Shanghai Archives of Psychiatry, 2014, Vol. 26, No. 2

•Biostatistics in psychiatry (20)•

Log-transformation and its implications for data analysis

Changyong FENG[1]*, Hongyue WANG[1], Naiji LU[1], Tian CHEN, Hua HE[1], Ying LU[2], Xin M. TU[1]

Summary: The log-transformation is widely used in biomedical and psychosocial research to deal with skewed data. This paper highlights serious problems in this classic approach for dealing with skewed data. Despite the common belief that the log transformation can decrease the variability of data and make data conform more closely to the normal distribution, this is usually not the case. Moreover, the results of standard statistical tests performed on log-transformed data are often not relevant for the original, non-transformed data. We demonstrate these problems by presenting examples that use simulated data. We conclude that if used at all, data transformations must be applied very cautiously. We recommend that in most circumstances researchers abandon these traditional methods of dealing with skewed data and, instead, use newer analytic methods that are not dependent on the distribution the data, such as generalized estimating equations (GEE).

Key words: hypothesis testing, outliers, log-normal distribution, normal distribution, skewness

1. Introduction

The log transformation, a widely used method to address skewed data, is one of the most popular transformations used in biomedical and psychosocial research. Due to its ease of use and popularity, the log transformation is included in most major statistical software packages including SAS, Splus and SPSS. Unfortunately, its popularity has also made it vulnerable to misuse – even by statisticians – leading to incorrect interpretation of experimental results.[1] Such misuse and misinterpretation is not unique to this particular transformation; it is a common problem in many popular statistical methods. For example, the two-sample t-test is widely used to compare the means of two independent samples with normally distributed (or approximately normal) data, but many researchers take this critical assumption for granted, using t-tests without bothering to check or even acknowledge this underlying assumption. Another example is the Cox regression model used in survival analysis; many studies apply this popular model without even being aware of the proportionality assumption (i.e., the relative hazard of groups of interest is constant over time) required for valid inference.

In this article we focus on the log-transformation and discuss major problems of using this method in practice. We use examples and simulated data to show that this method often does not resolve the original problem for which it is being used (i.e., non-normal distribution of primary data) and to show that using this transformation can introduce new problems that are even more difficult to deal with then the problem of non-normal distribution of data. We conclude with recommendations of alternative analytic methods that eliminate the need of transforming non-normal data distributions prior to analysis.

2. Log-normal transformation

2.1 Using the log transformation to make data conform to normality

The normal distribution is widely used in basic and clinical research studies to model continuous outcomes. Unfortunately, the symmetric bell-shaped distribution often does not adequately describe the observed data from research projects. Quite often data arising in real studies are so skewed that standard statistical analyses of these data yield invalid results. Many methods have been developed to test the normality assumption of observed data. When the distribution of the continuous data is non-normal, transformations of data are applied to make the data as 'normal' as possible and,

doi: http://dx.doi.org/10.3969/j.issn.1002-0829.2014.02.009

[1] Department of Biostatistics and Computational Biology, University of Rochester, Rochester, New York, United States
[2] Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, United States
*correspondence: feng@bst.rochester.edu

A full-text Chinese translation will be available at www.saponline.org on May 15, 2014.

thus, increase the validity of the associated statistical analyses. The log transformation is, arguably, the most popular among the different types of transformations used to transform skewed data to approximately conform to normality.

If the original data follows a log-normal distribution or approximately so, then the log-transformed data follows a normal or near normal distribution. In this case, the log-transformation does remove or reduce skewness. Unfortunately, data arising from many studies do not approximate the log-normal distribution so applying this transformation does not reduce the skewness of the distribution. In fact, in some cases applying the transformation can make the distribution more skewed than the original data.

To show how this can happen, we first simulated data u_i which is uniformly distributed between 0 and 1, and then constructed two variables as follows:

$$x_i=100(\exp(\mu_i)-1)+1,\ y_i=\log(x_i).$$

Shown in the left panel in Figure 1 is the histogram of x_i, while the right panel is the histogram of y_i (the log-transformed version of x_i) based on a sample size of n=10,000. While the distribution of x_i is right-skewed, the log-transformed data y_i is clearly left-skewed. In fact, the log-transformed data y_i is more skewed than the original x_i, since the skewness coefficient for y_i is 1.16 while that for x_i is 0.34. Thus, the log-transformation actually exacerbated the problem of skewness in this particular example.

In general, for right-skewed data, the log-transformation may make it either right- or left-skewed. If the original data does follow a log-normal distribution, the log-transformed data will follow or approximately follow the normal distribution. However, in general there is no guarantee that the log-transformation will reduce skewness and make the data a better approximation of the normal distribution.

2.2 Using the log transformation to reduce variability of data

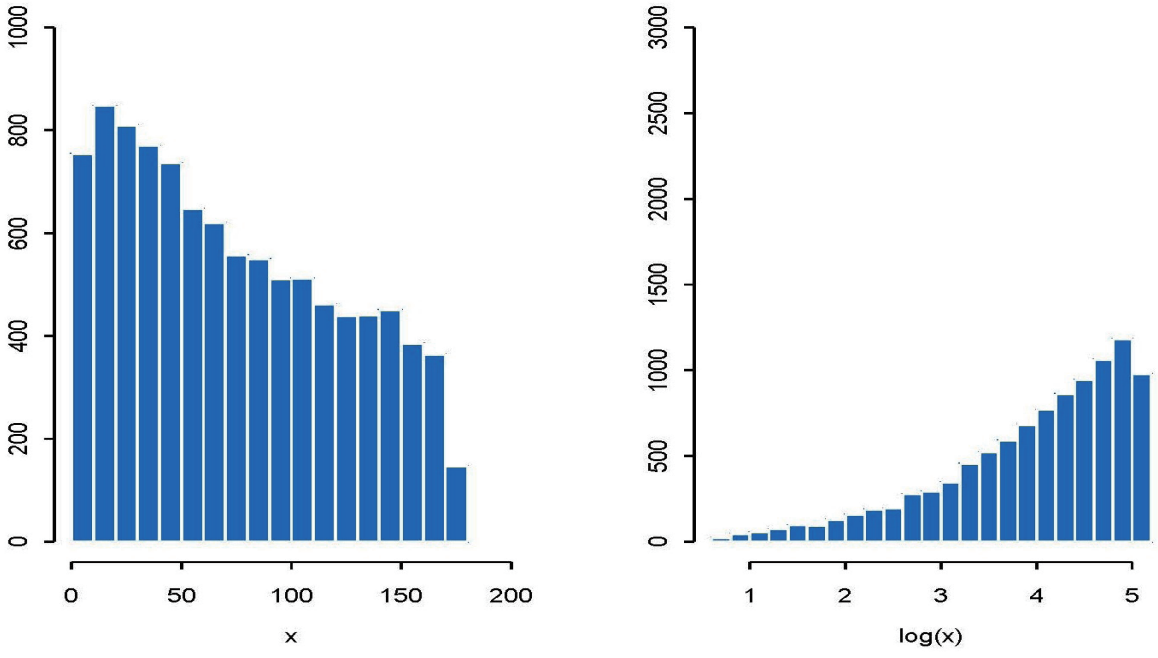
Another popular use of the log transformation is to reduce the variability of data, especially in data sets that include outlying observations. Again, contrary to this popular belief, log transformation can often increase – not reduce – the variability of data whether or not there are outliers.

For example, consider the following simple linear regression with only an intercept term:

$$y_i=\beta_0+\varepsilon i,\ \varepsilon i\sim U(-0.5, 0.5).$$

Unlike the ordinary regression analysis where the error term is assumed to have a normal distribution, the error term in this regression is uniformly distributed between -0.5 and 0.5. Thus y_i in the above model does not follow a log-normal distribution and the log-transformed y_i does not have a normal distribution. We then simulated data y_i for this model with a sample size of n=100 and a value of the β_0 parameter ranging from 0.5 to 5.5. Note that β_0 starts from 0.5, rather than from 0, to ensure $y_i>0$ and, thus, log (y_i) is correctly estimated when performing the log transformation on the data simulated from the linear regression of the original data. We fit two different linear models on the same data. The first model used the data without transformation, the second model used the log-transformed data. The

Figure 1. Histograms of original data (left plot) and log-transformed data (right plot) from a simulation study that examines the effect of log-transformation on reducing skewness.



ordinary least square method was used to estimate the intercepts in both models.

Table 1 shows the original and log-transformed estimates of β_0 and its standard errors averaged over 100,000 Monte Carlo (MC) simulations[1] from fitting the linear model to the original data. We use a large MC sample size to help reduce the sampling variability in the standard error estimates; thus the differences in the presented estimates from fitting the original and log-transformed data reflect true differences. The table shows that when β_0=0.5, the standard errors from the model fit to the original y_i were much smaller than those from fitting the log-transformed data. As β_0 increased towards 5.5, the standard errors from fitting the original data remained the same, while their counterparts from fitting the log-transformed data decreased. When β_0 increased past the value 1, the standard errors from fitting the log-transformed data became smaller than those from fitting the original data. Table 2 presents the same estimates of β_0 as those in Table 1, except that we introduced four outlying points (4, 6, 8 and 10) in the simulated data, thereby increasing the sample size to 104. As can be seen in Table 2, the estimates of β_0 and of the standard error of β_0 changed after introduction of the outliers, but the pattern of differences in these estimates between the model for the original data and for the log-transformed data remains the same. This example shows that the conventional wisdom about the ability of a log transformation of data to reduce variability especially if the data includes outliers, is not generally true. Whether the log transformation reduces such variability depends on the magnitude of the mean of the observations — the larger the mean the smaller the variability.

A more fundamental problem is that there is little value in comparing the variability of original versus log-transformed data because they are on totally different scales. In theory we can always find a transformation for any data to make the variability of the transformed version either smaller or larger than that of the original data. For example, if the standard deviation of variable x is σ, then the standard deviation of the scale transformation x/K (K>0) is σ/K; thus by selecting a sufficiently large or small K we can change the standard deviation of the transformed variable x/K to any desired level.

3. Difficulty of interpreting model estimates from log-transformed data

3.1 Estimation of model parameters

Once the data is log-transformed, many statistical methods, including linear regression, can be applied to model the resulting transformed data. For example, the mean of the log-transformed observations (log y_i), $\hat{\mu}_{LT} = (1/n)*\Sigma_{i=1}^{n} \log y_i$ is often used to estimate the population mean of the original data by applying the anti-log (i.e., exponential) function to obtain $\exp(\hat{\mu}_{LT})$. However, this inversion of the mean log value does not usually result in an appropriate estimate of the mean of the original data. For example, as shown by Feng and colleagues,[2] if y_i follows a log-normal distribution (μ, σ^2), then the mean of y_i is given by $E(y_i)=\exp(\mu + \sigma^2/2)$. If we log-transform y_i, the transformed log y_i follows a normal distribution with a mean of μ. Thus, the sample mean of the log-transformed data, $\hat{\mu}_{LT}=(1/n)*\Sigma_{i=1}^{n} \log y_i$ is an unbiased estimate of the mean μ of log y_i , and the exponential function of $\hat{\mu}_{LT}$, that is, $\hat{\mu}=\exp(\hat{\mu}_{LT})$, is an

Table 1. Simulation results for simple linear regression without outliers (n=100; 100,000 simulations)

β_0	original data		log-transformed data	
	Estimated Intercept	SE	Estimated Intercept	SE
0.50	0.5000	0.0288	−0.9999	0.0998
0.51	0.5100	0.0289	−0.9440	0.0887
0.55	0.5499	0.0289	−0.7993	0.0718
0.60	0.6001	0.0290	−0.6647	0.0608
0.70	0.7002	0.0289	−0.4591	0.0480
0.80	0.8000	0.0288	−0.2977	0.0401
0.90	0.8999	0.0288	−0.1626	0.0347
1.00	1.0001	0.0288	−0.0451	0.0307
1.50	1.5000	0.0289	0.3863	0.0198
5.50	5.5000	0.0289	1.7034	0.0053

Table 2. Simulation results for simple linear regression with outliers (n=104; 100,000 simulations)

β_0	original data		log-transformed data	
	Estimated Intercept	SE	Estimated Intercept	SE
0.50	0.7501	0.0277	−0.8886	0.0960
0.51	0.7599	0.0277	−0.8350	0.0849
0.55	0.7999	0.0277	−0.6956	0.0689
0.60	0.8500	0.0278	−0.5660	0.0585
0.70	0.9500	0.0278	−0.3678	0.0461
0.80	1.0499	0.0277	−0.2119	0.0386
0.90	1.1500	0.0278	−0.0811	0.0335
1.00	1.2501	0.0277	0.0323	0.0296
1.50	1.7499	0.0278	0.4497	0.0190
5.50	5.7501	0.0278	1.7328	0.0051

estimate of exp(μ). However, the mean of the original data y_i is exp($\mu+\sigma^2/2$), not exp(μ). Thus, even in this ideal situation, estimating the mean of the original y_i using the exponent or anti-log of the sample mean of the log-transformed data can generate inaccurate estimates of the true population mean of the original data.

3.2 Hypothesis testing with log-transformed data

It is also more difficult to perform hypothesis testing on log-transformed data. Consider, for example, the two sample t-test, which is widely used to compare the means of two normal (or near normal) samples. If the two samples have the same variance, the test statistic has a t-distribution. For skewed data (when the variance of samples is usually different), researchers often apply the log-transformation to the original data and then perform the t-test on the transformed data. However, as demonstrated below, applying such a test to log-transformed data may not address the hypothesis of interest regarding the original data.

Let y_{1i} and y_{2i} denote two samples. If the data from both samples follow a log-normal distribution, with log-normal (μ_1, σ_1^2) for the first sample and (μ_2, σ_2^2) for the second sample, then the first sample has the mean exp($\mu_1+\sigma_1^2/2$) and the second has the mean exp($\mu_2+\sigma_2^2/2$). If we apply the two-sample t-test to the original data, we are testing the null hypothesis that these two means are equal, H_0: exp($\mu_1+\sigma_1^2/2$)= exp($\mu_2+\sigma_2^2/2$).

If we log-transform the data, the transformed data have the mean μ_1 and variance σ_1^2 for the first sample and mean μ_2 and variance σ_2^2 for the second sample. Thus, if we apply the two-sample t-test to the transformed data, the null hypothesis of the equality of the means becomes, H_0: $\mu_1=\mu_2$.

The two null hypotheses are clearly not equivalent. Although the null hypothesis based on the log-transformed data does test the equality of the means of the two log-transformed samples, the null hypothesis based on the original data does not, since the mean of the original data also involves the parameters, σ_1^2 and σ_2^2. Thus, even if no difference is found between the two means of the log-transformed data, it does not mean that there is no differences between the means in the original data of the two samples. For example, if the null hypothesis for the log-transformed data, H_0: $\mu_1=\mu_2$, is not rejected for the log-transformed data, it does not imply that the null hypothesis for comparing the means of the original data of the samples, H_0: exp($\mu_1+\sigma_1^2/2$)= exp($\mu_2+\sigma_2^2/2$), is true, unless the variances of the two samples are the same.

3.3 Effect of adding a small constant to data when performing log transformations of data

Since the log transformation can only be used for positive outcomes, it is common practice to add a small positive constant, *M*, to all observations before applying this transformation. Although appearing quite harmless, this common practice can have a noticeable effect on the level of statistical significance in hypothesis testing.

We examine the behavior of the p-value resulting from transformed data using a simulation. We simulated data from two independent normal distributions, with sample size n=100. The data is generated in the following way: (1) generate two independent random numbers u_i and v_i (i=1, …, n), where u_i has a standard normal distribution and v_i has a normal distribution with mean of 1 and a standard deviation of 2; (2) generate y_{i1} and y_{i2} according to the following formulas:

$$y_{i1}=\exp(\mu_i)+15,\ y_{i2}=\exp(y_i)+13.$$

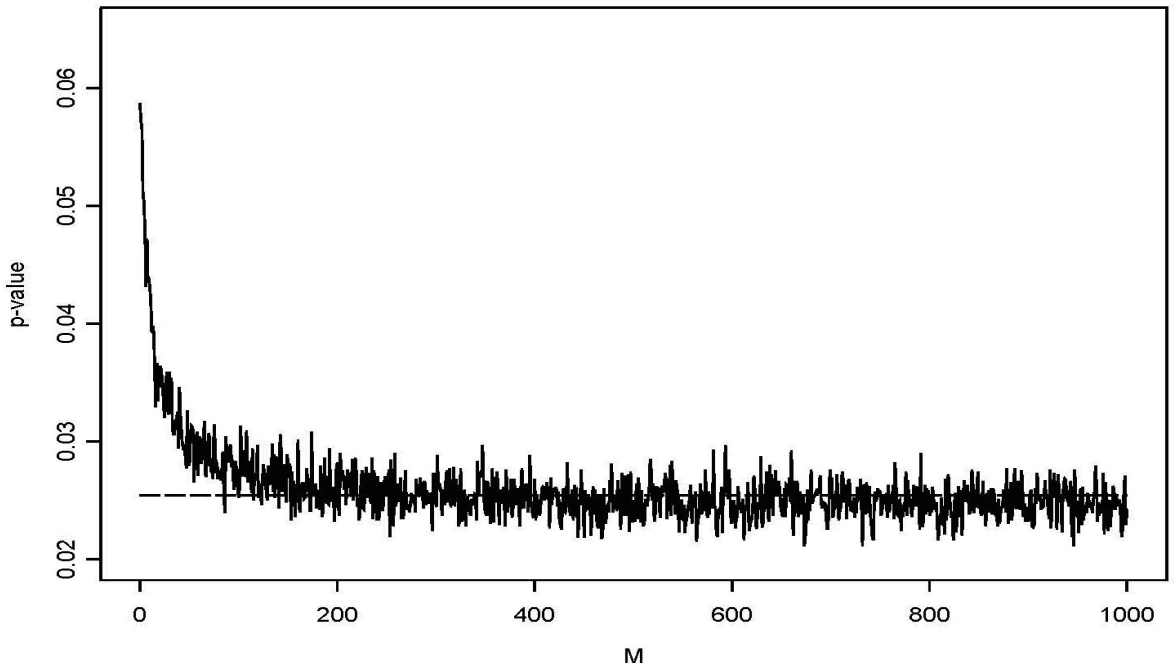
We then added a constant, *M*, to each observation of y_{i1} and y_{i2} before the data were log transformed. Figure 2 shows the p-values from comparing the means of the log-transformed data from the two samples, based on using different values of M. When *M*=0, the p-value for the difference in the means of the two samples of log-transformed data is 0.058, that is, the difference was not statistically significant at the usual type I error level of α=0.05. However, as *M* increases the p-values dropped and fell below the 0.05 threshold for statistical significance after it rose above 100. This simulation study indicates that the p-value of the test depends on what value is added to the data before applying the log-transformation, potentially making conclusions about differences between groups dependent on the somewhat arbitrary decision of the researcher about the size of *M* to be used in the analysis.

4. Discussion

Using transformations in general and log transformation in particular can be quite problematic. If such an approach is used, the researcher must be mindful about its limitations, particularly when interpreting the relevance of the analysis of transformed data for the hypothesis of interest about the original data. For example, we have demonstrated that in most circumstances the log transformation does *not* help make data less variable or more normal and may, in some circumstances, make data more variable and more skewed. Furthermore, log-transformed data cannot usually facilitate inferences concerning the original data, since it shares little in common with the original data.

For many applications, rather than trying to find an appropriate statistical distribution or transformation to model the observed data, it would probably be better to abandon this classic approach and switch to modern distribution-free methods. For example, a popular approach that can avoid many of these problems is the generalized estimating equations, or GEE.[3,4] This approach forgoes the distribution assumption, providing valid inference regardless of the distribution of the data. However, this is only appropriate for skewed data, if the data can be reasonably modeled by a parametric distribution such as the normal distribution, it is preferable to use the classic statistical methods because they usually provide more efficient inference than GEE.

Figure 2. P-values as a function of values added to the data before applying log-transformation.



Conflict of interest

The authors report not conflict of interest related to this manuscript.

Funding

This research was supported in part by the Novel Biostatistical and Epidemiologic Methodology grants from the University of Rochester Medical Center Clinical and Translational Science Institute Pilot Awards Program.

数据分析中的对数转换和意义

Changyong FENG, Hongyue WANG, Naiji LU, Tian CHEN, Hua HE, Ying LU, Xin M. TU

摘要：对数转换的方法在生物医学和社会心理研究中处理非正态数据时被广泛应用。本文重点介绍该传统方法在处理非正态数据时存在的严重问题。尽管通常认为对数转换可以减少数据的变异性，使数据更符合正态分布，但是通常并非如此。此外，对数转换后的数据得出的标准统计测试结果往往和未转化的原始数据不相关。我们通过使用模拟数据示例来说明这些问题。我们认为如果采用数据转换，必须非常谨慎应用。我们建议研究者在大多数情况下摒弃这些处理非正态数据的传统方法，选择采用较新的不依赖于数据分布的方法：如广义估计方程（GEE）。

关键词：假设检验，离群值，对数正态分布，正态分布，偏度

本文全文中文版从 2014 年 5 月 15 日起在 www.saponline.org 可供免费阅览下载

References

1. Robert CP, Casella G. *Monte Carlo Statistical Methods (2nd edition)*. New York: Springer. 2004
2. Feng C, Wang H, Lu N, Tu XM. Log-transformation: applications and interpretation in biomedical research. *Statistics in Medicine*. 2012; **32**: 230-239. doi: http://dx.doi.org/10.1002/sim.5486
3. Kowalski J, Tu XM. *Modern Applied U Statistics*. New York: Wiley. 2007
4. Tang W, He H, Tu XM. *Applied categorical and count data analysis*. FL: Chapman & Hall/CRC. 2012

Changyong Feng received his BSc in 1991 from the University of Science and Technology of China and subsequently obtained a PhD in statistics from the University of Rochester in 2002. He is currently an associate professor in the Department of Biostatistics and Computational Biology at University of Rochester. The main focus of his research is on survival analysis.