TO DETERMINE SKEWNESS, MEAN AND DEVIATION WITH A NEW APPROACH ON CONTINUOUS DATA

Article · February 2018

DOI: 10.21506/j.ponte.2018.2.5

CITATIONS
4

READS
90,083

3 authors:

Mehmet Guven Gunver
Istanbul University
18 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Mustafa Senocak
İstanbul University-Cerrahpaşa
78 PUBLICATIONS **1,881** CITATIONS
SEE PROFILE

Suphi VEHİD
İstanbul University-Cerrahpaşa
62 PUBLICATIONS **540** CITATIONS
SEE PROFILE

Vol. 74 | No. 2/1 | Feb 2018
DOI: 10.21506/j.ponte.2018.2.5

PONTE
Florence, Italy
International Journal of Sciences and Research

TO DETERMINE SKEWNESS, MEAN AND DEVIATION WITH A NEW APPROACH ON CONTINUOUS DATA

MEHMET GUVEN GUNVER (Corresponding Author)

(Department of Biostatistics/Cerrahpasa Faculty of Medicine Istanbul University/Turkey/+905325939078/gunver@gmail.com)

MUSTAFA SUKRU SENOCAK

(Department of Biostatistics/Cerrahpasa Faculty of Medicine Istanbul University/Turkey/+905353596506/mssenocak@gmail.com)

SUPHI VEHID

(Department of Public Health/Cerrahpasa Faculty of Medicine Istanbul University/Turkey/+905337271572/vehid@istanbul.edu.tr)

Abstract

In light of the deeply embedded controversy surrounding the skewness and its potential exaggerated effects on standard deviation and other statistical measures, this study attempts to suggest an alternative approach, the "***G**olden **R**atio **i**n **S**tatistics*". The GRiS method acknowledges the effects of skewness by accounting for each data elements' contribution to the center point based on its specific location in the data stack. The "Golden Ratio" is chosen as the typical tool for this system. This study recommends a more realistic expression of the standard deviation, the GRiS deviations, which overcomes the effect of extreme values in skewed data stacks. A formula for the coefficient of skewness is also put forward to quantitatively measure the symmetry of a data stack around its median. The practical benefits of the GRiS approach are demonstrated through a data analysis of the characteristics of wine. Finally, the implications behind this approach are discussed.

Keywords: Arithmetic Mean, GRiS Deviations, GRiS Mean, Skewness, Standard Deviation

Introduction

Contrary to mainstream thought, skewness is observed in numerous distribution patterns across diverse fields such as economics [Davis (2004)], medicine [Akbani et al (2004)], and engineering [Kotsiantis et al (2006)]. Statistical literature over the course of history has made many attempts to tackle the concept of skewness. For example log-normal distribution [Witte & Schmidt (1977)], Box-Cox transformation [Logothetis (1990)], outliers [Barnett (1978)], trimmed mean [Kim (1992)], winsorised mean [Yuen (1971)] and non-parametric methods [Anderson (2006)], have all been proposed to manage skewness with specific data stack conditions. None of these proposed methods, however, can be used indiscriminately; each of

them has been proposed to resolve a very specific set of data stack conditions. Other major concerns with the current statistical approaches to skewness is that they either a) deform the distribution, b) damage the nature of the distribution, c) disregard the distribution, or d) are unclear when and where they are to be used [Anscombe (1973)]. The ambiguity [Silberzahn & Uhlmann (2015)] generated within the statistics environment by these approaches have resulted in a loss of confidence in the results. Statistics is a quantitative discipline, and there should be no room for questioning or debating which approach should be applied in any specific distribution pattern [Huff (1993), Best (2004)]. The approach recommended in this article attempts to remove the ambiguity associated with the management of skewness, and enhance the perception of confidence amongst its practitioners. In this article, we will recommend a method of managing skewness that does not a) alter, b) disregard, nor c) transform any of the components of the data stack. Most importantly, the recommended method is valid for any data stack, irrespective of the number of components, the values of these components (such as the components not being less than 0, or whether there are components between 0 and 1), the nature of the distribution, or even the degree or direction of skewness.

Rationale

The most frequently used method for explaining continuous data in statistics is the normal distribution function, which rarely exists in nature [Pearson (1920)]. This raises concerns since a lot of the observable data around us tends to be skewed. In this section we will explain the rationale behind this study, the purpose of which is to eliminate the weaknesses of the skewness, the arithmetic mean and the standard deviation. We will show that these concepts have inherent weaknesses that can have a major impact on the validity of statistical analysis.

Skewness

Skewed distributions have a long history in statistics literature. In 1895 Karl Pearson proposed the gamma distribution as a solution for skewed data stacks [Pearson (1895)]. However, Pearson was aware that the solution he was proposing was inadequate, and he clearly stated his skepticism that a suitable solution could be found for skewness. Economists such as Vilfredo Pareto also showed a special interest in skewed data stacks because many of the variables used in economics (for example salary structure) are skewed [Pareto (1897)]. Observations made in other disciplines such as engineering, the natural sciences [Boyd & Crawford (2012)], medicine, psychology [James et al (1984)] and the social sciences [Osborne (2010)] have also revealed that many data stacks housed skewness within their very nature. Even with all the interest it has generated in the statistics field, skewness remains a complicated concept with no universally acceptable solution [Groeneveld & Meeden (1984)]. The skewness formula; which is used in software such as Minitab, SPSS, Excel and SAS, is as follows [Doane & Seward (2011)] [1] ;

[1] Adjusted Fisher-Pearson standardized moment coefficient.

$$G_1 = \left(\frac{n}{(n-1)(n-2)}\right)\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\hat{s}}\right)^3$$

While the equation generates a precise and net output concerning the direction of the skewness, the result it generates concerning the cardinality of the skewness is variant. This is because the formula has normality assumption and it contains sample size. As having normality assumption, it contains a conroversy that, tries to describe skewness by depending on normality. On the other hand, sample size is very effective on this formula that, outcomes vary so much for two data stacks which have similar histograms.

Arithmetic Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The arithmetic mean is derived by summing up the values of all the components within a data stack and then dividing this total by the number of data inputs. The formula totally disregards the location of each element within the data stack, however, which is of utmost importance in order to be able to refer to a 'distribution'. The definition of the arithmetic mean does not bear any reference to the median. For example, consider an employer with 20 employees whose salaries total $20.000. Based on the above formula, the arithmetic mean for employee salary is $1.000 . The mean salary will always be $1.000 whether one employee is paid $19.050 and the other 19 employees are paid $50 each, or each of the employees is paid $1.000 . When each employee receives a salary of $1.000, the arithmetic mean and the median overlap at $1.000. However, when 19 employees receive a salary of $50 and one receives a salary of $19.050, the median of the data stack will be $50, but the arithmetic mean will still be $1.000. In other words, the arithmetic mean pays no attention to personal gain.

Standard Deviation

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Conceptually, the standard deviation is more 'sensitive' than the arithmetic mean because it accounts for the data's specific location within the data stack. However, it is weak when it comes to characterizing the distribution for two reasons:

1) While the arithmetic mean is used as a measure of position, it is calculated without using any measures concerning position. Therefore, this has a negative effect on the standard deviation calculation.

2) The expression, "the squared differences from the arithmetic mean", generates suitable solutions for data stacks which do not contain outliers, but causes great problems in data stacks which contain outliers. That is because, as the data moves away from the arithmetic mean in terms of its location, its contribution to standard deviation is higher than desired, because of the 'square' being used.

Methodology

In this section we will lay the framework for our recommended approach, the GRiS method, which acknowledges the effects of skewness by accounting for each data elements' contribution to the center point based on its specific location in the data stack. We built the framework in a step by step manner, have starting with the coefficient of skewness. We will use this new proposed formula for the coefficient of skewness to introduce the GRiS method and then apply it to derive the GRiS mean and GRiS deviations.

Coefficient of Skewness (G)

The objective when calculating this coefficient is to observe whether the load distributions of both sides of the data stack, according to the median (med), are balanced. The recommended formula for G is as follows :

$$G = \frac{for\ all\ x_i < med\ \Rightarrow\ \sum(x_i - med)}{else\ \Rightarrow\ \sum(x_i - med)}$$

If the data stack is symmetrical around the median, G should be equal to -1. If the data stack is skewed towards the left of the median, G will be smaller than -1, and if the data stack is skewed towards the right of the median, it will be bigger than -1. When evaluating a data stack with normal distribution function, the G is accepted as being -1. Where the G moves away from -1, the skewness contained within the data stack does not comply with the formation of the normal distribution function, and it will be necessary to evaluate the relevant data stack using parameters with a different approach.

GRiS

As shown above, neither the arithmetic mean nor the standard deviation are capable of validly characterizing distribution, particularly in skewed data stacks. In order to be able to eliminate this problem, we propose using alternative formulas, the GRiS mean and the GRiS deviations. The anchor point upon which our model was designed is as follows:

If the arithmetic mean is sensitive to extreme values, then under these circumstances, the contribution of the extreme values to the GRiS mean needs to be low, and the contribution of the values which close to median (those are not extreme) needs to be high.

This anchor point prescribes that the data should contribute to the center point calculation based on its location within the data stack. This is in contrary to the calculation of the arithmetic mean where the contribution of each component to the arithmetic mean is the same. To overcome the challenge of assigning a quantitative definition to the data's location within the data stack, we started off by making certain empirical recommendations for the purposes of determining 'low' and 'high' contributions:

- 'Low' 0.75 , 'high' 1.25
- 'Low' 0.66 , 'high' 1.33
- 'Low' 0.50 , 'high' 1.50

With these 'low' and 'high' values, we created "coefficient masks" which were to increase either in a linear or a parabolic manner, and again attempted to define a measure of central location. At this stage we were not particularly successful. However, as we worked with different data stacks, and as we scrutinized more and more, we realized that the 'low' and 'high' values we stated above are empirical, that is to say they are based only on our individual observations and experiences. They are completely artificial. However, we already have a 'low' and a 'high' value, the existence of which has been known for centuries, and which has been discovered many times over in nature: the "Golden Ratio", which is defined as $\frac{1+\sqrt{5}}{2}$. The coefficient masks which are constituted by depending on golden ratio ($1/\varphi$ = as 'low' and φ as 'high'), phenomenally generate proper outcomes.

GRiS Mean

According to the new approach we are proposing, the components which are close to the median should make a higher contribution to the mean, and those which are further away should make a lower contribution. To ensure this, a linear "GRiS mean - coeffcient mask", which takes φ at the median and $1/\varphi$ at the extremes, is built into the formula. 'O' is proposed as the symbol for the GRiS Mean.

Our aim is to evaluate each element's contribution to the mean, depending on a standard coefficient sequence. As we match each element in the data stack with its own coefficient, we first sort the data stack in an ascending order. So, X_1 represents the smallest and X_n represent the biggest element of data stack. To calculate the coefficients of each element, we use their locations in ascending order of the data stack.

Table 1. GRiS MEAN

index (i)	data stack X_i (ascending)	weighting coefficients M_c	weighted distance from median
1	X_1	DYNAMIC MEAN COEFFICIENT MASK	$M_{C1} * (X_1 - Med)$
2	X_2		$M_{C2} * (X_2 - Med)$
3	X_3		$M_{C3} * (X_3 - Med)$
...	...		...
n-2	X_{n-2}		$M_{C(n-2)} * (X_{(n-2)} - Med)$
n-1	X_{n-1}		$M_{C(n-1)} *(X_{(n-1)} - Med)$
n	X_n		$M_{Cn} * (X_n - Med)$

The coefficients assigned to the other elements in the data stack are calculated using the formula below :

$$M_{C_i} = \begin{cases} if\ x_i < med \ \Rightarrow 1/\varphi + 2 * (i-1)/(n-1) \\ else \ \Rightarrow 1 + \varphi - 2 * (i-1)/(n-1) \end{cases}$$

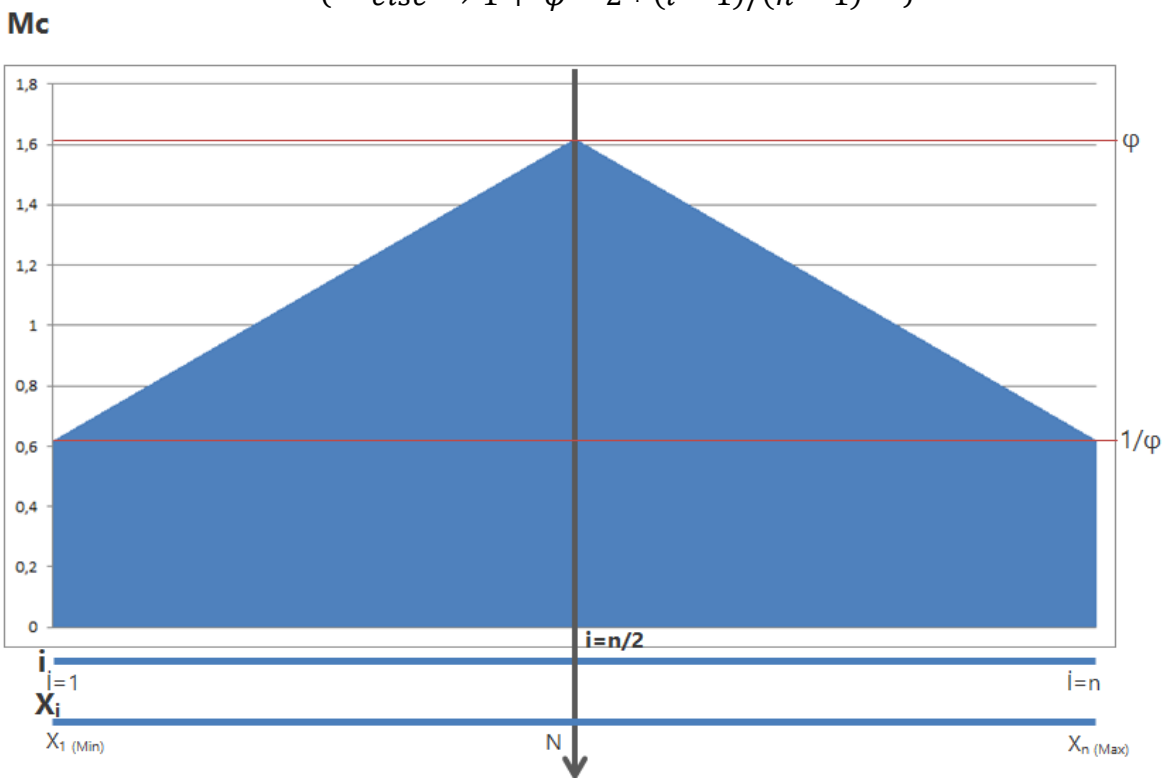

Fig.1 DYNAMIC MEAN COEFFICIENT MASK

The mean coefficient mask is dynamic, because the coefficient for each element changes depending on n (sample size). It is also symmetrical to the median. The next step is to calculate the weighted distance from the median of each element, and it is formulated as follows:

$$weighted\ distance\ from\ median = M_{C_i} * (X_i - Med)$$

The final step required for the calculation of the GRiS mean is calculating the deviation from the median:

$$deviation\ from\ median = \frac{\sum(M_{C_i} * (X_i - Med))}{\sum M_{C_i}}$$

The deviation from the median is always placed on the side of the median, which has already been set by the coefficient of skewness. If the data stack is skewed (as evidenced by the coefficient of skewness) to the right, the deviation from the median is positive. If the data stack is skewed to the left, the deviation from the median is negative.

The summation of the median and the deviation from the median gives us the GRiS Mean:

$$GRiS\ Mean\ (\boldsymbol{O}) = med + deviation\ from\ median$$

The anchor point of the GRiS mean (O) ensures that as the contribution of the far elements are reduced, O always occurs close to the median, and closer than the arithmetic mean.

GRiS Deviations

Overcoming the problem of the effect of extreme values in skewed data stacks on the arithmetic mean, by primarily altering our logic and secondarily relying on the "assistance of nature", was an encouraging first step. We then considered whether it would be possible to develop a different definition for deviation, which is an important parameter in terms of characterizing the data stack, but one which tends to lose its functionality in skewed data stacks.

The formula for standard deviation was defined around the end of the 19th century, when calculation tools had not advanced as much as today. The formula consists of taking the 2nd exponential in order to avoid the (-) sign, which the component of the data stacks' distance from the arithmetic mean may have depending on its position. Merely by taking the 2nd exponential, a significant problem occurs in the lack of definition of direction of the standard deviation. Any number which is squared turns into a positive number, so its sign (and therefore its direction) is lost.

To overcome this problem we used a similar version of the coefficient mask we had used in the GRiS mean calculation described above. The GRiS mean set as center point for GRiS deviations calculation, because O is located in accordance of skewness. This time, our mask needed to take the maximum value at the extreme values and the minimum value at the values close to the GRiS mean. That is because the 'deviation' is generated not by values which are close to the mean, but by the extreme values. This approach has two major advantages:

1. With this technique, the increase in the deviation in excess of its function (in excess of the norms of the normal distribution function), caused by the extreme values is prevented.

2. This technique has enabled the opportunity to perform something which has not been possible previously; the ability to calculate two separate deviations for both sides of the mean. With this method, any data stack can be expressed with four (G, O, D Left, D Right) parameters instead of two ($\bar{x}$, SD).

D Left and **D Right** are proposed as the symbol for the GRiS deviations.

Table 2. GRiS DEVIATIONS

index (i)	data stack X_i (ascending)	weighting coefficients M_c	weighted distance from GRiS MEAN
1	X_1	DYNAMIC DEVIATION COEFFICIENT MASK	$D_{C1} * (X_1 - O)$
2	X_2		$D_{C2} * (X_2 - O)$
3	X_3		$D_{C3} * (X_3 - O)$
...	...		...
n-2	X_{n-2}		$D_{C(n-2)} * (X_{(n-2)} - O)$
n-1	X_{n-1}		$D_{C(n-1)} *(X_{(n-1)} - O)$
n	X_n		$D_{Cn} * (X_n - O)$

This time, the first and last coefficient of the dynamic deviation coefficient mask are set to φ. And the coefficients, which are located closest to O from both sides are set to 1/φ. In order to create the dynamic deviation coefficient mask, we need to know the count of the elements of the data stack which are smaller than O.

k = count of elements of the data stack which are smaller than O

The coefficients assigned to other elements in the data stack are calculated by the formula below:

$$D_{C_i} = \varphi - \begin{Bmatrix} if\ X_i < O \Rightarrow (i-1)/(k-1) \\ else \Rightarrow 1 - (i-k-1)/(n-k-1) \end{Bmatrix}$$

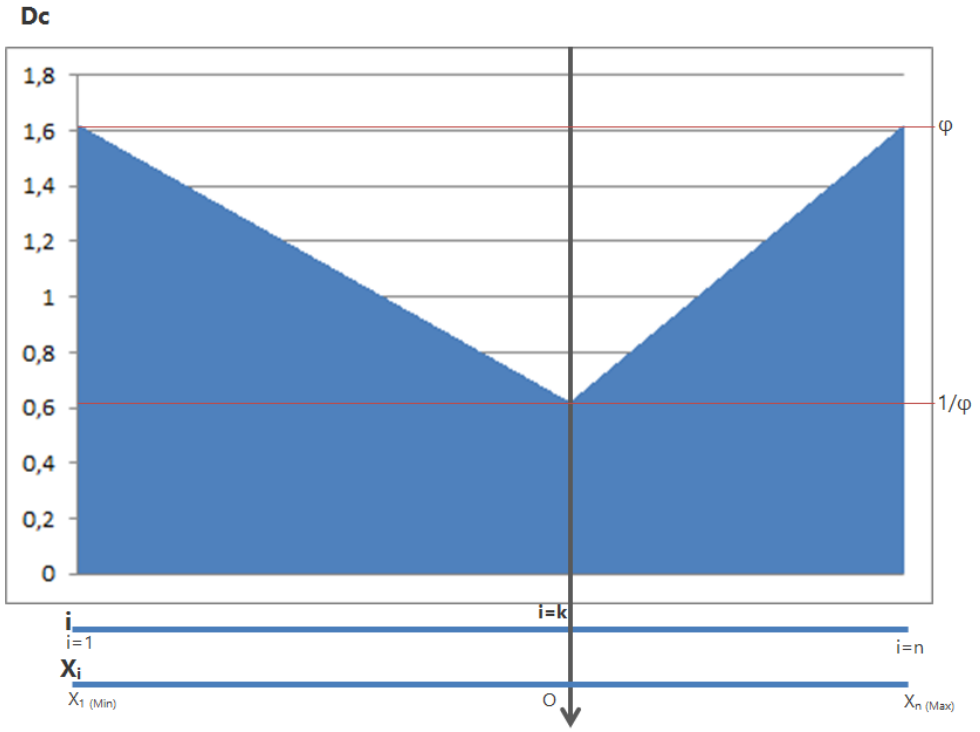

Fig.2 DYNAMIC DEVIATION COEFFICIENT MASK

The deviation coefficient mask is dynamic. Since the attended coefficients to each element change depending on both n and k, it is not symmetrical to the median nor to the GRiS mean.

The next step is to calculate the weighted distance of each element from the GRiS mean and it is defined as:

$$weighted\ distance\ from\ O = D_{C_i} * (X_i - O)$$

By doing so, we can now divide the data stack into two pieces by a special point (O). We can also 'divide' our calculation into two separate parts in order to define two independent parameters (deviations) for the both sides of the data stack.

$$GRiS\ Deviations = \begin{Bmatrix} for\ all\ X_i < O \Rightarrow \dfrac{\sum(D_{C_i} * (X_i - O))}{\sum D_{C_i}}\ \boldsymbol{D\ Left} \\ for\ all\ X_i > O \Rightarrow \dfrac{\sum(D_{C_i} * (X_i - O))}{\sum D_{C_i}}\ \boldsymbol{D\ Right} \end{Bmatrix}$$

Proof of Concept

This article is based on a post graduate thesis [Gunver (2014), Gunver et al (2014)], which applied the method proposed to all HDL[2], LDL[3], Cholesterol and Triglyceride examinations carried out at the Central Laboratory of the Cerrahpaşa Faculty of Medicine (Istanbul - Turkey), in 2012. To further demonstrate the validity of the method proposed, we applied it to a publicly available and previously published data set related to the characteristics of wine[4] [Cortez et al (2009)]. This data set in particular offered a number of the advantages we were seeking: 1) it contained more than one continuous variable, 2) parameters which are frequently used in biology such as density and pH were measured, 3) the data set has been used and referred to in many publications and studies, and most importantly 4) the fact that it was published in a manner that provides open access to everyone. The parameters of 4,898 white and 1,599 red wines were included. We calculated the coefficient of skewness (G), the GRiS Mean (O), and the GRiS Deviations (D Left and D Right) as outlined in the methodology section above, using a calculator published online[5].

[2] High-density lipoprotein
[3] Low-density lipoprotein
[4] http://archive.ics.uci.edu/ml/datasets/Wine+Quality
[5] http://www.goldenratioinstatistics.com/

Table 3. Proof of Concept

Data Stack	Example		Classic				New			
			Med	$\bar{x}$	SD	G_1	G	O	D Left	D Right
1	White wines	Fixed acidity	6.800	6.855	0.843	0.648	-0.842	6.836	-0.690	0.860
2		Volatile acidity	0.260	0.278	0.101	1.576	-0.598	0.273	-0.072	0.108
3		Citric acid	0.320	0.334	0.121	1.281	-0.713	0.329	-0.087	0.124
4		Residual sugar	5.200	6.391	5.072	1.077	-0.555	6.063	-3.935	5.830
5		Chlorides	0.043	0.046	0.022	5.023	-0.602	0.045	-0.011	0.018
6		Free sulfur dioxide	34.00	35.31	17.01	1.406	-0.818	34.82	-14.09	17.09
7		Total sulfur dioxide	134.0	138.4	42.50	0.390	-0.773	137.5	-37.17	43.75
8		Density	0.993	0.994	0.003	0.978	-0.789	0.994	-0.003	0.003
9		pH	3.180	3.188	0.151	0.457	-0.869	3.185	-0.130	0.152
10		Sulfates	0.470	0.490	0.114	0.977	-0.628	0.485	-0.090	0.120
11		Alcohol rate	10.40	10.51	1.230	0.487	-0.800	10.47	-1.063	1.342
12	Red wines	Fixed acidity	7.900	8.320	1.741	0.983	-0.518	8.223	-1.250	2.060
13		Volatile acidity	0.520	0.528	0.179	0.672	-0.896	0.524	-0.158	0.178
14		Citric acid	0.260	0.270	0.195	0.318	-0.875	0.267	-0.177	0.205
15		Residual sugar	2.200	2.539	1.410	4.541	-0.333	2.415	-0.509	1.666
16		Chlorides	0.079	0.087	0.047	5.680	-0.402	0.084	-0.016	0.045
17		Free sulfur dioxide	14.00	15.87	10.46	1.251	-0.620	15.20	-7.338	12.45
18		Total sulfur dioxide	38.00	46.47	32.90	1.516	-0.484	44.00	-22.09	38.96
19		Density	0.997	0.997	0.002	0.071	-1.004	0.997	-0.002	0.001
20		pH	3.310	3.311	0.154	0.194	-0.982	3.310	-0.138	0.148
21		Sulfates	0.620	0.658	0.170	2.429	-0.500	0.647	-0.106	0.186
22		Alcohol rate	10.20	10.42	1.066	0.861	-0.588	10.36	-0.828	1.228

Findings

Our assessment of the data stacks as a result of the calculations performed with the classical method (arithmetic mean and standard deviation) and the GRiS method is as follows:

Skewness

Almost all the data stacks were found to be skewed. Only two of the data stacks were almost perfectly symmetrical as their G value was very close to -1; namely the 19th data stack which tracked the density of the red wines, and the 20th data stack which tracked the pH of the red wines. However, despite their perfect symmetry, the arithmetic mean and standard deviation

generated results are distant from the empirical rule [Doane and Seward (2005)] (especially at the range of one standard deviation from the arithmetic mean). As in skewed data stacks, the GRiS which we have proposed has generated results which are close to the empirical rule, in symmetrical data stacks too. In particular, the 15th and 16th data stacks are sequences where the skewness is very apparent and which need to be especially examined.

A Value of Below Zero for the Value which is Three Standard Deviations Smaller than the Arithmetic Mean

In the 2nd, 3rd and 5th data stacks, the value which is three standard deviations smaller than the arithmetic mean has given results of less than zero. These parameters were volatility, citric acid and chloride, respectfully, and due to the very nature of wine, they are parameters which should be present, and it is therefore not theoretically possible for them to be smaller than zero. As location, direction and skewness are not taken into account while carrying out the calculations of the arithmetic mean and standard deviation, the standard deviation has increased unnecessarily, and now covers values which are theoretically impossible.

Lower and Upper Boundaries

While the 7th data stack appears to be perfectly compliant with the empirical rule when analyzed with the classical method, this appearance is actually quite misleading. There are outliers in both directions of the data stack which cannot be removed with Tukey's outlier definition. When the arithmetic mean and standard deviation are calculated and the density of the components at a standard deviation range is looked at, a result which is almost perfectly compliant with the empirical rule is obtained. However, the data stack is actually skewed. By running the numbers again using the GRiS approach, the data stacks' lower and upper boundaries changed.

The Mode being Distant from the Median

The mode is the component which is repeated the most within a data stack. In the event that the mode is distant from the median, or a large section of data is concentrated in a narrow area, which is distant from the median, it is difficult to talk about a 'symmetry' in that data stack. In such cases the concentration is at a point in the data stack which is not specific, but rather is stated. The 4th, 11th, 14th, 17th, 18th, and 22nd data stacks are examples of this situation. When analyzing the 4th data stack in particular, it seems as if the classical method has generated a perfect result that fully complies with the empirical rule. The GRiS method, on the other hand, generated results which are distant from the empirical rule. Upon further analysis it is found that the density of the data is in fact not around the mean, but in a region which is distant to the mean (and the median). In data stacks where this situation, which we are referring to as "distant concentration", is observed, the GRiS generates results which are at a range of one GRiS deviations each from the GRiS mean, and are smaller than the empirical rule.

Conclusions

The proposed GRiS method overcomes many of the weaknesses prevalent in the normal distribution, which is currently the most frequently used method of parameterization in all disciplines of science. With the proposed GRiS method skewness can be defined quantitatively, whereas it tends to be overlooked in the classical approach. The skewness of distributions of the same cardinality in different groups can be very different from each other. For instance, the density of the red wines (the 19th data stack) possesses a perfect symmetry, and the skewness coefficient is valued at -1, which points to complete symmetry. However, the density of the white wines (the 8th data stack) is not symmetrical, and is skewed to the right. Its skewness coefficient is at a value of -0.79, which points to a skewness to the right. A generic statement that the density of the wines is symmetrical, therefore, cannot be made. Each group which has been observed contains its own specific skewness characteristics, and by using the proposed GRiS method, skewness can be quantitatively assessed and defined. Herein, we shall also point out that, G_1 generated variant outcomes but G avoided this situation. The robustness of G is visible on the figures below; G_1 claims that pH of white wines is somehow symmetric, but G (correctly) proves the opposite.

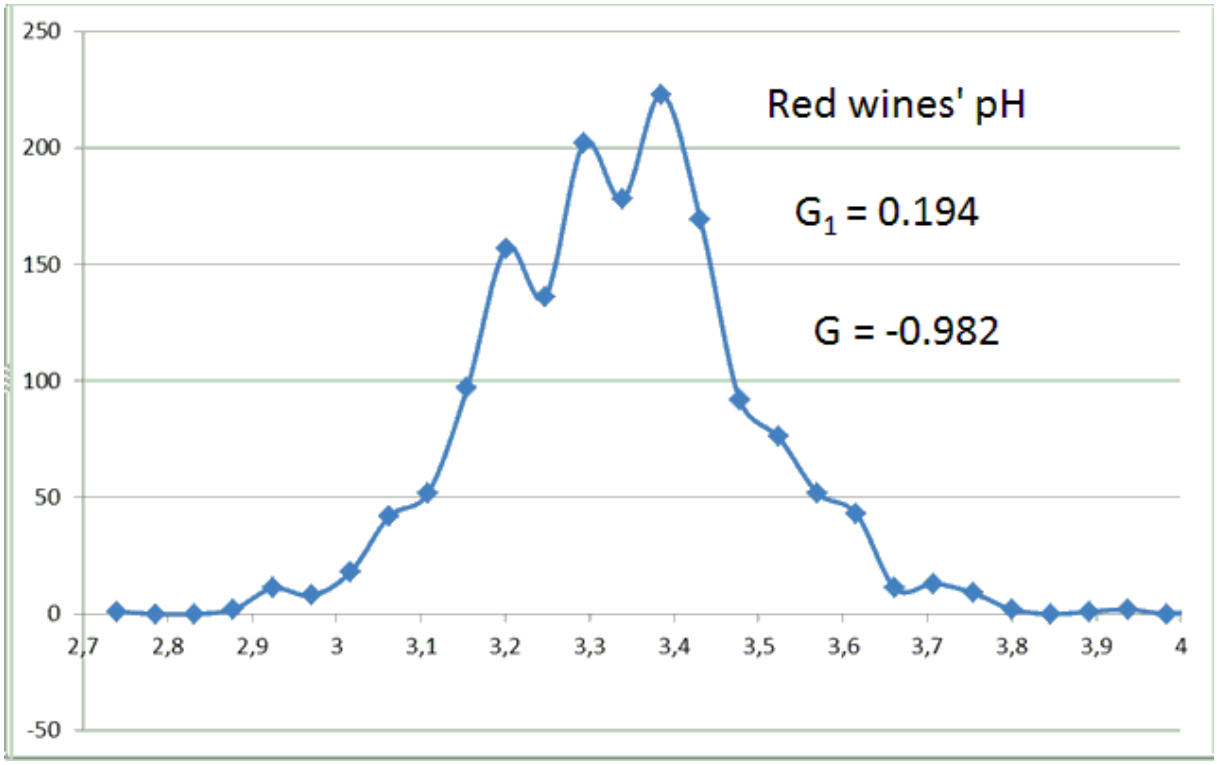

Fig.3 Red Wines' pH

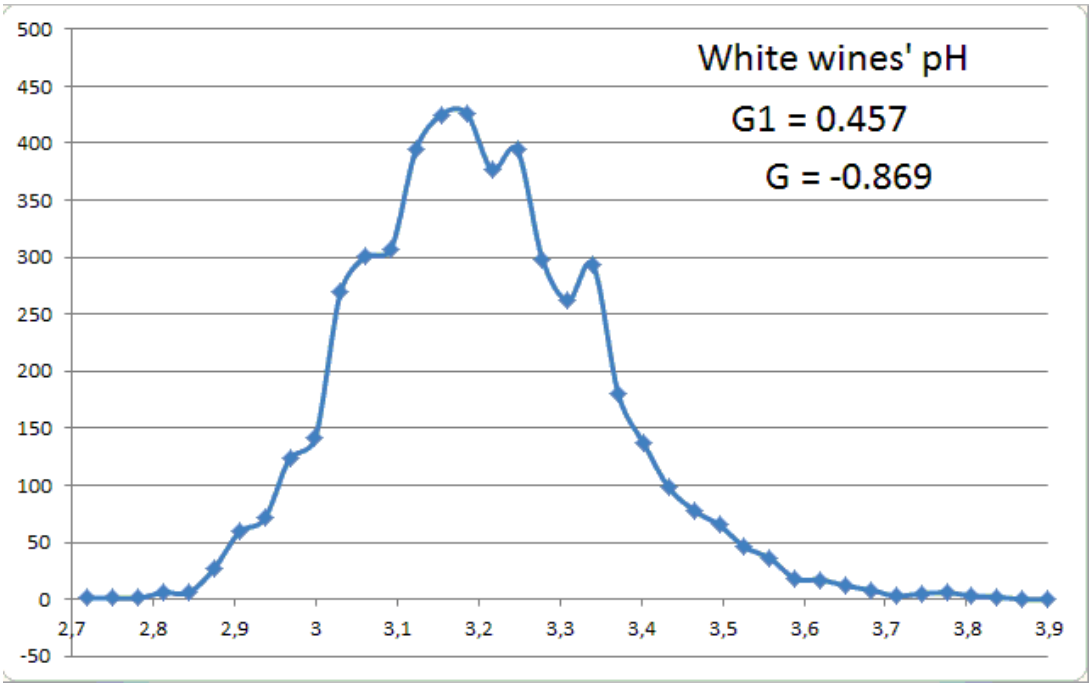

Fig.3 White Wines' pH

With the proposed GRiS method, parametrization is less misleading and more reflective of the actual nature of the data stack. In order for any data stack to be represented with a parametric function, the presumed mean needs to be close to the median. With the classical approach, this is sometimes achieved by removing the extreme values in the data stack which then raises concerns over the lower and upper boundaries. John Wilder Tukey was aware that the lower and upper boundaries determined based on the arithmetic mean and standard deviation values obtained following the removal of components, were subject to debate [Tukey (1977)]. Tukey was the first person to criticize the solution which himself had proposed, but he also clearly stated that his solution can serve as the foundation stone - as the first step. This often misguides practitioners, particularly in the biostatistics and epidemiology fields [Moyhinan (2006)]. The unnecessary increase in standard deviation in skewed data stacks affects the determination of the lower boundary, which can guide clinicians in the wrong direction. By using the GRiS method, however, new lower and upper boundaries can be defined for the use of clinicians in biostatistics and epidemiology, within the framework of commercial interest.

It has been observed that as the coefficient of skewness moves away from the value -1, which points to symmetry, the standard deviation also increases and the cover of standard deviation moves away from the empirical rule. A comparison of the classical approach with the proposed GRiS approach shows that while both methods scan approximately the same area with regards to the direction of the skewness (the tail), the arithmetic mean and standard deviation scan an unnecessarily large area towards the direction to the opposite side of the skewness. Another observation on the skewness coefficient is that the ratio of the left standard deviation to the right standard deviation is of a close value with the skewness coefficient. The method proposed generates a GRiS mean and GRiS deviations for the two different directions

of the GRiS mean separately. By taking into account the skewness contained within the data stack, the method is more successful in parameterizing both skewed and non-skewed data stacks than the classical method.

The solution proposed with the GRiS approach achieves a more successful parametrization than the classical method which is currently used. It produces a solution for skewness, and a solution for situations of distant concentration, which the classical method could not. One of the limitations of the GRiS approach, however, is that it is as powerless as the classical method for twin peaks [Maes (1998)]. Even with this limitation, the method proposed is valid and capable of reducing various uncertainties commonly faced in the field of statistics. Over a period of two years extensive testing was conducted on data stacks in various disciplines (meteorology [Lucas (2010)], economics, biology [McGill (2003)], chemistry [Periwal et al (2012)], human behavior [Faloutsos & Kamel (1994)]) etc., the results were evaluated, and the findings used to further develop our understanding and design of the GRiS method. Over the course of this two-year study, results obtained from different types of data stacks were similar to what was shared above in the proof of concept section, which shows the validity of the proposed approach. The GRiS method proposed in this article can be used to reevaluate the observations carried out in any discipline where descriptive statistics is used, and not just in the fields of biostatistics and epidemiology. For example, the medical sciences, economics, engineering, natural sciences, as well as the social sciences will all benefit from the proposed approach.

References

ANSCOMBE, Francis J. (1973)., Graphs in statistical analysis. The American Statistician, 27.1: 17-21.

AKBANI, R., KWEK, S., JAPCOWICZ, N. (2004)., Applying support vector machines to imbalanced datasets. In European conference on machine learning (pp. 39-50). Springer Berlin Heidelberg.

ANDERSON, M. J. (2006)., A new method for non-parametric multivariate analysis of variance, Centre for Research on Ecological Impacts of Coastal Cities, Marine Ecology Laboratories A11, University of Sydney, New South Wales, Australia.

BARNETT, V., (1978)., The Study of Outliers: Purpose and Model, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 27, No. 3, pp. 242-250.

BEST, J. (2004)., More Damned Lies and Statistics: How Numbers Confuse Public Issues., University of California Press.

BOYD, D., CRAWFORD, K. (2012)., Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Information, communication & society, 15(5), 662-679.

CORTEZ, P., CERDEIRA, A., ALMEIDA, F., MATOS, T., REIS, J. (2009)., Modelling Wine Preferences by Data Mining From Physicochemical Properties in Decision Support Systems, Elsevier, 47(4):547-553.

DAVIS, P. M. (2004)., For electronic journals, total downloads can predict number of users. portal: Libraries and the Academy, 4(3), 379-392.

DOANE, D. P., SEWARD, L. E. (2005)., Applied statistics in business and economics. USA: Irwin.

DOANE, D. P., SEWARD, L. E. (2011)., Measuring Skewness: A Forgotten Statistic?, Journal of Statistics Education, Volume 19, Number 2.

FALOUTSOS, C., KAMEL, I. (1994)., Beyond uniformity and independence: Analysis of R-trees using the concept of fractal dimension. In Proceedings of the thirteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (pp. 4-13). ACM.

GROENEVELD, R. A., MEEDEN, G. (1984)., Measuring Skewness and Kurtosis, The Statistician 33 391-399 Institute of Statisticians.

GUNVER, M. G. (2014)., A New Procedure for Determining Asymmetric Central Indefinite Distributions' Regulation by Using "Golden Ratio", MSc Thesis, Istanbul University, Institute of Health Science, Department of Biostatistic and Medical Informatics, Istanbul.

GUNVER, M. G., SENOCAK, M. S., VEHID, S. (2014)., Istatistikte Altin Oran, Turkmen Kitabevi, ISBN : 9786054749409, Istanbul.

HUFF, D., (1993)., How to Lie with Statistics., W. W. Norton & Company.

JAMES, L. R., DEMAREE, R. G., WOLF, G. (1984)., Estimating within-group interrater reliability with and without response bias. Journal of applied psychology, 69(1), 85.

KIM, S. J. (1992)., The Metrically Trimmed Mean as a Robust Estimator of Location, The Annals of Statistics, Vol. 20, No. 3, pp. 1534-1547.

KOTSIANTIS, S., KANELLOPOULOS, D., & PINTELAS, P. (2006)., Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering, 30(1), 25-36.

LOGOTHETIS, N. (1990)., Box-Cox Transformations and the Taguchi Method, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 39, No. 1, pp. 31-48.

LUCAS, C. (2010)., On developing a historical fire weather data-set for Australia. Australian Meteorological and Oceanographic Journal, 60(1), 1.

MAES, M. (1998)., Twin peaks: The histogram attack to fixed depth image watermarks. In International Workshop on Information Hiding (pp. 290-305). Springer Berlin Heidelberg.

McGILL, B. J. (2003)., Does Mother Nature really prefer rare species or are log-left-skewed SADs a sampling artefact?. Ecology Letters, 6(8), 766-773.

MOYHINAN, R. (2006)., Selling Sickness: How the World's Biggest Pharmaceutical Companies Are Turning Us All Into Patients, Nation Books.

OSBORNE, J. W. (2010)., Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research & Evaluation, 15(12), 1-9.

PARETO, V. (1897)., The New Theories of Economics, The Journal of Political Economics, Volume 5, Issue 4, p 485 - 502.

PEARSON, K. (1895)., Mathematical Contributions to the Theory of Evolution, II: Skew Variation in Homogeneous Material. Philos. Trans. Roy. Soc. London (A) 186 343–414.

PEARSON, K. (1920)., The fundamental problem of practical statistics. Biometrika, 13(1), 1-16.

PERIWAL, V., KISHTAPURAM, S., SCARIA, V., Open Source Drug Discovery Consortium. Computational models for in-vitro anti-tubercular activity of molecules based on high-throughput chemical biology screening datasets. BMC pharmacology, 12(1), 1-7.

SILBERZAHN, R., UHLMANN, E. L. (2015). Many hands make tight work. Nature, 526(7572), 189.

TUKEY, J. W. (1977)., Exploratory Data Analysis, Addison Wesley.

WITTE, A. D., SCHMIDT, P. (1977)., An Analysis of Recidivism, Using the Truncated Lognormal Distribution, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 26, No. 3, pp. 302-311.

YUEN, K. K. (1971)., A Note on Winsorized t, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 20, No. 3, pp. 297-304.