STATISTICS UNIT

Report No. 3/87

# EFFECT OF OUTLIERS IN ACCOUNTING THE VARIABILITY IN CHICKPEA TRIALS.

P.Venkateswarlu, G.Swaminathan & M.Singh

ICRISAT

International Crops Research Institute for the Semi-Arid Tropics

ICRISAT Patancheru P.O. Andhra Pradesh 502 324, India

# Effect of outliers in accounting the variability in Chickpea trials

P.VENKATESWARLU, G.SWAMINATHAN & M. SINGH
International Crops Research Institute for the Semi-Arid Tropics, Patancheru 502324, A.P.

## SUMMARY

For a set of 64 chickpea trials administered by ICRISAT, the effect of outliers in field variation was examined. Outliers were detected using a Q-Q plot and a statistical test. We compared the coefficients of variation with and without outliers. In almost every trial an outlier was successfully detected and its removal resulted into reduced coefficient of variation.

## INTRODUCTION

Outliers are the observations other than majority or the large number of observations generated through a systematic process. Outliers in agricultural experiments may arise due to rare fertility patches, undue low or high plant stand in a plot, incorrect (wrong) level of treatment assigned to a plot , very high (unbalanced) competition between short and tall genotypes happened to be (sown) in neighbours etc. The analysis of variance or method of fitting constants will tend to provide biased and inefficient estimates of treatment effects and field precision if there are some outlying observations amidst the data values.

The purpose of this paper is to illustrate the detection of outliers with a worked out example on real data and compare the estimates of error variances from the data with and without outliers. Based on the analysis of 64 trials of chickpea, the coefficients of variation are presented.

## MATERIALS AND METHODS

The data base is the yield records of 64 trials conducted in randomised complete block designs with 16 genotypes of chickpea conducted at various locations over three years administered by chickpea breeding unit of ICRISAT. Data from all the 64 trials were examined for detecting the presence of outliers using the technique described below.

### Procedure for detection of outliers

We shall first illustrate the detection of outliers from the data of one of the trials (Table 1).

Table 1. Yield (kg/ha) of 16 genotypes evaluated in four randomized blocks and least square residuals.

| Genotype | Yield | | | | Least Square residuals | | | |
|---|---|---|---|---|---|---|---|---|
| | Block | | | | Block | | | |
| | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| 1 | * | 314.2 | 242.8 | 176.1 | * | -16.8 | 58.2 | -41.4 |
| 2 | * | 433.1 | 433.2 | 57.1 | * | 38.7 | 185.1 | -223.9 |
| 3 | 109.5 | 318.9 | 190.4 | 147.6 | -134.7 | 58.3 | 76.1 | 0.3 |
| 4 | 252.3 | 295.1 | 756.8 | 599.8 | -276.3 | -250.0 | 358.1 | 168.1 |
| 5 | 147.6 | 485.5 | 523.6 | 257.0 | -258.4 | 63.0 | 247.5 | -52.1 |
| 6 | 142.8 | 547.4 | 304.6 | 342.0 | -244.0 | 144.1 | 47.7 | 52.1 |
| 7 | 357.0 | 595.0 | 214.2 | 290.4 | -59.7 | 161.8 | -72.6 | -29.5 |
| 8 | 238.0 | 890.1 | 333.2 | 180.9 | -225.1 | 410.5 | -0.1 | -185.4 |
| 9 | 666.4 | 357.0 | 442.7 | 609.3 | 95.0 | -230.9 | 1.1 | 134.8 |
| 10 | 533.1 | 466.5 | 399.8 | 476.0 | 11.7 | -71.5 | 8.3 | 51.5 |
| 11 | 904.4 | 856.8 | 647.4 | 595.0 | 101.0 | 36.8 | -26.2 | -111.6 |
| 12 | 1190.0 | 1099.6 | 856.8 | 913.9 | 122.4 | 15.4 | -81.0 | -56.8 |
| 13 | 656.9 | 623.6 | 276.1 | 785.4 | 18.8 | -31.0 | -232.1 | 244.2 |
| 14 | 542.6 | 585.5 | 618.8 | 690.2 | -119.2 | -92.9 | -86.8 | 125.3 |
| 15 | 661.6$_a$ | 871.1 | 476.0$_b$ | 909.2 | -120.4$_a$ | 72.5 | -176.2$_b$ | 224.0 |
| 16 | 2023.4 | 642.6 | 323.7 | 537.9 | 1089.0 | -308.4 | -480.9 | -299.7 |

* - missing values; a,b - suspected outliers

The occurrence of outliers in the data can be visualized with the help of a Q-Q plot (Figure 1) where the ordered normal residuals from Table 1 have been plotted against the respective quantile values of standard normal distribution, q. Corresponding to i-th ordered residual of the data, say, $r_{i,n}$, the associated quantile from standard normal distribution is given by $q_{i,n}$ where

$$\int_{-\infty}^{q_{i,n}} \phi(t)dt = i/n,$$

where n is the total number of residuals and $\phi(t) = (1/(2\pi)^{\frac{1}{2}})e^{-t^2/2}$, the probability density function of standard normal distribution. It is very obvious from the plot $(r_{i,n}, q_{i,n})$ i=1...n, n=62) in Figure 1 that the residuals corresponding to genotype 16 in 1st and 3rd blocks (say plots A and B) are lying too far from the points which lie nearly on a straight line. Thus the corresponding observations of these plots, i.e. 2023 and 323 are suspected outliers. A test described in Tiku et. al., (1986) for testing data anamoly and detecting outliers has been applied here. We got the following analysis of variance tables using FIT directives in GENSTAT and fitting blocks and genotypes factors on the yield data with and without the suspected outlier plots A and B for the genotype 16.

Table 2. Analysis of Variance

| | (i) With outliers | | | (ii) Without outliers | | |
|---|---|---|---|---|---|---|
| Source | df | SS | MS | df | SS | MS |
| Rep | 3 | 301489 | | 3 | 172636 | |
| Genotypes (Adj.for reps) | 15 | 3087037 | | 15 | 2621227 | |
| Residuals | 43 | 2925357 | 68032 | 41 | 1200409 | 29278 |
| TOTAL | 61 | 6313882 | | 59 | 3994272 | |

In order to test that atleast one of the two suspected plots A and B is an outlier, compute the change in residual sum of squares (d.f.=2) C = 2925357-1200409 = 1724948. The residual MS (d.f.=41) = 29278 is an unbiased estimate of $\hat{\sigma}^2$ of $\sigma^2$ (error variance per plot) whether outliers are present or not. Now we compute the test statistics (Tiku et.al., 1986).

$$F = (C/2)/\hat{\sigma}^2 = (174948/2)/29278 = 29.46$$

which is higher than tabulated values of F-distribution with 2 and 41 degrees of freedom at probability level P=.001. Therefore atleast one of A and B is an outlier. We then follow a sequential procedure to detect outliers one by one. On the basis of higher residual for plot A (also supported by too off position of A in Figure 1), we first test whether A is an outlier. Ignoring plot A, we get the following analysis of variance (table 3).

Table 3. Analysis of Variance ignoring plot A

| Source | df. | SS | MS |
|---|---|---|---|
| Rep | 3 | 189177 | |
| Genotype(Rep) | 15 | 2613409 | |
| Residual | 42 | 1222659 | 29111 |
| TOTAL | 60 | 4025246 | |

We get change in residual sums of squares due to omission of A from Table 2(ii) (d.f.=1) 2925357-1222659 = 1702698 and hence

$$F = (1702698/1)/29278 = 58.49$$

which is more than P=.001 probability level point of F distribution with 1 and 41 degrees of freedom. Thus A is a strong outlier. Now, to test whether B is also an outlier, we use the analyses of variance in Table 2(ii) (ignoring plots A and B) and Table 3 (ignoring plot A).

The change in residual sums of squares due to omission of B(d.f.=1) = 1222659-1200409 = 22250 and the statistic val—

$$F = (22250/1)/29278 = 0.38$$

which is less than P=.05 probability level point of F-distribution with 1 and 41 degrees of freedom. Therefore, there is no evidence of B to be an outlier. We get the Q-Q plot without A in Figure 2 which appears to be a reasonable straight line.

## RESULTS AND DISCUSSION

When individual trial data from chickpea experiments was subjected to the above analysis, we found that each data set had an outlier. Table 4 gives the values of residual mean squares, coefficient of variation (CV%) for each of the trials analysed

with and without outlier detected. It can be noticed that the reduction in residual mean square due to deletion of outlier is enormous. This fact has also been exhibited by cumulative distributions of CV(X)s presented in Figure 3.

The analyses of yield from 64 trials of chickpea have indicated the presence of outliers. Once an outlier is detected it is worth attempting to analyse the data after ignoring the outlier plot, since it provides more precise estimate of error variance and hence that of treatment contrasts also. It is therefore recommended that data from designed experiments must be subjected for the exploration of the presence of outliers as they have remarkable effect on the inferences on treatment comparisons.

Table 4. Residual means squares (MS) and coefficient of variation (CV%) when analysed with and without outliers in various trials

| Trial No. | With outlier | | Without outlier | | Percent reduction in MS |
|---|---|---|---|---|---|
| | MS | CV% | MS | CV% | |
| 1 | 60630 | 21.3 | 40743 | 17.1 | 32.8 |
| 2 | 155485 | 31.4 | 111775 | 26.4 | 28.1 |
| 3 | 22983 | 13.3 | 18522 | 11.3 | 19.4 |
| 4 | 90940 | 52.6 | 47655 | 41.2 | 47.6 |
| 5 | 73775 | 45.0 | 45413 | 35.7 | 38.4 |
| 6 | 15083 | 31.3 | 11758 | 27.0 | 22.0 |
| 7 | 46464 | 22.4 | 30584 | 18.2 | 34.2 |
| 8 | 89253 | 15.4 | 62410 | 12.7 | 30.1 |
| 9 | 90260 | 19.0 | 61507 | 15.3 | 31.9 |
| 10 | 33953 | 16.5 | 26470 | 14.6 | 22.0 |
| 11 | 10196 | 39.0 | 7856 | 35.7 | 23.0 |
| 12 | 331772 | 60.9 | 269261 | 58.4 | 18.8 |
| 13 | 35205 | 14.2 | 29363 | 12.8 | 16.6 |
| 14 | 213333 | 17.4 | 129780 | 13.8 | 39.2 |
| 15 | 68032 | 50.4 | 29278 | 34.6 | 57.0 |
| 16 | 85018 | 32.6 | 60892 | 28.3 | 28.4 |
| 17 | 39950 | 36.2 | 30670 | 32.9 | 23.2 |
| 18 | 78278 | 13.8 | 63895 | 12.5 | 18.4 |
| 19 | 1663753 | 39.2 | 956579 | 29.5 | 42.5 |
| 20 | 356736 | 42.9 | 236223 | 36.6 | 33.8 |
| 21 | 48163 | 36.5 | 38571 | 33.9 | 19.9 |
| 22 | 4808 | 24.9 | 3739 | 21.9 | 22.2 |
| 23 | 16862 | 10.7 | 11497 | 8.7 | 31.8 |
| 24 | 4389 | 5.7 | 1102 | 2.8 | 74.9 |
| 25 | 80966 | 19.0 | 61344 | 16.5 | 24.2 |
| 26 | 348560 | 32.1 | 268979 | 28.9 | 22.8 |
| 27 | 445644 | 35.5 | 328701 | 30.3 | 26.2 |
| 28 | 171098 | 29.8 | 118323 | 25.7 | 30.8 |
| 29 | 122005 | 19.8 | 98966 | 17.7 | 18.9 |
| 30 | 169556 | 35.3 | 86631 | 26.0 | 48.9 |
| 31 | 79777 | 39.6 | 68615 | 36.9 | 14.0 |
| 32 | 282722 | 59.3 | 224964 | 54.1 | 20.4 |
| 33 | 388491 | 52.4 | 206945 | 40.8 | 46.7 |
| 34 | 23046 | 47.8 | 18642 | 43.6 | 19.1 |
| 35 | 51735 | 80.0 | 23620 | 60.4 | 54.3 |
| 36 | 159554 | 15.5 | 132966 | 14.2 | 16.7 |
| 37 | 112020 | 35.9 | 82275 | 31.1 | 26.6 |
| 38 | 67924 | 16.2 | 52614 | 14.4 | 22.5 |
| 39 | 614023 | 50.4 | 288059 | 39.9 | 53.1 |
| 40 | 41633 | 19.4 | 33726 | 17.3 | 19.0 |
| 41 | 323156 | 24.9 | 130582 | 15.4 | 59.6 |
| 42 | 87102 | 11.2 | 69620 | 9.9 | 20.1 |
| 43 | 31019 | 17.5 | 24496 | 15.5 | 21.0 |
| 44 | 258304 | 55.4 | 213984 | 52.9 | 17.2 |
| 45 | 199165 | 21.5 | 169334 | 17.9 | 15.0 |
| 46 | 13004 | 19.8 | 10430 | 17.7 | 19.8 |
| 47 | 200086 | 30.5 | 141517 | 25.2 | 29.3 |

| | | | | | |
|---|---|---|---|---|---|
| 48 | 63891 | 51.8 | 53900 | 49.4 | 15.6 |
| 49 | 56624 | 74.0 | 37972 | 60.8 | 32.9 |
| 50 | 444848 | 27.8 | 351162 | 25.2 | 21.1 |
| 51 | 34214 | 39.3 | 22862 | 30.9 | 33.2 |
| 52 | 68623 | 12.0 | 55058 | 10.9 | 19.8 |
| 53 | 206960 | 17.9 | 164414 | 15.9 | 20.6 |
| 54 | 301201 | 23.4 | 230641 | 20.4 | 23.4 |
| 55 | 27124 | 12.0 | 15844 | 8.7 | 41.6 |
| 56 | 86542 | 25.0 | 62411 | 22.6 | 27.9 |
| 57 | 460131 | 40.4 | 357603 | 35.4 | 22.3 |
| 58 | 32636 | 33.7 | 22295 | 28.8 | 31.7 |
| 59 | 106825 | 17.4 | 85009 | 15.7 | 20.4 |
| 60 | 152078 | 48.2 | 124840 | 45.6 | 17.9 |
| 61 | 14870 | 28.0 | 11806 | 25.0 | 20.6 |
| 62 | 40113 | 12.1 | 28186 | 10.3 | 29.7 |
| 63 | 45739 | 36.1 | 32558 | 31.7 | 28.8 |
| 64 | 228688 | 26.4 | 188558 | 24.4 | 17.5 |

ACKNOWLEDGEMENT

Authors are thankful to Dr. Jagdish Kumar, Chickpea Breeder, Mr. J.H.Miranda, Senior Research Associate, Chickpea Breeding Unit of ICRISAT, for providing the data for illustration in this paper.

REFERENCE

Tiku M.L., Tan, W.Y., Balakrishnan, N. 1986. Robust Inference. Marcel Dekker Inc. New York, pp. 178.

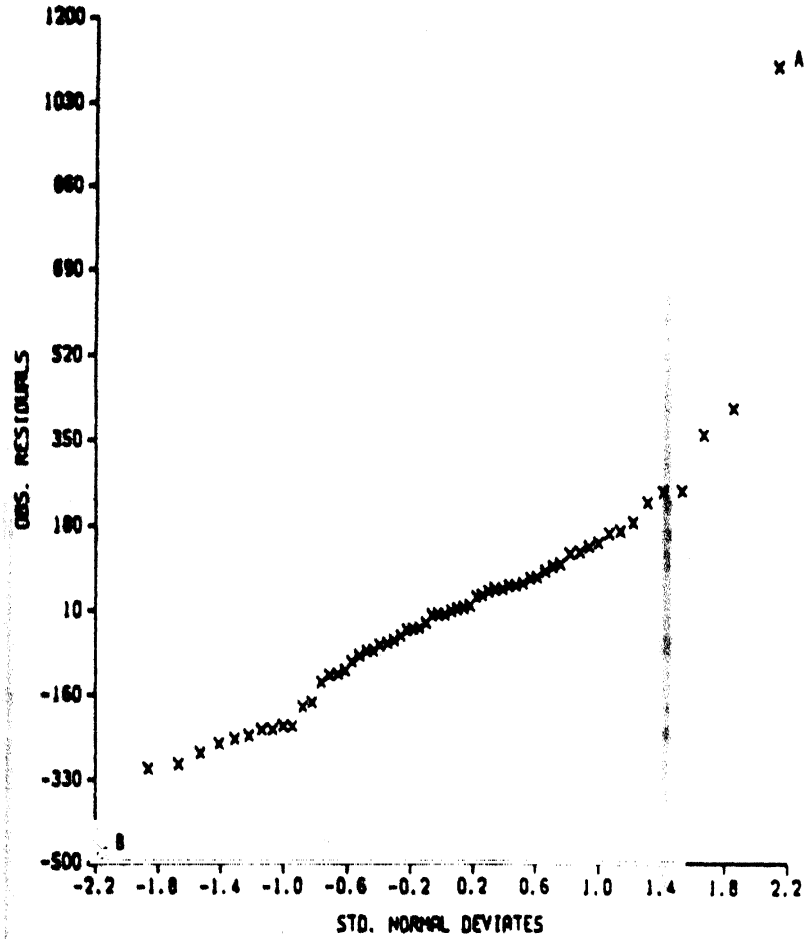


Figure 1. Q-Q plot constructed with the presence of outliers.

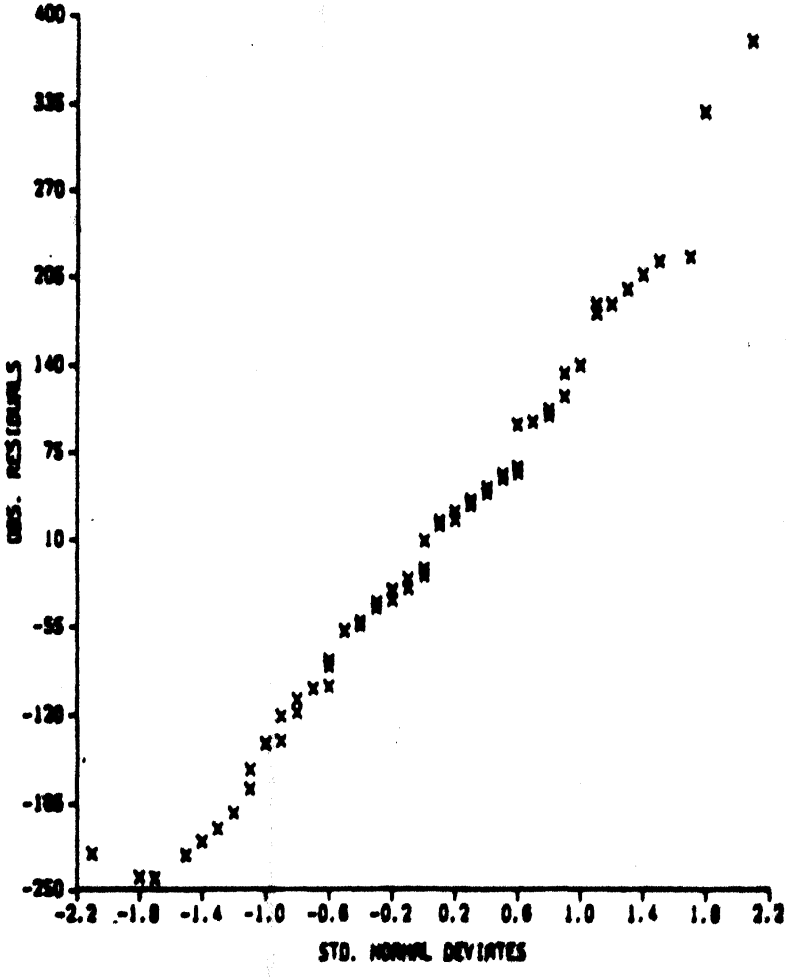


Figure 2. Q-Q plot constructed after ignoring the suspected outlier.

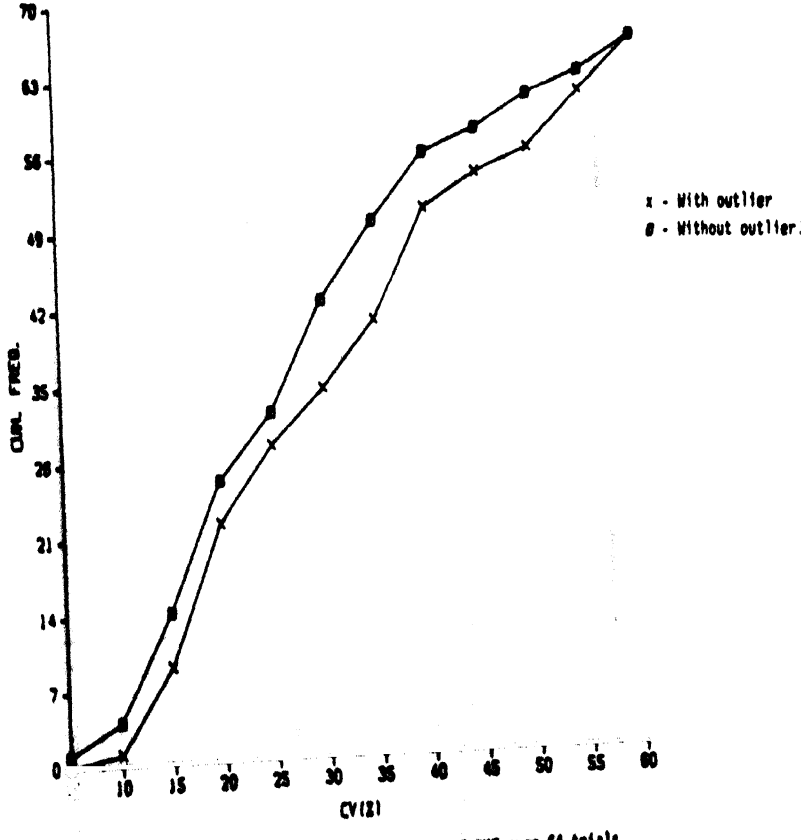


Figure 3. Distribution of CV% over 64 trials.