CHOOSING THE MOST REPRESENTATIVE AND EFFICIENT AVERAGES OF NUMERIC UNIVARIATE DATA SETS: VOTING AND BOOTSTRAPPING TECHNIQUES

Kayode Ayinde, Brian Haile, David Vlieger, and Taylor Harrison

Department of Mathematics and Statistics, Northwest Missouri State University, Maryville, Missouri, USA.

Email Address: ayindek@nwmissouri.edu

Abstract

Numeric univariate data set exhibits different characteristics which are expected to be summarily provided by a typical value or a representative of a set of values called averages. These characteristics change as data set departs from being symmetric to asymmetric with and without outliers resulting into a challenge of acceptance of each average to the subjects being represented. In this research, the voting and bootstrapping techniques are adopted as methods through which every data set can choose its best averages in terms of representativeness and efficiency. While bootstrapping method provides the most efficient average as one having least standard error, the voting technique provides opportunity for every subject in a data set to choose one and only one of the averages as its best representative and thereafter, the most representative average of the data set as one having the highest counts. The techniques were illustrated with eighteen (18) data sets of different characteristics sourced from https://artofstat.com/web-apps. Results show that the most representative average could be any of mode, mid-range, median, Lehmer mean and harmonic mean, and that the most efficient average could be any of harmonic mean, geometric mean, arithmetic mean, quadratic mean, Lehmer mean, mid-range and median. The study, therefore, recommends that every numeric data set should be allowed to choose its most representative using voting technique and its most efficient average using the bootstrapping method as both techniques provide better opportunity for the averages to interact with the data set and compete for their choice as the best averages.

Keywords: Averages, Voting Technique, Most Representative Average, Bootstrapping Technique, Most Efficient Average.

1. Introduction

An average of a data set is a representative of the data set which attempts to summarize and provide the characteristics of the data by a value (Mokros and Russell,1995; Mokros and Russell,1996; De

Carvalho, 2016; Emovon and Okechukwu, 2017). The commonest ones include the Mid-range (MR), Arithmetic Mean (AM), Geometric Mean (GM), Harmonic Mean (HM), Quadratic Mean (QM), Cubic Mean (CM), Quartic Mean (QTM), Median (MED), and Mode (MOD) which can now be obtained from the generalized, power or holder mean (GEM) defined as:

$$\bar{X}_{GEM(p)} = \left[\frac{\sum_{i=1}^{n} X_i^p}{n} \right]^{\frac{1}{p}} \quad (1)$$

Arranging $X_{1,} X_{2,} \cdots_{,} X_n$ in order of magnitude as $X_{[1],} X_{[2],} \cdots_{,} X_{[n]}$, $\bar{X}_{\substack{GEM(p) \\ p \to -\infty}} = X_{Min,} = X_{[1]}$ and $\bar{X}_{\substack{GEM(p) \\ p \to \infty}} = X_{Max,} = X_{[n]}$. Therefore, the mid-range and the median can be obtained respectively as:

$$\bar{X}_{MR} = \frac{X_{Max.} + X_{Min.}}{2} = \frac{X_{[1].} + X_{[n].}}{2} \quad (2)$$

$$\bar{X}_{MED} = \begin{cases} Mid-value\ of\ \left[X_{[1],} X_{[2],} \cdots_{,} X_{[n]}\right] = X_{\left[\frac{n+1}{2}\right]}, & if\ n\ is\ odd \\ Arithemetic\ \ mean\ of\ two\ mid-values\ of\ \left[X_{[1],} X_{[2],} \cdots_{,} X_{[n]}\right] = \frac{X_{\left[\frac{n}{2}\right]} + X_{\left[\frac{n+1}{2}\right]}}{2}, & if\ n\ is\ even \end{cases} \quad (3)$$

When p=-1, p=1, p=2, p=3, and p=4; the $\bar{X}_{GEM(p)}$ respectively becomes $\bar{X}_{HM}$,$\bar{X}_{AM}$,$\bar{X}_{QM}$, $\bar{X}_{CM}$, and $\bar{X}_{QTM}$; and with $\bar{X}_{\substack{GEM(p) \\ p \to 0}} = \bar{X}_{GM}$.

Furthermore, the mode symbolically is:

$$\bar{X}_{MOD} = Most\ f\ requent\ value\ of \left[X_{1,} X_{2,} \cdots_{,} X_n\right] \quad (4)$$

(Goodchild, 1988; Dor and Zwick 1999; Emovon and Okechukwu, 2017; Vogel, 2020; Mukhopadhyay et.al, 2021).

Another average also found in literature is Lehmer Mean (LM) which is defined as:

$$\bar{X}_{LM(p)} = \frac{\sum_{i=1}^{n} X_i^p}{\sum_{i=1}^{n} X_i^{p-1}} \tag{5}$$

When p=0, p=$\frac{1}{2}$ for any two values (say, $X_1 and X_2$), p=1, and p=2; the $\bar{X}_{LM(p)}$ respectively becomes $\bar{X}_{HM}$, $\bar{X}_{GM}$, $\bar{X}_{AM}$, and $\bar{X}_{Contra\ HM}$; and when $\bar{X}_{\substack{LM(p) \\ p\to-\infty}} = X_{Min,} = X_{[1]}$, and when $\bar{X}_{\substack{LM(p) \\ p\to\infty}} = X_{Max,} = X_{[n]}$ (Bullen,1987; Halley, 2004; Kennedy and Stanley, 2009).

Data sets especially numeric ones do exhibit different features ranging from being symmetric to being asymmetric (positively skewed and negatively skewed data) with and without outliers. These features often affect the representativeness of data sets by these averages. The mid-range is the simplest but only make sure of the two extreme values. The arithmetic mean has been some good statistical properties, but it is affected by outliers (Ajiboye et al, 2017; Alao, 2019; Vogel, 2022). The geometric and harmonic mean are less affected by outliers but have computational challenges with zero and/or negative value(s). The median is robust, but each value of observation is not used maximally and hence may not account for preferences. The mode is the only average that can be used for both numeric and non-numeric data, but at times it may not exist and if it does exist, it may not be unique (Kennedy and Stanley, 2009; Muthuvalu et. al, 2015; De Carvalho, 2016; Vogel, 2020; Mukhopadhyay et.al, 2021).

The uses of some of these averages have been restricted to specific situations while some are even becoming unpopular despite their knowledge of being representatives of the same data set. The arithmetic mean has been reported to be most suitable to represent data that are symmetric and the median for skewed and data sets with outliers (Crump, 1998; Casella and Berger, 2002; Hinkle et.al, 2003; Brase et.al, 2023). Others are less spoken about. Curto (2022) argued that there is nothing preventing any of these averages to be used as a representative of the data set provided it can be used justifiably (Mukhopadhyay et.al, 2021; Takacs and Bourrat, 2022). Moreover, there arises a question of how agreeable or acceptable each of these averages is to all the subjects as their representative as each may not be representing the majority efficiently well. In this research, efforts are not only made to overcome the challenge of getting the most representative average as mode may not always be but also to adopt the voting technique through which all the averages can compete for their acceptance by each subject. This concept of voting technique is now being embraced in various fields of study including statistics and probability to provide solutions to some challenges (Andrew et.al, 2002; Kun and Jiang, 2010; Diss and Merlin, 2021; Awde et.al, 2023). Furthermore, this research also provides opportunity for the averages to compete for their efficiency through the technique of bootstrapping.

2. Materials And Methods

2.1 The Voting Technique

Voting technique is adopted into measuring the discrepancy between each subject of the data set $X_1, X_2, \ldots, X_n$ and each average using the absolute deviation measure. The measure requires all the averages to contest for their acceptance by each of the observations/subjects of the data while each subject is expected to vote for one and only one of the averages as its best average; the average with the smallest discrepancy in absolute value (discrepancy closest to zero). The average that is voted for is then scored 1 while other averages are scored 0. The number of times each average scores 1 is the added together and the average with highest frequency (mode) is declared winner of the contest and the most representative average of the data set being considered. Alternatively, the frequency can be converted to relative frequency and when this happens, the average with the highest probability (relative frequency) is declared the winner and the most representative average of the data set under consideration.

Mathematically, the statistic is represented as:

$$p_j = \frac{\sum_{i=1}^{n} \gamma_{ij}}{n} \tag{6}$$

where

$$\gamma_{ij} = \begin{cases} 1 \ \text{if} \ \left|X_{ij} - \bar{X}_j\right| \ \text{is the minimum} \\ \\ 0 \ \ \text{otherwise} \end{cases} \quad for\ i = 1, 2, \ldots, n;\ and\ \ j = MR, AM, GM, HM, QM, CM, QTM, MED, MOD, LM(p).$$

and the most representative average for any data set is the one having the highest relative frequency defined as:

$$\underset{j}{Max}.\left[p_j\right]=Max.\left[\frac{\sum_{i=1}^{n}\gamma_{ij}}{n}\right] \quad (7)$$

This approach does provide equal opportunity to all the averages to be chosen as the most representative average and so, the mode may not necessary be the most representative average but shall only be if 50% of all the observations have the same value. Moreover, the challenges of non-existence and/or non-uniqueness of mode are overcome for any data set as the data set must produce at least one of the averages as the most representative average.

2.2 The Bootstrapping Technique

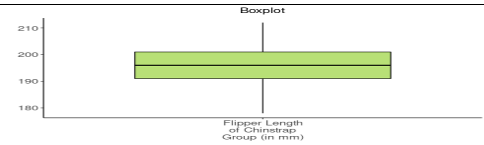
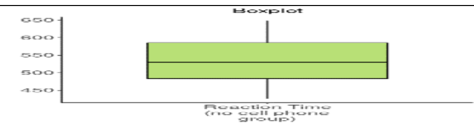
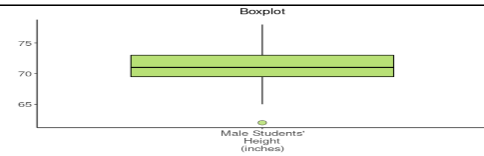
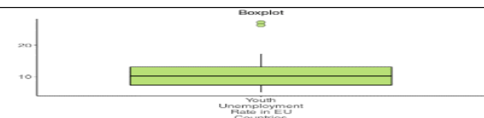
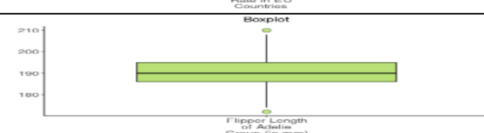
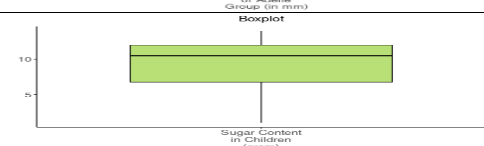
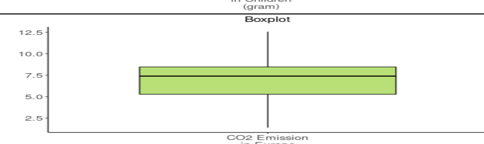
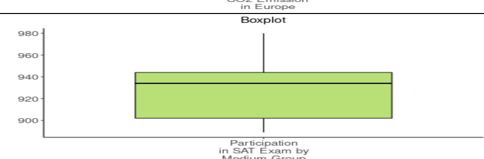
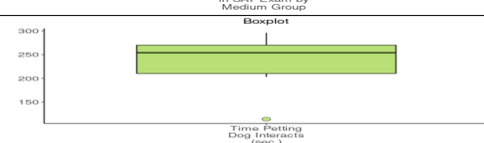
Bootstrapping is a versatile statistical resampling technique introduced by Efron (1979) for estimating standard errors, constructing confidence intervals and testing hypotheses. It is a procedure that enables the distribution of an estimator to be empirically investigated through resampling. The principle of its resampling involves sampling with replacement from a known (original) dataset to create several or multiple simulated data sets to allows variability of almost any statistic or model to be estimated (Efron, 2003; Horowitz, 2019). Furthermore, it assigns measure of accuracy in terms of bias, variance and confidence intervals to sample estimates and (Efron and Tibshirani,1993; Efron, 2003). Various other developments were noticeable as the technique becomes more useful and relevant in various fields (Bickel and Freedman, 1981; Singh, 1981; DiCiccio and Efron,1992; Shoemaker, Owen and Pathak, 2001; Good, 2006; Kleiner et. al, 2014; Ayinde et. al, 2023).

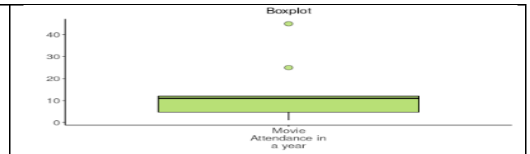
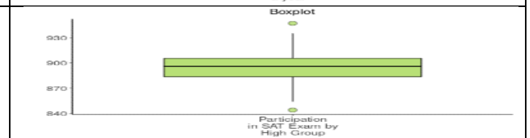
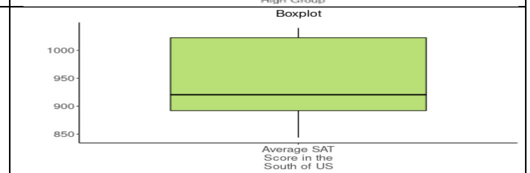
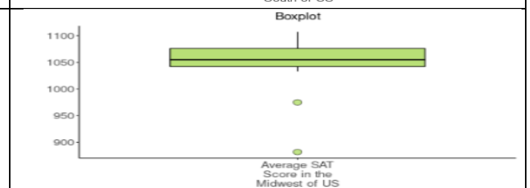
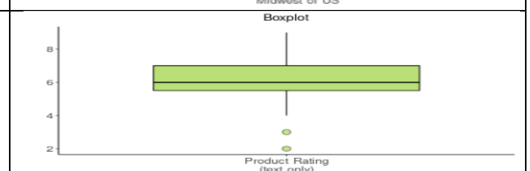
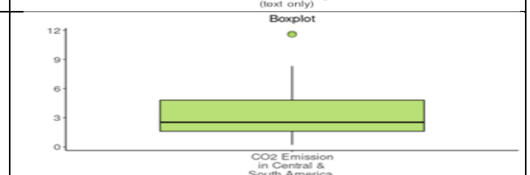
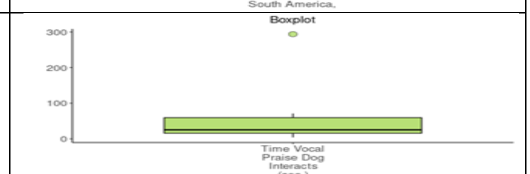
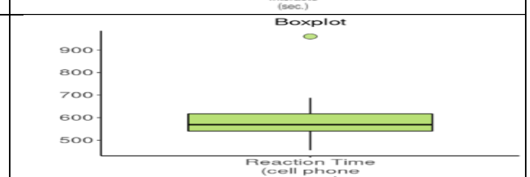
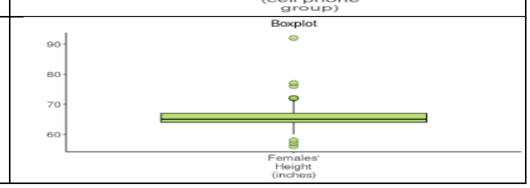
The basic procedures and concepts for bootstrapping include collecting or getting the original data set, resampling with replacement to create multiple bootstrap samples, estimating the statistic (in this case the averages already discussed in equation (1) to (5)) for each of the bootstrap samples, and analyzing the distribution of the estimated statistics across the bootstrap samples. Eighteen (18) data sets used in this study exhibit different characteristics ranging from being symmetric to being asymmetric (negatively and positively skewed data sets) with and without outliers as well as having no mode, one or more than one modes status. The datasets were sourced from this website (https://artofstat.com/web-apps). For each of the data set, bootstrap simulation study was further conducted 10,000 times on the averages to provide estimates for their biases and standard errors using R package. The average with the least standard error is thus identified as the most efficient average.

3. Results and Discussion

The summary of the nature of the eighteen (18) data sets classified as either symmetric, left skewed, or right skewed data as well as their outlier status (no outlier, outlier(s) in the left direction, outlier(s) in the right direction), their mode(s) and their pictorial representation using boxplot is provided in Table 1.

Table 1: The Nature of the Data Set Used and their pictorial representation.

Data Nature	Outlier Status	Variable Name	Mode(s) in the data set	The Boxplot
Symmetric	No	Palmer Pinguins: Flipper Length of Chinstrap Group (in mm)	Mode 1=187 Mode 2=195	
		Reaction Time (No cell phone group)	Mode 1=485 Mode 2=626	
	Left	Male Students' Height (inches)	Mode 1=70	
	Right	Youth Unemployment Rate in EU Countries	Mode 1=7.0 Mode 2=10.3	
	Both	Palmer Pinguins: Flipper Length of Chinstrap Group (in mm)	Mode 1=190	
Left Skewed Data	No	Sugar Content in Children (gram)	Mode 1=12 Mode 2=14	
		CO2 Emission in Europe	No Mode	
		Participation in SAT Exam by Medium Group	No Mode	
	Left	Time Petting Dog interacts (sec.)	No Mode	

	Right	Movie Attendance in a year	Mode 1=12	
	Both	Participation in SAT Exam by High Group	Mode =896	
Right Skewed Data	No	Average SAT Score in the South of US	No Mode	
	Left	Average SAT Score in the Midwest of US	No Mode	
		Product rating (text only)	Mode =7	
	Right	CO2 Emission in Central & South America	No Mode	
		Time Vocal Praise Dog Interacts (sec.)	No Mode	
		Reaction Time (cell phone group)	Mode =554	
	Both	Female Students' Height (inches)	Mode =64	

From Table 1, it can be observed that the mode of the variables varies from none (no mode) to one mode and to two modes.

The results obtained by adopting the voting technique to get the most representative average and using the bootstrapping approach for examining the efficiency of the averages are presented in Tables 2, 3, and 4 and the summary is also provided in Table 5. From these tables, the most representative average (MRA) in the data sets is observed not to be the mode all the time as there are instances when data sets have two modes of which none is the MRA. Moreover, whenever the MRA is not mode, the mode is either found in the second or at most the third preference (rank) competing very keenly with the MRA. Other averages observed to be MRA include the Lehmer Mean 54, the mid-range, the median, and the harmonic mean especially when the data set is left skewed and right skewed with outlier on the right direction. Similarly, from the tables, the most efficient average is among Lehmer Mean 54, quadratic mean, mid-range, arithmetic mean, harmonic mean, geometric mean, and the media.

Table 2: Results of Voting and Bootstrapping Techniques with Symmetric Data Sets

Outlier	Variable	Measures of Location		Voting Approach		Bootstrapping Approach	
		Name	Value	Relative Frequency	Rank	Bias	Standard Error
No	Palmer Pinguins: Flipper Length of Chinstrap Group (in mm)	Mid-range	195	0.26471	3	0.26095	1.1467614
		Arithmetic Mean	195.8235	0	10	0.007407353	0.8534081
		Geometric Mean	195.6953	0	10	0.009302293	0.854326
		Harmonic Mean	195.5669	0	10	0.011214874	0.8557782
		Quadratic Mean	**195.9514**	0	10	**0.005523186**	**0.8530137**
		Cubic Mean	196.0791	0	10	0.003643151	0.8531308
		Quartic Mean	196.2065	0	10	0.001760842	0.8537459
		Lehmer Mean 21	196.0795	0	10	0.003638956	0.8531655
		Lehmer Mean 32	196.3346	0	10	-0.00011718	0.854983
		Lehmer Mean 43	196.589	0	10	-0.00388678	0.8587792
		Lehmer Mean 54	**196.8427**	**0.45588**	**1**	-0.00769442	0.86446
		Median	196	0.05882	5	-0.19835	0.9034179
		Mode 1	187	0.26471	3		
		Mode 2	195	0.26471	3		
	Reaction Time (no cell phone group)	**Mid-range**	**537**	0	11.5	**0.1434**	**5.553001**
		Arithmetic Mean	533.5938	0.0625	5	-0.18876562	11.280801
		Geometric Mean	529.7217	0.03125	7	-0.06712712	11.218837
		Harmonic Mean	525.874	0.09375	4	0.05259729	11.135686
		Quadratic Mean	537.4576	0.03125	7	-0.31067513	11.319601
		Cubic Mean	541.2818	0	11.5	-0.43114181	11.334328
		Quartic Mean	545.0367	0	11.5	-0.54852826	11.325191
		Lehmer Mean 21	541.3495	0	11.5	-0.43357759	11.393193
		Lehmer Mean 32	549.0118	0	11.5	-0.67498155	11.463845
		Lehmer Mean 43	556.4585	0.03125	7	-0.90619917	11.487419
		Lehmer Mean 54	563.584	0.15625	3	-1.12138202	11.462907
		Median	530	0	11.5	-1.2073	19.14692
		Mode 1	**485**	**0.375**	**1**		
		Mode 2	626	0.21875	2		
Left	Male Students' Height (inches)	**Mid-range**	**70**	**0.46154**	**1.5**	0.547425	0.7426387
		Arithmetic Mean	**70.93162**	0	9.5	**0.003497436**	**0.2628719**
		Geometric Mean	70.87415	0	9.5	0.003978673	0.2632353
		Harmonic Mean	70.81651	0.00855	5	0.004468904	0.2639877
		Quadratic Mean	70.98895	0	9.5	0.003020833	0.2628824
		Cubic Mean	71.04618	0	9.5	0.002544748	0.2632523
		Quartic Mean	71.10332	0	9.5	0.002065276	0.2639673
		Lehmer Mean 21	71.04633	0	9.5	0.002544496	0.2631726
		Lehmer Mean 32	71.16076	0	9.5	0.001593314	0.2648096
		Lehmer Mean 43	71.27503	0	9.5	0.000628185	0.2677011
		Lehmer Mean 54	71.38926	0.36752	3	-0.00036573	0.2717649
		Median	71	0.16239	4	-0.22515	0.4068772
		Mode 1	**70**	**0.46154**	**1.5**		
Right	Youth Unemployment Rate in EU Countries	Mid-range	16.1	0.07143	5	-0.60771	1.6869511
		Arithmetic Mean	11.12857	0	12.5	0.007475357	1.0290635
		Geometric Mean	10.07363	0.03571	8.5	0.036772963	0.8239245
		Harmonic Mean	9.24585	0.10714	3	**0.052005582**	**0.69518**
		Quadratic Mean	12.40608	0.17857	2	-0.05663453	1.3024179
		Cubic Mean	13.81971	0	12.5	-0.16704126	1.5896442
		Quartic Mean	15.23121	0	12.5	-0.30592447	1.8398091
		Lehmer Mean 21	13.83023	0.03571	8.5	-0.12576212	1.6792018
		Lehmer Mean 32	17.14859	0.03571	8.5	-0.43562505	2.4290189
		Lehmer Mean 43	20.39108	0	12.5	-0.85058113	2.9719348
		Lehmer Mean 54	22.91226	0.07143	5	-1.1765512	3.2698551
		Median	10.15	0.03571	8.5	-0.38072	1.164678
		Mode 1	**7**	**0.35714**	**1**		
		Mode 2	10.3	0.07143	5		
Both	Palmer Pinguins: Flipper Length of Chinstrap Group (in mm)	**Mid-range**	**191**	**0.44371**	**1**	0.116	1.3481935
		Arithmetic Mean	189.9536	0	9	0.00119	0.5396862
		Geometric Mean	**189.8419**	0	9	**0.001970309**	**0.5395078**
		Harmonic Mean	189.7301	0.43046	2	0.002742319	0.539717
		Quadratic Mean	190.0654	0	9	0.00039781	0.5402594
		Cubic Mean	190.1772	0	9	-0.00041033	0.5412348
		Quartic Mean	190.2891	0	9	-0.00123832	0.5426206
		Lehmer Mean 21	190.1773	0	9	-0.0003944	0.5411758
		Lehmer Mean 32	190.4011	0	9	-0.00202648	0.5442245
		Lehmer Mean 43	190.625	0	9	-0.0037221	0.5488715
		Lehmer Mean 54	190.8493	0	9	-0.00549782	0.555156
		Median	190	0.12583	3.5	0.0346	0.3914309
		Mode 1	190	0.12583	3.5		

Table 3: Results of Voting and Bootstrapping Techniques with Left Skewed Data Sets

		Measures of Location		Voting Approach		Bootstrapping Approach	
Outlier	Variable	Name	Value	Relative Frequency	Name	Bias	Standard Error
No	Sugar Content in Children (gram)	Mid-range	7.5	0	10.5	0.445	1.025222
		Arithmetic Mean	9.2	0.1	5	0.00726	1.3374125
		Geometric Mean	7.415708095	0	10.5	0.24588337	1.8316503
		Harmonic Mean	**4.735870977**	**0.3**	**1**	0.8979829	2.3828148
		Quadratic Mean	10.13903348	0.1	5	-0.05291831	1.0653611
		Cubic Mean	10.69929247	0	10.5	-0.07637289	0.9207106
		Quartic Mean	11.0830884	0.1	5	-0.09093063	0.8368182
		Lehmer Mean 21	11.17391304	0	10.5	-0.09687869	0.8500249
		Lehmer Mean 32	11.91439689	0	10.5	-0.10770072	0.7327846
		Lehmer Mean 43	12.3190725	0	10.5	-0.12356828	0.6959358
		Lehmer Mean 54	**12.60194587**	0	10.5	**-0.1423496**	**0.6810764**
		Median	10.5	0	10.5	-0.3153	1.77247
		Mode 1	12	0.2	2.5		
		Mode 2	14	0.2	2.5		
	CO2 Emission in Europe	Mid-range	6.98046	0.03225806	8.5	0.16327455	0.6390809
		Arithmetic Mean	7.001402903	0	11.5	0.004383	0.4515703
		Geometric Mean	6.453066228	0.06451613	5.5	0.024046	0.5075594
		Harmonic Mean	**5.719252419**	**0.38709677**	**1**	0.08015239	0.6573299
		Quadratic Mean	**7.439982658**	0	11.5	**-0.00799422**	**0.4431117**
		Cubic Mean	7.815273514	0.03225806	8.5	-0.02031437	0.4563059
		Quartic Mean	8.149918089	0.03225806	8.5	-0.03478236	0.4803671
		Lehmer Mean 21	7.906035793	0.12903226	2.5	-0.02006595	0.4610955
		Lehmer Mean 32	8.623601887	0.09677419	4	-0.04555656	0.5345918
		Lehmer Mean 43	9.242307847	0.06451613	5.5	-0.08131237	0.6238592
		Lehmer Mean 54	9.793099762	0.12903226	2.5	-0.12794675	0.7074886
		Median	7.39317	0.03225806	8.5	-0.18447531	0.7425105
	Participation in SAT Exam by Medium Group	**Mid-range**	**934.5**	**0.44444444**	**1.5**	**-1.19565**	**6.666576**
		Arithmetic Mean	930.1111111	0	8	0.05728889	10.350011
		Geometric Mean	929.588872	0	8	0.11472382	10.31121
		Harmonic Mean	**929.0695546**	**0.44444444**	**1.5**	0.17123243	10.271552
		Quadratic Mean	930.6359236	0	8	-0.00102489	10.38791
		Cubic Mean	931.1629513	0	8	-0.0601662	10.424866
		Quartic Mean	931.691828	0	8	-0.12008008	10.460841
		Lehmer Mean 21	931.1610321	0	8	-0.0593736	10.428167
		Lehmer Mean 32	932.2179025	0	8	-0.17855658	10.505849
		Lehmer Mean 43	933.2802609	0	8	-0.30004272	10.582881
		Lehmer Mean 54	934.3466092	0	8	-0.42359698	10.659083
		Median	934	0.11111111	3	-5.4697	16.353685
Left	Time Petting Dog interacts (sec.)	**Mid-range**	**205**	**0.28571429**	**1.5**	12.39425	24.14061
		Arithmetic Mean	232	0	9	-0.1769857	21.27326
		Geometric Mean	223.0301002	0	9	1.1995389	24.5169
		Harmonic Mean	211.8060951	0.14285714	4	3.6375696	28.28654
		Quadratic Mean	238.9315503	0	9	-0.9107005	18.77224
		Cubic Mean	244.3057269	0	9	-1.335455	16.98757
		Quartic Mean	248.5704989	0	9	-1.6262492	15.75179
		Lehmer Mean 21	246.070197	0	9	-1.6073256	16.63089
		Lehmer Mean 32	255.4194356	0.14285714	4	-2.1049763	14.39712
		Lehmer Mean 43	261.8167292	0	9	-2.4093598	13.45509
		Lehmer Mean 54	**266.4934769**	**0.28571429**	**1.5**	**-2.7049388**	**12.97868**
		Median	254	0.14285714	4	-11.522	27.57742
Right	Movie Attendance in a year	Mid-range	23	0	10	-3.90105	6.108282
		Arithmetic Mean	13	0	10	-0.05125	3.950615
		Geometric Mean	8.011031303	0.1	4.5	0.4185667	2.87609
		Harmonic Mean	**4.338245421**	**0.3**	**1.5**	**0.7954802**	**2.415859**
		Quadratic Mean	18.03330253	0	10	-0.8522842	5.305885
		Cubic Mean	22.4633147	0	10	-1.8593036	6.475314
		Quartic Mean	25.99483802	0	10	-2.7845593	7.374
		Lehmer Mean 21	25.01538462	0.1	4.5	-2.0102536	7.719226
		Lehmer Mean 32	34.85547355	0	10	-4.950603	10.368727
		Lehmer Mean 43	40.28347596	0	10	-6.8797321	11.516635
		Lehmer Mean 54	42.74035661	0.1	4.5	-7.7422055	11.865059
		Median	11	0.1	4.5	-0.9022	3.16422
		Mode 1	**12**	**0.3**	**1.5**		
Both	Participation in SAT Exam by High Group	Mid-range	895.5	0	9	-0.8635	7.842593
		Arithmetic Mean	893.7777778	0	9	0.13423889	5.69456
		Geometric Mean	893.4512463	0	9	0.15043841	5.693226
		Harmonic Mean	**893.1246379**	**0.44444444**	**1.5**	0.16668347	5.694551
		Quadratic Mean	894.1042693	0	9	0.11803329	5.698574
		Cubic Mean	894.430759	0	9	0.10176942	5.705277
		Quartic Mean	894.7572855	0	9	0.08539472	5.714671
		Lehmer Mean 21	894.4308802	0	9	0.10183077	5.704885
		Lehmer Mean 32	895.0840961	0	9	0.06925093	5.725575
		Lehmer Mean 43	895.7375801	0	9	0.03628887	5.756586
		Lehmer Mean 54	**896.3914876**	**0.44444444**	**1.5**	0.00273305	5.797785
		Median	**896**	0.11111111	3.5	**1.10105**	**4.063203**
		Mode 1	896	0.11111111	3.5		

Table 4: Results of Voting and Bootstrapping Techniques with Right Skewed Data Sets

		Measures of Location		Voting Approach		Bootstrapping Approach	
Outlier	Variable	Name	Value	Relative Frequency	Rank	Bias	Standard Error
No	Average SAT Score in the South of US	**Mid-range**	**942**	0	8	**1.64725**	**5.671588**
		Arithmetic Mean	946	0	8	0.1052563	17.679387
		Geometric Mean	943.3527746	0	8	0.2679084	17.634348
		Harmonic Mean	940.7178277	0.0625	3	0.4279735	17.555907
		Quadratic Mean	948.6473923	0	8	-0.0578578	17.689956
		Cubic Mean	951.2828219	0	8	-0.2192543	17.665655
		Quartic Mean	953.8944606	0	8	-0.3768021	17.606755
		Lehmer Mean 21	951.3021934	0	8	-0.2216635	17.710783
		Lehmer Mean 32	956.5756657	0	8	-0.5440893	17.646287
		Lehmer Mean 43	961.7724752	0	8	-0.8533661	17.485437
		Lehmer Mean 54	966.8473521	0.4375	2	-1.1414849	17.231053
		Median	**920.5**	**0.5**	**1**	18.912	43.023644
Left	Average SAT Score in the Midwest of US	Mid-range	994.5	0.166666667	2	17.47055	28.92327
		Arithmetic Mean	1043.75	0	9.5	-0.154075	17.14268
		Geometric Mean	1041.975939	0	9.5	-0.00564466	17.94975
		Harmonic Mean	1040.076665	0.083333333	4.5	0.174434324	18.81246
		Quadratic Mean	1045.40642	0.083333333	4.5	-0.27610085	16.3918
		Cubic Mean	1046.953069	0	9.5	-0.3763783	15.69667
		Quartic Mean	1048.397875	0.083333333	4.5	-0.45895258	15.05607
		Lehmer Mean 21	1047.065469	0	9.5	-0.3976139	15.65869
		Lehmer Mean 32	1050.053236	0.083333333	4.5	-0.57546068	14.3611
		Lehmer Mean 43	1052.744264	0	9.5	-0.70392866	13.24513
		Lehmer Mean 54	**1055.169095**	**0.5**	**1**	-0.79637226	12.30112
		Median	**1055**	0	9.5	**1.6187**	**11.79372**
	Product rating (text only)	Mid-range	5.5	0.258064516	3	0.0566	0.5040554
		Arithmetic Mean	6.129032258	0	9	-0.0027129	0.2522464
		Geometric Mean	5.916901023	0	9	0.003511694	0.3055727
		Harmonic Mean	5.620143885	0	9	0.021484616	0.399955
		Quadratic Mean	6.288750838	0	9	-0.00589572	0.2263125
		Cubic Mean	6.417680446	0	9	-0.00864751	0.2157177
		Quartic Mean	6.528272069	0	9	-0.01171281	0.2137521
		Lehmer Mean 21	**6.452631579**	0	9	**-0.00869626**	**0.2103675**
		Lehmer Mean 32	6.683523654	0	9	-0.01372884	0.2137964
		Lehmer Mean 43	6.871613376	0	9	-0.02066231	0.2330668
		Lehmer Mean 54	7.039569495	0.096774194	4	-0.03008538	0.260029
		Median	6	0.290322581	2	0.291	0.4562234
		Mode 1	**7**	**0.35483871**	**1**		
Right	CO2 Emission in Central & South America	Mid-range	5.9367	0.026315789	9.5	-0.14725453	0.6115761
		Arithmetic Mean	3.633978947	0.052631579	5.5	-0.00157811	0.4468154
		Geometric Mean	**2.674506248**	0.026315789	9.5	**0.020325045**	**0.3663471**
		Harmonic Mean	**1.766090944**	**0.394736842**	**1**	0.072509446	0.3856605
		Quadratic Mean	4.573762859	0.157894737	2	-0.03272961	0.5550814
		Cubic Mean	5.424432809	0.052631579	5.5	-0.07556351	0.656362
		Quartic Mean	6.15918491	0.026315789	9.5	-0.12327476	0.7387933
		Lehmer Mean 21	5.756584448	0	12	-0.07097269	0.7566494
		Lehmer Mean 32	7.629844997	0.052631579	5.5	-0.19250977	1.0323019
		Lehmer Mean 43	9.016333521	0.026315789	9.5	-0.32427606	1.1823758
		Lehmer Mean 54	9.955475342	0.052631579	5.5	-0.42008742	1.2496212
		Median	2.539485	0.131578947	3	0.229180628	0.62949
	Time Vocal Praise Dog Interacts (Sec)	Mid-range	149	0	9	-37.88755	55.18881
		Arithmetic Mean	67.57142857	0.142857143	4	-0.4875143	35.59031
		Geometric Mean	28.89877654	0.142857143	4	3.6416773	17.55744
		Harmonic Mean	**13.6494012**	**0.285714286**	**1.5**	**3.5951278**	**10.69443**
		Quadratic Mean	116.5252885	0	9	-13.5433551	53.19612
		Cubic Mean	154.6890212	0	9	-26.4350304	65.6393
		Quartic Mean	180.937603	0	9	-35.8881043	74.02196
		Lehmer Mean 21	200.9450317	0	9	-39.4225508	86.46546
		Lehmer Mean 32	272.6076888	0	9	-70.863769	106.83625
		Lehmer Mean 43	289.5588584	0	9	-78.0641308	110.68672
		Lehmer Mean 54	293.0445885	0.142857143	4	-79.0805997	110.88975
		Median	**25**	**0.285714286**	**1.5**	10.4154	31.84487
	Reaction Time (cell phone group)	Mid-range	708	0.0625	5	45.9121	66.6744
		Arithmetic Mean	585.1875	0	12.5	0.0739625	15.37451
		Geometric Mean	579.5073595	0.03125	9.5	0.2365095	13.68614
		Harmonic Mean	**574.5255901**	0.0625	5	**0.3486053**	**12.59238**
		Quadratic Mean	591.8020678	0.03125	9.5	-0.2007227	17.82891
		Cubic Mean	599.6288467	0.0625	5	-0.6956964	21.17556
		Quartic Mean	608.9532243	0.03125	9.5	-1.5758379	25.44598
		Lehmer Mean 21	598.4914023	0	12.5	-0.4641764	20.59392
		Lehmer Mean 32	615.5943086	0.09375	2	-1.6445238	28.78975
		Lehmer Mean 43	637.8053894	0.0625	5	-4.1478189	40.06857
		Lehmer Mean 54	666.2250018	0.0625	5	-8.9445676	54.00851
		Median	569	0.03125	9.5	1.9049	13.57787
		Mode 1	**554**	**0.46875**	**1**		
Both	Female Students' Height (inches)	Mid-range	74	0.053435115	5	-2.58E+00	3.8944077
		Arithmetic Mean	65.38549618	0	10	-4.82E-04	0.206635
		Geometric Mean	65.30325897	0	10	-2.15E-04	0.1985787
		Harmonic Mean	**65.2246935**	0	10	**3.05E-05**	**0.1929753**
		Quadratic Mean	65.47235883	0.019083969	6	-7.97E-04	0.2180091
		Cubic Mean	65.56504596	0	10	-1.20E-03	0.2338589
		Quartic Mean	65.66506405	0	10	-1.77E-03	0.2556733
		Lehmer Mean 21	65.55933687	0	10	-1.11E-03	0.2308345
		Lehmer Mean 32	65.75081404	0	10	-1.99E-03	0.2705669
		Lehmer Mean 43	65.96603471	0.129770992	3	-3.40E-03	0.3320159
		Lehmer Mean 54	66.21306387	0.267175573	2	-5.89E-03	0.4224066
		Median	65	0.125954198	4	1.04E-01	0.2599695
		Mode 1	**64**	**0.419847328**	**1**		

Table 5: Summary Results of the Most Representative and Efficient Averages with Data Sets

Nature of the Data	Variable	Outlier	Voting Approach: Measure of location	Voting Approach: Value	Voting Approach: Standard Error	Voting Approach: Rank	Bootstrapping Approach: Measure of Location	Bootstrapping Approach: Value	Bootstrapping Approach: Bias	Bootstrapping Approach: Standard Error
Symmetric	Palmer Pinguins: Flipper Length of Chinstrap Group (in mm)	No	Lehmer Mean 54	196.8427	0.45588	1	Quadratic Mean	195.9514	0.0055	0.8530
	Reaction Time (no cell phone group)		Mode 1	485	0.375	1	Mid-range	537	0.1434	5.5530
	Male Students' Height (inches)	Left	Mid-range	70	0.46154	1.5	Arithmetic Mean	70.9316	0.0035	0.2629
			Mode 1	70	0.46154	1.5				
	Youth Unemployment Rate in EU Countries	Right	Mode 1	7	0.35714	1	Harmonic Mean	9.2459	0.0520	0.6952
	Palmer Pinguins: Flipper Length of Chinstrap Group (in mm)	Both	Mid-range	191	0.44371	1	Geometric Mean	189.8419	0.0020	0.5395
Left Skewed	Sugar Content in Children (gram)	No	Hamonic Mean	4.735870977	0.3	1	Lehmer Mean 54	12.6019	-0.1423	0.6811
	CO2 Emission in Europe		Hamonic Mean	5.719252419	0.38709677	1	Quadratic Mean	7.4400	-0.0080	0.4431
	Participation in SAT Exam by Medium Group		Mid-range	934.5	0.44444444	1.5	Mid-range	934.5	-1.1957	6.6666
			Hamonic Mean	929.0695546	0.44444444	1.5				
	Time Petting Dog interacts (sec.)	Left	Mid-range	205	0.28571429	1.5	Lehmer Mean 54	266.4935	-2.7049	12.9787
			Lehmer Mean 54	266.4934769	0.28571429	1.5				
	Movie Attendance in a year	Right	Hamonic Mean	4.338245421	0.3	1.5	Harmonic Mean	4.3382	0.7955	2.4159
			Mode 1	12	0.3	1.5				
	Participation in SAT Exam by High Group	Both	Hamonic Mean	893.1246379	0.44444444	1.5	Median	896	1.10105	4.0632
			Lehmer Mean 54	896.3914876	0.44444444	1.5				
Right Skewed	Average SAT Score in the South of US	No	Median	920.5	0.5	1	Mid-range	942	1.6473	5.672
	Average SAT Score in the Midwest of US	Left	Lehmer Mean 54	1055.169095	0.5	1	Median	1055	1.6187	11.7937
	Product rating (text only)		Mode 1	7	0.35483871	1	Lehmer Mean 21	6.4526	-0.0087	0.2104
	CO2 Emission in Central & South America	Right	Hamonic Mean	1.766090944	0.39473682	1	Geometric Mean	2.6745	0.0203	0.3663
	Time Vocal Praise Dog Interacts (Sec)		Hamonic Mean	13.6494012	0.28571426	1.5	Hamonic Mean	13.6494012	3.5951	10.6944
			Median	25	0.28571426	1.5				
	Reaction Time (cell phone group)		Mode 1	554	0.46875	1	Harmonic Mean	574.5255	0.3486	12.5924
	Female Students' Height (inches)	Both	Mode 1	64	0.41984738	1	Harmonic Mean	65.2246935	3.05 E-05	0.1930

Consequently, in view of the above findings, every data set needs to be allowed to choose its most representative average and most importantly, the most efficient average as its representative. The idea of using either the arithmetic mean or the median as often being said (Dor and Zwick, 1999; Julious and Debarnot, 2000) may not be truly representative as can be seen from the results obtained. Even when data sets are symmetric in the data sets used, the most efficient average is not the arithmetic mean. The arithmetic mean is the most efficient average only when the data set is symmetric and have outlier in the left direction. Similarly, the median is the most efficient average when data set is left skewed and have outlier(s) in both directions, and when right skewed and have outlier in the left direction. The emergency of other averages as most efficient average is a strong indication that there is need for caution in choosing or emphasizing a particular average as a representative of a data set (Jacquier et al, 2003). Every data needs to be freely allowed to choose it best representative by itself rather than specifying a particular one to avoid lying with statistics (Fleming and Wallace, 1986; Curto, 2022).

Conclusion

Numeric univariate data sets do exhibit different characteristics often summarized by averages. These characteristics change as the nature of the data sets changes, living a challenge of which average is be used and considered as best representative of the data set. This research has adopted the voting technique to choosing the most representative data sets and thereby provides solution to the challenge of the challenge of non-existence and lack of uniqueness of the mode, and further utilized the bootstrapping technique to choosing the most efficient average. The research also emphasized and advocated for the use of both techniques to select its best average in terms of representativeness and efficiency as they provide better opportunity for the averages to interact with the data set and compete with one another to be the best. Based on the eighteen (18) data sets used in this study, ranging from symmetric to asymmetric, with and without outliers, results clearly reveal that the most representative average may not necessarily be the mode but could be any of mid-range, median, Lehmer mean and harmonic mean, and that the most efficient average could be any of harmonic mean, geometric mean, arithmetic mean, quadratic mean, Lehmer mean, mid-range and median. Consequently, the study suggests that every dataset needs to be allowed to choose its most representative and efficient averages; and with these findings, caution is needed on the frequent use of the any averages as a representative of a data set without verification.

References

Ajiboye, A. S., Adejumo, T. J. and Ayinde, K. (2017): A Study on Sensitivity and Robustness of Matched-Pairs Inferential Test Statistics to Outliers. FUTA Journal of Research in Sciences, 15, 203- 208.

Alao, A. N., Ayinde, K., Solomon, G. S. (2019). A Comparative Study on Sensitivity of Multivariate Tests of Normality to Outliers, ASM Sc. J., 12(5), 65-71.

https://www.akademisains.gov.my/asmsj/?mdocs-file=4181

Andrew, G., Jonathan N. K. and Francis, T. (2002). The Mathematics and Statistics of Voting Power, *Statistical Science*, 17(4), 420 – 435.

Awde, A, Diss M., Kamwa, E., Rolland, J. Y., Tlidi, A. (2023). Social unacceptability for simple voting procedures **(2023)**, In Kurz, S., Maaser, N., and Mayer, A.

(editors), Advances in Collective Decision Making: Interdisciplinary Perspectives for the 21st Century. Springer, Berlin. [DOI: 10.1007/978-3-031-21696-1_3]

Ayinde, K., Daniel, J., Adepetun, A. and Ewemooje, O. S. (2023). Moving Block Bootstrap with Better elements' Representation for Univariate Time Series Data. Reliability: Theory & Applications, 3(74),18, 671-688. doi.org/10.24412/1932-2321-2023-374-671-688

Bickel P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*. **9** (6): 1196–1217. doi:10.1214/aos/1176345637

Brase, C. H., Brase, C. P., Dolor, J., and Seibert, J. (2023). Understanding Basic Statistics. 9th Edition, Cengage Publisher.

Bullen, P. S. (1987). Handbook of means and their inequalities. Springer.

Casella, G. and Berger, R.L. (2002) Statistical Inference. 2nd Edition, Duxbury Press, Pacific Grove.

Crump, K. S. (1998). On summarizing group exposures in risk assessment: is an arithmetic mean or a geometric mean more appropriate? Risk Anal., 18(3):293-7. doi: 10.1111/j.1539-6924.1998.tb01296.x. PMID: 9664725.

Curto, J. D. (2022). Averages: There is still something to learn. *Computational Economics*. 60(2), 755-779.

De Carvalho, M. (2016). Mean, what do you Mean?. *The American Statistician*, 70 (3), 764–776. DIO= https//doi:10.1080/00031305.2016.1148632.

DiCiccio, T. and Efron, B. (1992). More accurate confidence intervals in exponential families. *Biometrika*. **79** (2):231–245. doi:10.2307/2336835.

Diss, M. and Merlin, V. (2021). Introduction. In: Diss, M., Merlin, V. (eds) Evaluating Voting Systems with Probability Models. Studies in Choice and Welfare, 1-12. Springer, Cham. https://doi.org/10.1007/978-3-030-48598-6_1

Dor, D., and Zwick, U. (1999). Selecting the median. *SIAM Journal on Computing*, *28*(5), 1722-1758. https://doi.org/10.1137/S0097539795288611

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*. 7(1), 1-26. doi:10.1214/aos/1176344552

Efron, B (2003). Second thoughts on the bootstrap. *Statistical Science*. **18** (2): 135–140.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC. ISBN 0-412-04231-2

Emovon, I. and Okechukwu, O. M. (2017). Comparative Study of the use of Arithmetic Mean and Geometric Mean for Data Aggregation in FMEA Analysis. *Covenant Journal of Engineering Technology*, 1(2). Retrieved from *https://journals.covenantuniversity.edu.ng/index.php/cjet/article/view/704*

Fleming, P. J. and Wallace, J. J. (1986). How not to lie with statistics: The correct way to summarize benchmark results. *Communications of the ACM* 29 (3) :218–221. doi:https://doi.org/10.1145/5666.5673.

Goodchild, S. (1988). School Pupils' Understanding of Average. *Teaching Statistics*, 10, 77-81.

Good, P. (2006) Resampling Methods. 3rd Ed. Birkhauser.

Halley, R. M. (2004). Measures of Central Tendency, Location, and Dispersion in Salary Survey Research. *Compensation & Benefits Review*, *36*(5), 39-52.

https://doi.org/10.1177/0886368704268598

Hinkle, D. E., Wiersma, W., and Jurs, S. G. (2003). Applied Statistics for the Behavioral Sciences. Boston, MA: Houghton Mifflin Company.

Horowitz, J. L. (2019). Bootstrap methods in econometrics. *Annual Review of Economics*. 11: 193–224. arXiv:1809.04016. doi:10.1146/annurev-economics-080218-02565

Jacquier, E., Kane, A. and Marcus, A.J. (2003) Geometric or Arithmetic Mean: *A Reconsideration*, *Financial Analysts Journal*, 59(6), 46-53, DOI: 10.2469/faj.v59.n6.2574

Julious, S.A. and Debarnot, C. A. M. (2000). Why Are Pharmacokinetic Data Summarized By Arithmetic Means?, *Journal of Biopharmaceutical Statistics*, 10:1, 55-71, DOI: 10.1081/BIP-100101013

Kennedy, C. and Stanley, J. (2009). On "Averages". Mind, 118(471): 583-646.

Kleiner, A; Talwalkar, A; Sarkar, P; and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*. **76** (4): 795–816

Kun A. and Jiang, M. (2010). Voting-Averaged Combination Method for Regressor Ensemble *Advanced Intelligent Computing Theories and Applications*, 6215, ISBN: 978-3-642-14921-4.

Mokros, J., and Russell, S.J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, 26, 20-39.

Mokros, J., and Russell, S. J. (1996). What do children understand about average? *Teaching Children Mathematics*, 2, 360-364.

Mukhopadhyay,S. Amlan J., Das,A. J., Basu, A., Chatterjee, A. and Bhattacharya, S. (2021). Does the generalized mean have the potential to control outliers?, Communications in Statistics - Theory and Methods, 50:8, 1709-1727, DOI: 10.1080/03610926.2019.1652320

Muthuvalu, M. S., Asirvadam, V. S., and Mashadov,G. (2015). Performance analysis of Arithmetic Mean method in determining peak junction temperature of semiconductor device, *Ain Shams Engineering Journal*, 6(4), 1203-1210, https://doi.org/10.1016/j.asej.2015.04.007.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*. 9: 1187–1195.

Shoemaker, Owen J. and Pathak, P. K. (2001). "The sequential bootstrap: a comparison with regular bootstrap". *Communications in Statistics - Theory and Methods*. **30** (8–9): 1661–1674. doi:10.1081/STA-100105691

Takacs, P. and Bourrat, P. (2022). The arithmetic mean of what? A Cautionary Tale about the Use of the Geometric Mean as a Measure of Fitness. *Biol Philos* **37**, 12 (2022). https://doi.org/10.1007/s10539-022-09843-4

Vogel, R.M. (2020). The geometric mean? Communications in Statistics - Theory and Methods, 51:1, 82-94, DOI: 10.1080/03610926.2020.1743313