16. Transformations

A. Log

B. Tukey's Ladder of Powers

C. Box-Cox Transformations

D. Exercises

The focus of statistics courses is the exposition of appropriate methodology to analyze data to answer the question at hand. Sometimes the data are given to you, while other times the data are collected as part of a carefully-designed experiment. Often the time devoted to statistical analysis is less than 10% of the time devoted to data collection and preparation. If aspects of the data preparation fail, then the success of the analysis is in jeopardy. Sometimes errors are introduced into the recording of data. Sometimes biases are inadvertently introduced in the selection of subjects or the mis-calibration of monitoring equipment.

In this chapter, we focus on the fact that many statistical procedures work best if individual variables have certain properties. The measurement scale of a variable should be part of the data preparation effort. For example, the correlation coefficient does not require the variables have a normal shape, but often relationships can be made clearer by re-expressing the variables. An economist may choose to analyze the logarithm of prices if the relative price is of interest. A chemist may choose to perform a statistical analysis using the inverse temperature as a variable rather than the temperature itself. But note that the inverse of a temperature will differ depending on whether it is measured in °F, °C, or °K.

The introductory chapter covered linear transformations. These transformations normally do not change statistics such as Pearson's r, although they do affect the mean and standard deviation. The first section here is on log transformations which are useful to reduce skew. The second section is on Tukey's ladder of powers. You will see that log transformations are a special case of the ladder of powers. Finally, we cover the relatively advanced topic of the Box-Cox transformation.

Log Transformations

by David M. Lane

Prerequisites

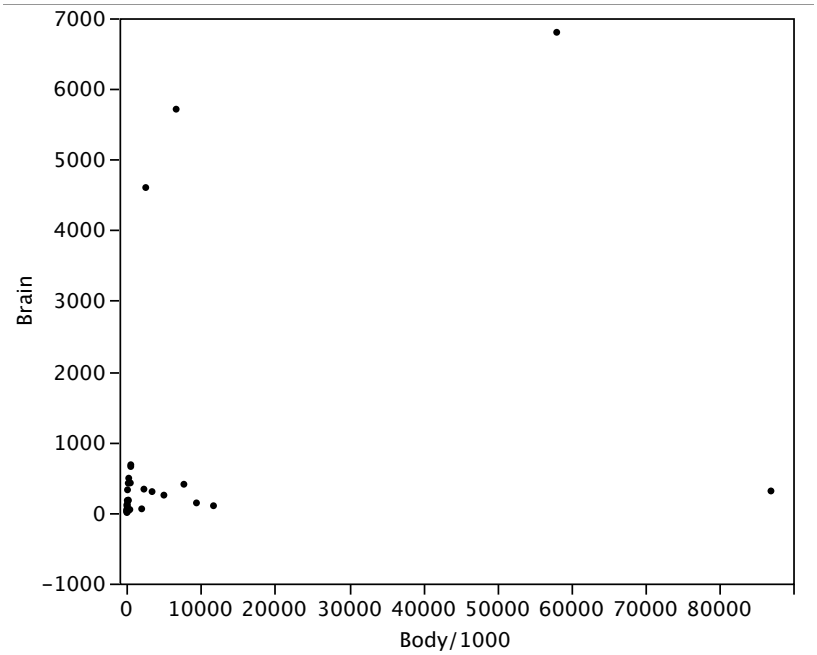
- Chapter 1: Logarithms
- Chapter 1: Shapes of Distributions
- Chapter 3: Additional Measures of Central Tendency
- Chapter 4: Introduction to Bivariate Data

Learning Objectives

1. State how a log transformation can help make a relationship clear
2. Describe the relationship between logs and the geometric mean

The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.

Figure 1 shows an example of how a log transformation can make patterns more visible. Both graphs plot the brain weight of animals as a function of their body weight. The raw weights are shown in the upper panel; the log-transformed weights are plotted in the lower panel.



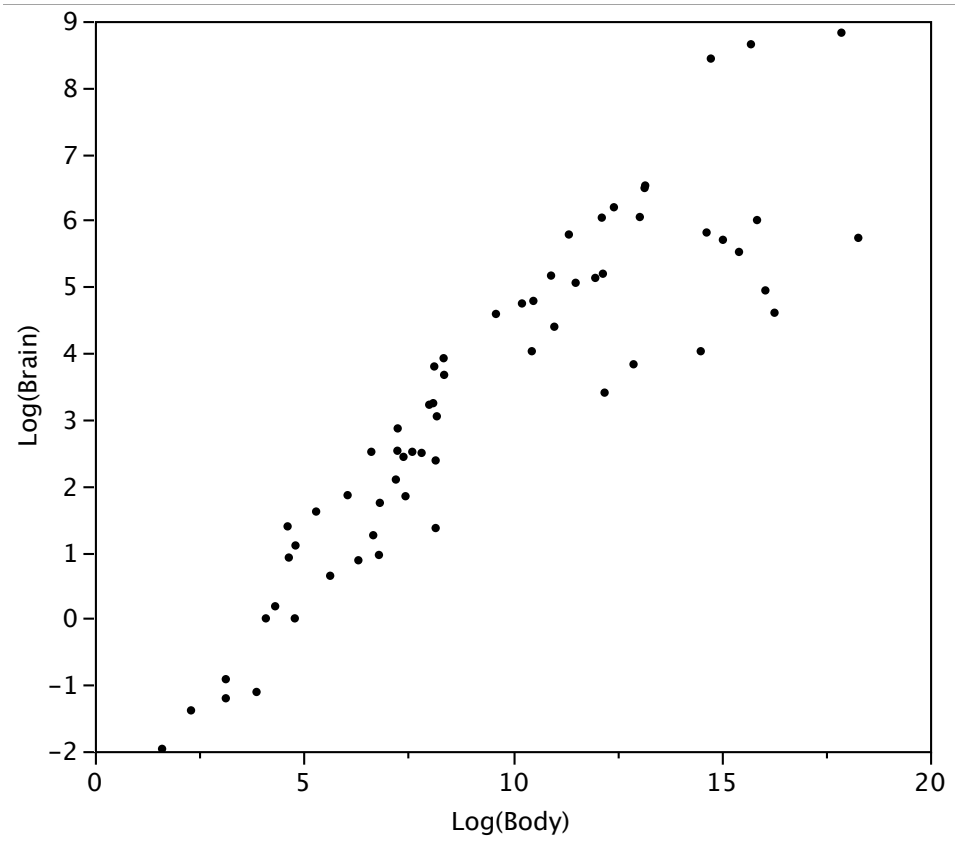


Figure 1. Scatter plots of brain weight as a function of body weight in terms of both raw data (upper panel) and log-transformed data (lower panel).

It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel.

The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.

Table 1 shows the logs (base 10) of the numbers 1, 10, and 100. The arithmetic mean of the three logs is

$$(0 + 1 + 2)/3 = 1$$

The anti-log of this arithmetic mean of 1 is:

$$10^1 = 10$$

which is the geometric mean:

$$(1 \times 10 \times 100)^{.3333} = 10.$$

Table 1. Logarithms.

X	$Log_{10}(X)$
1	0
10	1
100	2

Therefore, if the arithmetic means of two sets of log-transformed data are equal then the geometric means are equal.

Tukey Ladder of Powers

by David W. Scott

Prerequisites

- Chapter 1: Logarithms
- Chapter 4: Bivariate Data
- Chapter 4: Values of Pearson Correlation
- Chapter 12: Independent Groups t Test
- Chapter 13: Introduction to Power
- Chapter 16: Tukey Ladder of Powers

Learning Objectives

1. Give the Tukey ladder of transformations
2. Find a transformation that reveals a linear relationship
3. Find a transformation to approximate a normal distribution

Introduction

We assume we have a collection of bivariate data

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

and that we are interested in the relationship between variables x and y. Plotting the data on a scatter diagram is the first step. As an example, consider the population of the United States for the 200 years before the Civil War. Of course, the decennial census began in 1790. These data are plotted two ways in Figure 1. Malthus predicted that geometric growth of populations coupled with arithmetic growth of grain production would have catastrophic results. Indeed the US population followed an exponential curve during this period.

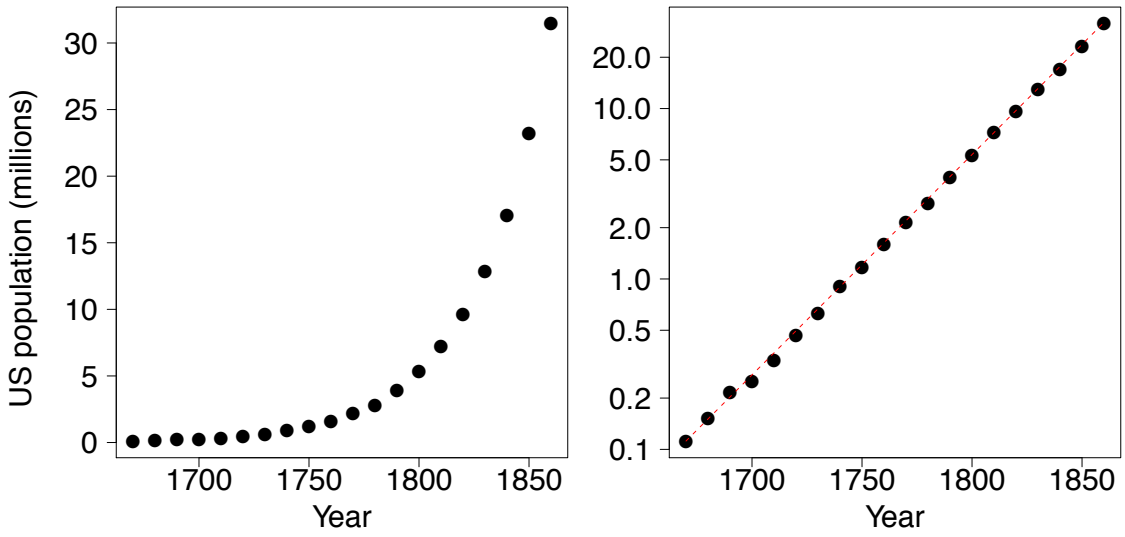


Figure 1. The US population from 1670 - 1860. The Y-axis on the right panel is on a log scale.

Tukey's Transformation Ladder

Tukey (1977) describes an orderly way of re-expressing variables using a power transformation. You may be familiar with polynomial regression (a form of multiple regression) in which the simple linear model $y = b_0 + b_1X$ is extended with terms such as $b_2x^2 + b_3x^3 + b_4x^4$. Alternatively, Tukey suggests exploring simple relationships such as

$$y = b_0 + b_1X^{\lambda} \text{ or } y^{\lambda} = b_0 + b_1X \quad \text{(Equation 1)}$$

where λ is a parameter chosen to make the relationship as close to a straight line as possible. Linear relationships are special, and if a transformation of the type x^{λ} or y^{λ} works as in Equation (1), then we should consider changing our measurement scale for the rest of the statistical analysis.

There is no constraint on values of λ that we may consider. Obviously choosing $\lambda = 1$ leaves the data unchanged. Negative values of λ are also reasonable. For example, the relationship

$$y = b_0 + b_1/x$$

would be represented by $\lambda = -1$. The value $\lambda = 0$ has no special value, since $X^0 = 1$, which is just a constant. Tukey (1977) suggests that it is convenient to simply define the transformation when $\lambda = 0$ to be the logarithm function rather than the

constant 1. We shall revisit this convention shortly. The following table gives examples of the Tukey ladder of transformations.

Table 1. Tukey's Ladder of Transformations

λ	-2	-1	-1/2	0	1/2	1	2
Xfm	$\frac{1}{x^2}$	$\frac{1}{x}$	$\frac{1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

If x takes on negative values, then special care must be taken so that the transformations make sense, if possible. We generally limit ourselves to variables where $x > 0$ to avoid these considerations. For some dependent variables such as the number of errors, it is convenient to add 1 to x before applying the transformation.

Also, if the transformation parameter λ is negative, then the transformed variable x^{λ} is reversed. For example, if x is increasing, then 1/x is decreasing. We choose to redefine the Tukey transformation to be $-(x^{\lambda})$ if $\lambda < 0$ in order to preserve the order of the variable after transformation. Formally, the Tukey transformation is defined as

$$\tilde{x}_{\lambda} = \begin{cases} x^{\lambda} & \text{if } \lambda > 0 \\ \log x & \text{if } \lambda = 0 \\ -(x^{\lambda}) & \text{if } \lambda < 0 \end{cases} \qquad (2)$$

In Table 2 we reproduce Table 1 but using the modified definition when $\lambda < 0$.

Table 2. Modified Tukey's Ladder of Transformations

λ	-2	-1	-1/2	0	1/2	1	2
Xfm	$\frac{-1}{x^2}$	$\frac{-1}{x}$	$\frac{-1}{\sqrt{x}}$	$\log x$	$\sqrt{x}$	x	x^2

The Best Transformation for Linearity

The goal is to find a value of λ that makes the scatter diagram as linear as possible. For the US population, the logarithmic transformation applied to y makes the relationship almost perfectly linear. The red dashed line in the right frame of Figure 1 has a slope of about 1.35; that is, the US population grew at a rate of about 35% per decade.

The logarithmic transformation corresponds to the choice λ = 0 by Tukey's convention. In Figure 2, we display the scatter diagram of the US population data for λ = 0 as well as for other choices of λ.

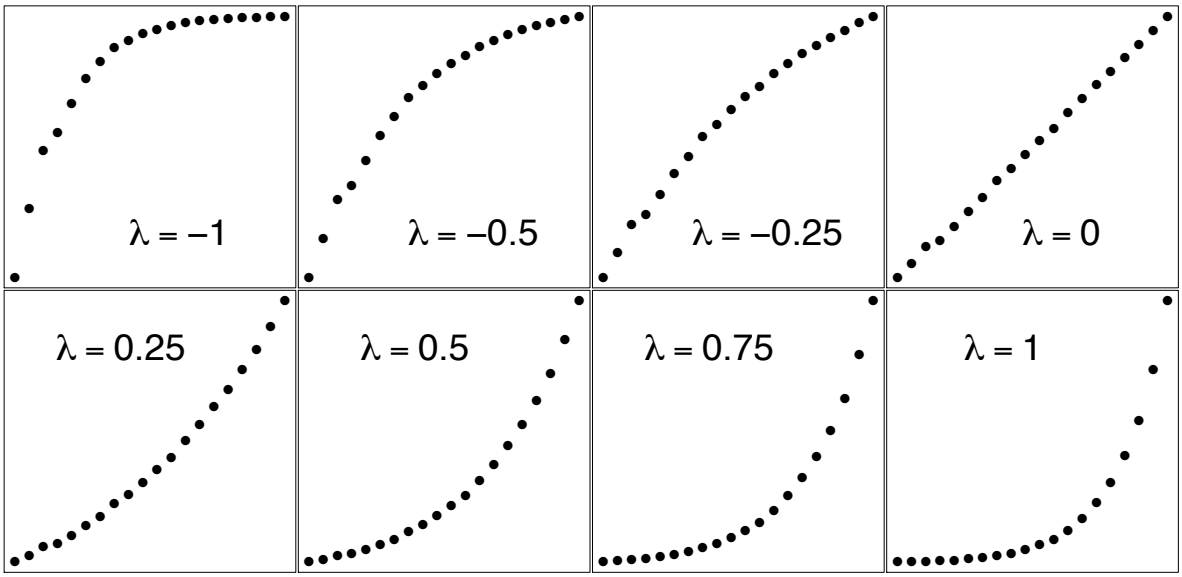


Figure 2. The US population from 1670 to 1860 for various values of λ.

The raw data are plotted in the bottom right frame of Figure 2 when λ = 1. The logarithmic fit is in the upper right frame when λ = 0. Notice how the scatter diagram smoothly morphs from convex to concave as λ increases. Thus intuitively there is a unique best choice of λ corresponding to the “most linear” graph.

One way to make this choice objective is to use an objective function for this purpose. One approach might be to fit a straight line to the transformed points and try to minimize the residuals. However, an easier approach is based on the fact that the correlation coefficient, r, is a measure of the linearity of a scatter diagram. In particular, if the points fall on a straight line then their correlation will be r = 1. (We need not worry about the case when r = −1 since we have defined the Tukey transformed variable x_λ to be positively correlated with x itself.)

In Figure 3, we plot the correlation coefficient of the scatter diagram $(x, \tilde{y}_\lambda)$ as a function of λ. It is clear that the logarithmic transformation (λ = 0) is nearly optimal by this criterion.

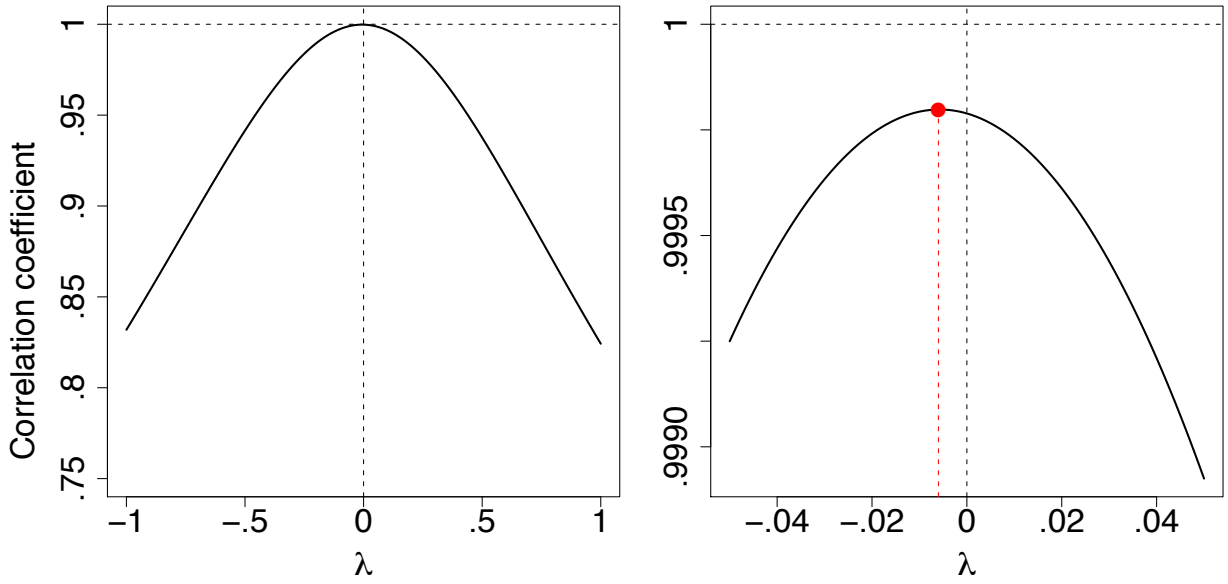


Figure 3. Graph of US population correlation coefficient as function of λ.

Is the US population still on the same exponential growth pattern? In Figure 4 we display the US population from 1630 to 2000 using the transformation and fit used in the right frame of Figure 1. Fortunately, the exponential growth (or at least its rate) was not sustained into the Twentieth Century. If it had, the US population in the year 2000 would have been over 2 billion (2.07 to be exact), larger than the population of China.

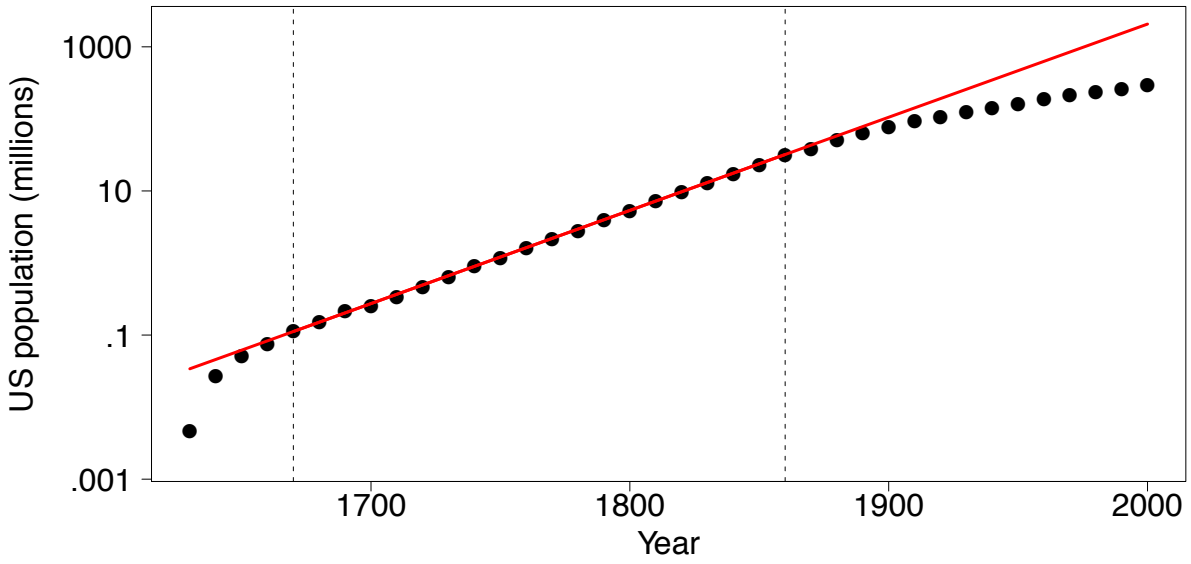


Figure 4. Graph of US population 1630-2000 with $\lambda = 0$.

We can examine the decennial census population figures of individual states as well. In Figure 5 we display the population data for the state of New York from 1790 to 2000, together with an estimate of the population in 2008. Clearly something unusual happened starting in 1970. (This began the period of mass migration to the West and South as the rust belt industries began to shut down.) Thus, we compute the best λ value using the data from 1790-1960 in the middle frame of Figure 5. The right frame displays the transformed data, together with the linear fit for the 1790-1960 period. The value of $\lambda = 0.41$ is not obvious and one might reasonably choose to use $\lambda = 0.50$ for practical reasons.

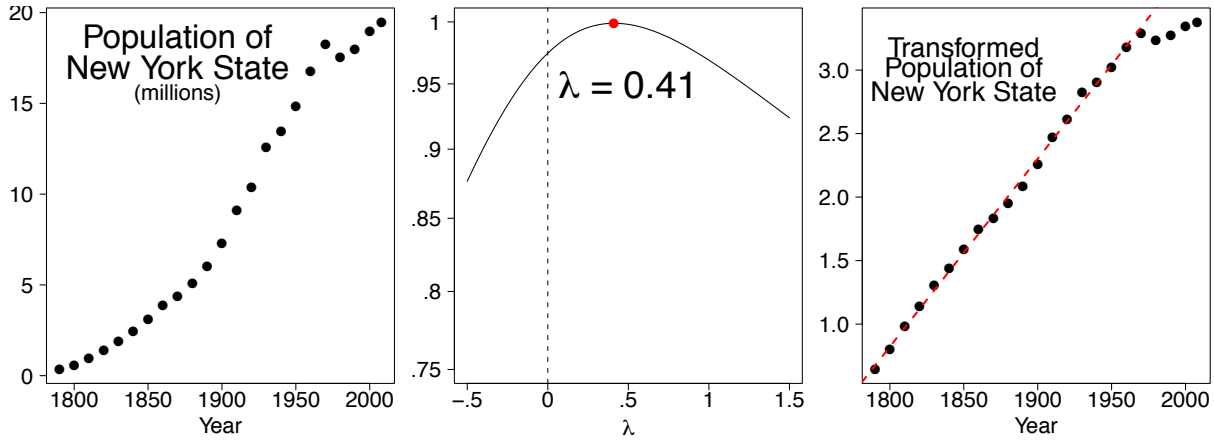


Figure 5. Graphs related to the New York state population 1790-2008.

If we look at one of the younger states in the West, the picture is different. Arizona has attracted many retirees and immigrants. Figure 6 summarizes our findings. Indeed, the growth of population in Arizona is logarithmic, and appears to still be logarithmic through 2005.

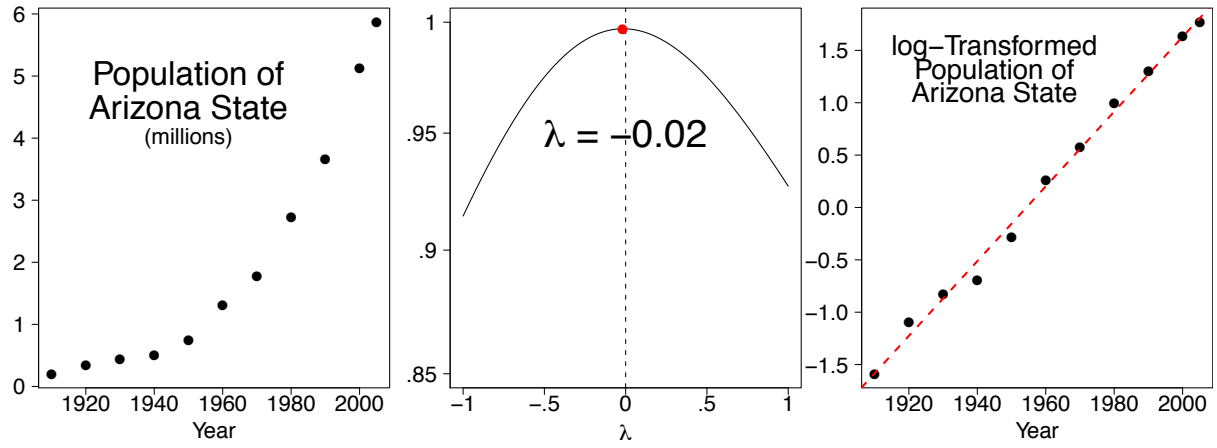


Figure 6. Graphs related to the Arizona state population 1910-2005.

Reducing Skew

Many statistical methods such as t tests and the analysis of variance assume normal distributions. Although these methods are relatively robust to violations of normality, transforming the distributions to reduce skew can markedly increase their power.

As an example, the data in the "Stereograms" case study is very skewed. A t test of the difference between the two conditions using the raw data results in a p value of 0.056, a value not conventionally considered significant. However, after a log transformation ($\lambda = 0$) that reduces the skew greatly, the p value is 0.023 which is conventionally considered significant.

The demonstration in Figure 7 shows distributions of the data from the Stereograms case study as transformed with various values of λ. Decreasing λ makes the distribution less positively skewed. Keep in mind that $\lambda = 1$ is the raw data. Notice that there is a slight positive skew for $\lambda = 0$ but much less skew than found in the raw data ($\lambda = 1$). Values of below 0 result in negative skew.

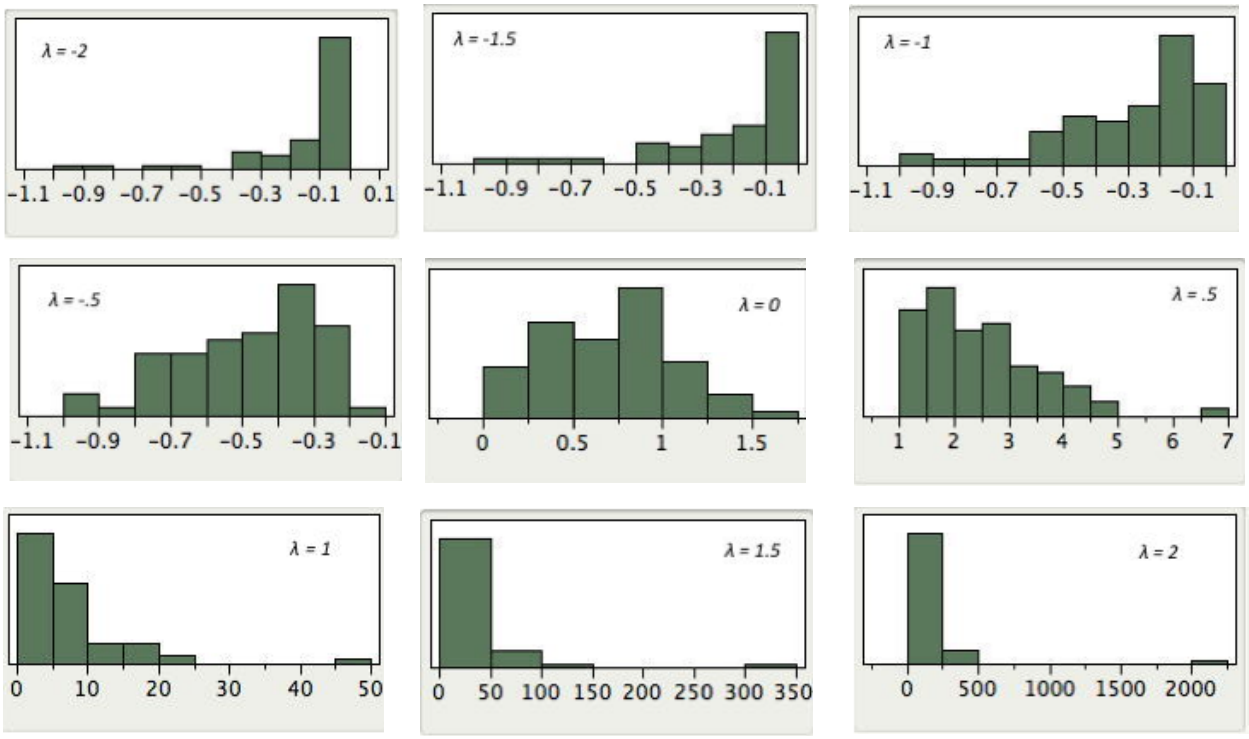


Figure 7. Distribution of data from the Stereogram case study for various values of λ.

Box-Cox Transformations

by David Scott

Prerequisites

This section assumes a higher level of mathematics background than most other sections of this work.

- Chapter 1: Logarithms
- Chapter 3: Additional Measures of Central Tendency (Geometic Mean)
- Chapter 4: Bivariate Data
- Chapter 4: Values of Pearson Correlation
- Chapter 16: Tukey Ladder of Powers

George Box and Sir David Cox collaborated on one paper (Box, 1964). The story is that while Cox was visiting Box at Wisconsin, they decided they should write a paper together because of the similarity of their names (and that both are British). In fact, Professor Box is married to the daughter of Sir Ronald Fisher.

The Box-Cox transformation of the variable x is also indexed by λ, and is defined as

$$x'_\lambda = \frac{x^\lambda - 1}{\lambda} \,. \qquad \texttt{(Equation 1)}$$

At first glance, although the formula in Equation (1) is a scaled version of the Tukey transformation x^λ, this transformation does not appear to be the same as the Tukey formula in Equation (2). However, a closer look shows that when $\lambda < 0$, both x_λ and x'_λ change the sign of x^λ to preserve the ordering. Of more interest is the fact that when $\lambda = 0$, then the Box-Cox variable is the indeterminate form 0/0. Rewriting the Box-Cox formula as

$$x'_\lambda = \frac{e^{\lambda \, \log(x)} - 1}{\lambda} \approx \frac{\left(1 + \lambda \, \log(x) + \frac{1}{2}\lambda^2 \log(x)^2 + \cdots\right) - 1}{\lambda} \rightarrow \log(x)$$

as $\lambda \rightarrow 0$. This same result may also be obtained using l'Hôpital's rule from your calculus course. This gives a rigorous explanation for Tukey's suggestion that the

log transformation (which is not an example of a polynomial transformation) may be inserted at the value $\lambda = 0$.

Notice with this definition of x'_λ that $x = 1$ always maps to the point $x'_\lambda = 0$ for all values of λ. To see how the transformation works, look at the examples in Figure 1. In the top row, the choice $\lambda = 1$ simply shifts x to the value $x-1$, which is a straight line. In the bottom row (on a semi-logarithmic scale), the choice $\lambda = 0$ corresponds to a logarithmic transformation, which is now a straight line. We superimpose a larger collection of transformations on a semi-logarithmic scale in Figure 2.

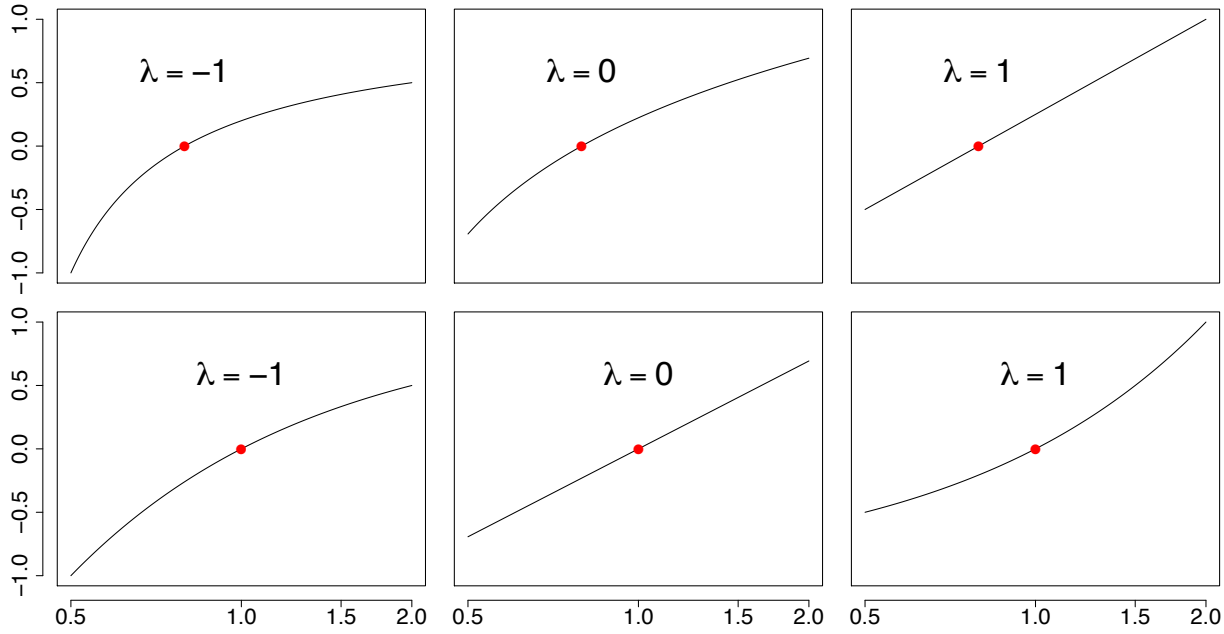


Figure 1. Examples of the Box-Cox transformation x'_λ versus x for $\lambda = -1, 0, 1$. In the second row, x'_λ is plotted against log(x). The red point is at $(1, 0)$.

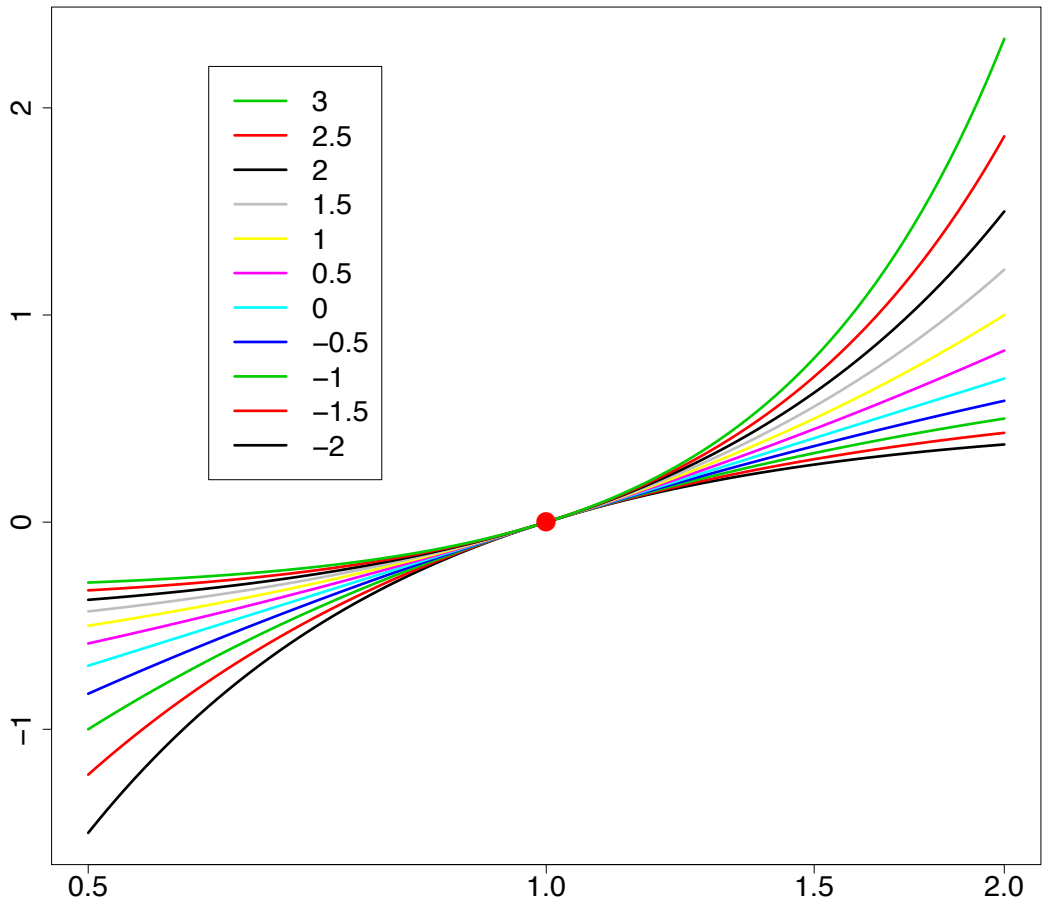


Figure 2. Examples of the Box-Cox transformation versus log(x) for $-2 < \lambda < 3$. The bottom curve corresponds to $\lambda = -2$ and the upper to $\lambda = 3$.

Transformation to Normality

Another important use of variable transformation is to eliminate skewness and other distributional features that complicate analysis. Often the goal is to find a simple transformation that leads to normality. In the article on q-q plots, we discuss how to assess the normality of a set of data,

$$x_1, x_2, \ldots, x_n.$$

Data that are normal lead to a straight line on the q-q plot. Since the correlation coefficients maximized when a scatter diagram is linear, we can use the same approach above to find the most normal transformation.
Specifically, we form the n pairs

$$\left(\Phi^{-1} \left(\frac{i - 0.5}{n} \right), x_{(i)} \right) , \qquad \text{for } i = 1, 2, \ldots, n ,$$

where Φ^{-1} is the inverse CDF of the normal density and $x_{(i)}$ denotes the i^{th} sorted value of the data set. As an example, consider a large sample of British household incomes taken in 1973, normalized to have mean equal to one (n = 7,125). Such data are often strongly skewed, as is clear from Figure 3. The data were sorted and paired with the 7125 normal quantiles. The value of λ that gave the greatest correlation (r = 0.9944) was λ = 0.21.

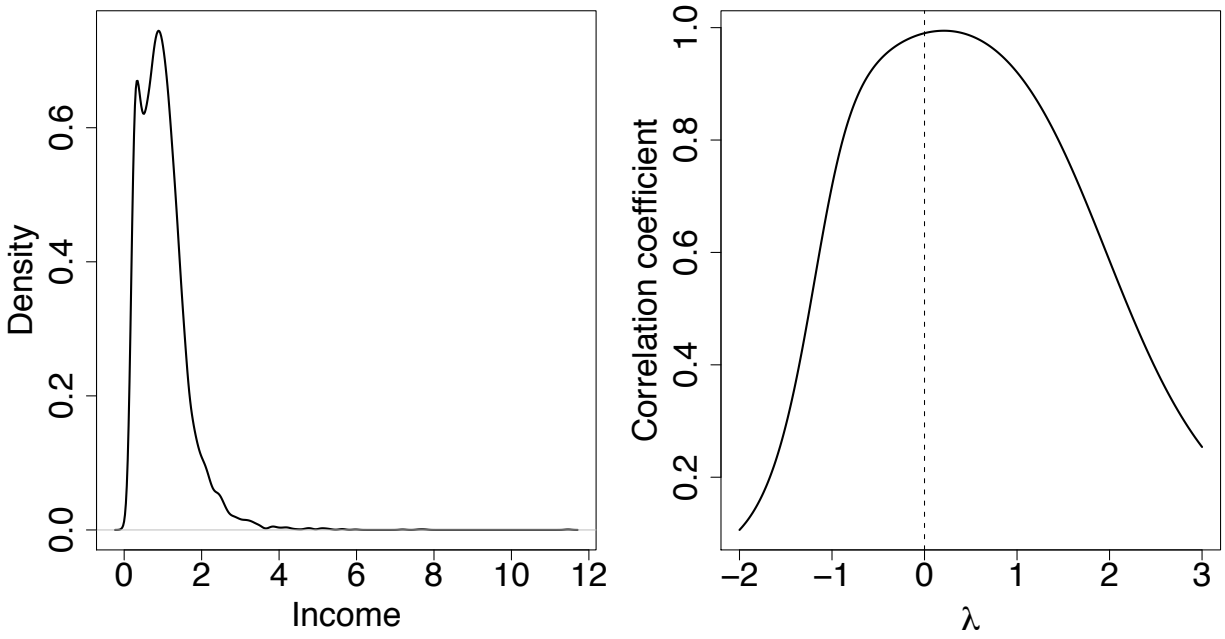


Figure 3. (L) Density plot of the 1973 British income data. (R) The best value of λ is 0.21.

The kernel density plot of the optimally transformed data is shown in the left frame of Figure 4. While this figure is much less skewed than in Figure 3, there is clearly an extra "component" in the distribution that might reflect the poor. Economists often analyze the logarithm of income corresponding to λ = 0; see Figure 4. The correlation is only r = 0.9901 in this case, but for convenience, the log-transform probably will be preferred.

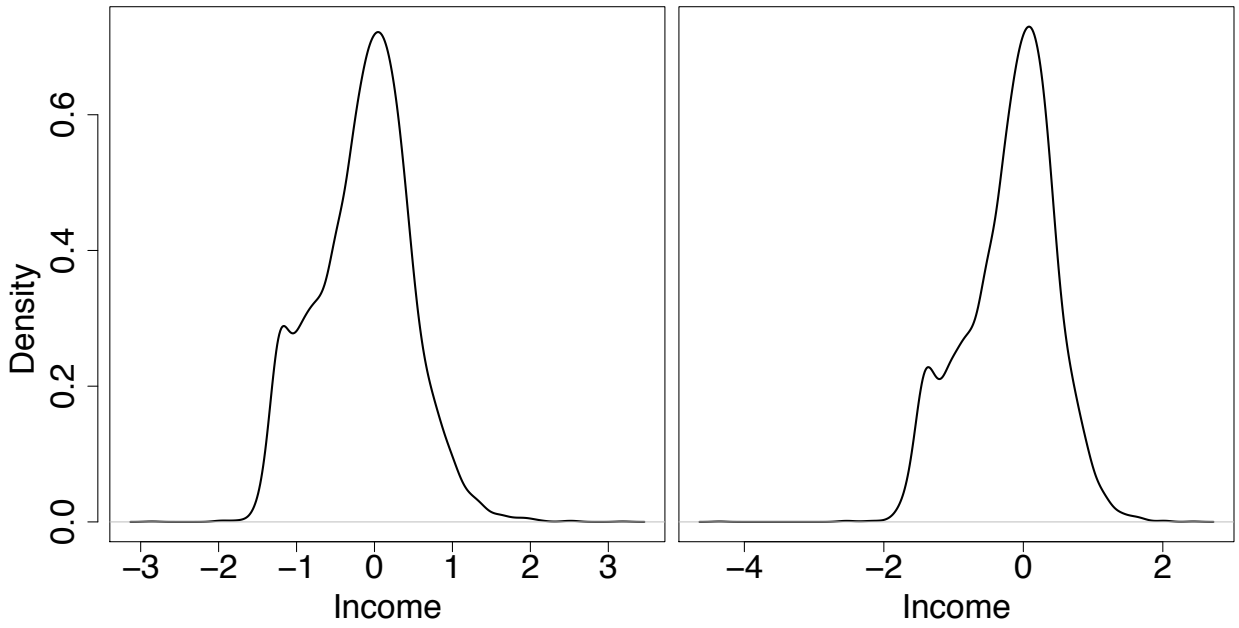


Figure 4. (L) Density plot of the 1973 British income data transformed with $\lambda = 0.21$. (R) The log-transform with $\lambda = 0$.

Other Applications

Regression analysis is another application where variable transformation is frequently applied. For the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

and fitted model

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \, ,$$

each of the predictor variables x_j can be transformed. The usual criterion is the variance of the residuals, given by

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \, .$$

Occasionally, the response variable y may be transformed. In this case, care must be taken because the variance of the residuals is not comparable as λ varies. Let $\bar{g}_y$ represent the geometric mean of the response variables.

$$\bar{g}_y = \left(\prod_{i=1}^{n} y_i \right)^{1/n}$$

Then the transformed response is defined as

$$y'_\lambda = \frac{y^\lambda - 1}{\lambda \cdot \bar{g}_y^{\lambda - 1}}$$

When λ = 0 (the logarithmic case),

$$y'_0 = \bar{g}_y \cdot \log(y)$$

For more examples and discussions, see Kutner, Nachtsheim, Neter, and Li (2004).

Statistical Literacy

by David M. Lane

Prerequisites

- Chapter 16: Logarithms

Many financial web pages give you the option of using a linear or a logarithmic Y-axis. An example from Google Finance is shown below.

What do you think?

To get a straight line with the linear option chosen, the price would have to go up the same amount every time period. What would result in a straight line with the logarithmic option chosen?

> The price would have to go up the same proportion every time period. For example, go up 0.1% every day.

References

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society*, Series B, 26, 211-252.

Kutner, M., Nachtsheim, C., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models*, McGraw-Hill/Irwin, Homewood, IL.

Tukey, J. W. (1977) Exploratory Data Analysis. Addison-Wesley, Reading, MA.

Exercises

Prerequisites

All Content in This Chapter

1. When is a log transformation valuable?
2. If the arithmetic mean of $\log_{10}$ transformed data were 3, what would be the geometric mean?
3. Using Tukey's ladder of transformation, transform the following data using a λ of 0.5: 9, 16, 25
4. What value of λ in Tukey's ladder decreases skew the most?
5. What value of λ in Tukey's ladder increases skew the most?
6. In the ADHD case study, transform the data in the placebo condition (D0) with λ's of .5, 0, -.5, and -1. How does the skew in each of these compare to the skew in the raw data. Which transformation leads to the least skew?