Statistical Models for Yield, Pest Abundance and Disease Incidence of Tomato

Juma William Yabeja

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in Research Methods in the Jomo Kenyatta University of Agriculture and Technology

2013

DECLARATION

I, **Juma William Yabeja** declare that this dissertation is my original work and has not been submitted for a degree in any other university. All sources of materials used in this dissertation have been fully acknowledged.

Signature …….. Date …….

Juma William Yabeja

This dissertation has been submitted for examination with our approval as University supervisors.

1. Signature …… Date …….

Dr Elijah M Ateka

JKUAT, Kenya

2. Signature…….. Date ……

Dr Joseph C Ndunguru

MARI, Tanzania

3. Signature …… Date ……..

Dr Daisy Salifu

ICIPE, Kenya

DEDICATION

This dissertation is dedicated to my mother Sele Muyuga, my late dad William Yabeja, my lovely wife Happyness G. Yabeja, my son William Yabeja, our family supervisor Dr Joseph Ndunguru, crew from JKUAT, ARI-Mikocheni and IITA-Tanzania: thanks for the endless help and encouragements during the hard time I faced in my studies and normal life. Finally thanks to almighty God who continue to give me life.

AKNOWLEDGEMENTS

I am grateful to RUFORUM for offering me a full scholarship for my MSc Studies in Research Methods and to my supervisors Dr Elijah M. Ateka (JKUAT), Dr Joseph Ndunguru (ARI-MIKOCHENI) and Dr Daisy Salifu (ICIPE) for their constructive criticism, continuous encouragement and support during the whole process of writing up of this dissertation.

I would like to thank Dr Peter Sseruwagi for his an endless help in refining the objectives and other parts of the dissertation. My deepest thanks extend to my loved wife Happyness Gabriel Yabeja for her support and encouragements throughout my MSc studies and to all ARI-Mikocheni staff for who in one way or another contributed to accomplishment of this dissertation.

I would like to thank Dr Chris Ojiewo from AVRDC, Arusha for supplying tomato seeds for my experiment and other supports, Dr James Legg from IITA-TANZANIA for his appreciated support for internet services and other assistances on my dissertation and last but not least I would like to thank the crew of lecturers in MSc Research Methods program at Jomo Kenyatta University of Agriculture and technology, Nairobi, Kenya who helped me to acquire Research Methods skills.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS/ACRONYMS

AIC	Akaike Information Criteria
ANOVA	Analysis of variance
ARI	Agricultural Research Institute
AVRDC	Asian Vegetable Research and Development Center
GLM	Generalized linear models
ICIPE	International Center of Insect Physiology and Ecology
IITA	International Institute of Tropical Agriculture
JKUAT	Jomo Kenyatta University of Agriculture and Technology
LM	General linear models
TYLCD	Tomato Yellow Leaf Curl Disease

ABSTRACT

Analysis of variance (ANOVA) is one of the general linear models used to nearly unimaginable range of problems in many different disciplines and has been a fundamental method used by plant pathologists and other researchers for analysis of continuous data, disease incidence and insect pest abundance data. However, disease incidence and pest abundance data usually violate the assumptions of ANOVA because they are discrete data. Most researchers often transform the data using arcsine for disease incidence, square root for pest abundance and other forms of transformation although most researchers finally do not check if the transformation was effective to correct for the violated assumptions. Hence, the objectives of this dissertation is to (1) to determine the performance of ANOVA on continuous data (tomato fruit weight) including the validity of statistical inferences, (2) to assess the performance of logistic regression and ANOVA on tomato yellow leaf curl disease incidence including the validity of statistical inferences, (3) to evaluate the performance of Poisson regression and ANOVA on pest abundance including the validity of statistical inferences. Tomato fruit weight data was analyzed assuming only normal distribution while tomato yellow leaf curl disease incidence data were analyzed assuming normal (ANOVA), binomial distribution (logistic regression). Whitefly population data were analyzed assuming normal (ANOVA), Poisson, and negative binomial error distribution. On the basis of multiple R Square (higher value) and small residual standard error close to zero, ANOVA model on tomato fruit weight confirmed better goodness of fit to the data. The greater p-value for deviance (p=0.1207) and Pearson (0.0896) statistics showed that logistic regression model performed better compared to ANOVA on tomato yellow leaf curl disease. Also the greater p-value for deviance (0.0077) and Pearson (0.2796) statistics, decreased AIC

value (475.22 to 300.11) indicated that negative binomial model was most appropriate compared to Poisson regression models. It was concluded that GLMs could be alternative models for discrete data.

CHAPTER ONE

1.0 INTRODUCTION

1.1 Background information

Currently, the field of plant pathology and its statistical applications continue to develop, providing chances for the use of statistics in the biological sciences and new demands for statistical attitudes in plant pathology. It is very familiar in plant pathology to estimate the relationship between disease responses to a number of environmental and other explainable variables (Sanogo and Yang, 2004). Despite the fact that common statistical methods such as ANOVA which fall under parametric tests being very well known and convenient, their assumptions are not every time met in contexts studied by plant pathologists and other biologists.

ANOVA, as one of the general linear models, has been applied to an almost unimaginable range of problems in many different disciplines. Has been a fundamental method used by plant pathologists and other biologists for analysis of disease incidence and pest abundance data. However, naturally disease incidence and pest abundance data often violates the assumptions of ANOVA of homogeneity of variance and normal distribution because they are discrete data (Madden and Hughes, 1995; Garrett *et al*., 2004). Statistical inference from pest counts data poses a number of challenges. For example in ecological count datasets (Fletcher *et al.,* 2005; Martub *et al*., 2005; Warton, 2005), pest counts frequently unveil two features: skewed distribution and large proportion of the values being zero. Also insect pest counts data reveal heterogeneity of variances among observational groups or populations (Taylor, 1961).

ANOVA throughout its application, a continuous distribution with a normal distribution has been assumed for the response variables (Y), and a linear model is fitted to the data to determine the coefficients using an ordinary least square methodology which minimizes the sum of squared distances of data points to the parameter estimates (Schabenberger and Pierce, 2002). The equations employed are known as general linear models. Many dependent variables of interest to plant pathologist are discrete data, such as disease incidence (number or proportion or percentage of diseased individuals in a total sampled plant population) or count of lesions or spores (Madden and Hughes, 1995; McRoberts *et al*., 2003).

In the application of analysis of variance (ANOVA), a standard method that has been used is the transformation of dependent variable (Y) which effect in approximated variable with a normal distribution. However, most researchers do not check if transformation was effective to correct the problem of normality and equal variance. In a real logic, this is forcing the data to fit a model that was developed for analysis of continuous data, rather than using an appropriate statistical method for the data at hand (Hughes and Madden, 1995; Madden *et al*., 2002). Additionally, variance-stabilizing transformations could not fully stabilize variances in counts data (McArdle and Anderson, 2004) or incidence data when some of the means are nearby to 0 or 100% (Madden, 2002). It is very familiar that departures from the assumption of homogeneity may result in inflated error rates (Cochran, 1947). Test of standard errors, significance and differences of the means may be affected if ANOVA is used for discrete data.

Despite the transformation of the dependent variable (Y) to meet the assumptions, ANOVA through arcsine transformation is still not an effective statistical tool for analyzing discrete data such as disease incidence and pest abundance for the

following weakness; the equalization of variance in proportional data when using arcsine transformation requires the number of trials to be equal for each data point, while the effectiveness of arcsine transformation in normalizing proportional data depends on sample size (n), and does not perform well at extreme ends of the distribution (Worton and Hui, 2010; Hardy, 2002). Another argument against arsine transformation is that it does not confine with proportional data between 0 and 1, resulting in the extrapolation of proportional values that are not biologically sensible.

Statistical methods which are unpopular in plant pathology institutional research at present, but that have potential for improving analysis of disease incidence and pest abundance are generalized linear models (logistic, Poisson regression and negative binomial models). These are usually a better alternative where one cannot assume the models for continuous data, they are appropriate for discrete data (Collett, 2002).

Fitting generalized linear models to the data broadly is implemented using maximum like-lihood, a method based on finding parameter estimates that result in the highest probability of observing the actual data obtained. When generalized linear model is used it is straightforward to account for the properties of data from discrete distributions such as the binomial and Poisson which are appropriate theoretical distributions to consider for proportions and counts respectively (Agresti, 2002; Collett, 2002). In the application of GLMs, one usually chooses the logit-link function for proportion data and the log-link function for counts. These link functions are used as natural transformation of the data because the residuals will not be normally distributed and cannot be constant across values of predictors (Turechek, 2004). For example in binomial model, dependent variable (Y) has only two possible values 0 and 1, present or absent where the residual has only two possible values for each predictor (X).With only two possible values, the residual cannot be normally

distributed. For proportions, the analysis is commonly known as logistic regression while for counts the analysis is commonly known as Poisson regression. The approach on these models is useful for designed experiment as well as observation (survey) studies where one is relating qualitative factors (e.g. fungicide treatment and cultivar) and quantitative factors (e.g. soil temperature) to responses (De Wolf *et al*., 2003; Hughes *et al*., 1998). Furthermore, GLM-based analysis of lesion counts and disease incidence from observation (survey) studies can be of direct benefit in developing efficient sampling protocols for either estimating mean disease levels or testing hypothesis about mean level (Hughes and Gottwald, 1998; Hughes and Gottwald, 1999; Madden and Hughes, 1999).

1.2 Problem statement and justification

Knowledge on the most appropriate statistical method based on model evaluation and goodness of fit test used in analysis of epidemiological data, is a vital aspect in drawing valid statistical inferences. These epidemiological studies provide useful information for understanding the ecology and biology of the pest. This information normally is used by plant pathologists and entomologists as the basis for the planning, establishment and monitoring of effective disease and pest management strategies.

When statistical assumptions are violated and an inappropriate statistical method used to analyze data, it systematically over-or under-estimate coefficients, results to larger standard error and finally resulting to inaccurate statistical insignificance. In epidemiological studies, commonly determined parameters relating diseases with biotic and abiotic factors are the analysis of disease incidence and pest abundance. Many publications on plant disease measurements (incidence, severity) and vector counts have used ANOVA in quantifying diseases.

An alternative model to disease incidence data apart from ANOVA therefore is the logistic regression, an analytical method that is designed to deal with proportional or percentage data (Steel and Torrie, 1997). Logistic regression allows for binomially distributed proportional data, unlike arcsine transformation that attempts to stabilize variance while the data may remain non-normal (Worton and Hui, 2010). The logit link function used in logistic regression provides a more biologically relevant analysis, where the proportional data never falls outside of 0 and 1 (Worton and Hui, 2010).This link also can deal with unbalanced data, whereas the arcsine transformation can only effectively equalize variance if proportional data points have an equal number of trials (Jaeger, 2008; Worton and Hui, 2010). In addition, logistic regression produces easily interpretable and biologically relevant coefficients, unlike arcsine transformation (Worton and Hui, 2010). On the other hand, Poisson regression model is the basic model for insect abundance or counts data, apart from ANOVA (McCullagh and Nelder, 1989).

For stakeholders (researchers and others) to develop rational and economical control measures, whether by use of pesticides or breeding resistant cultivars, it is not sufficient to state that disease causes crop losses; the magnitude of disease needs to be clearly and precisely quantified using appropriate statistical methods. This study compares the use of ANOVA and logistic regression on tomato yellow leaf curl disease incidence, ANOVA and Poisson regression on pest abundance data.

1.3 OBJECTIVES

1.3.1 General objective.

To provide alternative statistical models for analyzing disease incidence and pest abundance data based on model evaluation and goodness of fit test as the criteria for testing appropriateness of the data analysis model.

1.3.2 Specific objectives.

- To determine the performance of ANOVA on continuous data (tomato fruit weight) including the validity of statistical inference;
- To assess the performance of logistic regression and ANOVA on tomato yellow leaf curl disease incidence including the validity of statistical inference;
- To evaluate the performance of Poisson regression and ANOVA on pest abundance including the validity of statistical inference.

CHAPTER TWO

2.0 LITERATURE REVIEW

2.1 The concept of disease incidence

Disease incidence refers to the number of plant entities which are visually diseased out of the total number of plant units measured (Campbell and Madden, 1990). Another definition of disease incidence is the proportion (0 to 1) or percentage (0 to 100) of diseased entities within a sampling unit (Seem, 1984). The key factor to these and other related definition on disease incidence (Chellemi *et al*., 1988; Kranz, 1988; Nutter *et al*., 1993) is that incidence data are binary. Referring to Seem's (1984) terminology, plant could take only one of the two possible forms either a plant is diseased or it is not.

In plant epidemiological studies, disease incidence can be determined from fruits, tillers, leaves, flowers or seeds. Therefore whenever one wants to determine and record the disease status of individual observations, disease incidence can be calculated. Plant pathologist and entomologist frequently collect disease and pest incidence data since in many instances, especially with plant diseases caused by viruses, it is impractical to assess disease on the basis of pathogen abundance (McRoberts *et al*., 1996). Similarly with small arthropods such as aphids, thrips, mites, psyllidds and leafhoppers, presence or absence is often easier to establish than estimating abundance by counting individuals.

2.2 Distribution of Disease Incidence

Naturally, disease incidence is not normally distributed (Madden and Hughes, 1995; Garrett *et al*., 2004). Disease incidence is a binary variable because each observed

individual plant is either visibly affected or not, or damage symptoms are present or absent (Madden, 2002). Therefore it is characterized by a binomial outcome or beta binomial (Madden and Hughes, 1995; Collett, 2002). Regardless of many advantages of using the binomial distribution (Collett, 2002), this distribution only occasionally describes actual disease incidence data. Diseased individual characteristically are clustered in nature, resulting in a greater heterogeneity of disease incidence than would be expected for a random pattern (Madden, 2002). More typically the variance is larger more skewed than that expected by the binomial distribution (Hughes and Madden, 1995).

Disease incidence assesses the probability (π) of a plant or other plant entity being diseased. This probability is clearly a function of the pathogen, host and environment. The probability of plant not being diseased is ($1- \pi$). This probability comprises the two states of the Bernoulli distribution for describing the probability of individual observations taking on one of the two classes (e.g. diseased or healthy).

2.3 Disease Assessment.

Disease assessment or phytopathometry usually involves the measurement and quantification of plant disease. Therefore it is a primary significance in the study and analysis of plant disease epidemics (Nutter *et al*., 2006), also distinguished disease assessment and phytopathometry, the former being as the process of quantitatively measuring disease intensity and the second as the theory and practice of quantitative disease assessment. It is very important to have accurate disease assessment methods as identified early by (Chester, 1950; Kranz, 1988) who stated that without quantification of disease no studies in epidemiology, no assessment of crop losses and plant disease surveys and their applications would be possible. Also the idea was advanced by Lucas (1998) that disease assessment includes a number of

interconnected activities, such as the future progress of the disease, disease diagnosis, forecasting and crop loss. Strange (2003), mentioned that the measurement of plant disease and its effects on crop yield, quality and value are vital for control strategies.

Prior to modeling the change in disease intensity (dy) with change in time (dt), it is required first to obtain accurate and precise measurements of disease intensity (Nutter *et al*., 1991). Disease intensity is a general term for the amount of disease (injury) present in a host population (Nutter *et al*., 1991), while the most common types of disease intensity measures are prevalence, severity and disease incidence.

Prevalence is a term that is often used interchangeably with incidence but strictly defined as the number of fields within a specific geographical area (Country, state, or region) where a disease has been visually observed (symptoms) divided by the number of fields sampled and assessed (Campbell and Madden, 1990; Nutter *et al*., 1991; Zadoks and Schein, 1976). Normally, getting information concerning pathogen prevalence, individual plants or plant parts are sampled from a host population (fields) and these are tested for the presence of the disease such as virus using a reliable methods of indexing (infectivity, Enzyme-linked immunosorbent assay, Polymerase Chain Reaction and presence of inclusion bodies. (Kapa and Waterworth, 1981). Therefore, pathogen prevalence may include quantitative information concerning the presence of a disease in asymptomatic as well as symptomatic fields. Prevalence data are frequently multiplied by 100 to give the percentage of fields in which a disease or pathogen is present. Obviously, prevalence data does not give quantitative information about the relative amount of disease within the individual fields sampled. Within-field disease intensity measurements can be obtained by assessing plant population within fields for disease incidence or disease severity.

2.4 The inadequate of ANOVA over categorical dependent variables.

The problem with ANOVA and more commonly general linear models over binary or categorical outcome has been known for a long time (Rao, 1960; Winer *et al*., 1971; Cochran, 1940). ANOVA compares the means of different experimental or observational sample groups and determine whether to reject the null hypothesis that the groups have the same population means given the observed sample variances within and between the sample groups based on continuous scale (normal distributed) and application of ANOVA to discrete data compromise the results.

2.5 Concept and rationale of data transformation.

Data transformation is used by researchers to generate new variables from existing variables according to the mathematical functions. It has been used in statistical procedures as a vital instrument to serve many purposes including improving normality of a distribution, equalizing variance to meet assumptions and improve effect sizes, thus constituting important aspects of data cleaning and preparing for statistical analyses. Some commonly transformation tools includes: logarithmic, square root, arcsine, adding constants and trigonometric (Box and Cox, 1964).

Researchers have used transformation technique routinely as data cleaning before data analysis (Box and Cox, 1964; Sakia, 1992) provides a family of transformations which approximately normalize a particular variable, eliminating the need to randomly try different transformation to determine the best option. Box and Cox (1964) originally envisioned this transformation as a panacea for simultaneously correcting normality, linearity and homoscedasticity. These transformations often improve all of these aspects of a distribution or analysis, (Sakia, 1992) and others have noted that it does not always accomplish these challenging goals.

2.6 Data Transformation tools used for proportion and counts data

2.6.1 Arcsine transformation

This transformation has customarily been used for proportions, (which range from 0.00 to 1.00), and involves taking the arcsine of the square root of a number, with the resulting transformed data reported in degree of radians. This transformation is of the form

$$Y = \text{arcsine}(p) = \sin^{-1} p,$$

where *p* is the proportion and *Y* is the output of the transformation.

Due to the mathematical properties of this transformation, the variable must be transformed to the range of -1.00 to 1.00. Despite the fact that it is believed to be perfectly valid transformation technique the use of arcsine also known as inverse transformation (Rao, 1998) or angular transformation (Snedecor and Cochran, 1989), has been open for debate as to the usefulness in analysis of proportional data that tends to be skewed when the distribution is not normal.

Even if arcsine transformation is useful tool in stabilizing variances and normalizing proportional data, there are a number of reasons why this method can be problematic. The equalization of variance in proportional data when using arcsine transformations needs the number of trials to be equal for each data point, while the efficacy of arcsine transformation in normalizing proportional data is dependent on sample size (n), and does not perform well at extreme ends of the distribution (Worton and Hui, 2010; Hardy, 2002). An additional argument against arcsine transformation is that it does not confine proportional data between 0 and 1, resulting in the extrapolation of proportional values that are not biologically sensible (Hardy, 2002).

2.6.2 Square root transformation

It is useful for count data (data that follow a Poisson distribution) and more appropriate for data consisting of small whole numbers. If most of the values in the data set are less than 10, especially if zeros are present, the transformation to use is $(Y+0.5)^{½}$ instead of $Y^{1/2}$. In this, the square root of every value is taken. However, as one cannot take the square root of a zero number, a constant must be added to move the minimum value of the distribution above 0, preferably to 1.00 (Osborne, 2002). It reflect the fact that numbers above 0.00 and below 1.0 behave differently than numbers 0.00, 1.00 and those larger than 1.00. The square root of 1.00 and 0.00 remain 1.00 and 0.00 respectively, while numbers above 1.00 always become smaller, and numbers between 0.00 and 1.00 become larger (the square root of 4 is 2, but the square root of 0.40 is 0.63). Therefore if square root transformation is used to continuous variable that contains values between 0 and 1 as well as above 1, you are treating some numbers differently than others which may not be desirable.

2.7.0 Models, Normality and equal variance test Reviews

2.7.1 Logistic Regression Model for Binary Data.

Logistic regression is the statistical model which observes the influence of different factors on categorical rather than continuous outcome. The model estimates the probability of an event of binary outcome (Menard, 1995). The fundamental mathematical concept that underlies logistic regression is the logit-the natural logarithm of an odds ratio. It is part of categorical statistical models called generalized linear models. This broad class of models includes ordinary regression and ANOVA, as well as multivariate statistics such as analysis of covariance and log- linear regression (Agresti, 1996). The predicted value of the dependent variable is a probability. Logistic regression currently has increased popularity as a modern

statistical technique used to model the probability of discrete (i.e. binary or multinomial) outcomes. When correctly applied, logistic regression analyses yield very powerful insights to the attributes (i.e. variables) which are more or less likely to predict event outcome in a population of interest. These models also explain the extent to which changes in the values of the attributes may increase or decrease the predicted probability of event outcome. Generally in logistic regression the dependent or response variable is dichotomous, such as presence/absence or success/failure. Logistic regression model has been used as an alternative to ANOVA through arcsine transformation that is becoming more prevalent in today's biological data (Steel and Torrie, 1997). When logistic regression model is implemented, the test of significance of the model coefficients are performed most commonly with the Wald χ^2 statistic (Menard, 1995) which is based on the change in the likelihood function when an independent variable is added to the model. The Wald χ^2 statistics serves the same role as the t or F test of ordinary least square partial regression coefficients. There are numerous likelihood function statistics also available to assess goodness of fit (Cox and Snell, 1989). The logistic model (Agresti, 1996) has the following form.

$$\text{Logit } (Y) = \text{In } \left(\frac{\pi}{1-\pi} \right) = \alpha + \beta_1 x_1 \ldots + \varepsilon$$

where, π is the probability of event, α is the *Y* intercept, β is regression coefficient and *X* is a predictor. The logit function is the link function for the binomial distribution. β_i stands for the estimated coefficients and X_i stands for predictors in the model.

2.7.2 Poisson Regression model for counts data

Poisson regression analysis is a tool which allows modeling of dependent variables that are counts (Cameron and Trivedi, 1998; Kleinbaum *et al.*, 1998). It is usually applied to study the occurrence of small number of counts or events such as whitefly

counts as a function of a set of independent variables like cultivar and altitude, in an experiment and observational study in many fields, including Biology and Medicine (Gardner *et al*., 1995). This model is based upon the generalized Poisson distribution which has been comprehensively studied by researchers. In some disciplines the model has been used to model a household fertility data set (Wang and Famoye, 1997) and to model injury data (Wulu *et al*., 2002). Counts data with too many zeros are frequent in a number of applications. (Ridout *et al*., 1998) cited examples of data with too many zeros from diverse disciplines including agriculture and species abundance. Poisson regression models for count data assume an equality of variance and mean for each observation. This assumption breaks down if over-dispersion is present in the counts data. Therefore we discuss another model which accommodates data whose variance larger than the mean. Poisson regression model (Agresti, 1996) summarized as

$$\text{Log}_e(\mu) = \beta o+\beta_1 x_1+\beta_2 x_2+ \beta_n x_n +\varepsilon$$

The variance in Poisson model is equal to the mean

$$\text{Var}(y) =\mu$$

The Poisson regression model is modeled as

$$E(Y) =\mu=\exp(\beta o+\beta_1 x_1+\beta_2 x_2+ \beta_n x_n \ldots+\varepsilon)$$

2.7.3 Negative binomial Model (NBM)

Negative binomial distribution is the mixture of Poisson distribution in which the expected values of the Poisson distribution differ according to a gamma (type III) distribution (Johnson and Kotz, 1969). This support one of the four derivations of the negative binomial model (Anscombe, 1949). So far it has revealed that the limiting distribution of the NBD as the dispersion parameter (k) approaches zero, is the Poisson. Once k is an integer, the NBD turn into the Paschal distribution, and the

geometric distribution corresponds to k=1. The log series distribution arises when zeros are missing and such as k→∞ (Saha and Paul, 2005). The negative binomial statistic (Agresti, 1996) is given as;

$$f(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{k}{\mu+k}\right)^k \left(1-\frac{k}{\mu+k}\right)^k \quad y= 0,1, 2$$

where *k* and μ are parameters. The variance of negative binomial is as follows;

$$\text{Var } (y) = \mu + (\mu^2/k)$$

where *k* range from zero to infinity, μ is the mean.

2.7.4 Shapiro-Wilk test

Shapiro-Wilk test checks the normal assumption by creating W statistic, which is the ratio of the best estimator of the variance (based on the square of a linear combination of the order statistics) to the usual corrected sum of squares estimator of the variance, where $0 < Wn \leq 1$ and $7 \leq n \leq 2{,}000$. Shapiro-Wilk statistic (Shapiro and Wilk, 1965) is given as

$$W_n = \frac{(\sum_{i-1}^{n} a_i X_{(i)})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Where $x_{(i)}$ is the i^{th} largest order statistic, $\bar{x}$ is the sample mean and n is the number of observations.

The hypotheses used are:

H_o: Sample data has a normal distribution

H_1: Sample data does not have a normal distribution

Shapiro -Wilk test conclusion is that;

• P-value > 0.05 means the sample data is normally distributed.

• P-value < 0.05 means the sample data is not normally distributed.

2.7.5 Anderson-Darling test

Anderson-Darling Statistic was developed by Anderson and Darling in 1954. It is based on empirical distribution function. Its test statistic is called statistic which is the square of the difference of histogram width and area width below the normal curve. The Anderson-Darling statistic (Anderson and Darling in 1954) is calculate as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n} [(2i-1)log\ U_{(i)} + (2n+1-2i)\log\left(1-U_{(i)}\right)]$$

The Anderson-Darling hypotheses test:

H_0: Data is normally distributed.

H_1: Data is not normally distributed.

Anderson-Darling test interpretation;

• P-value > 0.05 means the data is normally distributed.

• P-value < 0.05 means the data is not normally distributed.

2.7.6 Skewness test

The skewness value offers a sign of either departure or no departure from symmetry in a given distribution. A data set is symmetric if the median divides the left side and the right side into two identical areas. Skewness is measured with the following equation (Kenney & Keeping 1962): The Skewness statistics is written as;

$$\frac{\sum_{i=1}^{N}(X-\bar{x})^3}{(N-1)S^3}$$

where, $\bar{x}$ is the mean, *N* is the number of data points and *s* is the standard deviation.

A symmetric distribution which is an indication of normally distributed data has a skewness value of zero. Negative values show data that are left skewed and positive values show data that are right skewed.

2.7.7 Kurtosis test

Kurtosis is a measure of whether the data are peaked or flat relative to a normal distribution. This means that data sets with high kurtosis incline to have a distinct peak near the mean, decline rather rapidly and have heavy tails. Data sets with low kurtosis incline to have a flat top near the mean, rather than a sharp peak. Kurtosis is measured with the following equation (Miles & Shevlin 2001): The Kurtosis statistics is written as;

$$\frac{\sum_{i=1}^{N}(X-\bar{x})^4}{(N-1)S^4}$$

Where $\bar{x}$ is the mean, *N* is the number of data points and *s* is the standard deviation.

The kurtosis for a standard normal distribution has a value of zero. If the distribution is perfectly normal, skewness and kurtosis values of zero will be obtained. Positive kurtosis shows a leptokurtic distribution. The word 'leptokurtic' is derived from the Greek word 'leptos', meaning small or slender. Negative kurtosis shows a platykurtic distribution. The term 'platykurtic' is derived from the French word 'plat', meaning flat (Miles & Shevlin 2001).

2.7.8 Histogram plot

Histogram is the most widely used graphical methods, which is the simplest and possibly the oldest method which divides the range of data into classes and plot bars equivalent to each bin or class. The height of each bar reveals the number of data points present in the equivalent bin. The method summarizes the data distribution into shape, standard deviation, skewness, kurtosis and presence of the extreme values (outliers). For a normal distributed variable the histogram display a mean of 0 and standard deviation of two. Also display a symmetric, single-peaked and bell- shaped (Armitage and Colton, 1998).

2.7.9. Box plot

Box plot or "box and whiskers" depict an excellent summary of distribution of data based on minimum, first quartile (25^{th} percentile), median (50^{th} percentile), third quartile (75^{th} percentile) and maximum of the data (Tukey, 1977). If data comes from a normally distributed population the 25^{th} percentile and 75^{th} percentile become symmetry, median and mean are located at the middle point. If the percentiles are not symmetry, it implies the possibility of skewed distribution.

2.8.0 Normal quantile quantile plot (Q-Q PLOT)

The normal Q-Q plot is a plot of the expected normal values against the corresponding of observed data. It aid to show how distribution of data could be normal or deviate from a normal distribution. A normal distributed data show a Q-Q plot with actual data points align along the straight line which originate from the right angle with a positive slope.

2.8.1 Levene's test

Levene's test is used to test if samples have equal variance. The equal variance across samples is so called homogeneity of variance. Common statistical models such as analysis of variance assume that variance is equal across groups or samples. Therefore, the Levene's test is used to verify this assumption prior to the use of analysis of variance model (Levenes, 1960).

2.8.2 Fligner-Killen test

This test used for testing homogeneity of variance of a sample or population (Fligner and Killeen, 1976). It is based on ranking the absolute values and assigning the increase scores. It is test among other widely used test for conformity of equal variance of groups or samples.

2.9.0 Tomato Yellow Leaf Curl Disease

Tomato (*Lycopersicon esculentum* Mill.) is one of the main vegetable crops grown in Tanzania. It is highly popular due to its high nutritive value, taste and multipurpose use in various food items such as salad as well as processed products like tomato sauce, pickle, ketchup, puree, dehydrated and whole peeled tomatoes. It is a good source of vitamins (A and C) and minerals (Hobson and Davies, 1971; Kalloo, 1991). In Tanzania, tomato is a major fruit vegetable crop that is cultivated by commercial and small scale farmers and used for both fresh and processing industries. Despite it is importance, it is seriously attacked by a number of diseases caused by fungi, bacteria and viruses. Among the viral diseases, whitefly-transmitted germiniviruses (WTGs) such as tomato yellow leaf curl diseases are of significant constraint whenever the crop is grown as they causes poor fruit yield and quality. The amount of losses due to these viruses often reaches 100% (Green and Kalloo, 1994). Losses vary depending on virus strain, the variety of the tomato, the age of the plant at infection time, temperature during disease development, presence of other diseases, and the extent that virus has spread in the plant.

CHAPTER THREE

3.0 MATERIALS AND METHODS

3.1.1 Study area

The field experiment was conducted in experimental station at Chambezi, Bagamoyo in Tanzania located on latitude 6 º 32´5´ longitudes 38º58´ and 34m above sea level.

3.1.2 Experimental design

The experimental was laid out as a completely randomized design. Seven tomato cultivars CNL3070J, CNL3078G, CNL3125P, CNL3125E, CNL3125L, Tanya and Tengeru from Asian Vegetable Research and Development Center in Arusha Tanzania were used as treatments in this experiment. The treatments were replicated three times in plots each measuring (6m by 1.2m). Plants were spaced within rows at 0.6m and 0.5m between rows. Each plot had two rows each with six plants which gives a total of twelve plants in a plot.

April 14, 2012 seeds were sowed in trays. They were left until the seedlings have three to four full expanded leaves. One week before transplanting manure was applied one the plots. Fifteen days old seedlings were transplanted in the field in May 2012.

3.1.3 Data collection

Two harvests were considered, first fruit harvest was on August 10, 2012 and the second harvest was on August 31, 2012. However, only marketable fruits weight was measured. Since each plot consisted of twelve plants, only six plants three from each row within a plot were considered for fruit weight measurements. All the seven cultivars were harvested and finally total fruit weights per cultivar in each of the three replications were measured using beam balance.

3.2 Field Survey

A field survey was carried out to obtain tomato yellow leaf curl disease incidence and whitefly abundance data to address second and third objectives.

3.2.1 Tomato Yellow Leaf Curl Disease

A survey on tomato yellow leaf curl disease was conducted in June, 2012 in Arusha as one of the major tomato growing areas in Tanzania. The districts surveyed included Arumeru East, Arumeru West and part of Arusha Township. Fields with a 2 to 4 month-old tomato crop were sampled on transect along rural roads at approximately 3 to 4km interval. An X-shaped transects stretching between opposing corners of each field was used. A total of 40 tomato plants were visually assessed in each field where the symptomatically plants were recorded as positive (+) and finally counted to get a total number of infected plants among the 40 assessed plants to determine the TYLCD incidence. Disease incidence was calculated as follows.

Disease incidence = $(\frac{n}{N})$ x100%

where *n* is the number of plants affected by disease and *N* is the total number of plants assessed.

3.2.3 Whitefly counts

Adult whitefly population was determined by counting the number of whiteflies on the five top most expanded leaves of a representative shoot on the 40 tomato plants randomly selected along diagonals of each field (Sseruwagi *et al*., 2004). This was used to determine the whitefly populations/counts.

3.4 Statistical analysis of the data

For the tomato fruit weight data the test for normality and homogeneity of variance were carried out through statistical tests; Shapiro-Wilk statistics, Anderson-darling

test, kurtosis, skewness and graphic methods using histogram, box plot, q-q plots. The same tests were also performed on tomato yellow leaf curl disease and whitefly counts before and after data transformation. The assumption of equality of variance on data was tested using statistical test; Levene's test and fligner test.

For analysis of tomato fruit weight data, only ANOVA was fitted to the data with the assumption that fruit weight as a continuous data has homogeneous variance and approximate log-normal distribution (Perry *et al*., 2003). Upon statistical test ANOVA showed to meet the assumptions (Table 1). ANOVA as one of the general linear model applied ordinary least square method. Hence, ANOVA was an appropriate statistical tool to analyze tomato fruit weight data.). Model evaluation performed using coefficient of determination (R^2) and residual standard errors. LSD was used to compare variety means after the ANOVA null hypothesis of equal means was rejected using the ANOVA F-test.

For Analysis of disease incidence data, the general linear model (ANOVA) and binomial model were fitted to the same data. Since the assumption of normal distribution in disease incidence data was not met, now the data is taken care of by assuming a binomial distribution for the data, and by using maximum likelihood methods to estimate the p-value and the parameters of the model. The general linear/normal distribution model applied ordinary least squares (OLS) on arcsine transformed TYLCD incidence. In this study arcsine transformation was used due to its popularity among many plant pathologists, breeders and other researchers. The probabilistic model using OLS assumes that the underlying errors of the transformed data are uncorrelated with homogeneous variance, and hence follow an approximate log-normal distribution (McArdle and Anderson, 2004; Warton, 2005). Binomial

distribution model (logistic regression) applied maximum-likelihood method with a natural logit transformation.

Logistic regression analysis was used to examine the effects of two treatments cultivar and altitude on tomato yellow leaf curl disease incidence. The probability of success (TYLCD infection) was modeled as a function of both cultivars and altitude. The regression coefficients were interpreted as the rate of change in the odds ratio of TYLCD per unit change in the altitude .In this logistic regression model, the proportion of success (infected plants), p, was modeled and logit transformation assumed to be a linear combination function of the cultivars and altitude.

$$\text{logit}(Y) = \text{In}\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 \ldots + \varepsilon$$

where, logit is the Link function for the binomial distribution and used as a natural transformation for categorical data. For TYLCD incidence, β_0 represents intercept, $\beta 1$, β_2 are the maximum likelihood estimates of the logistic regression coefficients and the X_1, X_2 stands for cultivar and altitude respectively, The TYLCD incidence was analyzed by assuming binomial distribution of diseased individual.

The effect of cultivar and altitude on the number of whitefly counts was tested using a linear model, (ANOVA), Poisson regression and negative binomial model. The general linear/normal distribution model applied ordinary least squares (OLS) on square root transformed whitefly counts. The probabilistic model assumed the underlying errors of the transformed whitefly counts data are all uncorrelated with homogeneous variance, and an approximate log-normal distribution (Perry *et al*., 2003).

Poisson distribution was considered as it arises under the assumption that insect pest are distributed randomly in space and the variance equals to the mean. However,

insect pest count data usually exhibit over dispersion, with a variance larger than the mean (Taylor, 1961). A Poisson regression model relating the whitefly population mean (μ) to the explanatory variables cultivars and altitude took a form of

$$\log_e(\mu) = \alpha + \beta_1 x_1 + \ldots + \varepsilon$$

where b_0 is an intercept and b_i is a parameter coefficients to be estimated for the $_i^{th}$ covariate and X_i stands for cultivar and altitude

When Poisson regression model which applied Maximum-likelihood approach using log link as a link function revealed over dispersion, negative binomial model was considered because one significant characteristic of the NBD is that it naturally accounts for over dispersion because its variance is often greater that the variance of a Poisson distribution with the same mean. The NBD can be derived from the Poisson distribution when the mean parameter is not identical for all members of the population, but itself is distributed with gamma distribution.

The Goodness-of –fit was based on Pearson and deviance statistics while model selection was based on Akaike Information Criteria (AIC). All these analyses were performed in R version 2.15.0.

CHAPTER FOUR

4.0 RESULTS AND DISCUSSION

This chapter displays the statistical and graphic methods used for normality and equal variance test of all data in this study, data analysis, interpretation and finally discussion of the findings from this study in association with the other existing literatures.

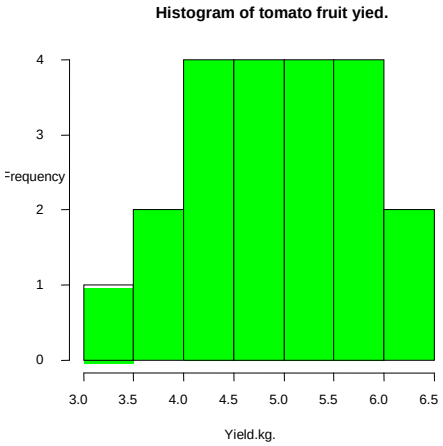
Table 4.1: Normality and equal variances test on tomato fruit weight data

Statistical test	p-value	Calculated value
Shapiro-Wilk test	0.6501	
Anderson-Darling test	0.7767	
Skewness		-0.2321
Kurtosis		-1.1201
EQUAL VARIANCE O TEST		
Levene's test	0.7803	
Fligner test	0.6155	

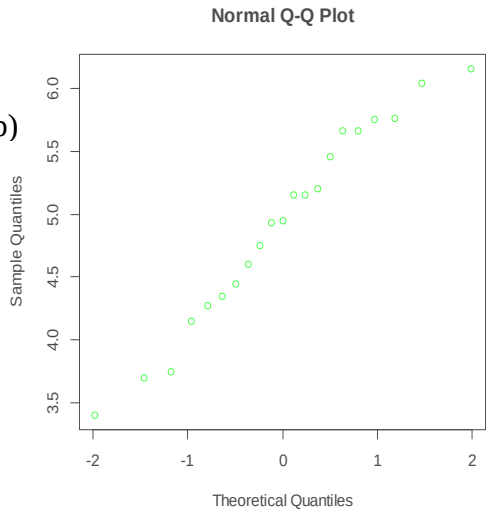
From (Table 4.1) above, Shapiro-Wilk test (p=0.6501) as well as Anderson-Darling test (p=0.7767) did not reject the null hypothesis that the fruit weight variable was normally distributed. The skewness (-0.2321) and kurtosis (-1.1201) also indicated an

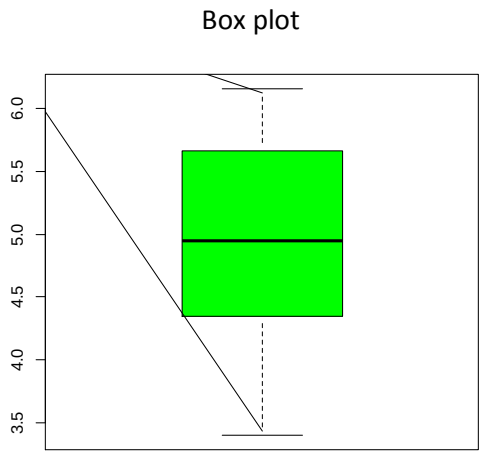
almost normal distribution. Levene's test (0.7803) as well as Fligner test (0.6155) also did not reject the nully hypothesis that the fruit weight variable had equal variance. Therefore, we concluded that fruit weight variable was normally distributed and had equal variance.

(a)



(b)





(c)

Figure 4.1: Graphic methods for normality test of tomato fruit weight

From (Fig 4.1) the histogram plot above, box plot with a median line divided the two quartiles equally (symmetry) and normal probability plots (q-q) with an almost

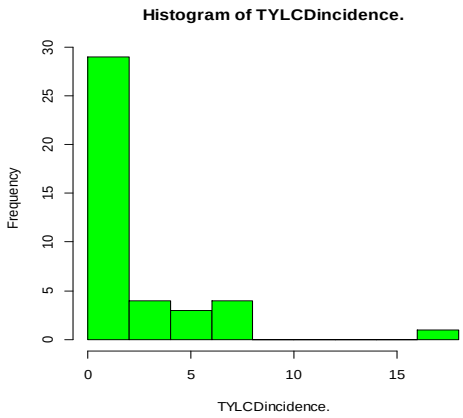
straight line from the right angle all indicated that the tomato fruit weight variable was normally distributed. Therefore, there was no evidence that the raw tomato fruit weight deviated from the normal distribution and equal variances.

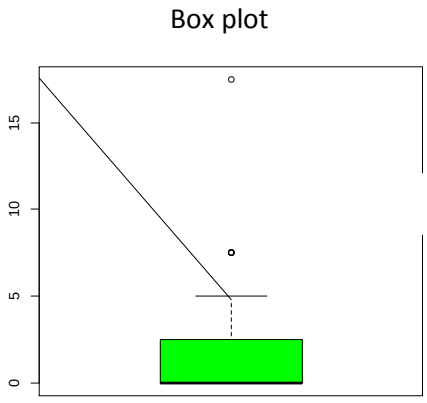
Table 4.2: Normality and equal variance test of Tomato Yellow Leaf Curl disease incidence before and after transformation of the data using arcsine function

	TYLCD incidence before transformation		TYLCD incidence after transformation	
Statistical test	p-value	Calc value	p-value	Calc value
Shapiro-Wilk test	9.118e-10		1.86e-09	
Anderson-Darling test	2.2e-16		2.2e-16	
Skewness		2.5097		0.9010
Kurtosis		7.2055		-1.1858
EQUAL VARIANCE TEST				
Levene's test	0.9319		0.8992	
Fligner test	0.7113		0.7495	

Shapiro-Wilk test (p=9.118e-10), Anderson-Darling test (p=2.2e-16), Skewness (2.5097) and Kurtosis (7.2055) in raw data as well as arcsine transformed data where Shapiro-Wilk test (p=1.86e-09) and Anderson-Darling test (p=2.2e-16) all rejected the null hypothesis that the raw and arcsine transformed incidence data were normally distributed. Therefore, from this (Table 4.2) we concluded that arcsine transformation did not ensure normality, hence the data remained non- normal.

(a)





(c)

(b)

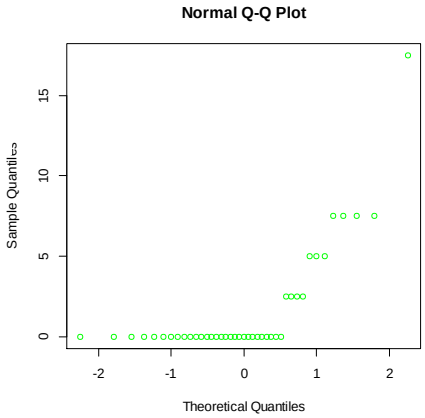


Figure 4.2: Normality test on tomato yellow leaf curl disease incidence data before arcsine transformation.

From the diagnostic plots (Fig 4.2), the histogram plot, box plot and normal probability plots (q-q) all indicated that the raw tomato yellow leaf curl disease incidence was not normally distributed. Most observations are highly concentrated on the left side of the distribution (a) while in box plot the data is right skewed(c). The observations in (b) produced s-shape which is the deviation from the line of fit.

Therefore, based on the statistical and graphic methods above, it was concluded that raw tomato yellow leaf curl disease incidence was neither normally distributed nor equal variance.

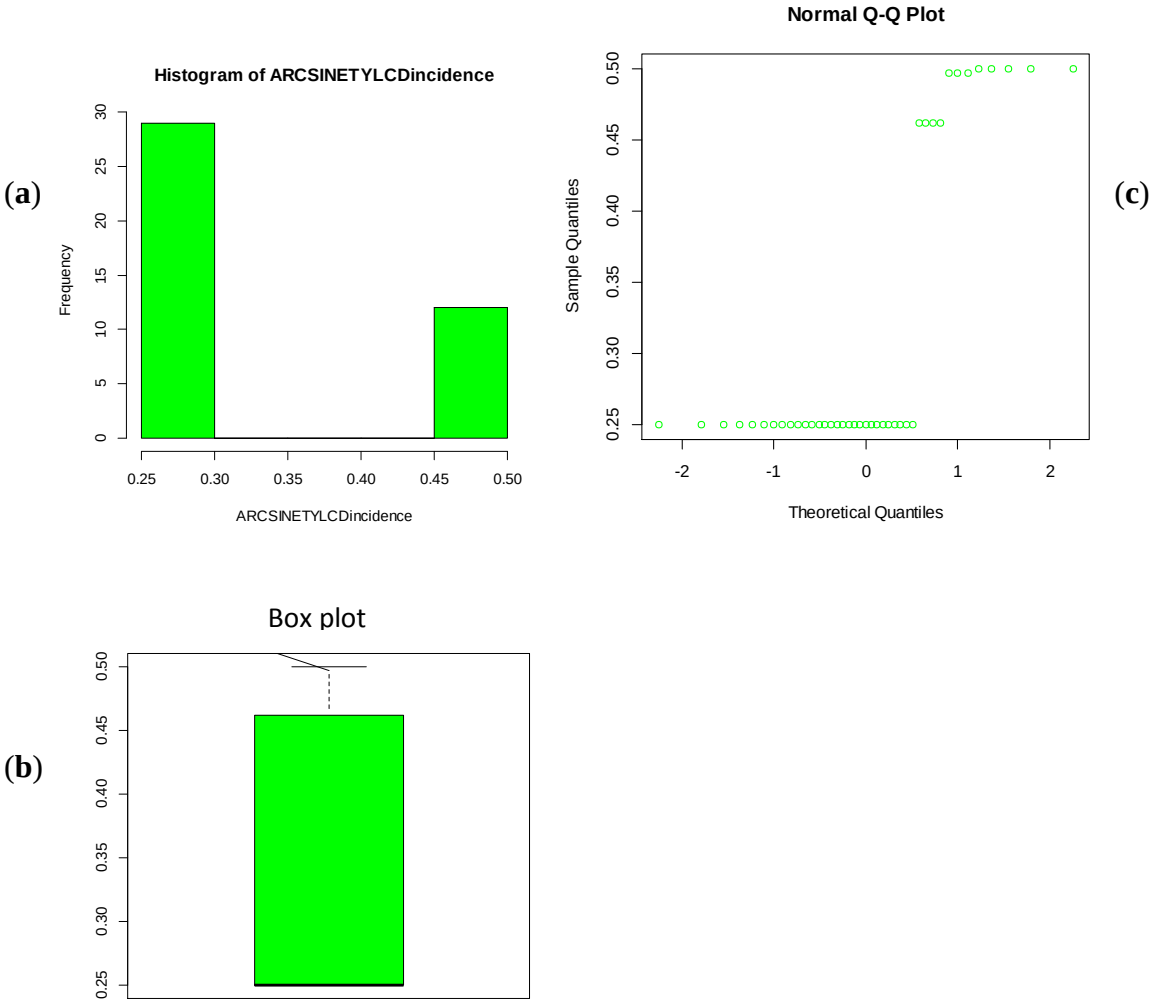


Figure 4.3: Normality test on tomato yellow leaf curl disease incidence data after arcsine transformation.

In (Fig 4.3) histogram plot, box plot, and normal probability plots (q-q) all showed that the arcsine transformed tomato yellow leaf curl disease incidence was not normally distributed. Therefore, based on the statistical test (Table 4.2) and graphic methods (Fig 4.3) all proved that arcsine transformation did not ensure normality. So

it was concluded that raw and arcsine transformed tomato yellow leaf curl disease incidence was not normally distributed.

	Whitefly counts before transformation		Whitefly counts after transformation	
Statistical test	p-value	Calc value	p-value	Calc value
Shapiro-Wilk test	5.71e-07		0.0066	
Anderson-Darling test	5.47e-07		0.0339	
Skewness		2.177926		0.7056
Kurtosis		5.614029		0.1459
EQUAL VARIANCE O TEST				
Levene's test	0.4386		0.5064	
Fligner test	0.1933		0.3818	

Table 4.3: Normality and equal variance test of whitefly population before and after transformation using square root function

From (Table 4.3) above, Shapiro-Wilk test (p=5.71e-07), Anderson-Darling test (p=5.47e-07), Skewness (2.177926) and Kurtosis (5.614029) in raw whitefly counts data as well as square root transformed whitefly counts data where Shapiro-Wilk test (p=0.0066) and Anderson-Darling test (p=0.0339) all rejected the null hypothesis that the raw whitefly counts and square root transformed data were normally distributed. Therefore it was concluded that square root transformation did not ensure normality in whitefly counts.

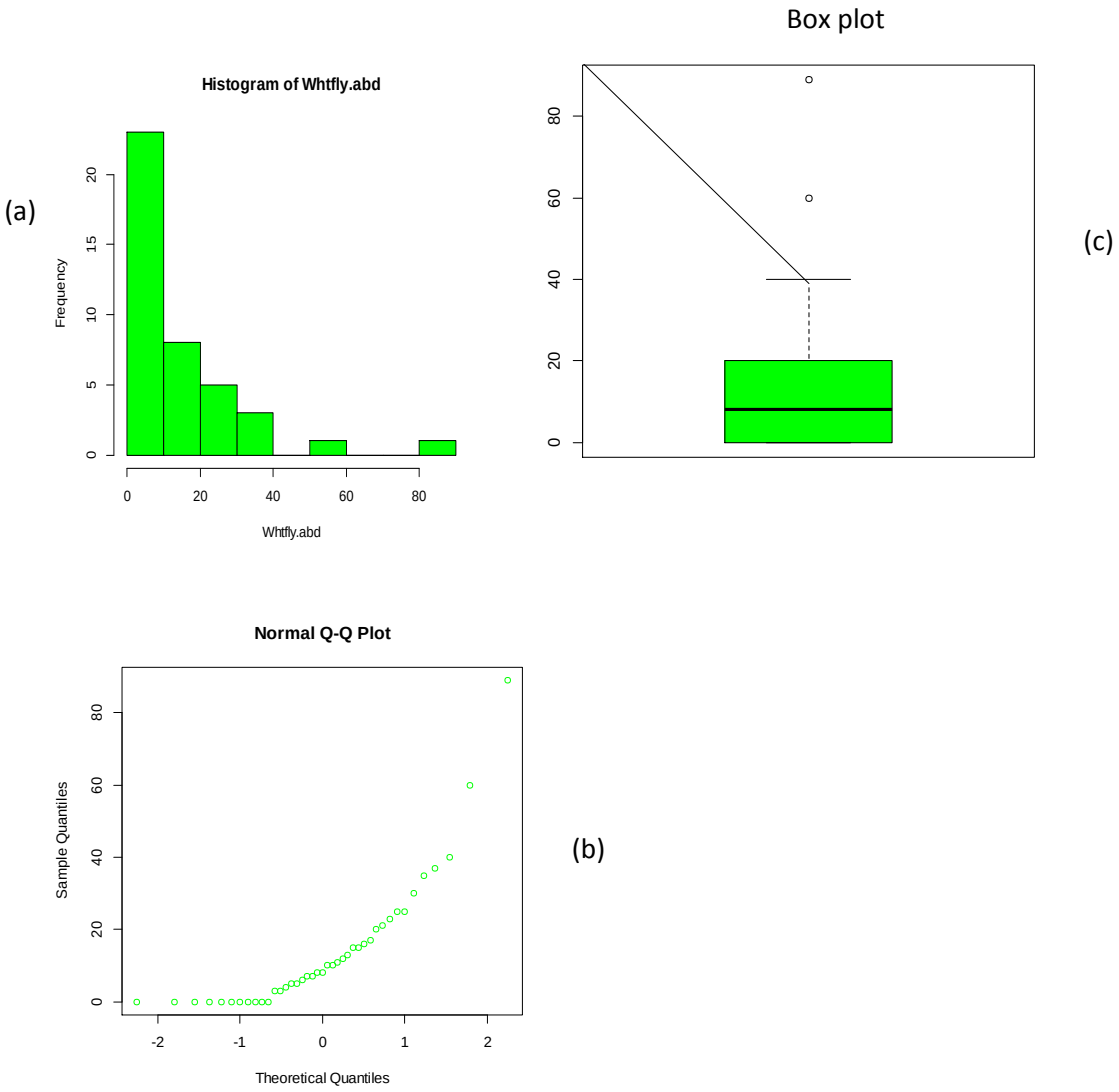


Figure 4.4: Normality test of whitefly abundance before transformation using square root function.

The histogram plot, box plot, and normal probability plots (q-q) all indicated that the raw whitefly counts data was not normally distributed (Fig 4.4). Most observations are highly concentrated on the left side of the distribution (a) and right skewed in box plot (c). The observations in (b) produced s-shape which is the deviation from the line of fit. Therefore, based on the statistical test (Table 4.3) and graphic methods (Fig 4.4) above, was concluded that raw whitefly counts data was not normally distributed.

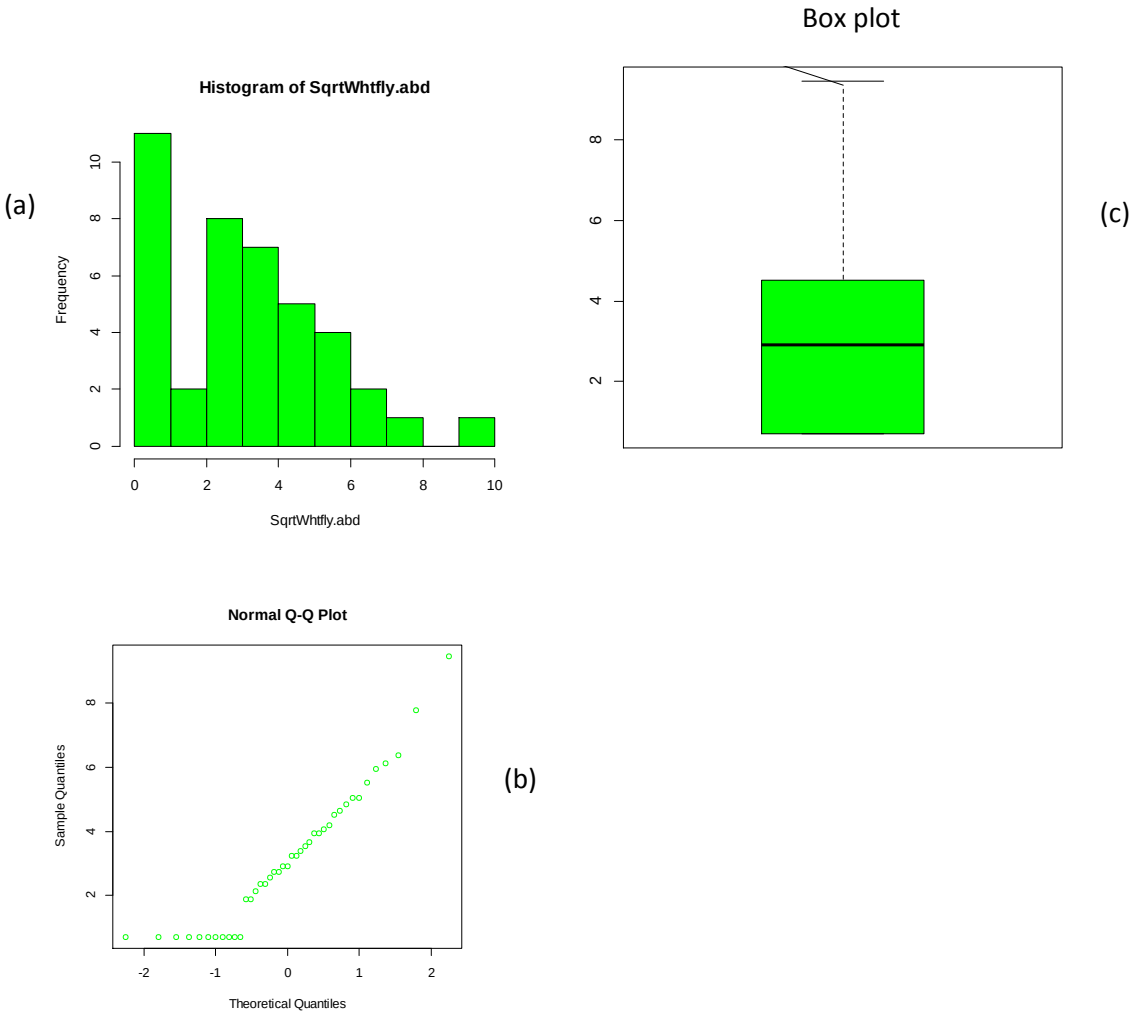


Figure 4.5: Normality test of whitefly abundance after transformation using square root function

In (Fig 4.5) the histogram plot, box plot, and normal probability plots (q-q) all indicated that the square root transformed whitefly counts data was not normally distributed. Most observations are highly concentrated on the left side of the distribution (a) while in the box plot data is left skewed(c). The observations in (b) produced s-shape which is the deviation from the line of fit. Therefore, based on the statistical test (Table 4.3) and graphic methods (Fig 4.5) above it was concluded that raw and square root transformed whitefly counts data was not normally distributed.

Table 4.4: Estimated p-values of Analysis of variance on tomato fruit weight

PREDICTOR	p-value	Calculated value
Cultivar	0.03789 *	
Goodness of fit test		
Multiple R-squared:		0.5708
Residual standard error:		0.622

Analysis of variance (ANOVA) was performed to tomato fruit weight data as indicated in (Table 4.4).There was a significant different of cultivar (p=0.0379). The goodness of fit test multiple R squared (57%) being larger and residual standard error (0.622) being small closer to zero indicated that AVONA was an appropriate model to analyze this data.

Table 4.5: Means in (kg) for the tomato fruit weight

Treatments	Means
TANYA	5.83a
CNL3078G-01009	5.39ab
CNL3070J-010010	5.35abc
CNL3125P-01005	4.95abcd
CNL3125L-01003	4.44bcd
TENGERU	4.28cd
CNL3125E-01002	4.19d
LSD	1.09

Means with the same letter are not significantly different.

Table 5 above provides the post hoc analysis of the cultivar where Tanya had the highest fruit weight yield compared to the rest tomato cultivar.

Table 4.6: Estimated the p-values of Analysis of variance before and after Arcsine Transformation of TYLCD incidence.

Predictors	p-value before Transformation	p-value arcsine- Transformation
Cultivars	0.00197 **	0.1447
Altitude	0.0777	0.0957
GOODNESS OF FIT TEST		
Adjusted R-squared	0.4723	0.1906
Residual standard error	2.576	0.0984

In Table 4.6 analysis of variance (ANOVA) was performed to reveal the effect of cultivar and altitude in two forms of data. In raw (untransformed) data cultivar indicated statistical significant different (p=0.002) on the influence of tomato yellow leaf curl disease incidence as main effect while it showed that there was no significant difference (p=0.096) on arcsine transformed tomato yellow leaf curl disease incidence data. On the other hand, altitude had no significant effect (p=0.078) on untransformed and transformed (p=0.098) incidence data. After analysis, goodness of fit of the model was determined by examining the coefficient of determination (R^2), which is the proportion of the variation in the disease incidence accounted for by cultivar and altitude. Residual standard error was also examined in both untransformed and transformed data. Lower value of Adjusted R-squared on arcsine transformed TYLCD disease incidence data indicated poor fit.

Table 4.7: Estimated parameter values of the logistic Regression model on Tomato Yellow Leaf Curl Disease incidence

Cultivar	β	Std Error	z value(χ^2)	p-value	odds ratio
Intercept	7.827e+00	3.815e+00	2.052	0.04020 *	2.51e+03
Honex	-2.143e+01	3.120e+03	-0.007	0.99452	4.93e-10
Maglobu	-2.064e+01	7.062e+03	-0.003	0.99767	1.09e-09
Mandeli	-2.170e+01	4.987e+03	-0.004	0.99653	3.76e-10
Meru	5.033e-01	1.523e+00	0.331	0.74096	1.65e+00
Mshumaa	-2.457e+00	8.402e-01	-2.925	0.00345 **	8.57e-02
Sadiki	-2.146e+01	7.062e+03	-0.003	0.99758	4.78e-10
Tanya	-3.507e+00	7.292e-01	-4.809	1.51e-06***	3.00e-02
Tengeru	-2.827e+00	1.132e+00	-2.497	0.01253 *	5.92e-02
Tengeru 97	-3.683e+00	1.266e+00	-2.910	0.00362 **	2.52e-02
Top harvest	-2.093e+01	7.062e+03	-0.003	0.99764	8.13e-10
Unknown	-3.430e+00	8.242e-01	-4.161	3.17e-05 ***	3.24e-02
Wonex	-3.030e+00	1.066e+00	-2.841	0.00449 **	4.83e-02
Altitude	-8.306e-03	3.359e-03	-2.473	0.01340 *	9.92e-01

Null deviance: 84.128 on 40 degrees of freedom

Residual deviance: 35.758 on 27 degrees of freedom

AIC: 94.618

Table 4.8: Analysis of Deviance of logistic regression model

	Df	Deviance Resid.	Df	Resid. Dev	Pr(>Chi)
NULL			40	84.128	
Cultivar	12	39.618	28	44.511	8.321e-05 ***
Altitude	1	8.753	27	35.758	0.003092 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

GOODNESS OF FIT TEST	p-value
Deviance	0.1207
Pearson	0.0896

Logistic regression model was fitted to the TYLCD incidence data by the maximum likelihood method to the explanatory variables cultivar and altitude. The results (Table 4.8) showed that the two independent variables cultivar and altitude were significant different (χ^2=39.62, p=0.001; χ^2=8.75, p=0.003 respectively. Deviance and Pearson statistics are both types of residuals where "The larger p-value the better fit of the model to the data".

Evaluation of logistic regression model was performed using Deviance and Pearson statistics (Chi-square distribution). The insignificant results (p=0.1207) for Deviance and (p=0.0896) for Pearson at a 0.05 significant level indicated that the model fitted well the data (Barrett, 2007; Hosmer and Lemeshow, 2000; McCullagh and Nelder, 1989).

Table 4.9: Estimated p-values from analysis of variance on whitefly abundance data

Predictors	p-value before Transformation	p-value of Square root Transformation
Cultivar	2.014e-05 ***	0.005461 **
Altitude	0.7517	0.7994
GOODNESS OF FIT TEST		
Adjusted R-squared	0.6296	0.3932
Residual standard error	10.97	1.67

When analysis of variance performed on whitefly abundance, raw data indicated cultivar were significantly different (p=2.014e-05) while altitude was non-significant different (p=0.7517; Table 7). Residual standard error being larger than zero (Table 4.9) indicated poor goodness of fit. When whitefly counts data were transformed using square root technique, residual standard error continued to be greater than zero which also indicated poor goodness of fit.

Table 4.10: Estimated parameter values of Poisson Regression Model on whitefly counts

Cultivar	β	Std Error	z- value(χ^2)	p-value
(Intercept)	5.290e+00	6.550e-01	8.076	6.70e-16 ***
Honex	-3.009e+00	2.370e-01	-12.699	< 2e-16 ***
Maglobu	-1.985e+00	3.079e-01	-6.445	1.15e-10 ***
Mandeli	-1.210e+00	1.792e-01	-6.754	1.43e-11 ***
Meru	-3.166e+00	6.141e-01	-5.156	2.52e-07 ***
Mshumaa	-2.503e+00	2.801e-01	-8.936	< 2e-16 ***
Sadiki	-4.454e-01	1.721e-01	-2.589	0.00963 **
Tanya	-1.798e+00	1.503e-01	-11.963	< 2e-16 ***
Tengeru	-1.985e+01	1.276e+03	-0.016	0.99
Tengeru 97	-1.993e+01	1.276e+03	-0.016	0.99
Top harvest	-1.786e+00	2.791e-01	-6.399	1.56e-10 ***
Unknown	-2.406e+00	1.586e-01	-15.175	< 2e-16 ***
Wonex	-2.231e+00	3.389e-01	-6.581	4.66e-11 ***
Altitude	-7.097e-04	5.725e-04	-1.240	0.21511
Null deviance: 763.72 on 40 degrees of freedom				
Residual deviance: 313.40 on 27 degrees of freedom				
AIC: 475.22				

Table 4.11: Estimated parameter values of negative binomial model on whitefly abundance

Cultivar	β	Std Error	z- value(χ^2)	p-value	odds ratio
(Intercept)	5.569e+00	2.325e+00	2.395	0.01662 *	2.62e+02
Honex	-3.064e+00	1.176e+00	-2.606	0.00916 **	4.67e-02
Maglobu	-1.978e+00	1.511e+00	-1.309	0.19047	1.38e-01
Mandeli	-1.230e+00	1.305e+00	-0.943	0.34569	2.92e-01
Meru	-3.089e+00	1.693e+00	-1.825	0.06804	4.56e-02
Mshumaa	-2.510e+00	1.312e+00	-1.912	0.05583	8.13e-02
Sadiki	-4.632e-01	1.494e+00	-0.310	0.75647	6.30e-01
Tanya	-1.842e+00	1.123e+00	-1.641	0.10082	1.59e-01
Tengeru	-3.536e+01	2.946e+06	0.000	0.99999	4.39e-16
Tengeru 97	-3.546e+01	2.946e+06	0.000	0.99999	3.98e-16
Top harvest	-1.788e+00	1.505e+00	-1.189	0.23460	1.67e-01
Unknown	-2.356e+00	1.116e+00	-2.111	0.03477 *	9.48e-02
Wonex	-2.279e+00	1.555e+00	-1.466	0.14276	1.02e-01
Altitude	-9.570e-04	1.837e-03	-0.521	0.60245	9.99e-01

Null deviance: 77.575 on 40 degrees of freedom

Residual deviance: 47.985 on 27 degrees of freedom

AIC: 300.11

Table 4.12: Analysis of Deviance of negative binomial model

	Df	Deviance Resid.	Df	Resid. Dev	Pr(>Chi)
NULL			40	77.575	
Cultivar	12	29.4200	28	48.155	0.003 **
Altitude	1	0.1695	27	47.985	0.68

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

GOODNESS OF FIT TEST	**p-value**
Deviance	0.0077
Pearson	0.2796

Poisson regression model was fitted to the whitefly abundance where over-dispersion was observed (Table 4.10). Due to over-dispersion, the negative binomial model was performed on the same data. According to the analysis of deviance (Table 4.12), there was a significant difference between cultivar (χ^2=29.4200, p=0.003412) in a number of whitefly abundance accumulation while there was non-significant different of altitude (χ^2= 0.1695, p= 0.680516) in influencing whitefly abundance accumulation. Akaike Information Criteria (AIC), Deviance and Pearson statistics were performed to evaluation the negative binomial model. Pearson (0.2796) and Deviance (0.0077) with insignificant results indicated that the model fitted well to the data (Barrett, 2007; Hosmer and Lemeshow, 2000; McCullagh and Nelder, 1989). When AIC from Poisson regression model was compared with AIC from negative binomial model "the model with the lowest AIC is being the best model". AIC value

of Poisson model of 475.22 to 300.11 of negative binomial indicated that negative binomial model was the best model compared to Poisson regression.

4.2 DISCUSSION

Raw tomato fruit weight data never deviated from the normal distribution and equal variance (Table 4.1; Fig 4.1). Despite of data transformation, raw and arcsine transformed tomato yellow leaf curl disease incidence data were non-normal (Table 4.2; Fig 4.2; Fig 4.3). Also raw and square root transformed whitefly abundance data were not normally distributed (Table 4.3; Fig 4.4; Figure 4.5). This is in conformity with many growing literatures that transformation tools like arcsine transformation, does not always ensure normality (Martub *et al*., 2005; Warton, 2005; Fletcher *et al*., 2005; McArdle and Anderson, 2004). Still if the approximate normality is indicated by the statistics like p-value and graphs on the transformed data, if the data come from some supplementary distribution than normal then the significant test expected to give over-or under- estimated coefficients, larger standard errors and biased p-value (Menard,1995). This statement is supported by the results provided at arcsine transformed column (Table: 4.6) where cultivar (p=0.1447) and altitude (p=0.0957) had non-significant different results compared to logistic regression model (Table 4.8) with cultivar (p=8.321e-05) which is significant different. The same variable which is cultivar has two different p-values when subjected to two different models. Therefore, the results obtained from normality and equal variance test using tomato fruit weight data revealed that continuous data continue to adhere to the assumptions of normality and equal variance (Perry *et al*., 2003) while arcsine and square root transformation does not necessary each time ensure normality and equal variance.

ANOVA (Table 4.4) was fitted to the tomato fruit weight data, there was significant different in cultivar (p=0.0379). The Model evaluation (goodness of fit test) indicated that the model fitted well the tomato fruit weight data based on Multiple R square (higher value the better model) and residual standard errors (small closer to zero is the

better model). These findings are in agreements with that ANOVA is the best statistical tools for analysis of continuous data which normally holds the assumptions of normally distribution and homogeneity of variance (Perry *et al*., 2003).

Analysis of variance (Table 4.6) was performed to reveal the effect of cultivar and altitude in two forms of data. In raw (untransformed) data cultivar indicated statistical significant different (p=0.0019) on the influence of tomato yellow leaf curl disease incidence as main effect while it showed that there was no significant difference (p=0.1447) on non-normal arcsine transformed tomato yellow leaf curl disease incidence data. On the other hand, altitude had no significant effect (p=0.4723) on untransformed and transformed incidence data (0.0957). After analysis, goodness of fit of the model was determined by examining the coefficient of determination (R^2), which is the proportion of the variation in the disease incidence accounted for by cultivar and altitude. Residual standard error was also examined in both untransformed and transformed data. Lower value of Adjusted R-squared on arcsine transformed TYLCD disease incidence data indicated poor fit.

When logistic regression model was used to the same tomato yellow leaf curl disease incidence data, cultivar (p=8.321e-05) and altitude (p=0.0031) in (Table 4.8) were significant different in influencing tomato yellow leaf curl disease incidence which is in disagreement with the ANOVA results of the same arcsine transformed data which showed cultivar (p=0.1447) and altitude (p=0.0957) being insignificant. Goodness of fit tests Deviance (p=0.1207) and Pearson (p=0.0896) statistics indicated that logistic regression model was the best and appropriate statistical tool in modeling tomato yellow leaf curl disease incidence.

When ANOVA was used on raw and square root transformed whitefly abundance data, cultivar was significantly different (p=2.014e-05) and (p=0.0055) respectively

with residual standard error greater than one. This indicated that the ANOVA was not an appropriate statistical method to analyze whitefly abundance data due to heterogeneity of variance and non-normal of many counts data (Taylor, 1961). Clearly, the study also has established that square root transformation of whitefly abundance data do not necessary ensure normality. In many cases, the transformation applied to normalize the data may lead to heterogeneity of variance. This is for the reason that one transformation might be best for ensuring homogeneity of variance, while another might be best for ensuring normality. In this circumstance statistical requirement cannot be met with linear models (Garrett, Madden, Hughes and Pfender, 2004).

Poisson regression model (Table 4.10) was used on the whitefly abundance data to test the effect of cultivar and altitude on influencing the accumulation of whitefly abundance. Over dispersion, implying variance being larger than the mean or variance exceeds the theoretical variance (Mullay, 1997) was obtained in Poisson regression model, solving this negative binomial model (Table 4.11) was used as suggested by some researchers as an alternative to the Poisson when there is evidence of over-dispersion (Paternoster and Brame, 1997; Osgood, 2000). Only cultivar (Table 4.12) was significantly different (p=0.003). Akaike Information Criteria was used for model selection between Poisson regression and negative binomial model because the use of AIC provides a consistent result and is independent of the order in which the models are computed (Anderson *et al.,* 2000; Burnham and Anderson, 2002). Reduction in Akaike Information Criteria (AIC) from (475.22 for Poisson to 300.11for negative binomial model) and the results of goodness of fit test Deviance and Pearson (0.2796) statistics, all suggested that negative binomial model was better and a more convenient model for modeling whitefly abundance with over-dispersion (McRoberts

et al., 1996; Anscombe, 1949; Sileshi *et al* (a)., 2006; Sileshi *et al* (b)., 2006) compared to the Poisson regression model. Information criteria such as AIC offer a more objective way of defining which model amongst a set of models is the best appropriate for analysis of the data which available at hand to researchers (Akaike, 1973; Yang, 2007).

Generalized linear models allow an appropriate analysis of skewed frequency or binary data. Furthermore with GLMs, the properties of data from discrete distributions such as binomial distribution, Poisson and negative binomial can be accounted for (Hughes and Madden, 1995; Collett, 2002). For Incidence and counts data generalized linear models present tremendous opportunities for improvement of statistical inference.

CHAPTER FIVE

5.0 CONCLUSION AND RECOMMENDATIONS

5.1 General Conclusion

ANOVA was fitted into three different data sets, tomato fruit weight, tomato yellow leaf curl disease incidence and pest abundance. Then it was compared to logistic regression for tomato yellow leaf curl disease incidence and Poisson regression for pest abundance. ANOVA was an appropriate model and fitted well the fruit weight data because the data indicated normally distribution. For arcsine and square root transformed data, ANOVA showed to be inappropriate model and did not fit well the data because the transformation did not ensure normality of the data. Poisson regression model for raw pest abundance was inappropriate models due to poor goodness of fit and over-dispersion. Solving the over-dispersion issue, a negative binomial model was applied for the pest abundance which revealed more sensitive analysis compared to standard ANOVA and Poisson regression model for analysis of pest abundance.

5.2 General Recommendation

Application of generalized linear models could improve statistical inference of disease incidence and pest abundance when compared to standard ANOVA. Based on the findings of this study, it was recommended that the disease incidence and pest abundance data should be analyzed using generalized linear models instead of standard ANOVA. The results obtained from application of ANOVA to disease incidence and pest abundance showed that the model was inappropriate and did not fit well the data used in this study.

REFERENCES

Agresti, A. 2002. Categorical Data Analysis Wiley-Interscience, New York.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. 2nd International symposium on Information theory, Ed. B. N. Petrov and F. Csaki. Budapest; Akademia Kiado.267-281.

Anderson, D.R., Burnham, K.P and Thompson, W.L. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64:912-923.

Anderson, T.W and Darling, D.A. 1952. Asymptomatic theory of certain "goodness-of-fit" criteria based on stochastic processes.*Ann.Math.Stat*.23:193-212.

Anscombe, F.J. 1949. The analysis of insect counts based on the negative binomial distribution. *Biometrics* 5:165-173.

Armitage, P and Colton, T. 1998. Encyclopedia of Biostatistics. Wiley. New York.

Barrett, P. 2007. "Structural Equation Modeling: Adjudging Model Fit," Personality Individual Differences, 42(5): 815-24.

Box, G.E.P., and Cox D.R. 1964. An analysis of transformations. *Journal of the Statistical Society*, 26:211-234.

Burnham, K.P., and Anderson, D.R. 2002. Model Selection and Multimode Inference: practical information-theoretic approach, 2nd edition. Springer-Verlag, New York.

Cameron, A.C., Trivedi, P.K. 1998. Regression analysis for count data. Econometrics Society Monographs No. 30. Cambridge University Press. Cambridge (UK).

Campbell, C.L and Madden, L.V. 1990. *Introduction to plant Disease Epidemiology*. New York: Wiley Interscience. 532pp.

Chellemi, D.O., Rohrbach, K.G., Yost, R.S and Sonoda, R.M. 1988. Analysis of the spatial pattern of plant pathogens and diseased plants using Geostatistics. Phytopathology 78:221-226

Chester, K.S. 1950. Plant disease loss: their appraisal and interpretation. Plant Disease Reporter Supplement No. 193:189-362.

Cochran, W.G. 1940. The analysis of variances when experimental errors follow the Poisson or binomial laws. *The Annals of Mathematical Statistics 11:* 335-347.

Cochran, W.G. 1947. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics* 3: 22.38.

Collett, D. 2002. *Modeling Binary Data*. 2nd edition.CRS Press, Boca Raton, FL

Cox, D.R and Snell, E.J. 1989. The Analysis of Binary Data. 2nd ed. London. Chapman and Hall. p. 236.

De Wolf, E.D., Madden, L.V., and Lipps, P.E. 2003. Risk assessment models for wheat Fusarium head blight epidemics based on within-season weather data. *Phytopathology*, 93: 428–435.

Fletcher, D., MacKenzie, D and Villouta, E. 2005. Modeling skewed data with many zeros: a simple approach combining ordinary and logistic regression.

Fligner, M.A and Killeen, T.J. 1976. Distribution-free two-sample tests for scale. 'Journal of the American Statistical Association.71 (353), 210-213.

Gardner, W., Mulvey, E.P., Shaw, E.C. 1995. Regression analysis of counts and rates: Poisson over dispersed Poisson and negative binomial models. Psychological Bulletin, 118(3):392-404.

Garrett, K.A., Madden, L.V., Hughes, G and Pfender, W.F. 2004. New applications of statistical tools in plant pathology. Phytopathology 94:999-1003.

Green, S.K., Kalloo, G. 1994. Leaf curling and yellowing virus of pepper and tomato: an overview. Technical Bulletin, Asian Vegetable Research and Development Center.21: 151.

Hardy, I.C.W. 2002. Sex ratios: Concepts and research methods. Cambridge University Press.

Hobson, G.E. and Davies, J.N. 1971. The tomato. In: Hulme AC (eds.). The biochemistry of fruits and their products. Vol. 2. Academic press, New York London. pp. 337– 482.

Hosmer, D.W and Lemeshow, Jr. 2000. Applied logistic regression (2nd ed.).New York. Janik, J., & Kravitz, H.M 1994. Linking work and domestic problem with police suicide. Suicide and Life Threatening Behavior.24:267-274.

Hughes, G and Gottwald, T.R. 1998. Survey methods for assessment of citrus tristeza virus incidence. *Phytopathology* 88:715-723.7

Hughes, G and Madden, L.V. 1995. Some methods allowing for aggregated patterns of diseases incidences in the analysis of data from designed experiment. *Plant Pathology* 44:927- 943.

Hughes, G., and Gottwald, T.R. 1999. Survey methods for assessment of citrus tristeza virus incidence when *Toxoptera citricida* is the predominant vector. Phytopathology 89:487-494.

Hughes, G., Munkvold, G.P., Samita, S. 1998. Application of the logistic-normal-binomial distribution to the analysis of Eutypa dieback disease incidence. *International Journal of Pest Management* 44: 35-42.

Jaeger, T.F. 2008. Categorical data analysis: away from ANOVAs (transformation or not) and towards Logit Mixed Models.J.Mem.Lang.59: 434-446.

Johnson, N.I and Kotz, S. 1969. *Discrete Distributions*. Houghton Mifflin Company, Boston.

Kalloo, G. 1991. Genetic improvement of tomato. Springer Verlag, Berlin Heidelberg, Germany. p. 358.

Kapa, J.M and Waterworth, H.E. 1981. Handbook of plant virus infections and Comparative Diagnosis. Ed .E. Kursta,pp.257-332.Elsevier/North Holland, New York.

Kenney, J.F and Keeping, E.S. 1962. Skewness.7.10 *in Mathematics of statistics*, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E., Nizam, A. 1998. Applied regression analysis and other multivariable methods-Third Edition. Brooks/Cole Publishing Company, Duxbury Press, Pacic Grove (CA).

Kranz, J. 1988. Measuring plant disease. *In Experimental Technique in Plant Disease Epidemiology*, eds, J Kranz, J Rotem. Springer-Verlag, Berlin, pp.35-50.

Levene, H. 1960. Contributions to Probability and Statistics.

Lucas, J.A. 1998. Plant pathology and plant pathogens (3rd edn). Blackwell Science, UK, p.274.

Madden, L.V. 2002. A population-dynamic approach to assess the threat of plant pathogens as biological weapons against annual crops. BioScience 52: 65-74.

Madden, L.V and Hughes, G. 1995. Plant disease incidence: Distributions, heterogeneity, and temporal analysis. *Annual Review of Phytopathology* 33:529-564

Madden, L.V and Hughes, G. 1999. An effective sample size for predicting plant disease incidence in a spatial hierarchy. *Phytopathology* 89: 770-781.

Madden, L.V., Turechek, W.W., and Nita, M. 2002. Evaluation of generalized linear mixed models for analyzing disease incidence data obtained in designed experiments. Plant Dis.86:316-325.

Martub, T.G., Wubtek, B.A., Kuhnert, J.R., Field, P.M., Low-Choy, S.A., Tyre, S.J Possingham, H.P. 2005. Aero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology Letters* 8: 1235-1246.

McArdle, B.H and Anderson, M.J. 2004. Variance heterogeneity, transformations, and models of species abundance: a cautionary tale. *Canadian Journal of Fisheries and Aquatic Sciences* 61: 1294-1302.

McCullagh, P and Nelder, J.A 1989. *Generalized Linear Models*, 2nd edition, Longo, Chapman and Hall).

McRoberts, N., Hughes, G and Madden, L.V. 1996. Incorporating spatial variability into simple disease progress models for crop pathogens. *Aspects of Applied Biology* 46: 1-8.

McRoberts, N., Hughes, G and Madden, L.V. 2003. The theoretical basis and practical application of relationships between different disease intensity measurements in plants. Ann.Appl.Biol.142:191-211.

Menard, S. 1995. Applied logistic regression analysis. Sage University Paper series on Quantitative Applications in the Social Sciences series no. 07-106.Thousand Oak(CA).

Miles, J and Shevlin, M. 2001. *Applying Regression and Correlation*. London: Sage.

Mullahy, J. 1997. Heterogeneity, excess zeros, and the structure of count data models. Journal of Applied Econometrics 12: 337–350.

Nutter, F.W. Jr., Teng, P.S and Shokes. 1991. Disease assessment terms and concepts. *plant Dis*. 75: 1187-1188.

Nutter, F.W., Esker, P.D and Netto, R.A. C. 2006. Disease assessment concepts and the advancements made in improving the accuracy and precision of plant disease data. *European Journal of Plant Pathology,* 115:95-1371.

Nutter, F.W., Teng, P.S and Royer, M.H. 1993. Terms and concepts for yield, crop los and disease thresholds. Plant Disease 77:211-215.

Osgood, W. 2000. Poisson-based Regression Analysis of Aggregate Crime Rates." Journal of Quantitative Criminology 16: 21-43.

Paternoster, R and Brame, R. 1997. Multiple Routes to Delinquency? A Test Developmental and General Theories of Crime," Criminology 35: 45-84.

Perry, J.N., Rothery, P., Clark, S.J., Heard, M.S and Hawes, C. 2003. Design, analysis and statistical power of the farm scale evaluation of genetically modified herbicide-tolerant crops. Journal of Applied Ecology 40: 17–31. Phytopathology78:221-226.

Rao, M.M. 1960. *Some asymptotic results on transformations in the analysis of variance*. ARLTechnical Note, 60-126. Aerospace Research Laboratory, Wright-Patterson Air Force Base.

Ridout, M., Demetrio, C.G. B and Hinde, J. 1998. Models for counts data with many zeros Invited paper presented at the Nineteenth International Biometric Conference, Cape Town, South Africa, 179-190.

Saha, K and Paul, S. 2005. Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* 61: 179.185.

Sakia, R.M. 1992. The Box-Cox transformation technique: A review. The

statistician, 41:169-178.

Schabenberger, O and Pierce, F.J. 2002. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press, Boca Raton, FL.

Seem, R.C. 1984. Disease incidence and severity relationships. *Annual Review of Phytopathology*. 22:137-50.

Shapiro, S.S and Wilk, M.B.1965. Analysis of variance test for normality. *Biometrika* 52:591-611

Sileshi, G., Girma, H and Mafongoya, P.L. 2006 (a). Occupancy-abundance models For predicting Efficient Analysis of Abundance and Incidence Data densities of three leaf beetles damaging the multipurpose tree *Sesbania sesban* in eastern and southern Africa. *Bulletin of Entomological Research* 96:61-69.

Sileshi, G., Mafongoya, P.L and Kuntashula, E. 2006 (b). Legume improved fallows reduce weed problems in maize in eastern Zambia. *Zambian Journal of Agriculture*.8:12.

Snedecor, G.W and Cochran, W.G. 1989. *Statistical methods*. 8th edition, Iowa State University, Ames.

Sseruwagi, P., Sserubombwe, W.S., Legg, J.P., Ndunguru, J., Thresh, J.M. 2004. Methods of Surveying the incidence and Severity of Cassava Mosaic Disease and Whitefly Vector Populations on Cassava in Africa: a review. Available online at www.sciencedirect.com,2004

Steel, R.G.D and Torrie, J.H. (1997). Principles and procedures of statistics. McGraw Hill Book Co., NY. USA.

Strange, R.N. 2003. Introduction to plant pathology. John Wiley & Sons Ltd., UK, 464pp.

Taylor, L.R. 1961. Aggregation, variance and the mean. Nature *189*: 732–735.

Tukey, J.W. 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA

Turechek, W.W. 2004. Nonparametric tests in plant disease epidemiology: characterizing disease associations. *Phytopathology* 94:1018-1021.

Wang, W and Famoye, F. 1997. Modeling household fertility decisions with generalized poisson regression. *Journal of population Economics* 10:273-283

Warton, D.I. 2005. Many zeros does not mean zero inflation: comparing the goodness-of- fit of parametric models to multivariate abundance data. *Environ metrics* 16 275–289.

Warton, D.I. and Hui, F.K.C. 2010. The arcsine is asinine: the analysis of proportions in ecology, *Ecology*.

Winer, B.J., Brown, D.R and Michels, K.M. 1971. Statistical *principles in experimental design*. New York: McGraw-Hill.

Wulu, J.T., Singh, K.P., Famoye, F and McGwin, G. 2002. Regression analysis of count data. *Journal of the Indian Society of Agricultural* Statistics 55:220-231www.sciencedirect.com, 2004.

Yang, Y. 2005. Can the strengths of AIC and BIC be shared? A conict between identication and regression estimation. Biometrika, 92: 937-950.

Zadoks, J.C and Schein, R.D. 1976. *Epidemiology and plant Disease Management*. Oxford University Press, New York, 427pp.

APPENDICES

Appendix 1: TYLCD incidence and Whitefly counts survey data sheet

FIELD NO	CULTIVAR NAME	ALTITUDE	TOTAL PLANT	INFECTED PLANTS	WHITEFLY COUNTS
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					

Appendix 2: Tomato fruit weight data sheet

CULTIVAR NAME	REPLICATION	Fruit weight(KG)
CNL 3125E	1	
CNL 3125L	1	
CNL3125P	1	
CNL3078G	1	
CNL 3070J	1	
TENGERU	1	
TANYA	1	
CNL 3125E	2	
CNL 3125L	2	
CNL3125P	2	
CNL3078G	2	
CNL 3070J	2	
TENGERU	2	
TANYA	2	
CNL 3125E	3	
CNL 3125L	3	
CNL3125P	3	
CNL3078G	3	
CNL 3070J	3	
TENGERU	3	
TANYA	3	