DATA ANALYTICS AND APPLIED MATHEMATICS
E-ISSN: 2773-4854
VOLUME 1, ISSUE 1, 2020, 31 – 36
DOI: https://doi.org/10.15282/daam.v1i1.5085

RESEARCH ARTICLE

Descriptive analysis of circular data with outliers using Python programming language

N.S. Zulkipli*, S.Z. Satari and W.N.S. Wan Yusoff

Centre for Mathematical Sciences, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Kuantan, Pahang, Malaysia

ABSTRACT - Descriptive statistics are commonly used in data analysis to describe the basic features of raw data. Descriptive summaries enable us to present the data in a more simple and meaningful way so that the interpretation will be easier to understand. The descriptive analysis of circular data with outliers is discussed in this study. Circular data is different from linear data in many aspects such as statistical modeling, descriptive statistics and etc. Hence, unlike linear data, the availability of statistical software specialises in analysing circular data is very limited. Python is a programming language which frequently used by data analysts nowadays. However, the package for circular statistics is not fully developed and it is not ready to use like in Splus or R programming language. In this study, the descriptive analysis of circular data is performed using the in-demand programming language, Python. Descriptive statistics of the circular data especially with the existence of outliers are discussed and the proposed Python code is available to use.

ARTICLE HISTORY
Received : 19/08/2020
Revised : 17/09/2020
Accepted : 27/11/2020
Published : 31/12/2020

KEYWORDS
Circular data
Descriptive analysis
Python
Programming language
Outlier

1. INTRODUCTION

Data analysis can assist people to make decisions and predictions. Data can change the world but bear in mind to be careful in choosing the appropriate techniques which suitable with the nature of data. Every day, data such as daily expenses, sales profit and driving distance are recorded and these kinds of data are known as linear data. Aside from linear data, there is another data type that has a direction which refers to circular data and spherical data, respectively [1]. Since early 1970, many researchers in applied sciences such as [1], [2], [3] and [4] are interested to explore circular data. This type of data commonly found in the area of meteorology and biology where researchers are interested to investigate the direction of wind and animals. Recently, [5] has reviewed circular biological data and revealed that there is vast opportunity to be explored in a biological field especially on the biomedical data that involve circular or angular values. According to [1], circular data have many unique and novel characteristics both in terms of modeling and their statistical procedures. Until today, researchers still working hard to develop statistical procedures that specialised for circular data. Not only the statistical procedures are not fully discovered, the user-friendly statistical software specialised for circular data is also limited [6]. It is indispensable to introduce user-friendly circular statistical software or package like well-known ORIANA, Splus and R programming language.

Two major procedures of statistical analysis involved when analysing the data which is descriptive statistics and inferential statistics. [7] said that it is very useful to summarise the circular data by appropriate descriptive statistics and cannot simply apply the conventional summary statistics on the line. To see this, consider there are two observations on the circle which consist of 1° and 359°. If the circular approach is applied, it would give the sample mean as 0°. Unfortunately, if the linear approach is applied, the sample mean will get 180° which is not correct. Hence, due to the dissimilar features between a circle and a straight line, the descriptive analysis for linear data cannot be simply applied for the circular data. Moreover, the issue regarding outliers in circular data is contrasting from that in the linear case [1]. Some observations in a dataset which are not consistent with the others are defined as outliers. According to [8], between 5% and 10% of any statistical dataset are outliers or also known as surprising points.

In the current era, there are many high-level programming languages available to use. Most of them such as Python, R, C++, Java, JavaScript and many more are open source programming languages. Python and R are very familiar among data analysts and statisticians. R programming language is frequently used by the researchers to analyse circular data since the package for circular statistics which also known as ‘circular’ and ‘CircStats’ has been released in 2017 and 2018, respectively. [9] stated that, the usage of Python language has been soaring since the early 2000s in industrial applications and research, while R still a popular language for traditional data analytical procedures. Recently, Python rank to be the top followed by Java, C, C++, JavaScript and R [10]. Python has rapidly developed huge libraries for data science such as Numpy, Pandas, Scipy, Matplotlib, StatsModel and Seaborn. These Python libraries extremely popular among data analyst and statisticians. Unfortunately, the library that specialised for circular statistics is not fully developed such as ‘pycircstat’ and ‘spicy.stats’. Until now, the ‘pycircstat’ package is under development due to some functions are not

***CORRESPONDING AUTHOR | S.Z. Satari** | ✉ zanariah@ump.edu.my

working. Meanwhile, the 'spicy.stats' package only provides functions of circular mean, circular variance and circular standard deviation. Here, we notice that the existing packages for circular data are limited to a few functions. Therefore, this study aim to generate the descriptive statistics analysis for circular data using Python language with available Python libraries. At the end of this study, the Python coding will be proposed and compared with other softwares. This study will be beneficial for those who are started exploring circular data and choose Python as their programming language.

2. METHODOLOGY

Analysing data always starts with descriptive statistics before proceeding with drawing conclusions about population based on a sample by applying appropriate inferential statistics. The descriptive analysis provides detailed summaries about sample data. It can be a summary statistic, for example, mean, median, standard deviation and also can be in graphical forms such as graphs and plots. From the descriptive analysis, the distribution of data can be identified and inferential analysis can be conducted according to the distribution of the data. In this study, we focus on proposing coding for descriptive statistics for circular data using Python programming language. Hence, the descriptive statistics used in this study are described.

Let a sample of n angles $\theta_1, \theta_2,\ldots, \theta_n$ be a sample of circular data. It is assumed that these angles are in degrees unit. The mean direction can be defined as the angle made by the resultant vector with the horizontal line. The mean direction, $\bar{\theta}$ is given by [1] as in (1).

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

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

A sample median direction, $\tilde{\theta}$ of n angles $\theta_1, \theta_2,\ldots, \theta_n$ can be defined as any point $\emptyset$, where half of the data lie in the arc $[\emptyset, \emptyset + \pi)$ and the other points are nearer to $\emptyset$ than to $\emptyset + \pi$ [7]. [2] mention that, an algorithm for linear data can be used if the data are concentrated on an arc of the circle substantially less than the complete circle. If the number of data is odd, $\tilde{\theta}$ is one of the data points and if the number of data is even, $\tilde{\theta}$ is the midpoint of two appropriate adjacent data points. [2] defined the median directions for any circular sample as the observation $\emptyset$ which minimises the summation of circular distances to all observations as in (2).

$$d(\emptyset) = \pi - \sum_{i=1}^{n} \left|\pi - |\theta_i - \emptyset|\right| \ for\ i = 1,2,\ldots,n. \tag{2}$$

The resultant length of the vector resultant, R lies in the range $(0,n)$ given by

$$R = \sqrt{C^2 + S^2}. \tag{3}$$

The mean resultant length $\bar{R}$ associated with the mean direction $\bar{\theta}$ is defined by

$$\bar{R} = R/n. \tag{4}$$

The mean resultant length is also called the measure of concentration of dataset. It describes how concentrated the data is towards the center [1, 3]. If the value of $\bar{R}$ is approaching 1, the dispersion of the dataset will be approaching 0. The sample circular variance, V, measures the variation in the angles about the mean direction. V lies in the range of [0,1] and the formula for V is given as

$$V = 1 - \bar{R}, \qquad 0 \leq V \leq 1. \tag{5}$$

The smaller the value of circular variance indicate that the more concentrated the data sample. The circular standard deviation, v is defined by (6)

$$v = \sqrt{-2\ ln\ (1 - V)} \tag{6}$$

and it can be simplified as

$$v = \sqrt{-2\ ln\ \bar{R}}, \qquad 0 < v < \infty. \tag{7}$$

The circular dispersion is used to find the confidence interval and it is defined by [1] as given by

$$\hat{\delta} = \frac{1-\bar{R}_2}{2\bar{R}^2}. \tag{8}$$

$\bar{R}_2$ is the mean resultant length of double angles $2\theta_1, \ldots, 2\theta_n$ given by

$$\bar{R}_2 = 1/n\sqrt{C_2^2 + S_2^2} \tag{9}$$

where; $C_2 = \cos 2\theta = (\cos\theta)^2 - (\sin\theta)^2$, $S_2 = \sin 2\theta = 2sin\theta cos\theta$. Lastly, the concentration parameter, κ can be approximated by [4] and given as in (10). It is also available in [2] and [7].

$$\hat{\kappa} = \begin{cases} 2\bar{R} + \bar{R}^3 + \dfrac{5\bar{R}^5}{6} & \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + \dfrac{0.43}{1-\bar{R}} & 0.53 \leq \bar{R} < 0.85 \\ \dfrac{1}{3\bar{R} - 4\bar{R}^2 + \bar{R}^3} & \bar{R} \geq 0.85 \end{cases} \tag{10}$$

The range of the concentration parameter, κ is between 0 and ∞. Thus, the higher the value of the concentration parameter, the less the dispersion and the circular sample will be concentrated towards the mean direction, $\bar{\theta}$.

There are a bunch of libraries in Python available for data analysis. For descriptive circular statistics, the libraries that will be used in this study are Numpy, Math, Statistics, Scipy and Pandas. The documentation of these libraries can be referred from [11]. Python version 3.8 is used in this study.

3. DATA APPLICATION

In this study, historical data is used to illustrate the application of descriptive analysis for circular data. The data involves is the direction of northern cricket frogs, *Acris Crepitans* data defined by [12]. A data of homing ability of the 14 northern cricket frogs, *Acris Crepitans* was recorded in a series experiment by [12]. The frogs were collected from the mudflats of abandoned stream meander near Indianola, Mississippi. After 30 hours of enclosure within a dark environmental chamber, the frogs were released and the directions taken by the frogs were recorded. The data are given in Table 1.

Table 1. A dataset of 14 directions of northern cricket frogs, *Acris Crepitans*

Observation, θ_i	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8	θ_9	θ_{10}	θ_{11}	θ_{12}	θ_{13}	θ_{14}
Frog's direction (in degrees)	104	110	117	121	127	130	136	145	152	178	184	192	200	316

The frog direction data are visualised in circular plot by using ORIANA as in Figure 1. It can be seen there is an observation that deviates far away from the others and may affect the frog data distribution. The other 13 observations are found to be concentrated towards the center of the dataset. Based on previous literature, observation 14 is already detected as an outlier by many researchers such as in [13-15]. This observation can affect the dispersion of data [1], [2] and [7], and it is located at the fourth quadrant which is 5.5152 radians or 316°. Hence, this can be interesting information to study further and to see the effect of descriptive analysis when observation 14 is excluded. Descriptive analysis for the frogs data is analysed using proposed Python coding and it will be discussed in the following section.

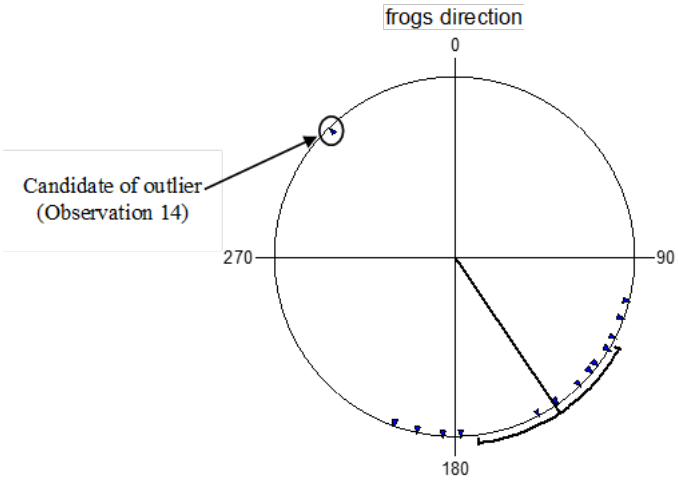

Figure 1. Circular plot of 14 directions of northern cricket frogs, *Acris Crepitans* with a candidate of outlier

4. PROPOSED PYTHON CODING

In this section, coding for descriptive analysis is proposed using Python programming language. The proposed coding and libraries are summarised in Table 2 with the corresponding formula for each circular descriptive statistic. All listed libraries must be imported into Python before the code is run.

Table 2. Proposed python coding for descriptive analysis of circular data

Descriptive Statistics	Equation	Python Coding	Library

Mean Direction	(1)	`C=round(sum(numpy.cos(fdirect_rad)),4)`
		`s=round(sum(numpy.sin(fdirect_rad)), 4)`
		
		`if c>0 and s>=0:`
		`  theta_bar=round(numpy.arctan(s/c),4)`
		`if c==0 and s>0:`
		`  theta_bar=round(numpy.pi/2,4)`
		`if c<0:`
		`  theta_bar=round(numpy.arctan(s/c) + np.pi, 4)`
		`if c>=0 and s<0:`
		`  theta_bar=round(numpy.arctan(s/c)+(2*np.pi), 4)`
		`if c==0 and s==0:`
		`  print("undefined")`
		
		`theta_bardeg=round(numpy.rad2deg(theta_bar), 4)`
			Pandas
			Numpy

Median Direction	(2)	`theta_tilda=round(statistics.median(fdirect_rad), 4)`	Scipy Statistics
Mean Resultant Length	(3), (4)	`R=math.sqrt(c**2+s**2)` `Rbar=round(R/n, 4)`	Math Numpy
Circular Variance	(5)	`V=round(1-Rbar,4)`	Numpy
Circular Standard Deviation	(6)	`v=round(math.sqrt(-2*numpy.log(1-V)),4)`	Math Numpy

Sample Circular Dispersion	(8), (9)	`c2=round(sum(((numpy.cos(fdirect_rad)) **2)-((numpy.sin(fdirect_rad))**2)),4)`
		`s2=round(sum(2*numpy.sin(fdirect_rad) *numpy.cos(fdirect_rad)), 4)`
		`R2=round(math.sqrt(c2**2+s2**2),4)`
		`delta_hat=round((1-R2bar)/(2*Rbar**2), 4)`
			Math
			Numpy

Estimated Concentration Parameter	(10)	`if rbar<0.53:`
		`  estkappa=(2*rbar)+(rbar**3)+((5*rbar**5)/6)`
		`if rbar>=0.53 and rbar<0.85:`
		`  estkappa= (1.39*rbar)+(0.43/(1-rbar)-0.4)`
		`if rbar>=0.85:`
		`  estkappa=1/(3*rbar4*rbar**2+rbar**3)`
			Numpy

4. RESULTS AND DISCUSSION

This study uses historical data to illustrate the descriptive analysis of circular data using Python programming language and will be compared with results from ORIANA software and R programming language. The results from the manual calculation are also included in this study to make sure the proposed coding from Python is validated. In the previous Data Application section, the dataset has been presented in graphical form using circular plot generated from ORIANA. Figure 1 shows that, observation 14 is not consistent with the others where this observation is located far away from the other 13 observations. Thus, it is interesting to see what is the effect on descriptive analysis if the suspected observation 14 is excluded from the dataset. The descriptive analysis for two different dataset are summarised and presented in Table 3 by using manual calculation, ORIANA software, R and proposed Python programming language. Let define the dataset that include observation 14 as Dataset 1 and dataset that exclude observation 14 as Dataset 2.

Table 3. Summary of descriptive analysis for Dataset 1 and Dataset 2 of northern cricket frogs by manual calculation, ORIANA, R and Python

Descriptive Statistics	Manual calculation	ORIANA	R	Python
Mean direction				
Dataset 1	145.9725°	145.9740°	145.9744°	145.9725°
Dataset 2	145.0843°	145.0830°	145.0843°	145.0843°
Median direction				
Dataset 1	140.5000°	****	140.5000°	140.5000°
Dataset 2	136.0000°	136.0000°	136.0000°	136.0000°
Mean resultant length				
Dataset 1	0.7252	0.7250	0.7252	0.7252
Dataset 2	0.8568	0.8570	0.8568	0.8568
Sample circular variance				
Dataset 1	0.2748	0.2750	0.2748	0.2748
Dataset 2	0.1432	0.1430	0.1432	0.1432
Sample standard deviation				
Dataset 1	45.9283°	45.9310°	45.9283°	45.9283°
Dataset 2	31.8564°	31.4890°	31.8564°	31.8564°
Circular dispersion				
Dataset 1	0.4383	****	0.4383	0.4383
Dataset 2	0.3373	****	0.3373	0.3373
Concentration parameter				
Dataset 1	2.1728	2.1730	2.1728	2.1728
Dataset 2	3.8029	3.8030	3.8029	3.8029

****Indicates that a result could not be generated by the software*

Generally, the values of descriptive statistics calculated by manual calculation, ORIANA software, R and Python programming language are not vastly different. We can see that ORIANA generated the value of descriptive statistics with three decimal places and others are standardised by taking four decimal places. It is worth to note that there is a limitation in ORIANA since some of the value of descriptive statistics is not given. Meanwhile, most of the calculated value from R is the same as in the proposed Python coding. The descriptive analysis is discussed further by looking at the generated value for Dataset 1 and Dataset 2. Based on Table 3, for Dataset 1, the mean and median frog's directions are 145.9725° and 140.0843°, respectively. This explains that the frogs are going to the direction of 145.9725° on average. The mean and median directions from Dataset 2 are 145.0843° and 136°, respectively. It can be seen the median of the frog's direction is moving to 136° from 140.0843° when observation 14 is removed. Meanwhile, the mean direction of Dataset 2 is moving very slightly to 145.0843°.

While the mean resultant length measured the concentration of the dataset. Mean resultant length Dataset 1 and Dataset 2 are 0.7252 and 0.8568, respectively. The concentration of data is increase when observation 14 is excluded. This indicates that all observations in Dataset 2 are more concentrated towards the center compared to observations in Dataset 1. The circular variance, standard deviation and circular dispersion of Dataset 1 are 0.2748, 45.9283° and 0.4383, respectively. By comparing with Dataset 1, the circular variance, standard deviation and circular dispersion of Dataset 2 are decrease to 0.1432, 31.8564° and 0.3373, respectively. Hence, we found that by excluding the observation 14, the dispersion of data is decreased which indicates that the observations are becoming more consistent.

The estimated concentration parameter, $\hat{\kappa}$ is 2.1728 and 3.8029 for Dataset 1 and Dataset 2. Since the $\hat{\kappa}$ value for Dataset 2 gets larger, it indicates the dispersion of data is less and it will be more concentrated towards the mean direction. Thus, it shows that observation 14 interesting to be investigated further since it affects the concentration of data.

5. CONCLUSIONS

In summary, it is found that Python programming is applicable to use for calculating the descriptive statistics for circular data. It is easy to write coding for descriptive circular statistics and generate the results since the library specialised in statistical computing is available to use. The descriptive analysis for circular data can be calculated using the proposed coding and it is beneficial for those who are started exploring circular data and decided to use Python as their programming language.

ACKNOWLEDGEMENTS

Institution(s)

All the authors would like to express their gratitude to Universiti Malaysia Pahang for the support/ facilities.

Fund

The authors would like to thank the Ministry of Higher Education for providing financial support under Fundamental Research Grant Scheme (FRGS) No. FRGS/1/2019/STG06/UMP/02/6) and Universiti Malaysia Pahang as additional financial support under Internal Research Grant RDU190363.

Individual Assistant

Authors would like to thank all the associate editors and referees for their thorough reading and valuable suggestions which led to the improvement of this paper.

DECLARATION OF ORIGINALITY

The authors declare no conflict of interest to report regarding this study conducted.

REFERENCES

[1] Hasan Jammalamadaka SR, Sengupta A. Topics in circular statistics. Singapore: World Scientific Publishing; 2001.

[2] Fisher NI, Lewis T, Embleton BJ. Statistical analysis of spherical data. New York: Cambridge university press; 1993 Aug 19.

[3] Mardia KV. Statistics of directional data. Journal of the Royal Statistical Society Series B: Statistical Methodology. 1975 Jul;37(3):349-71.

[4] Best DJ, Fisher NI. The BIAS of the maximum likelihood estimators of the von Mises-Fisher concentration parameters: the BIAS of the maximum likelihood estimators. Communications in Statistics-Simulation and Computation. 1981 Jan 1;10(5):493-502.

[5] Satari SZ, Khalif KM. Review on outliers identification methods for univariate circular biological data. Advances in Science, Technology and Engineering Systems. 2020;5(2):95-103.

[6] Hassan SF, Hussin AG, Zubairi YZ. Analysis of Malaysian wind direction data using ORIANA. Modern Applied Science. 2009 Mar;3(3):115-9.

[7] Mardia KV, Jupp PE. Directional statistics. London: John Wiley & Sons; 2009 Sep 25.

[8] Rousseeuw PJ, Hampel FR, Ronchetti EM, Stahel WA. Robust statistics: the approach based on influence functions. New York: Wiley; 1986.

[9] Chen CP, Zhang CY. Data-intensive applications, challenges, techniques and technologies: A Survey on Big Data. Information Sciences. 2014 Aug 10;275:314-47.

[10] Cass S. The top programming languages: Our latest rankings put Python on top-again-[Careers]. IEEE Spectrum. 2020 Jul 28;57(8):22-22.

[11] Python. The Python Standard Library [Online]. Retrieved from https://docs.python.org/3/library/index.html: 18 August 2020.

[12] Ferguson DE, Landreth HF, Mckeown JP. Sun compass orientation of the northern cricket frog, Acris crepitans. Animal Behaviour. 1967 Jan 1;15(1):45-53.

[13] Collett D. Outliers in circular data. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1980 Mar;29(1):50-7.

[14] Abuzaid AH, Hussin AG, Rambli A, Mohamed I. Statistics for a new test of discordance in circular data. Communications in Statistics-Simulation and Computation. 2012 Nov 1;41(10):1882-90.

[15] Badarisam F, Rambli A, Sidik M. A comparison on two discordancy tests to detect outlier in von mises (VM) sample. Indonesian Journal of Electrical Engineering and Computer Science. 2020 Jul;19(1):156.