POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Mean and standard deviation for open-ended grouped data

Açık uçlu gruplanmış veriler için ortalama ve standart sapma

***Yazar(lar) (Author(s))**: Ayfer Ezgi YILMAZ[1], Serpil AKTAS ALTUNAY[2]*

ORCID[1]: 0000-0002-6214-8014

ORCID[2]: 0000-0003-3364-6388

Bu makaleye şu şekilde atıfta bulunabilirsiniz(To cite to this article): Yilmaz A. E. ve Aktas Altunay S., "Mean and standard deviation for open-ended grouped data", ***Politeknik Dergisi***, 25(4): 1603-1611, (2022).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

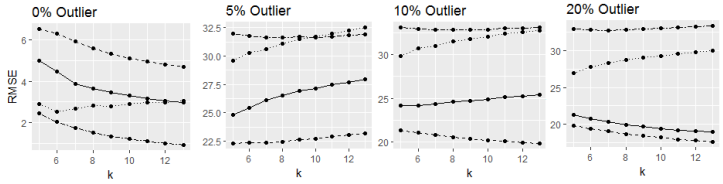
DOI: **10.2339/politeknik.836087**

Mean and Standard Deviation for Open-Ended Grouped Data

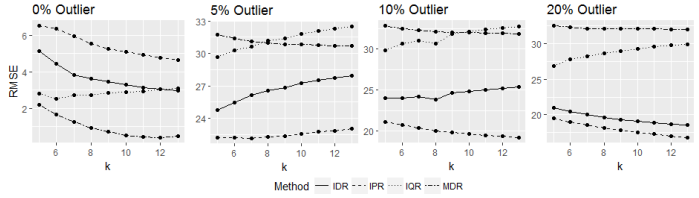
Highlights

- *Estimation of open-ended boundaries for grouped data*
- *Interdecile, interpercentile, and the mid-distance ranges for estimating the open-ended boundaries*
- *Calculating adjusted mean and standard deviation via estimated boundaries*
- *A suggested modification on MDR*

Graphical Abstract



(a) The RMSE results of LB_1



(b) The RMSE results of UB_k

Figure. The RMSE results of the unknown boundaries by percentage of outliers where $\sigma = 6$

Aim

It is suggested using IQR, IDR, IPR, and MDR to calculate the population mean and standard deviation in open-ended grouped data. This approach will be a kind of modification of the mid-distance range. It is aimed to compare these methods by their performance to estimate unknown category boundaries, to estimate mean and standard deviation.

Design & Methodology

A Monte Carlo simulation study is conducted to evaluate and compare four different methods to estimate the open-ended boundaries of frequency tables. The simulation space of the Monte Carlo study is composed of 216 different combinations.

Originality

This paper proposes four modified methods to estimate the population mean and standard deviation in open-ended group data.

Findings

The simulation results show that in case of no outliers, all methods perform well. When the number of categories increases, the value of RMSE and MAE of mean and standard deviation also increase; this case became more obvious for 20% percent outliers.

Conclusion

While the most appropriate measures of central tendency in open-ended data seem to be median, the proper application of mean among the proposed methods would be more useful.

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Mean and Standard Deviation for Open-Ended Grouped Data

Araştırma Makalesi / Research Article

Ayfer Ezgi YILMAZ*, Serpil AKTAS ALTUNAY

Faculty of Science, Department of Statistics, Hacettepe University, Turkey

(Geliş/Received : 04.12.2020 ; Kabul/Accepted : 07.07.2021 ; Erken Görünüm/Early View : 30.07.2021)

ABSTRACT

Frequency table of continuous quantitative data is arranged so that to describe the data better and convenience of numerical calculations. There are some difficulties to calculate the descriptive statistics of open-ended grouped data. Because the formulations of mean and standard deviation are based on midpoints, and midpoints are based on class intervals, it is necessary to know the lower-most and upper-most categories. In the previous studies, the interquartile, interdecile, interpercentile, and mid-distance ranges were used to estimate the unknown boundaries. This paper proposes four methods to estimate the population mean and standard deviation in open-ended group data. We conduct an extensive Monte Carlo simulation to compare these methods and the results are discussed over CO_2 emission data.

Keywords: Frequency tables, open-ended categories, mean, standard deviation, CO_2 emission.

Açık Uçlu Gruplanmış Veriler için Ortalama ve Standart Sapma

ÖZ

Verilerin daha iyi açıklanması ve hesaplamaların daha kolay olması amacıyla, sürekli nicel verilerin frekans tabloları düzenlenmektedir. Açık uçlu gruplanmış verinin tanımlayıcı istatistiklerini hesaplamada bazı zorluklar ortaya çıkmaktadır. Ortalama ve standart sapmanın formüllerinin sınıf değerlerine, sınıf değerlerinin de sınıf aralıklarına dayanmasından dolayı, ilk sınıfın alt ve son sınıfın üst sınır değerini bilmek gerekir. Önceki çalışmalarda, bilinmeyen sınırları tahmin etmek için çeyrekler, ondalıklar, yüzdeler ve sınıf ara değerleri arası aralıklar kullanılmıştır. Bu çalışmada, açık uçlu gruplanmış verilerinde kitle ortalaması ve standart sapmasını tahmin etmek amacıyla dört yöntem önerilmiştir. Bu yöntemleri karşılaştırmak amacıyla kapsamlı bir Monte Carlo benzetimi uygulanmış ve sonuçlar CO_2 emisyon verileri üzerinden tartışılmıştır.

Anahtar Kelimeler: Frekans tabloları, açık uçlu düzeyler, ortalama, standart sapma, CO_2 emisyonu.

1. INTRODUCTION

The frequency tables in which either no lower or upper limit are called open-ended. Open-ended categories are usually chosen by the researchers and depend on the type of research. Income, IQ scores, SAT scores, number of children, number of cigarettes smoked per day, number of households and number of the living room in a house are some examples having open-ended categories.

Suppose n is the number of observations and the data is grouped into k categories. LB_i denotes the lower boundary and UB_i denotes the upper boundary of the i*th* class. LB_1 is the minimum value and LB_2 is calculated by adding LB_1 to the class interval (c). f_i is the frequency of the i*th* group interval and m_i is the midpoint of i*th* class $[m_i = (LB_i + UB_i)/2]$. Then, the formulation of mean for the grouped data is [1]

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i m_i}{n}, \qquad (1)$$

the formulation of standard deviation for the grouped data is

$$s = \left[\frac{\sum_{i=1}^{k} f_i m_i^2 - \left(\sum_{i=1}^{k} f_i m_i\right)^2 / n}{n-1}\right]^{1/2}. \qquad (2)$$

To use Equations 1 and 2 for open-ended data, it is assumed that the first and last classes (open-ended class) have the same class interval as the other classes. In that case, the midpoints open-ended classes are $m_1 = m_2 - c$ and $m_k = m_{k-1} + c$, where the lower bound of the first and the upper bound of the last categories are ignored. If the values less than calculated m_1 are extremely small or the values greater than calculated m_k are extremely large, these values may be assumed as outliers. Consider an illustrative data example of 40 subjects classified into k=5 categories and summarized in Table 1. The class interval is c=10. Suppose that the true minimum value is 80, then the true value of m_1 is equal to 90. When the minimum value is unknown by the researcher, the first class's midpoint is assumed as $m_1 = 105.5 - 10 = 95.5$.

*Sorumlu Yazar (Corresponding Author)
e-posta : ezgiyilmaz@hacettepe.edu.tr

Table 1. A frequency table

i	LB_i	UB_i	f_i	m_i
1	-	100	5	95.5
2	101	110	8	105.5
3	121	120	10	115.5
4	131	130	13	125.5
5	141	140	4	135.5

The values of midpoints directly affect mean and also standard deviation. The effect of first-class on the mean is $(f_1 m_1/n)$. This value is 11.25 when $m_1 = 90$ and 11.94 when $m_1 = 95.5$. The difference increases when the frequency of the first-class increases. In that case, the calculated mean and standard deviation diverge from the true value. In this study, we discussed interquartile, interdecile, interpercentile, and mid-distance ranges to estimate the unknown boundaries. By these methods, midpoints of open-ended classes may be calculated. Then, mean and standard deviation may be estimated.

Yilmaz and Saracbasi [2] suggested interquartile, interdecile, interpercentile, and mid-distance ranges to estimate the unknown boundaries. They calculated the score values in log-linear models by using these four ranges. In this study, we suggest using these four methods to estimate the population mean and standard deviation in open-ended grouped data. This approach will be a kind of modification of the mid-distance range. It is aimed to compare these methods by their performance to estimate unknown category boundaries, to estimate mean and standard deviation.

In Section 2, the methods to estimate the open-ended boundaries are introduced. The Monte Carlo simulation results are summarized in Section 3. The CO_2 emission data is discussed in Section 4, followed by the conclusion in Section 5.

2. ESTIMATION METHODS OF OPEN-ENDED BOUNDARIES

By classical calculation of mean and standard deviation, $LB_1 = LB_2 - c$ and $UB_k = UB_{k-1} + c$ are used to estimate the unknown open-ended boundaries.

Interquartile range (IQR) is a measure of dispersion and useful to identify the outliers. IQR is the difference between the first and the third quartiles [3,4]. Let's P_{25} be the 25th percentile, and P_{75} be the 75th percentile. The k*th* percentile of grouped data is

$$P_k = LB + c\frac{nk/100 - cf_b}{f_{P_k}}. \quad (3)$$

Here, LB is the lower class boundary of P_k, cf_b is the cumulative frequency of the class before the percentile class, and f_{P_k} is the frequency of the percentile class. Then, interquartile range is $IQR = P_{75} - P_{25}$.

Under the normality assumption, the values less LB_1 and the values greater than UB_k are defined as outliers [5].

$$LB_1 = P_{25} - 1.5IQR \quad \text{and} \quad UB_k = P_{75} + 1.5IQR. \quad (4)$$

Yilmaz and Saracbasi [2] suggested using interdecile range (IDR) and interpercentile range (IPR) as alternatives to IQR to estimate the open-ended boundaries. When IDR is the difference between 10% and 90% percentiles ($IDR = P_{90} - P_{10}$), and IPR is the difference between 5% and 95% percentiles ($IPR = P_{95} - P_5$).

Under the normality assumption, the estimates of open-ended categories by IDR and IPR are

$$LB_1 = P_{10} - 0.78IDR \quad \text{and} \quad UB_k = P_{90} + 0.78IDR, \quad (5)$$

$$LB_1 = P_5 - 0.61IPR \quad \text{and} \quad UB_k = P_{95} + 0.61IPR. \quad (6)$$

Mid-distance range (MDR) is suggested as an alternative to IQR [2]. A mid-distance represents the midpoint of two classes. The mid-distance of the i*th* class is calculated as, $MD_i = (LB_i + UB_{i-1})/2$. Then, MDR is the difference between the mid-distances of 2*nd* and k*th* classes ($MDR = MD_k - MD_2$).

We suggested a modification on MDR. The estimations of open-ended boundaries are suggested as follows the normal distribution.

$$LB_1 = MD_2 - \frac{2}{|Z_1|k}MDR \quad \text{and} \quad UB_k = MD_k - \frac{2}{|Z_k|k}MDR. \quad (7)$$

As f_i is the frequency of the i*th* group interval, the probability of the i*th* category will be $p_i = f_i/n$. $Z_1 = \Phi^{-1}(p_1)$ and $Z_k = \Phi^{-1}(p_k)$ are the inverse of the standard normal cumulative distribution function.

3. SIMULATION STUDY

3.1. The Scope

In this article, a Monte Carlo simulation study is performed to compare the four methods to estimate the open-ended boundaries of frequency tables. The simulation space of the Monte Carlo study is composed of 216 different combinations of variances, the number of categories, and the percentage of outliers.

Data is generated from normal distributions with $X \sim N(\mu = 25, \sigma^2 = 6^2)$ and $X \sim N(\mu = 25, \sigma^2 = 10^2)$. The outlier percentage of the data is assumed as 0%, 5%, 10%, and 20%. After generating the normally distributed data, the values between $[min - 6\sigma;\ min]$ and between $[max; max + 6\sigma]$ are also generated randomly, added to the first data, and these values are assumed as outliers [6,7,8]. To estimate the mean and standard deviation, the number of outliers in the first and last categories is selected randomly. Then, data is grouped with the number of categories (k) from 5 to 13.

Equations 3-7 are used to estimate the values of open-ended categories. These estimates are used to calculate the midpoints. Equations 1 and 2 are used to calculate the means and variances.

3.2. Monte Carlo Simulation

For each of these 50,000 samples, we estimated the minimum and maximum values using four methods. Accordingly, we used these estimates to calculate the mean and the standard deviation of grouped data. We compared the methods using mean absolute error (MAE) and root mean squared error (RMSE). For mean,

$$MAE = \frac{1}{r}\sum_{i=1}^{r}|\mu - \bar{X}_i|, \quad (8)$$

and the root mean square error is

$$RMSE = \sqrt{\frac{1}{r}\sum_{i=1}^{r}(\mu - \bar{X}_i)^2}. \quad (9)$$

Here in Equations 8 and 9, r is the number of replications, $\bar{X}_i$ is the estimated mean by using one of the five methods and $\mu = 25$ is the true value of mean. MAE and RMSE for the unknown categories and standard deviation are calculated in the same way.

The results are represented in three parts: (1) Comparison of methods by estimation of unknown boundaries, (2) Effects of the percentage of outliers to the methods, and (3) Comparison of the methods by mean and standard deviation.

3.3. Results

The results related to RMSE are given in Figures 1-4 as MAE results are very similar to the ones obtained from RMSE. The results are inferred over the methods, percentage of outliers, number of categories, and standard deviation.

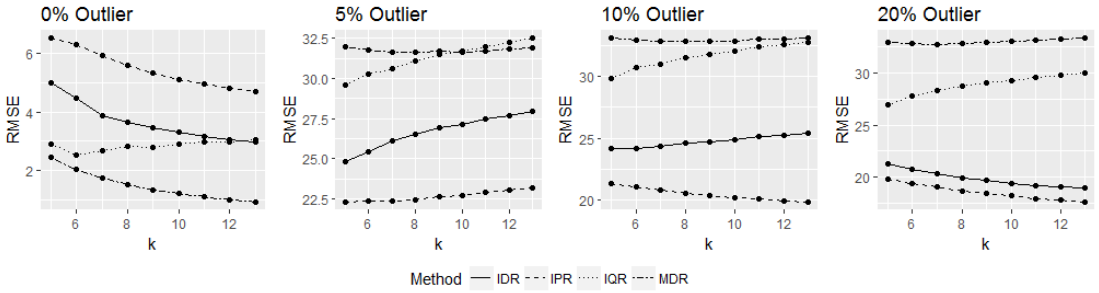
Figures 1 and 2 show the RMSE results for the first-class lower and the last class upper boundaries by the percentage of outliers, respectively. Both for the lower and upper boundaries, the MDR method performs better when there are no outliers. The IPR method performs better when data consist of 5%, 10%, and 20% outliers. When the percentage of outliers is increased, all the methods give larger RMSE and MAE values.

Figures 3 and 4 show the RMSE results for the estimated mean and standard deviation by the percentage of outliers where $\sigma = 6$ and $\sigma = 10$, respectively. The main findings of Figures 3 and 4 are summarized as follows:

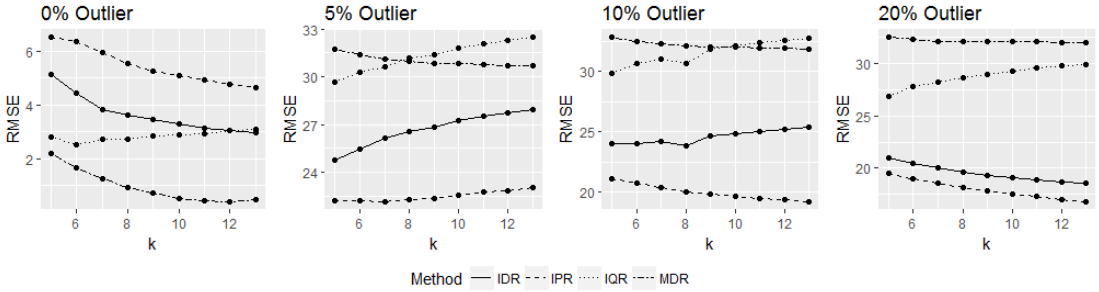
- Both for mean and standard deviation, all methods perform better when there are no outliers and worse when there are 20% percent outliers. When the percentage of outliers is increased, all the methods give larger RMSE and MAE values.
- For 0% percent outliers, when the number of categories increases, the value of RMSE and MAE of mean and standard deviation decrease.
- For 5% percent outliers, the value of RMSE and MAE of mean are not affected by the changes in *k*. When *k* increases, the value of RMSE and MAE of classical and MDR's standard deviations increase and the value of RMSE and MAE of IQR, IDR, and IPR's standard deviations decrease.
- For 10% percent outliers, when the number of categories increases, the value of RMSE and MAE of classical, IDR, IPR, and MDR's means also increase, except for IQR's. IQR's mean is not affected by the changes in *k*. When *k* increases, the value of RMSE and MAE of classical, IPR, and MDR's standard deviations increase. The value of RMSE and MAE of IQR and IDR's standard deviations are not affected by the changes in *k*.
- For 20% percent outliers, when the number of categories increases, the value of RMSE and MAE of mean and standard deviation also increase, except for IQR's standard deviation.
- When the standard deviation is increased, all the methods give larger RMSE and MAE values.

Figures 5 and 6 show the RMSE results for the estimates of mean and standard deviation by five methods (classical, IQR, IDR, IPR, MDR) where $\sigma = 6$ and $\sigma = 10$, respectively. The main findings of the comparison of the methods are summarized as follows:

- For 0% percent outliers, the performances of mean with the classical and MDR methods are similar. The standard deviation with the MDR method performs better where $k \leq 9$ and MDR, IQR, and classical methods perform similarly where $k \geq 10$. When *k* increases, the values of RMSE and MAE of the five methods are getting similar.
- For 5% percent outliers, the mean with the classical method performs better where $k \leq 9$ and the IQR method performs better where $k \geq 10$. The standard deviation with the MDR method performs better where $k \leq 9$ and IQR method performs better where $k \geq 10$.
- For 10% percent outliers, mean with the classical method perform better. The standard deviation with the MDR method performs better where $k = 5$ and the classical method performs better where $k \geq 6$.
- For 20% percent outliers, both mean and standard deviation with the classical method performs better.
- When the standard deviation is increased, all the methods give larger RMSE and MAE values.

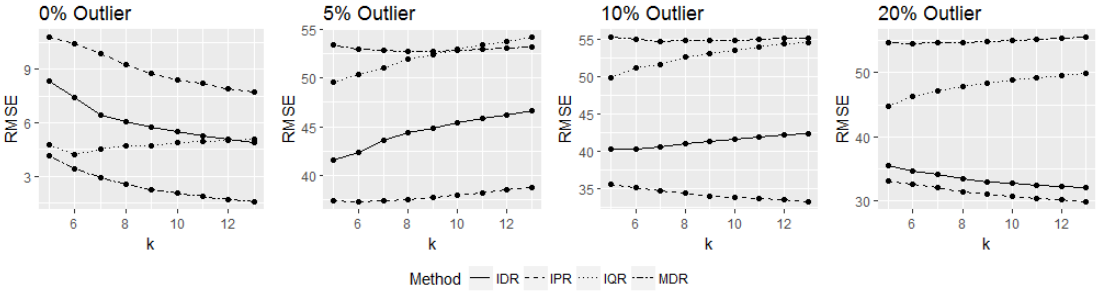


(a) The RMSE results of LB_1

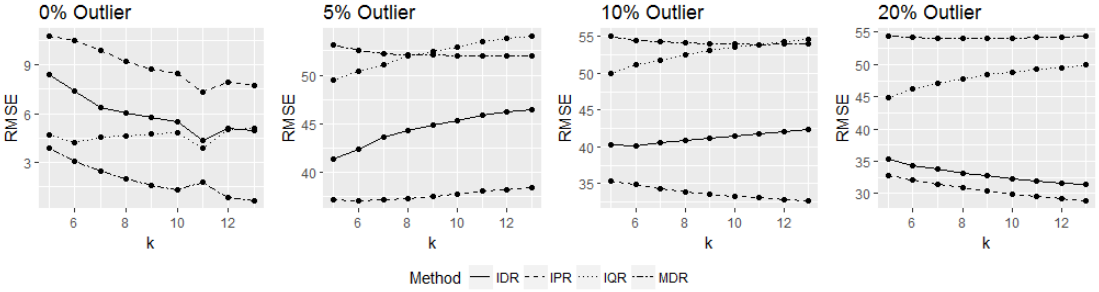


(b) The RMSE results of UB_k

Figure 1. The RMSE results of the unknown boundaries by percentage of outliers where $\sigma = 6$



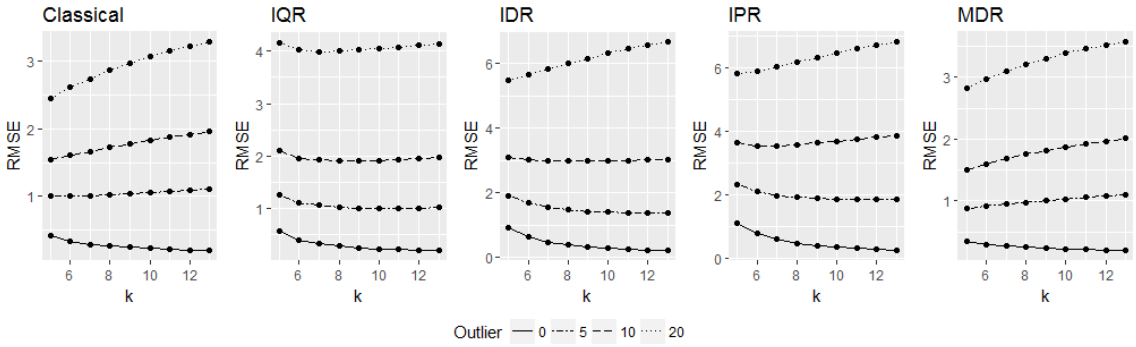
(a) The RMSE results of LB_1



(b) The RMSE results of UB_k

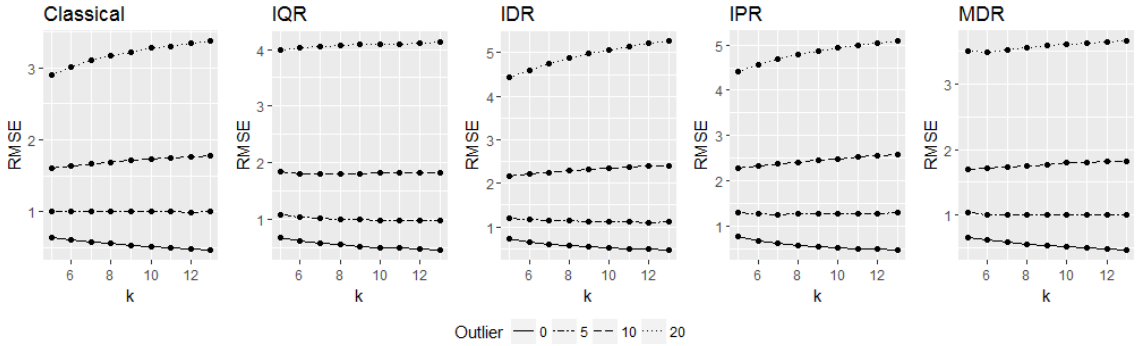
Figure 2. The RMSE results of the unknown boundaries by percentage of outliers where $\sigma = 10$

(a) The RMSE results of mean

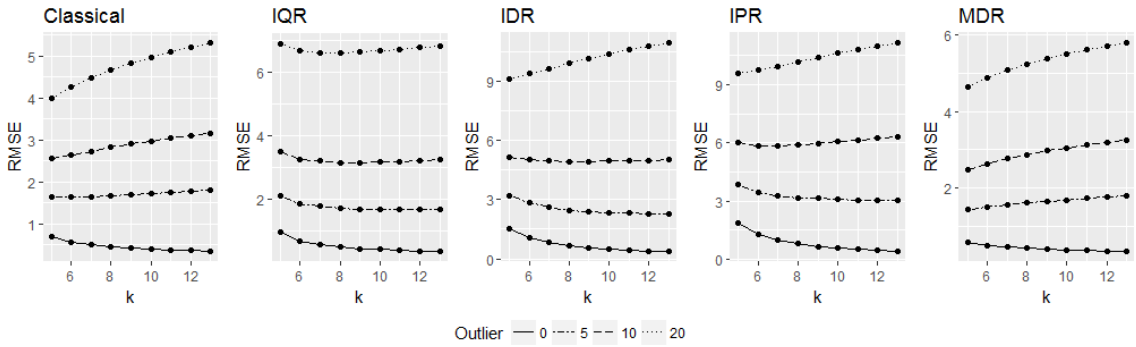


(b) The RMSE results of standard deviation

Figure 3. The RMSE results of mean and standard deviation by methods where $\sigma = 6$



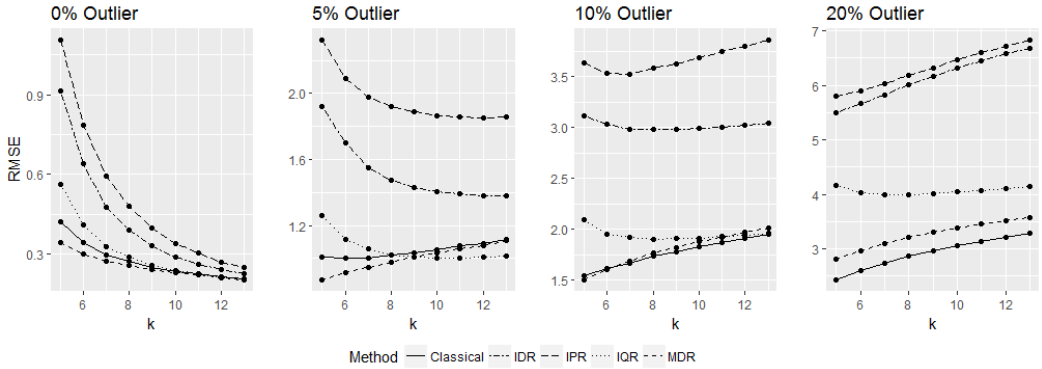
(a) The RMSE results of mean



(b) The RMSE results of standard deviation

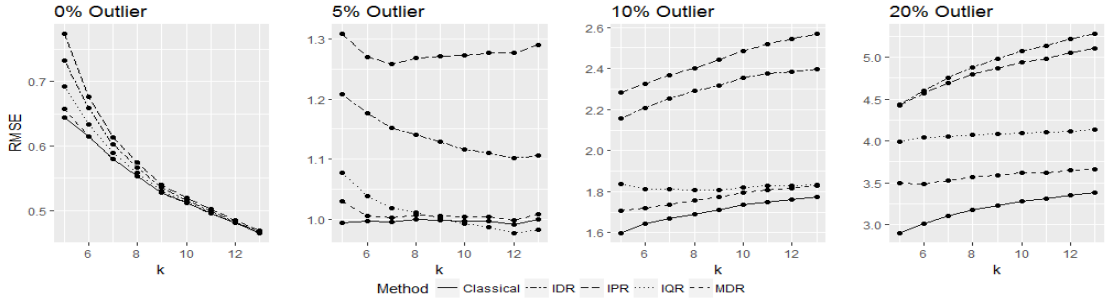
Figure 4. The RMSE results of mean and standard deviation by methods where $\sigma = 10$

(a) The RMSE results of mean



(b) The RMSE results of standard deviation

Figure 5. The RMSE results of mean and standard deviation by percentage of outliers where $\sigma = 6$



(a) The RMSE results of mean

(b) The RMSE results of standard deviation

Figure 6. The RMSE results of mean and standard deviation by percentage of outliers where $\sigma = 10$

4. CO_2 EMISSION DATA

CO_2 emissions (metric tons per capita) of 192 countries data is taken from the World Bank database for 2014 [9]. The mean and standard deviation of CO_2 emission variable in 2014 are 4.44 and 6.07 respectively, and they change up to 0.04 and 45.40. It can be seen from Figure 7; the data are highly positively skewed and consist of many outliers.

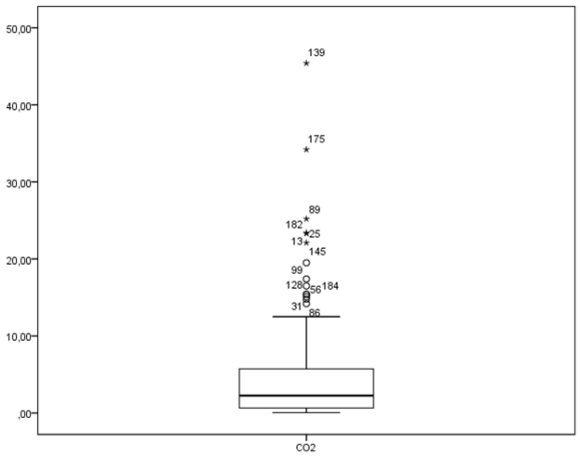


Figure 7. Box plot for CO_2 emission data

The goal is here to group data into five, nine, and ten class intervals and afterward estimate the mean, standard deviation, and unknown first-class lower and last class upper boundaries, and compare the results regarding the five methods given in Section 2. In the interpretations, "Better" means that the method gives the estimation close to the real values.

The data is grouped into 9 categories where the class interval is 3.1. The frequency table is summarized in Table 2.

Table 2. The frequency table of CO_2 emission ($k = 9$)

i	LB_i	UB_i	f_i
1	-	3.0	105
2	3.1	6.1	45
3	6.2	9.2	20
4	9.3	12.3	7
5	12.4	15.4	6
6	15.5	18.5	2
7	18.6	21.6	1
8	21.7	24.7	3
9	24.8	-	3

The estimated first-class lower and last class upper boundaries, means, standard deviations, and absolute errors by the methods given in Section 2 are summarized in Table 3.

The results represented in Table 3 show that the IQR method estimates the unknown lower-most boundary and the MDR method estimates the unknown upper-most boundary better than the others. The classical method estimates mean better and the IDR method estimates standard deviation better.

Suppose the data are grouped into 5 categories where the class interval is 1.5. The frequency table is summarized in Table 4.

Table 4. The frequency table of CO_2 emissions ($k = 5$)

i	LB_i	UB_i	f_i
1	-	1.9	89
2	2.0	3.4	20
3	3.5	4.9	27
4	5.0	6.4	17
5	6.5	-	39

The estimated first-class lower and last class upper boundaries, means, standard deviations, and absolute errors are summarized in Table 5.

The results represented in Table 5 show that the IDR method estimates the unknown lower-most boundary; the IQR method estimates the unknown upper-most boundary better than the others. The classical method estimates the mean better and the MDR method estimates the standard deviation better.

Suppose the data is grouped into 10 categories where the class interval is 1. The frequency table is summarized in Table 6.

The estimated first-class lower and last class upper boundaries, means, standard deviations, and absolute errors are summarized in Table 7.

The results represented in Table 7 show that the IPR method estimates the unknown lower-most boundary; the IDR method estimates the unknown upper-most boundary better than the others. The classical method estimates the mean better and the MDR method estimates the standard deviation better.

Table 3. The estimated first-class lower and last class upper boundaries, means, standard deviations, and absolute errors of CO_2 emission values ($k = 9$)

	Lower Boundary		**Upper Boundary**		**Mean**		**Standard Deviation**	
Method	LB	Abs. Error	UB	Abs. Error	Mean	Abs. Error	SD	Abs. Error
Classical	-	-	-	-	4.584	**0.143**	5.110	0.956
IQR	-5.2	**5.2**	12.3	33.1	3.041	1.400	5.649	0.417
IDR	-7.3	7.3	18.3	27.1	2.514	1.927	6.214	**0.149**
IPR	-8.9	8.9	24.2	21.1	2.122	2.318	6.695	0.629
MDR	-44.0	44.0	27.0	**18.4**	16.609	12.168	8.483	2.417

Table 5. The estimated first-class lower and last class upper boundaries, means, standard deviations, and absolute errors of CO_2 emission values ($k = 5$)

	Lower Boundary		**Upper Boundary**		**Mean**		**Standard Deviation**	
Method	LB	Abs. Error	UB	Abs. Error	Mean	Abs. Error	SD	Abs. Error
Classical	-	-	-	-	3.395	**1.045**	2.412	3.654
IQR	-5.3	5.3	12.3	33.1	2.498	1.943	4.389	1.677
IDR	-4.2	**4.2**	12.2	33.2	2.743	1.698	4.130	1.936
IPR	-3.6	3.6	11.8	33.6	2.841	1.599	3.993	2.113
MDR	-17.7	17.7	8.6	36.8	-0.752	5.192	6.804	**0.739**

Table 6. The frequency table of CO_2 emissions ($k = 10$)

i	LB_i	UB_i	f_i
-	1.9	89	-
2.0	2.9	15	2.0
3.0	3.9	14	3.0
4.0	4.9	18	4.0
5.0	5.9	11	5.0
6.0	6.9	9	6.0
7.0	7.9	5	7.0
8.0	8.9	6	8.0
9.0	9.9	6	9.0
10.0	-	19	10.0

Table 7. The estimated first-class lower and last class upper boundaries, means, standard deviations, and absolute errors of CO_2 emission values ($k = 10$)

	Lower Boundary		**Upper Boundary**		**Mean**		**Standard Deviation**	
Method	LB	Abs. Error	UB	Abs. Error	Mean	Abs. Error	SD	Abs. Error
Classical	-	-	-	-	3.934	**0.506**	3.109	2.957
IQR	-4.8	4.8	12.0	33.4	2.645	1.796	4.404	1.665
IDR	-5.7	5.7	16.7	**28.7**	2.668	1.772	5.063	1.003
IPR	-4.7	**4.7**	16.2	29.2	2.879	1.565	4.802	1.264
MDR	-15.5	15.5	11.2	34.2	0.125	4.316	6.789	**0.723**

5. CONCLUSIONS

Grouping data is a simple way that makes it easier for researchers to understand the raw data. Open-ended classes occur most frequently in many types of research. An open-ended distribution means that it has no boundary or both boundaries. Even though the open-ended classes may cause problems with calculations and interpretation in practice but they are unavoidable because of the nature of data. The measures of central tendencies and dispersion are calculated over the midpoints in grouped data. Finding out the midpoint for open-ended categories is a consideration for calculating the descriptive statistics. For instance, to calculate the mean of grouped data, the first step is to determine the midpoint of each class. But the disadvantage of the classic method is that the midpoint values directly affect the mean and standard deviation. In this study, we discuss interquartile, interdecile, interpercentile, and mid-distance ranges to estimate the unknown boundaries. Employing these methods, midpoints of open-ended classes, then mean and standard deviation can be estimated more accurately. The new methods enable us to calculate the midpoints considering the open-ended limits.

The simulation results show that in case of no outliers, entire methods perform well. When the number of categories increases, the value of RMSE and MAE of mean and standard deviation also increase; this case became more obvious for 20% percent outliers. In most cases, the value of k has no remarkable effect on the RMSE and MAE's. The MDR and IDR methods outperform classical methods when estimating the LB and UB values. Estimating the standard deviation with the MDR method in open-ended grouped data ($k \leq 9$) gives closer results to true values. Similarly, the IQR method performs better for $k \geq 10$. While the most appropriate measures of central tendency in open-ended data seem to be median, the proper application of mean among the proposed methods would be more useful.

DECLARATION OF ETHICAL STANDARDS

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Ayfer Ezgi YILMAZ: Design of the simulation study, analysis of results, interpretation of results, wrote the manuscript.

Serpil AKTAŞ ALYUNAY: Interpretation of results, wrote the manuscript.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

KAYNAKLAR (REFERENCES)

[1] Freund J.E., "Modern elementary statistics" 11th Ed., ***Pearson Education***, Upper Saddle River, New Jersey, (2004).

[2] Yilmaz A.E. and Saracbasi T., "The effect of changing scores for multi-way tables with open-ended ordered categories", ***Hacettepe Journal of Mathematics and Statistics***, 45(6): 1881-1890, (2016).

[3] Nick T.G., "Descriptive Statistics", Topics in biostatistics, ***Humana Press***, Totowa, New Jersey, (2007).

[4] Tukey J.W., "Exploratory data analysis", ***Addison-Wesley***, Massachusetts, (1977).

[5] Frigge M., Hoaglin D.C. and Iglewicz, B., "Some implementations of boxplot", ***The American Statisticians***, 43(1): 50-54, (1989).

[6] Yang J., Rahardja S. and Fränti P., "Outlier detection: How to threshold outlier scores?", ***International Conference on Artificial Intelligence, Information Processing and Cloud Computing***, 1-6, (2019).

[7] Simmons J.P., Nelson L.D. and Simonsohn U., "False positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant", ***Psychological Science***, 22(11): 1359–1366, (2011).

[8] Miller J., "Reaction time analysis with outlier exclusion: Bias varies with sample size", ***The Quarterly Journal of Experimental Psychology***, 43(4): 907– 912, (1991).

[9] ***https://data.worldbank.org/indicator/EN.ATM.CO2E.P***, The World Bank Data, "CO2 Emissions (metric tons per capita)", (2019).