TRANSFORMATION OF DATA

Rajender Parsad
I.A.S.R.I., Library Avenue, New Delhi-110 012

The interpretation of data based on analysis of variance (ANOVA) is valid only when the following assumptions are satisfied:

1. **Additive Effects:** Treatment effects and block (environmental) effects are additive.
2. **Independence of errors:** Experimental errors are independent.
3. **Homogeneity of Variances:** Observations have common variance.
4. **Normal Distribution:** Character under study follows a normal distribution.

Also the statistical tests t, F, z, etc. are valid under the assumption of independence of errors and normality of character under study.

The departures from these assumptions make the interpretation based on these statistical techniques invalid. Therefore, it is necessary to detect the deviations and apply the appropriate remedial measures.

- The assumption of independence of errors, *i.e*., error of an observation is not related to or depends upon that of another. This assumption is usually assured with the use of proper randomization procedure. However, if there is any systematic pattern in the arrangement of treatments from one replication to another, errors may be non-independent.
- The assumption of additive effects can be defined and detected in the following manner:

The effects of two factors, say, treatment and replication, are said to be additive if the effect of one-factor remains constant over all the levels of other factors. A hypothetical set of data from a randomized complete block design, with 2 treatments and 2 replications, with additive effects is given in Table 1.

Table 1

Treatment	Replication		Replication Effect
	I	II	I - II
A	190	125	65
B	170	105	65
Treatment Effect (A-B)	20	20	

Here, the treatment effect is equal to 20 for both replications and replication effect is 65 for both treatments.

When the effect of one factor is not constant at all the levels of other factor, the effects are said to be non-additive. A common departure from the assumption of additivity in biological experiments is one where the effects are multiplicative. Two factors are said to have multiplicative effects if their effects are additive only when expressed in terms of percentages. Table 2 illustrates a hypothetical set of data with multiplicative effects.

Table 2

Treatment	Replication		Replication Effect	
	I	II	I - II	100(I - II)/II
A	200	125	75	60
	(2.30103)	(2.09691)	(0.20412)	
B	160	100	60	60
	(2.20412)	(2.0000)	(0.20412)	
Treatment Effect (A-B)	40	25		
	(0.09691)	(0.09691)		
100 (A - B)/B	25	25		

In this case, the treatment effect is not constant over replications and the replication effect is not constant over treatments. However, when both treatment effect and replication effect are expressed in terms of percentages, an entirely different pattern emerges.

For such violations of assumptions, Logarithmic transformation is quite suitable. For illustration, the Logarithmic transformation of data in Table 2 is given in brackets.

This is, however a crude method for testing the additivity, statistical tests is available for testing the additivity of effects.

The assumptions of homogeneity of variances and normality are generally violated together. To test the validity of normality of character under study, one can take help of normal Probability Plot, D'Augstino's Test, Saphiro - Wilk's Test, etc. In general moderate departures from normality are of little concern in the fixed effects ANOVA as F - test is slightly affected but in case of random effects, it is more severely impacted by non - normality. The significant deviations of errors from normality, makes the inferences invalid. Hence, it is necessary to convert the data in some scale so that it follows a normal distribution, before being analysed.

To test the assumption of homogeneity of variances, we shall compute the variance and mean for each treatment across the replications (the range can be used in place of variance). The equality of variances is, then, tested using Bartlett's test for homogeneity of variances (see Appendix I).

Let Y_{ij} is the observation pertaining to i^{th} treatment $\left(i = 1(1)v\right)$ in the j^{th} replication $\left(j = 1(1)r_i\right)$, then

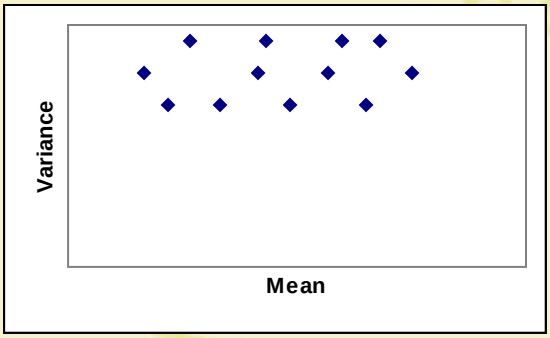
$$\text{Mean} = \bar{Y}_{i.} = \frac{1}{r_i}\sum_{j=1}^{r_i} Y_{ij}$$

$$\text{Variance} = S_i^2 = \frac{1}{r_i - 1}\sum_{j=1}^{r_i}\left(Y_{ij} - \bar{Y}_{i.}\right)^2$$

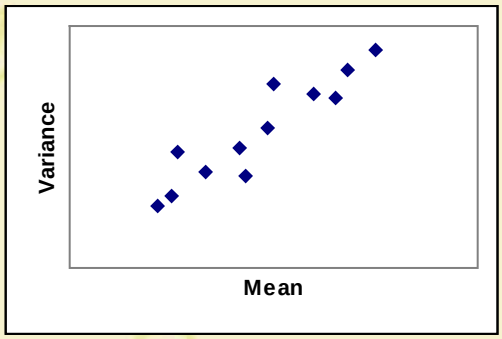
Now, if $S_{i.}^2$'s $(i = 1(1)v)$ are equal (constant), then, the variances are homogeneous. However, if Bartlett's test reject the hypothesis of equality of variances, then, variances are heterogeneous. The heterogeneity of variances can be classified into two types:

1. Where the variance is functionally related to mean.
2. Where there is no functional relationship between the variance and the mean.

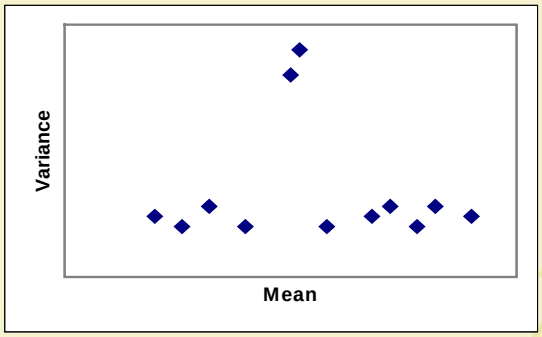
The above detection can also be done with the help of a scatter - diagram of mean and variances (or range).



(a) Homogeneous variance



(b) Heterogeneous variance where Variance is proportional to mean



(c) Heterogeneous variance without any functional relationship between variance and mean

The first of variance heterogeneity is usually associated with the data whose distribution is non-normal *viz.*, negative binomial, Poisson, binomial, etc. Data transformation is the most appropriate remedial measure, in such situations. With this technique, the original data are converted to a new scale resulting into a new data set that is expected to satisfy the homogeneity of variances. Because a common transformation scale is applied to all observations, the comparative values between treatments are not altered and comparison between them remain valid.

The second kind of variance heterogeneity usually occurs in experiments, where, due to the nature of treatments tested some treatments have errors that are substantially higher

(lower) than others. For example, in varietal trials, where various types of breeding material are being compared, the size of variance between plots of a particular variety will depend on the degree of genetic homogeneity of material being tested. The variance of F_2 generation, for example, can be expected to be higher than that of F_1 generation because genetic variability in F_2 is much higher than that in F_1. The variances of varieties that are highly tolerant of or highly susceptible to, the stress being tested are expected to be smaller than those of having moderate degree of tolerance. Also in testing yield response to a chemical treatment, such as, fertilizer, insecticide or herbicide, the non-uniform application of chemical treatments may result to a higher variability in the treated plots than that in the untreated plots. *Error partitioning is the remedial measure of such kind of heterogeneity.*

Here, we shall concentrate on those situations where character under study is non-normal and variances are heterogeneous and some function of means. Depending upon the functional relationship between variances and means, suitable transformation is adopted. The transformed variate should satisfy the following:

1. The variances of the transformed variate should be unaffected by changes in the means. This is also called the variance stabilizing transformation.
2. It should be normally distributed.
3. It should be one for which real effects are linear and additive.
4. The transformed scale should be done for which an arithmetic average from the sample is an efficient estimate of true mean.

The following are the three transformations, which are being used most commonly, in biological research.

a) Logarithmic Transformation
b) Square root Transformation
c) Arc Sine or Angular Transformation

a) Logarithmic Transformation

This transformation is suitable for the data where the variance is proportional to square of the mean or the coefficient of variation (S.D./mean) is constant or where effects are multiplicative. These conditions are generally found in the data that are whole numbers and cover a wide range of values. This is usually the case when analyzing growth measurements such as trunk girth, length of extension growth, weight of tree or number of insects per plot, number of eggmass per plant or per unit area etc.

For such situations, it is appropriate to analyze *logX* instead of actual data, *X*. When data set involves small values or zeros, $log(X+1)$, $log(2X+1)$ or $log\left(X+\frac{3}{8}\right)$ should be used instead of *logX*.

This transformation would make errors normal, when observations follow negative binomial distribution like in the case of insect counts.

b) Square-Root Transformation

This transformation is appropriate for the data sets where the variance is proportional to the mean. Here, the data consists of small whole numbers, for example, data obtained in counting rare events, such as the number of infested plants in a plot, the number of insects caught in traps, number of weeds per plot, parthenocarpy in some varieties of mango. This data set generally follows the Poisson distribution and square root transformation approximates Poisson to normal distribution.

For these situations, it is better to analyze $\sqrt{X}$ than that of *X*, the actual data. If X is confirmed to small whole numbers then, $\sqrt{X+\frac{1}{2}}$ or $\sqrt{X+\frac{3}{8}}$ should be used instead of $\sqrt{X}$.

This transformation is also appropriate for the percentage data, where, the range is between 0 to 30% or between 70 and 100%.

c) Arc Sine Transformation

This transformation is appropriate for the data on proportions, *i.e.*, data obtained from a count and the data expressed as decimal fractions and percentages. The distribution of percentages is binomial and this transformation makes the distribution normal. Since the role of this transformation is not properly understood, there is a tendency to transform any percentage using arc sine transformation. But only that percentage data that are derived from count data, such as % barren tillers (which is derived from the ratio of the number of non-bearing tillers to the total number of tillers) should be transformed and not the percentage data such as % protein or % carbohydrates, %N, etc. which are not derived from count data. The value of 0% should be substituted by $\left(\frac{1}{4n}\right)$ and the value of 100% by $\left(100-\frac{1}{4n}\right)$, where *n* is the number of units upon which the percentage data is based.

It is interesting to note here that not all percentage data need to be transformed and even if they do, arc sine transformation is not the only transformation possible. The following rules may be useful in choosing the proper transformation scale for percentage data derived from count data.

Rule 1: The percentage data lying within the range 30 to 70% is homogeneous and no transformation is needed.

Rule 2: For percentage data lying within the range of either 0 to 30% or 70 to 100%, but not both, the square root transformation should be used.

Rule 3: For percentage that do not follow the ranges specified in Rule 1 or Rule 2, the Arc Sine transformation should be used.

The other transformations used are reciprocal square root [$\frac{1}{\sqrt{X}}$, when variance is proportional to cube of mean], reciprocal [$\frac{1}{X}$, when variance is proportional to fourth power of mean] and tangent hyperbolic transformation.

The transformation discussed above are a particular case of the general family of transformations known as Box-Cox transformation.

d) Box-Cox Transformation

By now we know that if the relation between the variance of observations and the mean is known then this information can be utilize in selecting the form of the transformation. We now elaborate on this point and show how it is possible to estimate the form of the required transformation from the data. Box-Cox transformation is a power transformation of the original data. Let y_{ut} is the observation pertaining to the u^{th} plot, then the power transformation implies that we use y_{ut}'s as

$$y_{ut}^{*} = y_{ut}^{\lambda}.$$

Box and Cox (1964) have shown how the transformation parameter λ in $y_{ut}^{*} = y_{ut}^{\lambda}$ may be estimated simultaneously with the other model parameters (overall mean and treatment effects) using the method of maximum likelihood. The procedure consists of performing, for the various values of λ, a standard analysis of variance on

$$y_{ut}^{(\lambda)} = \begin{cases} \dfrac{y_{ut}^{\lambda} - 1}{\lambda \dot{y}_{ut}^{\lambda-1}} & \lambda \neq 0 \\ \dot{y}_{ut} \ln y_{ut} & \lambda = 0 \end{cases} \qquad \text{(A)}$$

where $\dot{y}_{ut} = \ln^{-1}\left[(1/n)\sum_{u=1}^{N}\sum_{t=1}^{n_u} \ln y_{ut}\right]$ is the geometric mean of the observations. The maximum likelihood estimate of λ is the value for which the error sum of squares, say $SS_e(\lambda)$, is minimum. Notice that we cannot select the value of λ by directly comparing the error sum of squares from analysis of variance on y^{λ} because for each value of λ the error sum of squares is measured on a different scale. Equation (A) rescales the responses so at error sums of squares are directly comparable.

Therefore, the λ can be estimated in three different ways *i.e.* by minimizing these error sum of squares.

This is a very general transformation and the commonly used transformations follow as particular cases. The particular cases for different values of λ are given below.

λ	Transformation
1	No Transformation
½	Square Root
0	Log
-1/2	Reciprocal Square Root
-1	Reciprocal

If any one of the observations is zero then the geometric mean is undefined. In the expression A, geometric mean is in denominator so it is not possible to compute that expression. For solving this problem, we add a small quantity to each of the observations.

Note: It should be emphasized that transformation, if needed, must take place right at the beginning of the analysis, all fitting of missing plot values, all adjustments by covariance etc. being done with the transformed variate and not with the original data. At the end, when the conclusions have been reached, it is permissible to 're-transform' the results so as to present them in the original units of measurement, but this is done only to render them more intelligible.

As a result of this transformation followed by back transformation, the means will rather be different from those that would have been obtained from the original data. A simple example is that without transformation, the mean of the numbers 1, 4, 9, 16 and 25 is 11. Suppose a square root transformation is used to give 1, 2, 3, 4 and 5, the mean is now 3, which after back- transformation gives 9. Usually the difference will not be so great because data do not usually vary as much as those given, but logarithmic and square root transformation always lead to a reduction of the mean, just as angles of equal formation usually lead to its moving away from the central value of 50%.

However, in practice, computing treatment means from original data is more frequently used because of its simplicity, but this may change the order of ranking of converted means for comparison. Therefore, to avoid such complexities, the procedure of converting the transformed means is preferred.

Although transformations make possible a valid analysis, they can be very awkward. For example, although a significant difference can be worked out in the usual way for means of the transformed data, none can be worked out for the treatment means after back transformation.

Some Useful References

Anderson,V.L. and McLean,R.A.(1974). *Design of Experiments: A realistic approach*. Marcel Dekker Inc., New York.

Bartlett,M.S.(1947). The use of transformation. *Biometrics*, **3**, 39-52.

Box,G.E.P. and Cox,D.R.(1964). An analysis of transformation. *J. Roy. Stastist. Soc*. **B**, **26**, 211-252.

Conover,W.J. and Iman,R.L.(1981). Rank transfomations as a bridge between parametric and non-parametric statistics (with discussion). *American Statistician*, **35**, 124-133.

D'Agostino, R.B. and Stephens, M. A.(1986). *Goodness-of-fit Techniques*. Marcel Dekkar, Inc., New York.

Dean, A and Voss, D (1999). *Design and Analysis of Experiments*. Springer, New York.

Dolby,J.L.(1963). A quick method for choosing a transformation. *Technometrics*, **5**, 317-326.

Draper,N.R. and Hunter,W.G.(1969). Transformations: Some examples revisited. *Technometrics*, **11**, 23-40.

Royston, P. (1992), Approximating the Shapiro-Wilk W-Test for non-normality. *Statistics and Computing*, **2**, 117 -119.

Shapiro, S. S. and Wilk, M. B.. (1965). An analysis of variances test for normality (Complete Samples). *Biometrika*, **52**, 591-611

Tukey, J.W.(1949). One degree of freedom for non-additivity. *Biometrics,* **5**, 232-242.

Appendix - I

Bartlett's Test for testing Homogeneity of Variances

Let there are m - independent samples drawn from a same population and i^{th} sample is of size n_i and $(n_1 + n_2 + ... + n_m) = N$. The null hypothesis for this test is $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_m^2$.

The alternative hypothesis is $H_1 : above\ \ not\ \ true\ \ for\ \ at\ \ least\ \ one\ \sigma_i^2$

For this test S_i^2 (sample variances) is taken as unbiased estimate of σ_i^2. The procedure involves computing a statistic whose sampling distribution is closely approximated by the χ^2 distribution with *m-1* degrees of freedom. The test statistic is

$$\chi_0^2 = 2.3026\frac{q}{c}$$

and null hypothesis is rejected when

$$\chi_0^2 > \chi_{\alpha,m-1}^2$$

where $\chi_{\alpha,m-1}^2$ is the upper α percentage point of χ^2 distribution with *m-1* degrees of freedom.

To compute χ_0^2, follow the steps:

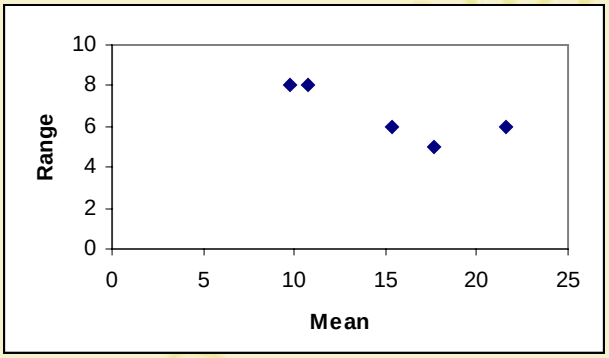
1. Compute mean and variance of all m-samples.
2. Obtain pooled variance $S_p^2 = \dfrac{\sum_{i=1}^{m}(n_i - 1)S_i^2}{N - m}$
3. Compute $q = (N - m)log_{10} S_p^2 - \sum_{i=1}^{m}(n_i - 1)log_{10} S_i^2$
4. Compute $c = 1 + \dfrac{1}{3(m-1)}\left(\sum_{i=1}^{a}(n_i - 1)^{-1} - (N - m)^{-1}\right)$
5. Compute χ_0^2.

Heterogeneity of variances can also be detected by using scatter plots of means and variance or range, residual Vs fitted values.

Example 1: Let there be five treatments each replicated 5 times and experiment is conducted using a randomized complete block design. The data obtained is presented alongwith means and variances of treatments as:

Treatment	Replication					Mean	Variance	Range
	I	II	III	IV	V	$\overline{Y_i}$	S_i^2	
A	7	7	15	11	9	9.8	11.2	8
B	12	17	12	18	18	15.4	9.8	6
C	14	18	18	19	19	17.6	4.3	5
D	19	25	22	19	23	21.6	6.8	6
E	7	10	11	15	11	10.8	8.2	8

A scatter plot of mean and range is given as follows:



It indicates the homogeneity of variances.

Bartlett's test:

Pooled Variance $\left(S_p^2\right) = \dfrac{4(11.2+9.8+4.3+6.8+8.2)}{20} = 8.06$

$$q = 20\log_{10} 8.06 - 4\left[\log_{10} 11.2 + \log_{10} 9.8 + \log_{10} 4.3 + \log_{10} 6.8 + \log_{10} 8.2\right]$$
$$= 0.45$$

$$c = 1 + \frac{1}{3(4)}\left(\frac{5}{4} - \frac{1}{20}\right) = 1.10$$

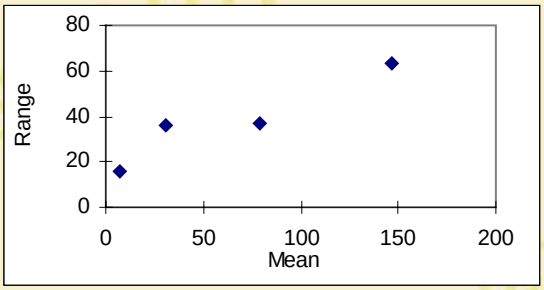
and the test statistic is $\chi_0^2 = 2.3026\,\dfrac{(0.45)}{1.10} = 0.93$.

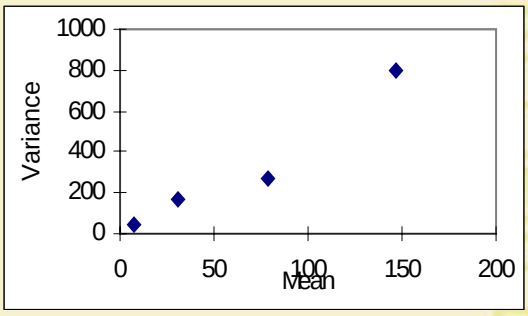
Since $\chi^2_{0.05,4} = 9.49$, we cannot reject the null hypothesis and conclude that all the five variances are same. The same conclusion is drawn from the scatter plots

Example 2: Suppose an entomologist is interested in determining whether four different kinds of traps caught equivalent insects when applied to same field. Each of the traps is used six times on the field and resulting data (number of insects per hour) are as shown below alongwith mean, variance and range.

Treatment	Replication						Mean	Variance	Range
	I	II	III	IV	V	VI	$\overline{Y_i}$	S_i^2	
A	3	1	12	7	17	2	7	40.4	16
B	9	29	21	24	28	45	31	168.4	36
C	63	84	97	61	98	71	79	270.8	37
D	172	118	109	172	143	168	147	798.8	63

A scatter plot of mean and variance and mean Vs range are given as follows:





Both plots indicate that variances are heterogeneous and variance is proportional to mean.

Bartlett's test:

Pooled Variance $\left(S_p^2\right) = \dfrac{5(40.4+168.4+270.8+798.8)}{20} = 319.5$

$$q = 20\log_{10}319.5 - 5\left[\log_{10}40.4 + \log_{10}168.4 + \log_{10}270.8 + \log_{10}798.8\right]$$
$$= 4.251$$

$$c = 1 + \frac{1}{9}\left(\frac{6}{5} - \frac{1}{20}\right) = 1.128$$

$\chi_0^2 = 8.678$.

Since $\chi_{0.05,3}^2 = 7.81$, therefore, we reject the null hypothesis and conclude that the variances are unequal. The $\dfrac{S_i^2}{\overline{Y}_{i.}}$ are 5.77, 5.43, 3.42 and 5.43, indicating that variance is proportional to mean. Therefore, squa`re root transformation should be used.