Log transforms From "Research methods: background and review" by Kerby Shedden

Prepared by Edward Ionides

Department of Statistics, University of Michigan

2018-03-01

License and copyright for the complete document

Log transforms (E.g., Homework 6 in STATS 401 W18)

Some quantities that vary over several orders of magnitude are best analyzed on the log scale.

For example, if we observe these values:

$$14, 28, 3, 60, 39, 13, 1, 9, 3, 55$$

We can take $\log_2$ to get their approximate values as powers of 2:

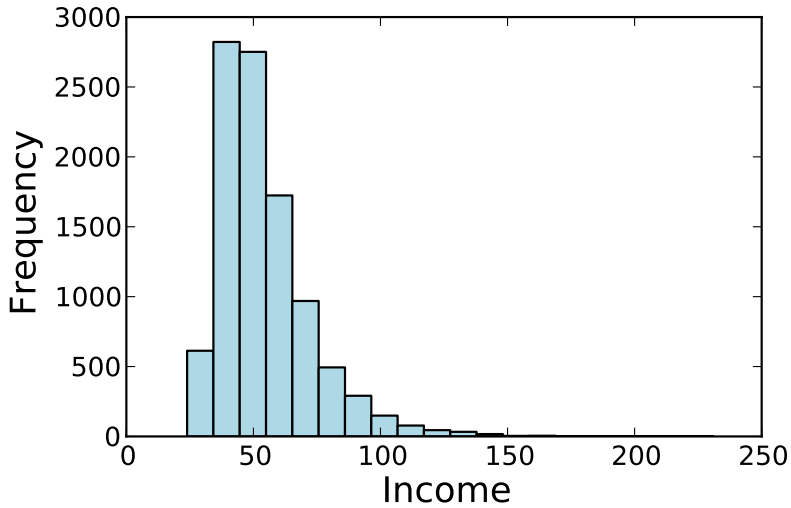
$$3.8, 4.8, 1.6, 5.9, 5.3, 3.7, 0, 3.2, 1.6, 5.8.$$

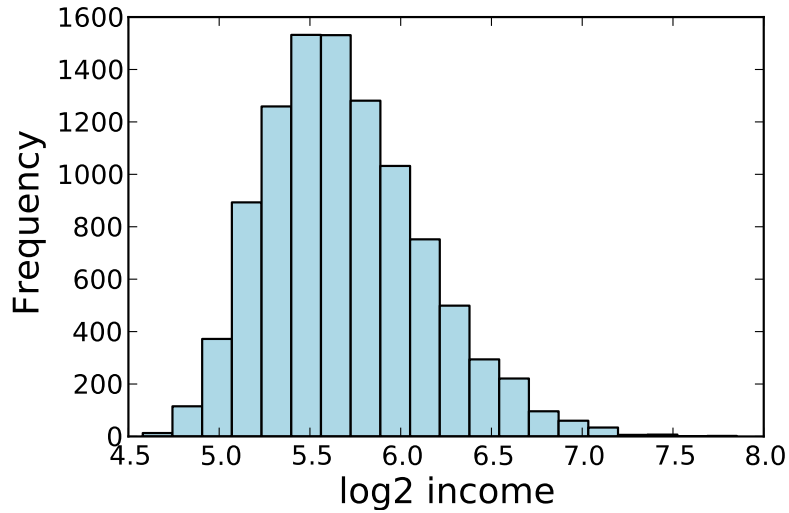
It usually doesn't matter what base is used, since we can convert from one base to another by scaling:

$$\log_b(x) = \log_a(x)/\log_a(b)$$

Symmetrizing effect of log transforms

The log transform symmetrizes right-skewed distributions:





It's common to transform data to make it more symmetric, and usually that's the right thing to do (but don't overdo it...).

Properties of log transforms

Remember the key properties of logarithms:

$$\log(ab) = \log(a) + \log(b) \qquad \log(a^b) = b\log(a).$$

As a consequence, if we take data $X_1, \ldots, X_n$ and scale it to get $Z_i = cX_i$, then

$$\log(Z_1), \ldots, \log(Z_n) = \log(c) + \log(X_1), \ldots, \log(c) + \log(X_n)$$

Thus changing the units of the original data becomes a shift by $\log(c)$ units for the log-transformed data.

Mean values and log transforms

If we observe data $X_1, \ldots, X_n$ and take a log transform to get $Y_i = \log X_i$, then the mean value of the logged data is:

$$\begin{aligned}
\bar{Y} &= n^{-1} \sum_i Y_i \\
&= n^{-1} \sum_i \log X_i \\
&= n^{-1} \log(X_1 \cdot X_2 \cdots X_n) \\
&= \log\left((X_1 \cdot X_2 \cdots X_n)^{1/n}\right).
\end{aligned}$$

$(X_1 \cdot X_2 \cdots X_n)^{1/n}$ is called the **geometric mean** of the X_i, so we see that the usual (arithmetic) mean of the log transformed data is the log of the geometric mean of the untransformed data.

Log transforms

We generally take the log of positive data that is substantially right skewed. If the data are roughly symmetrically distributed, there is no need to take a log transform, and you cannot take a log transform if any of the data values are less than or equal to zero.

Examples: We generally would log-transform income but not age.