Vol.12.Issue.3.2024 (July-Sept)

http://www.bomsr.com

Email:editorbomsr@gmail.com

RESEARCH ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

A NEW APPROACH FOR THE CALCULATION OF CLASS WIDTH FOR A FREQUENCY DISTRIBUTION

RAMNATH TAKIAR

Flat N0.11, 3rd Floor, Building # 9, 1st Khoroo, Ulaanbaatar district, Ulaanbaatar, Mongolia- 14241 &
Scientist G – (Retired)
National Centre for Disease Informatics and Research (NCDIR), Indian Council of Medical Research (1978-2013) Bangalore – 562110, Karnataka, India
Email: ramnathtakiar@gmail.com, ramnath_takiar@yahoo.co.in
DOI:10.33329/bomsr.12.3.9

Ramnath Takiar

ABSTRACT

To draw any meaningful inferences from the raw data, it is essential that it is summarized in a meaningful manner. One of the popular statistical tools to deal with the management of the huge data is to present it in the form of a frequency distribution. In a frequency distribution, often, the size of the class is arbitrarily decided. In literature, a few popular rules which are available to determine the Optimal number of class intervals are the Sturges rule, Square root rule or the Rice rule. In all these rules the Optimal number of classes are shown be a function of the sample size N, alone. However, in the present study, an attempt is made to develop an approach to decide the Optimal number of classes to form a frequency distribution being a function of the Standard deviation, the Range, and the sample size n.

For the study purposes, the normal populations of six sizes are considered namely the size of 50, 100, 150, 200, 500 and 750. For each population size, ten populations are generated using the function available on Excel. For each sample size and ten populations, five frequency distributions are generated using the number of classes as K, K+1, K+2, K+3 and K+4 where K = 1.72 + 0.73*ln(N). In general, (K+3)

is adjudged as the Optimal number of classes which yields the least % sum of Absolute Errors.

On comparison of the sum of % Absolute Errors, arising due to four methods namely, Sturges Rule, Rice Rule, Square root Rule and the Takiar Rule, the Takiar Rule is adjudged as the best Rule to decide the Optimum number of Class intervals. According to Takiar Rule, the formula for the calculation of Optimal number of classes is given by n = 4.72 + 0.73*ln(N).

Keywords: Frequency distribution, Optimal number of classes, Sturges Rule, Square root Rule, Rice Rule, Takiar Rule.

INTRODUCTION

In research studies, laboratory investigations, or routine surveys, we may come across a huge set of data. Depending on the objectives of the study, the data may be related to individuals, households, or industries. Further, the data collected may be related to ages, occupations, monthly incomes, or educational status of the subjects covered under the study. Such a data is termed as the raw data. To draw any meaningful inferences from the raw data, it is essential that it is well organized and tabulated in a meaningful way. One of the popular statistical tools to deal with the huge data is to present it in the form of a frequency distribution. In a frequency distribution, the data is grouped in various classes with number of observations belonging to each class, popularly known as the class-frequency. Generally, all the class intervals are chosen to be of the same size. The size of the class is known as the width of the class interval. Thus, in forming a frequency distribution, the width of the class interval plays a significant role. In most cases, arbitrarily decided number of class intervals form the basis of arriving at the width of the class interval by dividing the range by the desired number of classes. In literature, the formula due to Sturges is available for the determination of the number of class intervals (Gupta SC 2012, Keller G, Warrack B 2003, Statistic How to 2024, Statology 2021). Two other rules are also available in literature to determine the Optimal number of class intervals like the Square root rule and Rice rule (Statology 2021, Brian EB 2016). A careful observation of these rules will suggest that the Optimal number of classes is necessarily a function of the sample size N, alone. However, a recent study (Takiar R 2023) has shown that the Standard deviation (SDP) and the Range in a data set are related with the sample size, characterized by the equation: K = Range/SDP = 0.73*ln(N) + 1.72 where SDP $= \frac{1}{N}\boldsymbol{\Sigma}(\boldsymbol{x_i} - \bar{\boldsymbol{x}})^{\boldsymbol{2}}$**.** This finding led to the hypothesis that the Optimal number of classes should be a function of the SDP, Range, and the sample size N.

OBJECTIVE

To develop an approach to decide the Optimal number of classes to form a frequency distribution being a function of the SDP, the Range, and the sample size n.

MATERIALS AND METHODS

SOURCES OF DATA

To form a frequency distribution, it is essential that the raw data is available. For the study purposes, the required sets of data, with the predefined means and the standard

deviations, are generated using the Excel function key "Random Number Generation" and used as the source of the data.

GENERATION OF NORMAL POPULATION

For the study purposes, the normal populations of six sizes are considered namely the size of 50, 100, 150, 200, 500 and 750. For each population size, ten populations are generated as shown in Table 1.

Table 1: The Generation of Normal Populations with Predefined Mean and the Standard Deviation

Population size	50	100	150	200	500	750
No. of Populations	10	10	10	10	10	10

In selection of the normal populations, the efforts are made to choose populations with $\beta 1$and $\beta 2$, close to 0 and 3, respectively.

DETERMINATION OF NUMBER OF CLASSES TO BE CONSTRUCTED FOR A GIVEN SET OF RAW DATA

For each raw set of data, five sets of frequency distributions are formed with the number of classes as "n" equal to K, K+1, K+2, K+3 and K+4. To arrive at the value of "K" , the following formula due to Takiar (Takiar R 2023) is used: K = Range/SDP = 0.73*ln(N) + 1.72. It is also evident that each frequency distribution will give different estimates of the mean, SDP and Kurtosis. A frequency distribution whose estimates are closer to the parameter values of the raw data is adjudged as the best frequency distribution and the value of "n" corresponding to it is considered as the Optimal number of classes to be attempted.

CONSTRUCTION OF FREQUENCY DISTRIBUTIONS

For the formation of a frequency distribution, it is essential to know the number of class intervals and the class width. When the number of class intervals are chosen to be equal to the "n," the Class width is calculated using the formula: Class width = W = Range/n. In case of W being a fraction, the value of W is determined as W = INT(W)+1.

The lowest class interval is taken as ≤ L, where L is the integer part of the minimum value in the raw data. The subsequent class intervals are taken as L to (L+W), (L+W) to (L+2W)…… and (L+(K-1)*W) to (L+K*W). The above procedure is repeated when number of class intervals are taken to be as (K+1), (K+2), (K+3) and (K+4). Thus, systematically, for each raw normal set of data, five frequency distributions are formed.

ERROR CALCULATIONS AND COMPARISON OF THE FREQUENCY DISTRIBUTIONS

For each frequency distribution, the following three parameters are calculated using the available Excel keys: Mean, SD, and Kurtosis. The values of these parameters are expected to be different from those obtained from the raw data. The absolute difference between (Mean-mean) is treated as the Absolute Error in the estimation of the Mean where Mean is derived from the raw data and mean from the grouped data. To derive the % Absolute Error,

the formula used is $\frac{|Mean-mean|*100}{Mean}$. Similarly, the Absolute Errors in the estimation of SDP and Kurtosis are calculated. Thus, the sum of the % Absolute Errors due to Mean, SDP and Kurtosis is calculated for each frequency distribution and compared. The number of Classes corresponding to the frequency distribution with the least sum of % Absolute Errors is taken as the Optimal number of Classes for the associated sample size.

RULES AVAILABLE IN LITERATURE TO DETERMINE THE OPTIMAL NUMBER OF CLASS INTERVALS

There are three popular rules in the literature which are available to determine the Optimal number of class intervals to form a frequency distribution.

- **Sturges Rule :** Optimal number of classes = n = $|\log_2 N + 1|$ = 1 + 3.3*Log(N)
- **Square root Rule:** Optimal number of Classes **=** $n = \sqrt{N}$; and
- **Rice Rule:** Optimal number of Classes **=** $n = 2*\sqrt[3]{N}$

  Where N is the number of observations in the data set.

COMPARISON OF RULES TO DETERMINE THE OPTIMAL NUMBER OF CLASS INTERVALS

The Takiar rule developed in the current study to determine the Optimal number of classes are compared with the class intervals arrived using the above three rules. It is apparent that each rule will lead to different estimates of the parameters and thereby different quantum of errors. The % Absolute sum of Errors by all the four rules are compared. The rule with the minimum % Absolute Sum of Errors is adjudged as the best rule to determine the Optimal number of class intervals.

RESULTS

The details of the parameters of selected ten normal populations by the selected sample sizes of 50, 100, 150, 200, 500 and 750 are provided in Table 2 to Table 4.

Table 2: The details of Parameters - Samples of size 50 and 100

Population	Sample size = 50				Sample size =100			
	MEAN	SDP	SKEW	KURT	MEAN	SDP	SKEW	KURT
P1	56.88	16.585	0.043	2.876	57.49	15.186	0.035	2.977
P2	44.78	10.596	0.039	2.949	47.03	14.306	0.039	2.905
P3	62.51	16.667	0.020	2.817	68.9	16.434	0.008	3.014
P4	78.34	16.082	0.107	2.844	79.88	23.691	0.082	2.941
P5	55.56	10.340	0.028	2.866	49.34	15.344	0.056	2.872
P6	66.97	20.847	0.084	2.834	76.42	20.844	0.009	2.986
P7	35.72	8.271	0.090	2.819	31.71	10.006	0.041	2.895
P8	38.18	7.285	0.023	2.923	37.68	9.616	0.04	2.986
P9	41.72	10.844	0.022	2.888	88.32	23.467	0.015	2.935
P10	43.29	15.192	0.031	2.863	121.65	33.336	0.027	2.923

Table 3: The details of Parameters - Samples of size 150 and 200

SAMPLE	Sample size = 150				Sample size =200			
	MEAN	SDP	SKEW	KURT	MEAN	SDP	SKEW	KURT
S1	55.73	17.34	0.020	2.897	55.50	16.013	0.020	2.898
S2	45.61	12.199	0.029	2.918	44.21	11.697	0.107	2.999
S3	66.34	18.542	0.049	2.955	65.77	16.786	0.057	2.970
S4	108.52	27.042	0.072	3.000	76.14	17.946	0.002	3.023
S5	120.65	34.643	0.003	3.013	51.92	12.861	0.041	2.952
S6	127.56	36.884	0.049	3.018	70.64	17.161	0.011	2.933
S7	138.23	41.783	0.010	3.040	36.24	8.368	0.013	2.972
S8	150.4	46.614	0.005	2.884	38.18	9.054	0.025	2.950
S9	166.98	51.878	0.065	3.047	40.80	11.674	0.014	2.906
S10	171.57	62.989	0.062	2.868	43.77	12.545	0.040	2.985

Table 4: The details of Parameters - Samples of size 500 and 750

SAMPLE	Sample size = 500				Sample size =750			
	MEAN	SDP	SKEW	KURT	MEAN	SDP	SKEW	KURT
S1	200.59	61.253	0.083	3.011	299.92	78.719	0.038	3.036
S2	210.53	63.046	0.025	2.972	317.58	90.273	0.023	2.902
S3	222.85	66.876	0.020	3.005	334.75	91.99	0.02	3.061
S4	230.19	68.417	0.028	2.939	357.72	110.192	0.064	2.968
S5	237.74	72.543	0.016	2.933	368.17	119.297	0.042	2.998
S6	249.32	67.252	0.044	2.94	378.95	120.379	0.024	2.979
S7	256.83	78.384	0.006	2.945	392.35	121.566	0.014	2.889
S8	270.21	80.136	0.035	3.041	394.84	118.084	0.014	3.03
S9	278.99	89.94	0.042	2.774	418.19	135.233	0.007	2.958
S10	297.37	87.972	0.033	2.932	436.78	143.482	0.043	2.962

The Examples of Frequency Distributions, arrived based on the selected five number of Class Intervals for the Sample size of 200, is shown in Table 5.

Table 5: Frequency Distributions arrived based on the Selected Number of Five Class Intervals for the Samples of size of 200

Number of Class Intervals									
K		K+1		K+2		K+3		K+4	
Class Interval	Freq	Class Interval	Freq	Class Interval	Freq	Class Interval	Freq	Class Interval	Freq
< 16	0	< 16	0	< 16	0	< 16	0	< 16	0
16-28	11	16-27	9	16-26	7	16-25	6	16-24	6
28-40	22	27-38	19	26-36	17	25-34	13	24-32	11
40-52	49	38-49	37	36-46	29	34-43	21	32-40	16
52-64	60	49-60	62	46-56	49	43-52	42	40-48	29
64-76	35	60-71	39	56-66	47	52-61	53	48-56	40
76-88	17	71-82	24	66-76	28	61-70	29	56-64	40
88-100	6	82-93	8	76-86	16	70-79	20	64-72	29
100-112	-	93-104	2	86-96	6	79-88	10	72-80	15
112-124	-	104-115	-	96-106	1	88-97	5	80-88	8
124-136	-	115-126	-	106-116	-	97-106	1	88-96	5
136+	-	126+	-	116+	-	106+	-	96-104	1
Total	200	Total	200	Total	200	Total	200	Total	200

The example of the calculations of the Sum of % Absolute Error in Estimation of the Mean, SDP, and Kurtosis, occurring due to Selected five number of Classes, for the Sample size of 200, are shown in Table 6. For the Mean, the % absolute error ranged from 0.04% to 0.54% while for the SDP, it ranged from 0.51% to 3.96%. Similarly, for the Kurtosis, the % absolute error ranged from 1.06% to 6.73%. For Mean and SDP, the % absolute error remained below 5% while for the Kurtosis, it remained below 7%.

Table 6: Calculation of % Absolute Errors occurring due to Grouping of data into different Frequency Distributions - Sample size - 200

Description		MEAN	SDP	KURT	TOTAL
POPULATION		55.50	16.023	2.926	-
No. of Classes	K	55.66	16.658	2.729	-
Error	%	0.29	3.96	6.73	10.98
No. of Classes	K+1	55.44	16.158	2.817	-
Error	%	0.11	0.84	3.73	4.68

No. of Classes	K+2	55.80	16.400	2.729	-
Error	%	0.54	2.35	6.73	9.62
No. of Classes	K+3	55.47	16.104	2.957	-
Error	%	0.05	0.51	1.06	1.62
No. of Classes	K+4	55.52	16.290	2.847	-
Error	%	0.04	1.67	2.70	4.41

The Sum of % Absolute Errors in the estimation of Mean, SDP, and the Kurtosis, occurring due to Selected five number of Classes, for the Sample size of 50, 100, 150, 200, 500 and 750 are provided in Table 7. For each sample size, the number of classes with the least sum of % Absolute Errors is identified. The number of classes identified for the sample size of 50, 100 and 150 are observed to be (K+4), (K+3) and (K+3) and the corresponding % sum of Absolute Errors are observed to be 83.6, 53.15 and 39.6.

Table 7: Sum of % Errors Over 10 Samples in the Estimation of Mean, SDP and Kurtosis, Occurring due to Selected number of Classes and varying Sample size

Sample Size	NO. OF CLASSES	MEAN	SD	KURT	TOTAL
50	K	7.99	17.24	91.67	116.9
	K+1	6.71	18.35	99.31	124.37
	K+2	8.06	19.79	61.38	89.23
	K+3	6.04	24.31	61.45	91.8
	K+4	4.45	12.36	66.79	83.6
100	K	4.87	32.65	81.53	119.05
	K+1	8.22	29.87	58.79	96.88
	K+2	7.06	21.19	52.62	80.87
	K+3	8.26	5.72	39.17	53.15
	K+4	5.76	11.69	70.47	87.92
150	K	5.5	20.73	67.26	93.49
	K+1	4.3	15.17	37.23	56.7
	K+2	3.6	13.04	73.11	89.75
	K+3	1.73	8.88	28.99	39.6
	K+4	4.71	10.85	40.38	55.94
200	K	3.45	30.71	60.97	95.13

	K+1	4.29	15.17	49.33	68.79
	K+2	3.98	12.5	47.16	63.64
	K+3	3.54	11.17	40.34	55.05
	K+4	2.27	16.14	45.05	63.46
500	K	1.5	21.96	35.84	59.3
	K+1	2.73	14.52	50.33	67.58
	K+2	2.43	12.35	25.1	39.88
	K+3	1.05	11.29	22.98	35.32
	K+4	1.83	10.47	30.28	42.58
750	K	1.7	27.19	25.03	53.92
	K+1	1.1	19.89	14.54	35.53
	K+2	0.89	11.78	10.7	23.37
	K+3	0.75	15.1	8.72	24.57
	K+4	1.37	16.04	10.86	28.27

The number of classes for which the sum of % Absolute Errors is minimum is identified for the sample size of 200, 500 and 750 to be (K+3), (K+3) and (K+2), respectively. The corresponding sum of % Absolute Errors are observed to be 55.05, 35.32 and 23.37.

For the sample size of 100, 150, 200 and 400, the number of classes with the minimum Sum of % Absolute Error corresponded to (K+3) while in the case of sample size of 50 and 750, it corresponded to (K+4) and (K+2) number of classes, respectively.

The Sum of % Absolute Errors in the Estimation of Mean, SDP, and the Kurtosis, pooled over all the samples and sample sizes by the Number of Classes are shown in Table 8. The values in the last column, under the heading of Total, displays the Sum of % Absolute Errors for the selected parameters.

Table 8: The Sum of % Errors in the Estimation of Mean, SDP and the Kurtosis, - Pooled Over All the Samples and the Sample sizes by the Number of Classes

NO. OF CLASSES	MEAN	SDP	KURT	TOTAL
K	25.01	150.48	362.3	537.79
K+1	27.35	112.97	309.53	449.85
K+2	26.02	90.65	270.07	386.74
K+3	21.37	76.47	201.65	299.49
K+4	20.39	77.55	263.83	361.77

The Sum of % Absolute Errors, in the Estimation of Mean, SD and the Kurtosis By the Number of Classes, Pooled over all Sample sizes are shown in Fig. 1.

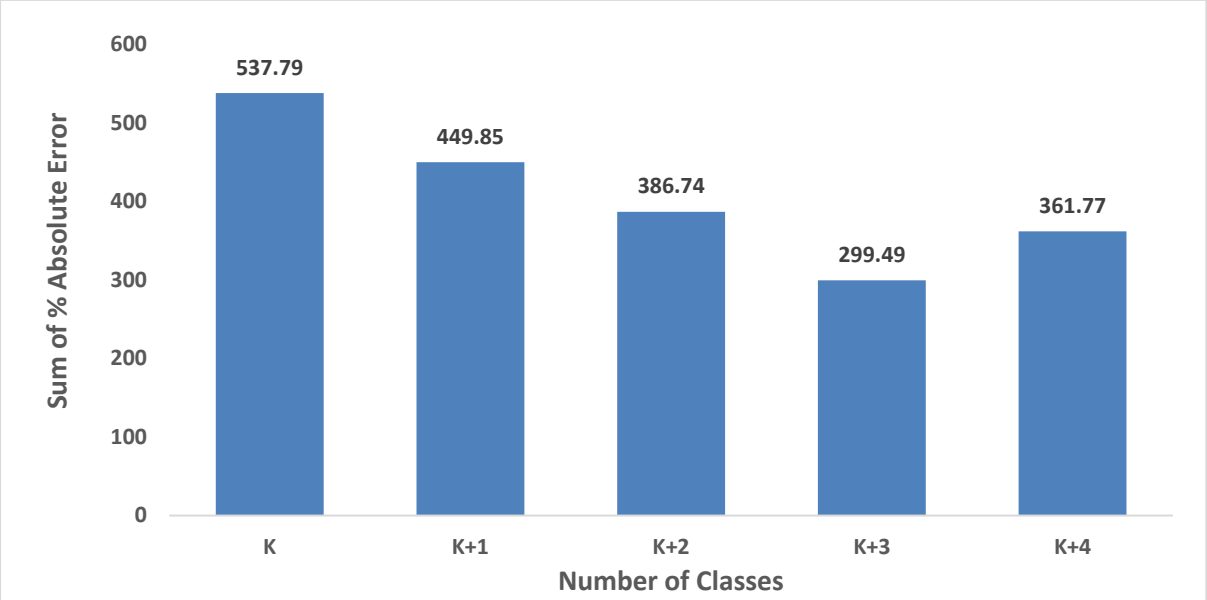

Fig. 1: Sum of % Absolute Errors in the Estimation of Mean, SD and the Kurtosis By the Number of Classes Pooled over all Sample sizes.

It is to be seen that the Sum of % Absolute Errors when pooled over all the samples and sample sizes, corresponded to (K+3) number of classes. Hence, for any given sample size, (K+3) should be taken as the number of classes to form a frequency distribution. It is evident that the selection of number of classes automatically decides the width of the Class Interval.

A comparison was attempted between different methods with respect to the Sum of % Absolute Errors and shown in Table 9.

Table 9: The Comparison of Different Methods Based on the Sum of % Absolute Errors by Selected Samples and Sample Sizes

SAMPLE SIZE	No. of Samples	TAKIAR	STURGUS	SQUARE ROOT	RICE
50	10	91.8	124.3	89.2	89.2
100	10	55.1	80.9	87.9	87.9
150	10	43.7	55.94	43.7	41.7
200	10	55.1	63.6	41.8	46.8
500	10	35.3	39.9	57.5	27.3
750	10	24.6	24.6	17.6	17.1
TOTAL	60	305.6	389.24	337.7	310.0

It is evident that the method suggested by the Takiar attains the minimum % Absolut Error. Thus, can be claimed as a better method, in comparison to prevalent methods, for

suggesting the Optimum number of Classes and thereby a proper width of the Class interval, to summarize raw data.

DISCUSSION

To draw any meaningful inferences from the raw data it is essential that the raw data is summarized and presented in a meaningful way. One of the popular statistical tools to deal with the huge raw data is to present it in the form of a frequency distribution. In a frequency distribution, the data is grouped into various classes with number of observations belonging to each class, popularly known as the class-frequency. In literature, there are few methods like Sturges method, Rice method and Square root method, available to suggest the Optimal number of Classes. In the current study, a new method is developed to arrive at the Optimal number of Classes using the relationship between the SDP and the range. For every raw data, five types of Class intervals are utilized. Based on each Class interval suggested, the Mean, SDP and Kurtosis are calculated and compared with the values arrived using the raw data. The number of classes for which the sum of % Absolute Error is minimum is identified. It is seen that in general, (K+3) number of classes yields the minimum sum of % Absolute Errors. It is to be noted that K is derived using the formula given as K = Range/SDP = 0.73*ln(N) + 1.72

To assess the efficacy of the current method, named as the "Takiar method" the sum of % Absolute Errors in estimation of the Mean, SDP, and Kurtosis of 60 samples of varying sample sizes, drawn from different normal populations, is employed. It is seen that the "Takiar method" yields the minimum Sum of % Absolute Errors as compared to other three methods. Thus, suggesting that the "Takiar method" is the best method in defining the Optimum number of Class intervals.

SUMMARY OF THE OBSERVATIONS

- In literature, there are few rules available to decide the Optimal number of Class intervals so that the raw data can be meaningfully tabulated and analyzed.
- The prevalent methods to define the Optimum number of Class intervals are Sturges method, Rice method and Square root method.
- In the current study, a new rule called the "Takiar Rule" is developed to define the Optimum number of Class intervals. This method utilizes the relationship between SDP and the Range to arrive at the Optimum number of Class intervals.
- A method which gives the least sum of % Absolute Errors in the estimation of Mean, SDP and Kurtosis is considered as the best.
- On comparison of the sum of % Absolute Errors, arising due to four methods namely, Sturges Rule, Rice Rule, Square root Rule and Takiar Rule, the Takiar Rule is adjudged as the best rule to decide the Optimum number of Class intervals.
- The Optimal number of classes (C) can be calculated using the Takiar formula: C = 4.72 + 0.73*ln(N)

RECOMMENDATIONS

Among the prevalent rules, the Takiar Rule, is shown to be the best in arriving at the Optimal number of Class intervals to summarize the raw data in the form of a frequency distribution.

REFERENCES

[1]. Blank B E 2016: Elementary Statistics, first president university press, Page 37 https://www.math.wustl.edu/~brian/stats/2200-02.pdf

[2]. Bobbitt Z 2021: Statology - What is Sturges' Rule? (Definition & Example) https://www.statology.org/sturges-rule/

[3]. Gupta SC 2012: Fundamental of Statistics, Seventh Edition, Himalaya Publishing House; Page 3.9-3.10 .

[4]. Keller G, Warrack B 2003: Statistics for Management and Economics, Thomson, Brooks/Cole; Page 35.

[5]. Microsoft Corporation, 2019. Microsoft Excel, Available at: https://office.microsoft.com/excel.

[6]. Statistics How to 2024: Class Width: Definition & Examples https://www.statisticshowto.com/probability-and-statistics/descriptive-statistics/frequency-distribution-table/class-width/

[7]. Takiar R 2023: The Relationship between the SD and the Range and a method for the Identification of the Outlier, *Bulletin of Mathematics and Statistics Research*, Vol. 11(4), Page 62-75

Biography

Dr. Ramnath Takiar

I am a Postgraduate in Statistics from Osmania University, Hyderabad. I did my Ph.D. from Jai Narain Vyas University of Jodhpur, Jodhpur, while in service, as an external candidate. I worked as a research scientist (Statistician) for Indian Council of Medical Research from 1978 to 2013 and retired from the service as Scientist G (Director Grade Scientist). I am quite experienced in large scale data handling, data analysis and report writing. I have sixty-six research publications in national and International Journals related to various fields like Nutrition, Occupational Health, Fertility and Cancer epidemiology. During the tenure of my service, I attended three international conferences, namely in Goiana (Brazil-2006), Sydney (Australia-2008), and Yokohama (Japan-2010), and I presented a paper on each. I also attended the Summer School related to Cancer Epidemiology (Modul I and Module II) conducted by International Agency for Research in Cancer (IARC), Lyon, France from 19th to 30th June 2007. After my retirement, I joined my son at Ulaanbaatar, Mongolia. I worked in Ulaanbaatar as a professor and consultant from 2013 to 2018, and I was responsible for teaching and guiding Ph.D. students. I also taught Mathematics to undergraduates and Econometrics to MBA students. During my service there, I also functioned as the Executive Editor for the in-house Journal "International Journal of Management." I am still active in research and have published thirteen research papers during 2021-24.