# Topic 8: Transformations of the data [S&T Ch. 9.16]

## 8. 1. The assumptions of ANOVA

For ANOVA, the linear model for the RCBD is:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

There are four key assumptions implicit in this model.

## 8. 1. 1. Additive effects

The treatment terms $\tau_i$ and error terms $\varepsilon_{ij}$ are assumed to be *added* (as opposed to, for example multiplied). A common way to say this is that "The treatment remains constant over replications and the replication effect remains constant over treatments". In terms of the model above, the treatment effect $\tau_i$ is independent of j, the index of replications, and the replication effect, which is the error term $\varepsilon_{ij}$, acts the same for all treatments.

If multiplicative effects were present, the model would more appropriately be:
$$Y_{ij} = \mu(\tau_i)(\beta_j) + \varepsilon_{ij}$$

*Tukey's 1-df Test for non-additivity* (Syllabus Topic 6.4.1) is a tool for testing if the departure from additivity is statistically significant in an RCBD with one observation per cell (treatment x block combination). If there is a significant multiplicative effect, a transformation can be used to restore the additive effects and generate a more sensitive ANOVA (smaller MSE by eliminating the interaction component).

## 8. 1. 2. Independence of errors

For the ANOVA to be valid, the values of the $\varepsilon_{ij}$ (i.e. the *errors* or *residuals*) must be statistically independent from one another. Failure to meet this assumption is often the result of failing to randomize properly.

Similarity among experimental units adjacent in space or time is called **positive autocorrelation**. The regular alternation of positive and negative errors is a manifestation of **negative autocorrelation**. While independence of errors in a sequence of continuous variates may be tested using the differences between adjacent values (Biometry page 394), the process of randomly allocating the treatments to the experimental units usually ensures that the $\varepsilon_{ij}$ will be independent. One ordinarily does not need to do more than a true randomization to satisfy this requirement. If another assumption is violated, however, it is often found that the assumption of independence of errors is also violated.

## 8. 1. 3. Normally distributed errors

For the ANOVA to be valid, the error terms $\varepsilon_{ij}$ (i.e. residuals) must be normally distributed. If not, the probabilities provided in the F Table will not be completely accurate. Although this is the least influential assumption on the F test, it is good practice to verify normality since all

statements of probability are based on normal distributions. For small to medium sample sizes, the Shapiro-Wilk test is recommended. If the sample size is $\geq 2000$, the Kolmogorov-Smirnov statistic (ST&D 564-568) is recommended.

### 8. 1. 4. Homogeneity of Variance

For the ANOVA, and particularly the mean comparisons to be valid, the variance among experimental units must be uniform across all levels of treatment. *This is the most influential assumption on the F test, and no conclusions should be drawn from the ANOVA if this assumption is violated.* The reason this assumption is so influential on the F test is because MSE, the denominator of the F test, is the *average* within-treatment variance. The very act of taking an average assumes that all within-treatment variances are estimates of the same value. If different groups of treatment levels have different variances, the average variance will be an inadequate standard to make comparisons between means.

It is possible to fail this assumption. For example, the variance among EUs within a treatment level may scale with the treatment mean. This is often the case when the treatments have substantial effects on the response variable. The variances may also be unequal but have no apparent relation to the treatment means. The following example, based on one from Little and Hills, shows what can happen when homogeneity of variance is violated.

| Treatment | Replicate | | | | | Total | Mean | $s^2$ |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | | |
| A | 3 | 1 | 5 | 4 | 2 | 15 | 3 | 2.5 |
| B | 6 | 8 | 7 | 4 | 5 | 30 | 6 | 2.5 |
| C | 12 | 6 | 9 | 3 | 15 | 45 | 9 | 22.5 |
| D | 20 | 14 | 11 | 17 | 8 | 70 | 14 | 22.5 |

Little and Hills first analyze the data using Fisher's protected LSD

| Source of variation | df | SS | MS | F |
|---|---|---|---|---|
| Treatments | 3 | 330 | 110 | 8.8* |
| Error | 16 | 200 | 12.5 | |

Since the F statistic is significant, we compute the LSD for $\alpha$ 0.05 significance level as LSD = 4.74. Therefore the difference between means A and B is not significant while that between C and D is. Analyzing A and B separately from C and D yields the opposite result. We do a two-sample test between A and B and between C and D:

| Source of variation | df | SS | MS | F |
|---|---|---|---|---|
| Treatments A B | 1 | 22.5 | 22.5 | 9* |
| Error | 8 | 20.0 | 2.5 | |

| Source of variation | df | SS | MS | F |
|---|---|---|---|---|
| Treatments C D | 1 | 62.5 | 62.5 | 2.8 NS |
| Error | 8 | 180 | 22.5 | |

When analyzed separately, the difference between A and B is significant while that between C and D is not.  The reason for this is that the variance for treatments C and D is larger than that for A and B. More specifically, by taking an average of all the variances (the MSE), those treatments with smaller variances were being unfairly penalized and those treatments with larger variances were being tested with an unreasonably small error.

> The consequences of moderate heterogeneity of variances are not too serious for the overall test of significance, but are very dangerous for contrasts and mean comparisons.

Among the different tests for homogeneity of variances, Bartlett's test (S&T, p. 471) and Levene's test are the most widely used. Bartlett's test is a chi-square goodness of fit test and Levene's test is an ANOVA of the square of the residuals.

If Levene's test is significant and the problem of heterogeneity of variances cannot be corrected by transformation, you can perform a **Welch's variance-weighted ANOVA** to test for differences among group means. This option is available in SAS means statement (means trtmt / welch;) or Analyst (One-Way ANOVA → Tests) but only for CRD (one–way ANOVA). The other alternative in this case is to analyze the data using **non-parametric statistics.**

## 8. 2. Transformations

If the evidence indicates that the assumptions for an analysis of variance cannot be satisfied, two courses of action are available. The first is to carry out a different test which does not require the rejected assumptions, such as a non-parametric test or a Welch's variance-weighted ANOVA. The second is to transform the response variable in such a manner that the resulting *transformed variable* meets the assumptions of the analysis. Once this is accomplished, you can perform the analysis on the transformed variable.

When a statistical test that was not significant with the original data suddenly becomes significant after a data transformation, people may feel suspicious.

**What is the justification for transforming the data?**

To begin, there is really no scientific necessity to employ the common linear or arithmetic scale to which we are accustomed.  If a relationship is multiplicative on a linear scale, it may make much more sense to think of it as an additive relationship on a logarithmic scale.  The square root of a variable is another frequent transformation.  The square root of the surface area of an organism may be a more appropriate measure of the effects of physiological and evolutionary forces on some biological variable than is the area itself.  Evidence of this is reflected in the fact that, often, the square root of the surface area is normally distributed while the surface areas themselves exhibit a skewed distribution.  In many cases, experience has shown that is better to express some variables in terms of logarithms (pH values), square roots (areas), or reciprocals (microbiological titrations).  As soon as one accepts the idea that *the scale of measurement is arbitrary*, it becomes clear that is valid to transform a variable in order to satisfy the assumptions of the analysis.

A fortunate fact about transformations is that very often several departures from the assumptions of ANOVA are cured simultaneously by a single transformation to a new scale. Simply by making the data homoscedastic (i.e. having uniform variances), normality and additivity of main effects may also be restored.

Four transformations will be discussed here: the logarithmic transformation, the square root transformation, the angular or arcsine transformation and the power transformation.

## 8. 2. 1 The log transformation (Little and Hill)

Whenever the **standard deviations** (not the variances) of samples are roughly proportional to the means, an effective transformation may be a log transformation.  Another criterion for deciding on this transformation is the evidence of multiplicative rather than additive main effects, as evidenced by significant Tukey's test. Finally, frequency distributions skewed to the right are often made more symmetrical by transformation to a logarithmic scale.

While logarithms to any base can be used, common logarithms (base 10) or natural logarithms (base e) are generally the most convenient.  Data with values $\leq 0$ cannot be transformed using logarithms, because the log of negative numbers is undefined and the log of zero is -∞.  In such cases, one may add a constant number to every data point so that the lowest value in the dataset becomes 1.  Data containing a large number of zeros would probably be handled better by some other method.

The following data set (Little and Hills 1978) will be used as an example of the effects of the log transformation on the assumptions of the ANOVA. The dependent variable is weight (lbs) of vitamin-treated and control animals; the design is an RCBD.

**Data table for an RCBD with 4 blocks and 6 treatments.**

| Species—Treatment | I | II | III | IV | Total | Mean | $s_i^2$ |
|---|---|---|---|---|---|---|---|
| **Block** | | | | | | | |
| Mice—control | 0.18 | 0.30 | 0.28 | 0.44 | 1.2 | 0.3 | 0.0115 |
| Mice—vitamin | 0.32 | 0.40 | 0.42 | 0.46 | 1.6 | 0.4 | 0.0035 |
| Subtotals | **0.50** | **0.70** | **0.70** | **0.90** | **2.8** | **0.35** | |
| Chickens—control | 2.0 | 3.0 | 1.8 | 2.8 | 9.6 | 2.40 | 0.3467 |
| Chickens—vitamin | 2.5 | 3.3 | 2.5 | 3.3 | 11.6 | 2.90 | 0.2133 |
| Subtotals | **4.5** | **6.3** | **4.3** | **6.1** | **21.2** | **2.65** | |
| Sheep—control | 108.0 | 140.0 | 135.0 | 165.0 | 548.0 | 137.0 | 546.0 |
| Sheep—vitamin | 127.0 | 153.0 | 148.0 | 176.0 | 604.0 | 151.0 | 425.3 |
| Subtotals | **235.0** | **293.0** | **283.0** | **341.0** | **1152.0** | **144.0** | |
| Totals | 240.0 | 300.0 | 288.0 | 348.0 | 1176.0 | 49.0 | |

In SAS a new transformed variable is generated after the input statement

```
data XXX;
input trtmnt $ block weight @@;
logwt= log10(weight);
cards;
```

In this data, there are no zeros, but since the lowest value is 0.18, the data was multiplied by 10 before taking logs. The next table compares the results of the analyses of the original and transformed response variables.

| | Original data | Transformed data |
|---|---|---|
| **Shapiro-Wilks (residuals)** | W= 0.95,  p= 0.32 | W= 0.97,  p= 0.56 |
| **Tuckey's test of nonadditivity** | F=  545,  **p= 0.0001** | F= 1.74 ,  **p= 0.21** |
| **Levene's test of homogeneity of $s^2$** | F= 2.5,  **p= 0.07** | F= 1.78,  **p= 0.17** |
| **Treatment effect** | F= 174.4, p= 0.0001 | F= 1959.4, p= 0.0001 |
| **Species effect** | F= 434.5, p= 0.0001 | F= 4883.0, p= 0.0001 |
| **Vitamin effect** | F=  1.14, **p= 0.30** | F=  16.4, **p= 0.0011** |
| **Species * Vitamin[a]** | F= 1.0, p= 0.39 | F= 1.58,  p= 0.24 |

[a] The calculation of the P value for the interaction was obtained coding the 6 treatments into a factorial design with animal and vitamin as classification variables. We will see this in the next lecture.

When this data was analyzed without any transformation, no significant effect of the vitamin treatment was detected. This seems very strange because every animal in every replicate receiving the vitamin showed a greater weight than the corresponding control animal. It also seems strange that no significant differences were detected for the interaction between vitamin effects and species, since the apparent response to vitamins is so different in the different species.

These unexpected results suggest that there is something wrong with the ANOVA, and indeed there is. When the assumptions of the ANOVA are checked, significant Tukey's and Levene's tests demonstrate that there is significant non-additivity of main effects as well as significant heterogeneity of variances. A quick look at the residuals confirms these violations of the ANOVA assumptions:
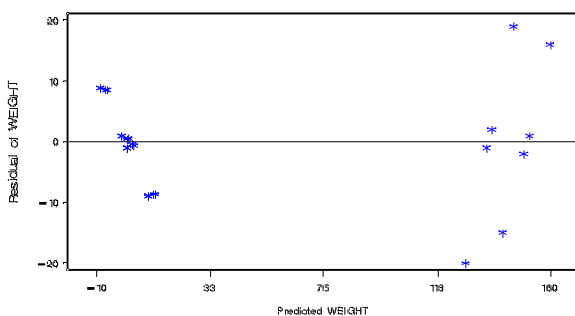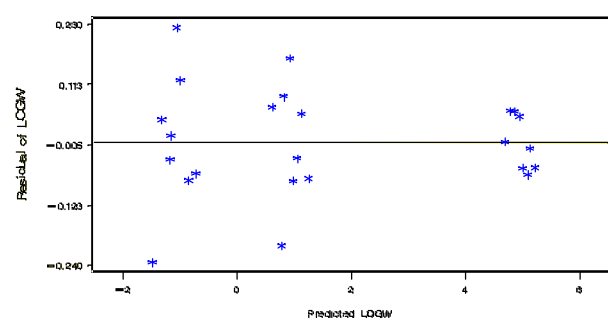


Residuals vs. predicted (**original data**).        Residual vs. predicted (**transformed data**).

The presence of multiplicative effects (see the smiley pattern in the above plot) and a rough proportionality between standard deviations and means suggest that a logarithmic transformation may be appropriate. Indeed, the transformed data meet all the ANOVA assumptions. Also, the results using the transformed data make more sense, particularly in terms of the significant effect of the vitamin treatment.

The interaction between species and vitamins is still not significant, which may seem puzzling. But recognize that the question is now different. With the original data, the interaction question was, "Does the amount of change in weight due to vitamins vary from species to species?"

With the log-transformed data, the question is now, "Does the ***proportion*** or ***percent*** change in weight due to vitamins vary from species to species?"

Did we get more satisfying results with the transformed data because we were simply playing with the numbers, fishing around until we got a result we liked? Or was the transformation we used justified and is the new analysis valid? Values for Tukey's and Levene's tests were significant for the original data but non-significant for the transformed data (even Shapiro-Wilk improved under the log transformation). The new analysis is valid because the transformed data satisfies *all* the assumptions of the ANOVA. But as you see with the interaction question, *interpretation* of results using transformed data can be tricky.

**Effect of log transformation on means and variances**

The data below illustrate the effect of the *ln* transformation on a multiplicative effect of a treatment that increases control values by 50% (*1.5). Notice the similar variance of 0.2740 after the *ln* transformation.

|  |  | **Control** |  |  |  | **Mean** | **Variance** |
|---|---|---|---|---|---|---|---|
| *Y* | 20 | 40 | 50 | 60 | 80 | 50 | 500 |
| *ln(Y)* | 2.9957 | 3.6889 | 3.9120 | 4.0943 | 3.820 | 3.8146 | 0.2740 |

|  |  | **Treatment (1.5*control)** |  |  |  | **Mean** | **Variance** |
|---|---|---|---|---|---|---|---|
| *Y* | 30 | 60 | 75 | 90 | 120 | 75 | 1125 |
| *ln(Y)* | 3.4012 | 4.0943 | 4.3175 | 4.4998 | 4.7875 | 4.2201 | 0.2740 |

## 8. 2. 2. The square root transformation (Little and Hill)

Whenever the response variable is a count of relatively rare events (e.g. insect counts on a leaf, blood cells within a gridded region of a hematocytometer, etc.) the data tend to follow a special distribution called a *Poisson distribution*. In this context, a "rare event" is an event that has a very low probability of occurring in any individual. For example, suppose that in a lot of lettuce seed, 0.1% of the seed carries mosaic disease virus. The probability that any individual seed carries the virus is then only 1/1000; so as far as a single seed is concerned, this is a very rare event. If we take 100 samples of 1000 seeds each from such a lot, we will obtain approximately these results:

> 37　samples will contain 0 infected seeds
> 37　samples will contain 1 infected seed
> 18　samples will contain 2 infected seeds
> 6　samples will contain 3 infected seeds
> 2　samples will contain 4 infected seeds

It is obvious that this looks very little like a normal distribution. This Poisson distribution has a very interesting characteristic—**the variance is equal to the mean**. In actual practice, the variance is generally somewhat larger than the mean because other factors, in addition to

sampling variation, are affecting the occurrence of the events being counted. At any rate, the variance tends to be proportional to the mean, thus **violating the assumption that the variances and means are not correlated**.

Another example of data of this kind is found in insect counts, such as those made from a standard number of sweeps with a net. The probability of finding an insect at a particular spot selected at random at one particular time is indeed a rare event.

Data of this kind can be made more nearly normal and at the same time the variances can be made relatively independent of the means by using a square root transformation. Actually, for count values near or less than 10, it is better to use:

$$Y_{transformed} = \sqrt{Y_{original} + \frac{1}{2}} \quad [ \quad \textbf{\textcolor{red}{SAS: SQRT(variable+0.5)}}$$

**Example.** Number of lygus bugs per 50 sweeps with a net. The experiment is an RCBD, testing 10 insecticides (A-J) and a control (K).

| Treatment | Block | | | | Total | Mean | $s_i^2$ |
|---|---|---|---|---|---|---|---|
| | I | II | III | IV | | | |
| A | 7 | 5 | 4 | 1 | 17 | 4.25 | 6.25 |
| B | 6 | 1 | 2 | 1 | 10 | 2.50 | 5.67 |
| C | 6 | 2 | 1 | 0 | 9 | 2.25 | 6.92 |
| D | 0 | 1 | 2 | 0 | 3 | 0.75 | 0.92 |
| E | 1 | 0 | 1 | 2 | 4 | 1.00 | 0.67 |
| F | 5 | 14 | 9 | 15 | 43 | 10.75 | 21.58 |
| G | 8 | 6 | 3 | 6 | 23 | 5.75 | 4.25 |
| H | 3 | 0 | 5 | 9 | 17 | 4.25 | 14.25 |
| I | 4 | 10 | 13 | 5 | 32 | 8.00 | 18.00 |
| J | 6 | 11 | 5 | 2 | 24 | 6.00 | 14.00 |
| K | 8 | 11 | 2 | 6 | 27 | 6.75 | 14.25 |

In this case, the variable is transformed with the following code:

```
Data Example;
   Do Trtmt = 1 to 11;
      Do Block = 1 to 4;
         Input Count @@;
            TransCount = sqrt(Count+0.5);
         Output;
      End;
   End;
Cards;
```

**A comparative summary of the results:**

|  | Original data | Transformed data |
|---|---|---|
| **Shapiro-Wilks (residuals)** | W= 0.98,  p= 0.73 | W= 0.987,  p= 0.95 |
| **Tuckey's test of nonadditivity** | F=  0.63,  p= 0.44 | F= 0.07 ,  p= 0.80 |
| **Levene's test of homogeneity of s²** | F= 1.61,  p= 0.15 | F= 0.90,  p= 0.55 |
| **Treatment effect** | F= 3.7,  p= 0.0026 | F= 4.04,  p= 0.0014 |

Though the data do not violate any of the assumptions, we can see that the transformed data meet the assumptions better than the original data (look particularly at Levene's Test, which moved from a low p-value of 0.1463 to a much stronger 0.6245). This improvement in the data can be seen in the scatterplots of residual vs. predicted values, shown below. Evidence of a correlation between treatment means and treatment variances manifests itself in the plot of the original data as a funnel shape. This funnel pattern is not visible in the plot of the transformed data.
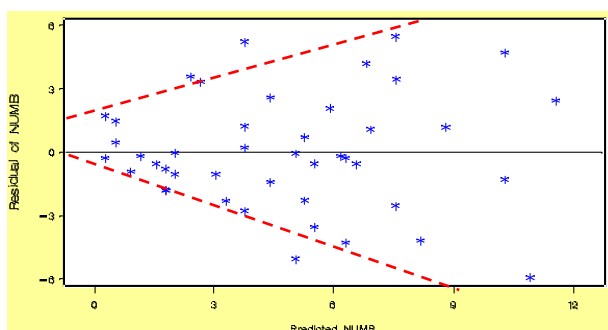


Residuals vs. Predicted: **Original data.**
Correlation mean-variance **r=0.89 \*\*** (*P*=0.001)
(*P*=0.1)
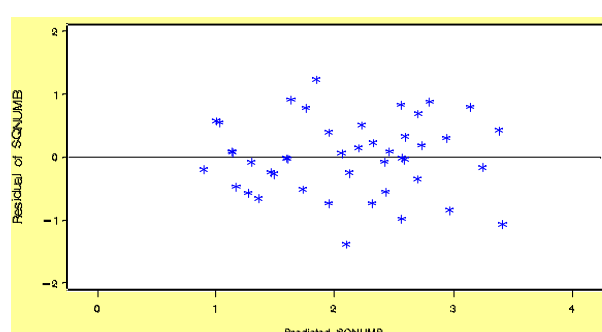
Residuals vs. Predicted: **Transformed data.**
Correlation mean-variance **r=0.37 NS**

The two analyses are not very different from one another, since both show highly significant treatment effect. However, the F value is about 10% higher after transformation and some differences appear when separating means.

The *weighted means* shown below are obtained by "**detransforming**" the means of the transformed data back to the original units. This is done by squaring the transformed means and subtracting one-half (i.e. by performing the inverse of the transformation function). The means obtained in this way are **smaller** than those obtained directly from the raw data because more weight is given to the smaller variates. This is as it should be, since in a Poisson distribution the smaller variates are measured with less sampling error (i.e. contain more information) than the larger ones. In presenting the results of such an experiment, it is best to use the *detransformed* means, making it clear in the report how they were obtained. Here are the Tukey mean separations performed on the original data and the transformed data ($\alpha = 0.05$):

```
Original Data
Tukey Grouping              Means        N      Trtmt

            A              10.750        4      F
    B       A               8.000        4      I
    B       A               6.750        4      K
    B       A               6.000        4      J
    B       A               5.750        4      G
    B       A               4.250        4      H
    B       A               4.250        4      A
    B                       2.500        4      B
    B                       2.250        4      C
    B                       1.000        4      E
    B                       0.750        4      D
```

```
Transformed Data    Detransformed
    Tukey Grouping              Means        N      Trtmt

            A                 10.3445        4      F
    B       A                  7.5957        4      I
    B       A     C            6.3084        4      K
    B       A     C            5.6073        4      J
    B       A     C            5.5851        4      G
    B       A     C            3.9416        4      H
    B       A     C            2.2060        4      B
    B             C            1.7970        4      C
    B             C            0.9028        4      E
                  C            0.6130        4      D
```

Note the differences in significance groupings in this example. In this case, the groupings derived from the transformed data are more reliable because the transformed data satisfied the assumptions of the analysis better than the original data.

**F** and **B** are declared different in the original data, but not in the transformed data.

**D** and **I** are declared different in the transformed data, but not in the original data.

This is a nice illustration of the fact that **the general effect of the square root transformation is to increase the precision with which we measure differences between *small* means.** This is a direct result of the shape of the square root function, shown below.
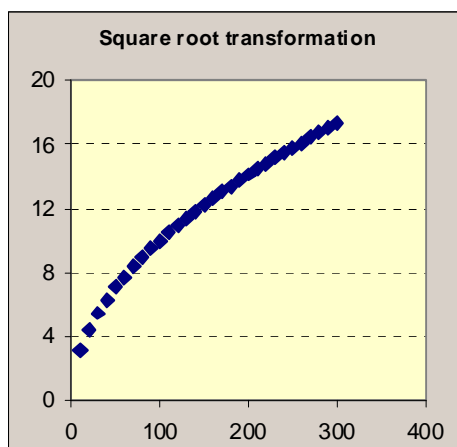
Small values are spread out relative to large values, making them easier to discriminate from one another. This is highly desirable in insect control work, where we are generally not as interested in differences between two relatively ineffective treatments as we are in comparing treatments that give good control.

Note that the values of the *detransformed* means are smaller than the original means. This is also due to the shape of the curve and is what makes this transformation appropriate for data following a *Poisson* distribution. The square root transformation essentially gives higher weight to smaller values, those values with higher information content (lower variance).

> In general, we can say that data requiring the square root transformation do not violate the ANOVA assumptions nearly as drastically as data requiring a log transformation. Consequently, the changes in the analysis brought about by the transformation are not nearly as spectacular.

## 8. 2. 3. The arcsine or angular transformation (Little and Hill)

Number of lettuce seeds germinating in samples of 50.

| Treatment | Blocks | | | Mean | $s_i^2$ |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | | |
| 1 | 0 | 0 | 1 | 0.33 | 0.33 |
| 2 | 0 | 1 | 0 | 0.33 | 0.33 |
| 3 | 0 | 0 | 1 | 0.33 | 0.33 |
| 4 | 0 | 2 | 0 | 0.67 | 1.33 |
| 5 | 2 | 0 | 0 | 0.67 | 1.33 |
| 6 | 0 | 2 | 3 | 1.67 | 2.33 |
| 7 | 7 | 10 | 7 | 8.00 | 3.00 |
| 8 | 11 | 12 | 15 | 12.67 | 4.33 |
| 9 | 13 | 18 | 18 | 16.33 | 8.33 |
| 10 | 22 | 16 | 13 | 17.00 | 21.00 |
| 11 | 24 | 13 | 18 | 18.33 | 30.33 |
| 12 | 23 | 21 | 29 | 24.33 | 17.33 |
| 13 | 24 | 29 | 29 | 27.33 | 8.33 |
| 14 | 37 | 28 | 27 | 30.67 | 30.33 |
| 15 | 42 | 41 | 40 | 41.00 | 1.00 |
| 16 | 39 | 41 | 45 | 41.67 | 9.33 |
| 17 | 41 | 45 | 40 | 42.00 | 7.00 |
| 18 | 47 | 41 | 43 | 43.67 | 9.33 |
| 19 | 45 | 42 | 48 | 45.00 | 9.00 |
| 20 | 46 | 42 | 48 | 45.33 | 9.33 |
| 21 | 49 | 46 | 48 | 47.67 | 2.33 |
| 22 | 48 | 49 | 48 | 48.33 | 0.33 |
| 23 | 50 | 49 | 48 | 49.00 | 1.00 |
| 24 | 49 | 49 | 50 | 49.33 | 0.33 |

Another kind of data that may require transformation is that based on counts expressed as percentages or proportions of the total sample. Such data generally exhibit what is called a

*binomial distribution* rather than a normal distribution. One of the characteristics of such a distribution is that the variances are related to the means, though in quite a different way than the types of data discussed before. Data discussed in 8.2.1 and 8.2.2 showed a positive correlation between means and variances. In binomial data, variances tend to be small at the two ends of the range of values (close to zero and 100%) but larger in the middle (around 50%).

The data above are from a completely randomized experiment on lettuce seed with 24 treatments, each replicated three times. Treatments are arranged in order of magnitude of their means. Note that there is a strong tendency for the variances at the extremes to be smaller than those in the middle of the range. This is typical of binomial data.

The arcsine transformation involves taking the inverse sine of the square root of each observation, expressed as a proportion of a total:

$$Y_{transformed} = \arcsin\left(\sqrt{\frac{Y_{original}}{Total\ Possible\ Counts}}\right)$$

In this case, since the total number of seeds in each replication is 50, the arcsine transformation would be:

$$Y_{transformed} = \arcsin\left(\sqrt{\frac{Y_{original}}{50}}\right)$$

In SAS, the transformation would be coded as follows:

```
Data Example;
   Do Trtmt = 1 to 24;
      Do Block = 1 to 3;
         Input Count @@;
            TransCount = arsin(sqrt(Count/50));
         Output;
      End;
   End;
Cards;
```

A comparative summary of the results:

**Comparison of transformed and original data**

|  | Original data | Transformed data |
|---|---|---|
| **Levene's test of homogeneity of $s^2$** | F= 2.43,     p= 0.0048 | F=0.99,     p= 0.49 |
| **Treatment effect** | F= 148.12,  p= 0.0001 | F= 100.14,  p=0.0001 |

As indicated in the above table, the arcsine transformation is appropriate for this kind of data because it tends to <mark>spread the values at both ends of the distribution</mark> compared with the central region (i.e. it tends to homogenize variances).  This effect is due to the shape of the arcsine function, a function which returns the number (in radians) whose sine is in the range $0 \rightarrow 1$:
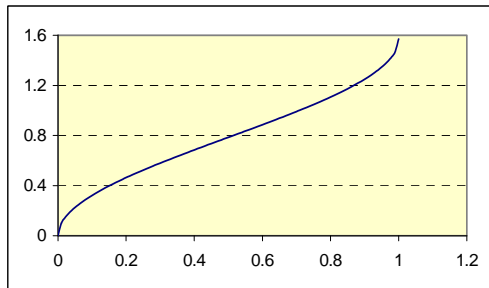


$$f(x) = arcsine(x) = sine^{-1}(x)$$

In general, proportion data of the sort described above should be transformed with the arcsine function if the range of percentages is greater than 0.40.  Otherwise, it is scarcely necessary. In this case, the arcsine transformation was successful is removing the correlation between treatment means and variances.  Indeed, the pattern of the residuals observable in the raw data is no longer apparent in the transformed data:
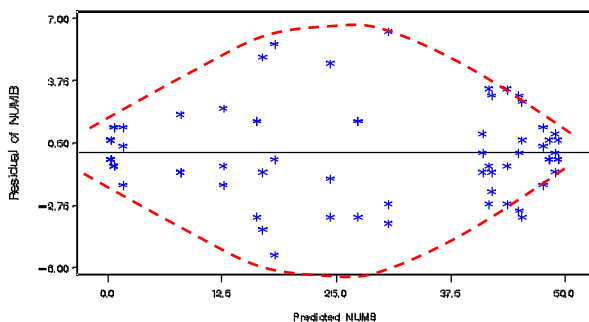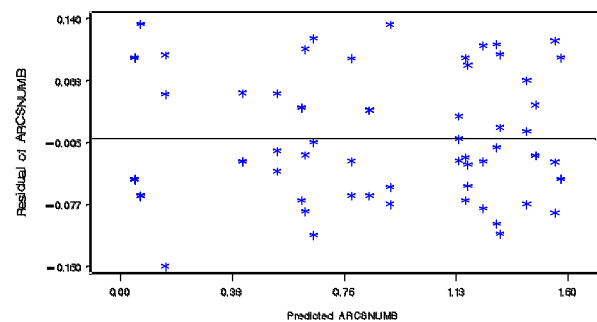


Residuals vs. predicted. **Original data**          Residual vs. pred. **Transformed data**

As you've seen with other transformations, the ANOVA results of arcsine-transformed data are not necessarily different from those of the original data, at least at the level of the overall ANOVA F test.  The important differences emerge in individual comparisons (e.g. contrasts and mean separations).  A comparison of Tukey mean separations is presented in the next page:

| | Original Data | | | Transformed Data | |
|---|---|---|---|---|---|
| **Original group** | **Original Means** | **Trtmt** | **Detransformed Means** | **Transformed group** | |
| A | 49.333 | 24 | 49.554 | A | |
| A | 49.000 | 23 | 49.348 | A | |
| A | 48.333 | 22 | 48.370 | A B | |
| A | 47.667 | 21 | 47.827 | A B | |
| A | 45.333 | 20 | 45.637 | A B | |
| A | 45.000 | 19 | 45.302 | A B | |
| A | 43.667 | 18 | 43.914 | A B | |
| A | 42.000 | 17 | 42.134 | B C | |
| A | 41.667 | 16 | 41.834 | B C | |
| A | 41.000 | 15 | 41.015 | B C | |
| B | 30.667 | 14 | 30.789 | D C | |
| B | 27.333 | 13 | 27.341 | D E | |
| B C | 24.333 | 12 | 24.333 | D E F | |
| D C | 18.333 | 11 | 18.216 | D E F G | |
| D C | 17.000 | 10 | 16.905 | D E F G | |
| D E C | 16.333 | 9 | 16.285 | E F G | |
| D E | 12.667 | 8 | 12.631 | F G | |
| E F | 8.000 | 7 | 7.952 | H G | |
| F | 1.667 | 6 | 1.111 | H I | |
| F | 0.667 | 5 | 0.225 | I | |
| F | 0.667 | 4 | 0.225 | I | |
| F | 0.333 | 1 | 0.112 | I | |
| F | 0.333 | 3 | 0.112 | I | |
| F | 0.333 | 2 | 0.112 | I | |

Which set of conclusions should we accept?

The answer is simple: we should accept the conclusions based on the more valid analysis. In this case, it is the analysis of the transformed data.

> **We do not transform data to give us results more to our liking**. We transform data so that the analysis will be *valid* and the conclusions *correct*.

## 8. 2. 4. The power transformation

(Hinz, P. N. and H. A. Eagles. 1976. Estimation of a transformation for the analysis of some agronomic and genetic experiments. Crop Science 16: 280-283.)

Field experiments are commonly conducted using replicated trials over a broad range of locations and environmental conditions; and often the means and the residual variances differ markedly across environments. Furthermore, there tends to be a positive correlation between the mean and the residual variance. In such instances, an analysis of the combined data over environments must contend with this heterogeneity in the residual variance. The transformation

of the dependent variable is a valid technique for reducing this relationship between the mean and the residual variance.

The choice of an optimal transformation from the many possible alternatives is not always obvious, especially if the functional relationship between mean and variance is unknown. The power transformation method provides a means of selecting an appropriate transformation from the broad class of power transformations by employing the data itself to estimate the exponent needed to transform the original measures.

A power transformation is any transformation in which the original measure is raised to some power "**a**":

$$Y_{transformed} = \left( Y_{original} \right)^a$$

The purpose of Hinz and Eagles' power transformation method is to find the value of "**a**" most appropriate for the data (i.e. most effective at minimizing the mean-variance correlation and variance heterogeneity).

Generally, if the variances and means are <u>positively correlated</u> (i.e. as means increase, variance increases), the optimal value of "**a**" will be < 1.

If the variances and means are <u>negatively correlated</u> (i.e. as means increase, variance decreases, or vice versa), the optimal value of "**a**" will be > 1.

The square root transformation discussed previously is one specific case of a power transformation, where **a** = 1/2.  If "**a**" is found to be approximately <u>equal to 0</u>, the <u>log</u> transformation is recommended.  So:

$$Y_{transformed} = \left( Y_{original} \right)^a \text{ if } \mathbf{a} \neq 0$$
$$Y_{transformed} = \log \left( Y_{original} \right) \text{ if } \mathbf{a} \approx 0$$

And note that
$$Y = \sqrt{X} \quad \text{if } \mathbf{a} = \tfrac{1}{2}$$
$$Y = \log X \quad \text{if } \mathbf{a} = 0$$
$$Y = 1/X \quad \text{if } \mathbf{a} = -1$$

A potentially effective value of the power of the transformation '**a**' can be determined by obtaining the slope '**m**' of the linear regression of $\log(s_i^2)$ versus $\log(\overline{Y}_i)$ (i.e. the best-fit slope of a log(Variance) * log(Mean) scatterplot), and then solving for **a** = 1 - **m**/2.

Suppose there are t groups of observations and $s_i^2$ and $\overline{Y}_i$ are the sample variance and sample mean, respectively, of the i$^{th}$ group.  We want to find the slope **m** that best satisfies the linear equation:

$$\log(s_i^2) = m * \log(\overline{Y}_i) + b$$

Use SAS, R, Excel, or any other program to perform a linear regression of the logarithms of the observed variances on the logarithms of the means in order to find the least squared estimate (**m**) of the slope.

Then, the data can be transformed using a power **a** = 1 - **m** /2.
**Examples of power transformation**

Using the means and variances from the lygus bug table (**Topic 8.2.2. Square root transformation**), we can write a small SAS program:

```
Data Power;
   Input Mean Variance;      * Treatment means and variances from original data;
   LogMean = Log10(Mean);              * Calculate the log of treatment means;
   LogVar = Log10(Variance);        * Calculate the log of treatment variances;
Cards;
4.25   6.25
2.50   5.67
2.25   6.92
0.75   0.92
1.00   0.67
10.75 21.58
5.75   4.25
4.25  14.25
8.00  18.00
6.00  14.00
6.75  14.25
;
Proc GLM;
   Model LogVar = LogMean;
Run;
Quit;
```

This code performs a linear regression of log(Variance) on log(Mean) of the original data.  The output includes an estimate (**m**) of the slope:

|  | Parameter | Estimate | Standard Error | t Value | Pr > |t| |
|---|---|---|---|---|---|
|  | Intercept | 0.108611900 | 0.12798558 | 0.85 | 0.4181 |
| Estimate of m → | LogMean | 1.238144871 | 0.19299582 | 6.42 | 0.0001 |

This output tells us that the best-fit line for this data is:

$$\log(s_i^2) = 1.2381 * \log(\overline{Y}_i^2) + 0.1086$$

So, **m**= 1.238q and the estimated power for transformation is:

**a** = 1 - (1.2381/ 2) = 0.38. Notice that this value is quite close to 0.5, the square root transformation.  To apply this recommended transformation, the following code could be used (SAS uses "**\*\***" to signify an exponent):

```
Data Example;
   Do Trtmt = 1 to 11;
      Do Block = 1 to 4;
         Input Count @@;
            TransCount = Count**(0.38);
         Output;
      End;
   End;
Cards;
```

Using $Y_{transformed} = (Y_{original})^{0.38}$ , Levene's test improves from p=0.15 in the untransformed data to p = 0.32 in the transformed data. The fact a square-root transformation gives better results for this particular dataset than the power transformation (Levene's test p=0.55) illustrates the point that this power transformation method does not render the optimal value of **a,** but only an estimate of that optimal value.

**Example 2**: Using the means and variances from the vitamin dataset (Topic 8.2.1 Logarithmic Transformation), we obtain:

$$\log(s_i^2) = 1.8557 * \log(\overline{Y}_i^2) - 1.3373$$

In this case, **m** = 1.8557 and the estimated power for transformation **a** = 1 – (1.8557/ 2) = 0.072, which is close to 0. In this instance, the power transformation yields consistently better results than the log transformation (see table below); though with **a** being so close to zero, the logarithmic transformation is also appropriate in this case.

|  | Log Transformed data | | Power Transformed data | |
|---|---|---|---|---|
|  | F | p | F | p |
| **Shapiro-Wilk** | W = 0.97 | 0.56 | W = 0.97 | 0.61 |
| **Levene's Test** | 1.78 | 0.17 | 1.10 | 0.40 |
| **Tukey's Nonadditivity Test** | 1.74 | 0.21 | 0.01 | 0.94 |

## 8.3. Reporting results from transformed data

When a transformation is applied, tests of significance are performed on the transformed data, but estimates of means are usually given in the familiar, untransformed scale. It is best to calculate means using the transformed data and then detransform them to the original units. In this way, we obtain correctly weighted means.

Since all the transformations discussed here are non-linear, confidence limits computed in the transformed scale will become asymmetrical when detransformed into the original scale. Stating the standard error in the original scale is therefore misleading. So, in reporting results of research with variables that require transformation, it is better to report detransformed means (i.e. estimates of means in the original scale) followed by their asymmetrical confidence limits, rather than by a standard error.

**Example**: Using log transformed data, the mean of Treatment B is found to be 1.176. The Tukey minimum significant difference is found to be 0.023. In the transformed scale, the confidence interval about the mean of Treatment B is symmetric:

$$\mu_B = \overline{Y}_B \pm MSD = 1.176 \pm 0.023 = \textbf{[1.153 , 1.199]}$$

The proper way to report this confidence interval is by detransforming these confidence limits:

$$\mu_B = [10^{1.153}, 10^{1.199}] = \textbf{[14.223 , 15.812]}$$

Notice that this is an asymmetric interval about the detransformed mean $10^{1.176} = 14.997$.

## 8.3.1 Detransforming *ln* transformed means back to the original scale

      The following example shows that when the mean of the logarithms is transformed back to the original scale, the result is the *Geometric Mean* of the original data:

|         |        |        |        |        |        | Mean   | Variance |
|---------|--------|--------|--------|--------|--------|--------|----------|
| Y       | 20     | 40     | 50     | 60     | 80     | 50     | 500      |
| *ln(Y)* | 2.9957 | 3.6889 | 3.9120 | 4.0943 | 4.3820 | **3.8146** | 0.2740   |

Geometric mean G $= (Y_1 * Y_2 * ... * Y_n)^{1/n}$
Geometric mean G $= (20*40*50*60*80)^{1/5} = 45.3586$

Note that when you detransform the mean of the log-transformed data back to the original scale, $e^{3.8146} = \textbf{45.3586}$.  This is the geometric mean of the original data, NOT the arithmetic mean (50).