Transformations

Summary

Many statistical methods require the data to fit assumptions of Normal distribution and uniform variance. When data do not fit these, one approach is to make them do so by a mathematical transformation. The most frequently used are the square root, the logarithm, and the reciprocal. These all reduce positive skewness and the extent to which variation increases with magnitude, the square root having the least effect and the reciprocal the greatest for each. The logarithm is the most often used. For a single, simple sample all will give us interpretable confidence intervals after transformations back to the original scale. For comparisons of means, only the logarithm can do this. Concentrations in blood are often analysed on the logarithmic scale, counts on the square root scale. There are other transformations in use, but they are seen rarely in the health research literature. Some data cannot be transformed satisfactorily and some, such as cost data, should not be. If data cannot be transformed, there are other strategies available, which do not require assumptions of Normal distribution or uniform variance.

1. The need for transformations

In Week 4 I described statistical methods in which we have to assume that data follow a Normal distribution with uniform variance. Later we shall meet other methods, regression and correlation, which require similar assumptions to be made about the data. Most analyses of continuous data in the health research literature are of this type. We should always check these assumptions. If the data meet the assumptions we can analyse the data as described. If they are not met, we have two possible strategies: we can use a method which does not require these assumptions, such as a rank-based method, or we can transform the data mathematically to make them fit the assumptions more closely. In this lecturer I describe the second approach. Instead of analysing the data as observed, we carry out a mathematical transformation first.

For example, Figure 1 shows serum cholesterol in stroke patients. As we have noted before, this does not follow a Normal distribution closely. This is shown by both the shape of the histogram and the Normal plot, in which the points should be close to the straight line. Figure 2 shows the same plots for the logarithms of cholesterol measurements. (See separate document *Logarithms*.) The logarithm follows a Normal distribution more closely than do the cholesterol measurements themselves. We could analyse the logarithm of serum cholesterol using methods which required the data to follow a Normal distribution. We call the logarithm of the cholesterol a **logarithmic transformation** of the data, or **log transformation** for short. We call the data without any transformation the **raw data**.

Even if a transformation does not produce a really good fit to the Normal distribution, it may still make the data much more amenable to analysis. Figure 3 shows a histogram and Normal plot for the area of venous ulcer at recruitment to the VenUS I trial, with the same for the log transformed area. The raw data have a very skew distribution and the small number of very large ulcers might lead to problems in analysis. Although the log transformed data are still skew, the skewness is much less and the data much easier to analyse.

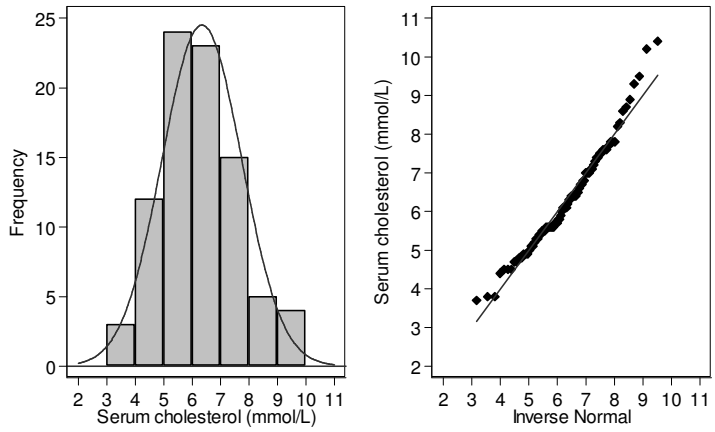

Figure 1. Histogram with Normal distribution curve and Normal plot for serum cholesterol in 86 stroke patients (data of Markus *et al.*, 1995)

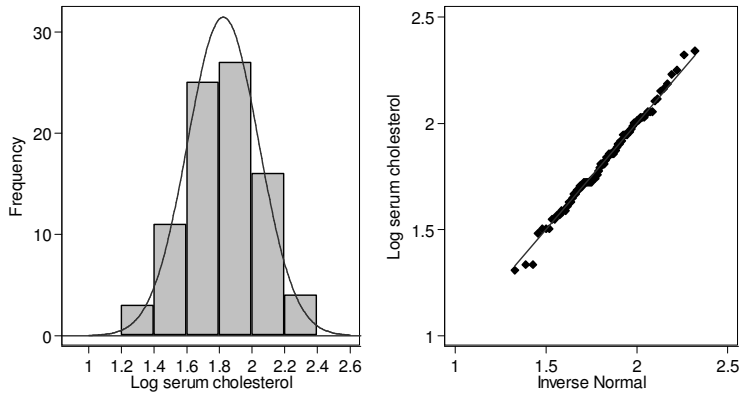

Figure 2. Histogram with Normal distribution curve and Normal plot for log transformed serum cholesterol in 86 stroke patients

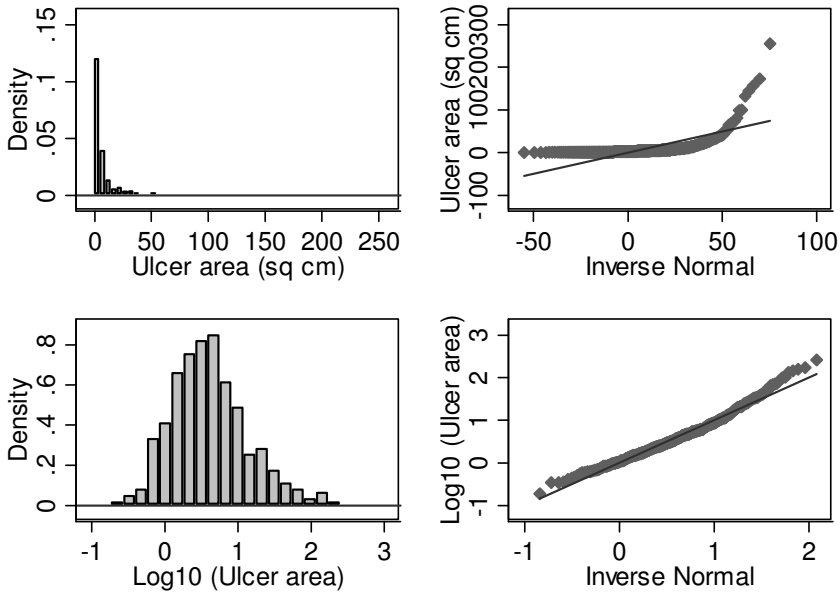

Figure 3. Histogram and Normal plot for area of venous ulcer at recruitment and log transformed ulcer area, VenUS I trial

Making a distribution more like the Normal is not the only reason for using a transformation. Figure 4 shows prostate specific antigen (PSA) for three groups of prostate patients: with benign conditions, with prostatitis, and with prostate cancer. One very high value makes it very difficult to see the structure in the rest of the data, although, as we would expect, we can see that the cancer group have the highest PSA values. A log transformation of the PSA gives a much clearer picture, shown in Figure 5. The variability is now much more similar in the three groups. Figure 6 shows a histogram and Normal plot for the raw data and the log transformed data. This shows the distribution of the within-group residuals, the difference between observed values and the mean of the group. The log transformation not only makes the variability more uniform but also makes the distribution closer to the Normal, thus meeting both assumptions: Normal distribution and uniform variance.

If we want to use the transformation only to make the scatter diagram easier to see, we sometimes use a logarithmic scale rather than the actual logarithms of the data (see *Logarithms*). Figure 7 shows this for the PSA data. The picture looks like the scatter plot for the log transformed PSA in Figure 5, but the vertical scale shows the original units.

The logarithm is not the only transformation used in the analysis of continuous data. Figure 8 shows arm lymphatic flow in patients with and without rheumatoid arthritis and oedema. The distribution is positively skew and the variability is clearly greater in the groups with greater lymphatic activity. A square root transformation has the effect of making the data less skew and making the variation more uniform. In these data, a log transformation proved to have too great an effect, making the distribution negatively skew, and so the square root of the data was used in the analysis (Kiely *et al.*, 1995).

2. Commonly used transformations for quantitative data

There are three commonly used transformations for quantitative data: the logarithm, the square root, and the reciprocal. (The **reciprocal** of a number is one divided by that number, hence the reciprocal of 2 is ½.) There are good mathematical reasons for these choices, Bland (2000) discusses them. They are based on the need to make variances uniform. If we have several groups of subjects and calculate the mean and variance for each group, we can plot variability against mean. We might have one of these situations:

- Variability and mean are unrelated. We do not usually have a problem and can treat the variances as uniform. We do not need a transformation.
- Variance is proportional to mean. A square root transformation should remove the relationship between variability and mean.
- Standard deviation is proportional to mean. A logarithmic transformation should remove the relationship between variability and mean.
- Standard deviation is proportional to the square of the mean. A reciprocal transformation should remove the relationship between variability and mean.

We call these transformations **variance-stabilising**, because their purpose is to make variances the same. For most data encountered in healthcare research, the first or third situation applies.

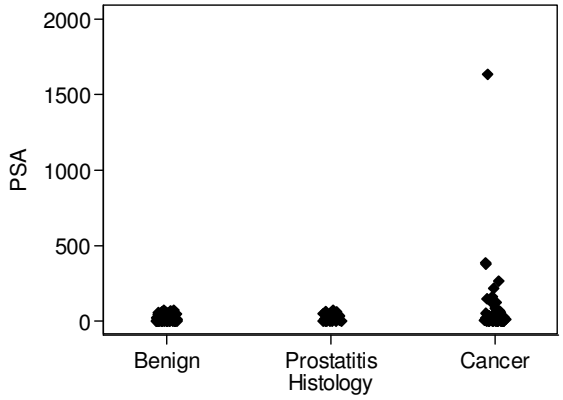

Figure 4. Prostate specific antigen (PSA) by prostate diagnosis (data of Cutting *et al.*, 1999)

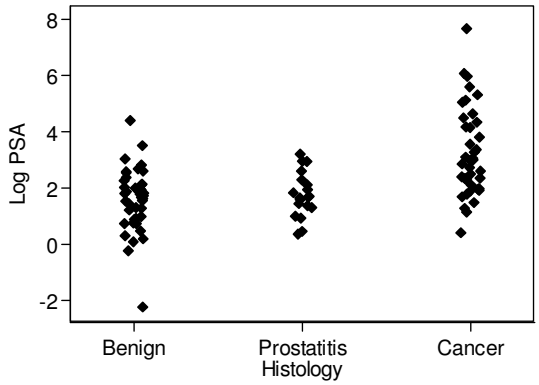

Figure 5. Log transformed PSA by prostate diagnosis

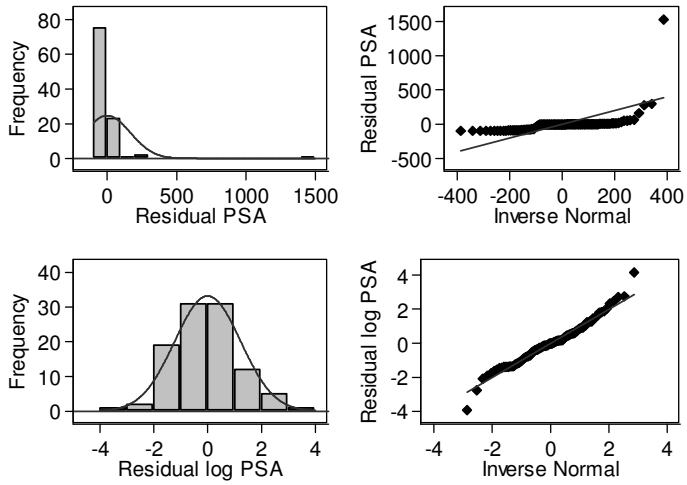

Figure 6. Histograms and Normal plots for the within-group residuals for the raw PSA data and the log transformed PSA data

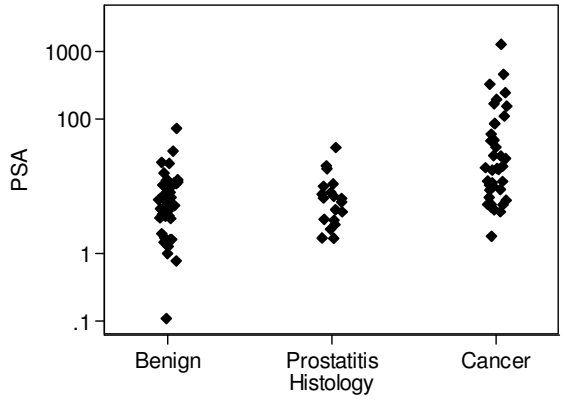

Figure 7. PSA by prostate diagnosis, shown on a logarithmic scale.

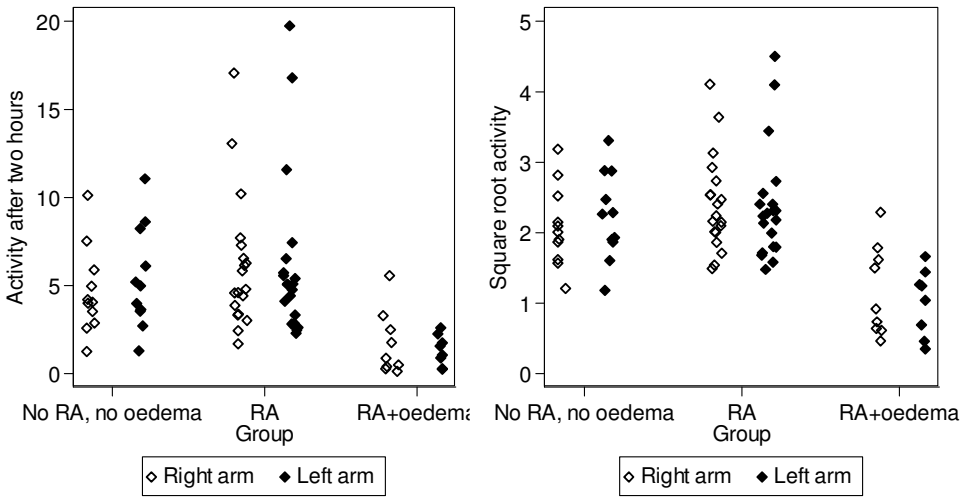

Figure 8. Arm lymphatic flow in rheumatoid arthritis with oedema (data of Kiely *et al.*, 1995)

Table 1. Biceps skinfold thickness (mm) in two groups of patients

Crohn's Disease				Coeliac Disease	
1.8	2.8	4.2	6.2	1.8	3.8
2.2	3.2	4.4	6.6	2.0	4.2
2.4	3.6	4.8	7.0	2.0	5.4
2.5	3.8	5.6	10.0	2.0	7.6
2.8	4.0	6.0	10.4	3.0	

Variance-stabilising transformations also tend to make distributions Normal. There is a mathematical reason for this, as for so much in statistics. It can be shown that if we take several samples from the same population, the means and variances of these samples will be independent if and only if the distribution is Normal. This means that uniform variance tends to go with a Normal Distribution. A transformation which makes variance uniform will often also make data follow a Normal distribution and *vice versa*.

There are many other transformations which could be used, but you see them very rarely. We shall meet one other, the logistic transformation used for dichotomous data, in Week 7. By far the most frequently used is the logarithm. This is particularly useful for concentrations of substances in blood. The reason for this is that blood is very dynamic, with reactions happening continuously. Many of the substances we measure are part of a metabolic chain, both being synthesised and metabolised to something else. The rates at which these reactions happen depends on the amounts of other substances in the blood and the consequence is that the various factors which determine the concentration of the substance are multiplied together. Multiplying and dividing tends to produce skew distributions. If we take the logarithm of several numbers multiplied together we get the sum of their logarithms. So log transformation produces something where the various influences are added together and addition tends to produce a Normal distribution. The square root is best for fairly weak relationships between variability and magnitude, i.e. variance proportional to mean or standard deviation proportional to the square root of the mean. The logarithm is next, for standard deviation proportional to the mean, and the reciprocal is best for very strong relationships, where the standard deviation is proportional to the square of the mean. In the same way, the square root removes the least amount of skewness and reciprocal the most.

The square root can be used for variables which are greater than or equal to zero, the log and the reciprocal can only be used for variables which are strictly greater than zero, because neither the logarithm nor the reciprocal of zero are defined. We shall look at what to do with zero observations in Section 6.

Which transformation should we use for what kind of data? For physical body measurements, like limb length or peak expiratory flow, we often need use only the raw data. For concentrations measured in blood or urine, we usually try the log first, then if this is insufficient try the reciprocal. For counts, the square root is usually the first thing to try. There are methods to determine which transformation will best fit the data, but trial and error, with scatter plots, histograms and Normal plots to check the shape of the distribution and relationship between variability and magnitude, are usually much quicker because the computer can produce them almost instantaneously.

3. Are transformations cheating?

At about this point, someone will ask ‘Aren’t transformations cheating?’. Data transformation would be cheating if we tried several different transformations until we found the one which gave the result we wanted, just as it would be if we tried several different tests of significance and chose the one which gave the result nearest to what we wanted, or compared treatment groups in a clinical trial using different outcome variables until we found one which gave a significant difference. Such approaches are cheating because the P values and confidence intervals we get are wrong. However, it is not cheating if we decide on the analysis we want to use before we see its result and then stick to it.