E

TGP/8.3 Draft 1
ORIGINAL: English
DATE: May 24, 2002

UPOV

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/8.3: Experimental Design Practices

Document prepared by experts from Germany and the Netherlands

to be considered by the

Technical Working Party on Automation and Computer Programs (TWC), at its twentieth session to be held in Texcoco, Mexico, from June 17 to 20, 2002

Technical Working Party for Vegetables (TWV), at its thirty-sixth session to be held in Tsukuba, Japan, from September 9 to 13, 2002

Technical Working Party for Agricultural Crops (TWA), at its thirty-first session to be held in Rio de Janeiro, Brazil, from September 23 to 27, 2002

Technical Working Party for Ornamental Plants and Forest Trees (TWO), at its thirty-fifth session to be held in Quito, from November 18 to 22, 2002

Technical Working Party for Fruit Crops (TWF), at its thirty-third session to be held in San Carlos de Bariloche, Argentina, from November 25 to 29, 2002

SECTION 8.3

EXPERIMENTAL DESIGN PRACTICES

8.3.1 Introduction

1. As stated in TGP/8.1 “Introduction”, DUS trials are experiments with usually two goals: the comparison of varieties and the absolute determination of the characteristics. In these experiments a certain number of characteristics of the plants are observed in order to assess Distinctness, Uniformity and Stability (DUS). The measurements or observations of the characteristics (see ‘TGP 8.4 “Type of characteristics and their scale levels”) are analyzed and using the results of the analyses (see TGP/8.5 “Statistical Methods”) statements are made about DUS. This section addresses a number of issues concerning the basics of statistics and experimental design.

8.3.2 Elementary statistical notions

Population

2. A population is a collection of units about which statements will be made. In DUS trials a population consists of plots of one or more plants of the same variety (or candidate variety) that could have been planted in the experiment(s).

Experimental units

3. An experimental unit in DUS trials is a plot with one or more plants. It is the smallest subdivision of the field about which the varieties are randomized. If there are more plants within a plot, the observations of a certain characteristic at each plant are used for estimating the variability of the variety. The mean (or other function) of the observations can be considered as the plot measurement for that characteristic.

Types of characteristics and their scale levels

4. The characteristics (as described in the Test Guidelines) can be divided into quantitative, qualitative and pseudo-qualitative ones. The choice of procedure for the assessment of distinctness, uniformity and stability depends on the scale level of the data which are recorded for a characteristic. The scale level of data depends on the type of expression of the characteristic and on the method of recording this expression. The type of scale may be quantitative or qualitative. Quantitative data can have a continuous or a discrete distribution. Qualitative data are data which can be arranged in discrete qualitative different categories. Usually they result from visual assessment. Subgroups of qualitative scales are ordinal and nominal scales (see also for TGP/8.4 “Types of Characteristics and Their Scale Levels”).

Distribution of data for evaluation of characteristics

To describe the variation of the data of a characteristic within a variety a frequency distribution or histogram can be used. This distribution describes for each possible expression of the characteristic the number of plots in the population with this value. Instead of a frequency distribution a relative frequency distribution can be calculated by dividing the numbers by the total numbers. This gives fractions instead of numbers.

5. In order to make a histogram, for quantitatively scaled characteristics, the expressions of the characteristic are divided into different classes. If the number of classes and of plots is large enough the (relative) frequency distribution can be approximated by a continuous function. In practice, for continuous quantitatively scaled characteristics, this distribution is very often the Normal distribution. For discrete quantitatively scaled characteristics (whose values result from counting) the Poisson or Binomial distribution can be used most of the time. When the number of plots is large enough and the mean is not too small the Normal distribution can also be considered.

Population mean, variance and standard deviation

6. For **quantitatively** scaled characteristics it is often sufficient to give some parameters of the distribution instead of specifying the whole distribution. Usually parameters for location and dispersion of the distribution are sufficient.

7. The parameter mostly used for location is the mean. The mean is the sum of all expressions of the characteristic in the population divided by the total number of plots.

8. The parameter most used for dispersion is the variance. The variance is the sum of all squared differences between the expressions of the characteristic and the mean divided by the total number of plots. The square root of the variance is called the standard deviation, which is expressed in the same unit as the measurement.

9. For **qualitatively** scaled characteristics it is not possible to describe the frequency distribution with a limited number of parameters.

Samples

10. In the preceding sections we have discussed the way a population can be described if the expressions of the characteristics of the total number of plots in the population are known. In practice it is almost never the case. Only a limited number of plots (a so-called sample) are investigated and the expressions of the characteristics of these plots are used to make conclusions about the whole population, for instance about the population mean.

Randomization

11. To be able to extrapolate the results of the sample to the population the first requirement is for the sample to be representative. This can be achieved by drawing the sample at random. This means that each plot of the population has an equal chance of being selected in the sample. In DUS trials randomization is used for allocating varieties/treatments to plots.

Stochastic or random variable

12. If a plot is selected from the population for the sample and the expression of the characteristic is determined, the result is called a stochastic or random variable. This is because it is unknown which plot is selected and therefore the value which is obtained. Such a random variable has a probability distribution, which is the same as the relative frequency distribution in the population.

Sample mean, variance and standard deviation

13. If the sample includes, say **n**, plots, the mean $\bar{y}$ of the expressions of the characteristic y can be calculated as:

$$\bar{y} = \frac{\sum_i y_i}{n}$$

14. This mean is an estimate of the mean in the whole population (μ), and is also a random variable with its own probability distribution. It can be proven that the variance of the mean equals the variance of the distribution of the characteristic observed on a plot divided by **n**, the number of plots in the sample.

15. The population variance (σ^2) can be estimated by the sample variance s^2:

$$s^2 = \frac{1}{n-1} \sum_i (y_i - \bar{y})^2 ,$$

where y_i is the expression of the characteristic of plot i and $\bar{y}$ the mean of all values y_i. The standard deviation is the square root of the variance.

Precision of the sample mean

16. As stated in 'Sample mean, variance and standard deviation' the variance of the mean equals the variance between plots divided by the number, **n**, of plots in the sample, so that the standard error of the mean (**sem**) equals the standard deviation divided by the square root of n:

$$\text{sem} = \frac{s}{\sqrt{n}}$$

17. As can be seen from the formula: the larger the number of plots the smaller the standard error of the mean.

Confidence interval

18. Besides giving the sample mean and its standard error it is also convenient to give a confidence interval for the population mean. The width of the interval depends on the choice

of the uncertainty level α, so the next interval is called a two-sided (1-α)100-% confidence interval:

$$(\bar{y} - \text{sem} * t_{n-1;1-\alpha/2},\ \bar{y} + \text{sem} * t_{n-1;1-\alpha/2}),$$

where $t_{n-1;1-\alpha/2}$ is the (1- α/2)100 percentage point of Student's t-distribution with n-1 degrees of freedom.

Comparison of two varieties

19. For quantitatively scaled, normally distributed characteristics the means of two varieties can be calculated. It can be proven that the difference of the two means is also normally distributed with a mean equal to the real difference between the two means and a standard deviation **sed**, the standard error of the difference of two means:

$$\text{sed} = s_{pooled} * \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where s is the pooled (residual) standard deviation and n_1 and n_2 are the number of plots of the two varieties.

20. The pooled (residual) standard deviation is the square root of the weighted mean of the 2 sample variances $s1^2$ and $s2^2$:

$$s_{pooled} = \sqrt{\frac{(n1-1) * s1^2 + (n2-1) * s2^2}{n1 + n2 - 2}}$$

If n1 and n2 are equal s_{pooled} is the square root of the mean of the 2 variances.

21. A two-sided (1-α)100-% confidence interval for the difference of two variety means is:

$$(\bar{y}_1 - \bar{y}_2 - \text{sed} * t_{n_1+n_2-1;1-\alpha/2},\ \bar{y}_1 - \bar{y}_2 + \text{sed} * t_{n_1+n_2-1;1-\alpha/2}).$$

22. The quantity **lsd**, the least significant difference between two means, is defined as:

$$\text{lsd} = \text{sed} * t_{n_1+n_2-1;1-\alpha/2}$$

If the absolute difference is greater than the lsd, the two means are said to be different.

Hypothesis testing

23. To decide whether two varieties are different, for example with respect to the means of a certain characteristic, a statistical test can be performed. Before discussing the procedure we will introduce some terminology.

24. By performing a test we have to consider two hypotheses:
- the **null hypothesis H0**, which states that both variety means are the same, and
- the **alternative hypothesis H1**, that they are different.
-

25. By using **a test statistic**, which is a formula of the observations of the plots of both varieties, we must decide to accept the null hypothesis H0 (and thus reject the alternative hypothesis H1) or vice versa. The decision to reject H0 occurs if the test statistic is greater than the (chosen) **critical value**, otherwise H0 is accepted. If H0 is rejected the test is called **significant**.

26. The correctness of the decision depends on the real difference of the two varieties (μ_1-μ_2). We can sketch the four possibilities in the next scheme:

		Real situation:	
		H0 true : $\mu_1 = \mu_2$	H1 true : $\mu_1 \neq \mu_2$
Decision:	H0 accept	Correct decision	**Type II error (β)**
	H0 reject	**Type I error (α)**	Correct decision

27. Firstly, the chance **α** of making a Type I error should not be too large. This can be achieved by the choice of the critical value (see above). The chance **α** is called the **size** of the test as well the nominal significance level.

28. Secondly the chance **β** of making a Type II error depends on the real difference between μ_1 and μ_2.

29. The greater the difference the smaller the chance **β** is. Furthermore, given a fixed difference between the two variety means the chance **β** is smaller with a greater number of observations. The chance **1 - β** of correctly rejecting H0 is called the power of the test.
There is also a relation between the chances **α** and **β**:

given a fixed number of observations the chance **β** increases as the chance **α** is chosen lower and the chance **β** decreases as the chance α is chosen higher.

30. For quantitatively scaled, normally distributed characteristics the most commonly used test statistic is Student's t-statistic t_ν :

$$t_\nu = \frac{\bar{y}_1 - \bar{y}_2}{\text{sed}}$$, which has a Student's t distribution with ν degrees of freedom.

31. Calculating this statistic and choosing the critical value (by means of choosing α) results in the decision of accepting or rejecting H0. If t_ν is greater than the critical value $t_{\nu;1-\alpha/2}$ H0 is rejected, which means that the two varieties are declared different with respect to this characteristic.

Significance

32. When performing a test the decision is acceptance or rejection of the null hypothesis. If the null hypothesis is rejected, the test is called **significant**. As you can see in 'Hypothesis testing' the result of the test depends on the choice of the chance **α**. So by speaking of a significant result in testing the choice of **α** must always be mentioned.

33. A more informative way of presenting the result of a test is to state the real **P-value**: the fraction of the distribution of the test statistic t_ν under H0 that is greater than the absolute value of t_ν.

8.3.3 Principles of experimental design

Experimental units and variance components

34. An experimental unit in a DUS trial corresponds to the smallest subdivision of the field so that different units may comprise different varieties i.e. each unit contains only one variety. These units are often called "plots". One of the most important requirements of experimental units is **independence**. This means that observations within one plot are not influenced by the varieties in other plots. For example if big varieties are planted next to a small one there could be a negative influence of the big ones to the small one. In such a case a row of plants on both sides of the plot can be planted in order to avoid dependency.

35. When the same variety is assigned to a number of different plots and there is only one observation for each plot, the observations in the different plots may vary. The variation between these observations will be called the 'between-plot variability'. This variability is a mixture of a number of sources of variation: different plots, different plants, different times of observation, different errors of measurement and so on. It is not possible to distinguish between these sources of variation.

36. When there are observations of more than one, say n, plants per plot it is possible to compute two variance components: the within-plot or plant component and the plot component. After calculating the mean of the n observations per plot the standard error of such a plot mean is:

$$sem_{plot} = \sqrt{s^2_{plot} + (s^2_{plant} / n)}$$

This sem_{plot} is the 'between-plot' standard deviation.

If there are **r** replicates of each variety in the experiment, the **sed** between 2 variety means is:

$$sed = \sqrt{2 * \left(\frac{s^2_{plot}}{r} + \frac{s^2_{plant}}{r * n} \right)}$$

Randomization

37. In designing an experiment to compare a number of varieties with each other it is important to randomize the varieties over the plots. If the varieties are arranged systematically in the field, the following can happen:

A	B	C	D	A	B	C	D
A	B	C	D	A	B	C	D

high fertility ----------> low fertility

38. If the fertility of the soil decreases from the left to the right of the field, the plants of variety A have grown on more fertile plots than the other 3 varieties B, C and D have. The same is true for variety B in comparison with varieties C and D, and for variety C in comparison with D.

39. Briefly stated: the comparison of the varieties is influenced by a difference in fertility of the plots. In this example fertility is taken as a systematic effect which can be replaced by any other systematic effect.

40. To avoid such a systematic error it is advisable to randomize the varieties across the site:

A	C	C	B	D	B	C	B
D	A	B	A	A	C	D	D

Blocking

41. Blocking will be introduced here by means of differences in fertility. As stated further on several other systematic sources of variation could have been used for blocking. To be able to compare the varieties, the standard deviations of the replicates of each variety are used. If there are for example great fertility differences between the plots and this fertility has a great influence on the expressions of the characteristic being observed, the standard deviations may become too large to compare the varieties with enough power. In this case it is beneficial to arrange the plots in blocks. The blocks are arranged in such a way that the plots within blocks have comparable fertilities and the fertility differences are between the blocks.

42. After blocking and randomization the lay-out could be:

Block	I		II		III		IV	
	B	C	C	D	D	B	C	A
	D	A	B	A	A	C	B	D

43. To eliminate the variation because of the fertility differences from one side of the block to another from the standard deviation it would be even better to arrange the plots in the following way:

Block	I	II	III	IV
	B	A	B	D
	C	C	A	B
	D	B	D	C
	A	D	C	A

44. Although it is not always clear how heterogeneous the field is, and therefore it is unknown how to arrange the blocks, it is usually a good idea to create blocks for many other reasons. When there are different sowing machines, different observers, different observation days, all these effects are included in the residual standard deviation if they are randomly assigned to the plots. However, these effects can be eliminated from the residual standard deviation if all the plots within each block will have the same sowing machine, the same observer, the same observation day, and so on.

Types of experimental design

45. Because of the presence of only one treatment factor (variety) in DUS trials (per year, per site) simple designs can be used. The next 3 types of experimental design are most used:

- Completely randomized design

46. When several varieties are examined in a number of replications the varieties can be completely randomized over all plots in the field as shown in the last table of the Randomization section. This design is only recommended if the total number of plots is small.

- *Randomized complete block design*

47. In a randomized, complete block design the number of plots per block equals the number of varieties and all varieties are present in each block. The advantage of a randomized complete block design is that the standard deviation between plots does not contain variation due to differences in blocks. As we have seen in section Blocking blocks can be used to control a number of sources of variation. Because there is usually no cost associated with blocking so it is better to block.

- *Randomized incomplete block design, for example Alpha design*

48. In a randomized incomplete block design the number of plots per block is less than the number of varieties. Such a design may be appropriate when the number of varieties is very large. In this case the block size for a randomized complete block design would become so large that the plots within a block would be too heterogeneous. There are many types of incomplete block designs: the most familiar are balanced incomplete block designs, partially balanced incomplete block designs and alpha designs. In a balanced incomplete block design the standard errors of differences are equal for each pair of varieties whereas in partially balanced incomplete block designs and alpha designs they are not.

49. The main advantage of alpha designs is flexibility: they are available whenever the number of varieties is a multiple of the block size, and they can be easily adapted even when it is not. An alpha design has the extra property that groups of the incomplete blocks can be arranged into complete blocks so that all varieties are present in the complete block.

50. As often as possible it is recommended to randomize plots according to the 3 types of experimental design above. Nevertheless in practice the DUS expert wishes to plant varieties that are difficult to distinguish side by side in order to be able to compare them directly at different stages of growth. This practice should be limited to specific difficult cases and should not considered as good practice per se as it can induce bias in the decisions. Especially in this case it is very important that the DUS expert doesn't have any idea about the coding of the varieties.

8.3.4 General remarks

Trial elements

51. The plot is the smallest subdivision of the trial and the unit on which the varieties and the soil and plant conditions should be focussed. Therefore the trial elements:

- plot size
- shape of the plots
- alignment of the plots
- barrier rows and border strips
- protective strips

should be arranged accordingly.

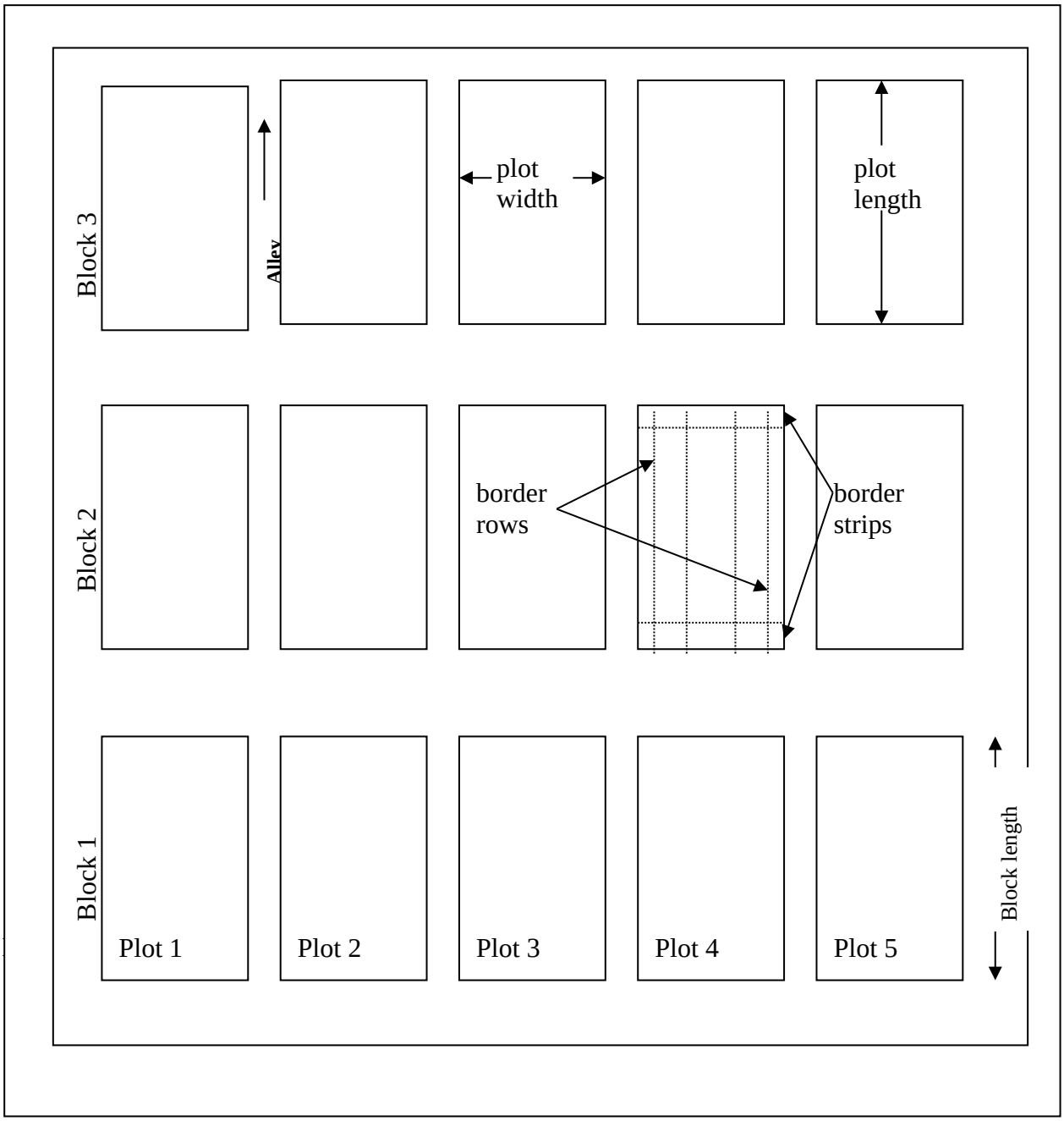
52. The following Figure 1 may be helpful to give some explanations of the particular trial elements.

53. In general there are four classes of methods of observations that are used in the newest Test Guidelines:

MG: measurement of a group of plants or part of plants
MS: measurement of a number of individual plants or parts of plants
VG: visual assessment of a group of plants or parts of plants
VS: visual assessment of a number of individual plants or parts of plants

54. The highest requirements on planning of the trial are based on characteristics that are observed on individual single plants (MS and VS). These characteristics determine the number of single plants and therefore the size of the plot. But some other factors have also an influence on the plot size: in some cases it is necessary to have border rows and strips to minimise the interplot interference (interaction of genotypes) and other special border effects.

Figure 1: Trial elements in a trial layout



Plot size and shape of the plot

55. Plot size and the shape of the plots depend also on the soil conditions and on sowing and harvesting machinery. The shape of the plot can be defined as the ratio of plot length divided by plot width. This ratio can be important for compensation of the difference in soil within the blocks.

56. Plots with square shape have the smallest total length of the borders (circumference). From the theoretical point of view the quadratic shape is optimal to minimise the interaction of genotypes. Grouping of the varieties can have the same effect.

57. Narrow and long plots are preferred from the technological point of view. The best length to width ratio lies between 5:1 and 15:1 and depends on the plot size and the number of the varieties. The larger the number of varieties in the trial the narrower the plots.

58. The aim of DUS testing is to get averages of characteristics for each variety and to judge the within variety variability by calculating the standard deviation. The averages will be used for determining the distinctness of the varieties; the standard deviations are the basis for examination of uniformity in the case of quantitative characteristics. For qualitative characteristics the number of off-types will be determined. Off-types are plants that are different from the other ones of the same variety.

59. For distinctness unbiased and precise estimation of averages is necessary. The bias is difficult to calculate. Nevertheless it is common to reduce the bias by suitable precautions which are the exclusion of external influences by lay down of protection strips on the border of the trial. Additionally often it is necessary to exclude border rows and strips of the plot from average calculation. The rest of the plot without border rows and strips (effective plot size) are the basis of unbiased and precise averages.

60. The plot size depends also upon the sample size.

Sample size

The recommendation to put 60 plants (3 times 20) into a DUS trial is not a general one.
The question is not:
What is the optimal sample size in DUS testing?
The question is:
What is the optimal sample size in DUS testing for a specific crop over all characteristics?

61. There are three stages of precision in DUS trials:

- precision from individual single plants (within a plot)
- precision from replications (over the plots)
- precision from years or cycles (over the years or cycles)

62. The survey of uniformity of a variety needs all data from individual single plants of the plot. The procedures for distinctness use the averages of the variety for each year or cycle only.

63. At first the crop expert has to know what are the characteristics with the highest variation within the variety over the plot and/or over the years or cycles.

Quantitative characteristics:

Stage 1: Precision from individual single plants (within a plot)

not yet ready

Stage 2: Precision from replications (over the plots)

The number of replications or blocks (n) can be calculated from the following formula:

$$n = 2\, s^2 / d^2 * (t_{df,1-\alpha/2} + t_{df,1-\beta})^2$$

where:

- s is the 'between-plot' standard deviation sem_{plot}
- d is the given difference between the variety means which has to be significant
- t is the percentage point of Student's t-distribution
- df are the degrees of freedom in correspondence to the 'between-plot' standard deviation (df=(a-1)*(n-1) for a one factorial block design)
- a is the number of varieties
- n is the number of blocks
- α is the type I error
- β is the type II error

Stage 3: Precision from years or cycles (over the years or cycles)

not yet ready

Here I will include an example for calculation the sample size from the uniformity point of view (ANOVA, comparison of variances).

Qualitative characteristics:

64. For qualitative characteristics distinctness procedures are not the basis to determine the optimal sample size up to now.

65. From the uniformity point of view the optimal sample size can be calculated. The number of accepted off-types depends on:

- the population standard (P%)
- the total number of plants (n)
- the acceptance probability (100-α)
- the type I error (α)
- the type II error (β for 2P, 5P or 10P)

66. All these elements are coherent by the binomial distribution (see TWC/11/16). The whole number of plants and therefore the plot size depend on all the other elements. In other words: It is possible to determine the whole number of plants and therefore the plot size if all the other elements are known.

67. At the end the maximum of determined sample sizes is from the statistical point of view the optimal one.

68. Conclusion: It is difficult to determine the optimal sample size. Per characteristic, per stage and per type of testing it is possible to give formulas for calculating the number of plants or the number of plots, but it is not clear how to combine all these individual calculations. Another difficulty is that the crop expert has not enough information about the varieties when he starts to work with a new crop and when he has to establish a new Test guideline. Especially the standard deviation, the difference between the varieties and the

population standard are unknown at this time. So in this case the crop expert has to look to other similar crops and he can plan the trial as for a similar crop. Later the crop expert will get more experience with the new crop and he can use standard deviations and differences between the averages for the evaluations and he can define an own population standard if necessary and then he is able to calculate the whole number of plants.

Analysis over years or cycles

69. The comparison between candidate and reference varieties is mostly based on observations from 1 to 3 years or cycles. Therefore, the number of replicates and the number of plants per plot in a single trial have an indirect effect on the variability which is used in the COYD and COYU analyses. Before performing these analyses the means of the variety means and (log) standard deviations per year or cycle are calculated and then the analysis is performed on these means in the two-way variety by year or cycle layout. The residual variation in these analyses is the variety by year or cycle interaction. More refined techniques based in REML can be used, which allow for, e.g., between-trial heterogeneity in error variance.

70. The precision of the variety means in one experiment, when used for COYD for example, is only used indirectly, because the standard deviation in that analysis is the interaction between the varieties and the years or cycles. If the differences between the varieties over the years or cycles are very different, the precision of the means per experiment are less important.

[End of document]