# Session 2(c): Statistical standard, methodology and application in data management and usage

## Estimation of Confidence Intervals for Concentration Parameter in Von Mises Distribution Using A New Statistic Based on Circular Distance.

Siti Fatimah binti Hassan

# Outline

- Circular Statistics
- Von Mises Distribution
- New Statistic of Circular Distance
- Confidence Intervals (CI) Based on Circular Distance
- Simulation Result & Discussion
- Illustrative Example
- Conclusion

# Circular Statistics

- Circular or directional statistics is a branch of statistics deals with data points distributed on a circle.
- It uses angle as the measurements of directions in the range of $(0, 2\pi)$ radians or $(0^o, 360^o)$.
- Examples of circular data: compass direction, wind direction, wave direction, orientation of animal
- This type of data occurs in many fields: meteorology, biology, geology, geography and medicine.
- The von Mises distribution is also known as the circular normal distribution and is a continuous probability distribution.
- It may be thought of as the circular version of the normal distribution, since it describes the distribution of a random variate with period $2\pi$.

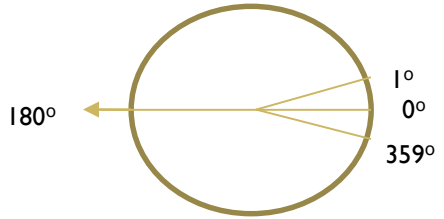



Figure 1: Arithmetic mean pointing the wrong way

- For example, as shown in Figure 1 for the measurements of wind direction data, the calculated arithmetic mean for 1° and 359° using conventional linear techniques is 180° .
- Whereas using circular statistics, the mean direction should be 0°.

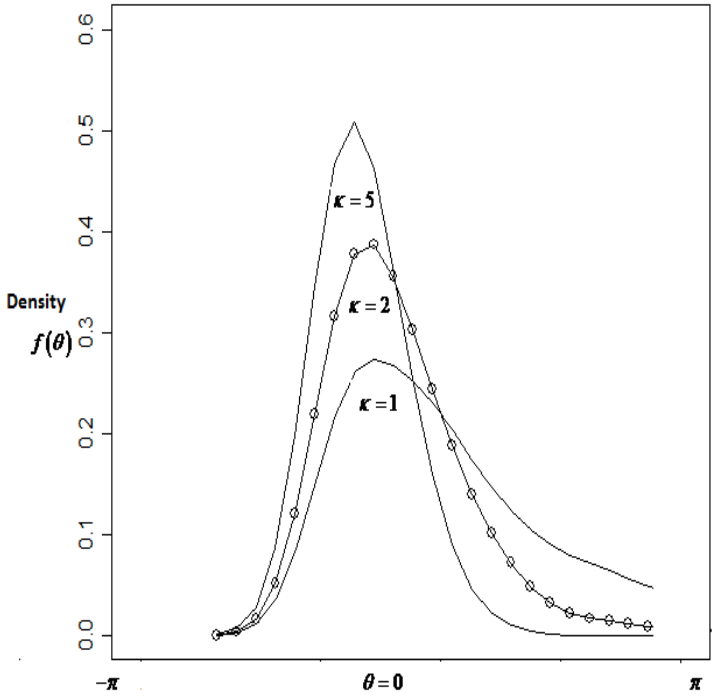
# Von Mises distribution

- For a notation, the von Mises distribution is denoted by $VM(\mu_0, \kappa)$ where $\mu_0$ $(0 \leq \mu_0 \leq 2\pi)$ is the mean direction, the $\kappa$ is known as the concentration parameter and has probability density function given by

$$g(\theta; \mu_0, \kappa) = \{2\pi I_0(\kappa)\}^{-1} \exp\{\kappa \cos(\theta - \mu_0)\}$$

- $I_0$ denotes as the modified Bessel function of the first kind and order zero, Mardia & Jupp (2000).

- the larger the value of concentration parameter, the greater the clustering around the mode or the higher concentration towards the population mean direction.



# New Statistic of Circular Distance

*Proposition 1*

Suppose $\theta_1,\ldots,\theta_n$ be *i.i.d* observation from von Mises distribution with mean direction, $\mu$ and concentration parameter $\kappa$.

Then for $j=1,\ldots,n$:

$$G_j = \kappa\left[n - C\cos\theta_j - S\sin\theta_j\right] \sim \chi^2_{n-1} \text{ as } \kappa \to \infty \qquad (I)$$

where $C=\sum_{i=1}^{n}\cos\theta_i$ and $S=\sum_{i=1}^{n}\sin\theta_i$.

# The proof of Proposition 1

Suppose $\theta_i,\ldots,\theta_n$ is a random variable from $VM(\mu,\kappa)$. For any observation $\theta_i$ and large $\kappa$, it is shown by Jammalamadaka & SenGupta (2001) that,

$$\sqrt{\kappa}(\theta_i-\mu)\rightarrow N(0,1) \text{ as } \kappa\rightarrow\infty \quad (2)$$

Since $\theta_i$ and $\theta_j$ are independent observations, $\sqrt{\frac{\kappa}{2}}(\theta_i-\theta_j)\rightarrow N(0,1)$ .

From the properties of standard normal distribution, it can be approximate to Chi squared distribution as below

$$\frac{\kappa}{2}(\theta_i-\theta_j)^2\rightarrow\chi_1^2 \quad (3)$$

For large value of concentration parameter, the distribution of von Mises distribution is said to be more concentrated. This highly concentrated distribution will lead to shorter circular distance between two points.

# The proof of Proposition 1

From the second Taylor series expression,

we have $\cos\alpha \approx 1-\frac{\alpha^2}{2}$ or $\frac{\alpha^2}{2} \approx 1-\cos\alpha$

Substitute for $\alpha = \theta_i - \theta_j$, we have,

$$\frac{\left(\theta_i-\theta_j\right)^2}{2} = 1-\cos\left(\theta_i-\theta_j\right) = 1-\cos\theta_i\cos\theta_j-\sin\theta_i\sin\theta_j \qquad (4)$$

Hence, substituting (4) in (3),

$$\kappa\left(1-\cos\theta_i\cos\theta_j-\sin\theta_i\sin\theta_j\right) \sim \chi_1^2 \qquad (5)$$

Further due to independent of $\theta_i$ and $\theta_j$, for $i \neq j$,

$$\sum_{i \neq j} \kappa\left(1 - \cos\theta_i \cos\theta_j - \sin\theta_i \sin\theta_j\right) \sim \chi^2_{n-1} \text{ or}$$

$$\kappa\left[(n-1) - \cos\theta_j \sum_{i \neq j} \cos\theta_i - \sin\theta_j \sum_{i \neq j} \sin\theta_i\right] \sim \chi^2_{n-1} \quad (6)$$

where $C = \sum_{i=1}^{n} \cos\theta_i = \sum_{i \neq j}^{n} \cos\theta_i + \cos\theta_j$ and $S = \sum_{i=1}^{n} \sin\theta_i = \sum_{i \neq j}^{n} \sin\theta_i + \sin\theta_j$

Thus, $\kappa\left[(n-1)-\cos\theta_j\{C-\cos\theta_i\}-\sin\theta_j\{S-\sin\theta_j\}\right]$.

Hence,

$$G_j = \kappa\left[n - C\cos\theta_j - S\sin\theta_j\right] \sim \chi^2_{(n-1)}$$

$$= \kappa\left[n-1-C\cos\theta_j+\cos^2\theta_j - S\sin\theta_j+\sin^2\theta_j\right]$$

$$= \kappa\left[n - C\cos\theta_j - S\sin\theta_j\right] \qquad (7)$$

# The empirical proof of Proposition 1

For this purpose, the Kolmogorov-Smirnov test is used to identify the suitable samples as well as the concentration parameter that can be used in this approximation.

Table 1: The percentage of samples correctly approximated by the Chi Squared distribution with df $(n-1)$:

| $\kappa$ | Sample size, $n$ | | | | | |
|---|---|---|---|---|---|---|
| | 10 | 20 | 30 | 50 | 70 | 100 |
| 2 | 53.2 | 64.4 | 82.1 | 99.0 | 100.0 | 100.0 |
| 4 | 49.8 | 67.0 | 85.9 | 99.9 | 100.0 | 100.0 |
| 6 | 50.0 | 67.0 | 89.0 | 99.8 | 100.0 | 100.0 |
| 8 | 49.7 | 70.8 | 90.3 | 100.0 | 100.0 | 100.0 |
| 10 | 48.4 | 69.8 | 90.9 | 100.0 | 100.0 | 100.0 |

- For $n = 10$, the percentage is a decreasing function for all $\kappa$ .
- For the range $20 \leq n \leq 50$, the percentage is an increasing function for all $\kappa$, while constant for $70 \leq n \leq 100$.
- For any $\kappa$ and $n \geq 50$, more than 99% can be approximated to Chi Squared distribution with df $(n - 1)$.

**Conclusion**: For sample size $n \geq 30$, for von Mises distribution can be approximated by Chi Squared distribution with df $(n - 1)$.

# Confidence intervals (CI) based on circular distance

Recall that $G_j = \kappa\left[ n - C\cos\theta_j - S\sin\theta_j \right] \sim \chi^2_{(n-1)}$. Hence,

solving for $\kappa$, we get

$$\chi^2_{\left(n-1,\frac{\alpha}{2}\right)} < \kappa\left[n - C\cos\theta_j - S\sin\theta_j\right] < \chi^2_{\left(n-1,1-\frac{\alpha}{2}\right)}$$

$$\frac{\chi^2_{\left(n-1,\frac{\alpha}{2}\right)}}{\left[n - C\cos\theta_j - S\sin\theta_j\right]} < \kappa < \frac{\chi^2_{\left(n-1,1-\frac{\alpha}{2}\right)}}{\left[n - C\cos\theta_j - S\sin\theta_j\right]}$$

$$\frac{\chi^2_{\left(n-1,\frac{\alpha}{2}\right)}}{A_j} < \kappa < \frac{\chi^2_{\left(n-1,1-\frac{\alpha}{2}\right)}}{A_j} \quad \text{where } A_j = \left[ n - C\cos\theta_j - S\sin\theta_j \right] \qquad (8)$$

# Method of CI

- Method 1: Mean

$$\text{Lower limit} = \frac{1}{n}\sum_{j=1}^{n}\kappa_L^j, \qquad \text{Upper limit} = \frac{1}{n}\sum_{j=1}^{n}\kappa_U^j$$

Hence,

$$CI_{(mean)} = \left(mean\left(\kappa_L^j\right), mean\left(\kappa_U^j\right)\right) \tag{9}$$

- Method 2: Median

$$\text{Lower limit} = med\left(\kappa_L^j\right), \text{ Upper limit} = med\left(\kappa_U^j\right)$$

Hence,

$$CI_{(med)} = \left(med\left(\kappa_L^j\right), med\left(\kappa_U^j\right)\right) \tag{10}$$

# Method of CI

- Method 3: Percentile

The simulation will be carried out to identify the most potential percentile that can be used as the CI. From the simulation studies, these following steps must be follow:

- **Step 1**: All values of concentration parameter in lower limit and upper limit sets are sorted in ascending order. It then, will be divided into various percentages for further evaluation.
- **Step 2**: From the results, the most potential cut of point of percentile that will produce $(100-\alpha)100\%$ of target values is noted. We note that for $\alpha = 0.05$ or the 95% target values lie between 30th to 50th percentile.
- **Step 3**: Finally, each new percentage in Step 2 will be examined to asses how well they produce the target value of 0.95 or 95% of CI.

# Simulation result & discussion

Table 2: Coverage probability for various value of $\kappa$ for each sample size, $n$ = 30, 50, 70 and 100.

| Sample size, $n$ | Concentration parameter, $\kappa$ | Mean | Med | Percentile 34% |
|---|---|---|---|---|
| 30 | 2 | 0.945 | 0.913 | 0.968 |
| | 4 | 0.904 | 0.862 | 0.945 |
| | 6 | 0.884 | 0.841 | 0.939 |
| | 8 | 0.880 | 0.833 | 0.938 |
| | 10 | 0.876 | 0.831 | 0.932 |
| 50 | 2 | 0.919 | 0.856 | 0.970 |
| | 4 | 0.844 | 0.765 | 0.943 |
| | 6 | 0.819 | 0.739 | 0.943 |
| | 8 | 0.805 | 0.725 | 0.939 |
| | 10 | 0.799 | 0.726 | 0.936 |

- Median gives the poorest performance in which the coverage probability is far from the target values in comparison to the other methods.
- Percentile method is the best method because the coverage probability are consistently close to the target values.
- For $\kappa = 2$, 30th percentile gives the coverage probability that is close to the target value.
- For all sample sizes and $\kappa \geq 4$, 34th percentile consistently gives the best coverage probability with values close to the target value in comparison to other percentiles as well as the mean and median.

# Simulation result & discussion

Table 2 *cont'd*

| Sample size, $n$ | Concentration parameter, $\kappa$ | Mean | Med | Percentile 34% |
|---|---|---|---|---|
| 70 | 2 | 0.896 | 0.804 | 0.970 |
| | 4 | 0.796 | 0.689 | 0.946 |
| | 6 | 0.747 | 0.632 | 0.936 |
| | 8 | 0.726 | 0.621 | 0.936 |
| | 10 | 0.723 | 0.618 | 0.934 |
| 100 | 2 | 0.852 | 0.706 | 0.971 |
| | 4 | 0.713 | 0.564 | 0.946 |
| | 6 | 0.646 | 0.499 | 0.936 |
| | 8 | 0.622 | 0.478 | 0.929 |
| | 10 | 0.606 | 0.476 | 0.931 |

- Median gives the poorest performance in which the coverage probability is far from the target values in comparison to the other methods.
- Percentile method is the best method because the coverage probability are consistently close to the target values.
- For $\kappa = 2$, 30th percentile gives the coverage probability that is close to the target value.
- For all sample sizes and $\kappa \geq 4$, 34th percentile consistently gives the best coverage probability with values close to the target value in comparison to other percentiles as well as the mean and median.

# Simulation result & discussion

Table 3: Expected length for various value of $\kappa$ for each sample size, $n$ = 30, 50, 70 and 100.

| Sample size, $n$ | Concentration parameter, $\kappa$ | Mean | Med | Percentile 34% |
|---|---|---|---|---|
| 30 | 2 | 2.528 | 2.700 | 2.089 |
| | 4 | 5.405 | 5.708 | 4.277 |
| | 6 | 8.315 | 8.741 | 6.475 |
| | 8 | 11.177 | 11.747 | 8.657 |
| | 10 | 14.048 | 14.739 | 10.849 |
| 50 | 2 | 1.912 | 2.045 | 1.618 |
| | 4 | 4.080 | 4.313 | 3.308 |
| | 6 | 6.263 | 6.592 | 5.011 |
| | 8 | 8.429 | 8.859 | 6.707 |
| | 10 | 10.569 | 11.090 | 8.382 |

- Median gives the widest length in comparison to the other methods.
- Percentile method gives the narrowest expected length as compared to the other methods. Hence, it also can be concluded that percentile method is the superior method as compared to the other methods using expected length as the performance indicator.
- Using both measures of performance (coverage probability and expected length), it can be concluded that 34th percentile gives the most efficient CI in comparison to all methods as it give good coverage probability as well narrow expected length.

# Simulation result & discussion

Table 3 *cont'd*

| Sample size, $n$ | Concentration parameter, $\kappa$ | Mean | Med | Percentile 34% |
|---|---|---|---|---|
| 70 | 2 | 1.592 | 1.704 | 1.362 |
| | 4 | 3.391 | 3.588 | 2.781 |
| | 6 | 5.234 | 5.519 | 4.237 |
| | 8 | 7.041 | 7.405 | 5.664 |
| | 10 | 8.814 | 9.266 | 7.070 |
| 100 | 2 | 1.324 | 1.417 | 1.129 |
| | 4 | 2.811 | 2.975 | 2.306 |
| | 6 | 4.330 | 4.568 | 3.504 |
| | 8 | 5.833 | 6.138 | 4.685 |
| | 10 | 7.317 | 7.690 | 5.863 |

- Median gives the widest length in comparison to the other methods.
- Percentile method gives the narrowest expected length as compared to the other methods. Hence, it also can be concluded that percentile method is the superior method as compared to the other methods using expected length as the performance indicator.
- Using both measures of performance (coverage probability and expected length), it can be concluded that 34th percentile gives the most efficient CI in comparison to all methods as it give good coverage probability as well narrow expected length.

# Conclusion

- A new statistics for circular distance in von Mises distribution was proposed and CI for concentration parameter are developed based on mean, median and percentile.
- All the three proposed methods of obtaining CI provide alternate approaches and are appealing due to the simplicity of getting the CI.
- However, based on simulation studies, CI based on percentile is the most superior of the three proposed method.
- The CI based on percentile consistently gives good coverage probability as well as the smallest expected length.
- Median gives the poorest performance in which the coverage probability is far from the target values as well as having the widest length in comparison to the other methods.

# THANK YOU