Reasons to transform data

- to more closely approximate a theoretical distribution that has nice statistical properties
- to spread data out more evenly
- to make data distributions more symmetrical
- to make relationships between variables more linear
- to make data more constant in variance (*homoscedastic*)

Ladder of powers

A useful organizing concept for data transformations is the *ladder of powers* (P.F. Velleman and D.C. Hoaglin, *Applications, Basics, and Computing of Exploratory Data Analysis*, 354 pp., Duxbury Press, 1981). Data transformations are commonly power transformations, $x'=x^{\theta}$ (where x' is the transformed x). One can visualize these as a continuous series of transformations:

θ		transformation
3	x^3	cube
2	x^2	square
1	x^1	identity (no transformation)
1/2	$x^{0.5}$	square root
1/3	$x^{1/3}$	cube root
0	log(x)	logarithmic (holds the place of zero)
-1/2	$-1/x^{0.5}$	reciprocal root
-1	$-1/x$	reciprocal
-2	$-1/x^2$	reciprocal square

Note:
- higher and lower powers can be used
- fractional powers (other than those shown) can be used
- minus sign in reciprocal transformations can (optionally) be used to preserve the order (relative ranking) of the data, which would otherwise be inverted by transformations for $\theta<0$.

To use the ladder of powers, visualize the original, untransformed data as starting at $\theta=1$. Then if the data are *right-skewed* (clustered at lower values) move *down* the ladder of powers (that is, try square root, cube root, logarithmic, etc. transformations). If the data are *left-skewed* (clustered at higher values) move *up* the ladder of powers (cube, square, etc).

Special transformations

$x'=\log(x+1)$
- often used for transforming data that are right-skewed, but also include zero values.
- note that the shape of the resulting distribution will depend on how big x is compared to the constant 1. Therefore the shape of the resulting distribution depends on the units in which x was measured. One way to deal with this problem is to use $x'=\log(x/\text{mean}(x)+k)$, where k is a small constant ($k<<1$). In this transformation, the mean x will be transformed to near $x'=0$ and k will function as a shape factor (small k will make x' more left-skewed, larger k will make it less so). But most importantly, changing the units of measure will not change the shape of the distribution.

$x' = \sqrt{x + 0.5}$ -sometimes used where data are taken from a Poisson distribution (for example, counts of random events that occur in a fixed time period), or used for right-skewed data that include some x values that are very small or zero. As above, the resulting distribution of x' depends on the units used to measure x.

$x' = arcsin\sqrt{x}$ -used for data that are proportions (for example, fraction of eggs in a clutch that fail to hatch); converts the binomial distribution that often characterizes such data into an approximate normal distribution.

Important note

-in general, parameters (means, standard deviations, regression slopes, etc.) that are calculated on the transformed data and then are transformed back to the original units, will <u>*not*</u> equal the same parameters calculated on the original, untransformed data.

Symmetry plots (a precise visual tool for displaying departures from symmetry)

How to: -sort the data set x_i, i=1..n into ascending order, and find the median

-for each pair of points surrounding the median (which will be the the points x_i and $x_{(n+1-i)}$, plot:

-on the horizontal axis, the distance x_{median}-x_i
-on the vertical axis, the distance $x_{(n+1-i)}$-x_{median}

-if the points lie consistently above the 1:1 line, then the data are right-skewed.
-if the points lie consistently below the 1:1 line, then the data are left-skewed.
-if the points lie close to the 1:1 line, then $x_{median}-x_i \approx x_{(n+1-i)}-x_{median}$ and the distribution is approximately symmetrical.

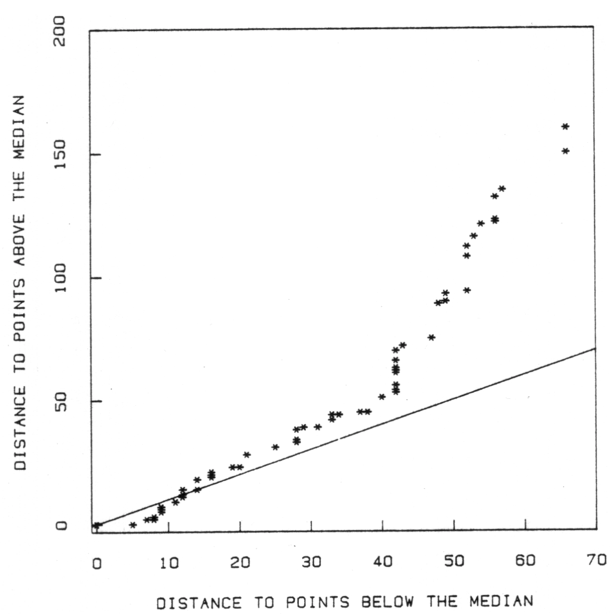

Figure 2.15 A symmetry plot of the ozone data.

Reference: Chambers, J. M., W. S. Cleveland, B. Kleiner and P. A. Tukey, *Graphical Methods for Data Analysis*, 395 pp., Wadsworth & Brooks/Cole Publishing Co., 1983.