" can also be used to explain dispersion of the data. In a circular population, K denotes the concentration parameter which is related to ρ as follows: $I_1(K)/I_0(K) = \rho$, where $I_m(K)$ is a modified Bessel function of the first kind of order m.

[1] Department of Statistics, Faculty of Science, Chiang Mai University, Chiang Mai, 50200, Thailand

2. PROBABILITY DENSITY FUNCTION

Let ϕ be a circular random variable with a von Mises distribution. Von Mises (Batschelet 1981, p.34) developed the modified Bessel function of the first kind for the distribution of ϕ.

The modified Bessel function of the first kind of order m is

$$I_m(K) = \frac{1}{2\pi}\int_0^{2\pi} e^{Z\cos\alpha}\cos m\alpha\, d\alpha \quad \ldots \tag{1}$$

Set $m = 0$, $\alpha = \phi - \theta$ and $Z = K$, then

$$I_0(K) = \frac{1}{2\pi}\int_0^{2\pi} e^{K\cos(\phi-\theta)} d(\phi-\theta)$$

$$1 = \int_0^{2\pi} \frac{1}{2\pi I_0(K)} e^{K\cos(\phi-\theta)} d\phi .$$

So,

$$f(\phi) = \frac{1}{2\pi I_0(K)} e^{K\cos(\phi-\theta)}, 0 < \phi < 2\pi \quad \ldots. \tag{2}$$

f is the probability density function of the random variable ϕ, where θ is the mean direction of population and K is the parameter of concentration.

3. STATISTICS FOR CONCENTRATION

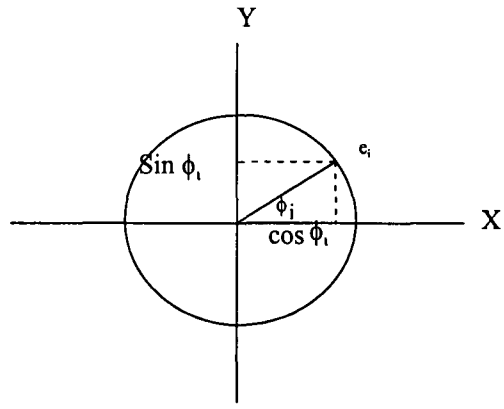
Given $\phi_1, \phi_2, \ldots, \phi_n$ representing a set of sample angular data of size n, we can transform the data to directional data in terms of, say, the unit vector $e_1, e_2, \ldots, e_n$. Thus, the mean vector, $\bar{e} = \sum_{i=1}^{n} \frac{e_i}{n}$. If r denotes the length of the sample mean vector, that is $r = |\bar{e}|$, then the length of the sample mean vector can be written in the following forms (Tragreattikul,1993, pp.24-31):

$$r = \frac{1}{n}\left[(\sum_{i=1}^{n}\cos\phi_i)^2 + (\sum_{i=1}^{n}\sin\phi_i)^2 \right]^{1/2} \tag{3}$$

$$r = \frac{1}{n}\sum_{i=1}^{n}\cos(\phi_i - \bar{\phi}) \quad , \ \bar{\phi} \text{ is the mean angle} \tag{4}$$

Proof:

Figure 1
Diagram showing the unit vector e_i at its angle ϕ_i



Let e_i be the unit vector, i = 1, 2, 3, ... , n. Then

$$x_i = \cos\phi_i \ , \quad \bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{\left(\sum_{i=1}^{n} \cos\phi_i\right)}{n}$$

and

$$y_i = \sin\phi_i \ , \quad \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n} = \frac{\left(\sum_{i=1}^{n} \sin\phi_i\right)}{n}$$

$$r = \left[(\bar{x})^2 + (\bar{y})^2\right]^{1/2}$$

$$= \frac{1}{n}\left[\left(\sum_{i=1}^{n} \cos\phi_i\right)^2 + \left(\sum_{i=1}^{n} \sin\phi_i\right)^2\right]^{1/2}.$$

This proves relation (3). To prove relation (4)

$$\sum_{i=1}^{n} \cos(\phi_i - \bar{\phi}) = \cos\bar{\phi}\sum_{i=1}^{n} \cos\phi_i + \sin\bar{\phi}\sum_{i=1}^{n} \sin\phi_i$$

$$= \frac{\bar{x}}{r} n\bar{x} + \frac{\bar{y}}{r} n\bar{y}$$

$$= \frac{n}{r}\left[(\bar{x})^2 + (\bar{y})^2\right]$$

$$= \frac{nr^2}{r} = nr,$$

that is, $r = \frac{1}{n}\sum_{i=1}^{n} \cos(\phi_i - \bar{\phi})$.

If r tends to 1, this implies that the sample tends to have little dispersion or have more concentration; whereas, if r tends to 0, it implies the opposite. Hence, we can explain sample dispersion by the statistic r. Batschelet (1981) transformed r to $s = [2(1\text{-}r)]^{1/2}180/\pi$ and called it the sample circular standard deviation (unit : degree), $0 \leq s \leq 180\sqrt{2}/\pi$.

4. ESTIMATING THE PARAMETER OF CONCENTRATION

From (1)

$$I_0(K) = \frac{1}{2\pi}\int_0^{2\pi} e^{K\cos\alpha} d\alpha$$

$$I_0'(K) = \frac{1}{2\pi}\int_0^{2\pi} e^{K\cos\alpha}\cos\alpha d\alpha$$
$$= I_1(K)$$

Define the ratio $I_1(K)/ I_0(K) = \rho$. By the Maximum Likelihood Method, we can find an estimator for parameter K, say k, as follows. We have the likelihood function

$$L(K) = \left(2\pi I_0(K)\right)^{-n} e^{K\sum\cos(\phi_i-\theta)}$$

$$\ln L(K) = -n\left(\ln 2\pi + \ln I_0(K)\right) + K\sum_{i=1}^{n}\cos(\phi_i - \theta)$$

$$\frac{\partial}{\partial K}\ln L(K) = \frac{-nI_0'(k)}{I_0(k)} + \sum_{i=1}^{n}\cos(\phi_i - \theta)$$

$$0 = \frac{-nI_1(k)}{I_0(k)} + \sum_{i=1}^{n}\cos(\phi_i - \theta)$$

$$\frac{I_1(k)}{I_0(k)} = \frac{1}{n}\sum_{i=1}^{n}\cos(\phi_i - \theta) \qquad (5)$$

Also by differentiating $\ln L(K)$ with respect to θ, we obtain the maximum likelihood estimator $\bar{\phi}$ for parameter θ. From relation (4),

$$\frac{I_1(k)}{I_0(k)} = \frac{1}{n}\sum_{i=1}^{n}\cos(\phi_i - \bar{\phi})$$

$$\frac{I_1(k)}{I_0(k)} = r, \qquad (6)$$

where (Sneddon, 1972, p.164)

$$I_0(k) = \frac{1}{2\pi}\int_0^{2\pi} e^{k\cos\alpha} d\alpha$$

$$= \frac{1}{2\pi}\int_0^{2\pi}\sum_{i=0}^{\infty}\frac{(k\cos\alpha)^i}{i!} d\alpha$$

$$= 2 + \frac{k^2}{2} + \frac{k^4}{32} + \frac{k^6}{1{,}152} + \ldots$$

$$= \sum_{t=0}^{\infty} \frac{(\frac{k}{2})^{2t}}{(t!)^2} \tag{7}$$

and

$$I_1(k) = \frac{1}{2\pi} \int_0^{2\pi} e^{k\cos\alpha} \cos\alpha d\alpha$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \sum_{i=0}^{\infty} k^i \frac{(\cos\alpha)^{i+1}}{i!} d\alpha$$

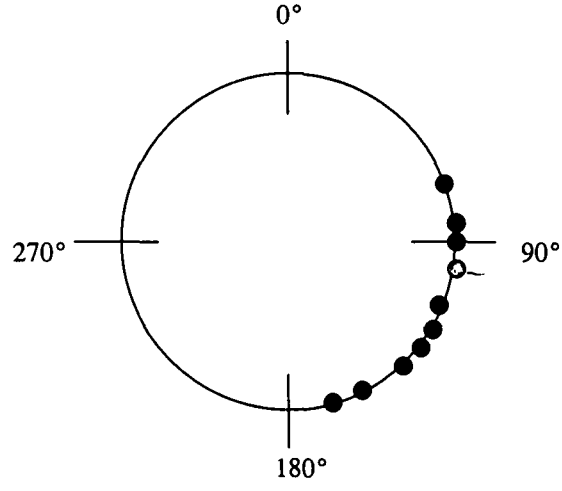
$$= k(1 + \frac{k^2}{8} + \frac{k^4}{192} + \frac{k^6}{9{,}216} + \ldots)$$

$$= \sum_{t=0}^{\infty} \frac{(\frac{k}{2})^{1+2t}}{t!(t+1)!} \tag{8}$$

Thus, r in (6) may be considered as an estimator of resultant ρ. From equation (8), we see that the relation between k and r is not simple. A solution can be obtained by either numerical analysis or computer analysis. Mardia (1972, pp.122-123) constructed a relationship between r and k for the extreme case, $r \approx (1 - k^2/8 + k^4/48)k/2$ or $k \approx (12 + 6r^2 + 5r^4)r/6$ for small values of r ($r < 0.45$) and $r \approx 1 - 1/2k - 1/8k^2 - 1/8k^3$ or $1/k \approx 2(1-r) - (1-r)^2 - (1-r^3)$ for large values of r ($r > 0.8$). For other values of r he showed (p.298) a table of k and r based on a table prepared by Batshelet in 1965.

EXAMPLE: Given an angular data sample, ϕ (degree) : 85, 90, 75, 98, 120, 130, 125, 137, 160, 150, as shown in Figure 2.

Figure 2
Data plot in the form of a circular diagram



$$r = \frac{1}{n}\left[(\sum_{i=1}^{n} \cos\phi_i)^2 + (\sum_{i=1}^{n} \sin\phi_i)^2 \right]^{1/2}$$

$$= \frac{1}{10}\left[(-4.0467)^2 + (7.9276)^2\right]^{1/2} = 0.8901.$$

From $s = [2(1-r)]^{1/2}180/\pi$ and $I_1(k)/I_0(k) = r$, we have $s = 26.86$ and $k = 4.84509$

5. HYPOTHESIS TESTING FOR CONCENTRATION

For testing the hypothesis concerning concentration, Ho : $K = c$ (c is constant), we transform K to ρ using Batschelet's table or when appropriate use Mardia's approximation. Then from ρ we get an expression for the population circular standard deviation σ by using the relation $\sigma = \left[2(1-\rho)\right]^{1/2} \frac{180}{\pi}$. We compute $\chi^2 = ns^2 / \sigma^2$ and accept Ho if the value falls within the critical values of the chi-square variate with n-1 degrees of freedom; otherwise, reject Ho.

From the previous example, we can test the null hypothesis, Ho : $K = 6$. First, $K = 6$ is transformed to $\rho = 0.91262$ then to $\sigma = 23.95$, and the test statistic, $\chi^2 = 10(26.86)^2/(23.95)^2 = 12.578$. For a level of significance, $\alpha = 10\%$ and degrees of freedom 9, we have a critical region $\chi^2_{.05} < 3.3$ and $\chi^2_{.95} > 16.9$. Thus, the null hypothesis is not rejected.

6. INTERVAL ESTIMATION FOR CONCENTRATION

From the Chi-square variable, a $(1-\alpha)100\%$ confidence interval for population circular variance is constructed as follows:

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

The lower and upper limits of the population circular variance are then transformed to the upper and lower limits of the population circular concentration, say k_1 and k_2 respectively, that is $k_1 < K < k_2$ via the relations between s and r, and then r and k.

From the previous example, a 90% confidence interval for population concentration is constructed via:

$$\begin{aligned} 10(26.86)^2/16.9 &< \sigma^2 < 10(26.86)^2/3.3 \\ 20.66 &< \sigma < 46.76 \\ 0.94 &> \rho > 0.67 \\ 8.6104 &> K > 1.8418 \end{aligned}$$

Applying Stephen's Formula (Mardia, 1972, pp.150-151) to the previous data when the mean direction is unknown, the 90% confidence interval for population concentration K, is given by $1.81 < K < 8.67$. This would seem to imply a more efficient interval estimate of the concentration parameter K if the Chi-square approximation is used except that the intervals cannot be validly compared because of the difference in the ranges of the standard deviations.

7. CONCLUSION

Estimating and testing the concentration parameter K in directional data can produce reasonable results by using some Chi-square approximations. However, this method may be more efficient only if the sample has a high concentration. This is because the sample circular variance by Batschelet's Formula does not exceed $180\sqrt{2}/\pi$.

ACKNOWLEDGMENT

The author would like to thank Dr. Robert Molloy for his reading of the manuscript and valuable suggestions for its improvement.

REFERENCES

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

MARDIA, K.V. (1972), *Statistics of Directional Data*, London and New York: Academic Press

SNEDDON, I.H. (1972), *The Use of Integral Transforms*, New York: McGraw-Hill

TRAGREATTIKUL, S. (1993), *Circular Distribution*, M.S. Thesis, Chiang Mai University