E

TGP/8.2 Draft 2
ORIGINAL: English
DATE: May 27, 2003

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

Associated Document
to the
General Introduction to the Examination
of Distinctness, Uniformity and Stability and the
Development of Harmonized Descriptions of New Varieties of Plants (document TG/1/3)

DOCUMENT TGP/8

"USE OF STATISTICAL PROCEDURES IN DISTINCTNESS, UNIFORMITY AND STABILITY TESTING"

Section TGP/8.2: Validation of Data and Assumptions

Document prepared by experts from Denmark and the Netherlands

to be considered by the

Technical Working Party on Automation and Computer Programs (TWC), at its twenty-first session to be held in Tjele, Denmark, from June 10 to 13, 2003

SECTION 8.2
VALIDATION OF DATA AND ASSUMPTIONS

8.2.1 Introduction

1. Most often statistical analyses are carried out in order to assist the crop expert when assessing candidate varieties for distinctness, uniformity and stability. In document TGP/8.3, "Experimental Design Practices", aspects of designing the experiments in which the data are recorded are discussed. In document TGP/8.4 "Types of Characteristics and Their Scale Levels", it is shown that the choice of which statistical methods to use, depends on the type of characteristic, its scale level and whether distinctness or uniformity is considered. In document TGP/8.5 "Statistical Methods for DUS Examination", the statistical methods are described. The statistical methods are based on some theory and in order to ensure that the results can be trusted the assumptions behind the theory have to be met - at least approximately. The purpose of this section is to describe the assumptions behind the most common statistical methods used in DUS testing and to show how these assumptions may be validated. It is important to note that the recommended methods for quantitative characteristics (COYD and COYU) are based on variety means per year for COYD, and variety means of the (logarithm of the) between plants standard deviation per year for COYU. Some methods for checking the data are described in 8.2.2 "Check on Data Quality" below. In 8.2.3 "Assumptions", the assumptions underlying the analysis of variance methods are given and in 8.2.4 "Validation", some methods for evaluating these assumptions are given. The assumptions and methods of validation are here described for the analyses of single experiments (randomized blocks). However, the principles are the same when analyzing data from several experiments over years. Instead of plot means, the analyses are then carried out on variety means per year (and blocks then become equivalent to years). The methods described here are intended for quantitative characteristics, but some of the methods may also be used for checking qualitative characteristics on the ordinal scale (pseudo-qualitative characteristics).

8.2.2 Check on data quality (before doing analyses)

2. Visual or automatic inspection of the data for values that are logically inconsistent or in conflict with prior information about the ranges likely to arise for the various characteristics.

3. Examination of frequency distributions of the characteristics to look for small groups of discrepant observations.

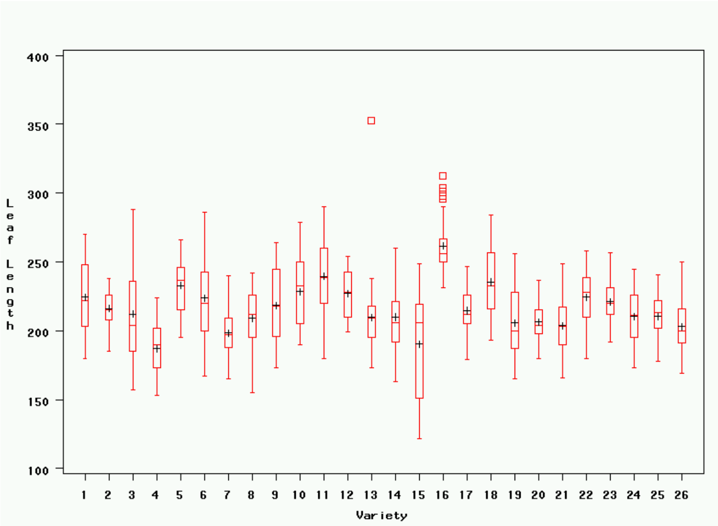

Figure 1. Box-plot for Leaf Length of 26 oil seed rape varieties. There are 60 observations for each variety.

4. Examination of scatter plots of pairs of characteristics likely to be highly related. This may often detect discrepant observations more sensitively than 2.

5. Other types of plot may also be used to validate the quality of the data. A so-called BOX plot is an efficient way to get an overview of the data. In a BOX plot a box is drawn for each group (plot or variety) (Figure 1). The box shows the range for the largest part of the individual observations (usually 75%). A horizontal line through the box and a symbol indicates the median and mean, respectively. At each end of the box, vertical lines are drawn to indicate the range of possible observations outside the box, but within a reasonable distance (usually 1.5 times the height of the box). Finally, observations more extreme than that are shown individually. In Figure 1, it is seen that one observation of variety 13 is clearly much larger than the remaining observations of that variety. Also it is seen that variety 16 has large leaf lengths and that about 4 observations are relatively far from the mean. Among other things that can be seen from the figure are the variability and the symmetry of the distribution. So it can be seen that the variability of variety 15 is relatively large and that the distribution is slightly skewed for this variety (as the mean and median are relatively far apart).

6. When discrepant observations are found, the next step will be to find out why the observations are deviating. In some cases, it may be possible to go back to the field and to check if the plant or plot is damaged by external factors (e.g. rabbits) or a measurement error has occurred. In the last case a correction is possible. In other cases, it may be necessary to look in previous notes (or on other measurements from the same plant/plot) in order to find the reason for the discrepant observation. Generally observations should only be removed when good reasons are present.

8.2.3 Assumptions

7. First of all, it is very important to design experiments in a proper way. The most important assumptions of analysis of variance methods are:

- independent observations
- variance homogeneity
- normally distributed observations (residuals)
- additivity of block and variety effects for a randomized block design and additivity of year and variety effects for COYD.

8. In addition, one could state that there should be no mistakes in the data. However, most mistakes (at least the biggest) will usually also mean that the observations are not normally distributed and that they have different variances.

9. The assumptions mentioned here are most important when the statistical methods are used to test hypotheses. When statistical methods are used only to estimate effects (means), the assumptions are less important and the assumption of normal distributed observations is not necessary.

Independent observations

10. This is a very important assumption. It means that no records may depend on other records in the same analysis (dependence between observations may be built into the model, but this is not so in the COYD and COYU or other UPOV recommended methods). Dependency may be caused e.g. by competitions between neighbouring plots, by lack of randomisation or by improper randomisation. More details on ensuring independence of observations may be found in TGP/8.3 "Experimental Design Practices."

Variance homogeneity

11. Variance homogeneity means that the variance of all observations should be identical apart from random variation. Typical deviations from the assumption of variance homogeneity fall most often into one of the following two groups:

- The variance depends on the mean, this maybe so that the larger the mean value the larger the standard deviation is. In this case, the data may often be transformed such that the variances on the transformed scale may be approximately homogeneous. Some typical transformations are: logarithm for characteristics (where the standard deviation is approximately proportional to the mean), the square-root (where the variance is approximately proportional to the mean, e.g. counts) and the angular transformation (where the variance is low at both ends of the scale and higher in between, typical for percentages).

- The variance depends on e.g. variety, year or block. If the variances depend on such variables in a way that is not connected to the mean value, it is usually not possible to obtain variance homogeneity by transformation. In such cases, it might be necessary either to use more complicated statistical methods that can take unequal variances into account or to exclude the group of observations with deviant variances (if only a few observations have deviant variances). To illustrate the seriousness of variance heterogeneity: imagine a small trial with 10 varieties where varieties A, B, C, D, E, F, G and H each has a variance of 5, whereas varieties I and J each has a variance of 10. The real probability of detecting differences between these varieties when they in fact have the same mean is shown in Table 1. In Table 1, the variety comparisons are based on the pooled variance as is normal in traditional ANOVA. If they are compared using the 1% level of significance, the probability that the two varieties with a variance of 10 become significantly different from each other is almost 5 times larger (4.6%) than it should be. On the other hand, the probability of significant differences between two varieties with a variance of 5 decreases to 0.5%, when it should be 1%. This means that it becomes more difficult to detect differences between two varieties with small variances and easier between varieties with large variances.

Table 1. Real probability of significant difference between two identical varieties in the case where variance homogeniety is assumed but not fulfilled (varieties A to H have a variance of 5 and varieties I and J have a variance of 10.)

Comparisons, variety numbers	Formal test of significance level	
	1%	5%
A and B	0.5%	3.2%
A and I	2.1%	8.0%
I and J	4.6%	12.9%

Normal distributed observations

12. The data should be approximately normally distributed. The ideal normal distribution means that the distribution of the data is symmetric around the mean value and with the characteristic bell-shaped form (see Figure 2). If the data are not approximately normally distributed, the actual level of significance may deviate from the nominal level. The deviation may be in both directions depending on the way the actual distribution of the data deviates from the normal distribution. However, deviation from normality is usually not as serious as deviations from the previous two assumptions.

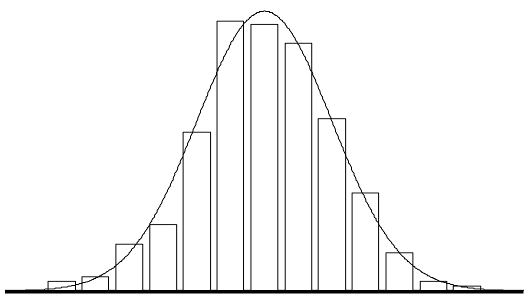

Figure 2. Histogram for normal distributed data with the ideal normal distribution shown as a curve

Additivity of block and variety effects

13. The effects of blocks and varieties are assumed to be additive because the error term is the sum of random variation and the interaction between block and variety. (For a formal description of the model see TGP 8-5 Two-way anova alinea 7). This means that the effect of a given variety is the same in all blocks. This is demonstrated in Table 2 where plot means of artificial data (of plant length in cm) are given for two small experiments with three blocks and four varieties. In experiment I the effect of blocks and varieties are additive because the difference between any two varieties are the same in all blocks, e.g. the difference between variety A and B are 4 cm in all three blocks. In experiment II the effect are not additive, e.g. the difference between variety A and B are, 2, 4 and 6 cm in the three blocks.

Table 2. Artificial plot means of plant length in cm from two experiments showing additive block and variety effects (left) and non-additive block and variety effects (right)

Experiment I			
Variety	Block		
	1	2	3
A	70	72	69
B	74	76	73
C	75	77	74
D	71	73	70

Experiment II			
Variety	Block		
	1	2	3
A	70	72	69
B	72	74	77
C	76	74	73
D	71	72	71

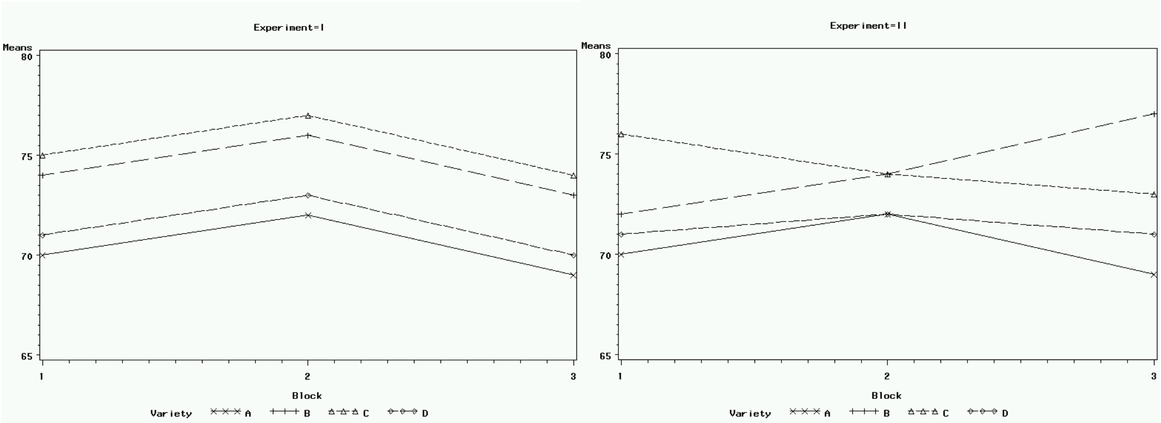

Figure 3. Artificial plot means from two experiments showing additive block and variety effects (left) and non-additive block and variety effects (right) using same data as in table 2

14. In Figure 3 the same data are presented graphically. Plotting the means versus block numbers and joining the observations from the same varieties by straight lines construct the graphs. Plotting the means versus variety names and joining the observations from the same blocks could also have been used (and may be preferred especially if many varieties are to be shown in the same figure). The assumption on additivity is fulfilled if the lines for the varieties are parallel (apart from random variation). As there is just a single data value for each variety in each block, it is not possible to separate interaction effects and random variation. So in practice the situation is not as nice and clear as here because the effects are masked by random variation.

8.2.4 Validation

15. The purpose of validation is partly to check that the data are without mistakes and that the assumptions underlying the statistical analyses are fulfilled.

16. There are different methods to use when validating the data. Some of these are:

- look through the data
- produce plots to verify the assumptions
- make formal statistical tests for the different types of wrong assumptions. In the literature several methods to test for outliers, variance homogeneity and normality may be found. Such methods will not be mentioned here partly because many of these depend on assumptions that do not affect the validity of COYD and COYU seriously and partly because the power of such methods depends heavily on the sample size (this means that serious lack of assumptions may remain undetected in small datasets, whereas small and unimportant deviations may become significant in large datasets)

<u>Looking through the data</u>

17. In practice, this method is only applicable when a few observations have to be checked. For large datasets this method takes too much time, is boring and the risk of overlooking suspicious data increases as one goes through the data. In addition, it is very difficult to judge the distribution of the data and to judge the degree of variance homogeneity when using this method.

Using Figures

18. Different kinds of figures can be prepared which are useful for the different aspects to be validated. Many of these consist of plotting the residuals in different ways. (The residuals are the differences between the observed values and the values predicted by the statistical model).

19. The plot of the residuals versus the predicted values may be used to judge the dependence of the variance on the mean. If not dependant, then the observations should fall approximately (without systematic deviation) in a horizontal band symmetric around zero (Figure 4). In cases where the variance increases with the mean, the observations will fall approximately in a funnel with the narrow end pointing to the left. Outlying observations, which may be errors, will be shown in such a figure as observations that clearly have escaped from the horizontal band formed by most other observations. In the example used, no observations seem to be outliers (the value at the one bottom left corner where the residual is about -40 mm may at first glance look so, but several observations have positive values of the same numerical size). Here it is important to note that an outlier is not necessarily an error and also that an error will not necessarily show up as an outlier.

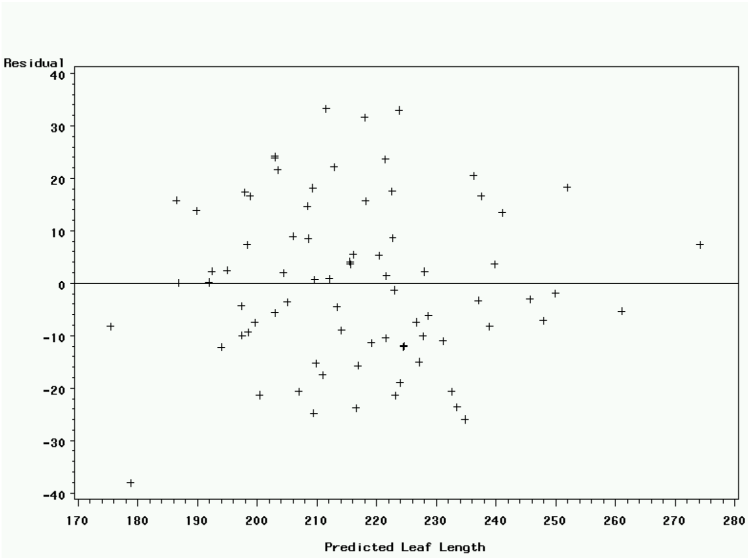

Figure 4. Plot of residuals versus plot predicted values for Leaf Length in 26 oil seed rape varieties in 3 blocks

20. The residuals can also be used to form a histogram, like Figure 2, from which the assumption about the distribution can be judged.

21. The range (max value minus min value) or standard deviation for each plot may be plotted versus some other variables such as the plot means, variety number or plot number. Such figures (Figure 5) may be useful to find plots were the variation is extremely large (which may be caused by a single plant) or to judge whether min-max range depends on something. It is clearly seen that the min-max range for one of variety 13's plot is much higher than in the other two plots. Also the min-max range in one of variety 3's plot seems to be relatively large.

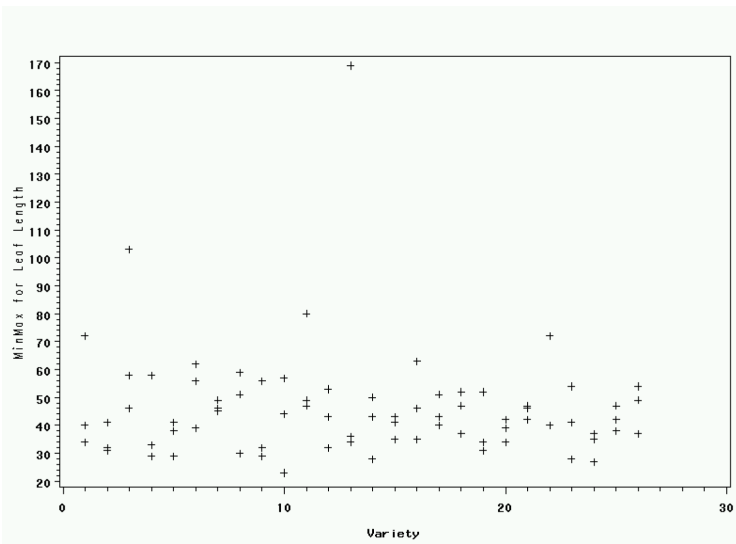

Figure 5. Differences between minimum and maximum of 20 leaf lengths for 3 plots of 26 oil seed rape varieties

22. A figure with the plot means (or variety adjusted means) versus the plot number can be used to find out whether the characteristic depends on the location in the field (Figure 6). This, of course, requires that the plots be numbered such that the numbers indicate the relative location. In the example shown here, there is a clear trend showing that the leaf length decreases slightly with plot number. However most of the trend over the area used for the trial will - in this case - be explained by differences between blocks (plot 1-26 is block 1, plot 27-52 is block 2 and plot 53-78 is block 3).

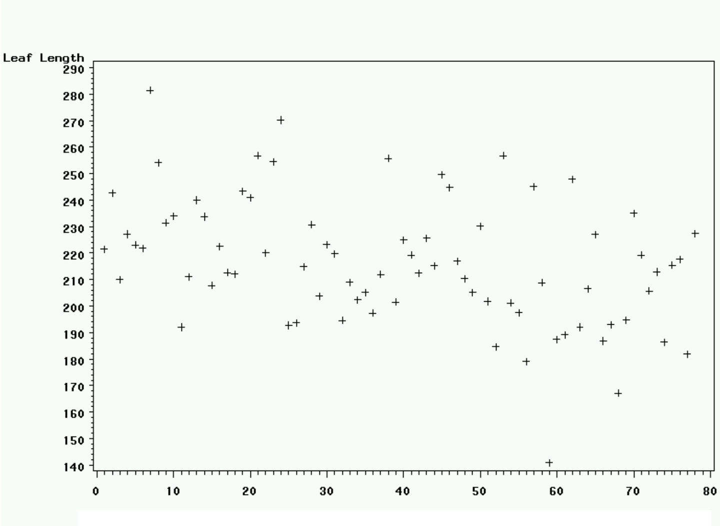

Figure 6. Plot means shown against plot numbers of 20 leaf lengths for 3 plots of 26 oil seed rape varieties

23. The plot means can also be used to form a figure where the additivity of block and variety effects can be visually checked at (see Figure 3).

24. Normal Probability Plots (Figure 7). This type of graph is used to evaluate the normality of the distribution of a variable, that is, whether and to what extent the distribution of the variable follows the normal distribution. The selected variable will be plotted in a scatterplot against the values "expected from the normal distribution." The standard normal probability plot is constructed as follows. First, the deviations from the mean (residuals) are rank ordered. From these ranks the program computes the expected values from the normal distribution, hereafter called z-values. These z-values are plotted on the X-axis in the plot. If the observed residuals (plotted on the Y-axis) are normally distributed, then all values should fall onto a straight line. If the residuals are not normally distributed, then they will deviate from the line. Outliers may also become evident in this plot. If there is a general lack of fit, and the data seem to form a clear pattern (e.g., an S shape) around the line, then the variable may have to be transformed in some way.

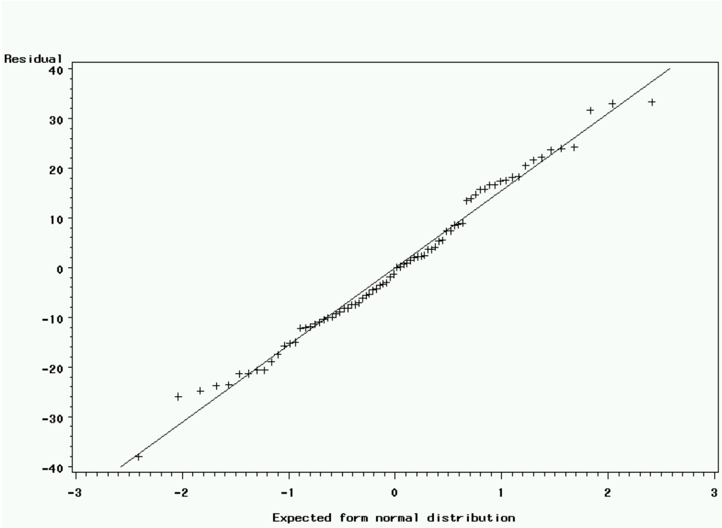

Figure 7. Normal probability plot for the residuals of Leaf Length in 26 oil seed rape varieties in 3 blocks

[End of document]