The Philippine Statistician
1992, Vol. 41,
Pages 18-34

VARIANCE STABILIZING AND NORMALIZING TRANSFORMATIONS FOR FIELD COUNT DAMAGE DATA IN RICE CAUSED BY THE STALK-EYED FLY (SEF), *DIOPSIS LONGICORNIS* MACQUART (DIPTERA: DIOPSIDAE)

by

M. B. De Ramos
Institute of Mathematical Sciences and Physics
University of the Philippines, Los Baños

C. J. Angela
National Crop Protection Center
University of the Philippines, Los Baños

R. C. Joshi
International Institute of Tropical Agriculture
Ibadan, Nigeria

M. A. Ynalvez
Los Baños Computer Center
University of the Philippines, Los Baños

Abstract

Three sets of field count damage data caused by the stalk-eyed fly (SEF) *Diopsis longicornis* Macquart, at three rice growth stages were statistically analyzed to determine if the Box-Cox power-shift transformation could stabilize the variances and normalize the distribution. The analyses of the residuals of the transformed data revealed that the kurtosis coefficients decreased to insignificant values, the skewness coefficients also decreased but remained significant, and the normality statistic increased but also remained significant. On the other hand, the transformation did homogenized the variances for some combinations of λ and c. The best combination of x and c that would satisfy homogeneity of variances assumption and that would make the distribution a little bit normal is $\lambda = -1$ and $c = 1$. Therefore, in analyzing field count SEF damage data in rice which is expressed as number of damaged tillers per hill, the use of the reciprocal transformation with shifting constant equal to 1 is recommended.

Keywords: Poisson distribution, power-shift transformation, residual analysis, rice, stalk-eyed fly, *Diopsis longicornis*.

1. INTRODUCTION

The analysis of variance (ANOVA) is a widely used statistical technique in analyzing comparative experiments. However, this technique is valid only when the mathematical and statistical assumptions are met. These assumptions are: (i) the treatment effects and environmental effects are additive, (ii) the experimental errors are independent, (iii) the experimental errors have common variance, and (iv) the experimental errors are normally distributed. Failure to meet any one of these assumptions would affect both the level of significance and sensitivity of the F- statistic (Cochran, 1947).

When there is sufficient reason to believe that the underlying conditions required in the ANOVA are not met in the data under analysis, alternatives or techniques to make valid tests on such data possible is greatly demanded. For data which fail to meet the test conditions there are two approaches recommended by Bartlett (1947) namely, (i) bend the data to fit the assumptions of the ANOVA by making nonlinear transformation, or (ii) develop new method of analysis which approximate the original form of the data.

However, data obtained from many experiments do not usually satisfy the assumptions required, particularly normality and homogeneity of variance, so the use of a transformation is needed. Hence, the reason for transforming data are to find a metric in which the theoretical assumptions made in the analysis are more readily satisfied and to make the analysis simpler than would be possible (Draper and Hunter, 1969). The use of a transformation may also be necessary to normalize the distribution of the errors or to achieve greater consistency of variance (Dolby, 1963).

In many statistical experiments, one usually encounters data that are not normally distributed. For instance, data that consist of integers such as field counts which are randomly distributed over a number of units usually follow Poisson distribution. Data that are distributed as Poisson have a variance equal to the mean (Harcourt, 1963). To be able to make the variance of such kinds of data independent of the mean, a transformation is required.

With the availability of many types of data transformations, one is confronted with the problem of choosing the best transformation to use. Rather than simply trying various transformations in order to find out which one works best, Box-Cox (1964) developed a procedure for estimating the best transformation to normality within the family of power transformation given as

$$y(\lambda) = \frac{(y^{\lambda} - 1)}{\lambda}, \quad \lambda \neq 0$$

or

$$y(\lambda) = \ln y, \quad \lambda = 0$$

The modified Box-Cox transformation included a shifting constant c defined as

$$y(\lambda) = \frac{\left[(y+c)^{\lambda} - 1\right]}{\lambda}, \quad \lambda \neq 0$$

or

$$y(\lambda) = \ln(y+c), \quad \lambda = 0$$

The Box-Cox procedure have been used by many researchers since it was developed in 1964. Guerrero (1982) applied it to the study of binary response model and De Ramos (1983) used it in comparing the arcsine transformation and Box-Cox transformation in analyzing percentage data. Carrol and Ruppert (1984) used the Box-Cox procedure in fitting theoretical models to data. Barlev (1988) gave a simpler method of obtaining a class of variance stabilizing transformations. Tsai (1988) used power transformations in a two-stage procedure to achieve normality and homogeneity of the errors and to remove non-linearity of the regression function. Hinkley (1988) extended Lawrance's results concerning test of transformations in regression. Logothetis (1990) assessed the applicability of Box-Cox transformation for simplifying and statistically validating a "Taguchi analysis".

The problem of this study focused on the analysis of the field count damage (*dead-hearts*) data on rice, caused by the stalk-eyed fly (SEF) *Diopsis longicornis* Macquart (Diptera: Diopsidae). The questions raised were:

(1) Does the field count SEF damage data measured as number of damaged tillers per hill at various plant growth stages follow a Poisson distribution?

(2) Can the Box-Cox power-shift transformations make the distribution of the errors in SEF damage normal?

(3) Can the Box-Cox power-shift transformations stabilize the variance of SEF damage among the various plant growth stages?

With these questions in mind, this study was conducted with the following objectives:

(1) to determine if the field count SEF damage data expressed as number of damaged tillers per hill in rice follows a Poisson distribution.

(2) to determine if the Box-Cox power-shift transformations can make the distribution of the residuals in a one-way analysis of variance model follow a normal distribution, and

(3) to determine if the Box-Cox power-shift transformations can stabilize the variances of the experimental errors in SEF damage among the various plant growth stages.

2. METHODOLOGY

The Data

The data used in this study were SEF damaged tillers on the most commonly cultivated rice variety, ITA 306 (FARO 37). The field experiment was conducted at the National Cereal Research Institute (NCRI), Badeggi, Nigeria during the 1990 wet season (WS). One week after transplanting, 10 sampling stations each measuring 1M x 1M were established randomly in the field. In each station, 25 hills of the rice plants were randomly sampled starting from 14 days after transplanting (DAT) up to 70 DAT, at weekly intervals. SEF damage was assessed by counting the number of damaged tillers and total tillers per hill on the 250 randomly selected hills on each sampling occasion.

In this study the SEF damage data from only three growth periods representing the rice's early vegetative (14 DAT), maximum tillering (42 DAT) and maturity (70 DAT) stages were used in the analysis. These growth stages also represent to the varying degrees of vulnerability against SEF. In addition, since the sample size per sampling time was quite large (equal to 250) it was thought that by using data from three growth periods was sufficient to obtain reliable results that would answer the objectives of this study.

Fitting the Poisson Distribution

Data on counts such as the number of tillers per hill in rice may follow approximately a Poisson distribution. Thus, the first part of this study was to determine if the data sets for each sampling time follow three different Poisson distributions. On a per hill basis, the number of SEF damaged tillers denoted by a variate x can have values 0, 1, 2, ...,τ, where $r \leq r_{max}$, r_{max} being the total number of tillers in a hill of the rice plant. It was observed that in the 750 hills randomly selected, the values of r_{max} ranged from 8 to 54. Thus to determine if in each sampling time, damage data set follows a Poisson distribution, the variate x was denoted by x_{ij} defined as the number of damaged tillers in the jth class value of x at the ith sampling time, and f_{ij} was defined as the frequency or number of hills corresponding to x_{ij}, where $i = 1, 2, 3$, and $j = 1, 2, ..., k_i$, k_i being the number of discrete classes of the variate x at the ith sampling time. Therefore, the data representation $[x_{ij}, f_{ij}]$ gave the class values and class frequencies of the variate x.

If the variate x_i in the ith sampling time follows a Poisson distribution with parameter m_i, then the probability function of x_i is given by

$$f(x_i) = \frac{e^{-m_i}(m_i)^{x_i}}{x_i!}, \quad x_i = 0, 1, \ldots, \tau_i$$

where e ≈ 2.7183 is the base of the natural logarithm. To determine whether the data x_i fitted a Poisson distribution, the parameter m_i was estimated using the maximum likelihood estimate $\hat{m}_i$ given by

$$\hat{m}_i = \frac{\sum_{j=1}^{k_i} x_{ij} f_{ij}}{\sum_{j=1}^{k_i} f_{ij}}$$

$$= \frac{\sum_{j=1}^{k_i} x_{ij} f_{ij}}{250}, \quad i = 1,2,3$$

To test the goodness of fit of the data x_i with the estimated parameter m_i, the Kolmogorov-Smirnov D statistic defined as

$$D_i = \max_j \left[F(x_{ij}) - S(x_{ij}) \right], \quad i = 1,2,3 \text{ was used,}$$

$$\text{where } F(x_{ij}) = e^{-m_i} \sum_{l=0}^{x_{ij}} \frac{\hat{m}_i^l}{l!}, \quad l = 0,1,\ldots,x_{ij}$$

$$\text{and } S(x_{ij}) = \sum_{i=0}^{x_{ij}} \frac{f_{il}}{250},$$

The quantity $F(x_{ij})$ was the estimate of the theoretical relative cumulative frequencies and $S(x_{ij})$ was the sample relative cumulative frequencies. The significance of the D_i statistic was determined by comparing D_i with the critical value $D_{\alpha(n)} = D_{.05(250)}$. When $D_i \geq D_{.05(250)}$, the test was declared significant, which meant that the data set x_i in the ith sampling time did not follow a Poisson distribution, and when $D_i < D_{.05(250)}$, then the data set x_i fitted a Poisson distribution with parameter estimated by m_i.

Estimation of the Parameter l in the Box-Cox Power-Shift Transformation

The fact that the three data sets were counts which may follow Poisson distributions, the use of analysis of variance to compare the mean SEF damage levels of the three sampling times will not be valid because of non-normality and inequality of the variances. Thus in practice the reasons for using the transformations $(x)^{1/2}$ or $(x+.5)^{1/2}$ and $\ln(x)$ or $\ln(x+1)$ for count data are to more or less normalize the data as well as to stabilize the variance. As an alternative to these commonly used transformations for count data, Box-Cox power-shift transformation has been used in many data analyses to attain normality and stability of variances. Since in the observed three data sets the values of the variate x predominantly consisted of 0's, the use of Box-Cox power shift transformation

$$x(\lambda) = \begin{cases} \dfrac{(x+c)^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(x+c) & \text{for } \lambda = 0 \end{cases}$$

was in order..

Let us denote by x_{ij} the number of damaged tillers at the jth hill of the ith sampling time, where i = 1,2,3, j=1,2,...,250. The data set x_{ij} were then transformed to

$$x_{ij}(\lambda) = \begin{cases} \dfrac{(x_{ij}+c)^{\lambda}-1}{\lambda} & \text{when } \lambda \neq 0 \\ \ln(x_{ij}+c) & \text{when } \lambda = 0 \end{cases}$$

for values of λ ranging from -4 to 2 at interval 0.25 and for fixed values of c equal to 0.25, 0.50, 0.75, 1, 2, and 3. The reason for using λ in the range from -4 to 2 was to include the case of $\lambda = -1$ (reciprocal transformation), $\lambda = 0$ (logarithmic transformation), and $\lambda = 1/2$ (square root transformation), while the reason for the use of $c = .25, .50, .75, 1, 2$, and 3 was to include the conventional constants normally added to x which are $c = .5$ and $c = 1$. With the 25 values of λ and 6 values of c, a total of 25 x 6 = 150 data sets of 750 observations each were created.

In order to determine the value of λ for a given c that will normalize and stabilize the variances of the three sets of data, the one-way classification model

$$x_{ij}(\lambda) = \mu + \tau_i + e_{ij}, \quad e = 1,2,3 \quad j = 1,2,\ldots,250$$

was fitted for each value of λ per given fixed value of c in order to generate 25 values of the mean square error (MSE) for every value of c. For each value of λ, the likelihood function $L(\lambda)$ defined as

$$L(\lambda) = -\frac{\gamma}{2}\ln \text{ MSE}(\lambda) + (\lambda - 1)\frac{\gamma}{n}\sum_{i}^{3}\sum_{j}^{250}\ln(x_{ij}+c)$$

where MSE (λ) is the mean square error of the transformed data using λ, γ is the degrees of freedom associated with MSE, and n is the total sample was computed. The estimate of λ, say λ, was determined as value of λ corresponding to the maximum value of $L(\lambda)$. Hence,

$$\hat{\lambda} \rightarrow \max_{\lambda} L(\lambda)$$

The above analysis of variance procedure also gave information as to what happened to the values of the F-statistic in testing the significance of τ_i in the model.

Test of Normality

The question of how normal could the distribution of the residuals be after transformation was answered by analyzing the residuals e_{ij} defined by

$$\hat{e}_{ij} = x_{ij}(\lambda) - \hat{\mu} - \hat{\tau}_i, \quad i = 1,2,3 \quad j = 1,2,\ldots,250$$

The test for the normality of the residuals e_{ij} was carried out for each value of λ and for each fixed value of c using the SAS package done on a mainframe IBM 4331. The results of the residual analysis indicated the values and significance of the Shapiro-Wilk W Statistics, as well as the coefficient of Skewness (g_1) and coefficient of kurtosis (g_2). When the distribution of the residuals follow approximately a normal distribution, the values of W are close to unity, while the value of g_1 and g_2 are close to 0. Since the significance of the g_1 and g_2 statistics were not indicated in the output, the approximate standard normal Z-statistic were computed using

$$Z(g_1) = \frac{g_1}{\sqrt{6/n}}, \qquad Z(g_2) = \frac{g_2}{\sqrt{24/n}}$$

where $n = 750$.

Test of Homogeneity of Variances

One of the assumptions of analysis of variance is that the errors must have homogeneous variances. Thus another question this study would like to answer was what power-shift transformation of the infestation data could make the within sampling time variances homogeneous. To get the answer, the Box-Cox power-shift transformation was applied to the raw data for values of λ from m-4 to 2 at an interval of .25, and for c values equal to .25, .50, .75, 1, 2, and 3. After the transformations were made, the Bartlett's homogeneity of variances x^2 test was applied to compare the within sampling time variances. The combinations of λ and c that gave insignificant x^2-values indicated the power-shift transformations that stabilized the variances.

The first step was to compute the sampling time sampling variance s_i and the pooled sample variance s_p^2 by the formulas

$$s_i^2 = \frac{\sum_{j}^{n_i} \left(x_{ij}^{(\lambda)} - \bar{x}_i\right)^2}{(n_i - 1)}$$

$$s_p^2 = \frac{\sum_{i=1}^{3} (n_i - 1) s_i^2}{\sum_{i=1}^{3} (n_i - 1)}$$

where $\bar{x}_i = \sum_{i=1}^{3} x_{ij}^{(\lambda)}$, n_i=250 for all i.

Then the χ^2 statistic was computed as

$$\chi^2_{(2\ df)} = \frac{1}{M}\left[\sum_{i}^{3} (n_i - 1)\log_{10} s_p^2 - \sum_{i=1}^{3} (n_i - 1)\log_{10} s_i^2\right]$$

where

$$M = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{3} \frac{1}{n_i - 1} - \frac{1}{\sum (n_i - 1)}\right], \quad k = 3$$

The computations of the χ^2 - statistic was also done using SAS programme on a mainframe IBM machine.

3. RESULTS AND DISCUSSION

Fitting of Poisson Distribution to the Data

The distribution of the SEF damage data which is expressed as of damaged tillers per hill were highly skewed to the right, with the highest frequency at $x = 0$ (no infestation) in all of the three plant growth stages (Table I). The pattern of frequency distributions were quite similar, $x = 0, f_0 = 176$; $x = 1, f_1 = 63$; and $x = 2, f_2 = 11$ for 14 DAT; $x = 0, f_0 = 204$; $x = 1, f_1 = 31$; $x = 2, f_2 = 12$; $x = 3, f_3 = 3$ for 42 DAT; and $x = 0, f_0 = 188$; $x=1, f_1 = 30$; $x = 2, f_2 = 12$, $x = 3, f_3 = 15$, $x = 4, f_4 = 5$ for 70 DAT. The number of damaged tillers had a mean of (x) of .340 and variance (s^2) of .31395 for 14 DAT; a mean of .256 and variance of .35990 for 42 DAT; and a mean of .476 and variance of .94922 for 70 DAT. It was noted that the mean and variance at 14 and 42 DAT did not show much difference which signified that each follow a Poisson distribution, but at 70 DAT the mean and variance were already much different, the variance being about twice that of the mean signifying that the distribution was no longer Poisson.

Using the mean x as estimate of the parameter m of a Poisson distribution, the estimate of the Poisson probability functions were:

$$f(x) = e^{-.340}(.340)^x/x! \text{ at 14 DAT,}$$

$$f(x) = e^{-.256}(.256)^x/x! \text{ at 42 DAT,}$$

$$\text{and } f(x) = e^{-.476}(.476)^x/x! \text{ at 70 DAT.}$$

These estimates of Poisson probability functions were then used to estimate the theoretical relative cumulative frequency distributions $F(x_{ij})$ Table II. For example, at 14 DAT (i=1),

$$F(x_{ij}) = e^{-.340}\sum_{l=0}^{x_{ij}} .340^l/l!, \quad l = 0,1,\ldots,x_{ij}$$

Hence,

$$F(0) = e^{-.340}.340^0/0!$$
$$= .712$$
$$F(1) = f(0) + f(1)$$
$$= e^{-.340}\left[.340^0/0! + .340^1/1!\right]$$
$$= .953$$
$$F(2) = fd(0) + f(1) + f(2)$$
$$= e^{-.340}\left[.340^0/0! + .340^1/1! + .340^2/2!\right]$$
$$= 1.0$$

The other values of $F(x_{1j})$, namely, $F(x_{2j})$ for 42 DAT and $F(x_{3j})$ for 70 DAT were computed by the same procedure.

To test the goodness of fit of Poisson distributions to the data, the sample relative cumulative frequency distribution $S(x_{ij})$ were also computed by the formula

$$S(x_{ij}) = \sum_{l=0}^{x_{ij}} f_{il}/250, \quad l = 0,2,\ldots,x_{ij}$$

For example, at 14 DAT (i=1),

$$S(x_{ij}) = \sum_{l=0}^{x_{ij}} f_{il} / 250$$

Hence,

$$S(0) = f_{10}/250$$
$$= 176/250$$
$$= .704$$

$$S(1) = (f_{10} + f_{11})/250$$
$$= \frac{176+63}{250}$$
$$= .956$$

$$S(2) = (f_{10} + f_{11} + f_{12})/250$$
$$= \frac{176+63+11}{250}$$
$$= 1.0$$

The test of goodness of fit, i.e., whether $S(x_{ij})$ was in close agreement with $F(x_{ij})$ for all i and j was done by the Kolmogorov-Smirnov D statistic which is also shown in Table II. The results indicated that the distributions of the number of damaged tillers per hill at 14 DAT and 42 DAT followed Poisson distributions, but that for 70 DAT it deviated from a Poisson type distribution. The reason for no longer satisfying a Poisson distribution at 70 DAT, was that higher damage occurred that stretched the curve further to the right giving a variance much higher than the mean.

Estimates for x in the Box-Cox Power-Shift Transformation

The mean square error values (MSE) obtained from the analysis of variance of the transformed variates $x^{(\lambda)}$ are shown in Table III. It will be noted that when c = .25, .50 and .75. The MSE values attained a minimum value when λ was varied from -4 to 2. For example, at c = .25, the minimum value of MSE was. 55 corresponding to λ =.50; at c=.5, the minimum value of MSE was .35 corresponding to λ = 0 and λ = .5; at c = .75 the minimum value of MSE was .103 corresponding to λ = -3.0. However, when c = 1, 2, 3, the MSE values increased exponentially as λ was varied from -4 to 2. These results indicated that the minimum values of MSE occurred at values of λ lower than -4.0.

The results shown in the Table II also indicated that when the power $\lambda < 1$, the values of MSE decreased as the value of c was increased from .25 to 3; the value of MSE was constant for all values of c when λ = 1; and when the power $\lambda > 1$, the values of MSE increased as the values of c were increased. For example, at λ = -4, the MSE values decreased from 729.9 to .0000103 when c was increased from .25 to 3, and when λ = 2, the MSE values increased from 3.08 to 14.7 when c increased from .25 to 3, respectively.

By substituting the values of MSE in the likelihood function

$$L(\gamma) = -\frac{\gamma}{2} \ln \text{ MSE}(\lambda) + (\lambda - 1)\frac{\gamma}{n} \sum_{i}^{3} \sum_{j}^{250} \ln(x_{ij} + c)$$

where γ= 747 and n = 750 for all the values of λ and c, the maximum values of $L(\lambda)$ was found to be 1205.8 corresponding to $\hat{\lambda}$ = -2.0 when c = .25; equal to 1042.7 corresponding to $\hat{\lambda}$ = -2.75 when c = .50; and

equal to 962.9 corresponding to $\hat{\lambda}$ = -3.50 when c = .75 (Table IV). As the minimum values of MSE were not attained by varying the values of $\hat{\lambda}$ for c = 1, 2, and 3, the maximum values of $L(\hat{\lambda})$ when c = 1, 2, and 3 were also not attained. Thus no estimates of λ were obtained for c = 1, 2, and 3. Therefore, the Box-Cox power-shift transformations for the number of damaged tillers were as follows:

$$x^{(\lambda)} = \frac{(x+.25)^{-2.0} - 1}{(-2.0)} \quad \text{for } c = .25$$

$$x^{(\lambda)} = \frac{(x+.50)^{-2.75} - 1}{(-2.75)} \quad \text{for c = 0.50}$$

$$\text{and} \quad x^{(\lambda)} = \frac{(x+.75)^{-3.50} - 1}{(-3.50)} \quad \text{for c = .75}$$

Effect of the Box-Cox Power-Shift Transformations on the Distributions of Residuals

The results of the analysis of the residuals e_{ij} were summarized in terms of the Shapiro-Wilk W statistic (Table V), coefficient of skewness g_1 (Table VI), and Kurtosis coefficient g_2 (Table VII),

$$\text{where} \quad \hat{e}_{ij} = x_{ij}^{(\lambda)} - \hat{\mu} - \hat{\tau}, \; i = 1,2,3, \; j = 1,2,\ldots,250$$

$$x_{ij}^{(\lambda)} = \begin{cases} \dfrac{(x_{ij}+c)^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln(x_{ij}+c) & \text{for } \lambda = 0 \end{cases}$$

$$\hat{\mu} = \sum_{i=1}^{350}\sum_{j=1}^{250} x_{ij}^{(\lambda)} \Big/ 250 - \mu$$

$$\text{and} \quad \hat{\tau}_i = \sum_{j=1}^{250} x_{ij}^{(\lambda)} \Big/ 250 - \hat{\mu}$$

As reference points, the values of normality statistics W, g_1 and g_2 were .633, 3.07, and 14.35, respectively, when λ = 1, *i.e.*, there were only shifting of constant transformations made on the raw data. All of these normality statistics were very significantly different from the critical values of $W_{.01(750)}$ = .930, $g_1[.01,750]$ = 0.175, and $g_2[.01,750]$ = .350.

As to what happened to the values of W when the values of λ where changed from -2.0 to 2.0 at fixed value of c can be seen in Table V. It was noted that the value of W attained a maximum value of .663 or .662 within the ranges of λ and c. So this maximum value was still highly significant compared to the critical value $W_{.01(750)}$ = .930. Note that $P[W \leq W_{.01(750)}] = .01$, hence $P(W \leq .663) << .01$. The maximum values of W were observed when x = .50 and c = .25 and .50; λ = .25 and c = .75 and 1.0; λ = 0 and c = 2; and λ = -.25 and c = 3.0. For the other combinations of λ and c, the values of W obtained were smaller than .662 which means that the distributions of the residuals for those transformations were even more non-normal.

The effects of the transformations on the skewness coefficient (g_1) can be seen in Table VI. It was noted that as λ was varied from -2.0 to 1 and the c value was varied from .25 to 3, the values increased to a constant value of 3.07. The lowest value of g_1 was 1.14 when λ = -2.0 and c = .25. The values of g_1 increased very rapidly for $\lambda > 1$, the rate of increase being faster for lower values of c.

There were profound effects of the power-shift transformations on the values of the Kurtosis coefficient (g_2). At any given fixed value of c, the values of g_2 increased exponentially as the values of λ was increased from -2.0 to 2.0. For example, at $c = .25$, $g_2 = -.636$ when $\lambda = -2.0$ and $g_2 = 120.96$ when $\lambda = 2.0$. The desirable values of g_2 were those which values were close to 0 or to the critical value $g_2[.01,750] = .350$. Therefore, the combinations of λ and c that made the values of g_2 not significant or almost not significant were:

($\lambda = -.25$, c = .25) giving $g_2 = -.131$
($\lambda = -.25$, c = .50) giving $g_2 = .322$
($\lambda = -.50$, c = .50) giving $g_2 = -.046$
($\lambda = -1.0$, c = .75) giving $g_2 = -.211$
($\lambda = -1.0$, c = 1.0) giving $g_2 = .003$
and ($\lambda = -2.0$, c = 2) giving $g_2 = .016$.

The Effects of the Power - Shift Transformation on the Stabilization of Variances

The results of the Bartlett's - X^2 test for homogeneity of variances are shown in Table VIII. For the seven values of λ from - 4 to .5 and six values of c from .25 to 3, the values of X^2 obtained ranged from .0224 to 31.26. Comparing these values of X^2 with the critical value $X^2_{.05(2)} = 5.991$, some values were significant [$X^2 \geq 5.991$] and some were not [$X^2 < 5.991$]. Those values of X^2 that were not significant are underlined to indicate the appropriate combinations of λ and c that made the within sampling time variances homogeneous.

Based on the results shown in Table VIII, some of the best combinations of x and c that stabilized the variances of the damage data were as follows:

(1) $\lambda = .25$, c = .25
(2) $\lambda = 0$, c = .25 or c = .50
(3) $\lambda = -1$, c = 1 or c = 2
(4) $\lambda = -2$, c = 3

Therefore, the power-shift transformation that can be applied to stabilize the variances of the number of SEF damaged tillers, is any of the following simple transformations:

(1) Logarithmic ($\lambda = 0$):

$$x^{(0)} = \ln (x + .25)$$
or $$x^{(0)} = \ln (x + .50)$$

(2) Reciprocal ($\lambda = -1$):

$$x(-1) = \frac{(x+1)^{-1} - 1}{-1} = \frac{x}{x+1}$$

or

$$x(-1) = \frac{(x+2)^{-1} - 1}{-1} = \frac{x+1}{x+2}$$

The Effects of the Power-Shift Transformations on the F Statistics

From the previous sections it was found out that the power-shift transformation was able to stabilize the variances of the SEF damage field count data. It was also found out from the residual analyses that the normality was not attained; the distribution remained skewed; and the distribution attained normal heights. The question therefore is, which of the F values given in Table IX are considered valid for testing the differences between the damage levels in three plant growth stages. But basing from the skewness coefficient (g_1), Kurtosis coefficient (g_2), normality test (w) and homogeneity test (X^2), the best combination for λ and c were $\lambda = -1$ and $c = 1.0$. For this combination, the F value was 3.50 at 2 and 747 degrees of freedom. The corresponding Type I probability level is about .05. On the other hand, the F value corresponding to no transformation ($\lambda = 1$) was 6.68 with probability level about .001. The statistical implication of these results is that on the average even small differences in damage levels will be declared significant more oftenly even though the actual damage levels in the populations are the same.

4. SUMMARY AND CONCLUSIONS

Three sets of field count SEF damage data on rice variety, ITA 306 for three plant growth stages were statistically analyzed with the main objective of determining whether the Box-Cox power-shift transformations could stabilize the variances and normalize the distribution of the data. The results are as follows:

(1) At 14 and 42 DAT the distributions of the SEF damaged tillers followed the Poisson distribution. The mean damage levels and variabilities were similar in magnitudes, thus made the data to fit the Poisson model. However, at 70 DAT, the distribution of the data did not follow a Poisson distribution because even though the mean damage level increased, the variability also increased to a magnitude that was about double of the mean.

(2) The values of the power of X that maximized the Box-Cox log likelihood function were - 2.0 for $c = .25$; -2.75 for $c = .5$; and - 3.50 for $c = .75$. These results differed very much from those obtained in residual analysis for which good choices for λ were 0, .25, and .5 based on the normality test w; -.25, -.5 and -1.0 based on the kurtosis statistic (g_2); and -1.0 and -2.0 based on the skewness statistic (g_1).

(3)The power-shift transformation had profound positive effects on stabilizing the variances. The good choices for the power of X were $\lambda = 0$ for any shifting constant c equal to $\lambda = .25$, $\lambda = .50$ and $\lambda = .75$ and $\lambda = -1$ for c equal to 1.

(4)From the results above, the use of the power $\lambda = -1$ and shifting constant $c = 1$ is recommended. This means that the appropriate transformation for field count SEF damaged tillers (*dead-hearts*) data in rice is reciprocal, which in simplified form is:

$$X(\lambda) = \frac{x}{x+1}$$

TABLE I

Frequency and percentage distributions of the number of SEF damaged tillers per hill after transplanting, ITA 306, 1990 WS.

Number of damaged tillers	Days after transplanting (DAT)					
	14		42		70	
(x)	(f)	%	(f)	%	(f)	%
0	176	0.704	204	0.816	188	0.752
1	63	0.252	31	0.124	30	0.120
2	11	0.044	12	0.048	12	0.048
3	0	0	3	0.012	15	0.060
4	0	0	0	0	5	0.020
Sum	250	1.0	250	1.0	250	1.0
Mean	.340		0.256		0476	
Variance	0.31365		0.35990		0.94922	

TABLE II

The Kolmogorov-Smirnov goodness of fit test (D) for fitting the Poissong distribution to the number of SEF damaged tillers per hill in three plant growth stages, ITA 306, 1990 WS.

Number of SEF damaged tillers	14 DAT		42 DAT		70 DAT	
(x)	F(x)	S(x)	F(x)	S(x)	F(x)	S(x)
0	0.712	0.704	0.774	0.816	.0621	.0752
1	0.953	0.956	0.972	0.940	0.917	0.920
2	1.0	1.0	0.996	0.988	0.998	0.980
3			1.0	1.0	0.999	0.988
4					1.0	1.0
Estimate of Parameter (m)	0.340		0.256		0.476	
D-value	0.008 NS		0.042 NS		0.131 **	

NS: Not significant **: Highly significant

$$F(x) = \bar{e}\sum_{i=0}^{mx} m^{i}/_{i!} \quad S(x) = \sum_{i=0}^{x} f_i/_{250} \quad D = \max|F(x) - S(x)|$$

TABLE III

Values of the mean square error (MSE) in the analysis of variance at fixed values of λ and c.

	c					
λ	.25	.50	.75	1	2	3
-4.0	729.90	2.92	.111	.0106	.0000317	.00000103
-3.0	83.65	1.25	.103	.0170	.00159	.0000127
-2.0	11.19	.82	.106	.0302	.00121	.0000166
-1.0	2.08	.39	.136	.0621	.00824	.00232
-.5	1.08	.35	.168	.0964	.0226	.00897
0	.68	.35	.227	.163	.0647	.0357
.5	.55	.42	.349	.299	.1963	.1478
1	.641	.641	.641	.641	.641	.641
2	3.08	3.74	1.48	5.29	9.355	14.7

The underlined values are the smallest at a given value of c.

TABLE IV

Estimates of the power corresponding to the maximum of log likelihood function L at fixed values of c.

c=.25		c=.50		c=.75	
λ	*L*	λ	*L*	λ	*L*
-2.50	1192.6	-3.25	1035.2	-4.0	958.6
-2.25	1202.5	-3.0	1040.5	-3.75	961.9
-2.0	1205.8	-2.75	1042.7	-3.50	962.9
-1.75	1201.8	-2.50	1038.2	-3.25	961.9
-1.50	1188.4	-2.25	913.3	-3.0	957.7

Underlined figures are estimated λ ($\hat{\lambda}$) and the corresponding to maximum log likelihood function L(max).

TABLE V

Values of the normality statistic W based on residuals at fixed values of λ and c.

	c						
λ	.25	.50	.75	1	2	3	4
-2.0	.609	.612	.617	.621	.631	.645	.653
-1.0	.615	.622	.627	.631	.651	.658	.659
-.5	.624	.630	.641	.647	.659	.661	.661
-.25	.629	.642	.650	.655	.661	.662	.661
0	.643	.653	.657	.660	.662	.661	.661
.25	.656	.661	.662	.663	.660	.657	.656
.50	.663	.663	.661	.660	.655	.656	.656
1.0	.633	.633	.633	.633	.633	.633	.633
1.5	.525	.544	.557	.566	.588	.599	.606

All computed values of W are highly significant.

TABLE VI

Values of the skewness coefficient (g1) based on residual analysis at fixed values of λ and c

	c					
λ	.25	.50	.75	1	2	3
-2.0	1.14	1.15	1.16	1.19	1.30	1.42
-1.0	1.16	1.20	1.24	1.30	1.48	163
-.5	1.21	1.28	1.36	1.43	1.64	1.81
0	1.36	1.48	1.58	1.64	1.90	2.06
.5	1.82	1.95	2.05	2.13	2.33	2.46
1	3.07	3.07	3.07	3.07	3.07	3.07
2	9.35	8.61	7.99	7.48	6.11	5.35

All values of g1 are significantly different from 0.

TABLE VII

Values of the kurtosis coefficient (g2) based on residual analysis at fixed values of λ and c

	c					
λ	.25	.50	.75	1	2	3
-2.0	-.636	-.605	-.545	-.459	.016	.552
-1.0	-.567	-.410	-.211	.003	.880	1.683
-.5	-.373	-.046	-.287	.614	1.785	2.776
-.25	-.131	.322	.747	1.143	2.488	3.561
0	.361	.943	1.473	1.946	3.469	4.611
.5	3.06	3.982	4.685	5.260	6.910	7.993
1	14.35	14.35	14.35	14.35	14.35	14.35
2	120.96	105.89	93.79	84.01	59.40	46.74

All values of g2 are significantly different from 0 except those which are underlined

TABLE VIII

Values of X^2 statistic (Bartlett) for testing equality of variances of fixed values of λ and c.

	c					
λ	.25	.50	.75	1	2	3
-4.0	16.36	15.97	15.19	14.10	8.74	4.27
-3.0	16.13	15.06	13.48	11.68	5.13	1.42
-2.0	14.99	12.36	9.62	7.16	1.27	.034
-1.0	10.00	5.58	2.77	1.12	.53	3.44
0	.0224	.778	2.52	4.51	11.92	17.61
.25	1.69	4.77	7.69	10.33	18.38	23.77
.50	9.25	13.55	16.85	19.53	26.83	31.26

$$X^2_{(2\ d.f)} = \frac{2.3026}{M}\left[\sum_{i=1}^{3}(n_i - 1)\log_{10} s^2 - \sum_{i=1}^{3}(n_i - 1)\log_{10} s_i^2\right]$$

$$M = 1 = \frac{1}{6}\left[\sum_{i=1}^{3}\frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{3} n_i - 1}\right]$$

All values of X^2 are significant ($p<.05$) except those which are underlined.

TABLE IX

Values of the F-statistic in the analysis of variance at fixed values of λ and c.

	c					
λ	.25	.50	.75	1	2	3
-4.0	3.96	3.93	3.88	3.81	3.55	3.44
-3.0	3.94	3.87	3.78	3.69	3.46	3.45
-2.0	3.88	3.73	3.61	3.53	3.47	3.63
-1.0	3.66	3.52	3.48	3.50	3.78	4.12
-.5	3.53	3.51	3.59	3.69	4.16	4.55
0	3.64	3.83	4.03	4.21	4.76	5.12
.5	4.55	4.84	5.05	5.21	5.67	5.85
1	6.68	6.68	6.68	6.68	6.68	6.68
2	9.54	9.38	9.48	9.08	8.59	8.24

All values of F are significant with 2 and 747 degrees of freedom [F(0.05) = 3.00].

REFERENCES

Barlev, S.K. (1988). On the classical choice of variance stabilizing transformations and application for a Poisson variate. **Biometrika 75**(4), pp. 803-804.

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

Box, G.E. and Cox, D.R. (1964). An analysis of transformations. **Journal of the Royal Statistical Society B-26**, pp. 211-243.

Carrol, J.B. and Ruppert, D. (1984). Power transformations when fitting theoretical models to data. **Journal of American Statistical Association 79**(386), pp. 321 - 328.

Cochran, W.G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. **Biometrics 3**(1), pp. 22-37.

De Ramos, M. (1983). The comparison between the arcsine and the Box-Cox transformations in normalizing percentage data. **The Philippine Statistician 32**(3-4), pp. 19-30.

Dolby, J.L. (1963). A quick method for choosing a transformation. **Technometrics 5**(3), pp. 317-325.

Draper, N.R. and Hunter, W.G. (1969). Transformations: Some examples revisited. **Technometrics 11**(1), pp. 23-41.

Guerrero, V.M. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika* **69**(2), pp. 309-314.

Harcourt, D.G. (1963). Population dynamics of Leptinotarsa decemlineata (Say) in Eastern Ontario. I. Spatial pattern and transformation of field counts. **The Canadian Entomologist 95**, pp. 813-820.

Hinkley, D.V. (1988). More on score tests for transformation in regression **Biometrika** 75(2), pp. 366-369.

Logothetis, N. (1990). Box-Cox transformations and the Taguchi method. **Applied Statistician** 39(1), pp. 31-48.

Tsai, C. (1988). Power transformations and reparameterizations in non-linear models. **Technometrics** 30(4), pp. 441-448.