Epidemiology: Open Access
ISSN: 2161-1165

Jain, Epidemiology (Sunnyvale) 2017, 7:1
DOI: 10.4172/2161-1165.1000297

Research Article OMICS International

Distributional Characteristics of Selected Chemical and Environmental Variables: Data from NHANES 2003-2004

Ram B Jain*

Private Consultant, Dacula, USA

***Corresponding author:** Ram B Jain, Private Consultant, Dacula, USA, Tel: +1-910-729-1049; E-mail: Jain.ram.b@gmail.com

Received date: January 25, 2017; **Accepted date:** February 18, 2017; **Published date:** February 25, 2017

Abstract

Objective: Log-transformations are commonly used to normalize chemical data. However, log-transformations do not always normalize the data. Thus, the objective of this study was to recursively use Tukey's exploratory techniques to erect fences towards the data extremes until normality or near normality was achieved for the data lying within these fences.

Design: Data from National Health and Nutrition Examination Survey for the period 2003–2004 for 27 variables were used to conduct this study. Some of the 27 variables included for this study were: serum folate, serum transferrin receptor, urinary perchlorate, serum polychlorobiphenyl (PCB) 44, PCB-28, PCB-87, and PCB-52. Tukey's exploratory techniques were recursively used to erect fences towards the data extremes until normality or near normality was achieved for the data lying within these fences. Following this, robust techniques were used to estimate statistical parameters for the reduced data lying within these fences. The statistical properties of the reduced data so obtained were evaluated and compared with the original log-transformed data.

Setting: Cross-sectional data from National Health and Nutrition Examination Survey (NHANES) for the period 2003–2004 for 27 variables.

Subjects: 1790 to 8363 depending up on the variable of interest who participated in NHANES 2003-2004.

Results: The use of non-normal data for statistical analysis can lead to under- or over- estimation of the measures of central tendency (means and geometric means) depending upon the comparative mix and magnitude of the observations that are identified as potential outliers and trimmed from the lower and upper tails of the original distributions to achieve normality. The standard deviations are always over-estimated and the widths of the confidence intervals around the means are over-estimated. Additional insights into the demographic characteristics of those which were trimmed from extreme tails can be very valuable.

Conclusion: To obtain correct estimates of descriptive data, it is worthwhile to temporarily trim certain percent data (probably, < 5%) to achieve normality or near normality. An evaluation of these trimmed data can provide insight into the characteristics for a given variable of the persons who have too low or too high concentrations of the chemicals of interest.

Keywords: Cross-sectional studies; Persistent organic pollutants; Nutritional variables; Blood metals; Urine metals; Phytoestrogens; Outliers; Data transformations

Introduction

The distributions of most, if not all, chemical and environmental variables are characterized by a few relatively large measurements, or in other words, distributions of chemical and environmental variables are positively skewed. For this reason, data for chemical and environmental variables are assumed to be log-normally distributed even though not all positively skewed distributions can be considered to be mathematically log-normal. Since most statistical techniques including t-test, analysis of variance, and regression analysis assume normality of the distribution, it is necessary to transform log-normally distributed variables to normality by taking logs of the original measurements.

However, as will be seen in this paper, log-transformations do not always achieve normality. Sometimes, log-normally distributed variables still remain positively skewed or can become negatively skewed after the log-transformations. Under these circumstances, it is necessary to search for other techniques to achieve normality, or, if not possible, to achieve near normality, since many statistical tests like t-tests are robust to non-normality to some degree.

The most often used and discussed methodology to normalize data in the literature is the power transformation methodology developed by Box et al. [1]. Using this methodology a non-normal variable y is transformed to a normal variable $x = (y^{\lambda}-1)/\lambda$, where λ is the power transformation parameter and must be estimated if not already known. After the transformation, all analyses are carried out for x in place of y. Clark et al. [2] have presented a simplified method to use Box-Cox transformations. Other authors who have evaluated the applicability of Box-Cox and other transformations are Coder et al. [3], Errecalde et al. [4], Gasser et al. [5], Kingman and Zion [6], Montez-Rath et al. [7],

Payton et al. [8], Ponikowski et al. [9], van Albada and Robinson [10], and Volkova et al. [11].

To the best of our knowledge, the inferences based on x cannot always be converted back to the original variable, y. For example, it is unclear how to convert the value of the mean of x back to the mean or a comparable parameter for the original variable y. The mean for y cannot simply be back-transformed from the mean of x.

The issue of converting inferences based on x in a regression or any other modeling situation may even be more complex. If the differences between x for males and females are statistically significant, there is no clear way to determine if the same occurs for y and if so, what is the magnitude of the differences in the original scale. Some of the often used transformations to normalize non-normal data are special cases of Box-Cox transformations. For example, when λ=-1, it is equivalent to reciprocal transformation, when λ= 0.5, it is equivalent to square root transformation, and when λ=1/3, it is equivalent to cube root transformation.

Mateu [12] used Box-Cox transformations to normalize three environmental datasets, namely, for wind direction, SO2, and particle concentrations. Normalization of data was achieved for wind direction when λ=2, for SO2 when λ=0.5, and for particle concentrations when λ=0.5. According to Mateu [13], if a transformation can achieve symmetry, it is sufficient for practical purposes. Mateu [12] recommended logit transformations for percents and proportions.

If the presence of outliers or extreme values is an issue, then log transformation is a better choice than square root transformation (http://www.unm.edu/~marcusj/datatransforms.pdf). However, square root transformation was shown to perform better in achieving constant variance of the residuals and normality of the distribution for percent data [14].

Square root transformation has also been shown to stabilize variance for the counts data [15]. However, log transformation as compared to square root transformation was found to be more effective in reducing the skew and leptokurtosis that characterize the untransformed inter-individual EEG amplitude distributions [16].

Estimates of statistical parameters for the data that are not normally distributed can be biased. In order to obtain unbiased estimates to the degree it is possible, robust statistical techniques to estimate location and scale parameters have been proposed. Computations of trimmed and Winsorized means (http://www.statisticalanalysisconsulting.com/measures-of-central-tendency-the-trimmed-mean-and-median/) are two of the many techniques that have been proposed to obtain robust estimates of location parameters. Trimmed means are computed by trimming x% observations from each tail of the ordered data. X can vary from 0.1% to as much as 25%. If X=25%, trimmed mean so computed is based on the middle 50% of the data.

If an ordered data of size 20 is written as Y1, Y2, Y3…, Y18, Y19, Y20, and if X=10%, then trimmed mean is based on observations Y3, …,Y18. On the other hand, in order to compute Winsorized mean, first observations Y1 and Y2 are set equal to Y3 and observations Y19 and Y20 are set equal to Y18 and the Winsorized mean is computed for all 20 observations after the values of the observations Y1, Y2, Y19, and Y20 have been modified. However, depending up the value of X, the modified distribution used to compute Winsorized mean may became fat tallied and as such, computation of Winsorized means may not always be a good idea.

In this paper, if the normality is not achieved for the log-transformed data, we approach the task of achieving normality or near normality as an outlier detection problem followed by robust estimation using trimmed means. However, instead of using same value of X for both lower and upper tails, the value of X is allowed to be different for lower and upper tails depending up on the results of the outlier analyses as described later on.

In other words, we try to achieve normality by temporarily trimming a certain number of the lowest and highest observations from the data. Estimates of statistical parameters are then based on the data that remains after certain observations have been trimmed from the tails. For the purpose of this communication, dataset that remains after certain observations have been trimmed from the tails of the original dataset is called a reduced dataset.

The dataset containing observations that are trimmed from the original dataset is called trimmed dataset for the purpose of this communication. Robust estimation procedures are used for the reduced dataset. For the purpose of this study, a modified trimmed mean is computed. The advantages and drawbacks of this technique are discussed. Recommendations are made about the applicability of this technique under specific circumstances. Additional insight into the data that can be achieved using this technique is also discussed

Material and Methods

We downloaded publically available data for about 100 chemical variables from the National Health Examination and Nutrition Survey (NHANES) for the years 2003–2004 (www.cdc.gov/nchs/nhanes/nhanes2003-2004/lab03_04.htm). Data were downloaded for persistent organic pollutants (POPS), nutritional variables, and urinary and blood metals. Since, the percent values below the limit of detection (LOD) imputed as LOD/$\sqrt{2}$ can affect computations of skewness and also the log-transformation process, we used only those variables which had less than one percent observations below the LOD.

This selection process provided data for 27 variables for analysis purposes. A majority of these variables, 17, were measured in serum; seven were measured in urine; two were measured in plasma; and one was measured in the whole blood.

The sample sizes, skewnesses, and p-values for the Shapiro-Wilk test of normality [17] for the log10-transformed data for these variables are given in Table 1. The skewnesses prior to log10-transformation are also given. Since, for all 27 variables used in the study the Shapiro-Wilk test of normality W was statistically significant ($p \leq 0.01$) for the log10-transformed data, we used Tukey's exploratory techniques [18] to identify potential outliers. It should be noted that the W test computes the skewness of the data and evaluates if the skewness of the dataset was statistically significantly different from zero.

In order to use Tukey's exploratory techniques, we computed, Q1, the first quartile; Q3, the third quartile; IQR, the interquartile range computed as IQR=Q3-Q1; K=M*IQR, where M is arbitrarily called the fence multiplier; and lower fence FL=Q1-K; and upper fence, FU=Q3+K. All observations in magnitude below FL and above FU were considered potential outliers. When M=1.5, FL and FU are called Tukey's lower inner and upper inner fences respectively.

Variable	Sample Size	Skewness for untransformed data	Skewness for log10 transformed data	p-value for Shapiro-Wilk test for normality for log10 transformed data	Skewness of log10 transformed data with p-value for the test of normality when the fence multiplier was					
					3	2.5	2	1.5	1	0.5
Blood Lead	8373	9.4	0.36	≤ 0.01	0.34 (0.01)	0.31 (0.01)	0.25 (0.01)	0.22 (0.01)	0.07 (0.01)	0.04 (0.01)
Plasma Homocysteine	7888	7	0.57	≤ 0.01	0.38 (0.01)	0.31 (0.01)	0.23 (0.01)	0.15 (0.01)	0.05 (0.01)	0.01 (0.01)
Plasma Methylmalonic Acid	7544	28.4	1.39	≤ 0.01	0.65 (0.01)	0.53 (0.01)	0.38 (0.01)	0.34 (0.01)	0.17 (0.01)	0.13 (0.01)
Urine Creatinine	4449	1.2	-0.7	≤ 0.01	-0.7 (0.01)	-0.69 (0.01)	-0.61 (0.01)	-0.5 (0.01)	-0.33 (0.01)	-0.23 (0.01)
Serum Vitamin B12	8267	27.8	0.43	≤ 0.01	0.12 (0.01)	0.11 (0.01)	0.06 (0.04)	0.07 (0.03)	0.01 (0.01)	0.04 (0.01)
Serum Folate	8268	26.6	0.35	≤ 0.01	0.2 (0.01)	0.13 (0.01)	0.09 (0.01)	0.03 (0.01)	0.04 (0.01)	-0.05 (0.01)
Red Blood Cell Folate	8296	2.9	0.32	≤ 0.01	0.31 (0.01)	0.28 (0.01)	0.25 (0.01)	0.15 (0.01)	0.11 (0.01)	0.1 (0.01)
Serum Transferrin Receptor	2831	3.9	0.5	≤ 0.01	0.37 (0.01)	0.27 (0.01)	0.18 (0.01)	0.03 (0.01)	-0.08 (0.01)	-0.01 (0.01)
Serum Vitamin C	7277	0.6	-1.86	≤ 0.01	-1.24 (0.01)	-1.09 (0.01)	-0.91 (0.01)	-0.7 (0.01)	-0.53 (0.01)	-0.42 (0.01)
Urinary Daidzein	2594	13.1	0.24	≤ 0.01	0.24 (0.01)	0.24 (0.01)	0.24 (0.01)	0.19 (0.01)	0.11 (0.01)	0.1 (0.01)
Urinary Equol	2590	21.5	0.48	≤ 0.01	0.3 (0.01)	0.08 (0.01)	-0.06 (0.01)	-0.03 (0.15)*	-0.03 (0.01)	-0.02 (0.01)
Urinary Enterolactone	2594	14.1	-0.91	≤ 0.01	-0.8 (0.01)	-0.74 (0.01)	-0.67 (0.01)	-0.52 (0.01)	-0.36 (0.01)	-0.27 (0.01)
Urinary Genistein	2594	24.2	0.33	≤ 0.01	0.33 (0.01)	0.31 (0.01)	0.29 (0.01)	0.22 (0.01)	0.15 (0.01)	0.13 (0.01)
Urinary Iodine	2526	30.1	0.16	≤ 0.01	-0.19 (0.01)	-0.27 (0.01)	-0.26 (0.01)	-0.15 (0.01)	-0.1 (0.02)	-0.06 (0.01)
Urinary Perchlorate	2522	15.3	0.04	≤ 0.01	0 (0.01)	-0.04 (0.01)	-0.08 (0.01)	-0.15 (0.01)	-0.08 (0.01)	-0.07 (0.01)
Serum PCB 28	1790	23.1	0.49	≤ 0.01	0.15 (<0.01)	0.08 (0.01)	0.03 (0.02)	-0.05 (<0.01)	-0.06 (<0.01)	0 (<0.01)
Serum PCB 44	1814	18	0.08	≤ 0.01	0.26 (<0.01)	0.26 (<0.01)	0.16 (<0.01)	0.03 (0.01)	0 (<0.01)	-0.01 (<0.01)
Serum PCB 49	1801	18.9	-0.43	≤ 0.01	0.1 (<0.01)	0.14 (<0.01)	0.02 (0.15)*	-0.05 (<0.01)	-0.07 (<0.01)	-0.05 (<0.01)
Serum PCB 52	1821	21.4	0.09	≤ 0.01	-0.03 (<0.01)	-0.03 (<0.01)	-0.08 (<0.01)	-0.16 (<0.01)	-0.12 (<0.01)	-0.1 (<0.01)
Serum PCB 66	1822	24.4	0.69	≤ 0.01	0.42 (<0.01)	0.35 (<0.01)	0.27 (<0.01)	0.31 (<0.01)	0.2 (<0.01)	0.15 (<0.01)
Serum PCB 74	1822	6.5	0.57	≤ 0.01	0.57 (<0.01)	0.57 (<0.01)	0.56 (<0.01)	0.52 (<0.01)	0.41 (<0.01)	0.33 (<0.01)
Serum PCB 87	1816	12.8	-0.88	≤ 0.01	-0.9 (<0.01)	-0.92 (<0.01)	-1.07 (<0.01)	0.06 (<0.01)	0.08 (<0.01)	-0.17 (<0.01)

Serum PCB 99	1801	15.3	0.66	≤ 0.01	0.62 (<0.01)	0.59 (<0.01)	0.6 (<0.01)	0.46 (<0.01)	0.32 (<0.01)	0.19 (<0.01)
Serum PCB 118	1811	8.4	0.74	≤ 0.01	0.74 (<0.01)	0.74 (<0.01)	0.69 (<0.01)	0.59 (<0.01)	0.45 (<0.01)	0.34 (<0.01)
Serum PCB 138/158	1820	7.3	0.23	≤ 0.01	0.23 (<0.01)	0.23 (<0.01)	0.23 (<0.01)	0.22 (<0.01)	0.17 (<0.01)	0.09 (<0.01)
Serum PCB 153	1820	6.8	0.13	≤ 0.01	0.13 (<0.01)	0.13 (<0.01)	0.13 (<0.01)	0.13 (<0.01)	0.09 (<0.01)	0.03 (<0.01)
Serum PCB 180	1820	4.7	-0.24	≤ 0.01	-0.24 (<0.01)	-0.24 (<0.01)	-0.24 (<0.01)	-0.24 (<0.01)	-0.07 (<0.01)	-0.09 (<0.01)
*The distribution was normal										

Table 1: Sample sizes, skewnesses for the untransformed and log10 transformed data with fence multipliers from 0.5 to 3.0 and p-values for Shapiro-Wilk test of normality for the log transformed data.

As an example, when M=1.5, in a sample S={1, 12, 13, 15, 16, 18, 21, 24, 29, 71} of size 10, Q1=12.5; Q3=26.5; IQR=Q3-Q1=14; FL=12.5-1.5*14=-8.5; FU=26.5+1.5*14=47.5. Thus, assuming M=1.5; observations below -8.5 and above 47.5 were considered potential outliers. For this sample, since there was no observation below -8.5, there was no potential outlier on the lower side of the sample. However, there was one observation 71 above 47.5 which was considered a potential outlier. When M=0.5, FL=12.5-0.5*14=5.5 and FU=26.5+0.5*14=33.5. Since there was one observation below 5.5 and one observation above 33.5 in the sample, S, there were a total of two potential outliers in the data. It should be noted that as M decreases, the number of observations identified as potential outliers increases. When M=1.5, there was only one potential outlier in the sample. When M=0.5, there were two potential outliers in sample S. Higher values of M lead to smaller number of observations identified as potential outliers. Thus, higher values of M, for example, 1, will likely leave the reduced dataset with larger variability than will a relatively smaller value of M, for example, 0.5.

In the procedure proposed here, a specific value of M was used for the original log10-transformed dataset. Potential outliers below FL and above FU were trimmed from the original dataset, and the reduced dataset was tested for normality by using the W test. If the reduced dataset was found to be normally distributed, that dataset was accepted. If not, a different value of M was used for the original log10-transformed dataset. This process continued until a reduced dataset was found to be normally distributed or near normally distributed, or a decision was made to discontinue testing for normality as described below.

The value of M we initially used varied from 3.0 to 0.5 in decrements of 0.5. The p-values for the W test for each of the 27 datasets before and after applying M are given in Table 1. The reduced dataset for which normality or near normality was achieved was evaluated further for the distributional characteristics. The subsets of the data that were below FL and above FU were also evaluated for their demographic characteristics.

SAS Proc 9.3 (www.sas.com) was used to do statistical analysis.

Results

In the results presented below and throughout the manuscript, percent observations trimmed refers to the percent observations trimmed from the original dataset. For example, if there were 100 observations in the original dataset, and five observations were identified as potential outliers and trimmed from the lower tail, and 10 observations were identified as potential outliers and trimmed from the upper tail; then it will be said that a total of "15% observations were trimmed, 10% were trimmed from the upper tail and 5% were trimmed from the lower tail". The use of the words "lower" and "upper" tail always refers to the tails of the original log10-transformed data.

Distributions were not normal (Table 1) for any of the 27 variables even after the log10-transformations ($p \leq 0.01$). We could not find any observable pattern in terms of the size of skewness before or after log10-transformations that could be attributed to the matrices in which these variables were measured. Log10-transformations did substantially reduce skewness for all variables. For example, the skewness of serum Vitamin B12 was reduced from 27.8 to 0.43 (Table 1). But, for six variables, namely, urinary creatinine, serum Vitamin C, urinary enterolactone, PCB 49, PCB 87, and PCB 180, the distributions became negatively skewed after the log10-transformations. As the value of the fence multiplier, M, decreased, the absolute values of the skewnesses also decreased. However, because of relatively large sample sizes, even the smallest departures from the skewness of zero caused the p-values for the Shapiro-Wilk test to remain below 0.05. For example, when M=0.5, for serum PCB 153, the sample skewness was 0.03 but the p-value for the Shapiro-Wilk1 test of normality was still <0.01 (Table 1). The values of M below 0.5 were not considered because of the possible trimming of a substantial amount of data, probably as much as 25% or more. For urinary equol, normality was achieved when M=1.5, and for serum PCB 49 when M=2.0 (Table 1). For serum folate and serum PCB 44, the distribution became negatively skewed as M was reduced from 1.0 to 0.5 (Table 1). For serum transferrin receptor, the distribution became negatively skewed as M was reduced from 1.5 to 1.0 (Table 1). For urinary perchlorate, the distribution became negatively skewed as M was reduced from 3.0 to 2.5 (Table 1). For serum PCB 28 and PCB 87, the distributions became negatively skewed from positively skewed or vice versa as M was reduced from 2.0 to 1.5 (Table 1).

For each of the variables for which skewness switched signs from positive to negative or vice versa, further attempts were made to find a value of M for which normality or near normality could be achieved. For example, for serum folate and serum PCB 44, the values of M were explored between 1.0 and 0.5 in decrements of 0.1. In addition, while

the value of skewness remained negative for PCB 52 both at M=3.0 and M=2.5, the skewness increased as M was decreased further. As such, for PCB 52, a value of M between 3.0 and 2.5 was considered in decrements of 0.1. The results are given in Table 2 for these variables. The values of M for the urinary equol and serum PCB 49 were accepted as given in Table 1 (1.5 for urinary equol and 2.0 for PCB 49). For the other 18 variables, near normality was considered to be achieved when M=0.5 or when M=1.0. Even though the absolute skewness was lowest at 0.5, the value of 1.0 was preferable because too much data may be trimmed before robust estimations when M=0.5.

Variable	Fence Multiplier	Skewness of the reduced dataset	p-value for Shapiro-Wilk test of normality
Serum folate	0.9	0.03	0.01
	0.8	0.014	0.01
	0.7	-0.014	0.01
	0.6	-0.03	0.01
Serum PCB 44	0.9	-0.009	<0.001
	0.8	-0.027	<0.001
	0.7	-0.019	<0.001
	0.6	-0.026	<0.001
Serum transferrin receptor	1.4	0.026	0.01
	1.3	0.013	0.01
	1.2	-0.048	0.01
	1.1	-0.074	0.01
Urinary perchlorate	2.9	0.003	0.01
	2.8	-0.02	0.01
	2.7	-0.02	0.01
	2.6	-0.038	0.01
Serum PCB 28	1.9	0.017	0.016
	1.8	-0.004	0.013
	1.7	-0.014	0.009
	1.6	-0.022	0.006
Serum PCB 87	1.9	-1.067	<0.001
	1.8	-0.741	<0.001
	1.7	-0.758	<0.001
	1.6	-0.127	<0.001
Serum PCB 52	2.9	-0.031	0.001
	2.8	-0.031	0.001
	2.7	-0.031	0.001
	2.6	-0.031	0.001

Table 2: Fence multipliers, skewness, and p-value for Shapiro-Wilk test of normality for selected variables.

For the variables given in Table 2 and for urinary equol and PCB 49, the weighted means with their confidence intervals and standard deviations before and after M were applied as well as the number and percent of observations trimmed due to the application of fence multipliers are given in Table 3. The means of the reduced data, i.e., the data remaining after certain observations potentially identified as outliers were trimmed by use of fence multipliers were higher or lower than the original log-transformed data depending upon the mix of observations trimmed from the lower and upper tails of the original

data. For example when M=1.6, for log PCB 87, 16.4% observations were trimmed but the mean of the reduced data was still higher.

The final mean was 0.797 ng/g compared to the original mean of 0.602 ng/g because, of the total of 16.4% observations that were trimmed from the original data, 16% were from the lower tail and only 0.4% was from the upper tail. As would be expected, the standard deviations of the reduced data were always lower than for the original data. For example, the standard deviation of the reduced data for log PCB 87 was 0.249, 49.6% lower than that of the original data, which was 0.494. For this reason, the widths of the confidence intervals of the means for the reduced data were always lower than that of the original data.

The percent of observations trimmed to obtain the reduced sample varied from 0.1% for urinary perchlorate to 16.4% for PCB 87 (Table 3). A relatively large percent of observations, 9.5%, were also trimmed for serum folate.

Variable	Original N	Mean and 95% confidence intervals for the original log10-transformed data	Standard deviation for the original log10-transformed data	Fence Multiplier	Sample sizes for the reduced data after applying the fence multipliers	Mean and 95% confidence intervals for the reduced data after applying the fence multiplier	Standard deviation for the reduced data after applying the fence multiplier	Number and percent observations trimmed from the lower tail	Number and percent observations trimmed from the upper tail
Urinary Equol (ng/ml)	2590	0.904 (0.850-0.959)	0.604	1.5	2503	0.875 (0.827 - 0.923)	0.504	33 (1.3%)	54 (2.1%)
Serum Folate (ng/ml)	8268	1.085 (1.068-1.102)	0.223	0.8	7481	1.076 (1.064 - 1.088)	0.163	341 (4.1%)	446 (5.4%)
Serum PCB 49 (ng/g)	1801	0.896 (0.864 - 0.928)	0.261	2	1782	0.898 (0.864 - 0.933)	0.239	12 (0.7%)	7 (0.4%)
Serum PCB 52 (ng/g)	1821	1.210 (1.169 - 1.250)	0.268	2.9	1820	1.209 (1.169 - 1.250)	0.267	0 (0.0%)	1 (0.1%)
Serum PCB 44 (ng/g)	1814	1.098 (1.070 - 1.127)	0.249	0.9	1720	1.091 (1.064 - 1.117)	0.21	38 (2.1%)	56 (3.1%)
Serum Transferrin Receptor (mg/ml)	2831	0.559 (0.549 - 0.570)	0.135	1.3	2786	0.555 (0.545 - 0.565)	0.122	5 (0.2%)	40 (1.4%)
Urinary Perchlorate (ng/ml)	2522	0.508 (0.466 - 0.550)	0.395	2.9	2519	0.509 (0.468 - 0.550)	0.391	1 (0.0%)	2 (0.1%)
Serum PCB 28 (ng/g)	1790	1.476 (1.449 - 1.504)	0.219	1.8	1780	1.473 (1.446 - 1.501)	0.213	0 (0.0%)	10 (0.6%)
Serum PCB 87 (ng/g)	1816	0.602 (0.546 - 0.657)	0.494	1.6	1518	0.797 (0.770 - 0.825)	0.249	290 (16.0%)	8 (0.4%)

Table 3: Weighted means with 95% confidence intervals and standard deviations with and without application of fence multipliers for selected variables.

Table 4 shows the weighted means, 95% confidence intervals of the weighted means, and standard deviations before and after the fence multipliers were used for the 18 variables not included in Tables 2 and 3. The number and percent observations which were trimmed to obtain the reduced samples are given in Table 5. Whether the means of the reduced data were higher or lower than the original log10-transformed data depended on the mix of observations trimmed from the lower and upper tails. For example, for blood lead (Table 4), while the mean of the original log10-transformed data was 0.179 µg/dl, the mean of the reduced data when M=0.5 was 0.150 µg/dL since 9.6% of the observations were trimmed from the upper tail (Table 5). On the other hand, for PCB 180 (Table 4), while the mean of the original log-transformed data was 1.827 ng/g, the mean of the reduced data when M=1.0 was 1.842 ng/g, since a majority of observations removed were from the lower tail. In general, standard deviations of the original log10-transformed data were greater than or equal to standard deviations of the reduced data. The standard deviations were higher when M=1 than when M=0.5.

The widths of the confidence intervals for the means of the original log10-transformed data were greater than or equal to the widths of the confidence intervals of the reduced data The widths were higher when M=1 than when M=0.5. When M=1.0, the percent observations trimmed to obtain the reduced samples varied from a very low of 0.3% for PCB 153 to a high of 11.8% for serum Vitamin C (Table 5). The percent observations trimmed for PCB congeners were much smaller than for non-PCB variables. When M=0.5, the percent observations trimmed to obtain reduced samples varied from 4.1% for PCB 180 to a high of 21.5% for serum Vitamin C (Table 4).

	Statistics for Original Log10-Transformed Data			Statistics for Reduced Log10-Transformed Data after Outliers have been Fenced Out with Fence Multiplier, M=0.5			Statistics for Reduced Log10-Transformed Data after Outliers have been Fenced Out with Fence Multiplier, M=1.0		
Variable Name	**Sample Size**	**Mean with 95% Confidence Interval**	**Standard Deviation**	**Reduced Sample Size**	**Mean with 95% Confidence Interval**	**Standard Deviation**	**Reduced Sample Size**	**Mean with 95% Confidence Interval**	**Standard Deviation**
Blood Lead (ug/dl)	8373	0.179 (0.173-0.185)	0.292	7122	0.150 (0.146-0.155)	0.206	8071	0.161 (0.155-0.167)	0.257
Plasma Homocysteine (umol/l)	7888	0.852 (0.849-0.856)	0.181	6522	0.840 (0.837-0.842)	0.122	7566	0.841 (0.837-0.844)	0.155
Plasma Methylmalonic Acid (umol/l)	7544	-0.900 (-0.904 - -0.895)	0.21	6036	0.931 (-0.934-0.920)	0.116	7008	-0.928 (-0.932 - -0.920)	0.151
Urine Creatinine (mg/dl)	4449	2.039 (2.030-2.049)	0.315	3641	2.100 (2.093-2.106)	0.2	4170	2.082 (2.075-2.09)	0.258
Serum Vitamin B12 (pg/ml)	8267	2.729 (2.724-2.733)	0.209	6800	2.724 (2.721-2.727)	0.133	7832	2.723 (2.719-2.727)	0.17
Red Blood Cell Folate (ng/ml)	8296	2.394 (2.390-2.397)	0.16	6683	2.383 (2.381-2.386)	0.099	7760	2.385 (2.382-2.388)	0.129
Serum Vitamin C (mg/dl)	7277	-0.073 (-0.081 - -0.060)	0.308	5713	0.018 (0.014-0.021)	0.13	6424	0.007 (0.003-0.011)	0.169
Urinary Daidzein (ng/ml)	2594	1.855 (1.828-1.882)	0.711	2164	1.817 (1.796-1.838)	0.502	2497	1.827 (1.802-1.852)	0.636
Urinary Enterolactone (ng/ml)	2594	2.495 (2.468-2.521)	0.679	2077	2.635 (2.618-2.652)	0.391	2392	2.602 (2.582-2.623)	0.511
Urinary Genistein (ng/ml)	2594	1.517 (1.490-1.543)	0.696	2132	1.447 (1.427-1.467)	0.472	2480	1.478 (1.454-1.502)	0.609
Urinary Iodine (ng/ml)	2526	2.199 (2.184-2.214)	0.385	2007	2.223 (2.213-2.233)	0.227	2350	2.215 (2.203-2.227)	0.299
Serum PCB 66 (ng/g)	1822	0.913 (0.900-0.927)	0.295	1489	0.875 (0.866-0.884)	0.181	1711	0.890 (0.879-0.901)	0.231
Serum PCB 74 (ng/g)	1822	1.402 (1.381-1.423)	0.458	1613	1.323 (1.306-1.341)	0.358	1785	1.377 (1.357-1.397)	0.428
Serum PCB 99 (ng/g)	1801	1.357 (1.338-1.376)	0.41	1521	1.286 (1.272-1.300)	0.283	1724	1.316 (1.299-1.332)	0.353
Serum PCB 118 (ng/g)	1811	1.514 (1.493-1.535)	0.46	1571	1.418 (1.402-1.435)	0.333	1750	1.473 (1.453-1.492)	0.408
Serum PCB 138/158 (ng/g)	1820	1.873 (1.848-1.897)	0.529	1696	1.834 (1.811-1.856)	0.466	1812	1.866 (1.842-1.890)	0.52
Serum PCB 153 (ng/g)	1820	1.982 (1.957-2.008)	0.553	1705	1.960 (1.937-1.984)	0.496	1815	1.978 (1.953-2.003)	0.547
Serum PCB 180 (ng/g)	1820	1.827 (1.797-1.857)	0.656	1745	1.843 (1.815-1.871)	0.599	1805	1.842 (1.813-1.871)	0.631

Table 4: Sample sizes, weighted means with 95% confidence intervals, standard deviations for original and reduced log transformed data, and number and percent observation removed from the lower and upper tails for the reduced data.

	Fence Multiplier = 0.5		Fence Multiplier = 1.0	
Variable Name	**Number (percent) Observations Removed from Lower Tail**	**Number (percent) Observations Removed from Upper Tail**	**Number (percent) Observations Removed from Lower Tail**	**Number (percent) Observations Removed from Upper Tail**
Blood Lead (ug/dl)	445 (5.3%)	806 (9.6%)	52 (0.6%)	250 (3%)
Plasma Homocystein (umol/l)	591 (7.5%)	775 (9.8%)	71 (0.9%)	251 (3.2%)
Plasma Methylmalonic Acid (umol/l)	556 (7.4%)	952 (12.6%)	79 (1%)	457 (6.1%)
Urine Creatinine (mg/dl)	572 (12.9%)	236 (5.3%)	260 (5.8%)	19 (0.4%)
Serum Vitamin B12 (pg/ml)	707 (8.6%)	760 (9.2%)	182 (2.2%)	253 (3.1%)
Red Blood Cell Folate (ng/ml)	715 (8.6%)	898 (10.8%)	187 (2.3%)	349 (4.2%)
Serum Vitamin C (mg/dl)	1194 (16.4%)	370 (5.1%)	790 (10.9%)	63 (0.9%)
Urinary Daidzein (ng/ml)	191 (7.4%)	239 (9.2%)	29 (1.1%)	68 (2.6%)
Urinary Enterolactone (ng/ml)	361 (13.9%)	156 (6%)	181 (7%)	21 (0.8%)
Urinary Genistein (ng/ml)	176 (6.8%)	286 (11%)	29 (1.1%)	85 (3.3%)
Urinary Iodine (ng/ml)	297 (11.8%)	222 (8.8%)	114 (4.5%)	62 (2.5%)
Serum PCB 66 (ng/g)	117 (6.4%)	216 (11.9%)	29 (1.6%)	82 (4.5%)
Serum PCB 74 (ng/g)	36 (2%)	173 (9.5%)	0 (0%)	37 (2%)
Serum PCB 99 (ng/g)	77 (4.3%)	203 (11.3%)	3 (0.2%)	74 (4.1%)
Serum PCB 118 (ng/g)	39 (2.2%)	201 (11.1%)	0 (0%)	61 (3.4%)
Serum PCB 138/158 (ng/g)	32 (1.8%)	92 (5.1%)	0 (0%)	8 (0.4%)
Serum PCB 153 (ng/g)	42 (2.3%)	73 (4%)	0 (0%)	5 (0.3%)
Serum PCB 180 (ng/g)	45 (2.5%)	30 (1.6%)	14 (0.8%)	1 (0.1%)

Table 5: Number and percent observation removed from the lower and upper tails for the reduced data.

Plasma methylmalonic acid was selected for a detailed demographic evaluation of subjects in the lower and upper tails, because, for this variable, a majority of the trimmed observations were in the upper tail (6.1% *vs.* 1% when M=1, Table 5).

Vitamin C was also selected for a detailed demographic evaluation of subjects in the lower and upper tails, because, for this variable, a majority of the trimmed observations were in the lower tail (10.9% *vs.* 0.9% when M=1, Table 5). The results are given in Table 6.

For plasma methylmalonic acid, those subjects in the lower tail of the distribution (Table 6), for both M=0.5 and M=1.0, were predominantly females (62.2% when M=0.5, 70.9% when M=1), non-Hispanic blacks and Mexican Americans (79% when M=0.5, 81% when M=1), and aged ≤ 29 years old (74.8% when M=0.5, 69.6% when M=1).

The distinction between the middle of the distribution and the lower tail was much sharper when M=1 than when M=0.5. This might influence the choice between M = 0.5 and M=1. On the other hand, those who were in the upper tail were predominantly non-Hispanic whites (64.8% when M=0.5, 68.3% when M=1), males (54.7% when M=0.5, 57.1% when M=1), and aged 50+ years (63% when M=0.5, 68.9% when M=1).

More specifically (data not shown), 63% (when M=0.5) of those who were in the lower tail were non-Hispanic black and Mexican American males and females aged ≤ 29 years. However, when M=1, 26.6% of those who were in the lower tail were Mexican American males aged 50+years (data not shown).

Thus, selection of M could bias the interpretation of the results. In the upper tail, 46% were non-Hispanic white males and females aged 50+years when M = 0.5, and 51.6% when M=1.

For serum Vitamin C (Table 6), for both M=0.5 and 1.0, the subjects who were in the lower tail were predominantly males (55.9% when M=0.5, 57% when M=1), non-Hispanic whites (52.5% when M=0.5, 55.8% when M=1), and aged ≤ 29 years or 50+years (71.5% when M=0.5, 62% when M=1). When M=1, the age group distribution in the lower tail was similar, 31.6% for aged ≤ 29 years, 30.4% for those aged 30-49 years, and 38% who were aged 50+years.

		M=0.5			M=1.0		
Variable	**Demographic Group**	**Number (%) Observations in the Middle of the Distribution**	**Number (%) Observations in the Lower Tail**	**Number (%) Observations in the Upper Tail**	**Number (%) Observations in the Middle of the Distribution**	**Number (%) Observations in the Lower Tail**	**Number (%) Observations in the Upper Tail**
Plasma Methylmalonic Acid	Males	2988 (49.5%)	210 (37.8%)	521 (54.7%)	3435 (49%)	23 (29.1%)	261 (57.1%)
	Females	3048 (50.5%)	346 (62.2%)	431 (45.3%)	3573 (51%)	56 (70.9%)	196 (42.9%)
	Non-Hispanic Whites	2456 (40.7%)	79 (14.2%)	617 (64.8%)	2835 (40.5%)	5 (6.3%)	312 (68.3%)
	Non-Hispanic Blacks	1632 (27%)	260 (46.8%)	101 (10.6%)	1918 (27.4%)	40 (50.6%)	35 (7.7%)
	Mexican Americans	1498 (24.8%)	179 (32.2%)	157 (16.5%)	1743 (24.9%)	24 (30.4%)	67 (14.7%)
	Others	450 (7.5%)	38 (6.8%)	77 (8.1%)	512 (7.3%)	10 (12.7%)	43 (9.4%)
	≤ 29 Years	3346 (55.4%)	416 (74.8%)	226 (23.7%)	3845 (54.9%)	55 (69.6%)	88 (19.3%)
	30-49 Years	1177 (19.5%)	108 (19.4%)	126 (13.2%)	1339 (19.1%)	18 (22.8%)	54 (11.8%)
	50+ Years	1513 (25.1%)	32 (5.8%)	600 (63%)	1824 (26%)	6 (7.6%)	315 (68.9%)
Serum Vitamin C	Males	2774 (48.6%)	667 (55.9%)	149 (40.3%)	3115 (48.5%)	450 (57%)	25 (39.7%)
	Females	2939 (51.4%)	527 (44.1%)	221 (59.7%)	3309 (51.5%)	340 (43%)	38 (60.3%)
	Non-Hispanic Whites	2242 (39.2%)	627 (52.5%)	234 (63.2%)	2608 (40.6%)	441 (55.8%)	54 (85.7%)
	Non-Hispanic Blacks	1571 (27.5%)	247 (20.7%)	62 (16.8%)	1729 (26.9%)	147 (18.6%)	4 (6.3%)
	Mexican Americans	1469 (25.7%)	245 (20.5%)	52 (14.1%)	1614 (25.1%)	149 (18.9%)	3 (4.8%)
	Others	431 (7.5%)	75 (6.3%)	22 (5.9%)	473 (7.4%)	53 (6.7%)	2 (3.2%)
	≤ 29 Years	3063 (53.6%)	413 (34.6%)	167 (45.1%)	3374 (52.5%)	250 (31.6%)	19 (30.2%)
	30-49 Years	1079 (18.9%)	341 (28.6%)	37 (10%)	1208 (18.8%)	240 (30.4%)	9 (14.3%)
	50+ Years	1571 (27.5%)	440 (36.9%)	166 (44.9%)	1842 (28.7%)	300 (38%)	35 (55.6%)

Table 6: Number and percent observations in the middle of the distribution and in lower and upper tails by demographic variables for plasma methylmalonic acid and serum Vitamin C when fence multiplier was 0.5 and 1.0.

This is another instance where the selection of M could lead to different interpretations of the results. Those who were in the upper tail (Table 5) were predominantly female (62.3% when M=0.5, 60.3% when M=1), non-Hispanic white (63.2% when M=0.5, 85.7% when M=1), and aged ≤ 29 years or 50+ years (90% when M=0.5, 85.8% when M=1). More specifically (data not shown), more than 53% of those who were in lower tail were male and female non-Hispanic whites in all three age groups for M=0.5 as well as M=1. Among those who were in the upper tail, 31.1% were non-Hispanic white females aged ≥ 30 and 11.4% were non-Hispanic black males aged ≥ 50 when M=0.5. When M=1, Mexican American males aged ≥ 50 and non-Hispanic black males and females aged 30-49 years formed 67.5% of all those who were in the upper tail.

From the variables in Table 2, we selected PCB 87 and serum folate for a detailed study of demographic characteristics of those who were in the lower and upper tails. For PCB 87, 16% of the subjects which were trimmed were in the lower tail and 0.4% in the upper tail. For serum folate, 4.1% of the subjects trimmed were in the lower tail and 5.4% were in the upper tail. The results are given in Table 7. For serum folate (M=0.8), the subjects in the lower tail were predominantly males (54.5%), non-Hispanic whites and blacks (71.5%), and those aged ≤ 29 years (45.5%). For serum folate (M=0.8), the subjects in the upper tail

were predominantly females (59.2%), non-Hispanic whites (72.9%), and those aged 50+years (63.9%). Specifically (data not shown), those who were in the lower tail were non-Hispanic whites and non-Hispanic males and females aged ≤ 29 years (32.8%). Those who were in the upper tail were predominantly non-Hispanic white males and females aged 50+ years (51.7%).

Variable	M	Demographic Group	Number (%) Observations in the Middle of the Distribution	Number (%) Observations in the Lower Tail	Number (%) Observations in the Upper Tail
Serum Folate	0.8	Males	3716 (49.7%)	186 (54.5%)	182 (40.8%)
		Females	3765 (50.3%)	155 (45.5%)	264 (59.2%)
		Non-Hispanic Whites	2965 (39.6%)	115 (33.7%)	325 (72.9%)
		Non-Hispanic Blacks	1996 (26.7%)	129 (37.8%)	50 (11.2%)
		Mexican Americans	1950 (26.1%)	66 (19.4%)	46 (10.3%)
		Others	570 (7.6%)	31 (9.1%)	25 (5.6%)
		≤ 29 Years	4315 (57.7%)	155 (45.5%)	127 (28.5%)
		30-49 Years	1324 (17.7%)	106 (31.1%)	34 (7.6%)
		50+ Years	1842 (24.6%)	80 (23.5%)	285 (63.9%)
Serum PCB 87	1.6	Males	797 (52.5%)	144 (49.7%)	4 (50%)
		Females	721 (47.5%)	146 (50.3%)	4 (50%)
		Non-Hispanic Whites	681 (44.9%)	156 (53.8%)	2 (25%)
		Non-Hispanic Blacks	355 (23.4%)	71 (24.5%)	4 (50%)
		Mexican Americans	361 (23.8%)	43 (14.8%)	1 (12.5%)
		Others	121 (8%)	20 (6.9%)	1 (12.5%)
		≤ 29 Years	655 (43.1%)	127 (43.8%)	4 (50%)
		30-49 Years	341 (22.5%)	71 (24.5%)	4 (50%)
		50+ Years	522 (34.4%)	92 (31.7%)	0 (0%)

Table 7: Number and percent observations in the middle of the distribution and in lower and upper tails by demographic variables for serum folate and serum PCB 87.

For serum PCB 87, while males and females (Table 7) were almost equally distributed in the lower and upper tails; there were 52.5% males as compared to 47.5% females in the middle of the distribution. While non-Hispanic whites were predominant in the lower tail (53.8%), non-Hispanic blacks were predominant in the upper tail (50%). The distribution of age groups was not substantially different in the lower tail than in the middle of the distribution. Non-Hispanic white males and females aged 50+years accounted for 23.8% of the subjects in the lower tail.

Discussion

We have described a simple method based on Tukey's fences to achieve normality or near normality when log10-transformations of chemical data do not achieve normality. This method involves identifying and trimming observations from the lower and upper tails of the distribution that may hinder achieving normality after log10-transformations. The reduced dataset obtained after trimming certain observations from the lower and upper tails had means which could be smaller or larger than the original log10-transformed data depending upon the percent mix and magnitude of the observations that are trimmed from the lower and upper tails. For example, when a large majority of observations were trimmed from the lower tail as compared to upper tail (16% *vs.* 0.4%), the means of the log10-transformed reduced data for PCB 87 (Table 3) was higher (mean=0.797 ng/g, geometric mean (GM)=6.3 ng/g) than for the original log10-transformed dataset (mean=0.602 ng/g, GM=4.0 ng/g); the GM for the trimmed dataset was more than 50% higher. On the other hand, when a majority of observations that were trimmed from the upper tail (Table 4, M=0.5) as compared to the lower tail (11.1% *vs.* 2.2% for PCB 99); the mean for the reduced dataset was lower than for the original log10-transformed dataset. For example, the mean for PCB 99 for the reduced dataset was 1.418 ng/g (GM=26.2 ng/g) as compared to the mean for the original data which was 1.357 ng/g

(GM=22.8 ng/g); the GM of the reduced dataset was about 20% lower than for the original dataset. Thus, statistical values from the data which are not normal can lead to under- or over-estimation of the measures of central tendency such as means or geometric means.

Conversely, as expected, we found the estimates of dispersion, for example standard deviations, were always lower for the reduced dataset than for the original non-normal dataset. For example, for PCB 87 (Table 3), standard deviation for the reduced dataset was 0.249 (geometric standard deviation=1.77) while for the original dataset, it was 0.494 (geometric standard deviation=3.12) or the geometric standard deviation of the reduced dataset was about 43% lower than that of the original dataset.

The principal issues with the approach we proposed to achieve normality or near normality are (i) what percent of data from the original dataset can be ignored/trimmed to obtain robust estimates, (ii) what is the cost if no further analysis can be done on the data that are trimmed from the lower and upper tails, (iii) what should be done with the data trimmed from the lower and upper tails, and (iv) what additional information or insight can be achieved by studying the observations that form trimmed data. It is not simple to determine the percent from the original dataset that can be ignored/trimmed to achieve normality or near normality. Individual researchers must use additional clinical insight and input to decide which variables and the percent of data that can be trimmed without unacceptably altering the dataset. For example, for the variables in Table 3, there were only a few observations that needed to be trimmed from PCB 49, PCB 52, serum transferrin receptor, urinary perchlorate, and PCB 28 to achieve near normality; in these cases, the outliers can be ignored. In general, loss of ≤ 5% of the original dataset may not be substantial. But it depends up on the individual research issues involved. However, if a large majority of observations are trimmed from one tail compared to the other, there may be some information contained in the trimmed data that should not be ignored. This was probably the case with PCB 87 (Table 3) for which 16% out of the total of 16.4% of the observations which were trimmed were from the lower tail. It may be important to understand the demographics, residential location, dietary habits, and risky behaviors of the subjects trimmed from the lower tail to understand which of these factors might lower the concentration of the variable under consideration. While we did not evaluate the residential conditions (for example, industrialized *vs.* non-industrialized areas), their dietary habits (for example, consumption of fatty fish that may have exposed them to excessive PCB levels), or behavior (for example, smoking and/drinking) of these 290 subjects, we did use 24 demographic groups (2 gender × 4 race/ethnicity × 3 age groups) to more accurately identify them. By doing this, we found that 69 (23.8%) of them were non-Hispanic white males and females aged 50+ years, and 87 (30%) were non-Hispanic white males and females aged ≤ 49 years; 44.9% of the total population were non-Hispanic whites in the middle of the distribution. It would be informative to investigate the differences in residential, dietary, and behavioral factors between those non-Hispanic whites in the middle of the distribution and those who are in the lower tail. Similar evaluations could be useful if there are a substantial number of subjects which are trimmed from the upper tail.

Another question concerns the differences in the outcome of statistical analysis when the non-normality of the log-transformed data is ignored and analyses are carried out as if the log-transformed data were normal. We have already shown that the GM of the data may be under or over-estimated and the geometric standard deviation will always be over-estimated. Statistically significant differences discovered for original non-normal log10-treansformed may become statistically insignificant for the normal or nearly normal for the reduced data and vice versa. It is not impossible for the direction of statistically significant differences to be different between original and reduced dataset. Also, it is difficult to generalize what will happen to the regression coefficients when non-normal log-transformed data are used in model fitting. The results of this occurrence will depend up on the degree of non-normality of the dependent variable, the number of covariates, the total number of cells in the data, the cell sizes, and other independent variables in the model and their distributions.

In this study, we implicitly assumed that a single distribution will be sufficient to describe all demographic groups. There may be circumstances when this may not be true, for example, different demographics groups, for example, non-Hispanic blacks and Mexican Americans may assume different distributions, or in other words the total population may be a mixture of several distributions. If that is the case, each individual distribution should be analyzed separately by using the methodology proposed here.

While we described the demographic characteristics of the distributional tails of methylmalonic acid, Vitamin C, PCB 87, and serum folate, other characteristics of tails should also be looked into, for example, how their dietary habits are different from those who are in the middle of the distribution.

The outlier detection methodology we proposed to normalize or nearly normalize the data is simple, but it is also crude. However, it affords us an opportunity to convert inferences which are based on normalized log-transformed reduced data.

Better methods to normalize data, for example, Box-Cox transformations have been proposed. However, until the issue of convertibility of estimated parameters for the transformed data using Box-Cox transformations to original scale can be resolved, data that remains non-normal after log-transformation can be normalized or nearly normalized using Tukey's exploratory procedures as defined here. This may, in fact, lead to additional insight into the data regarding the subjects at the tails of the distributions. Such, insight may not be possible using Box-Cox transformations.

The use of alternate non-parametric methods has been recommended when the distribution of the data to be analyzed is not normal (http://blog.minitab.com/blog/adventures-in-statistics-2/choosing-between-a-nonparametric-test-and-a-parametric-test). However, before succumbing to this temptation, what non-parametric methods actually do needs to be understood. Essentially, non-parametric methods rank order the data before attempting to do any analysis. For example, let us see we need to compare the means of two datasets, say, X with observations {3, 12, 21, 34, 231} and Y with observations {2, 9, 23, 39, 49}. Then, with use of a non-parametric method, they will be ranked XR={2, 4, 5, 7, 10} and YR={1, 3, 6, 8, 9}. The sum of ranks for the two datasets will be ∑XR=28 and ∑YR=27. Then, by one or the other non-parametric methods, for example, Wilcoxon Rank SumTest, ∑XR will be compared with ∑YR and p-value for the test of statistical significance will be computed which will inform whether or not the "means" of the two datasets are statistically significantly different. To the best of my knowledge, there is no way to indicate by what magnitude X and Y are different in the original scale. In the opinion of this author, drawing conclusions based on ranks rather than the original scale is a serious drawback. In the clinical sciences, it is essential to know the differences in the original scale than in the scale based on ranks. The comparative powers of parametric *vs.*

non-parametric tests is not an issue that should solely be used to make a judgment about the appropriateness of a statistical test of significance. Consequently, this author prefers to use parametric tests.

Transformations other than log transformation have been proposed to reduce right skewness of the data. It should be noted that the main issue with chemical and environmental data is the skewness of the data and only those transformations that reduce the skewness should be considered. Certain transformations like 1/X changes the skewness of the data from right skewed to left skewed and vice versa and as such are not of use for analyzing chemical and environmental data. Log transformations are not always capable of normalizing data but the use of Tukey's fences along with log transformations as described in this communication, can achieve near normality, if not normality of the data.

Acknowledgment

Some of the work presented here was completed when the author was with the Centers for Disease Control and Prevention in Atlanta, Ga, USA.

References

1. Box GEP, Cox DR (1964) An analysis of transformations. J R Statist Soc 26: 211-252.
2. Clark JE, Osborne JW, Gallagher P, Watson S (2016) A simple method for optimising transformation of non-parametric data: an illustration by reference to cortisol assays. Hum Psychopharmacol. 31: 259-267.
3. Coder DM, Redelman D, Vogt RF (1994) Computing the central location of immunofluorescence distributions: logarithmic data transformations are not always appropriate. Cytometry 18: 75-78.
4. Errecalde JO, Mestorino N, Mariño EL (1997) The effects of the method of calculation on the evaluation of the pharmacokinetic parameters of oxytetracycline after intravenous administration to calves. Vet Res Commun 21: 273-281.
5. Gasser T, Ziegler P, Seifert B, Prader A, Molinari L, et al. (1994) Measures of body mass and of obesity from infancy to adulthood and their appropriate transformation. Ann Hum Biol. 21: 111-125.
6. Kingman A, Zion G (1994) Some power considerations when deciding to use transformations. Stat Med 13: 769-783.
7. Montez-Rath M, Christiansen CL, Ettner SL, Loveland S, Rosen AK (2006) Performance of statistical models to predict mental health and substance abuse cost. BMC Med Res Methodol 6: 53.
8. Payton ME, Richter SJ, Giles KL, Royer TA (2006) Transformations of count data for tests of interaction in factorial and split-plot experiments. J Econ Entomol 99: 1002-1006.
9. Ponikowski P, Piepoli M, Amadi AA, Chua TP, Harrington D, et al. (1996) Reproducibility of heart rate variability measures in patients with chronic heart failure. Clin Sci (Lond) 91: 391-398.
10. van Albada SJ, Robinson PA. (2007) Transformation of arbitrary distributions to the normal distribution with application to EEG test-retest reliability. J Neurosci Methods. 161: 205-211.
11. Volkova N, Klapper E, Pepkowitz SH, Denton T, Gillaspie G, et al. (2002) A case-control study of the impact of WBC reduction on the cost of hospital care for patients undergoing coronary artery bypass graft surgery. Transfusion 42: 1123-1126.
12. Mateu J (1995) The problem of assessing and achieving normality: an application to environmental data. Transac Ecology Env 6: 267-274.
13. Mateu J (1997) Methods of assessing and achieving normality applied to environmental data. Env Mngmnt. 21: 767-777.
14. Van-Wyck DB, Giles I, Sharpe K (2012) Within-patient variation of hemoglobin and reticulocytes: implications for evaluating ESA responsiveness in dialysis patients. Int J Lab Hematol 34: 577-583.
15. Alexander N, Bethony J, Corrêa-Oliveira R, Rodrigues LC, Hotez P, et al. (2007) Repeatability of paired counts. Stat Med 26: 3566-3577.
16. Pollock VE, Schneider LS, Lyness SA (1990) EEG amplitudes in healthy, late-middle-aged and elderly adults: normality of the distributions and correlations with age. Electroencephalogr Clin Neurophysiol 75: 276-288.
17. Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52: 591-611.
18. Tukey JW (1977) Exploratory Data Analysis.