Lecture 5

Log Transformations

In this lecture, we will discuss the most common transformation made to data to eliminate skewness and improve normality - the logarithm transformation. The theoretical justification of this transformation is that your errors are occurring on a multiplicative rather than additive scale. This occurs all the time - salaries, for instance, are often percentage pay raises of your existing salary making their error from their trend, over time, multiplicative rather than linear. Many other phenomena naturally occur on the multiplicative rather then the linear scale. Such data are bounded below by zero and are right skewed. We will discuss the interpretation of the results after a log transformation has been made which, on the original scale, involves the fold increase (or decrease) in median value.

Mathematical facts for logarithms

All logarithms are multiples of each other, so, for mathematicians, there is but one logarithm, the natural logarithm: $\log_e(x) = \ln(x) = \log(x)$. When we use log we will mean the natural logarithm. When we want to use any other base, we will specify it. Sometimes, we will use ln for the natural logarithm, especially if there are logarithms involving other bases around.

When $y = \log x$ then $x = e^y$. This relationship gives us various rules for handling logarithms. For instance $\log(ab) = \log a + \log b$ because when you multiply numbers with the same base you add their exponents. That is, let $x = \log a$ and $y = \log b$. Then $a = e^x$ and $b = e^y$ and $ab = e^x e^y = e^{x+y}$ so that $\log(ab) = x + y = \log a + \log b$.

Similarly, $\log(a/b) = \log a - \log b$.

As for changing bases, if $x = \log_2 y$ then $y = 2^x$. But $2 = e^{\ln 2}$ so $y = \left(e^{\ln 2}\right)^x = e^{x \ln 2}$. Thus, $\ln y = x \ln 2$ and $x = \log_2 y = \frac{\ln y}{\ln 2}$. In general, $\log_b y = \frac{\ln y}{\ln b}$.

Multiplicative versus additive errors

Data which is normally distributed has an additive error structure. That is, if $X_1, X_2, \ldots, X_n$ is normal with mean μ and standard deviation σ, then $X_i = \mu + \epsilon_i$ where ϵ_i is the error and is normally distributed with mean 0 and standard deviation σ.

Data which is log-normally distributed (for which the logarithm is normally distributed) has a multiplicative error structure. That is, if $\log X_1, \log X_2, \ldots, \log X_n$ are normally distributed with mean μ and standard deviation σ then $\log X_i = \mu + \epsilon_i$ so that $X_i = e^{\mu+\epsilon_i} = e^{\mu} \times e^{\epsilon_i}$. Multiplicative error structures occur naturally - pay raises are often percentage increases over your current salary, population growth is a percentage increase over the current size, etc... Data which is log normally distributed is bounded below by zero and skewed to the right.

Statistical interpretation

Consider your data: initially you have two sets of data: $x_1, x_2, \ldots, x_{n_x}$ and $y_1, y_2, \ldots, y_{n_y}$. If a log transformation is appropriate, then after the transformation, you are comparing two new sets of data: $\log x_1, \log x_2, \ldots, \log x_{n_x}$ and $\log y_1, \log y_2, \ldots, \log y_{n_y}$. The question a t-test would answer on the transformed data is whether $\text{mean}(\log X) = \text{mean}(\log Y)$. How does this relate to the original data? By the rules of logarithms, $\text{mean}(\log X) = (\log x_1 + \log x_2 + \cdots \log x_{n_x})/n_x = \log(\sqrt[n_x]{x_1 x_2 \cdots x_{n_x}})$ which is the logarithm of the geometric mean, not of the arithmetic mean. We will return to this observation at the end of this lecture.

However, if the log transformation is successful and the logged data is symmetrical, then $\text{mean}(\log X) \approx \text{median}(\log X) \approx \log(\text{median}(X))$. So that the t-test for the equality of the means from two log transformed data sets can be interpreted as a test for the equality of the median values on the original scale.

Thus, when comparing the means of two log-transformed data sets, we have

$$\begin{aligned} a &= \text{mean}(\log X) - \text{mean}(\log Y) \\ &\approx \text{median}(\log X) - \text{median}(\log Y) \\ &= \log(\text{median}X) - \log(\text{median}Y) \\ &= \log \frac{\text{median}X}{\text{median}Y} \end{aligned}$$

so that

$$\frac{\text{median}X}{\text{median}Y} = e^a = C$$

or, equivalently,

$$\text{median}X = C \times \text{median}Y.$$

If $C > 1$, we say that the median of X is C-fold (or times) higher than the median of Y. If $C < 1$, we could say that X is C-fold (or times) lower than the median of Y or we could invert C and say that the median of Y is $1/C$-fold (or times) higher than the median of X.

The following example will involve the $100 \times (1-\alpha)\%$ confidence interval for the difference in two means based on the t-distribution. The general formula is:

$$(\bar{X} - \bar{Y} \pm t_{\frac{\alpha}{2},\text{d.f.}} \text{S.E.}(\bar{X} - \bar{Y})$$

where $t_{\frac{\alpha}{2},\text{d.f.}}$ is the cut off from the t-distribution with the given degrees of freedom that puts probability $\alpha/2$ in the right tail, and the degrees of freedom and the standard error come from one of the versions of the t-test given in Lecture 3.

Example 1 The following data are originally from Kapitulnik *et al.* 1976 and used in Biostatistics by van Belle *et al.* The data are the rate of metabolism (nmol $3H_2O$ formed/g per hour) of the drug zoxazolamine in the placentas of women who smoked and who didn't smoke during pregnancy.

Non-smoker	Smoker
0.18	0.66
0.36	0.60
0.24	0.96
0.50	1.37
0.42	1.51
0.36	3.56
0.50	3.36
0.60	4.86
0.56	7.50
0.36	9.00
0.68	10.08
	14.76
	16.50

The data from the smoking group just beg for a log transformation and the data from the non-smokers are not hurt by such a transformation. NOTE: it only makes sense to log transform both data sets if you log transform one of them. For the log transformed data, any statistical software package can spit out Welch's version of the confidence interval and you see that the confidence interval for the difference in the means of the logged data from smokers and the logged data from non-smokers is $(1.39, 2.86)$. Exponentiating, we see that the confidence interval for

$$\frac{\text{median rate for smokers}}{\text{median rate for non-smokers}} \text{ is } (e^{1.39}, e^{2.86}) = (4.0, 17.4)$$

In words, the median rate of metabolism for smokers is 4 to 17.4 times higher than the median rate for non-smokers (based on a 95% confidence interval of the logged data using Welch's t-statistic with 15 degrees of freedom.)

Another way of interpreting the result using words is to say that the rate of metabolism for smokers is 300-1640% higher than that for non-smokers. (For "percent higher," you subtract 1 from the fold-increase and then multiply by 100. This works when there is an increase.)

Yet another way of viewing the data is to look at the confidence interval for the ratio of the median rate for non-smokers to the median rate for smokers. This is the reciprocal of the confidence interval above: $(0.057, 0.25)$. In words, the median rate of metabolism for a non-smoker is between 0.057 and 0.25 times the rate for a smoker. That sound awkward. Another way is to day that the median rate of metabolism for a non-smoker is 75-94.3% lower than that of a smoker. (For "percent lower," you subtract the fold-decrease from 1 and then multiply by 100.) ■

The interpretations above for the data on the original scale involve the medians because, as noted before, if the log transformed worked, then the logged data is symmetrical and the mean of the logged data is approximately the median of the logged data and, in mathematical terms, "median" and "logarithm" commute (that is, the log of the median is the median of the logs.) Another possible interpretation is to work with the mathematical fact that the mean of the logged data is the geometric rather than arithmetic mean of the original data. There is dispute, however, whether geometric means have useful interpretations. IU SPEA faculty member, David Parkhurst, has written at least one paper on this subject. See the reference provided below.

References and Readings

[1] J. Kapitulnik, W. Levin, P. J. Poppers, J. E. Tomaszewski, D. M. Jerina, and A. H. Conney. Comparison of the hydroxylation of zoxazolamine and benzo[a]pyrene in human placenta: effect of cigarette smoking. *Clinical Pharmacology and Therapeutics*, 20:557–564, 1976.

[2] David F. Parkhurst. Arithmetic versus geometric means for environmental concentration data. *Environmental Science and Technology: News and Research Notes*, pages 92A–98A, February 1 1998.

[3] Gerald van Belle, Lloyd Fisher, Patrick J. Heagerty, and Thomas Lumley. *Biostatistics: A Methodology for the Health Sciences*. John Wiley and Sons, Inc., 2^{nd} edition, 2004.

Exercises for Lecture 5

1. –

2. –