Malaysia Statistics Conference 2016
15 November 2016
Sasana Kijang, Bank Negara Malaysia
Strengthening Statistical Usage for Decisions and Innovation

A NEW DISCORDANCY TEST IN CIRCULAR DATA USING SPACINGS THEORY

Adzhar Bin Rambli
Joint Research with:
Prof Dr Ibrahim Mohamed
Prof Dr Abdul Ghapor Hussin

A REVIEW: CIRCULAR STATISTICS

- Circular statistics is a branch of statistics deals with data points distributed around a unit circle.

Examples: Biology (Animals navigation), Meteorology (wind and wave directions).

- Due to the bounded range property of circular variables, special methods are required to analyze circular data.

Example 1

Consider the difference between two observations ; A=315° and B=45°.

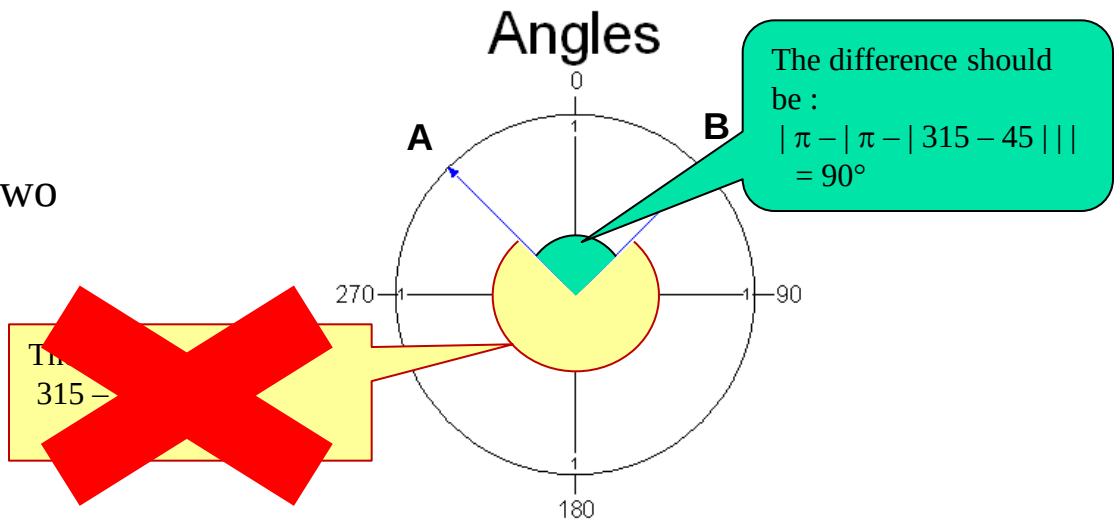


Fig1. Circular plot

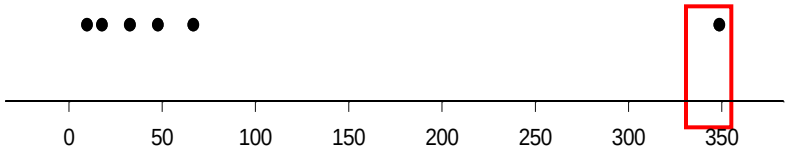
A REVIEW: OUTLIER

- Describe as a value with large circular distances from the value to the two neighbouring observations on a unit circle.

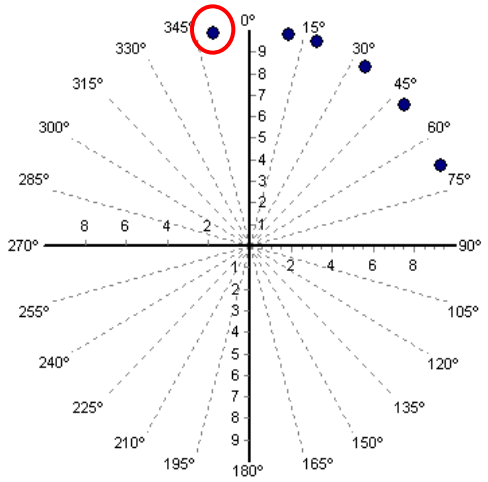
An example, consider the following data set | 1 , 5 , 11 , 22 , 35 , 90 , 350.

Linear Case: 350 is an outlier.

Circular Case: 350 is consistent with the other observations.



(a)



(b)

Fig 2. Linear and circular representation of data set

OBJECTIVES

1) To review circular samples and existed outlier detection method.
2) To review spacing's theory.
3) To develop a new test of discordance based on gaps between observations
4) To investigate the performance of new test.
5) To illustrate a practical example based on an eye data set.

Conclusion

The Circular Samples

- 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 stable distribution with the wrapped Cauchy and the wrapped normal distributions as the special cases.
- In this study, we use the von Mises distribution (also known as the circular normal distribution) which is the most commonly used which is a continuous probability distribution on a circle.

The Discordancy Tests

•There are four tests of discordancy in circular samples (*M*, *C*, *D* and *A* statistics).

Table 1

Statistic	Formula	Remarks
M Mardia (1975)	$M' = \min_i \left\{ \frac{n-1-R_{(-i)}}{n-R} \right\}$	$R = \sqrt{S^2 + C^2}$
C Collett (1980)	$C = \max_i \left\{ \frac{\bar{R}_{(-i)} - \bar{R}}{\bar{R}} \right\}$	$\bar{R} = \frac{R}{n}$
D Collett (1980)	$D_i = \frac{T_i}{T_{i-1}}, \quad D = \min\left(D_k, D_k^{-1}\right)$	$T_i = \theta_{(i+1)} - \theta_{(i)}, \quad (i = 1,..,n-1)$ $T_n = 2\pi - \theta_{(n)} - \theta_{(1)}$
A Abuzaid (2010)	$A = \max_j \left\{ \frac{D_j}{2(n-1)} \right\} \; (j = 1,..,n)$	$D_j = \sum_{i=1}^{n} \left(1 - \cos(\theta_i - \theta_j)\right)$

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

back

The Spacings Theory

1. Suppose $\theta_1, \theta_2, ..., \theta_n$ are $(i.i.d)$ circular observations located on the circumference of a unit circle while $\theta_{(1)}, \theta_{(2)}, ..., \theta_{(n)}$ are the corresponding ordered circular observations.
2. Following notation by Rao (1976), we define the one-step spacing for the i-th ordered observation as

$$G_{1i} = \theta_{(i+1)} - \theta_{(i)}, \quad i = 1, 2, ..., n, \qquad (5)$$

and $G_{1n} = 2\pi - \theta_{(n)} + \theta_{(1)}$.

3. Note that $\{G_{1i}, i = 1, 2, ..., n\}$ gives a sequence of distances between successive observations on the circumference.

The Spacing's Theory

4. The statistic (5) can be generalized to detect a patch of outliers in circular data.
5. For that, we define G_{ai} as the a-step spacing for the ith ordered observation, $a = 1,2,3,\ldots$ and $i = 1,2,\ldots,n$ such that

$$G_{ai} = \theta_{i+a} - \theta_i \quad \text{for} \quad i = 1,2,\ldots,n-a \qquad (6)$$

and $G_{ai} = 2\pi - \theta_i + \theta_{(i+a)-n}$ for $i = (n+1)-a,(n+2)-a,\ldots,n$.

6. We will use the statistics (6) in the development of a new discordancy test, denoted by G_a, for detecting a single, multiple as well as a patch of outliers.

back

A new statistics of discordance: The G_a Statistic

Suppose $\theta_1, \theta_2, ..., \theta_n$ are $(i.i.d)$ circular observations from a *VM* distribution. The steps to obtain the G_a statistic is described below:

Step 1 Order the observations as $\theta_{(1)}, \theta_{(2)}, ..., \theta_{(n)}$.

Step 2 For a choice of *a*-step spacing, calculate G_{ai}, $i = 1, 2, ..., n$ as given in equation (6).

Step 3 Define $G_i = \min\left(G_{ai}, G_{a,i-a}\right)$ for $i = 1, 2, ..., n,$ which is the smaller of the *a*-step spacing on either side of θ_i.

Step 4 Define $G_a = \max_{i=1,2,...,n} \left(G_i\right)$.

If the value of G_a exceeds a pre-determined cut-off point, say C_g, then the *i*th observation corresponding to $\max_{i=1,2,...,n} \left(G_{ai}\right)$ is identified as an outlier.

back

The Spacing's Theory

Let $\theta_3 = 32^\circ$, if we use one step spacing $a = 1$;

$$G_i = \min\left(G_{ai}, G_{a,i-a}\right)$$

$$G_{ai} = \theta_{i+a} - \theta_i, \qquad G_{a(i-a)} = \theta_i - \theta_{i-a}$$

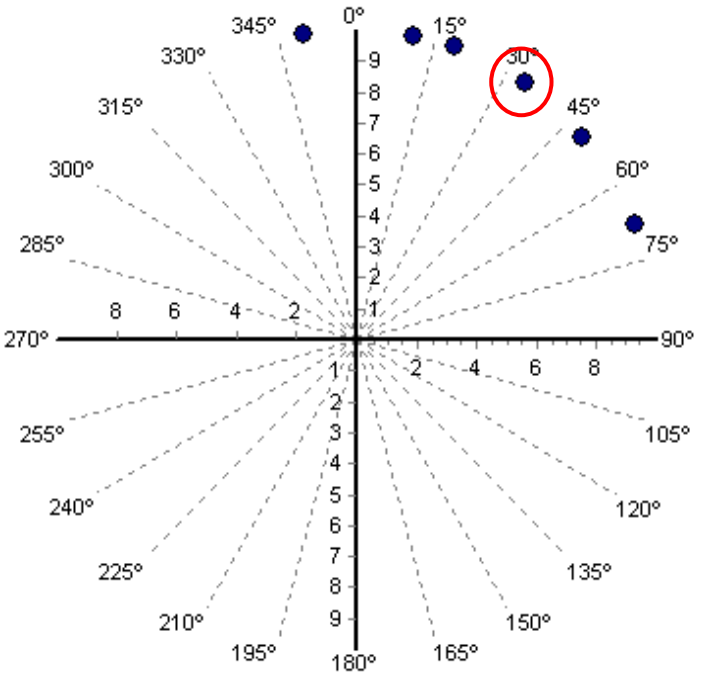
$$G_{13} = \theta_4 - \theta_3 \qquad G_{12} = \theta_3 - \theta_2$$

$$G_{13} = 47^\circ - 32^\circ \qquad G_{12} = 32^\circ - 16^\circ$$

$$= 15^\circ \qquad = 16^\circ$$

$$G_{(i=3)} = \min\left(G_{13}, G_{12}\right)$$

$$G_{(a=1)} = \max_{i=1,\ldots,n}\left(G_i\right)$$



The Spacing's Theory

Let $\theta_3 = 32^\circ$, if we use one step spacing $a = 2$;

$$G_i = \min\left(G_{ai}, G_{a,i-a}\right)$$

$$G_{ai} = \theta_{i+a} - \theta_i, \qquad G_{a(i-a)} = \theta_i - \theta_{i-a}$$

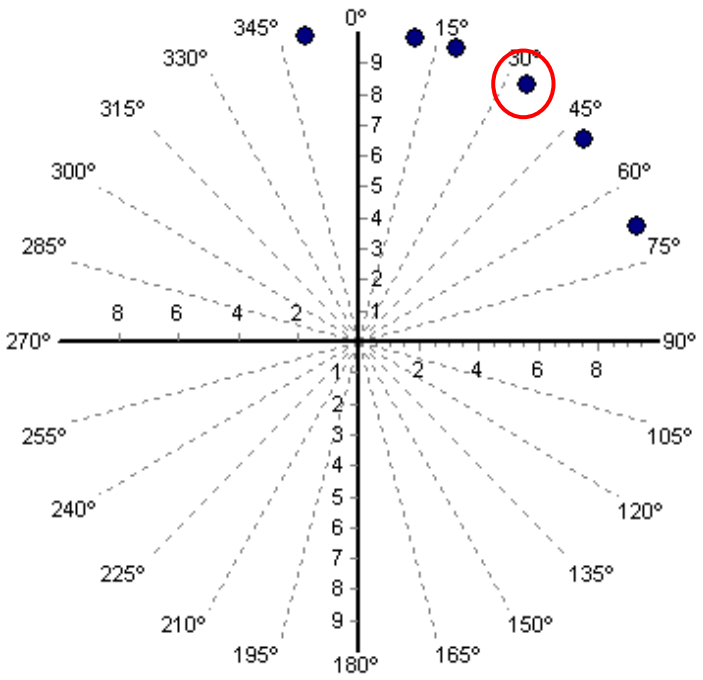
$$G_{23} = \theta_5 - \theta_3 \qquad G_{21} = \theta_3 - \theta_1$$

$$G_{23} = 73^\circ - 32^\circ \qquad G_{21} = 32^\circ - 8^\circ$$

$$= 41^\circ \qquad = 24^\circ$$

$$G_{(i=3)} = \min\left(G_{23}, G_{22}\right)$$

$$G_{(a=2)} = \max_{i=1,\ldots,n}\left(G_i\right)$$



back

Performance Of The G_1 Statistic

1. David (1970, p.185) and Barnett and Lewis (1984, p.64-68) stated that a good test should have
   (i) a high power function,
   (ii) a high probability of identifying a contaminating value as an outlier when it is indeed an extreme value, where an extreme value is defined as a point with a maximum circular distance from the mean direction of the data
   (iii) a low probability of wrongly identifying a good observation as discordance, which is an observation that does not belong to the pre-assumed VM distribution.

Performance Of The G_1 Statistic

1. Let

   i. $P1 = (1 - \beta)$ be the power function where β is the type-II error;

   ii. $P3$ the probability that the contaminant point is an extreme point and is identified as discordance; and

   iii. $P5$ the probability that the contaminant point is identified as discordance given that it is an extreme point.

2. A good test is expected to have (i) **high** $P1$, (ii) **high** $P5$ and (iii) **low** $P1 - P3$.

Performance Of The G_1 Statistic

1. Samples are generated in such a way that $(n-1)$ of the observations come from $VM(\alpha, \kappa)$ and one observation from $VM(\alpha + \lambda\pi, \kappa)$, where λ is the degree of contamination and $0 \leq \lambda \leq 1$.

2. We use different sample sizes n in the range [20, 100] and different values of κ in the range [5.29, 7.5].

3. The measures of performance of the G_1, C, D and A statistics are then calculated.

4. The process is carried out 4000 times.

Performance Of The G_1 Statistic

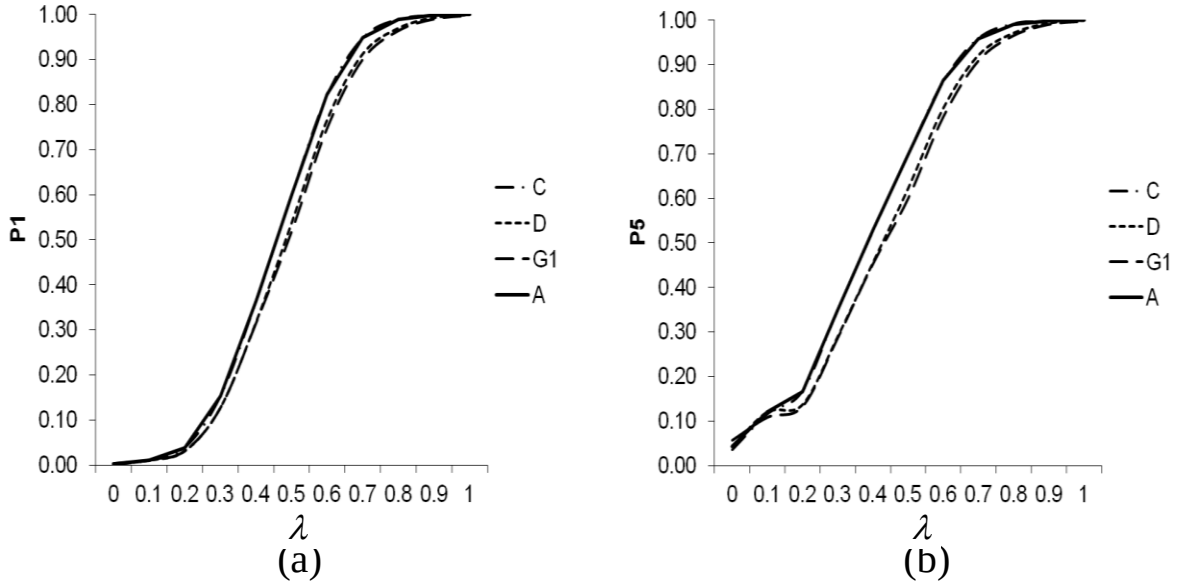


Figure 2 : Performance of C, D, G_1 and A statistics when n=20, $\kappa = 5.29$

Advantage Of The G_a Statistic

- All four statistics can also be used to detect multiple outliers that are well separated from each other.

- The proposed G_1 statistic holds an advantage above the others as the only statistic that can be generalized to detect a patch of outliers in a circular sample.

Advantage Of The G_a Statistic

- All four statistics can also be used to detect multiple outliers that are well separated from each other.

- The proposed G_1 statistic holds an advantage above the others as the only statistic that can be generalized to detect a patch of outliers in a circular sample.

back

Performance Of The G_a Statistic

1. To study the performance of the G_a statistic for a = 1, 2, 3, 4, we generate samples based on different sizes n = 5, 20 and 100 and concentration parameter values κ = 5.29, 7.42 and 10.27.
2. The samples are generated in such a way that $n-a$ of the observations come from $VM(\alpha,\kappa)$ and the remainder from $VM(\alpha+\lambda\pi,\kappa=10.27)$, $0\leq\lambda\leq1$.
3. We set $\kappa=10.27$ so that the outlying observations are clustered in a single patch.

back

Performance Of The G_a Statistic

4. The G_a statistic in each random sample is then calculated.
5. If G_a is greater than the corresponding cut-off point, then we have correctly detected the patch of a outliers.
6. We repeat the simulation 4000 times and obtain the proportion of correct detection of the patch of outliers introduced into the samples.

back

Performance Of The G_a Statistic

Table 5. Proportion of correct detection of patches of outliers

		Single outlier			A patch of 2 outliers		
	κ	5.29	7.42	10.27	5.29	7.42	10.27
n	λ						
20	0	0.00	0.00	0.00	0.00	0.00	0.00
	0.25	0.02	0.07	0.19	0.00	0.01	0.06
	0.5	0.46	0.74	0.89	0.19	0.62	0.86
	0.7	0.91	0.98	1.00	0.85	0.98	1.00
	1	1.00	1.00	1.00	1.00	1.00	1.00
50	0	0.00	0.00	0.00	0.00	0.00	0.00
	0.25	0.01	0.04	0.13	0.00	0.00	0.02
	0.5	0.28	0.65	0.85	0.11	0.46	0.78
	0.7	0.83	0.98	1.00	0.77	0.97	1.00
	1	0.99	1.00	1.00	0.99	1.00	1.00
100	0	0.00	0.00	0.00	0.00	0.00	0.00
	0.25	0.00	0.02	0.08	0.00	0.00	0.01
	0.5	0.16	0.57	0.80	0.04	0.33	0.71
	0.7	0.74	0.95	0.99	0.62	0.94	0.99
	1	0.98	1.00	1.00	0.99	1.00	1.00

back

Performance Of The G_a Statistic

		A patch of 3 outliers			A patch of 4 outliers		
	κ	5.29	7.42	10.27	5.29	7.42	10.27
n	λ						
20	0	0.00	0.00	0.00	0.00	0.00	0.00
	0.25	0.00	0.00	0.01	0.00	0.00	0.00
	0.5	0.13	0.53	0.80	0.07	0.49	0.78
	0.7	0.83	0.98	1.00	0.81	0.99	1.00
	1	0.99	1.00	1.00	0.99	1.00	1.00
50	0	0.00	0.00	0.00	0.00	0.00	0.00
	0.25	0.00	0.00	0.00	0.00	0.00	0.00
	0.5	0.03	0.39	0.74	0.01	0.24	0.71
	0.7	0.68	0.97	1.00	0.65	0.96	1.00
	1	0.99	1.00	1.00	1.00	1.00	1.00
100	0	0.00	0.00	0.00	0.00	0.00	0.00
	0.25	0.00	0.00	0.00	0.00	0.00	0.00
	0.5	0.01	0.20	0.64	0.00	0.17	0.57
	0.7	0.57	0.94	0.99	0.42	0.94	0.99
	1	0.99	1.00	1.00	0.99	1.00	1.00

back

The Practical Example

- We consider an eye data set obtained from a glaucoma clinic at the University of Malaya Medical Center, Malaysia.
- Images of the posterior segment of the eyes of 23 patients were taken using the Anterior Segment Optical Coherence Tomography **(**AS-OCT).
- The variable of our interest is the angle of posterior corneal curvature defined as follow.

The Practical Example

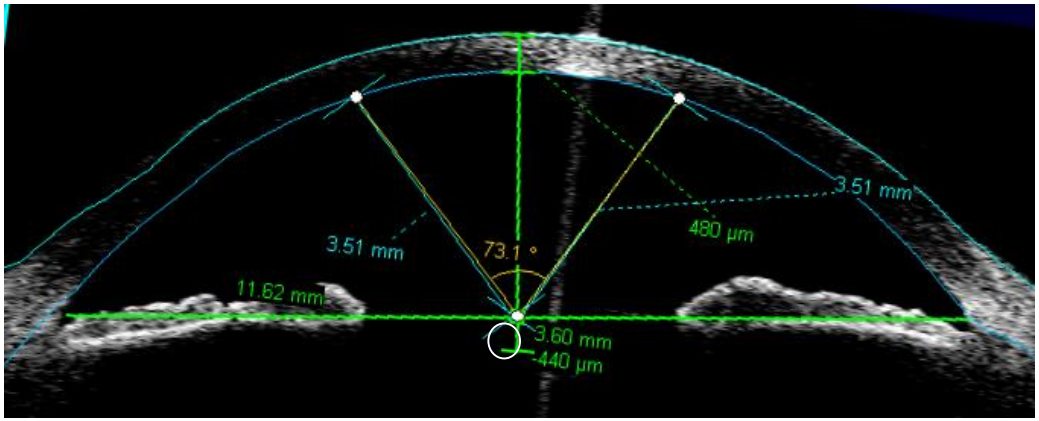


Figure 1: Posterior corneal curvature measurement

1. In the diagram, O is the intersection of the geometrical axis of the eye (horizontal line) with the line made between the nasal and temporal scleral spurs (vertical line).
2. From O, we draw radii to the posterior surface of the cornea in the range [3.49,3.51] mm.
3. Then, the angle of the area generated by radii is called the angle of posterior corneal curvature as shown in Figure 1.

The Practical Example

1. The circular plot of the Eye data is given in Figure 2.
2. It can be seen that a patch of two observations lies further away from the rest.
3. Furthermore, the P-P plot of angle of posterior corneal curvature given in Figure 3 indicates that the data follow a *VM* distribution.
4. We can therefore apply the proposed discordancy test on the data.

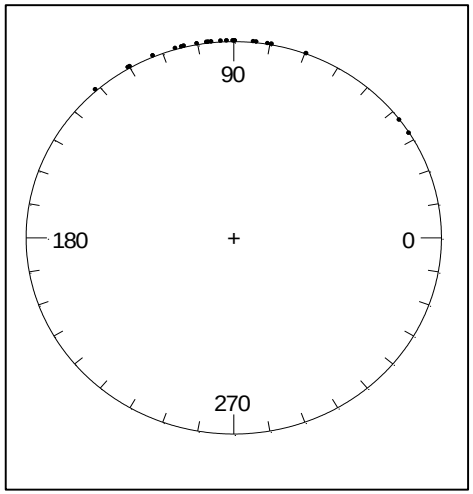


Figure 2: Circular plot of angle of posterior corneal curvature

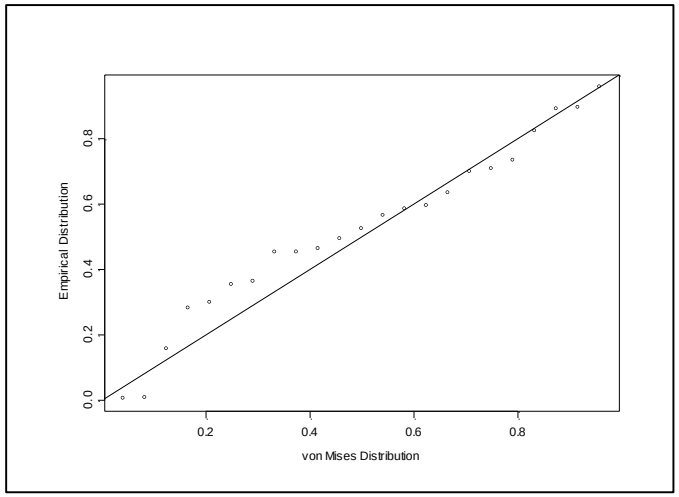


Figure 3: P-P plot of angle of posterior corneal curvature

back

The G_a-Statistic

1. Summary statistics for the given data set are calculated; the estimated mean direction is $\hat{\mu} = 1.61$ radians or 92° and the estimated concentration parameter is $\hat{\kappa} = 6.84$.

   Table 4. Result based on G_2 statistic

Observation	G_2 statistic	Cut-off Point
10	0.68	0.67
17	0.78	

2. Note that we also identify both observations as a patch of 2 outliers using G_3 and G_4 but not G_1.
3. Further, the deletion of these points from the original data changes the parameter esimates to $\hat{\mu} = 1.69$ or 97° and $\hat{\kappa} = 15.5$, indicating that the estimated $\hat{\kappa}$ is significantly affected by the existence of these outliers in the data as expected.

back

CONCLUSION

1) Have reviewed circular samples and existed outlier detection method.
2) Have reviewed spacing's theory.
3) Have developed a new test of discordance based on gaps between observations
4) Have investigated the performance of new test.
5) Have illustrated a practical example based on an eye data set.

THANK YOU


MyStats 2016
Malaysia Statistics Conference
15 November 2016
Sasana Kijang, Bank Negara Malaysia
2016
Strengthening Statistical Usage for Decisions and Innovation