Data Transforms: Natural Logarithms and Square Roots

Parametric statistics in general are more powerful than non-parametric statistics as the former are based on ratio level data (real values) whereas the latter are based on ranked or ordinal level data. Of course, non-parametrics are extremely useful as sometimes our data is highly non-normal, meaning that comparing the means is often highly misleading, and can lead to erroneous results. Non-parametrics statistics allow us to make observations on statistical patterning even though data may be highly skewed one way or another. However, by doing so, we loose a certain degree of power by converting the data values into relative ranks, rather than focus on the actual differences between the values in the raw data. The take home point here is that we always use parametric statistics where possible, and we resort to non-parametrics if we are sure parametrics will be misleading.

Parametric statistics work on ratio level data, that is data that has a true zero value (where zero means absence of value) and the intervals between data are consistent, independent of the data point value. The obvious case in point are the Roman numeral real values we are used to counting everyday {…, -4, -3, -2, -1, 0, 1, 2, 3, 4,…}. However, these are not the only values that constitute ratio level data. Alternatives are logged data, or square rooted data, where the intervals between the data points are consistent, and a true zero value exists.

The possibility of transforming data to an alternative ratio scale is particularly useful with skewed data, as in some cases the transformation will *normalize* the data distribution. If the transform normalizes the data, we can go ahead and continue to use parametric statistics in exactly the same way, and the results we get (*p* values etc.) are equally as valid as before.

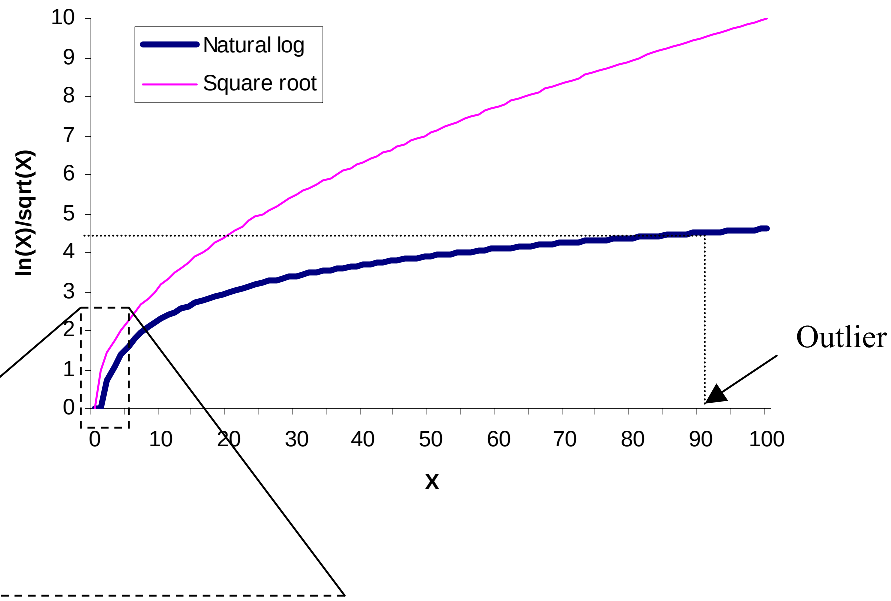
The way this works is that both the natural logarithm and the square root are mathematical functions meaning that they produce curves that affect the data we want to transform in a particular way. The shapes of these curves normalize data (if they work) by passing the data through these functions, altering the shape of their distributions. For example look at the figures below.

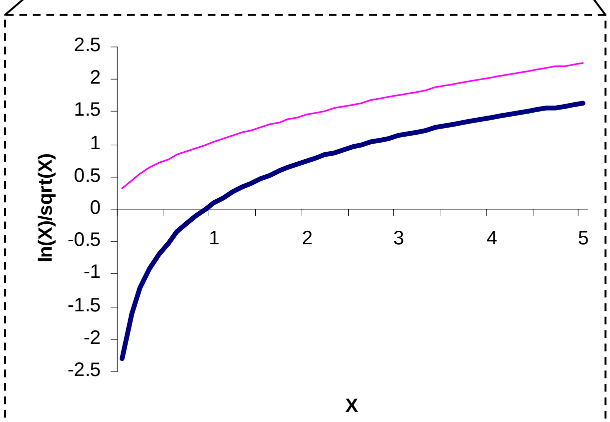
Mathematically, taking the natural logarithm of a number is written in a couple of ways:

$$X = \ln x \text{, or}$$
$$X = \log_e x$$

And taking the square root is written:

$$X = \sqrt{x}$$





Looking at the top figure we can see that the presence of any outliers on the X axis will be reduced on the Y axis due to the shape of the curves. This effect will be most effective with the log function as opposed to the square root function (√). We can extrapolate out by seeing that given the curve of the log function the more extreme the outlier, the greater the affect of log transforming.

Looking at the inset figure we can see that logging values that are less than 1 on the X axis will result in negative log values; even though this may seem to be a problem intuitively, it is not. This is because $\ln(1)=0$, therefore $\ln(<1)<0$. In fact $\ln(0)$ is undefined meaning that the log function approaches the Y axis asymptotically but never gets there. A usual method of dealing with raw data where many of the values are less than 1 is to add an arbitrary constant to the entire data set and then log transform; in this way we avoid dealing with negative numbers.

What does all this mean? Well, transforming data sets works most effectively for data distributions that are skewed to the right by the presence of outliers. However, transforming the data does not always work as it depends ultimately on the specific values involved. In general, it is best to attempt log transforming first, if that doesn't work try square root transforming, and if that doesn't work, go with a non-parametric test.

MINITAB EXAMPLE

It is very easy to transform data either in EXCEL or MINITAB (I usually use EXCEL). In EXCEL the code is simply =ln(X), where X is your data, and you can click and drag the formula down a whole column of data. In MINITAB you can use the CALCULATOR function under CALC on the toolbar and store the transformed variables in a new column.

An example comes from Binford (2001) using data on hunter-gatherer group sizes (*N*=227); I won't bother to list all 227 data points…

Reading the data into MINITAB, to look at the normality of the data we need to run the descriptive stats, do a normality test and look at the distribution. For the descriptive stats, in MINITAB procedure is:

>STAT

>BASIC STATISTICS

>DESCRIPTIVE STATISTICS

>Double click on the column your data is entered

>GRAPHS: choose BOXPLOT and GRAPHICAL SUMMARY,

>OK

>OK

The output reads:

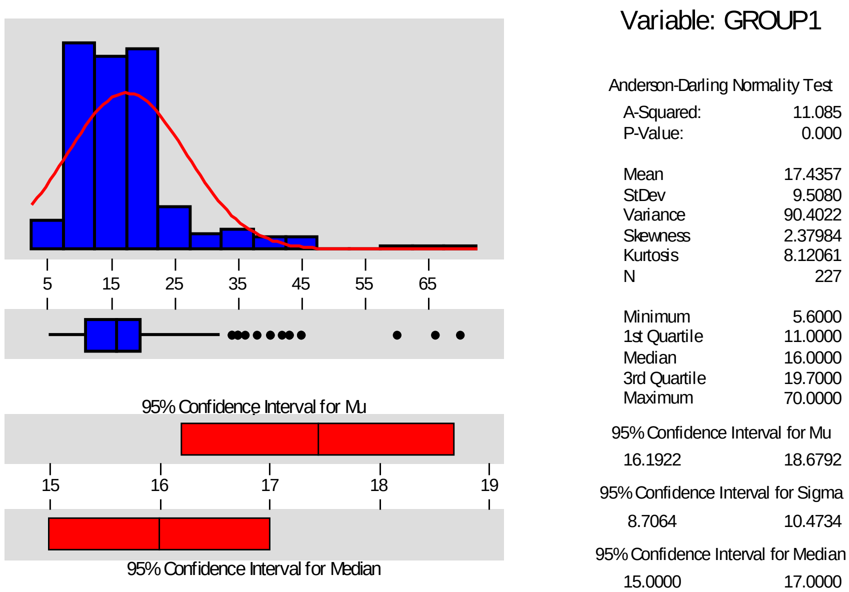
Descriptive Statistics

Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
GROUP1	227	17.436	16.000	16.358	9.508	0.631

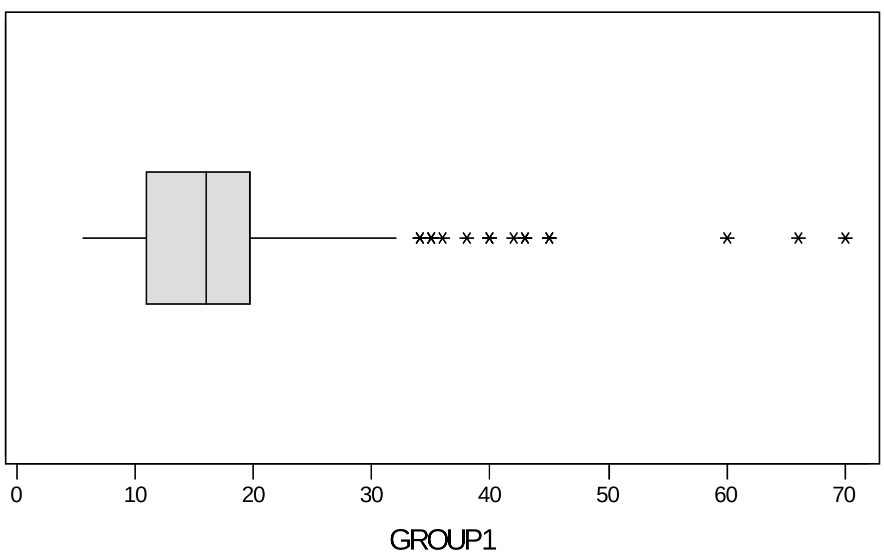
Variable	Min	Max	Q1	Q3
GROUP1	5.600	70.000	11.000	19.700

With the two graphics:

Descriptive Statistics



Boxplot of GROUP1



From the descriptive stats output we can see the mean and median are different, especially considering the standard error. We also see from the graphical output, the boxplot shows a bunch of outliers, and a heavily skewed distribution. The Anderson-Darling result on the graphical summary gives p=0.000, meaning that the data is very non-normal. Given the skewness of the data and the presence of outliers, log transforming is at least worth trying.

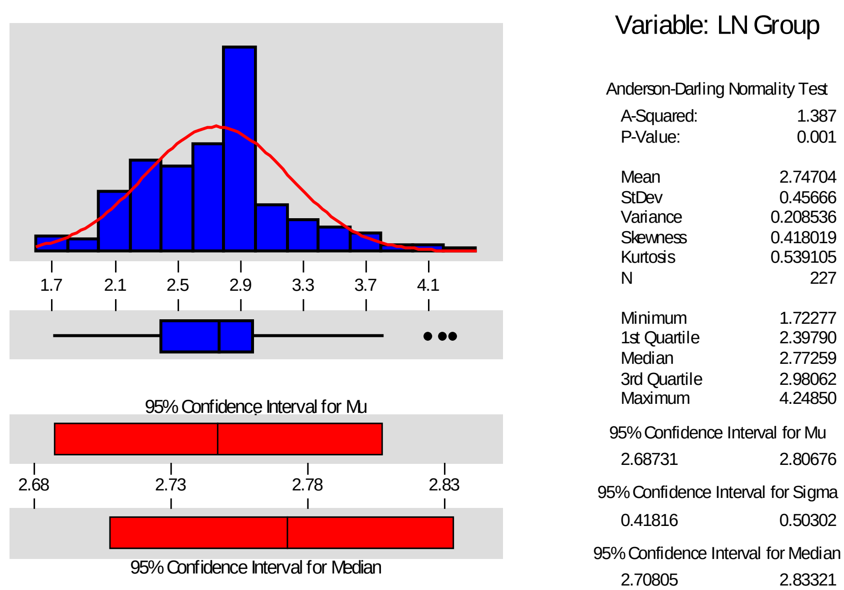
So, logging the data in EXCEL and transferring it into MINITAB we run the same set of procedures, leading to the following outputs:

Descriptive Statistics

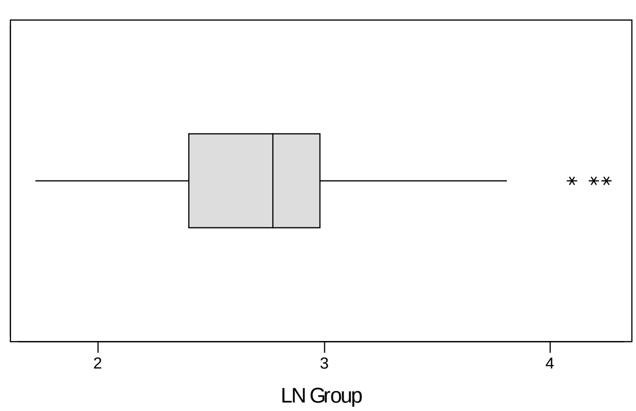
Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
LN Group	227	2.7470	2.7726	2.7339	0.4567	0.0303

Variable	Min	Max	Q1	Q3
LN Group	1.7228	4.2485	2.3979	2.9806

Descriptive Statistics



Boxplot of LN Group



Well, while it was a good idea to try a log transform, and we see from the descriptive statistics that the mean and median a very close, the Anderson-Darling result still tells us that the data is non-normal. We see from the boxplot that we still have a few stubborn outliers. We have made the data kind of symmetrical, but unfortunately it is still non-normal: we have to go ahead and use non-parametric statistics from here if we want to use this data statistically.

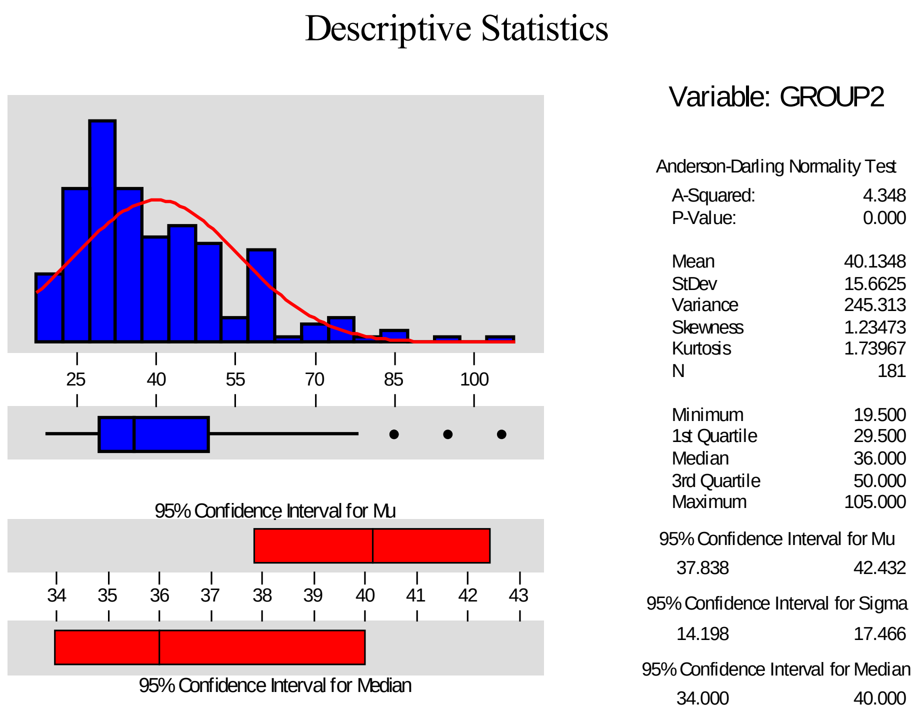
Let's try a second example. We'll take some more data from Binford (2001), this time referring to the mean annual aggregation size of terrestrial hunter-gatherers (N=181). Following the same procedures as above we find the following: For the raw data

Descriptive Statistics

Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
GROUP2	181	40.13	36.00	38.86	15.66	1.16

Variable	Min	Max	Q1	Q3
GROUP2	19.50	105.00	29.50	50.00

And,



We see that the median and means are not equal, and the Anderson-Darling stat is non-significant, so logging the data and putting it into MINITAB we get:

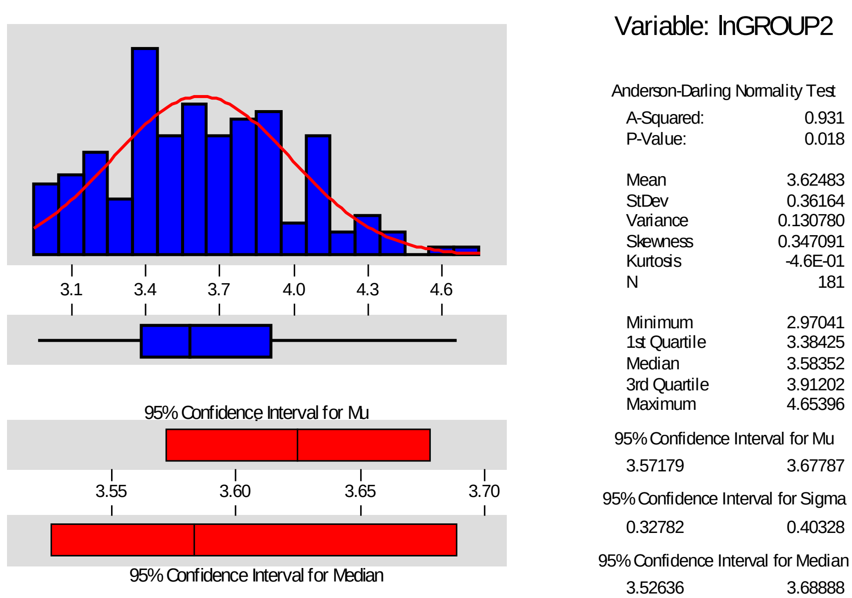
Descriptive Statistics

Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
lnGROUP2	181	3.6248	3.5835	3.6147	0.3616	0.0269

Variable	Min	Max	Q1	Q3
lnGROUP2	2.9704	4.6540	3.3842	3.9120

And,

Descriptive Statistics



In this case we see that the mean and median are now very similar, and the boxplot shows the presence of no outliers. The Anderson-Darling test shows a significance level of roughly 0.02 (98%), and while this is less than the usual α level of 0.05 (95%), this result is pretty strong. And here we come up against the subjectivity of statistics; it is up to the observer to decide whether this data is normal enough for parametric statistics. Most would argue that it is, given that, in reality, the Anderson-Darling test is very conservative in that it will detect the slightest deviation from normality, and that parametric statistics are remarkably robust, only being dramatically effected by highly non-normal data. I would accept the log-transformed data as close enough to normal to use parametric statistics.