Introduction to Log Transformation

Shanzhao Qiao
email: sq2224@columbia.edu

1 Background Information

1.1 Normal Distribution

The normal distribution is also named as Gaussian distribution, Gauss distribution, which is a continuous probability distribution. The normal distribution is widely used in many scientific studies. It can be defined with just two parameters, mean and variance. Its mean, median and mode have the same value. The normal distribution also has important implications, central limit theorem (CLT), which is key to probability theory. However, data in real life does not always strictly follow a normal distribution. They are usually skewed, which makes statistical analysis hard.

1.2 Logarithm

In mathematics, the logarithm is the inverse function of the exponential function. It can be defined with a given number x and a base b, where b is not 1. Logarithmic function can be written mathematically as $y = \log_b x$. This means that b to the power of y equals to x. More explicitly, the relationship between logarithm function and exponential function is: if $x = b^y$ and $x > 0$, $b > 0$, $b \neq 1$, define $y = \log_b x$ to be the logarithmic function with base b of x. Commonly used logarithm function usually have based 2, 10 or e(natural number).
Example:
Base 2, $y = \log_2 x$: $\log_2 8 = 3$
Base 10, $y = \log_{10} x$: $\log_{10} 100 = 2$
Base e, $y = \log_e x = \ln x$: $\ln 7.389 = 2$

2 Introduction

As mentioned previously in background information, real-life data are often skewed, which makes analytic research hard to perform. Therefore, Log transformation is usually introduced to reduce the skewness of real-life data. Due to its ease of use and popularity, log transformation is included in most major statistical software packages including SAS, Splus, and SPSS.[1] In statistics, data transformation is performed to every point in a data frame in order to achieve some specific goals. Each data point z is replaced with the transformed value $y_i = f(z_i)$. [2] In this case, where we apply log transformations, i.e. $y_i = \log z_i$, the goal is to reduce the skewness and improve the interpretability or appearance of graphs.

3 Definition of Terms

3.1 Continuous/Discrete variable

3.1.1 Discrete Variable

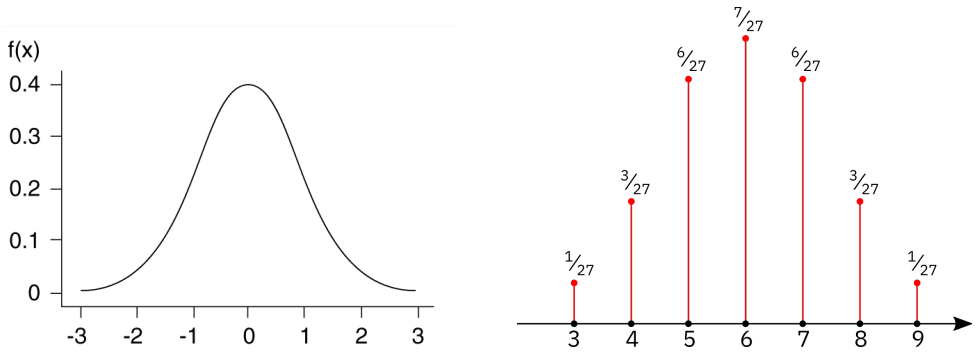
Discrete variables are countable in a finite amount of time.[5] Discrete variables are over a particular range of real values. The value that is permitted to take is either finite or countably infinite. In statistics, probability mass functions can be used to describe the probability distributions of discrete variables.

3.1.2 Continuous Variable

Continuous variables are impossible to count.[5] There are an infinite amount of numbers. Mathematically speaking, a continuous variable can take on an uncountable set of values. In statistics, the probability density function is used to describe the probability distributions of continuous variables.

3.1.3 Examples

As shown in figure 1, figure 1a is the PDF of a continuous normal distribution and figure 1b is the PMF of a discrete normal distribution taken using 7 points.



(a) PDF of continuous normal [8] distribution

(b) PMF of discrete normal distribution

Figure 1: Examples of continuous variables and discrete variables

3.2 Log transformation

Log transformation is a data transformation method in which it replaces each variable x with a log(x). The base of the logarithm is not fixed as it depends on the purposes of statistical modeling and usually is a choice of the analyst.

As mentioned, real-life data usually appears to be skewed and hard to perform analysis. Log transformation can reduce the skewness and transform the original data into a bell curve. One thing to note, the original curve needs to behave as a log-normal distribution. Otherwise, log transformation will not function as it is intended.

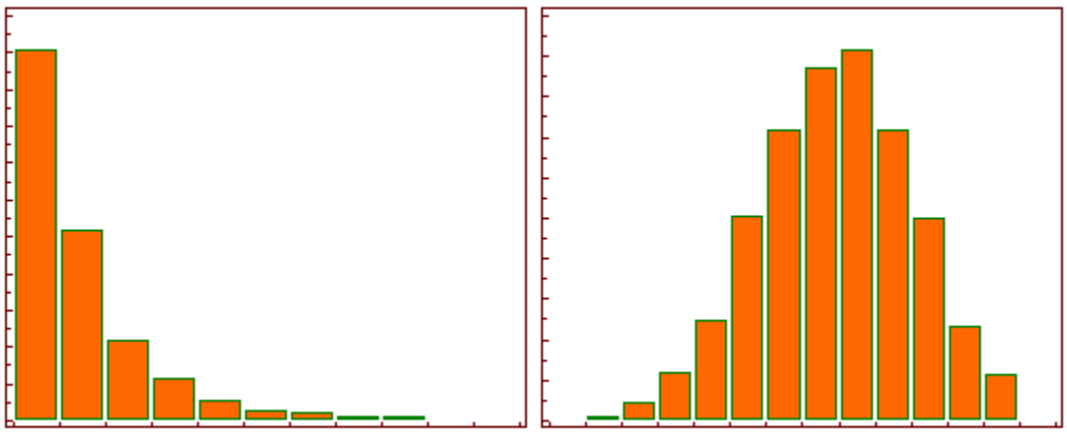
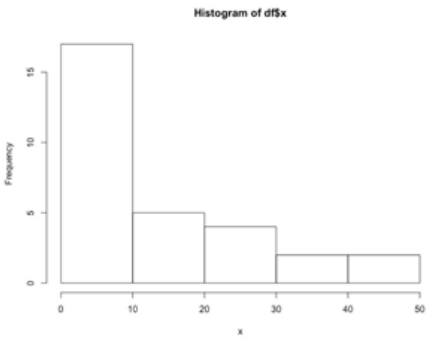


Figure 2: Log transformation on skewed data distribution [12]

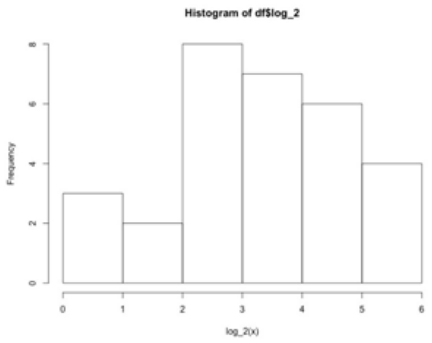
4 Example

4.1 Comparison between log transformed data and original data

The following is a data set containing a variable x, as well as data after log-transformation.

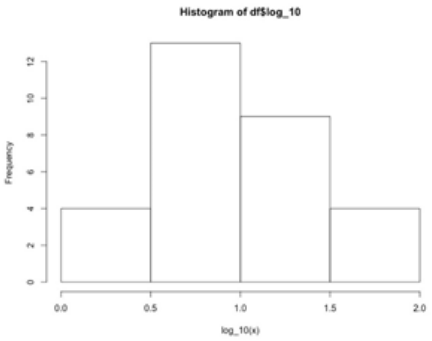


(a) Histogram of x

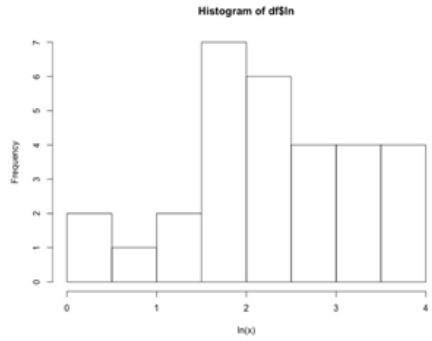


(b) Histogram of $\log_2 x$

Figure 3: Histogram of x and $\log_2 x$



(a) Histogram of $\log_1 0x$



(b) Histogram of $\ln x$

Figure 4: Histogram of $\log_{10} x$ and $\ln x$

5 Properties

5.1 Basic properties of logarithm

$$\log_b xy = \log_b x + \log_b y$$

$$\log_b \frac{x}{y} = \log_b x - \log_b y$$

$$\log_b x^n = n \log_b x$$

$$\log_b x = \log_a x + \log_a b$$

5.2 Change in natural log ≈ percentage change

From calculus, we are aware that function e^x is its own derivative, and the derivative of $\ln x$ is $1/x$. For business analysis, small changes in the natural log of a variable can be directly interpretable as percentage changes, since they are really close to each other. [13] The reason behind this is that $y = \ln x$ passes through (1,0) with a slope of 1. Figure 5 shows that the value around point (0,1) are very close for both functions. This implies that $\ln(1 + r) \approx r$.

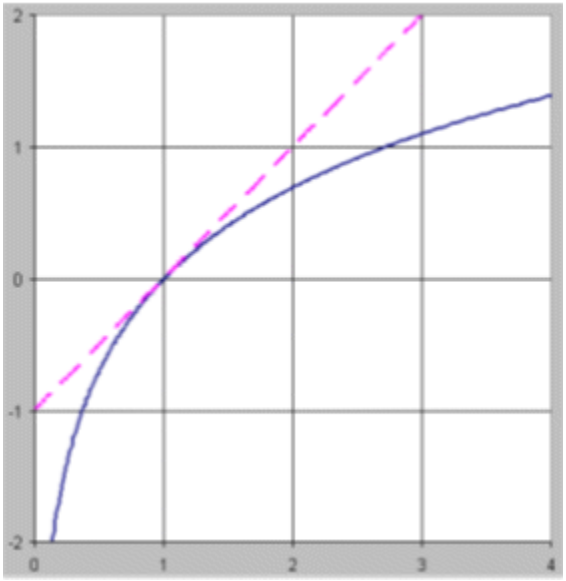


Figure 5: Log transformation on skewed data distribution
[13]

5.3 Linearization of exponential growth and inflation

From the basic property of logarithm, we know that $\log xy = \log x + \log y$ regardless of the base. This relationship converts multiplicative relationships into additive relationships. Applying log transformation on a series usually affects similar to deflating. It will deflate the exponential pattern and stabilize the variance. Also log transformations do not need any extra data for doing that.[13]

6 Estimations

6.1 Estimation of model parameters

Even though many statistical methods can be applied after data log-transformation, it will also impose many difficulties. For example, the mean of the log-transformed observations $\log y_i$, $\mu_{LT} = \frac{1}{n}\sum_{i=1}^{n} \log y_i$ is often used to estimate the population mean of the original data by applying the anti-log (i.e., exponential) function to obtain $e^{\mu_{LT}}$. However, this mean log value sometimes cannot be used as an appropriate estimate. For example, as shown by [10], if y_i follows a log-normal distribution (μ, σ_2), then the mean of y_i is given by $E(y_i) = e^{\mu + \frac{\sigma_2}{2}}$. The mean of the original data y_i is $e^{\mu + \frac{\sigma_2}{2}}$. However, applying the anti log function, the estimation mean is e^{μ}.

7 Interpretation and Applications

7.1 Outcome variable is log transformed

Usually, a linear relationship is hypothesized between a log transformed outcome variable and a group of predictor variables. Mathematically, the relationship should be:

$$\log y = \beta_0 + \beta_1 x_1 ... + \beta_k x_k$$

The interpretation of this should be easy since this is only a simple regression. For example, we have a relationship between x and y as $\ln y = 10 + 0.12x$. We also have two inputs, x_1 and x_2, and the corresponding output values are y_1 and y_2. Then we can do some quick calculations.

$$\ln y_2 - \ln y_1 = (10 + 0.12x_2) - (10 + 0.12x_1)$$

$$\frac{y_2}{y_1} = e^{0.12(x_2 - x_1)}$$

Therefore, one unit increase in x will increase the dependent variable by 1.127. [9] Equivalently, log-level regression can be seen as a special form of exponential regression. By taking the exponential of each side of the equation, we get:

$$y = e^{b_0 + b_1 x_1 + ... + b_k x_k}$$

Based on [6], it says that when the outcome variable is log-transformed, it is natural to interpret the exponentiated regression coefficients. These values correspond to changes in the ratio of the expected geometric means of the original outcome variable.

7.2 Some predictor variables are log transformed

When the outcome is not log transformed and Mathematically, the relationship should be:

$$y_i = \beta_0 + \beta_1 \log x_1 ... + \beta_k \log x_k$$

We can also use one simple example to see the relationship between y increase and x increase. The relationship between x and y is $\ln y = 10 + 0.12x$. We have two inputs, x_1 and x_2, and the corresponding output values are y_1 and y_2.

$$y_2 - y_1 = (10 + 0.12 \ln x_2) - (10 + 0.12 \ln x_1)$$

$$y_2 - y_1 = 0.12(\ln x_2 / x_1)$$

Based on these equations, we can see that for a 10 percent increase in x, y would increase by 0.011 unit. [9]

7.3 Both the outcome variable and some predictor variables are log transformed

If the outcome and predictors are all log transformed, using mathematical expression, the relationship should be:

$$\log y_i = \beta_0 + \beta_1 \log x_1 ... + \beta_k \log x_k$$

Similarly, log-log regression can be transformed into power regression. If taking exponential of both sides of the equation, we get:

$$y = e^{\beta_0} x_1^{\beta_1} x_2^{\beta_2} ... x_k^{\beta_k}$$

$$\Rightarrow y = C e^{\beta_0} x_1^{\beta_1} x_2^{\beta_2} ... x_k^{\beta_k}$$

We will also use an example for this type of transformation. Like what we did above, we have the function: $\ln y = 10 + 0.12 \ln x$.We also have two inputs, x_1 and x_2, and the corresponding output values are y_1 and y_2. Then we will also do some quick calculations.

$$\ln y_2 - \ln y_1 = (10 + 0.12 \ln x_2) - (10 + 0.12 \ln x_1)$$

$$\frac{y_2}{y_1} = (x_2 - x_1)^{0.12}$$

This shows us that for a 10 percent increase of x, y will increase 75 percent. [9]

8 Application Example

8.1 Log Transformation in image processing

Log transformation is widely used in many files. It is popular and useful in the field of image processing. Log transformation can be used to change the grayscale of a picture. Applying log transformation to gray images can expand the parts with a lower grayscale and compress parts with a higher grayscale. Figure 6 compares the same picture under different log transformations. It is easy to observe that with a larger base, parts with a lower grayscale will be emphasized more.

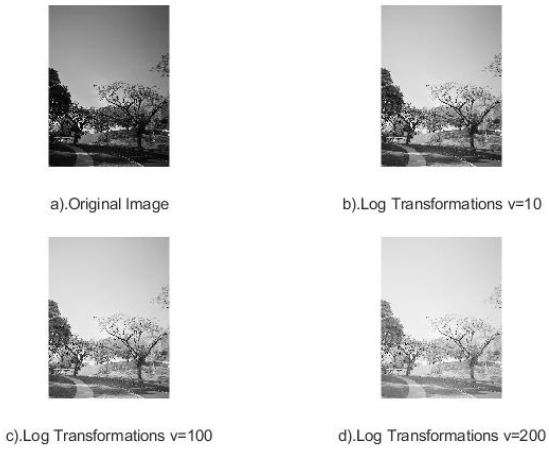
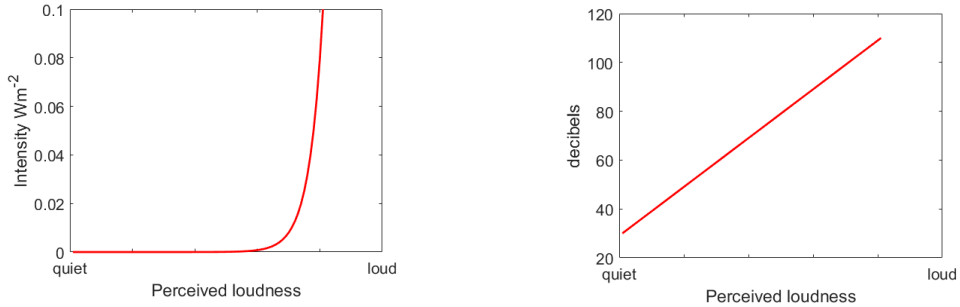


Figure 6: Gray picture with different log transformations
[4]

8.2 Sound intensity

In real life, there is another application of log transformation in our daily life. In high school, we learned about sound intensity, which is defined by $I = \frac{P}{A}$. [3] Yet another commonly variable that is used to describe the sound is sound intensity level. Sound intensity level is proportional to perceived loudness. As shown in figure 7, in order to find the gradient of the perceived loudness of a specific person. Log transformation is used to transform sound intensity into sound intensity level so that other analyses may be performed.



(a) Sound intensity vs perceived loudness (b) Sound intensity vs perceived loudness

Figure 7: Sound Intensity
[3]

9 History

The history of logarithms can be traced back to seventeenth-century Europe. Both John Napier and Joost Burgi invented the logarithms. But they are slightly different from each other. John Napier publicly announced the method of logarithms in 1614 in a book titled Mirifici Logarithmorum Canonis Descriptio.[11] Joost Burgi pushlished his method in 1620. Napier's approach was algebraic and Burgi's approach was geometric. Both methods were quite different from current definition of logarithms.The possibility of defining logarithms as exponents was recognized by John Wallis in 1685 and by Johann Bernoulli in 1694. [7]

10 Criticism and Limitations

Even though log transformation can be applied widely, it is not a good idea to always use log transformation regardless of the situation. If this technique is used improperly, the result dataset may become more skewed and harder to perform analysis.
As mentioned previously, if the original data follows the log-normal distribution, the log-transformed data will follow a normal or near-normal distribution, which reduces the skewness. However, if the original data does not follow a log-normal distribution, the result of transformation may be more skewed. Examples in [1] is a good illustration. The first step is to simulate u_i. Uniformly distribute u_i between 0 and 1. Then construct two variables as follows, $x_i = 100\exp(u_i - 1) + 1$,$y_i = \log(x_i)$. Figure 8 is the result of the simulated dataset. Based on the result, x_i has a skewness coefficient of 0.34. But y_i has a coefficient of 1.16. This can show that log transformation sometimes will cause more skewness in a particular example.

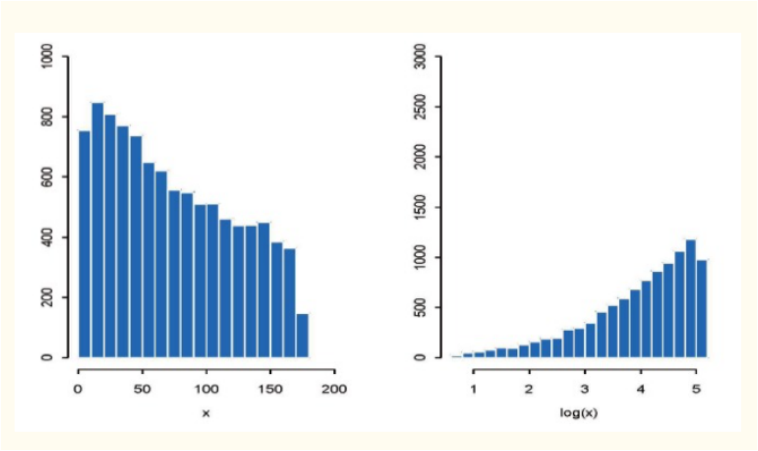


Figure 8: Result of simulated data
[1]

11 Ethical Considerations

Statistical analysis is widely used in many fields. The clinic is such an example. However, if the statistical analysis is not used properly, the result maybe misleading and result in further consequences.

12 References

References

[1] Naiji LU Changyong FENG Hongyue WANG. "Log-transformation and its implications for data analysis". In: *Shanghai Arch Psychiatry. 2014 Apr; 26(2)* (). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/.

[2] *Data transformation (statistics)*. URL: https://en.wikipedia.org/wiki/Data_transformation_(statistics).

[3] *Decibel Scale*. URL: http://salfordacoustics.co.uk/sound-waves/waves-transverse-introduction/decibel-scale.

[4] *Digital image processing*. URL: https://blog.csdn.net/d_turtle/article/details/79737873.

[5] *Discrete vs Continuous variables: How to Tell the Difference*. URL: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/discrete-vs-continuous-variables/.

[6] "FAQ HOW DO I INTERPRET A REGRESSION MODEL WHEN SOME VARIABLES ARE LOG TRANSFORMED?" In: (). URL: https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/.

[7] "History of Logarithms". In: (). URL: https://www.teachengineering.org/content/van_/lessons/van_bmd_less2/history_of_logarithms_handout.pdf.

[8] *Illustration of the central limit theorem*. URL: https://en.wikipedia.org/wiki/Illustration_of_the_central_limit_theorem#Probability_mass_function_of_the_sum_of_three_terms.

[9] *Log Transformation: Purpose and Interpretation*. URL: https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9.

[10] "Log-transformation: applications and interpretation in biomedical research". In: (). URL: https://doi.org/10.1002/sim.5486.

[11] *Logarithm*. URL: https://en.wikipedia.org/wiki/logarithm.

[12] *Logarithmic transformation*. URL: https://www.medcalc.org/manual/log_transformation.php.

[13] *The logarithm transformation*. URL: https://people.duke.edu/~rnau/411log.htm.