Research Article *Journal of Plantation Crops, 2012, 40(2): 132-134*

A robust estimate for location in experiments with plantation crops

C.T. Jose[1]*, S. Jayasekhar[2], K. Muralidharan[2], K.P. Chandran[2] and D. Jaganathan[1]

[1]*Central Plantation Crops Research Institute, Regional Station, Vittal-574 243, Karnataka*
[2]*Central Plantation Crops Research Institute, Karsaragod-671 121, Kerala*

(Manuscript Received: 02-03-12, Revised: 23-03-12, Accepted: 24-05-12)

Abstract

Presence of outliers or extreme values in the experimental data is a major concern for data analysis. Many times, the experimental data contains abnormal or extreme values due to genetic variations, loss of yield due to pest/disease infestation, errors in tabulation/data entry *etc*. In field experimental data analysis, the plot mean or treatment mean is usually considered for comparison. The sample mean, which is usually taken as an estimate of the population mean (location estimator) is highly affected by the presence of outliers or extreme values particularly when the sample size is small. In this paper, kernel weighted location estimator with weight proportional to the value of the estimated kernel density function is proposed, to handle the outliers or extreme values. The kernel weighted location estimator is robust even if the underlying distribution is non-symmetric. The performance of the proposed method is compared with the existing procedures through simulation study. The method is also applied to the analysis of yield data of arecanut and cocoa mixed cropping experiment at CPCRI Regional Station, Vittal.

Keywords: Location estimator, outliers, perennial crops, robust technique

Introduction

The fundamental objective of an agricultural experiment is to obtain data systematically and to make inferences or appropriate decisions based on the data. Presence of outliers or extreme values in the experimental data is a major concern for data analysis. Outlier is an observation that appears to be inconsistent with the remainder of the observations in the data set. The experimental yield data may contain abnormal or extreme values due to genetic variations among treatments (super trees/ very low yielders), yield loss due to pest/disease infestation, tabulation/data entry errors *etc*. These extreme values or outliers will not only increase experimental error, but sometimes it affects the inference also. The sample mean, which is usually taken as an estimate of the population mean (location estimator), is highly affected by the presence of outliers or extreme values particularly when the sample size is small. During the past several years various approaches have been proposed to deal with the lack of robustness of the sample mean as an estimate of the population mean when the distribution sampled is contaminated by gross errors, *i.e*., it has heavier tails than the normal distribution. Hodges and Lehmann (1956), proposed estimates related to the well-known robust Wilcoxon and normal scores tests. Huber (1964) considered essentially the class of maximum likelihood estimates and found those members of this class which minimize the maximum variance over various classes of contaminated distributions.

In this paper, a robust technique is proposed to handle the outliers or extreme values and to control the experimental error. In field experiments with perennial tree crops, the plot size or the number of trees per plot is usually taken as more than one and the plot average is considered for data analysis. Presence of an outlier in the plot may badly affect plot mean (location parameter) and it may give a wrong conclusion. In this paper, instead of plot average, a robust estimate for the location parameter

**Corresponding Author: ctjos@yahoo.com*

is proposed to handle the outliers or extreme values in the data. The performance of the proposed method is compared with the existing procedures in the SPSS package through simulation study.

Materials and Methods

Analysis of variance (ANOVA) technique is used to compare the effect of different treatments in field experiments and the statistical model for the simple ANOVA is of the form

$$Y = X\beta + \varepsilon \qquad (1)$$

where, $Y = [y_1, y_2, ..., y_n]$ is the observation vector, X is the design matrix, β is the vector of treatment effect and ε is the error term. When the plot size is more than one, plot mean is taken as y_i (i=1,...,n), but the mean is very sensitive to outliers. In this paper, a robust estimate for the plot value y_i is proposed which is not sensitive to outliers as plot average. Kernel weighted location estimator with weight proportional to the value of the estimated kernel density function is used as the plot value y_i. Let $Y_i = [y_{11}, y_{12}, ..., y_{1p}]$ be the observation vector of the i^{th} plot and p is the plot size. The estimate of the i^{th} plot value $\hat{y}_{i.}$ is obtained by the iterative procedure with starting value θ_0 = median (Y_i) in the following expression.

$$\theta_{ik} = \frac{\sum_{j=1}^{p} y_{ij} K\left(\frac{y_{ij} - \theta_{ik-1}}{h_{ik-1}}\right)}{\sum_{j=1}^{p} K\left(\frac{y_{ij} - \theta_{ik-1}}{h_{ik-1}}\right)}, \quad i=1,...n, \quad k=1,..., q$$

where, the symmetric kernel density function K and the bandwidth h_{ik} are taken as

$$K(x) = \begin{cases} \frac{3}{4}\left(1 - x^2\right), & |x| \le 1 \\ 0, & otherwise \end{cases}$$

$$h_{ik} = 3\sqrt{\text{med}\left[\left(Y_i - \theta_{ik-1}\right)^2\right]}$$

The location estimate $\hat{y}_{i.} = \theta_{iq}$ where q is fixed in such a way that the value of $|\theta_{iq} - \theta_{iq-1}|$ is negligible. Estimated plot value $(\hat{y}_{i.})$ is taken as the i^{th} plot value for the analysis. In robust M- estimators for location, the distance from the central value is generally considered for obtaining the weight of each observation and therefore it is more suitable for symmetric distributions. In the proposed procedure, the estimated value of the kernel density function corresponding to the observation is used as the weight and therefore it is suitable for any continuous distribution which need not be symmetric. The estimated value of the kernel density function corresponding to a rare observation (outlier) will be very low and its effect on the location estimate will be insignificant. A simulation study is carried out to see the performance of the proposed method for handling outliers in the data. The following ANOVA model is considered for the simulation study.

$$y_{ijk} = \mu + t_i + b_j + \varepsilon_{ijk}, \ i=1,..4 \ j=1,..,5, \ k=1,..7 \qquad (2)$$

where, $\mu = 9$; $t = [-4 \ 0 \ 0 \ 4]$ is the treatment effect vector, $b = [-2 \ -2 \ 0 \ 2 \ 2]$ is block effect, the error term ε_{ijk} follows N (0, 2) and

$$y_{ij8} = 3\mu + t_i + b_j + \varepsilon_{ij8}, \ i=1,..4 \ j=1,..,5,$$

are taken as the outliers. Based on the above model 100 sets of data are generated with 20 treatment combinations and 8 observations per plot. The average mean squared errors (AMSE) of the estimated values with the true values of 100 sets of simulated data is computed using the proposed method and it is compared with the robust estimates in the SPSS package (Ver.13). The above simulation study is extended by changing the number of observations per plot (25, 50, 75 and 100) with one observation per plot as outlier to see the effect of outliers in different location estimators as number of observations increases.

The experiment consists of 6 spacing and 2 manurial treatments with 4 replications. Yield data of 4, 6 and 8 arecanut palms were selected randomly from each plot to compare performance of the proposed method under different plot size. Yield data from 9^{th} to 16^{th} year of planting were considered for the analysis.

Results and Discussion

The average mean square errors (AMSE) of the estimated values with the true value obtained from the 100 set of simulated data (model 2) using the proposed and existing methods (SPSS) are given in Fig.1. The data generated for the simulation study contains one outlier in each plot of size 8. Note that in the simulation study (Fig. 1), the AMSE of the

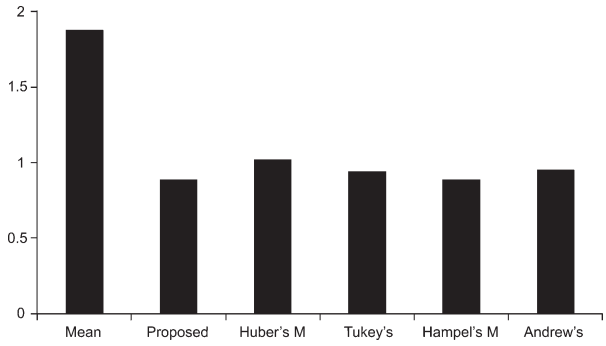

Fig. 1. Average MSE of the estimated values with the true values in the simulation study (plot size 8)

estimated values are comparatively less in the proposed method than the existing methods available in the SPSS package (Ver 13.0). Also note that in the presence of outliers, mean performs very badly. In the computation of mean, all the observations will have equal weight, whereas in the case of robust locations estimators, the weight of extreme values will be less than that of the observations which are near to the central value.

The average mean square errors (AMSE) of the estimated values with the true value obtained from the 100 set of simulated data (model 2) for different number of observations per plot (25, 50, 75 and 100) using the proposed and existing methods are given in Table 1.

The AMSE of the estimated values with the true value or the effect of outliers on different location estimators reduces as the number of observations or the plot size increases. Also, the performance of different estimators is more or less same when the number of observations is large (Table 1).

The proportion of unexplained variations (error sum of squares/ total sum of squares) in the analysis of arecanut yield data in the mixed cropping experiment at CPCRI Regional Station by taking different plot sizes (number of palms/plot) and different period (number of years) using the proposed robust technique as well as by taking plot average is given in Table 2.

Table 2. Proportion of unexplained variations in ANOVA

Period (no. of years)	8 palms/plot		6 palms/plot		4 palms/plot	
	robust	mean	robust	mean	robust	mean
2	0.39	0.41	0.45	0.47	0.48	0.54
3	0.33	0.38	0.40	0.46	0.45	0.53
4	0.26	0.28	0.32	0.36	0.33	0.36
5	0.26	0.28	0.29	0.31	0.32	0.33
6	0.26	0.26	0.27	0.28	0.31	0.32
7	0.27	0.27	0.29	0.29	0.34	0.33
8	0.28	0.27	0.31	0.30	0.34	0.33

Note that, if the number of years considered for the cumulative yield data is less than six, the proposed method performs better than the plot average irrespective of the plot size (Table 1). When the number of years is six or more, the proportion of unexplained variation in the analysis of the cumulative yield data using the robust estimate (proposed method) and the plot average is almost same. It is obvious that as the number of years or the number observation increases, the effect of extreme or abnormal value on the mean value will be less. The plant to plant variation in yield is very high in tree crops and many times the simple average will be misleading in the presence of some extreme values. The effect of these outliers or extreme values is more when the number of observation is comparatively less. It is better to use robust estimate for the plot value than the simple plot average for the analysis of yield data in perennial crops like arecanut particularly when the number of years is less.

Table 1. Average MSE of the estimated values with the true values in the simulation study for different plot size

Plot size	Mean	Proposed	Huber's M	Tukey's	Hampel's M	Andrew's
25	0.719	0.163	0.176	0.196	0.165	0.199
50	0.192	0.060	0.061	0.064	0.060	0.064
75	0.065	0.040	0.040	0.043	0.039	0.044
100	0.041	0.039	0.037	0.038	0.038	0.038

References

Hodges, J.L. and Lehmann, E.L. 1956. Efficiency of some nonparametric competition of the t-test, *Ann, Math. Statist*. **27**: 324-335.

Huber, P.J. 1964. Robust estimation of a location parameter, *Ann. Math. Statist*. **35**: 73-102.