Vol.3.Issue.2.2015

http://www.bomsr.com

RESEARCH ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

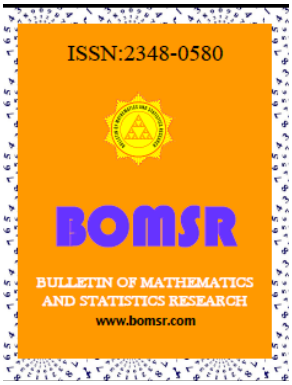
ISSN:2348-0580

SUMMARIZING UNOBTRUSIVE DATA WITH EXTREME VALUES USING MULTIPLE MEANS

JOHN KOMLA COKER AYIMAH

Lecturer (Ho Polytechnic) & PhD Student (UG-Legon), Department of Statistics

Email contact: johncokeroriginal@gmail.com



ABSTRACT

The arithmetic mean is one single most-used measure for describing data, perhaps due to the theoretical advantage it has over other measures. Despite this, a major setback to obtaining optimal mean is when the data is inundated with extreme values. When this happens, the importance of the mean is missed. The mean is termed upward bias in such times, since it tends to represent high data points only. Meanwhile, when data points are ordered, the mean corresponds to the centre of gravity of observations with the confidence interval forming one homogeneous set, being values that the mean can correctly represent. So, beside the mean, there could be two more sets for which unique means will correctly represent. Ultimately, these multiple means will better represent the data than what a common mean would have done. Therefore, the objective of this study is to propose theories to guide regrouping of such data in order to obtain multiple centres which best describe the large variability in the data. The main theories proposed are based on a lemma and two subsequent corollaries. In the end, eight cohorts of data were used to test the theory of multiple means. It was found that the multiple-mean theory would be a good alternative in describing data with extreme values. This is particularly because the theory has an advantage over the Chebyshev's theorem, used when the distribution is not mound shaped; and the Sturgis formula, used when regrouping widely spread out data. In all except one cohort, the percentage of observations accounted for by the three centres is greater than the 75% a Chebyshev's theorem would have produced for observations within two standard deviations, and 100% of all observations are accounted for eventually. This makes the multiple-mean theory a better tool in describing non-mound shape data. The other advantage is that, unlike the Sturgis formula, the multiple-mean theory is likely to have fewer and fixed number of groups for the same random samples.

Keywords: Arithmetic Mean; Extreme Data Points; Outliers; Summary Statistics

INTRODUCTION

Optimal estimates minimise the chance of having to deal with high inaccuracies in projections. Often, values of a random variable may contain two or more groups that are not originally associated with a pre-determined class of variables – hence the classification of such values cannot be conspicuously done. Unobtrusive data, many times than not, do have unsuspecting groups, being the reason why data points vary from one another. Naturally, if data points emanate from the same or similar characteristic, there would be no need to talk about extreme values, except for obvious errors in data collection are detected. In effect, the summary statistics for homogenous data converge to a single value. This value in fact, posses the summary information unique to those observations. Theoretically, the arithmetic mean common to the observations normally plays this role. The implication is that, when a set of observations contains extreme values, a common mean will not be optimal in adequately describing the set. In such cases, the data may be displaying unique characteristics if further explored, could better inform the user. This study explores a new approach of computing multiple means – where each mean corresponds to a unique group (characteristic) – to provide single representatives for each "constituent" in the data. Just as "an average is a value that is typical, or representative, of a set of data", according to Spiegel and Stephens (2011), even so, will these multiple means represent defined groups within a data.

Many areas of life do generate unobtrusive data; data whose existence is not likely to be influenced by manipulations from the onset, or for which a researcher has no chance of influencing in anyway. Of course what is critical in such cases is to be sure that the data are error-free. Many areas in life do generate data that could be termed as unobtrusive. Examples are:

1. Metric data on age of an unclassified subjects;
2. Metric data on the scores obtained after an examination;
3. Amount of rainfall for a particular month measured over a period;
4. Varied prices of a certain commodity across a number of outlets; *et cetera*.

In all these areas, one may not have a prior knowledge of the kind of characteristics inherent in the population. Interestingly, the dispersion of the data points does have information about possible groupings in the data whose characteristic become evident anyway. So, in all those areas mentioned above, extreme values may hold valuable information about particular characteristics, within the data, that a researcher needs to know. Therefore, discarding the extreme values may amount to discarding valuable information too. In fact, Keller and Warrack (2003) proposed that we do no such thing to an outlier when it's found out to be a legitimate member of the data. Essentially, the outlier in this case is just an extreme value with a valuable piece of information.

It is no doubt that the arithmetic mean is one most-used summary statistic frequently required to pass-on first-hand information about data. Indeed, Gordor and Howard (2006) argued that "the mean may be considered as a value each observation in a sample would have if the sum of all the values were shared out equally among the observations": but, in the face of extreme data points, using a common mean for estimates would be very misleading – especially when measures of variability are not considered carefully in those instances. In addition, a common mean may obstruct efforts aimed at unmasking possible new characteristics that the data may be portraying. However, finding unique groups, using clearly defined theories, would allow for calculating multiple means to correct for what would have otherwise be missed if a common mean was used.

Background of the Study

The act of summarising data, in particular, can pose serious challenges; especially when certain characteristics, like extreme values, pop up during data collection. The problem becomes even daunting when an extreme value is not actually an outlier that should be discarded.

Ramachandran and Tsokos (2009) confirm that it is often impossible to say whether an outlier is really an extreme value within a skewed population or whether it represents a value drawn from a different population. The decision as to whether to treat a value as an outlier (to be discarded) or just an extreme value and hence going ahead to deal with its impact on the final result, can trigger some discomforts for many statistics practitioners. If the interest is to identify some discordant observations, to be subsequently removed from the data set, then a number of methods are available, according to Solak (2009). However, many real life situations, as those mentioned above, do generate data whose values may not necessarily be a "contaminant", "discordant" nor an outlier; they may be legitimate members of the set just that they are extreme measurements. Meanwhile, the presence of these values unduly influence the most sought-for measure used in representing data, the arithmetic mean. Clearly, a mean obtained from such circumstances will not provide a good description of the set of values. When this happens, the margin of error, $M_E=kS_e$ is large hence very extreme values fall outside the confidence interval $CI = \bar{x} \pm kS_e$.

In dealing with cases such as this, a number of robust tools are used as pointed out by Ramachandran and Tsokos (2009). They are; to transform the data by taking the natural logarithm, or, to perform the analysis twice, with and without "outlier", and report both results. The former suggests that the extreme values are incorporated in the entire data set, which in fact, may conceal certain unique characteristic of those values; the application of the latter also stems from the fact that the so-called "outlier" is first identified – which also may not be easy, according to Ramachandran and Tsokos (2009). So, a critical question is: how do we deal with such data, so that in addition to the common mean we have other parameters which adequately describe the data set? Wouldn't such reduce the "injures" to estimates the mean would have caused if it was used singularly without accounting for possible huge variances?

The expected value of a population is often influenced by large variance in data. For a random variable, X, with finite expectation, $E(X) = \mu$, and variance, $Var(X)$, the Chebyshev's inequality suggests that, for any real number $\epsilon > 0$, the probability,

$$P(|X - \mu| < \epsilon) \geq 1 - \frac{Var(X)}{\epsilon^2} \qquad (1)$$

Hence, when the variance of the random variable becomes larger than ϵ^2, the Chebyshev's inequality provides trivial estimate of the probability of the event, $P(|X - \mu| < \epsilon)$, according to Nsowah-Nuamah (1999). In such cases, the probability estimate becomes meaningless and serves no purpose doing so. The overall essence of expression (1) is that the arithmetic mean converges to the expected value of the population; the arithmetic mean represents the data, hence the Khinchin's law of large numbers (continuous case). The law states that, for a sequence of independent and identically distributed continuous random variables, $X_1, X_2, \ldots, X_n$, with finite common mean, μ, and finite common variance, σ^2; and any real number $\epsilon > 0$ then

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| < \epsilon) = 1 \qquad (2)$$

Thus, equation (2) sought to convey that, by averaging an increasingly large number of observations (at least 30) of the value of a quantity, we can obtain more accurate measures of the expectation of that quantity (Nsowah-Nuamah, 1999). The combine effect of (1) and (2) is that, for optimal mean value, the variance shouldn't be large and the number of samples must be large as well. These notwithstanding, room must also be made for those other many real life situations where one may not have such large observations of the value of a quantity but yet a mean has to be found. Also, other real life situations may emerge where large number of observations is acquired

but the data is inundated with very extreme values, which may jeopardise equation (2). Obviously, these situations may require other skills in order to have even more accurate results.

In summary, describing data; especially when extreme values are involved can poses serious challenge but yet necessary when optimal discovery is to be made. Finding and applying more robust ways could ease the challenge thereby providing some headway. When data is skewed, there are unique groups whose values are identical and hence can be put together. Accordingly, finding unique centres to represent each group, which includes the common mean, could give a better explanation of the data as well as provide some clue about the various constituents in the data. So the common mean in this case as one of the boundaries, and then other means are determined so that beside the least and largest value, these account for the maximum number of the data points, and also clearly demarcates the data. These centres would then be used as representatives for the various demarcations in further understanding the distribution of the data.

The Problem

The arithmetic mean is often required due to its theoretical advantage over other measures of central tendency. However, the importance of a common mean is lost when describing a skewed univariate data. When data exhibit long tail(s), then there exist hidden characteristics within the data set whose values are responsible for this phenomenon. Therefore, summarising data into one single common mean does not allow for these characteristics to be understood, since it will not be a representative of the entire set of observations. This situation rubs the data mining process of information that would have help in subsequent endeavours. Also, situations such as this do not frequently merit the use of the empirical rule to obtained proportion of values that fall into various standard deviations from the mean. Unfortunately, the Chebyshev's theorem can only account for 89% of the values falling within three standard deviations from the mean; 11% unaccounted for. In fact, even with the Chebyshev's theorem, the 75% of values falling within two standard deviations results mostly in very wide range which is too wide hence has the tendency of placing very extreme values in the same group. This doesn't promote group analysis of data. An ideal situation is one that can account for maximum number of data points.

Objective of the Study

The general objective of the study is to promote a re-grouping method when performing a univariate analyses so that multiple means are computed for data inundated with extreme values. Specifically,

1. To put together theories to guide the regrouping of data for multiple centres to be computed;
2. To test the theories using empirical data.

Delimitation and Significance of the Study

The study is not about putting together theory to identify outliers; it's strictly to promote another way of regrouping data so that statistics computed better describe the data. It is believed that these additional pieces of information about the groups within data could aid good planning and decision making practices. Again, the boundaries to be determined could serve as bins in drawing histograms for studying the distribution of data. This method is also expected to provide a more compact and fewer groups than would have been if the Sturgis formula is use.

Review of Relevant Literature

The Common Mean as an Estimate

One advantage of the mean, which emphasises its importance, is perhaps the fact that it's "clear, precise and corresponds to the centre of gravity of the observations" (Nortey and Afrim 2013). Lots of literature on the importance of the mean exist. According to Kirk (1999), the mean has a number of mathematical properties that make it the preferred measure; it is expected to have sampling stability and mathematically tractable: this implies that, means of repeated samples are

suppose to be similar and verifiable respectively– It is therefore not recommended for markedly skewed distributions. Clarke and Cooke (1998) seem to explain why this is so. They explained that the mean uses the actual values of all the observations; particularly useful in detecting small differences between sets of observations. The mean has been found a convenient measure to use in the theory of statistics, but can be misleading if the distribution is not symmetrical or almost symmetric, they added. In view of this, a better approach, having found the mean of a set of observations, is to measure variability by seeing how closely the individual observations cluster round the mean (Clarke and Cooke, 1998). If we use a random sample to estimate parameter such as the mean, how good is the estimator? We investigate the sampling distribution of the sample mean to find unbiased estimators (Clarke and Cooke, 1998).

On the other hand, if we want to use the mean to represent a data, we may want to find how good is the data represented? In the case of estimating a population parameter using a sample, we may find an estimator that is unbiased but leads to estimates that are very widely spread; such an estimator is of little use. For an unbiased estimator to be regarded "good", we require also that the variance of its sampling distribution – and hence its standard error – should be small (Clarke and Cooke, 1998). The same can be said about a common mean; it's of "little use" when the data from which the mean was computed is very widely spread. Consequently, just as the bias of an estimator occurs when a sample does not accurately represent the population from which the sample is taken, according to Ramachandran and Tsokos (2009), so will there be bias in information when the mean does not accurately represent the data from which it was computed.

There are two types of estimators as defined by Keller and Warrack (2003), the point estimator and the interval estimator. The point estimator seeks to draw inferences about the population parameter by estimating the value of an unknown parameter using single value or point; the interval estimator on the other hand, is an estimator that seeks to draw inferences about the population by estimating the value of an unknown parameter using an interval. Therefore, both point and interval estimators draw inference about the population. Of what use is an estimate if it cannot accurately help the course of drawing inference? In this vein, a common mean is a point estimate which ought to represent the data much better.

The estimator satisfies the consistency property if the sample estimator has high probability of being close to the population value, θ, for large sample size (Ramachandran and Tsokos, 2009). The estimator, $\hat{\theta}$, is said to be a consistent estimator of θ if, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| > \varepsilon\right) = 0 \qquad (3)$$

When this happens, $\hat{\theta}$ is said to converge in probability to θ. That is, the sample estimator should have a high probability of being close to the population value. Therefore, a sufficient condition for consistency is that

$$\lim_{n\to\infty} Var\left(\hat{\theta}\right) = 0 \qquad (4)$$

So in estimation, one seeks to satisfy an efficient condition where the variance of the estimator and its estimates are small as possible, approaching zero. A drawback to achieving this condition is when samples are not representative enough; either certain clusters or certain strata are not well represented. This means that obtaining optimal estimates is directly linked to meeting the requirement of good representation. The clusters in a data equally deserve this treatment; they need to also satisfy the requirement of minimal variance. For a random sample, $x_{1j}, x_{2j}, \ldots, x_{nj}$,

from n different experimental units, the arithmetic mean (the mean) for the variable, X_j, is calculated as in equation (5).

$$\bar{X}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{5}$$

Equation (5) projects very interesting scenes about the random sample which needs to be brought to the fore. The right hand side of the equation is the ratio, $\sum_{i=1}^{n} x_{ij} : n$ indicating the sharing of the sum into n equal portions, as Gordor and Howard (2006) noted. So, for the mean to be a good representative of the samples, then any unit, x_{ij}, should be approximately equal to the ratio of the sum of all observations to the number of observations,

$$x_{i*} \approx \sum_{i=1}^{n} x_{i*} : n \tag{6}$$

Ultimately, what this means is that any data point is a likely candidate for the mean. A major setback to achieving the condition stated in equation (6) is the variability in the data collected. Situations such as this obviously require more than just finding the mean; finding other ways of reducing the variability in the data would greatly improve estimates. Practically, quantitative data generated in everyday activities are bound to have extreme values, values that are mostly responsible for bringing about such high variability in the data. Because of this seemingly weakness of the mean, a number of statistical measures are built to help improve on the accuracy of estimates.

The Mean and Measures of Variability

Apparently, in an attempt to overcome the shortcomings that are likely to be associated with the mean, and to also ensure a certain level of accuracy, a number of theoretical measures were developed to mitigate these effects. One of the measures that readily come to mind is the standard deviation. The standard deviation gives an estimated measure of how far the data points are spread away from the mean. Assume a random sample, $x_{1j}, x_{2j}, \dots, x_{nj}$ taken for some variable X_j; assume also that the mean and standard deviations are respectively, $\bar{X}_j$ and S_j. Then ideally the standard deviation

$$S_* \approx 0 \tag{7}$$

which result suggests that, in an ideal situation, the standard deviation should be as close to zero as possible; in fact the closer it is to zero, the better. This appears to be a critical condition for having a more accurate mean for projections. What this implies is that, the data in this case should be homogenous; means calculated from homogenous sets are optimal.

In order to get over the huge challenge of making accurate projections using the mean, the distribution of the sample statistic was introduced where the mean is calculated from repeated samples. The resulting distribution of the many means is approximately normally distributed for large sample size, that is $n \geq 30$, according to the central limit theorem. However, the problem may not be about how many repeated samples to take but rather how were the extreme values in the samples dealt with appropriately? It should not be forgotten that, the practitioner, in all these instances, seeks an approximation for the population parameter or a statistic from the sample. A population has a distribution called a population distribution, which is usually unknown. What has become increasingly known is that the sampling distribution, which is what we frequently know, is used many times as an estimate for the population distribution when samples are representative enough (Ramachandran and Tsokos, 2009). Clearly, the problem may not be how repeated samples

estimate a population mean, but how the statistic obtained from a sample is optimal in describing the data, hence the population.

The Weighted Mean

Sometimes we associate with data points certain weighting factors, depending on the significance or importance attached to the random samples (Spiegel and Stephens, 2011). This is done when the weighted mean is preferred. The implication is that since the mean has an upward bias, according Nortey and Afrim (2013), associating larger values with high weights will also result in additional problem peculiar to the weighted mean. To avoid this, a critical point to note in the construction of the weighted mean is to adopt methods that reduce the variances in the numbers by assigning weights that take care of the variability. Recall equation (5), this time for some fixed i^{th} experiment for j replicates: let the arithmetic mean be

$$\mu_i = \frac{\sum_{j=1}^{p} x_{ij}}{p} \tag{8}$$

Then assuming that some $w_j's$ are the weights assigned to the experimental units within experiment i of the j replicates. The weighted mean is calculated as

$$(\mu_w)_i = \sum_{j=1}^{p} \frac{w_j x_{ij}}{w_j} \tag{9}$$

Then the following four points are true of equations (8) and (9).

1. If the weights in equation (9) are such that the ratios, $\frac{w_j}{\sum_{j=1}^{p} w_j}$, are identical and approximately equal to $\frac{1}{p}$ in equation (8), then the results of the two expressions will be the same. Here, the mean and the weighted mean share the same problems; notably amongst them is the undue influence of extremes values.
2. In the light of point (1) above, the weighted mean would be a good estimate of the $x_{ij}'s$ only if it satisfies the condition of having the standard deviation as small as possible.
3. Following points (1) and (2) above, if there are some larger values of $x_{ij}'s$ and some $w_j's$ greater than one, the result will be unduly or overly influenced hence “exaggerates”.
4. Clearly, both equations are “upward bias” when data records high variance, a situation which tends to overestimate the results towards high $x_{ij}'s$ values. Therefore, in a typical data, both methods tend to represent higher values only.

Approximations to the Mean

When the variance of the values of a random variable is small, then the values are identical and the mean is very likely to be adequately approximated by other statistics, like the median, mode, harmonic mean and geometric mean. This is a unique condition under which the mean accurately represents the data. Situations such as this do not pose difficulties when describing the data set. All other situations require extra care when handling the data. Indeed in all other situations, Nortey and Afrim (2013) argued that the mean tends to “give greater importance to larger values and less importance to smaller values, hence loses its representative character and thereby concealing facts leading to distorted conclusions”. Despite these, the mean can still be put to meaningful use so that its role in describing the data is not entirely lost.

Dealing with Extreme Values in Data

The problems associated with mean and weighted mean, for which caution is to be exercised, stem from unrepresentative samples – samples which include extreme values far from the mean. Extreme values in data could be due to either the presence of outliers or a skewed population (Ramachandran and Tsokos, 2009). Because outliers may be an indicative of novel phenomenon (Ramachandran and Tsokos, 2009), excluding them or eliminating them from the data may not be an appropriate option, because it is impossible to say whether an outlier is really an extreme value within a skewed population or whether it represents a value drawn from a different population. In the case of a skewed population, Ramachandran and Tsokos (2009) proposed two possible ways of dealing with extreme values – one way is to transform the data such as taking natural logarithm; so as to reduce the undue influence of the outlier, or to perform the analysis twice; with and without the "outliers", and report both results. A deduction from the second method proposed above, or perhaps an extension, is to identify major groupings in the data; analyse them independently and report the results accordingly. This is the motivation behind this study.

The problems of having good estimates when data contain extreme values have led to the development of empirical methods like the Jackknife and Bootstrap methods. Ramachandran and Tsokos (2009) mentioned the Jackknife method as "very useful when outliers are present in the data or the dispersion of the distribution is wide". At least so far, it is also known that not all "outliers" are indeed "outliers"; they may contain valuable information once they did not come about as a result of error in measurement. In fact Keller and Warrack (2003) proposed that in such situations "we do nothing to the outlier – as in dropping them – for it would be judged to be valid". Hence discarding it or "leaving out" an outlier in computing for statistic may take some information out too.

Also, the reason why the most-sought-for normality requirement is not met is the presence of extreme values in samples. Bootstrap methods are used in such times when it is obvious that the sampling distribution is not normal and/or the data is not normal. To have the best possible estimate of a population parameter, it is extremely important to equally have the best population distribution (Ramachandran and Tsokos, 2009). This is because good estimates and the probability of making the various type errors, in hypothesis testing, depend on estimating a good distribution for the population. Therefore, having a good description of the population is key to making inferences about the population.

Research Methodology

The development of methods to be used in this study is based on a lemma and two corollaries.

Lemma (the mean as a centre of gravity): Let, $x_{11}, x_{21}, \dots, x_{np}$, be random observations associated with some random variables, $X_1, X_2, \dots, X_p$. Then for some fixed observation, i, the mean, $\bar{x}_i$, is equivalent to the "centre of gravity".

Proof: Suppose that a system of particles, p, consists of masses, $m_1, m_2, \dots, m_p$, placed at some points with position vectors, $r_1, r_2, \dots, r_p$, corresponding to the masses. Then by definition (according to Turner *et al.,* 1986), their point of balance, the centre of mass or centre of gravity, relative to the origin (O), is computed as

$$\bar{r_o} = \frac{m_1 r_1 + m_2 r_2 + \cdots + m_p r_p}{m_1 + m_2 + \cdots + m_p} = \sum_{j=1}^{p} \frac{m_j r_{oj}}{m_j} \qquad (10)$$

Subsequently, assume that $x_{i1}, x_{i2}, \dots, x_{ij}$ are the observations of some fixed i^{th} experiment for j replicates: then the arithmetic mean is as equation (8)

$$\bar{x}_i = \frac{\sum_{j=1}^{p} x_{ij}}{p} \tag{11}$$

Now, assuming that some $w_j's$ are the weights assigned to the experimental units within experiment i of the j replicates. Then the weighted mean would be

$$(\bar{x}_w)_i = \sum_{j=1}^{p} \frac{w_j x_{ij}}{w_j} \tag{12}$$

Assume further that $\bar{x}_i = (\bar{x}_w)_i$ and that the ratios, $\frac{w_j}{\sum_{j=1}^{k} w_j}$, are identical and approximately equal to $\frac{1}{p}$. In effect, $\sum_{j=1}^{p} w_j = p$ and $w_1 = w_2 = \cdots = w_p = 1$. Hence, equation (11) becomes

$$\bar{x}_i = \sum_{j=1}^{p} \frac{w_j x_{ij}}{w_j} \tag{13}$$

Clearly, equations (10) and (13) are identical, given the fixed points respectively. Therefore, in general, the mean is the point of balance of the data when the weights are unity.

Corollary 1: Consider the mean of some independent random observations to be $\bar{x}$, and the mean of all observations falling within the confidence interval of the mean to be $\bar{x}_*$; then $\bar{x} = \bar{x}_*$.

Proof: Let the observations be, $x_1, x_2, \dots, x_n$ and ordered such that, $x_1 < x_2 < \cdots < x_n$. Then their point of balance is the mean $\bar{x}$. Hence the set of all observations estimating the centre of gravity, $\{\dots, x_{n-\delta}, \bar{x}_*, x_{n-\delta+2}, \dots\} \subset \{x_1, x_2, \dots, x_{n-\delta}, \bar{x}, x_{n-\delta+2}, \dots, x_{n-1}, x_n\}$, for some δ, in which $[x_{n-\delta}, x_{n-\delta+2}]$ is the confidence interval for the mean. Following the lemma, $\bar{x}_* = \bar{x}$

Corollary 2: Subsequent to corollary 1, for ordered observations, $x_1, x_2, \dots, x_n$, with extreme values, there can be two other unique centres, at each side of $\bar{x}$, within the observations relative to cluster of observations so that these centres are closest to the respective clusters.

Proof: Let the observations be, $x_1, x_2, \dots, x_n$ and ordered such that, $x_1 < x_2 < \cdots < x_n$. Then their point of balance is the mean $\bar{x}$. Hence the set of all observations estimating the centre of gravity, $\{\dots, x_{n-\delta}, \bar{x}_*, x_{n-\delta+2}, \dots\} \subset \{x_1, x_2, \dots, x_{n-\delta}, \bar{x}, x_{n-\delta+2}, \dots, x_{n-1}, x_n\}$, for some δ. Similarly, let the centre of all observations less than $x_{n-\delta}$, and those greater than $x_{n-\delta+2}$ be respectively, $\bar{x}_L, \bar{x}_R$ so that $\bar{x}_L < \bar{x} < \bar{x}_R$. Therefore, $\bar{x}_L$ and $\bar{x}_R$ are closest to the respective halves than $\bar{x}$.

Summarising Extreme Data Points

So far, evidences from literature suggest that working with data containing extreme values require some care. It is also evident that not all extreme values are indeed outliers to be discarded. Clearly, a common mean will represent the centre of the entire data but will not be enough to represent the extreme ends. Hence the common mean tends to be bias and misleads when used singularly in describing the data. Thus, a common mean is not optimal in such circumstances.

To ensure that information from data is not misleading, and also that maximum pieces of information are extracted for whatever purpose, extra works on the data are required. This study is an effort in this direction. The theory being proposed here is based on the lemma and corollaries in computing multiple means, for describing metric data which have extreme values that are not

necessarily outliers, or for which there is no evidence that the extreme values are really outliers to be discarded. In such instances, as would be seen, the Chebyshev's theorem seems not adequate; adopting the Sturgis formula for regrouping random samples may also result in having too many groupings. The lemma and corollaries above provide good alternative in this circumstance. The following laws are hereby deduced.

Law 1(condition for optimal mean): Let, $x_{1j,}$ $x_{2j,\ldots\ldots}$ $x_{N,j}$ be random values associated with some random variables, $x_{1,}$ $x_{2,\ldots\ldots}x_{,j}$ for which at least one average, A_{nj} and a measure of variability, S_{Dj}, are computed. Then at a minimum value of S_D *Min* (S_D) approaching zero, A_n'*S* will be identical for some *n* averages.

The outcome of law 1 has a direct consequence on the confidence interval, $\bar{x}_* \pm kS_*$, estimating a common mean, $\bar{x}_*$. If the common mean is correctly estimated by a confidence interval, which is the probability of predicting the mean, then it will not be difficult to know the likely members at the centre of that data. Hence understanding the data becomes an easy task.

Law 2(optimal mean): If conditions in law 1 above are met, and that the arithmetic mean, which is the common mean, $\bar{x}_j \in \{A_{nj}\}$, then the common mean is optimal and is a good one, otherwise the common mean is a "bad" one and the data contain at least one extreme value, and these values will not be contained in the confidence interval.

The extreme values in the data are the reason for the "bad" mean. Hence putting values that fall outside the confidence interval in separate groups, sandwiching the common mean, will allow for two more centres of gravity to be obtained conforming to corollary 2.

Law 3(the multiple mean law): Let $X_{1j},X_{2j},\ldots X_{Nj,}$ be random values from the population of random variables,$X_1,X_2,\ldots X_{j,}$ for which at least one average, A_n and a measure of variability,S_D are computed. Assume also that the condition for optimal mean is violated and so the common mean $\bar{x}_{j,}$, is a bad one. Then there exist at least one random value, $x_{ij} \notin \{\bar{x}_j \pm kS_*\}$, whose inclusion maximises, S_{DJ} so that $A_{1j} \neq A_{2j} \ldots \neq A_{nj,}$ for some *n* averages. Then there exist unique groups with centres, $\bar{x}_{Lj,}\bar{x}_{Rj,}$ to the left and right of the common mean respectively, which are the means of all values excluding the least and largest values.

Definition 1 (Extreme Value): Subsequent to law 3 above: for a random variable,*j* all values $x_{ij} \notin \{\bar{x}_j \pm kS_*\}$, constitute extreme values for which unique centres, $\bar{x}_{Lj,}\bar{x}_{Rj,}$ could be determined.

Definition 2 (boundary statistics): Subsequent to law 3 above: for a random variable,*j* there shall be five boundaries statistics; being the least value, X_{1j} the extreme mean to the left of the common mean, $\bar{x}_{Lj}$, the common mean, $\bar{x}_j$, the extreme mean to the right of the common mean, $\bar{x}_{Rj}$ as well

as the largest value, X_{Nj}: where $X_{1j},X_{2j},\ldots X_{Nj,}$ is the order statistics.

Ultimately, the boundary statistics could form the basis for understanding the distribution of the values hence for making inferences, interventions, *et cetera*. These are the methods adopted in this study.

Information on Data Required

Data required are solely to test the theory so they were purposively obtained; data with the high possibility of having extreme values so defined in this work. The data are metric and from primary source and so are unobtrusive. They are scores of students in eight academic courses from the records department of Ho Polytechnic.

Sampling and Sample Size

Four different sample sizes were defined for which two sets of data were obtained for each sample size. So in all, eight cohorts of data were used to test the theories. The sample sizes were defined as:

1. Small sample size, $N = 20$
2. Large sample size, $N > 30$
3. Larger sample size, $N > 100$
4. Very large sample size, $N > 200$

Variables in the Study

In view of the above definitions, there are eight variables in this study:

1. Cohort 1, small sample size
2. Cohort 2, small sample size
3. Cohort 3, large sample size
4. Cohort 4, large sample size
5. Cohort 5, larger sample size
6. Cohort 6, larger sample size
7. Cohort 7, very large sample size
8. Cohort 8, very large sample size

Data Analysis and Procedures

Data analysis was done in two sections; preliminary and further analysis. The preliminary analysis is to ascertain that the data has extreme values and that the common mean is not optimal by finding standard deviation(S_D), the mean$(\bar{x}_*)$, median(MD_*), harmonic mean(H_*) and the mode(M_*) using law 1. The confidence interval is then found and hence the boundary statistics determined. Finally, the distribution of the data is assessed by drawing the histogram with the boundary statistics serving as abscissa and the frequency as ordinate. The main software applications used were the Microsoft excel and the R.

Results, Discussions and Implications

The results are organised in accordance with the data analysis and procedures. Table 1 is the summary of the preliminary analysis whiles Table 2 summarises the results of the first part of the further analysis. The final aspect is the discussion on the histograms, which also depicts the distribution of the cohorts.

Table 1: Summary Statistics for Cohorts

Cohort	N	Variability (S_D)	Averages(A_n)			
			$\bar{x}_*$	MD_*	H_*	M_*
Cohort 1	20	11	70	75	68	75
Cohort 2	20	**5**	**72**	**74**	**72**	**74**
Cohort 3	39	11	65	64	63	61
Cohort 4	39	10	61	59	59	54
Cohort 5	112	11	70	71	66	80
Cohort 6	112	**7**	**67**	**68**	**67**	**68**
Cohort 7	425	11	64	64	62	57
Cohort 8	425	15	57	56	51	50

Source: Data analysis of test data

Law 1 appears to be violated in all the cohorts. The cohorts that came closest to obeying the law are cohorts 2 and 6, where two pairs of averages record equal values. Clearly, these are also the two cases where the measure of variability is smallest. The implication of Table 1 is that, the data sets

have clusters, and that the common mean will not adequately represents all the data points, since it will not be optimal in describing the performance of students in the various cohorts. Finding different centres will be a better representative of all possible groupings in the different cohorts.

Table 2: Boundary Statistics for Cohorts

Cohort	Boundary Statistics					% of values within $[\bar{x}_{L\bullet}, \bar{x}_{R\bullet}]$
	$x_{1\bullet}$	$\bar{x}_{L\bullet}$	$\bar{x}_{\bullet}$	$\bar{x}_{R\bullet}$	$x_{N\bullet}$	
Cohort 1	50	56	70	78	83	70
Cohort 2	60	66	72	77	78	80
Cohort 3	39	56	65	76	93	85
Cohort 4	45	53	61	72	88	87
Cohort 5	10	61	70	80	92	85
Cohort 6	54	61	67	74	87	84
Cohort 7	34	54	64	74	85	78
Cohort 8	13	44	57	69	92	76

Source: Data analysis of test data

Table 2 presents the results using the methods put forth in this study. It is evident that in all the cases, the common mean would not have represented the data, using multiple centres however have. Here, a better understanding of the performances is seen with the percentage of values within the three centres being at least 70%; in fact, all except cohort 1, have percentages greater than 75%, which is a figure for values within two standard deviations using the Chebyshev's theorem. In each cohort, the centres are the mean of three important groups in the data. They are also the mean performances of the three main groups in the data, beside the observations that can be classified as the least and largest categories.

The histograms depict the distribution of scores over the boundary statistics. Thus, Table 2 and Figure 1 provide a comprehensive description of the data. From the histograms, it is clear that the method used here provided a smaller number of groups which depicted the distribution of the original data very concisely. A five-group is fixed unlike the several groups possible when the Sturgis method of grouping was used. The effect of larger groups is that it makes description very clumsy whiles too few groups leave obvious gaps in description due to the fact that group centres will be further away from one another. Ultimately, the following implications are true of the multiple mean approach in data summary.

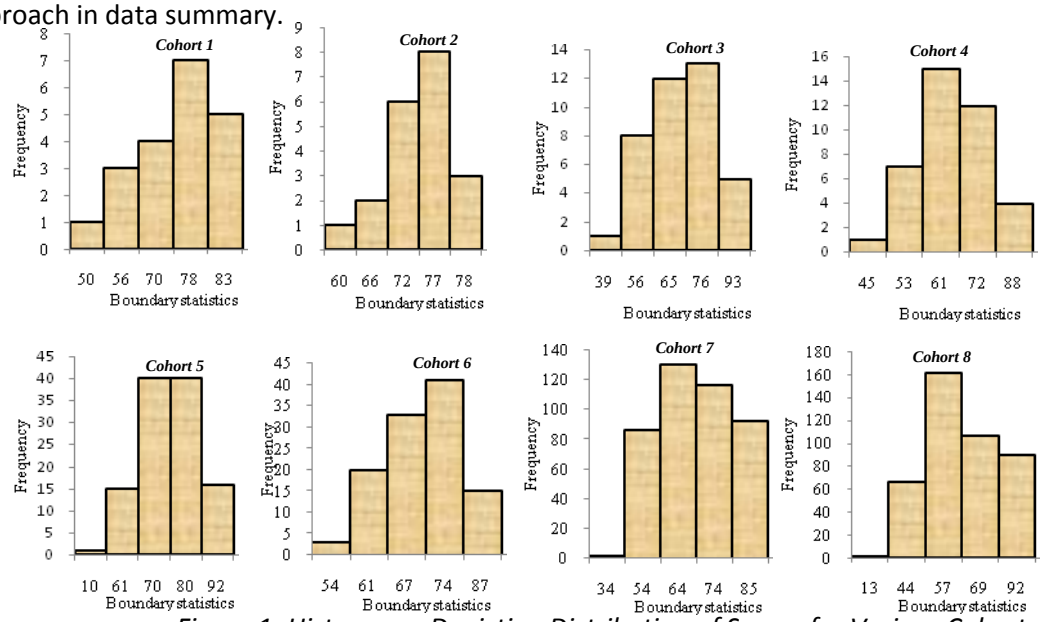


Figure 1: Histograms Depicting Distribution of Scores for Various Cohorts

1. It is easier in determining the percentage of values falling within the confidence interval of the mean.
2. Similarly, percentage of values in the neighbourhood of the two centres closest to the centre of the data could easily be determined too.
3. There is a very high tendency of having percentages in (1) and (2) above explaining the greatest proportion of the distribution of values.
4. Percentage of values at the very extremes, the least and largest values, are accordingly determined.
5. Data with extreme values, like in many cases of unobtrusive data, could be put into a maximum of five groups to provide for easy understanding of the data.

Conclusion and Recommendation

Using the arithmetic mean in describing data with extreme values frequently poses challenge and often meaningless. The drawbacks are due to the disadvantages inherent in the computation of the measure, notably is its upward bias and hence unrepresentative nature in such situations. In view of this, any attempt at making inference using the mean seriously distorts information and misleads. In addition, the result of such attempts also masks certain vital characteristics that ought to surface in the whole description process. Despite these challenges with the use of the mean, it is the single most used measure of central tendency.

When the distribution of data is not mound-shaped, the assumptions of normality cannot be applied and so estimates for the proportion of values falling within one, two and three standard deviations can't also be accurate when used. In such cases, the Chebyshev's theorem is applied in finding estimate of the proportion that fall within two and three standard deviations of the mean. These estimates are 75% and 89% respectively. Clearly, even at three standard deviations of the mean, about 11% of the data points would not be accounted for, ostensibly due to the possibility of very extreme data points. To accommodate all observations in the data description procedures, data are put into classes using the Sturgis method, class width and group limits are subsequently determined. Meanwhile, one drawback in this method is that, the number of groups/classes increases with increasing number of data points. The implication is that, there is the tendency of having too many groups to deal with. Obviously, this will create other problems associated with apathy in the data mining process. This is why an alternate method was explored in this study. The study was built on the theory of finding two more centres for the data, in addition to the centre of balance. The two extreme values were also included as lower and upper limits respectively. The method is thus called the multiple means method.

In the end, it appears that the multiple means method will provide a clearer, concise and yet comprehensive view of the data. Eight cohorts of data were used to test the theory. In all, except for the first cohort, the proportion of values within the two closest centres sandwiching the mean were at least 76%, which is already greater than the value for the Chebyshev's estimate. This suggests strongly that a five-limit or five boundary statistics would be enough in adequately describing data sets under similar conditions. In conclusion, the multiple means method could be a viable alternative to other methods in describing data which have extreme values.

REFERENCES

[1]. Clarke, G. M., and Cooke, D. (1998). *A Basic Course in Statistics (4th Edition).* New York, United States of America: Oxford University Press Inc.

[2]. Gordor, B. K., and Howard, N. K. (2006). *Introduction to Statistical Mathematics*. Cape Coast: Ghana Mathematics Group.

[3]. Keller, G., and Warrack, B. (2003). *Statistics for Management and Economics (6th Edition)*. Pacific Grove, United States of America: Thomson Learning.

[4]. Kirk, R. E (1999). Measures of Central Tendency. In: *Statistics, An Introduction (4th Edition).* Orlando, United States of America: Harcourt Brace and Company.

[5]. Ramachandran, K.M., and Tsokos, C. P. (2009). Mathematical Statistics with Applications (1st Edition). California, USA: Elsevier Academic Press.

[6]. Spiegel, M.R., and Stephens, L.J. (2011). Statistics (4th Edition). New York, USA: McGraw-Hills.

[7]. Solak, M. K. (2009). Detection of multiple outliers in univariate data sets. *Schering – Plough Research Institute*, New Jersey: paper SP06-2009, 1 – 7.

[8]. Turner, L. K., Knighton, D., and Budden, F. J. (1986). *Advanced Mathematics 2, 2nd edition; A Unified Course in Pure Mathematics, Mechanics and Statistics.* London: Longman Group Limited.

[9]. Nortey, E.N.N., and Afrim, J. (2013). *Numeracy Skills – The Basics and Beyond.* Accra, Ghana: Dieco Ventures.

[10]. Nsowah-Nuamah, N.N.N. (1999). A First Course in Probability Theory (Volume 2). Accra, Ghana: Ghana Universities Press.