
The logarithmic transformation and the geometric mean in reporting experimental IgE results: what are they and when and why to use them?

Jake Olivier, PhD*; William D. Johnson, PhD*; and Gailen D. Marshall, MD, PhD†

Background: Immunologic data, such as IgE and interleukin 4, tend to have positively skewed distributions with a long tail of larger values. This renders analyses based on normal distribution theory questionable (eg, *t* tests and analysis of variance) and distorts the sample mean as a measure of central tendency. These problems can be addressed through analysis of log-transformed data. Data analyzed in this fashion are summarized with the geometric mean.

Objective: To elucidate the use of the logarithmic transform and the geometric mean in the analysis of immunologic data.

Methods: The analysis may be conducted by transforming the data to a logarithmic scale to achieve a bell-shaped (approximately normal) distribution. The bell-shaped distribution needed to validate statistical inferences is only achieved in the transformed scale. In summarizing the research findings, the statistical analyst usually will transform means and confidence intervals from the logarithmic scale back to the original scale of measurement. Statistical inferences in the log scale remain valid for the data. The result of back transforming the mean of logarithmic values to the original scale is the geometric mean. This statistic is less subject to distortion by the unusually large values in the tail of the positively skewed distribution of the data.

Results: A brief example is used to illustrate this type of analysis.

Conclusions: Logarithmic transformation permits valid statistical inference for positively skewed immunologic data. A result of this analysis is the geometric mean, which is a better measure of central tendency of this data type than the usual sample mean.

Ann Allergy Asthma Immunol. 2008;100:333–337.

INTRODUCTION

The sample mean is the most frequently used statistic for summarizing research data in applications in which the response of interest is measured on a continuous scale. This use may be entirely appropriate if the distribution of the data is symmetric and bell shaped so that it resembles the Gaussian or normal distribution. However, if the response has a positively skewed distribution, so that some of the observed values are exceptionally large, the ordinary (arithmetic) mean may not be a wise choice for a summary statistic. This is because the disproportionate influence of a few large observations may inflate the mean to such a magnitude that it no longer provides a useful estimate of central tendency. Inferences based on an overinflated estimate of central tendency may not be valid.

This article suggests an approach to analyzing positively skewed data. Many researchers choose the geometric mean to summarize this type of data.^{1–8} The initial step in a geometric mean analysis is to transform positively skewed data to a

more symmetric and bell-shaped distribution. This allows the researcher to conduct statistical tests, such as analysis of variance (ANOVA), while yielding *P* values that are interpretable in the original scale. Confidence intervals computed from transformed data are also valid and remain interpretable when back transformed to the original scale.

METHODS

Transforming Data for Analyses

The statistical theory used to develop the classic techniques for constructing confidence intervals for population means and performing statistical tests of hypotheses begins with the assumption that the response of interest has a normal distribution. If this assumption is true, inferences from the results are valid (eg, statistical significance of *P* values and width of confidence intervals). If the assumption is not true, we may be able to transform the data to a scale in which we can draw valid inferences from our results. Statisticians use many types of transformations to accomplish this goal. One common type is the logarithmic transformation. If the data are continuous and nonnegative and have a positively skewed distribution, they may have a distribution that is referred to as log normal. This description refers to data that when logarithmically transformed are then normally distributed. The analyst may use any base for the logarithm when making the transformation, but the natural logarithm with base *e*, denoted *ln* or *log_e*,

Affiliations: * Division of Biostatistics, Department of Preventive Medicine, University of Mississippi Medical Center, Jackson, Mississippi; † Division of Clinical Immunology and Allergy, Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi.

Disclosures: Authors have nothing to disclose.

Received for publication July 13, 2007; Received in revised form October 2, 2007; Accepted for publication October 2, 2007.

is conventionally used, and all results presented herein assume that the data are transformed in this fashion.

To illustrate immunologic data with a positively skewed distribution that could be encountered in practice, we simulated a random sample of $n = 200$ IgE levels using the reported geometric mean (106.6 IU/mL) and confidence interval (101.5–112.0 IU/mL) found in the study by Borish et al.¹ The data were simulated by first computing the mean and SD of the data on the log scale ($\bar{x} = 4.7$ and $s = 1.7$, respectively) by log transforming the geometric mean and confidence limits. Then, a random sample was created using a normal distribution random number generator with the same mean and SD as the data found in the study by Borish et al. We then applied the reverse of the log transformation (exponentiation) to the data to obtain a simulation of data that we might encounter in experimental studies. A histogram of the resulting data is shown as Figure 1, as is a trace of the normal distribution that provides the best fit to the histogram. This histogram is fairly typical of positively skewed immunologic data. Clearly, any analyses of the data in this scale that assume the data come from a normal distribution will be flawed. Figure 2 then shows the data when log transformed back to the original simulated data, together with a trace of the normal distribution that best fits these data. The normal distribution assumption is more appropriate in the logarithmic scale.

Sample Mean and 95% Confidence Intervals

Let $x_1, x_2, \dots, x_n$ denote a sample of n observations or measurements (the original data), such as those in Figure 1. The arithmetic mean of these observations is defined as follows:

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

Let $y_i = \ln(x_i)$, for $i = 1, 2, \dots, n$, represent the sample of log-transformed observations such as those in Figure 2. The arithmetic mean of the log-transformed data is as follows:

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{\sum_{i=1}^n \ln(x_i)}{n} \quad (2)$$

Similarly, the SD of the log-transformed data is as follows:

$$s = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}} \quad (3)$$

If one were interested in a 2-sided 95% confidence interval for the population mean of the transformed data and if the log-transformed data have a reasonably bell-shaped distribution, then the interval is approximated as follows:

$$\left(\bar{y} - 1.96 \frac{s}{\sqrt{n}}, \bar{y} + 1.96 \frac{s}{\sqrt{n}} \right) \quad (4)$$

The unit of measurement for the sample mean and the upper and lower 95% confidence bounds are on the log scale.

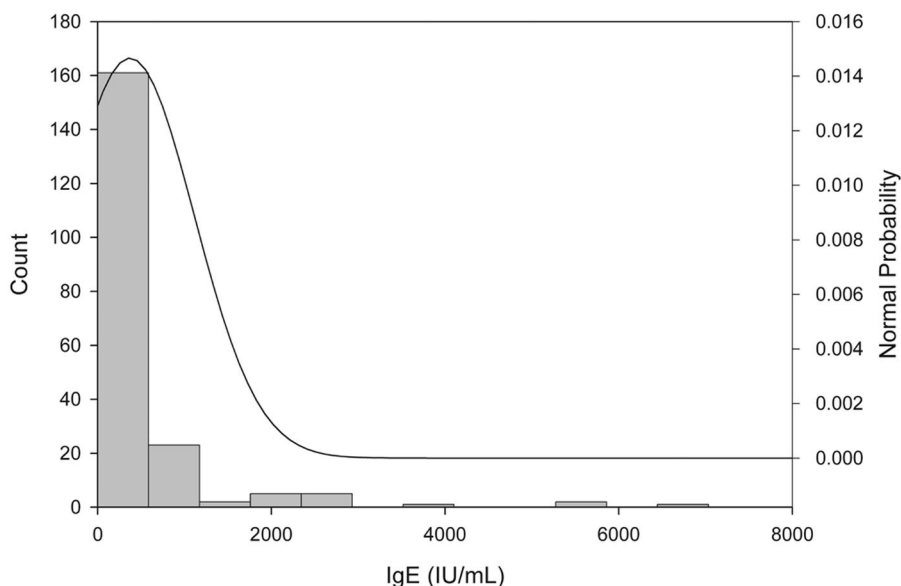


Figure 1. Simulated IgE levels from the data of Borish et al.¹

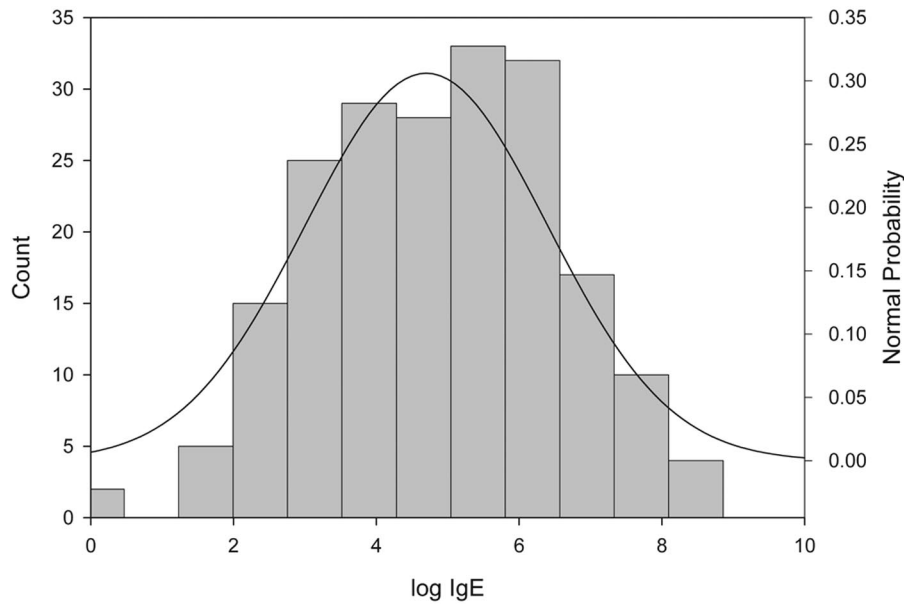


Figure 2. Simulated log-transformed IgE levels from the data of Borish et al.¹

The antilog of the arithmetic mean for log-transformed data is the geometric mean (ie, $\bar{x}_g = e^{\bar{y}}$). The geometric mean can also be computed directly from the original untransformed data as follows:

$$\bar{x}_g = n \sqrt{x_1 x_2 \cdots x_n} \quad (5)$$

The geometric and arithmetic means are measures of central tendency (ie, where the bulk of the data tends to be located). Both are estimates of the mean of the population from which the sample was drawn. However, the geometric mean has an advantage over the arithmetic mean in skewed populations. As the associated distribution of the data becomes progressively more skewed (ie, with more observations taking unusually large values), the arithmetic mean is pulled by these extreme values more and more out of the area where the bulk of the data is located. In contrast, under the same circumstances, the geometric mean is much less affected by this phenomenon and remains approximately equal to the median that, by definition, is in the exact middle of the data (half the data below it and half the data above it). Thus, as an estimate of the population mean, the geometric mean is less subject to distortion from a few unusually large values than the ordinary arithmetic mean.

The estimate of the SD based on the log-transformed data, on the other hand, has no reasonable interpretation in the original scale and consequently should be reported only for the transformed data.⁹ Instead, when summarizing results in the original scale, it is recommended to report a confidence interval. A 2-sided 95% confidence interval on the population mean of the data in the original untransformed scale is also easily computed from the log-transformed data. Just as the

geometric mean can be computed simply by exponentiation of the sample mean of the log-transformed data, a 95% confidence interval on the population mean in the original untransformed scale can be computed by exponentiation of each of the confidence bounds given in equation 4 for the log-transformed data, as follows: $(e^{\bar{y} - 1.96 \times s.e.}, e^{\bar{y} + 1.96 \times s.e.})$ or $(\bar{x}_g e^{-1.96 \times s.e.}, \bar{x}_g e^{1.96 \times s.e.})$, where $s.e. = s\sqrt{n}$ is the SEM for the log-transformed data. The lower bound of the interval cannot be less than 0, whereas the upper limit may be large, consistent with the domain of the original data. The confidence interval of the population mean in the original positively skewed data will not be symmetric about the geometric mean of the sample. In fact, the distance from the mean to the upper bound always will be larger than the distance from the mean to the lower bound and, in this context, the confidence interval will itself be positively skewed.

Two-Sample t Test and ANOVA

The hypothesis of equality of k population means (null hypothesis) can be stated as follows: $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$.

When $k = 2$, the null hypothesis is often tested using the 2-sample t test. An F test produced by ANOVA is frequently used for testing equality of means when k is greater than 2. The relevant test statistic may be computed from the data in the original scale. Alternatively, the test statistic could be computed from the data in the log-transformed scale. The resulting computations for these 2 test statistics will differ. Fortunately, this does not matter. The goals are to identify a scale in which the assumptions that underlie the test statistic are well supported and to test the hypothesis in that scale. The relevant scale is presumed in this work to be the logarithmic scale. The resulting P value will apply to the data in that scale

Table 1. Descriptive Statistics for IgE Levels by Age Category

Statistic	Adults (>17 y)	Adolescents (13–17 y)	Children (6–12 y)
$\bar{x}_g^a$	85.2	223.8	182.5
$\bar{y} = \ln(\bar{x}_g)$	4.45	5.41	5.21
s	1.69	1.49	1.82
n	3,474	496	767
95% Confidence interval			
Original scale ^a	80.5–90.1	196.4–255.1	160.4–207.6
Log scale	4.4–4.5	5.3–5.5	5.1–5.4

^a As reported in the study by Borish et al.¹

and will continue to apply to the data when back transformed by exponentiation into the original scale.

Specifically, the t test is appropriate in the situation in which the 2 samples to be compared have the same distribution except that the means, μ_1 and μ_2 , may differ. If the test statistic in the log scale rejects the hypothesis that $\mu_1 = \mu_2$, then the corresponding hypothesis is also rejected in the original scale. If the hypothesis of equality is not rejected in the log scale, then it is also not rejected in the original scale. The test statistic that may have been computed in the original scale to test this hypothesis is invalid and, therefore, not useful. Because the data after log transformation have a distribution that is better approximated by a normal distribution, that is the scale in which testing should be conducted.

The logarithmic transformation has the desirable property that for $x_1 < x_2$, which we transform by taking $y_1 = \log(x_1)$ and $y_2 = \log(x_2)$, then it is always true that $y_1 < y_2$. This implies the order of a data set is preserved after log transformation. In mathematical jargon, the logarithmic transform y is a monotonically increasing function of x . It is because of this property that (1) the P value computed after logarithmic transformation continues to apply when the data are transformed by exponentiation back to the original scale and (2) when 2-sided 95% confidence bounds computed after logarithmic transformation undergo exponentiation back to the original scale, they are valid 2-sided 95% confidence bounds on the geometric mean of the data in the original scale. There are other transformations that could be used to produce a suitable bell-shaped distribution (eg, the square root transformation). To determine whether a given transformation is appropriate and how to work with the data, a competent statistician should be consulted.

RESULTS

For illustrative purposes, we used IgE statistics found in the study by Borish et al,¹ stratified by age classes. The strata were adults (aged >17 years), adolescents (aged 13–17 years), and children (aged 6–12 years). Descriptive statistics are presented in Table 1. The sample mean, SD, and 95% confidence intervals on the log scale were deduced from the reported results. Pairwise t tests using the log-transformed data were performed on the 3 age groups, with the results listed in Table 2. For example, the test statistic for the 2-sample t test comparing IgE levels between adults and

children is computed from the information in Table 1 as follows:

$$t = \frac{\bar{x}_{\text{adult}} - \bar{x}_{\text{child}}}{s_p \sqrt{\frac{1}{n_{\text{adult}}} + \frac{1}{n_{\text{child}}}}} = -12.0 \quad (6)$$

where

$$s_p^2 = \frac{(n_{\text{adult}} - 1)s_{\text{adult}}^2 + (n_{\text{child}} - 1)s_{\text{child}}^2}{n_{\text{adult}} + n_{\text{child}} - 2} = 2.94 \quad (7)$$

is a weighted average of the variances from each strata.

DISCUSSION

For the statistical analysis of nonnegative positively skewed data, we recommend a logarithmic function or other monotonic function that transforms the original data to a scale that is suitable for statistical analysis. Traditional statistical methods based on normal distribution theory can then be applied. For descriptive purposes, the arithmetic mean and confidence intervals calculated from log-transformed data should be back transformed to their original scale through exponentiation. P values computed in the logarithmic scale remain valid in the original scale. The sample mean in the log scale transforms back to the geometric mean, and the confidence bounds around the log mean transform to corresponding bounds around the geometric mean.

Results analyzed in this fashion are improved by establishing a solid foundation for valid inferences. Validity is achieved because the assumption of a normal distribution that helps justify use of these statistical techniques often is better met when the analysis is performed using the log-transformed data.

Table 2. Pairwise t Tests for IgE Levels by Age Category^a

Age category	t Test value	P value
Adults vs adolescents	−12.0	<.001
Adults vs children	−11.1	<.001
Adolescents vs children	2.0	.04

^a Data were computed from log scale summary statistics given in Table 1.

REFERENCES

1. Borish L, Chipps B, Deniz Y, Gujrathi S, Zheng B, Dolan CM. Total serum IgE levels in a large cohort of patients with severe or difficult-to-treat asthma. *Ann Allergy Asthma Immunol*. 2005;95:247–253.
2. Koopmans JG, Lutter R, Jansen HM, van der Zee JS. Adding salmeterol to an inhaled corticosteroid reduces allergen-induced serum IL-5 and peripheral blood eosinophils. *J Allergy Clin Immunol*. 2005;116:1007–1013.
3. Karmaus W, Arshad SH, Sadeghnejad A, Twiselton R. Does maternal immunoglobulin E decrease with increasing order of live offspring? investigation into maternal immune tolerance. *Clin Exp Allergy*. 2004;34:853–859.
4. Green SA, Malice MP, Tanaka W, Tozzi CA, Reiss TF. Increase in urinary leukotriene LTE4 levels in acute asthma: correlation with air-flow limitation. *Thorax*. 2004;59:100–104.
5. Currie GP, Jackson CM, Lee DK, Lipworth BJ. Determinants of airway hyperresponsiveness in mild asthma. *Ann Allergy Asthma Immunol*. 2003;90:560–563.
6. Koh YI, Choi IS. Relationship between nasal and bronchial responsiveness in perennial allergic patients with asthma. *Int Arch Allergy Immunol*. 2002;129:341–347.
7. Langley SJ, Holden J, Derham A, Hedgeland P, Sharma RK, Woodcock A. Fluticasone propionate via the Diskhaler or hydrofluoroalkane-134a metered-dose inhaler on methacholine-induced airway hyperresponsiveness. *Chest*. 2002;122:806–811.
8. Chhabra SK, Gupta M. Prednisolone hastens recovery from histamine-induced bronchospasm in asthmatics. *J Asthma*. 2000;37:435–440.
9. Cole TJ. Sympercents: symmetric percentage differences on the 100 log(e) scale simplify the presentation of log transformed data. *Stat Med*. 2000;19:3109–3125.

Requests for reprints should be addressed to:

Jake Olivier, PhD

NSW Injury Risk Management Research Centre & Department of Statistics

University of New South Wales

Building G2, Western Campus

Western Campus Drive

WSW Sydney, Australia, 2052

E-mail: j.olivier@unsw.edu.au
