UPOV
International Union for the Protection of New Varieties of Plants

E

TGP/8/4

Original: English
Date: November 1, 2019

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

TRIAL DESIGN AND TECHNIQUES USED IN THE EXAMINATION OF DISTINCTNESS, UNIFORMITY AND STABILITY

adopted by the Council
at its fifty-third ordinary session
on November 1, 2019

TABLE OF CONTENTS PAGE

INTRODUCTION **5**

PART I: DUS TRIAL DESIGN AND DATA ANALYSIS **6**

1. DUS TRIAL DESIGN 6
1.1 Introduction *6*
1.2 Growing cycles *6*
1.2.1 Introduction *6*
1.2.2 Independent growing cycles *7*
1.3 Testing Place *7*
1.3.1 Purpose *7*
(a) Minimizing the overall testing period 7
(b) Reserve Trial 7
(c) Different agro-climatic conditions 8
1.3.2 Use of information from multiple locations *8*
(a) Additional tests 8
(b) DUS examined on the basis of data for the same characteristics examined at different locations 8
1.4 Conditions for conducting the examination *8*
1.5 Test Design *8*
1.5.1 Introduction *8*
1.5.2 Number of Plants in the trial *9*
1.5.3 Trial layout *9*
1.5.3.1 Introduction 9
1.5.3.2 Single plots 11
1.5.3.3 Replicate plots (statistical analysis) 11
1.5.3.3.1 Introduction 11
1.5.3.3.2 Replicate plots for statistical analysis of individual plant data 11
1.5.3.3.3 Randomization 11
1.5.3.3.4 Randomized incomplete block designs 14
1.5.3.3.5 Design for pair-wise comparisons between particular varieties 15
1.5.3.3.6 Further statistical aspects of trial design 15
1.5.3.3.6.1 Introduction 15
1.5.3.3.6.2 The hypotheses under test 15
1.5.3.3.6.3 Determining optimal sample size 17
1.5.3.3.7 Trial elements when statistical analysis is used 17
1.5.3.3.7.1 Introduction 17
1.5.3.3.7.2 Plots and blocks 18
1.5.3.3.7.3 Allocation of varieties to plots 18
1.5.3.3.7.4 Plot size, shape and configuration 19
1.5.3.3.7.5 Independence of plots 19
1.5.3.3.7.6 The arrangement of the plants within the plot/ Type of plot for observation 19
1.5.3.4 Blind Randomized Trials 19
1.6 Changing Methods *20*
2. DATA TO BE RECORDED 21
2.1 Introduction *21*
2.2 Types of expression of characteristics *21*
2.3 Types of scales of data *22*
2.3.1 Data from qualitative characteristics *22*
2.3.2 Data from quantitative characteristics *22*
(a) Ratio scale 23
(b) Interval scale 23
(c) Ordinal scale 24
2.3.3 Data from pseudo-qualitative characteristics *24*
2.3.4 Summary of the different types of scales *25*
2.3.5 Scale levels for variety description *25*
2.3.6 Relation between types of expression of characteristics and scale levels of data *25*
2.3.7 Relation between method of observation of characteristics, scale levels of data and recommended statistical procedures *26*
2.4 Different levels to look at a characteristic *28*
3. MINIMIZING THE VARIATION DUE TO DIFFERENT OBSERVERS OF THE SAME TRIAL 32
3.1 Introduction *32*
3.2 Training and importance of clear explanations of characteristics and method of observation *32*
3.3 Testing the calibration *32*
3.4 Testing the calibration for QN/MG or QN/MS characteristics *32*
3.5 Testing the calibration for QN/VS or QN/VG characteristics *33*
3.6 Trial design *33*
3.7 Example of Cohen's Kappa *34*

3.8 References ... *35*
4. VALIDATION OF DATA AND ASSUMPTIONS ... 36
4.1 Introduction ... *36*
4.2 Validation of data ... *36*
4.3 Assumptions for statistical analysis and the validation of these assumptions ... *37*
4.3.1 Assumptions for statistical analysis involving analysis of variance ... *38*
4.3.1.1 Introduction ... 38
4.3.1.2 Independent observations ... 38
4.3.1.3 Variance homogeneity ... 38
4.3.1.4 Normal distributed observations ... 39
4.3.1.5 Additivity of block and variety effects ... 39
4.3.2 Validation of assumptions for statistical analysis ... *40*
4.3.2.1 Introduction ... 40
4.3.2.2 Looking through the data ... 40
4.3.2.3 Using figures ... 41
5. CHOICE OF STATISTICAL METHODS FOR EXAMINING DISTINCTNESS ... 45
5.1 Introduction ... *45*
5.2 Statistical methods for use with two or more independent growing cycles ... *45*
5.2.1 Introduction ... *45*
5.3 Summary of selected statistical methods for examining distinctness ... *47*
5.4 Requirements for statistical methods for distinctness assessment ... *48*
6. CYCLIC PLANTING OF VARIETIES FROM THE VARIETY COLLECTION TO REDUCE TRIAL SIZE ... 49
6.1 Summary of requirements for application of method ... *49*
6.2 Summary ... *49*
6.3 Cyclic Planting of Established Varieties in Trial ... *49*
6.3.1 The assessment of distinctness by data compensation ... *50*
6.3.2 Method of analysis for distinctness assessment ... *50*
6.3.3 The assessment of uniformity ... *51*
6.4 Comparison of the cyclic planting system with the existing system ... *51*
6.5 Cyclic planting system software ... *51*
6.6 Additional technical detail and example of analysis for distinctness assessment ... *51*
6.6.1 Example of distinctness assessment ... *52*
6.7 References ... *52*

PART II: SELECTED TECHNIQUES USED IN DUS EXAMINATION ... 54

1. THE GAIA METHODOLOGY ... 54
1.1 Some reasons to sum and weight observed differences ... *54*
1.2 Computing GAIA phenotypic distance ... *54*
1.3 Detailed information on the GAIA methodology ... *55*
1.3.1 Weighting of characteristics ... *55*
1.3.2 Examples of use ... *56*
1.3.2.1 Determining “Distinctness Plus” ... 56
1.3.2.2 Other examples of use ... 56
1.3.3 Computing GAIA phenotypic distance ... *57*
1.3.4 GAIA software ... *58*
1.3.5 Example with Zea mays data ... *59*
1.3.5.1 Introduction ... 59
1.3.5.2 Analysis of notes ... 59
1.3.5.3 Electrophoresis analysis ... 60
1.3.5.4 Analysis of measurements ... 62
1.3.5.5 Measurements and 1 to 9 scale on the same characteristic ... 63
1.3.6 Example of GAIA screen copy ... *64*
2. PARENT FORMULA OF HYBRID VARIETIES ... 67
2.1 Introduction ... *67*
2.2 Requirements of the method ... *67*
2.3 Assessing the originality of a new parent line ... *67*
2.4 Verification of the formula ... *68*
2.5 Uniformity and stability of parent lines ... *68*
2.6 Description of the hybrid ... *68*
3. THE COMBINED OVER-YEARS CRITERIA FOR DISTINCTNESS (COYD) ... 69
3.1 Summary of requirements for application of method ... *69*
3.2 Summary ... *69*
3.3 Introduction ... *69*
3.4 The COYD method ... *70*
3.5 Use of COYD ... *70*
3.6 Adapting COYD to special circumstances ... *71*
3.6.1 Differences between years in the range of expression of a characteristic ... *71*
3.6.2 Small numbers of varieties in trials: Long-Term COYD ... *72*
3.6.3 Crops with grouping characteristics ... *72*

3.7 Implementing COYD ... 73
3.8 References ... 73
3.9 COYD statistical methods ... 76
3.9.1 Analysis of variance ... 76
3.9.2 Modified joint regression analysis (MJRA) ... 76
3.9.3 Comparison of COYD with other criteria ... 77
3.10 COYD software ... 77
3.11 Schemes used for the application of COYD ... 84
4. 2X1% METHOD ... 87
4.1 Requirements for application of method ... 87
4.2 The 2x1% Criterion (Method) ... 87
5. PEARSON'S CHI-SQUARE TEST APPLIED TO CONTINGENCY TABLES ... 89
6. FISHER'S EXACT TEST ... 92
6.1 Assessment of Distinctness ... 92
7. MATCH APPROACH ... 94
7.1 Requirements for application of method ... 94
7.2 Match Method ... 94
8. THE METHOD OF UNIFORMITY ASSESSMENT ON THE BASIS OF OFF-TYPES ... 95
8.1 Fixed Population Standard ... 95
8.1.1 Introduction ... 95
8.1.2 Using the approach to assess uniformity in a crop ... 95
8.1.3 Issues to be considered when deciding on the use of the method ... 96
8.1.4 Examples ... 97
8.1.5 Introduction to the tables and figures ... 100
8.1.6 Method for one single test ... 101
8.1.7 Method for more than one single test (year) ... 102
8.1.8 Note on balancing the Type I and Type II errors ... 102
8.1.9 Definition of statistical terms and symbols ... 102
8.1.10 Tables and figures ... 103
9. THE COMBINED-OVER-YEARS UNIFORMITY CRITERION (COYU) ... 110
9.1 Summary of requirements for application of method ... 110
9.2 Summary ... 110
9.3 Introduction ... 111
9.4 The COYU Criterion ... 111
9.5 Use of COYU ... 111
9.6 Mathematical details ... 112
9.7 Early decisions for a three-year test ... 114
9.8 Example of COYU calculations ... 115
9.9 Implementing COYU ... 116
9.10 COYU software ... 116
9.10.1 DUST computer program ... 116
9.11 Schemes used for the application of COYU ... 120
10. UNIFORMITY ASSESSMENT ON THE BASIS OF THE RELATIVE VARIANCE METHOD ... 123
10.1 Use of the relative variance method ... 123
10.2 Thresholds for different sample sizes ... 123
10.3 The relative variance test in practice ... 123
10.4 Example of relative variance method ... 124
10.5 Relationship between relative variance and relative standard deviation ... 124
11. EXAMINING CHARACTERISTICS USING IMAGE ANALYSIS ... 125
11.1 Introduction ... 125
11.2 Combined characteristics ... 125
11.3 Image recording: calibration and standardization ... 125
11.4 Conclusions ... 127
11.5 References ... 127
12. EXAMINING CHARACTERISTICS ON THE BASIS OF BULK SAMPLES ... 128

INTRODUCTION

The purpose of this document is to provide guidance on trial design and data analysis, and to provide information on certain techniques used for the examination of DUS. This document is structured as follows:

PART I: DUS TRIAL DESIGN AND DATA ANALYSIS: provides guidance on trial design, data validation, and assumptions to be fulfilled for statistical analysis.

PART II: TECHNIQUES USED IN DUS EXAMINATION: provides details on certain techniques referred to in documents TGP/9 “Examining Distinctness”, and TGP/10 Examining Uniformity where further guidance is considered appropriate.

An overview of the parts of the process of examining distinctness in which trial design and techniques covered in this document are relevant is provided in the schematic overview of the process of examining distinctness provided in document TGP/9 “Examining Distinctness”, section 1 “Introduction”.

PART I: DUS TRIAL DESIGN AND DATA ANALYSIS

1. DUS TRIAL DESIGN

1.1 Introduction

1.1.1 Guidance for conducting the examination is provided in the Test Guidelines where available. A number of Test Guidelines have been developed and there are continual additions, an up-to-date list of which is provided in document TGP/2, "List of Test Guidelines Adopted by UPOV" and on the UPOV website (http://www.upov.int/en/publications/tg_rom/). However, UPOV recommends the following procedure to provide guidance on the testing of distinctness, uniformity and stability where there are no Test Guidelines.

DUS Testing Experience of Other Members of the Union

1.1.2 The examining office is invited to consult document TGP/5, "Experience and Cooperation in DUS Testing," (http://www.upov.int/en/publications/tgp/) and the GENIE Database (http://www.upov.int/genie/en/) to ascertain whether other members of the Union have practical experience in the examination of DUS.

1.1.3 Where such experience is available experts are invited to approach the *members* of the Union concerned and, in accordance with the principles in the General Introduction, seek to harmonize their testing procedures as far as possible. As a next step, the members of the Union concerned are invited to inform UPOV of the existence of the harmonized testing procedure, according to the measures provided in document TGP/5, "Experience and Cooperation in DUS Testing," or, if appropriate, recommend that UPOV prepare Test Guidelines for the species concerned.

DUS Testing Procedures for New Species or Variety Groupings

1.1.4 Where practical DUS testing experience is not available in other members of the Union for the species or variety grouping concerned, experts will need to develop their own testing procedures.

1.1.5 When developing such testing procedures, offices are encouraged to align them on the principles set forth in the General Introduction (document TG/1/3), and the guidance for the development of Test Guidelines contained in document TGP/7, "Development of Test Guidelines." Further guidance is provided in document TGP/13 "Guidance for New Types and Species".

1.1.6 The testing procedure should be documented, in accordance with the requirements of Test Guidelines, to the extent that experience and information permit.

1.1.7 In accordance with the guidance in the General Introduction and document TGP/7, this section follows the structure of section 3 "Method of Examination" of the UPOV Test Guidelines.

1.2 Growing cycles[1]

1.2.1 Introduction

1.2.1.1 A key consideration with regard to growing trials is to determine the appropriate number of growing cycles. In that respect, document TGP/7, Annex I: TG Template, section 4.1.2, states:

> "4.1.2 Consistent Differences
>
> "The differences observed between varieties may be so clear that more than one growing cycle is not necessary. In addition, in some circumstances, the influence of the environment is not such that more than a single growing cycle is required to provide assurance that the differences observed between varieties are sufficiently consistent. One means of ensuring that a difference in a characteristic, observed in a growing trial, is sufficiently consistent is to examine the characteristic in at least two independent growing cycles."

1.2.1.2 The UPOV Test Guidelines, where available, specify the recommended number of growing cycles. When making the recommendation, the experts drafting the UPOV Test Guidelines take into account factors such as the number of varieties to be compared in the growing trial, the influence of the environment on the

[1] See Chapter 3.1 of the Test Guidelines (document TGP/7: Annex 1: TG Template)

expression of the characteristics, and the degree of variation within varieties, taking into account the features of propagation of the variety e.g. whether it is a vegetatively propagated, self-pollinated, cross-pollinated or a hybrid variety.

1.2.1.3 Where UPOV has not established individual Test Guidelines for a particular species or other group(s), the examination should be carried out in accordance with the principles established in the General Introduction, in particular, the recommendations contained in section 9 "Conduct of DUS Testing in the Absence of Test Guidelines" (see paragraphs 1.1.1 to 1.1.7).

1.2.2 Independent growing cycles

1.2.2.1 As indicated in section 1.2.1.1, one means of ensuring that a difference in a characteristic, observed in a growing trial, is sufficiently consistent is to examine the characteristic in at least two independent growing cycles.

1.2.2.2 In general, the assessment of independence is based on the experience of experts.

1.2.2.3 When a characteristic is observed in a growing trial in two independent growing cycles, it is generally observed in two separate plantings or sowings. However, in some perennial crops, such as fruit trees, the growing cycles take the form of one trial observed in two successive years.

1.2.2.4 When field or greenhouse crop trials are planted/sown in successive years, these are considered to be independent growing cycles.

1.2.2.5 Where the two growing trials are in the same location and the same year, a suitable time period between plantings may provide two independent growing cycles. In the case of trials grown in greenhouses or other highly controlled environments, provided the time between two sowings is not "too short", two growing cycles are considered to be independent growing cycles.

1.2.2.6 Where two growing cycles are conducted in the same year and at the same time, a suitable distance or a suitable difference in growing conditions between two locations may satisfy the requirement for independence.

1.2.2.7 The rationale for using independent growing cycles is that if the observed difference in a characteristic results from a genotypic difference between varieties, then that difference should be observed if the varieties are compared again in a similar environment but in an independent growing cycle.

1.3 Testing Place[2]

1.3.1 Purpose

1.3.1.1 Document TGP/7, "Development of Test Guidelines", (see Annex I, TG Template, section 3.2) clarifies that "Tests are normally conducted at one place". However, for example, it may be considered appropriate to conduct tests at more than one place for the following purposes:

(a) Minimizing the overall testing period

1.3.1.2 More than one location may be used on a routine basis, for example, as a means of achieving more than one independent growing cycle in the same year, as set out in section 1.2.2.6. This could reduce the overall length of the testing period and facilitate a quicker decision.

(b) Reserve Trial

1.3.1.3 Authorities may designate a primary location, but organize an additional reserve trial in a separate location. In general, only the data from the primary location would be used, but in cases where that location failed, the reserve trial would be available to prevent the loss of one year's results, provided there was no significant variety-by-location interaction.

[2] See Chapter 3.2 of the Test Guidelines (document TGP/7: Annex 1: TG Template)

(c) Different agro-climatic conditions

1.3.1.4 Different types of varieties may require different agro-climatic growing conditions. In such cases, the breeder would be required to specify the candidate variety type, to allow the variety to be distributed to the appropriate testing location. Section 1.3.2.2 "Additional Tests" addresses the situation where a variety needs to be grown in a particular environment for certain characteristics to be examined, e.g. winter hardiness. However, in such cases each variety will be tested in one location.

1.3.2 Use of information from multiple locations

1.3.2.1 Where more than one location is used, it is important to establish decision rules with regard to the use of data from the different locations for the assessment of DUS and for the establishment of variety descriptions. The possibilities include:

(a) Additional tests

1.3.2.2 Document TGP/7, "Development of Test Guidelines", explains that, in addition to the main growing trial, additional tests may be established for the examination of relevant characteristics (see document TGP/7: Annex 1: TG Template section 3.6). For example, additional tests may be carried out to examine particular characteristics e.g. greenhouse tests for disease resistance, laboratory tests for chemical constituents etc. In such cases, the data for particular characteristics can be obtained at a different location to the main growing trial.

(b) DUS examined on the basis of data for the same characteristics examined at different locations

1.3.2.3 In order to minimize the overall testing period where two independent growing cycles are recommended (see section 1.3.1 *(a)*), a second location might be used to check the consistency of a difference observed in the first location. Such cases would normally apply where the assessment of distinctness is based on Notes (see document TGP/9 sections 5.2.1.1(b) and 5.2.3) and the assessment of distinctness could be considered as based on the first location.

1.3.2.4 In cases where the assessment of distinctness is based on statistical analysis of growing trial data obtained in two or more independent growing cycles (see document TGP/9 sections 5.2.1.1(c) and 5.2.4) it might be considered desirable to combine data from different locations, instead of different years, in order to minimize the overall testing period or to be able to use data from a reserve trial. The suitability of such an approach would depend on the features of the crop concerned (see section 1.2). In particular, careful consideration would need to be given to check whether the necessary assumptions would be satisfied. In that respect, it should be noted that the COYD criterion was tested on data over different years and not tested on data from different locations.

1.4 Conditions for conducting the examination[3]

Document TGP/7 Development of Test Guidelines explains that "the tests should be carried out under conditions ensuring satisfactory growth for the expression of the relevant characteristics of the variety and for the conduct of the examination". Specific guidance, if appropriate, will be provided in the relevant Test Guidelines.

1.5 Test Design[4]

1.5.1 Introduction

In general, the DUS examination is mainly based on a growing trial. There may be additional growing trials for the examination of particular characteristics or particular aspects of DUS; e.g. ear-rows for examination of uniformity, or additional field trials with plants at different stages of development, such as young and mature trees. Furthermore, there may be characteristics which require examination by additional tests, e.g. disease

[3] See Chapter 3.3 of the Test Guidelines (document TGP/7: Annex 1: TG Template)

[4] See Chapter 3.4 of the Test Guidelines (document TGP/7: Annex 1: TG Template)

resistance. The explanations provided in the following sections are intended to provide guidance on the principles applied for growing trials.

1.5.2 Number of Plants in the trial

The number of plants in the trial is influenced by several factors such as genetic structure of the variety, way of reproduction of the species, the agronomic features and the "feasibility" of the trial. The most significant criteria to determine the number of plants are, the variability within and between varieties, and the method of assessment of distinctness and uniformity.

1.5.3 Trial layout

1.5.3.1 Introduction

1.5.3.1.1 The type of trial layout will be determined by the approaches to be used for the assessment of distinctness, uniformity and stability. The approaches to be used for the assessment of distinctness are explained in document TGP/9 Examining Distinctness, section 5.2.1:

> "5.2.1 Introduction
>
> 5.2.1.1 Approaches for assessment of distinctness based on the growing trial can be summarized as follows:
>
> (a) Side-by-side visual comparison in the growing trial (see section 5.2.2);
>
> (b) Assessment by Notes / single variety records ("Notes"): the assessment of distinctness is based on the recorded state of expression of the characteristics of the variety (see section 5.2.3);
>
> (c) Statistical analysis of growing trial data: the assessment of distinctness is based on a statistical analysis of the data obtained from the growing trial. This approach requires that, for a characteristic, there are a sufficient number of records for a variety (see section 5.2.4).
>
> 5.2.1.2 The choice of approach or combination of approaches for the assessment of distinctness, which is influenced by the features of propagation of the variety and the type of expression of the characteristic, determines the method of observation and type of record (VG, MG, VS or MS). The common situations are summarized by the table in section 4.5. […]."

1.5.3.1.2 The approaches to be used for the assessment of uniformity are explained in document TGP/10 Examining Uniformity, section 2.5.1:

> "2.5.1 The type of variation in the expression of a characteristic within a variety determines how that characteristic is used to determine uniformity in the crop. In cases where it is possible to "visualize" off-types, the off-type approach is recommended for the assessment of uniformity. In other cases, the standard deviations approach is used. Thus, the uniformity of a variety may be determined by off-types alone, by standard deviations alone, or by off-types for some characteristics and by standard deviations for other characteristics. Those situations are considered further in section 6."

1.5.3.1.3 Document TGP/7 Development of Test Guidelines ASW 5 Plot design identifies the following types of DUS trial

> ASW 5 (TG Template: Chapter 3.4) – Plot design
>
> *(a) Single plots*
> "Each test should be designed to result in a total of at least {…} [plants]/[trees]"
>
> *(b) Spaced plants and row plots*
> "Each test should be designed to result in a total of at least {…} spaced plants and {…} meters of row plot."
>
> *(c) Replicate plots (or Replicates)*
> "Each test should be designed to result in a total of at least {…} plants, which should be divided between {…} replicates."

Spaced plants and row plots form different trials and, in particular, do not constitute replicate plots (see section 1.5.3.3).

1.5.3.1.4 Single plot trials are suitable when distinctness is assessed on a side-by-side visual comparison or by notes/single variety records (see document TGP/9 section 4.3.2.3) and when uniformity is assessed by off-types. Common examples of this are vegetatively propagated ornamental and fruit varieties.

1.5.3.1.5 Replicate plots are suitable when the assessment of distinctness requires, for at least some characteristics, the calculation of a variety mean by observation or measurement of groups of plants (see document TGP/9 section 4.3.2.4). In such cases, uniformity is, in general, assessed on the basis of off-types. Common examples of this are self-pollinated agricultural crops (e.g. cereals).

1.5.3.1.6 Replicate plots are appropriate when records for a number of single, individual plants or parts of plants are required for statistical analysis of individual plant data for the assessment of distinctness, for at least some characteristics (normally quantitative characteristics) (see document TGP/9 section 4.3.3). In such cases, uniformity is assessed, for the relevant characteristics, in general, by standard deviation. Common examples of this are cross-pollinated varieties.

1.5.3.1.7 The following table summarizes common types of trial design according to the method of examining distinctness and uniformity:

		UNIFORMITY	
		Off-type approach	Standard deviation
DISTINCTNESS	Side-by-side visual comparison (VG)	Single plots (see section 1.5.3.2)	
	Notes / single variety records (VG/MG)	Single plots (see section 1.5.3.2)	
	Variety mean Statistical analysis of records for a group of plants [Replicate plots for group data records] (MG/MS)	Replicate plots	
	Statistical analysis of individual plant data (MS)		Replicate plots (see section 1.5.3.3.3)

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

(See documents TGP/9 “Examining Distinctness”, Section 4 “Observation of characteristics” and TGP/7, Annex I-TG template, ASW 7 (b)).

1.5.3.1.8 Occasionally, such as in the circumstances described in document TGP/9 section 6.4, it may be appropriate to conduct randomized "blind" testing. In such cases existing plots or parts of plants taken from the trial may be used (e.g. 'Randomized variety plots' and 'Parts of plants of varieties' mentioned in document TGP/9 section 6.4.4). In other cases, plants must be sown specifically for the randomized "blind" testing, such as plots containing plants of both varieties to be distinguished, with the plants sown in a random but known order. In this case these mixture-plots physically form a part of the trial in the field. Alternatively the randomized "blind" testing may take the form of a mixture of pots with the two varieties in a greenhouse, also considered to be an extension to the trial. The layout of these randomized "blind" testing trials is discussed in section 1.5.3.4.

1.5.3.2 Single plots

This trial design implies that for each variety included in the trial, there will be a single plot, and distinctness and uniformity will be assessed on the same plot.

1.5.3.3 Replicate plots (statistical analysis)

1.5.3.3.1 Introduction

Replicate plots are used when more than a single record per variety is required for the assessment of distinctness. The data from a group of plants can be used to calculate a variety mean, or the individual plant data can be used for statistical analysis

1.5.3.3.2 Replicate plots for statistical analysis of individual plant data

1.5.3.3.2.1 Where the assessment of distinctness and uniformity is based on statistical analysis of individual plant data, the trial will comprise of a number of plots. The plots will be grouped, in general, into replicates such that each replicate contains one plot of each variety. The allocation of varieties to plots will involve randomization (see section 1.5.3.3.3). Examples of trial designs used when such statistical analysis is used are:

- Completely randomized design and randomized complete block design (see section 1.5.3.3.3)
- Randomized incomplete block designs (see section 1.5.3.3.4)
- Design for pair-wise comparisons between particular varieties (see section 1.5.3.3.5)

1.5.3.3.2.2 Distinctness may be assessed by statistical analysis for all characteristics or for some characteristics by statistical analysis (in particular quantitative characteristics) and for other characteristics (in general pseudo-qualitative and qualitative characteristics) by side-by-side visual comparison or by notes/single variety records, as appropriate.

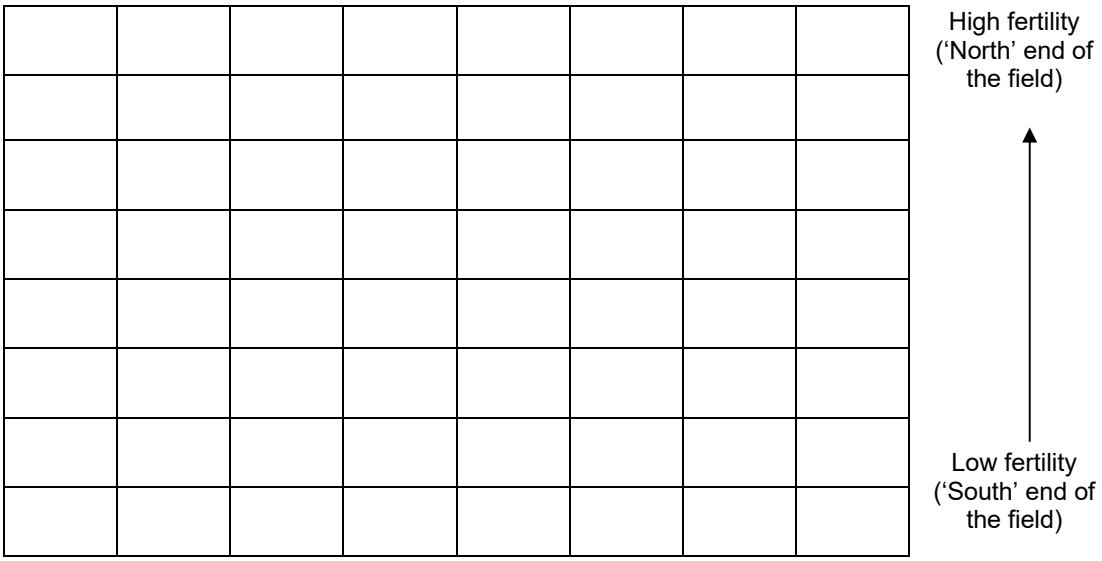
1.5.3.3.2.3 Uniformity can be assessed by standard deviation for all characteristics, or by standard deviation for some characteristics and by off-types for other characteristics, as appropriate (see document TGP/10/1, section 6.4).

1.5.3.3.3 Randomization

1.5.3.3.3.1 If there are to be replicate plots of each variety in the growing trial, decisions must be made as to whether the replicate plots should be grouped into blocks and how the plots should be aligned within a block, i.e. the Experimental Design. This determines how local, unwanted or nuisance variation is controlled and hence how precisely distinctness and uniformity can be assessed. Then there is the notion that variation arises from different sources, and how this can affect the choice of sample sizes, which again impacts on precision. Precision is important because it in turn impacts on the decision making. If data are relatively imprecise and decisions are based on this data, there is an appreciable chance that inappropriate or wrong decisions will get made. This is discussed below.

1.5.3.3.3.2 In designing an experiment it is important to choose an area of land that is as homogeneous as possible in order to minimize the variation between plots of the same variety, i.e. the random variation. Assume

that we have a field where it is known that the largest variability is in the 'north-south' direction, e.g. as in the following figure:



1.5.3.3.3.3 Let's take an example where four varieties are to be compared with each other in an experiment within this field where each of the varieties is assigned to 4 different plots. It is important to randomize the varieties over the plots. If varieties are arranged systematically, not all varieties would necessarily be under the same conditions (see following figure).

Variety A	Variety A	Variety A	Variety A	Variety B	Variety B	Variety B	Variety B	Higher fertility row
Variety C	Variety C	Variety C	Variety C	Variety D	Variety D	Variety D	Variety D	Lower fertility row

If the fertility of the soil decreases from the north to the south of the field, the plants of varieties A and B have grown on more fertile plots than the other varieties. The comparison of the varieties is influenced by a difference in fertility of the plots. Differences between varieties are said to be confounded with differences in fertility.

1.5.3.3.3.4 To avoid systematic errors it is advisable to randomize varieties across the site. A complete randomization of the four varieties over the sixteen plots could have resulted in the following layout:

Variety C	Variety A	Variety A	Variety B	Variety C	Variety D	Variety B	Variety C	Higher fertility row
Variety C	Variety A	Variety D	Variety A	Variety D	Variety B	Variety D	Variety B	Lower fertility row

1.5.3.3.3.5 However, looking at the design we find that variety C occurs three times in the top row (with high fertility) and only once in the second row (with lower fertility). For variety D we have the opposite situation. Because we know that there is a fertility gradient, this is still not a good design, but it is better than the first systematic design.

1.5.3.3.3.6 When we know that there are certain systematic sources of variation like the fertility gradient in the paragraphs before, we may take that information into account by making so-called blocks. The blocks should be formed so that the variation within each block is minimized. With the assumed gradients we may choose either two blocks each consisting of one row or we may choose four blocks – two blocks in each row with four plots each. In larger trials (more plots) the latter will most often be the best, as there will also be some variation within rows even though the largest gradient is between rows.

Block I				Block II				
Variety A	Variety C	Variety D	Variety B	Variety A	Variety C	Variety D	Variety B	Higher fertility row
Variety B	Variety C	Variety A	Variety D	Variety C	Variety A	Variety D	Variety B	Lower fertility row
Block III				Block IV				

An alternative way of reducing the effect of any gradient between the columns is to use plots that are half the width, but which extend over two rows, i.e. by using long and narrow plots:

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

In both designs above, the ‘north-south’ variability will not affect the comparisons between varieties.

1.5.3.3.3.7 In a randomized complete block design the number of plots per block equals the number of varieties. All varieties are present once in each block and the order of the varieties within each block is randomized. The advantage of a randomized complete block design is that the standard deviation between plots (varieties), a measure of the random variation, does not contain variation due to differences between blocks. The main reason for the random allocation is that it ensures that the results are unbiased and so represent the varieties being compared. In other words, the variety means will, on average, reflect the true variety effects, and will not be inflated or deflated by having been allocated to inherently better or worse plots. An interesting feature of the randomization is that it makes the observations from individual plots ‘behave’ as independent observations (even though they may not be so). There is usually no extra cost associated with blocking, so it is recommended to arrange the plots in blocks.

1.5.3.3.3.8 Blocking is introduced here on the basis of differences in fertility. Several other systematic sources of variation could have been used as the basis for blocking. 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 other reasons. When there are different sowing machines, different observers, different observation days, such 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 have the same sowing machine, the same observer, the same observation day, and so on.

1.5.3.3.3.9 Management may influence the choice of the form of the plots. In some crops it may be easier to handle long and narrow plots than square plots. Long narrow plots are usually considered to be more influenced by varieties in adjacent plots than square plots. The size of the plots should be chosen in such a way that the necessary number of plants for sampling is available. For some crops it may be necessary also to have guard plants (areas) in order to avoid large competition effects. However, overly large plots require more land and will often increase the random variability between plots. Growing physically similar varieties together, e.g. varieties of similar height may also reduce the competition between adjacent plots. If nothing is known about the fertility of the area, then layouts with compact blocks (i.e. almost square blocks) will often be most appropriate because the larger the distance between two plots the more different they will usually be. In both designs above, the blocks can be placed as shown or they could be placed ‘on top of each other’ (see following figure). This will usually not change the variability between plots considerably – unless one of the layouts forces the crop expert to use more heterogeneous soil.

Variety A	Variety C	Variety D	Variety B	Block I	Higher fertility row
Variety A	Variety C	Variety D	Variety B	Block II	
Variety B	Variety C	Variety A	Variety D	Block III	
Variety C	Variety A	Variety D	Variety B	Block IV	Lower fertility row

1.5.3.3.4 Randomized incomplete block designs

1.5.3.3.4.1 If the number of varieties becomes very large (>20-40), it may be impossible to construct complete blocks that would be sufficiently homogeneous. In that case it might be advantageous to form smaller blocks, each one containing only a fraction of the total number of varieties. Such designs are called incomplete block designs. Several types of incomplete block designs can be found in the literature for example, balanced incomplete block designs and partially balanced incomplete block designs such as lattice designs and row and column designs. One of the most familiar types for variety trials is a lattice design. The generalized lattice designs (also called α-designs) are very flexible and can be constructed for any number of varieties and for a large range of block sizes and number of replicates. One of the features of generalized lattice designs is that the incomplete blocks form a whole replicate. This means that such designs will be at least as good as randomized complete block designs, since the analysis can be performed using either a lattice model or a randomized complete block model. The lattice model should be preferred if conditions are fulfilled. Determining optimal sub-block size depends on different factors, such as the variability of the soil and the differing susceptibilities of characteristics to that variability. However, if there is no information available, e.g. from the first trial, the applicable number of sub-blocks could be calculated as a whole number close to the square root of the number of varieties, e.g. 100 varieties would require 10 sub-blocks.

1.5.3.3.4.2 Incomplete blocks need to be constructed in such a way that it is possible to compare all varieties in an efficient way. An example of an α-design is shown in the following figure:

Block I	Sub-block I	Variety F	Variety E	Variety O	Variety S
	Sub-block II	Variety M	Variety H	Variety J	Variety T
	Sub-block III	Variety B	Variety C	Variety D	Variety G
	Sub-block IV	Variety L	Variety A	Variety R	Variety N
	Sub-block V	Variety Q	Variety K	Variety P	Variety I

Block II	Sub-block I	Variety D	Variety P	Variety F	Variety A
	Sub-block II	Variety R	Variety E	Variety J	Variety B
	Sub-block III	Variety N	Variety G	Variety Q	Variety H
	Sub-block IV	Variety K	Variety S	Variety M	Variety C
	Sub-block V	Variety O	Variety I	Variety T	Variety L

Block III	Sub-block I	Variety D	Variety T	Variety E	Variety Q
	Sub-block II	Variety B	Variety M	Variety A	Variety I
	Sub-block III	Variety C	Variety F	Variety L	Variety H
	Sub-block IV	Variety R	Variety G	Variety K	Variety O
	Sub-block V	Variety P	Variety J	Variety N	Variety S

In the example above, 20 varieties are to be grown in a trial with three replicates. In the design the 5 sub-blocks of each block form a complete replicate. Thus each replicate contains all varieties whereas any pair of varieties occurs either once or not at all in the same sub-block. Note: in the literature, the blocks and sub-blocks are sometimes referred to as super-blocks and blocks.

1.5.3.3.4.3 The incomplete block design is most suitable for trials where grouping characteristics are not available. If grouping characteristics are available then some modification may be advantageous for trials with many varieties, such as using grouping characteristics to form separate trials rather than a single trial, see document TGP/9/1 section 2.3 Grouping varieties on the basis of characteristics.

1.5.3.3.5 Design for pair-wise comparisons between particular varieties

1.5.3.3.5.1 When a close comparison is needed between a pair of varieties by means of statistical analysis, it may be good to grow them in neighbouring plots. A similar theory to that used in split-plot designs may be used for setting up a design where the comparisons between certain pairs of varieties are to be optimized. When setting up the design, the pairs of varieties are treated as the whole plot factor and the comparison between varieties within each pair is the sub-plot factor. As each whole plot consists of only two sub-plots, the comparisons within pairs will be (much) more precise than if a randomized block design was used.

1.5.3.3.5.2 If, for example, four pairs of varieties (A-B, C-D, E-F and G-H) have to be compared very precisely, then this can be done using the following design of 12 whole plots each having 2 sub-plots:

Pair 1 variety A	Pair 3 variety E	Pair 4 variety H
Pair 1 variety B	Pair 3 variety F	Pair 4 variety G
Pair 3 variety F	Pair 2 variety D	Pair 1 variety A
Pair 3 variety E	Pair 2 variety C	Pair 1 variety B
Pair 4 variety G	Pair 1 variety B	Pair 2 variety C
Pair 4 variety H	Pair 1 variety A	Pair 2 variety D
Pair 2 variety D	Pair 4 variety H	Pair 3 variety E
Pair 2 variety C	Pair 4 variety G	Pair 3 variety F

In this design each column represents a replicate. Each of these is then divided into four incomplete blocks (whole plots) each consisting of two sub-plots. The four pairs of varieties are randomized to the incomplete blocks within each replicate and the order of varieties is randomized within each incomplete block. The comparison between varieties of the same pair is made more precise at the cost of the precision of the comparison between varieties of a different pair.

1.5.3.3.6 Further statistical aspects of trial design

1.5.3.3.6.1 Introduction

1.5.3.3.6.1.1 This section describes a number of concepts that are relevant when designing growing trials for which distinctness and/or uniformity are to be assessed by statistical analysis of the growing trial data.

1.5.3.3.6.2 The hypotheses under test

1.5.3.3.6.2.1 When statistical analysis of growing trial data is to be used to assess distinctness and/or uniformity, the purpose of the growing trial is to get precise and unbiased averages of characteristics for each variety and also to judge the within-variety variability by calculating the standard deviation. Assessments of the distinctness of varieties are made based on the characteristic averages. The type of variation in the expression of a characteristic within a variety determines how that characteristic is used to determine uniformity in the crop. In cases where it is possible to “visualize” off-types, the off-type approach is recommended for the assessment of uniformity. In other cases, the standard deviations approach is used.

1.5.3.3.6.2.2 In evaluating distinctness and uniformity we test a null hypothesis (H_0) and either accept or reject it. If we reject it, we accept an alternative hypothesis (H_1). The null and alternative hypotheses for the distinctness and uniformity decisions are given in the following table:

	Null Hypothesis (H_0)	Alternative Hypothesis (H_1)
Distinctness	two varieties are not distinct for the characteristic	two varieties are distinct
Uniformity	a variety is uniform for the characteristic	a variety is not uniform

1.5.3.3.6.2.3 We make each evaluation by computing a test statistic from the observations using a formula. If the absolute value of the test statistic is greater than its chosen critical value, the null hypothesis (H_0) is rejected, the alternative hypothesis (H_1) is accepted, and the test is called significant. If the test statistic is not greater than its chosen critical value, the null hypothesis (H_0) is accepted. The choice of the critical value that the test statistic is compared with is explained below.

1.5.3.3.6.2.4 Note that if the null hypothesis (H_0) is rejected for distinctness, this leads to the conclusion that the candidate variety is distinct.

1.5.3.3.6.2.5 On the other hand, if the null hypothesis (H_0) is rejected for uniformity, the candidate variety is considered not uniform.

1.5.3.3.6.2.6 The test statistic is based on a sample of plants, trialled in a sample of growing conditions. Thus if the process were to be repeated at a different time, a different value of the test statistic would be obtained. Because of this inherent variability, there is a chance that a different conclusion is arrived at compared to the conclusion which would be reached if the trial could be repeated indefinitely. Such "statistical errors" can occur in two ways, let us first consider distinctness conclusions:-

- The conclusions based on the test statistic, i.e. from the DUS trial, is that two varieties are distinct, when they would not be distinct if the trial could be repeated indefinitely. This is known as a Type I error and its risk is denoted by α.

- The conclusions based on the test statistic, i.e. from the DUS trial, is that two varieties are not distinct, when, if the trial could be repeated indefinitely, they would be distinct. This is known as a Type II error and its risk is denoted by β.

	Conclusion based on test statistic	
Conclusion if the trial could be repeated indefinitely	Varieties are not distinct (H_0 true)	Varieties are distinct (H_1 true)
Varieties are distinct (H_1 true)	Different result, Type II error, made with probability β	Same result
Varieties are not distinct (H_0 true)	Same result	Different result, Type I error, made with probability α

1.5.3.3.6.2.7 Likewise, it is possible when deciding on uniformity based on a test statistic, i.e. from the DUS trial, to decide that a variety is not uniform, when if the trial could be repeated indefinitely the variety would be uniform, i.e. a Type I error (α). Alternatively, a Type II error (β) is the conclusion based on a test statistic that a variety is uniform when, if the trial could be repeated indefinitely the variety would not be uniform. The following table shows the two types of statistical error that can be made when testing for uniformity:

	Conclusions based on test statistic	
Conclusion if the trial could be repeated indefinitely	Variety is uniform (H_0 true)	Variety is not uniform (H_1 true)
Variety is uniform (H_0 true)	Same result	Different result, Type I error, made with probability α
Variety is not uniform (H_1 true)	Different result, Type II error, made with probability β	Same result

1.5.3.3.6.2.8 The risk of making a Type I error can be controlled easily by choice of α, which determines the critical value that the test statistic is compared against. α is also known as the size of the test and the significance level of the test. The risk of making a Type II error is more difficult to control as it depends, for example in the case of distinctness, on the size of the real difference between the varieties, the chosen α, and the precision of the test which is determined by the number of replicates and the inherent variability of the measurements. The crop expert can reduce the risk of making a Type II error by increasing the precision, e.g. by increasing the number of replicates, by reducing the random variability by choice of number of plants per plot (or sample size), by controlling local, unwanted or nuisance variation through careful choice of experimental design, and by improving the way measurements/observations are made and so reducing the observer error.

1.5.3.3.6.3 Determining optimal sample size

1.5.3.3.6.3.1 The precision of a test does not depend on sample size alone. The precision of a test based on the observations of one experiment also depends, say for quantitative characteristics, on at least three sources of variation:

- the variation between individual plants within a plot, i.e. the “within-plot” or “plant” variance component: a mixture of different sources of variation such as different plants, different times of observation, different errors of measurement
- the variation between the plots within a block, i.e. the “between-plot” or “plot” variance component
- the variation caused by the environment, i.e. the variation in the expression of characteristics from year to year (or from location to location)

1.5.3.3.6.3.2 To estimate the optimal sample size for a quantitative characteristic it is necessary to know the standard deviations of the above sources of variation, expected differences between the varieties which should be significant, the number of varieties and the number of blocks in the trial. Additionally, it is necessary to determine the Type I (α) and Type II (β) error probabilities. Computing the optimal sample size for each characteristic enables a determination of the optimal sample size for this trial for all quantitative characteristics. Especially for the assessment of uniformity, the Type II error is sometimes more important than the Type I error. In some cases the Type II error could be greater than 50 % which may be unacceptable.

1.5.3.3.6.3.3 The precision of the variety means in one year’s or one cycle’s experiment depends on the number of replicates, the number of plants per plot, and the experimental design. When these means are used in the over-year or over-cycle analysis for COYD for example, their precision is only of benefit indirectly, because the standard deviation in that analysis is based on the interaction between the varieties and the years or cycles. Further, if the differences between the varieties over the years or cycles are very large, the precision of the means per experiment are relatively unimportant.

1.5.3.3.6.3.4 Where available, the UPOV Test Guidelines recommend an appropriate sample size for the trial as a whole, taking into account the factors explained above.

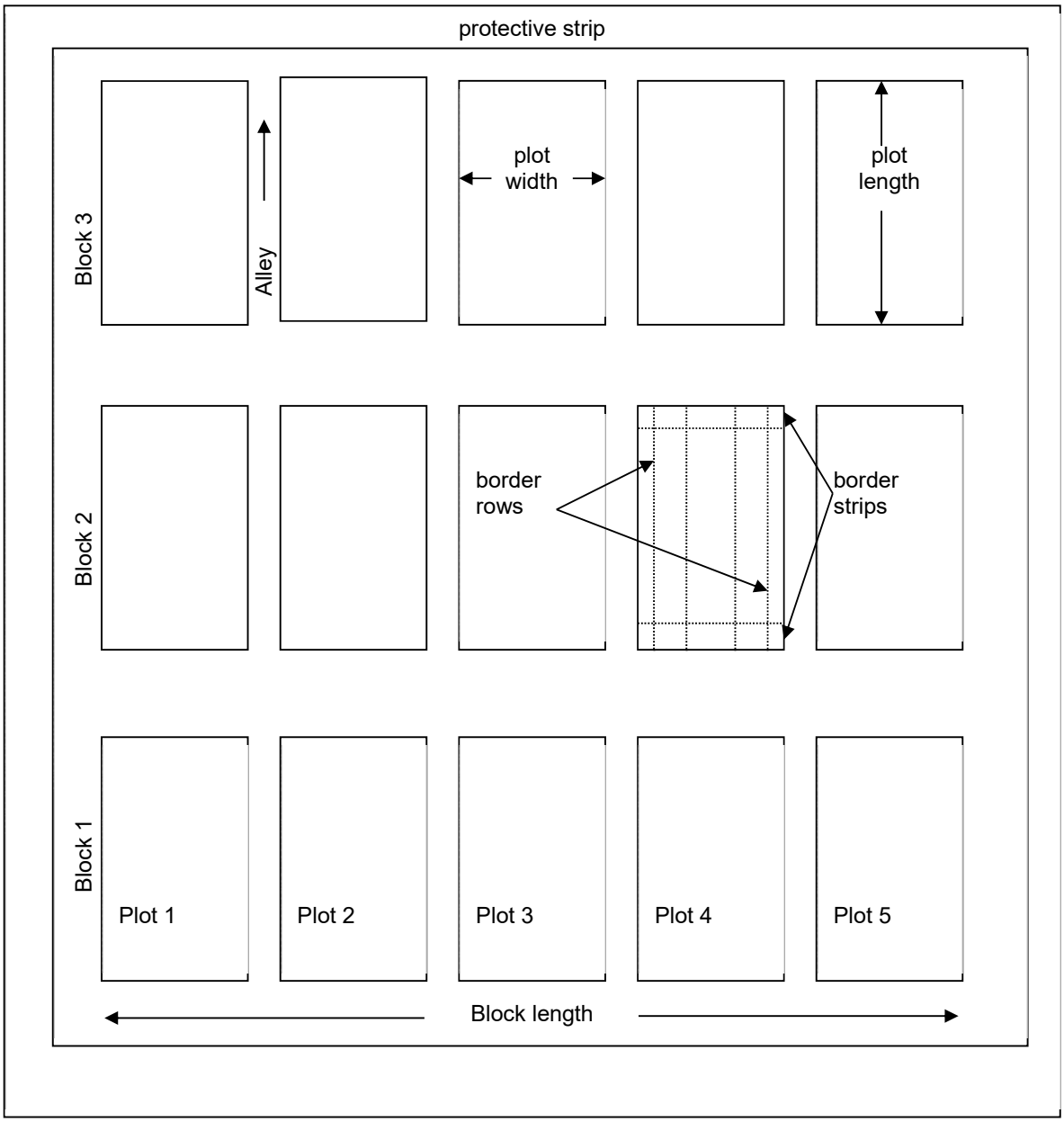
1.5.3.3.7 Trial elements when statistical analysis is used

1.5.3.3.7.1 Introduction

1.5.3.3.7.1.1 In deciding on trial layout, it is important that local variation in conditions is taken into account. For this, decisions on: plot size, shape of the plots, alignment of the plots, barrier rows and border strips and protective strips are needed.

1.5.3.3.7.1.2 For the assessment of distinctness unbiased observation of characteristics are necessary. In some cases it is necessary to have border rows and strips to minimize bias caused by inter-plot interference, i.e. interference between plants on different plots, and other special border effects, such as shading and soil moisture. Also, protective strips on the border of the trial are often used to reduce the chance of external influences biasing one plot in favour of another. When observing characteristics on the plants on a plot it is usual to exclude the plot’s border rows and border strips.

1.5.3.3.7.1.3 The following figure may be helpful to give some explanations of the particular trial elements:



1.5.3.3.7.2 Plots and blocks

A plot is the experimental unit to which the varieties are allocated. A plot contains plants from the same variety. Depending on the type of growing trial, a plot may be an area of land, or a group of plant pots. A block is a group of plots within which the varieties are allocated. A growing trial may contain just one block or it may contain more than one block.

1.5.3.3.7.3 Allocation of varieties to plots

1.5.3.3.7.3.1 Several factors will influence the decision on allocating varieties to plots: in particular the selected approach for distinctness (see section 1.5.3.1.1) and uniformity (see section 1.5.3.1.2).

1.5.3.3.7.3.2 When distinctness is assessed by statistical analysis of growing trial data, depending on the trial design either randomization or partial randomization must be used, as it ensures that there is no subjectivity in the allocation. Random allocation ensures that on average the effects of other factors influencing

the plants' characteristics, such as soil conditions, are expected to cancel out when the variety means are compared.

1.5.3.3.7.3.3 Sections 1.5.3.2 and 1.5.3.3.1 to 1.5.3.3.5 provide more details on different ways of allocating varieties in plots and blocks.

1.5.3.3.7.4 Plot size, shape and configuration

1.5.3.3.7.4.1 Section 3 of the Test Guidelines "Method of examination", provides information on the duration of the test, the testing place, the test design, number of plants/parts of plant to be examined as well as on additional tests which may be used for the assessment of relevant characteristics. The Test Guidelines may indicate the type of record required for the assessment of distinctness (single record for a group of plants or parts of plants (G), or records for a number of single, individual plants or parts of plants (S)). Uniformity, however, is assessed on the whole sample under examination by the off-type approach and/or by the standard deviation approach (see document TGP/10 section 3). These will determine the sample size, i.e. the number of plants which must be observed, and hence determine the minimum effective size of the plot. To decide on the actual plot size, allowance must be made for any necessary border rows and strips.

1.5.3.3.7.4.2 The plot size and the plot shape also depend on the soil and other conditions, irrigation equipment, or on the 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 to mitigate variation in conditions within the block (e.g. caused by soil variation).

1.5.3.3.7.4.3 Square plots have the smallest total length of the borders (circumference). From the theoretical point of view the square shape is optimal to minimize the interference of different phenotypes. Grouping the varieties can also help minimize this interference.

1.5.3.3.7.4.4 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 varieties. The larger the number of varieties in a block the narrower the plots - but not so narrow that the inter-plot competition becomes a problem.

1.5.3.3.7.5 Independence of plots

1.5.3.3.7.5.1 When distinctness and uniformity are to be assessed by statistical analysis of the growing trial data, one of the most important requirements of experimental units is independence.

1.5.3.3.7.5.2 Independence of plots means that observations made on a plot are not influenced by the circumstances in other plots. For example, if tall varieties are planted next to short ones there could be a negative influence of the tall ones interfering with the short ones and a positive influence in the other direction. In such a case, in order to avoid this dependency an additional row of plants can be planted on both sides of the plot, i.e. border rows and strips. Another possibility to minimize this influence is to grow physically similar varieties together.

1.5.3.3.7.6 The arrangement of the plants within the plot/ Type of plot for observation

The UPOV Test Guidelines may specify the type/s of plot for the growing trial (e.g. spaced plants, row plot, drilled plot, etc.) in order to examine distinctness as well as uniformity and stability.

1.5.3.4 Blind Randomized Trials

1.5.3.4.1 Part of a trial may consist of plots sown specifically for randomized "blind" testing, such as plots containing plants of both the varieties to be distinguished between, with the plants sown in a random but known order, or alternatively a mixture of pots with the two varieties in a greenhouse. The two varieties comprise the candidate plus the variety with which the distinctness of the candidate is in dispute. The principle of randomized "blind" testing is that a judge, sometimes the breeder, is presented with the plants and is asked to tell plant by plant which is the candidate, and which is the other variety.

1.5.3.4.2 To allow this, the plants must be presented or sown in a random order but such that the tester knows which is which variety, the judge judges each plant, and the tester counts the number of times the different varieties are correctly identified. In order to reinforce the blindness of the test, a different number of plants from each of the two varieties are presented, for instance 51 of the candidate and 69 of the other, rather than 60 of each. As differences may occur at different stages of growth, the judge can assess the plants on more than one occasion.

1.6 Changing Methods

Changes in the methods of assessing DUS may have a significant impact on decisions. Therefore, due consideration should be given to seeking to ensure that there is consistency in decisions to change the methods.

2. DATA TO BE RECORDED

2.1 Introduction

Document TGP/9 Examining Distinctness, sections 4.4 and 4.5, provide the following guidance on the type of observation for distinctness in respect of the type of characteristic and the method of propagation of the variety:

"4.4 Recommendations in the UPOV Test Guidelines

"The indications used in UPOV Test Guidelines for the method of observation and the type of record for the examination of distinctness, are as follows:

"Method of observation

"M: to be measured (an objective observation against a calibrated, linear scale e.g. using a ruler, weighing scales, colorimeter, dates, counts, etc.);
"V: to be observed visually (includes observations where the expert uses reference points (e.g. diagrams, example varieties, side-by-side comparison) or non-linear charts (e.g. color charts). "Visual" observation refers to the sensory observations of the expert and, therefore, also includes smell, taste and touch.

"Type of record(s)

"G: single record for a variety, or a group of plants or parts of plants;
"S: records for a number of single, individual plants or parts of plants

"For the purposes of distinctness, observations may be recorded as a single record for a group of plants or parts of plants (G), or may be recorded as records for a number of single, individual plants or parts of plants (S). In most cases, "G" provides a single record per variety and it is not possible or necessary to apply statistical methods in a plant-by-plant analysis for the assessment of distinctness.

"4.5 Summary

"The following table summarizes the common method of observation and type of record for the assessment of distinctness, although there may be exceptions:

	Type of expression of characteristic		
Method of propagation of the variety	QL	PQ	QN
Vegetatively propagated	VG	VG	VG/MG/MS
Self-pollinated	VG	VG	VG/MG/MS
Cross-pollinated	VG/(VS*)	VG/(VS*)	VS/VG/MS/MG
Hybrids	VG/(VS*)	VG/(VS*)	* *

* Records of individual plants only necessary if segregation is to be recorded.
** To be considered according to the type of hybrid."

2.2 Types of expression of characteristics

2.2.1 Characteristics can be classified according to their types of expression. The following types of expression of characteristics are defined in 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, the "General Introduction", Chapter 4.4):

2.2.2 "Qualitative characteristics" (QL) are those that are expressed in discontinuous states (e.g. sex of plant: dioecious female (1), dioecious male (2), monoecious unisexual (3), monoecious hermaphrodite (4)). These states are self-explanatory and independently meaningful. All states are necessary to describe the full range of the characteristic, and every form of expression can be described by a single state. The order of states is not important. As a rule, the characteristics are not influenced by environment.
2.2.3 "Quantitative characteristics" (QN) are those where the expression covers the full range of variation from one extreme to the other. The expression can be recorded on a one-dimensional, continuous or discrete,

linear scale. The range of expressions is divided into a number of states for the purpose of description (e.g. length of stem: very short (1), short (3), medium (5), long (7), very long (9)). The division seeks to provide, as far as practical, an even distribution across the scale. The Test Guidelines do not specify the difference needed for distinctness. The states of expression should, however, be meaningful for DUS assessment.

2.2.4 In the case of "pseudo-qualitative characteristics" (PQ) the range of expression is at least partly continuous, but varies in more than one dimension (e.g. shape: ovate (1), elliptic (2), circular (3), obovate (4)) and cannot be adequately described by just defining two ends of a linear range. In a similar way to qualitative (discontinuous) characteristics – hence the term "pseudo-qualitative" – each individual state of expression needs to be identified to adequately describe the range of the characteristic.

2.3 Types of scales of data

The possibility to use specific procedures 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 way of recording this expression. The type of scale may be nominal, ordinal, interval or ratio.

2.3.1 Data from qualitative characteristics

2.3.1.1 Data results from qualitative characteristics are nominal scaled data without any logical order of the discrete categories. They result from visually assessed (notes) qualitative characteristics.
Examples:

Type of scale	Example	Example number*
nominal	Sex of plant	1
nominal with two states	Leaf blade: variegation	2

* For description of the states of expressions, see Table 6.

2.3.1.2 A nominal scale consists of numbers which correspond to the states of expression of the characteristic, which are referred to in the Test Guidelines as notes. Although numbers are used for designation there is no logical order for the expressions and so it is possible to arrange them in any order.

2.3.1.3 Characteristics with only two categories (dichotomous characteristic) are a special form of a nominal scaled characteristic.

2.3.1.4 The nominal scale is the lowest classification of the scales (Table 1). Few statistical procedures are applicable for evaluations (see section 2.3.7).

2.3.2 Data from quantitative characteristics

2.3.2.1 Data results from quantitative characteristics are metric (ratio or interval) or ordinal scaled data.

2.3.2.2 Metric scaled data are all data which are recorded by measuring or counting. Weighing is a special form of measuring. Metric scaled data can have a continuous or a discrete distribution. Continuous metric data result from measurements. They can take every value out of the defined range. Discrete metric data result from counting.

Examples

Type of scale	Example	Example number*
Continuous metric	Plant length in cm	3
Discrete metric	Number of stamens	4

* For description of the states of expression, see Table 6.

2.3.2.3 The continuous metric scaled data for the characteristic "Plant length" are measured on a continuous scale with defined units of assessment. A change of unit of measurement e.g. from cm into mm is only a question of precision and not a change of type of scale.

2.3.2.4 The discrete metric scaled data of the characteristic “Number of stamens” are assessed by counting (1, 2, 3, 4, and so on). The distances between the neighboring units of assessment are constant and for this example equal to 1. There are no real values between two neighboring units but it is possible to compute an average which falls between those units.

2.3.2.5 Metric scales can be subdivided into ratio scales and interval scales.

(a) Ratio scale

2.3.2.6 A ratio scale is a metric scale with a defined absolute zero point. There is always a constant non-zero distance between two adjacent expressions. Ratio scaled data may be continuous or discrete.

2.3.2.7 The definition of an absolute zero point makes it possible to define meaningful ratios. This is a requirement for the construction of indexes, which are the combination of at least two characteristics (e.g. the ratio of length to width). In the General Introduction, this is referred to as a combined characteristic (see document TG/1/3, section 4.6.3).

2.3.2.8 It is also possible to calculate ratios between expressions of different varieties. For example, in the characteristic ‘Plant length’ assessed in cm, there is a lower limit for the expression which is ‘0 cm’ (zero). It is possible to calculate the ratio of length of plant of variety ‘A’ to length of plant of variety ‘B’ by division:

Length of plant of variety ‘A’ = 80 cm
Length of plant of variety ‘B’ = 40 cm
Ratio = Length of plant of variety ‘A’ / Length of plant of variety ‘B’
= 80 cm / 40 cm
= 2.

2.3.2.9 So it is possible in this example to state that plant ‘A’ is double the length of plant ‘B’. The existence of an absolute zero point ensures an unambiguous ratio.

2.3.2.10 The ratio scale is the highest classification of the scales (Table 1). That means that ratio scaled data include the highest information about the characteristic and it is possible to use many statistical procedures (see section 2.3.7).

2.3.2.11 The examples 3 and 4 (Table 6) are examples for characteristics with ratio scaled data.

(b) Interval scale

2.3.2.12 An Interval scale is a metric scale without a defined absolute zero point. There is always a constant non-zero distance between two adjacent units. Interval scaled data may be distributed continuously or discretely.

2.3.2.13 An example for a discrete interval scaled characteristic is ‘Time of beginning of flowering’ measured as date which is given as example 5 in Table 6. This characteristic is defined as the number of days from April 1. The definition is useful but arbitrary and April 1 is not a natural limit. It would also be possible to define the characteristic as the number of days from January 1.

2.3.2.14 It is not possible to calculate a meaningful ratio between two varieties which is illustrated by the following example:

Variety ‘A’ begins to flower on May 30 and variety ‘B’ on April 30

Case I) Number of days from April 1 of variety ‘A’ = 60
Number of days from April 1 of variety ‘B’ = 30

$$\text{Ratio}_{\text{I}} = \frac{\text{Number of days from April 1 of variety 'A'}}{\text{Number of days from April 1 of variety 'B'}} = \frac{60}{30} = 2$$

Case II) Number of days from January 1 of variety ‘A’ = 150
Number of days from January 1 of variety ‘B’ = 120

$$\text{Ratio}_{II} = \frac{\text{Number of days from January 1 of variety 'A'}}{\text{Number of days from January 1 of variety 'B'}} = \frac{150}{120} = 1.25$$

$$\text{Ratio}_{I} = 2 > 1{,}25 = \text{Ratio}_{II}$$

2.3.2.15 It is incorrect to state that the time of flowering of variety 'A' is twice that of variety 'B'. The ratio depends on the choice of the zero point of the scale. This kind of scale is defined as an "Interval scale": a metric scale without a defined absolute zero point.

2.3.2.16 The interval scale is classified lower than the ratio scale (Table 1). At the interval scale, no useful indexes can be formed such as ratios. The interval scale is theoretically the minimum scale to calculate arithmetic mean values.

(c) Ordinal scale

2.3.2.17 Discrete categories of ordinally scaled data can be arranged in an ascending or descending order. They result from visually assessed (notes) quantitative characteristics.

Example:

Type of scale	Example	Example number*
ordinal	Intensity of anthocyanin	6

* For description of the states of expressions, see Table 6

2.3.2.18 An ordinal scale consists of numbers which correspond to the states of expression of the characteristic (notes). The expressions vary from one extreme to the other and thus they have a clear logical order. It is not important which numbers are used to denote the categories. In some cases ordinal data may reach the level of discrete interval scaled data or of discrete ratio scaled data (see section 2.3.7).

2.3.2.19 The distances between the discrete categories of an ordinal scale are not exactly known and not necessarily equal. Therefore, an ordinal scale does not fulfil the condition to calculate arithmetic mean values, which is the equality of intervals throughout the scale.

2.3.2.20 The ordinal scale is classified lower than the interval scale (Table 1). Fewer statistical procedures can be used for ordinal scale than for each of the higher classified scale data (see section 2.3.7).

2.3.3 Data from pseudo-qualitative characteristics

2.3.3.1 Data results from pseudo-qualitative characteristics are nominal scaled data without any logical order of all discrete categories. They result from visually assessed (notes) qualitative characteristics.

Example:

Type of scale	Example	Example number*
nominal	Shape	7
nominal	Flower color	8

* For description of the states of expressions, see Table 6.

2.3.3.2 A nominal scale consists of numbers which correspond to the states of expression of the characteristic, which are referred to in the Test Guidelines as notes. Although numbers are used for designation there is no inevitable order for all of the expressions. It is only possible to arrange some of them in an order.

2.3.3.3 The nominal scale is the lowest classification of the scales (Table 1). Few statistical procedures are applicable for evaluations (see section 2.3.7).

2.3.4 Summary of the different types of scales

Table 1: Types of expressions and type of scales

Type of expression	Type of scale	Description	Distribution	Data recording	Scale Level
QN	ratio	constant distances with absolute zero point	Continuous	Absolute measurements	High
			Discrete	Counting	
	interval	constant distances without absolute zero point	Continuous	Relative measurements	
			Discrete	Date	
	ordinal	Ordered expressions with varying distances	Discrete	Visually assessed notes	
PQ or QL	nominal	No order, no distances	Discrete	Visually assessed notes	Low

2.3.5 Scale levels for variety description

The description of varieties is based on the states of expression (notes) which are given in the Test Guidelines for the specific crop. In the case of visual assessment, the notes from the Test Guidelines are usually used for recording the characteristic as well as for the assessment of DUS. The notes are distributed on a nominal or ordinal scale (see section 2.3). For measured or counted characteristics, DUS assessment is based on the recorded values and the recorded values are transformed into states of expression only for the purpose of variety description.

2.3.6 Relation between types of expression of characteristics and scale levels of data

2.3.6.1 Records taken for the assessment of qualitative characteristics are distributed on a nominal scale, for example “Sex of plant”, “Leaf blade: variegation” (Table 6, examples 1 and 2).

2.3.6.2 For quantitative characteristics the scale level of data depends on the method of assessment. They can be recorded on a metric (when measured or counted) or ordinal (when visually observed) scale. For example, “Length of plant” can be recorded by measurements resulting in ratio scaled continuous metric data. However, visual assessment on a 1 to 9 scale may also be appropriate. In this case, the recorded data are ordinal scaled because the size of intervals between the midpoints of categories is not exactly the same.

Remark: In some cases visually assessed data on metric characteristics may be handled as measurements. The possibility to apply statistical methods for metric data depends on the precision of the assessment and the robustness of the statistical procedures. In the case of very precise visually assessed quantitative characteristics the usually ordinal data may reach the level of discrete interval scaled data or of discrete ratio scaled data.

2.3.6.3 A pseudo-qualitative type of characteristic is one in which the expression varies in more than one dimension. The different dimensions are combined in one scale. At least one dimension is quantitatively expressed. The other dimensions may be qualitatively expressed or quantitatively expressed. The scale as a whole has to be considered as a nominal scale (e.g. “Shape”, “Flower color”; Table 6, examples 7 and 8).

2.3.6.4 In the case of using the off-type procedure for the assessment of uniformity the recorded data are nominally scaled. The records fall into two qualitative classes: plants belonging to the variety (true-types) and plants not belonging to the variety (off-types). The type of scale is the same for qualitative, quantitative and pseudo-qualitative characteristics.

2.3.6.5 The relation between the type of characteristics and the type of scale of data recorded for the assessment of distinctness and uniformity is described in Table 2. A qualitative characteristic is recorded on a nominal scale for distinctness (state of expression) and for uniformity (true-types vs. off-types). Pseudo-qualitative characteristics are recorded on a nominal scale for distinctness (state of expression) and on a

nominal scale for uniformity (true-types vs. off-types). Quantitative characteristics are recorded on an ordinal, interval or ratio scale for the assessment of distinctness depending on the characteristic and the method of assessment. If the records are taken from single plants the same data may be used for the assessment of distinctness and uniformity. If distinctness is assessed on the basis of a single record of a group of plants, uniformity has to be judged with the off-type procedure (nominal scale).

Table 2: Relation between type of characteristic and type of scale of assessed data

Procedure	Type of scale	Distribution	Type of characteristic		
			Qualitative	Pseudo-qualitative	Quantitative
Distinctness	ratio	Continuous	No	No	**Yes**
		Discrete	No	No	**Yes**
	interval	Continuous	No	No	**Yes**
		Discrete	No	No	**Yes**
	ordinal	Discrete	No	No	**Yes**
	nominal	Discrete	**Yes**	**Yes**	No
Uniformity	ratio	Continuous	No	No	**Yes**
		Discrete	No	No	**Yes**
	interval	Continuous	No	No	**Yes**
		Discrete	No	No	**Yes**
	ordinal	Discrete	No	No	**Yes**
	nominal	Discrete	**Yes**	**Yes**	**Yes**

2.3.7 Relation between method of observation of characteristics, scale levels of data and recommended statistical procedures

2.3.7.1 Established statistical procedures can be used for the assessment of distinctness and uniformity considering the scale level and some further conditions such as the degree of freedom or unimodality (Tables 3 and 4).

2.3.7.2 The relation between the expression of characteristics and the scale levels of data for the assessment of distinctness and uniformity is summarized in Table 6.

Table 3: Statistical procedures for the assessment of distinctness

Type of Scale	Distribution	Observation method	Procedure	Further Condition	Reference document
ratio	continuous	MS MG (VS) [1]	COYD	at least 10 and preferably at least 20 df[3]	TGP/8 and 9
	discrete				
interval	continuous		long term COYD	df<10	TGP/8
			2x1% method	at least 10 and preferably at least 20 df	TGP/8
	discrete				
ordinal	discrete	VS	Pearson's Chi-Square test	$E_{ij} \geq 5$ [4]	TGP/8
		VS	Fisher's Exact test	$E_{ij} < 10$	TGP/8
		VS	GLM models Threshold models		
		VG	See also explanation for QN characteristics in TGP/9 sections 5.2.2 and 5.2.3 See explanation for QN characteristics in TGP/9 section 5.2.4		TGP/9
nominal	discrete	(VS) [2]	Pearson's Chi-Square test	$E_{ij} \geq 5$	TGP/8
		VS	Fisher's Exact test	$E_{ij} < 10$	TGP/8
		VS	GLM models	$E_{ij} \geq 5$	
		VG	See explanation for QL and PQ characteristics in TGP/9 sections 5.2.2 and 5.2.3		TGP/9

1) see remark in section 2.3.2.18
2) normally VG but VS would be possible
3) df – degree of freedom
4) E_{ij} – expected value of a class

Table 4: Statistical procedures for the assessment of uniformity

Type of scale	Distribution	Observation method	Procedure	Further Conditions	Reference document
ratio	continuous	MS	COYU	df≥20	TGP/8 and 10
	discrete	MS VS	Relative variance method	$s^2_c \leq 1.47\ s^2$	TGP/8
interval	continuous				
	discrete				
ordinal	discrete	VS	Threshold model		
nominal	discrete	VS	Off-type procedure for dichotomous (binary) data	Fixed population standard	TGP/8 and 10

2.4 Different levels to look at a characteristic

2.4.1 Characteristics can be considered in different levels of process (Table 5). The characteristics as expressed in the trial (type of expression) are considered as process level 1. The data taken from the trial for the assessment of distinctness, uniformity and stability are defined as process level 2. These data are transformed into states of expression for the purpose of variety description. The variety description is process level 3.

Table 5: Definition of different process levels to consider characteristics

Process level	Description of the process level
1	characteristics as expressed in trial
2	data for evaluation of characteristics
3	variety description

From the statistical point of view, the information level decreases from process level 1 to 3. Statistical analysis is only applied in level 2.

2.4.2 Sometimes for DUS experts it seems that there is no need to distinguish between different process levels. The process level 1, 2 and 3 could be identical. However, in general, this is not the case.

Understanding the need for process levels

2.4.3 The DUS expert may know from UPOV Test Guidelines or his own experience that, for example, “Length of plant” is a good characteristic for the examination of DUS. There are varieties which have longer plants than other varieties. Another characteristic could be ‘Variegation of leaf blade’. For some varieties, variegation is present and for others not. The DUS expert has now two characteristics and he knows that “Plant: length” is a quantitative characteristic and “Variegation of leaf blade’” is a qualitative characteristic (definitions: see sections 2.2.2 and 2.2.3). This stage of work can be described as **process level 1.**

2.4.4 The DUS expert then has to plan the trial and to decide on the type of observation for the characteristics. For characteristic “Variegation of leaf blade”, the decision is clear. There are two possible expressions: “present” or “absent”. The decision for characteristic “Plant length” is not specific and depends on expected differences between the varieties and on the variation within the varieties. In many cases, the DUS expert will decide to measure a number of plants (in cm) and to use special statistical procedures to examine distinctness and uniformity. But it could also be possible to assess the characteristic “Plant length” visually by using expressions like ”short”, “medium” and “long”, if differences between varieties are large enough (for distinctness) and the variation within varieties is very small or absent in this characteristic. The continuous variation of a characteristic is assigned to appropriate states of expression which are recorded by notes (see document TGP/9, section 4). The crucial element in this stage of work is the recording of data for further evaluations. It is described as **process level 2**.

2.4.5 At the end of the DUS test, the DUS expert has to establish a description of the varieties using notes from 1 to 9 or parts of them. This phase can be described as **process level 3**. For “Variegation

of leaf blade" the DUS expert can take the same states of expression (notes) he recorded in process level 2 and the three process levels appear to be the same. In cases where the DUS expert decided to assess "Plant: length" visually, he can take the same states of expression (notes) he recorded in process level 2 and there is no obvious difference between process level 2 and 3. If the characteristic "Plant: length" is measured in cm, it is necessary to assign intervals of measurements to states of expressions like "short", "medium" and "long" to establish a variety description. In this case, for statistical procedures, it is important to be clearly aware of the relevant level and to understand the differences between characteristics as expressed in the trial, data for evaluation of characteristics and the variety description. This is absolutely necessary for choosing the most appropriate statistical procedures in cooperation with statisticians or by the DUS expert.

Table 6: Relation between expression of characteristics and scale levels of data for the assessment of distinctness and uniformity

Example	Name of characteristic	Distinctness				Uniformity			
		Unit of assessment	Description (states of expression)	Type of scale	Distri-bution	Unit of assessment	Description (states of expression)	Type of scale	Distri-bution
1	Sex of plant	1 2 3 4	dioecious female dioecious male monoecious unisexual monoecious hermaphrodite	nominal	discrete	True-type Off-type	Number of plants belonging to the variety Number of off-types	nominal	discrete
2	Leaf blade: variegation	1 9	absent present	nominal	discrete	True-type Off-type	Number of plants belonging to the variety Number of off-types	nominal	discrete
3	Length of plant	cm	assessment in cm without digits after decimal point	ratio	conti-nuous	cm	assessment in cm without digits after decimal point	ratio	continuous
						True-type Off-type	Number of plants belonging to the variety Number of off-types	nominal	discrete
4	Number of stamens	counts	1, 2, 3, ... , 40, 41, ...	ratio	discrete	counts	1, 2, 3, ... , 40, 41, ...	ratio	discrete
5	Time of beginning of flowering	Date	e.g. May 21, 51st day from April 1	interval	discrete	Date	e.g. May 21, 51st day from April 1	interval	discrete
						True-type Off-type	Number of plants belonging to the variety Number of off-types	nominal	discrete
6	Intensity of anthocyanin	1 2 3 4 5 6 7 8 9	very low very low to low low low to medium medium medium to high high high to very high very high	ordinal	discrete (with an underlying quanti-tative variable)	True-type Off-type	Number of plants belonging to the variety Number of off-types	nominal	discrete

Example	Name of characteristic	Distinctness				Uniformity			
		Unit of assessment	Description (states of expression)	Type of scale	Distri-bution	Unit of assessment	Description (states of expression)	Type of scale	Distri-bution
7	Shape	1 2 3 4 5 6 7	deltate ovate elliptic obovate obdeltate circular oblate	nominal	discrete	True-type Off-type	Number of plants belonging to the variety Number of off-types	nominal	discrete
8	Flower color	1 2 3 4 5 6 7 8 9 10	dark red medium red light red white light blue medium blue dark blue red violet violet blue violet	nominal	discrete	True-type Off-type	Number of plants belonging to the variety Number of off-types	nominal	discrete

3. MINIMIZING THE VARIATION DUE TO DIFFERENT OBSERVERS OF THE SAME TRIAL

3.1 Introduction

3.1.1 This section considers variation between observers of the same trial at the authority level. It has been prepared with QN/MG, QN/MS, QN/VG and QN/VS characteristics in mind. It does not explicitly deal with PQ characteristics like color and shape. The described Kappa method in itself is largely applicable for these characteristics, e.g. the standard Kappa characteristic is developed for nominal data. However, the method has not been developed for PQ characteristics and may also require extra information on calibration. As an example, for color calibration, you also have to take into account the RHS Colour chart, the lighting conditions and so on. Differences between observers on PQ characteristics could be tested using non-parametric methods, such as frequency of deviations. These aspects are not covered in this document.

3.1.2 Variation in measurements or observations can be caused by many different factors, like the type of crop, type of characteristic, year, location, trial design and management, method and observer. Especially for visually assessed characteristics (QN/VG or QN/VS) differences between observers can be the reason for large variation and potential bias in the observations. An observer might be less well trained, or have a different interpretation of the characteristic. So, if observer A assesses variety 1 and observer B variety 2, the difference observed might be caused by differences between observers A and B instead of differences between varieties 1 and 2. Clearly, our main interest lies with the differences between varieties and not with the differences between the observers. It is important to realize that the variation caused by different observers cannot be eliminated, but there are ways to control it.

3.1.3 It is recommended that, wherever possible, one observer should be used per trial to minimize variation in observations due to different observers.

3.2 Training and importance of clear explanations of characteristics and method of observation

3.2.1 Training of new observers is essential for consistency and continuity of plant variety observations. Calibration manuals, supervision and guidance by experienced observers as well as the use of example varieties illustrating the range of expressions are useful ways to achieve this.

3.2.2 UPOV test guidelines try to harmonize the variety description process and describe as clearly as possible the characteristics of a crop and the states of expression. This is the first step in controlling variation and bias. However, the way that a characteristic is observed or measured may vary per year, location or testing authority. Calibration manuals made by the local testing authority and example varieties are very useful for the local implementation of the UPOV test guideline. Where needed these crop-specific manuals explain the characteristics to be observed in more detail, and specify when and how they should be observed. Furthermore they may contain pictures and drawings for each characteristic, often for every state of expression of a characteristic.

3.2.3 The Glossary of Terms Used in UPOV Documents (document TGP/14) provides useful guidance for clarifying many characteristics, in particular PQ characteristics.

3.2.4 Once an observer is trained it is important to ensure frequent refresher training and recalibration.

3.3 Testing the calibration

3.3.1 After training an observer, the next step could be to test the performance of the observers in a calibration experiment. This is especially useful for inexperienced observers who have to make visual observations (QN/VG and QN/VS characteristics). If making visual observations, they should preferably pass a calibration test prior to making observations in the trial. But also for experienced observers, it is useful to test themselves on a regular basis to verify if they still fulfill the calibration criteria.

3.3.2 A calibration experiment can be set up and analyzed in different ways. Generally it involves multiple observers, measuring the same set of material and assessing differences between the observers.

3.4 Testing the calibration for QN/MG or QN/MS characteristics

3.4.1 For observations made by measurement tools, like rulers (often QN/MS characteristics), the measurement is often made on an interval or ratio scale. In this case, the approach of Bland and Altman (1986)

can be used. This approach starts with a plot of the scores for a pair of observers in a scatter plot, and compare it with the line of equality (where y=x). This helps the eye gauging the degree of agreement between measurements of the same object. In a next step, the difference per object is taken and a plot is constructed with on the y-axis the difference between the observers and on the x-axis either the index of the object, or the mean value of the object. By further drawing the horizontal lines y=0, y=mean(difference) and the two lines y = mean(difference) ± 2 x standard deviation, the bias between the observers and any outliers can easily be spotted. Similarly we can also study the difference between the measurement of each observer and the average measurement over all observers. Test methods like the paired t-test can be applied to test for a significant deviation of the observer from another observer or from the mean of the other observers.

3.4.2 By taking two measurements by each observer of every object, we can look at the differences between these two measurements. If these differences are large in comparison to those for other observers, this observer might have a low repeatability. By counting for each observer the number of moderate and large outliers (e.g. larger than 2 times and 3 times the standard deviation respectively) we can construct a table of observer versus number of outliers, which can be used to decide if the observer fulfills quality assurance limits.

3.4.3 Other quality checks can be based on the repeatability and reproducibility tests for laboratories as described in ISO 5725-2. Free software is available on the ISTA website to obtain values and graphs according to this ISO standard.

3.4.4 In many cases of QN/MG or QN/MS, a good and clear instruction usually suffices and variation or bias in measurements between observers is often negligible. If there is reason for doubt, a calibration experiment as described above can help in providing insight in the situation.

3.4.5 In the case of QN/MG observations consideration and allowance may need to be given to the possible random within plot variation.

3.5 Testing the calibration for QN/VS or QN/VG characteristics

3.5.1 For the analysis of ordinal data (QN/VS or QN/VG characteristics), the construction of contingency tables between each pair of observers for the different scores is instructive. A test for a structural difference (bias) between two observers can be obtained by using the Wilcoxon Matched-Pairs test (often called Wilcoxon Signed-Ranks test).

3.5.2 To measure the degree of agreement the Cohen's Kappa (κ) statistic (Cohen, 1960) is often used. The statistic tries to accounts for random agreement:$\kappa = (P(agreement) - P(e)) / (1-P(e))$, where P(agreement) is the fraction of objects which are in the same class for both observers (the main diagonal in the contingency table), and P(e) is the probability of random agreement, given the marginals (like in a Chi-square test). If the observers are in complete agreement the Kappa value $\kappa = 1$. If there is no agreement among the observers, other than what would be expected by chance (P(e)), then $\kappa = 0$.

3.5.3 The standard Cohen's Kappa statistic only considers perfect agreement versus non-agreement. If one wants to take the degree of disagreement into account (for example with ordinal characteristics), one can apply a linear or quadratic weighted Kappa (Cohen, 1968). If we want to have a single statistic for all observers simultaneously, a generalized Kappa coefficient can be calculated. Most statistical packages, including SPSS, Genstat and R (package Concord), provide tools to calculate the Kappa statistic.

3.5.4 As noted, a low κ-value indicates poor agreement and values close to 1 indicate excellent agreement. Often scores between 0.6-0.8 are considered to indicate substantial agreement, and above 0.8 to indicate almost perfect agreement. If needed, z-scores for kappa (assuming an approximately normal distribution) are available. The criteria for experienced DUS experts could be more stringent than for inexperienced staff.

3.6 Trial design

If we have multiple observers in a trial, the best approach is to have one person observe one or more complete replications. In that case, the correction for block effects also accounts for the bias between observers. If more than one observer per replication is needed, extra attention should be given to calibration and agreement. In some cases, the use of incomplete block designs (like alpha designs) might be helpful, and an observer can be assigned to the sub blocks. In this way we can correct for systematic differences between observers.

3.7 Example of Cohen's Kappa

In this example, there are three observers and 30 objects (plots or varieties).The characteristic is observed on a scale of 1 to 6. The raw data and their tabulated scores are given in the following tables:

Variety	Observer 1	Observer 2	Observer 3
V1	1	1	1
V2	2	1	2
V3	2	2	2
V4	2	1	2
V5	2	1	2
V6	2	1	2
V7	2	2	2
V8	2	1	2
V9	2	1	2
V10	3	1	3
V11	3	1	3
V12	3	2	2
V13	4	5	4
V14	2	1	1
V15	2	1	2
V16	2	2	3
V17	5	4	5
V18	2	2	3
V19	1	1	1
V20	2	2	2
V21	2	1	2
V22	1	1	1
V23	6	3	6
V24	5	6	6
V25	2	1	2
V26	6	6	6
V27	2	6	2
V28	5	6	5
V29	6	6	5
V30	4	4	4

Scores for variety	1	2	3	4	5	6
V1	3	0	0	0	0	0
V2	1	2	0	0	0	0
V3	0	3	0	0	0	0
V4	1	2	0	0	0	0
V5	1	2	0	0	0	0
V6	1	2	0	0	0	0
V7	0	3	0	0	0	0
V8	1	2	0	0	0	0
V9	1	2	0	0	0	0
V10	1	0	2	0	0	0
V11	1	0	2	0	0	0
V12	0	2	1	0	0	0
V13	0	0	0	2	1	0
V14	2	1	0	0	0	0
V15	1	2	0	0	0	0
V16	0	2	1	0	0	0
V17	0	0	0	1	2	0
V18	0	2	1	0	0	0
V19	3	0	0	0	0	0
V20	0	3	0	0	0	0
V21	1	2	0	0	0	0
V22	3	0	0	0	0	0
V23	0	0	1	0	0	2
V24	0	0	0	0	1	2
V25	1	2	0	0	0	0
V26	0	0	0	0	0	3
V27	0	2	0	0	0	1
V28	0	0	0	0	2	1
V29	0	0	0	0	1	2
V30	0	0	0	3	0	0

The contingency table for observer 1 and 2 is:

O1\O2	1	2	3	4	5	6	Total
1	3	0	0	0	0	0	3
2	10	5	0	1	0	1	17
3	2	1	0	0	0	0	3
4	0	0	0	1	0	0	1
5	0	0	0	1	0	2	3
6	0	0	1	0	0	2	3
Total	15	6	1	3	0	5	30

The Kappa coefficient between observer 1 and 2, κ(O1,O2) is calculated as follows:

- κ(O1,O2) = (P(agreement between O1 and O2) – P(e)) / (1 – P(e)) where:
- P(agreement) = (3+5+0+1+0+2)/30 = 11/30 ≈ 0.3667 (diagonal elements)
- P(e) = (3/30).(15/30) + (17/30).(6/30) + (3/30).(1/30) + (1/30).(3/30) + (3/30).(0/30) + (3/30).(5/30) ≈ 0.1867. (pair-wise margins)
- So κ(O1,O2) ≈ (0.3667-0.1867) / (1-0.1867) ≈ 0.22

This is a low value, indicating very poor agreement between these two observers. There is reason for concern and action should be taken to improve the agreement. Similarly the values for the other pairs can be calculated: κ(O1,O3) ≈ 0.72, κ(O2,O3) ≈ 0.22.Observer 1 and 3 are in good agreement. Observer 2 is clearly different from 1 and 3 and the reasons for the deviation requires further investigation (e.g. consider need for additional training).

3.8 References

Cohen, J. (1960) A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20: 37-46.

Cohen, J. (1968) Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4): 213-220.

Bland, J. M. Altman D. G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement, Lancet: 307–310.

http://www.seedtest.org/en/stats-tool-box-_content---1--1143.html (ISO 5725-2 based software)

4. VALIDATION OF DATA AND ASSUMPTIONS

4.1 Introduction

It is important that the data are correct, i.e. without mistake. This is the case irrespective of whether the data are notes obtained from visual observation (V) (see document TGP/9 section 4.2.1) or measurement (M) (see document TGP/9 section 4.2.2) and whether they result in a single record for a group of plants (G) (See document TGP/9 section 4.3.2) or whether they result in records for a number of single, individual plants or part of plants (S) (see document TGP/9 section 4.3.3) for statistical analysis. Section "Validation of data" describes how the data can be validated or checked. These preliminary checks can be done on all data, whether or not they are subsequently analyzed by statistical methods.

4.2 Validation of data

4.2.1 This section is concerned with validating the data to ensure that there are no (obvious) mistakes.

4.2.2 In order to avoid mistakes in the interpretation of the results the data should always be inspected so that the data are logically consistent and not in conflict with prior information about the ranges likely to arise for the various characteristics. This inspection can be done manually (usually visually) or automatically. When statistical methods are used, the validation of assumptions can also be used as a check that the data are without mistakes (see section 2.3.2.1.1.)

4.2.3 Table 1 shows an extract of some recordings for 10 plants from a plot of field peas. For 'Seed: shape' (PQ) the notes are visually scored on a scale with values 1 (spherical), 2 (ovoid), 3 (cylindrical), 4 (rhomboid), 5 (triangular) or 6 (irregular). For Seed: black color of hilum (QL), the notes are visually scored on a scale with values 1 (absent) or 9 (present). For 'Stem: length' (QN) the measurements are in cm and from past experience it is known that the length in most cases will be between 40 and 80 cm. The 'Stipule: length' is measured in mm and will in most cases be between 50 and 90 mm. The table shows 3 types of mistakes which occasionally occur when making manual recordings: for plant 4, 'Seed: shape' the recorded value, 7, is not among the allowed notes and must, therefore, be due to a mistake. It might be caused by misreading a hand-written "1". A similar situation is seen for plant 8 for characteristic 'Seed: black color of hilum', where note 8 is not allowed and must be a mistake. The 'Stem: length' of plant 6 is outside the expected range and could be caused by changing the order of the figures, so 96 has been keyed instead of 69. The 'Stipule: length' of 668 mm is clearly wrong. It might be caused by accidentally repeating the figure 6 twice. In all cases a careful examination needs to be carried out in order to find out what the correct values should be.

Table 1: Extract of recording sheet for field peas

Plant no	Seed: shape (UPOV 1) (PQ)	Seed: black color of hilum (UPOV 6) (QL)	Stem: length (UPOV 12) (QN)	Stipule: length (UPOV 31) (QN)
1	1	1	43	80
2	2	1	53	79
3	1	1	50	72
4	7	1	43	668
5	2	9	69	72
6	1	1	96	72
7	1	1	51	70
8	2	8	64	63
9	1	1	44	62
10	2	1	49	62

4.2.4 Graphical displays, or plots of the characteristics, may help to validate the data. For example, examination of the frequency distributions of the characteristics may identify small groups of discrepant observations. Also, in the case of quantitative characteristics, examination of scatter plots of pairs of characteristics that are likely to be highly related may detect discrepant observations very efficiently.

4.2.5 Other types of graphical 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 quantitative data. In a box-plot a box is drawn for each group (plot

or variety). In this case, data of 'Leaf: length' (in mm) are used from an experiment laid out in 3 blocks of 26 plots with 20 plants per plot. Within each block, 26 different oilseed rape varieties were randomly assigned to each plot. In Figure 1, all 60 'Leaf: lengths' of each of the 26 varieties are taken together. (If there are large block differences a better box-plot can be produced by taking the differences with respect to the plot mean). 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, more extreme observations 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).

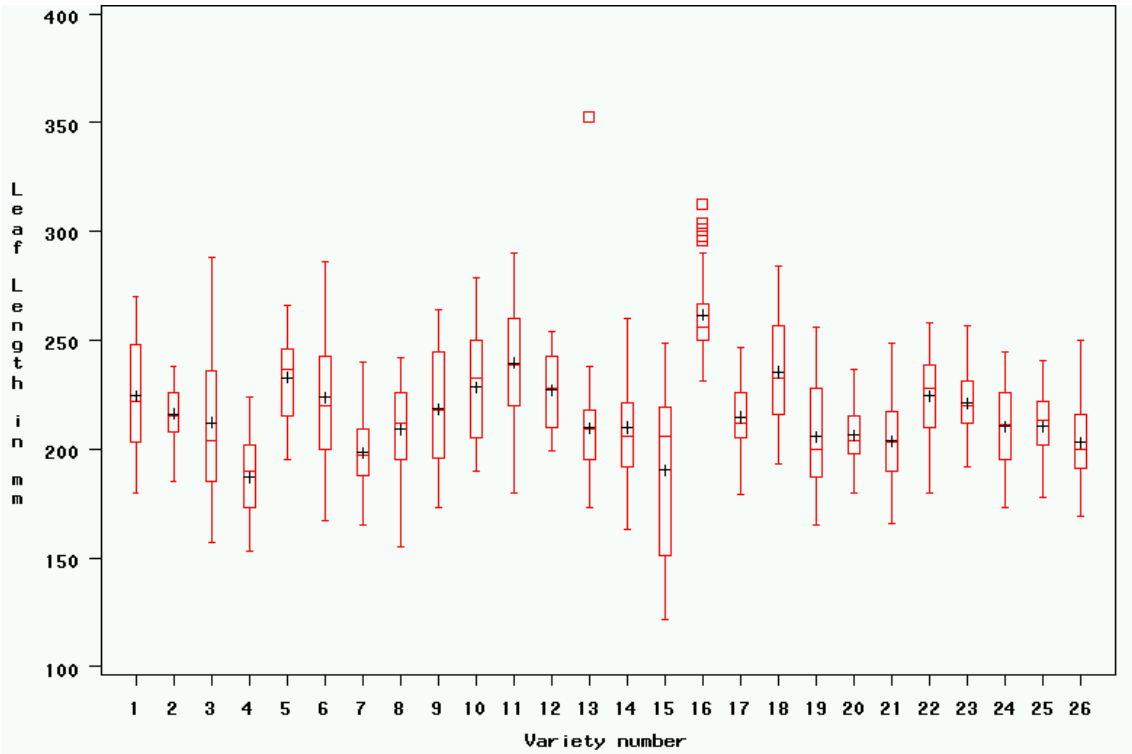


Figure 1. Box-plot for leaf length of 26 varieties of oil seed rape

4.2.6 When discrepant observations are found, it is important to try 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 mistake has occurred. In the latter 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 there are good reasons.

4.3 Assumptions for statistical analysis and the validation of these assumptions

If data are to be statistically analyzed, then the assumptions behind the theory on which the statistical methods are based must be met - at least approximately. This section describes the assumptions behind the most common statistical analysis methods used in DUS testing. It is followed by a section on the validation of the assumptions required for statistical analysis: it describes how they may be evaluated.

The methods described here for the validation of the assumptions behind the statistical methods are 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.

4.3.1 Assumptions for statistical analysis involving analysis of variance

4.3.1.1 Introduction

4.3.1.1.1 Firstly, it is essential that the growing trial/experiment is designed properly and involves randomization. The most important assumptions of analysis of variance methods are:

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

4.3.1.1.2 One could also state that there should be no mistakes in the data. However, it is not necessary to state this as an assumption. Firstly, because it is already covered in the previous section on validation of data, and secondly because if there are mistakes (or at least large ones) it will result in failure of the above assumptions, as the observations will not be normally distributed and they will have different variances (non-homogeneity of variances).

4.3.1.1.3 The assumptions mentioned here are most important when the statistical methods based on the Method of Least Squares are used to test hypotheses. When such statistical methods are used only to estimate effects (means), the assumptions are less important and the assumption of normally distributed observations is not necessary.

4.3.1.2 Independent observations

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 has not been built into COYD and COYU or the other methods included in document_TGP/8). Dependency may be caused by e.g. competition between neighboring plots, lack of randomization or improper randomization. More details on ensuring independence of observations may be found in Part I: section 1.5.3.3.7 “Trial elements when statistical analysis is used”.

4.3.1.3 Variance homogeneity

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:

(i) The variance depends on the mean, e.g. 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 of characteristics are: the logarithmic transformation (where the standard deviation is approximately proportional to the mean), the square-root transformation (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).

(ii) The variance depends on for example, variety, year or block. If the variances depend on such variables in a way that is not connected to the mean value, it is not possible to obtain variance homogeneity by transformation. In such cases it might be necessary either to use more sophisticated 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 trial with 10 varieties where varieties A, B, C, D, E, F, G and H each have a variance of 5, whereas varieties I and J each have a variance of 10. The real probability of detecting differences between these varieties when, in fact, they have the same mean is shown in Table 2. In Table 2, 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 too difficult to detect differences between two varieties with small variances and too easy to detect differences between varieties with large variances.

Table 2. Real probability of significant difference between two identical varieties in the case where variance homogeneity 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 names	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%

4.3.1.4 Normal distributed observations

The residuals should be approximately normally distributed. The residual is the part of an observation that remains unexplained after fitting a model. It is the difference between the observation and the prediction from the model. 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 residuals 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 residuals deviates from the normal distribution. However, deviation from normality is usually not as serious as deviations from the previous two assumptions.

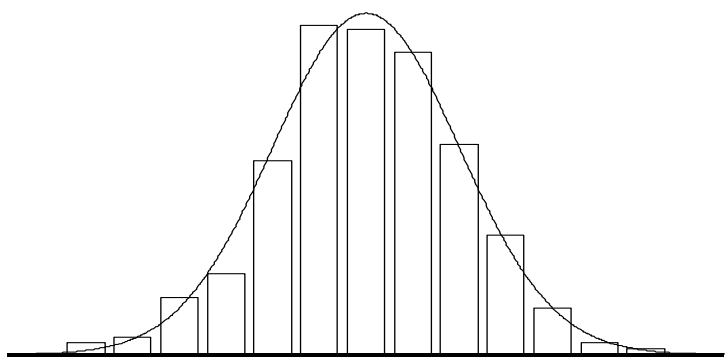


Figure 2. Histogram for normal distributed data with the ideal normal distribution shown as a curve

4.3.1.5 Additivity of block and variety effects

4.3.1.5.1 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. This means that the effect of a given variety is the same in all blocks. This is demonstrated in Table 3 where plot means of artificial data (of leaf length in mm) are given for two small experiments with three blocks and four varieties. In Experiment I, the effects of blocks and varieties are additive because the differences between any two varieties are the same in all blocks, e.g. the differences between variety A and B are 4 mm in all three blocks. In Experiment II, the effects are not additive, e.g. the differences between variety A and B are 2, 2 and 8 mm in the three blocks.

Table 3. Artificial plot means of leaf length in mm 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	240	242	239
B	244	246	243
C	245	247	244
D	241	243	240

Experiment II			
Variety	Block		
	1	2	3
A	240	242	239
B	242	244	247
C	246	244	243
D	241	242	241

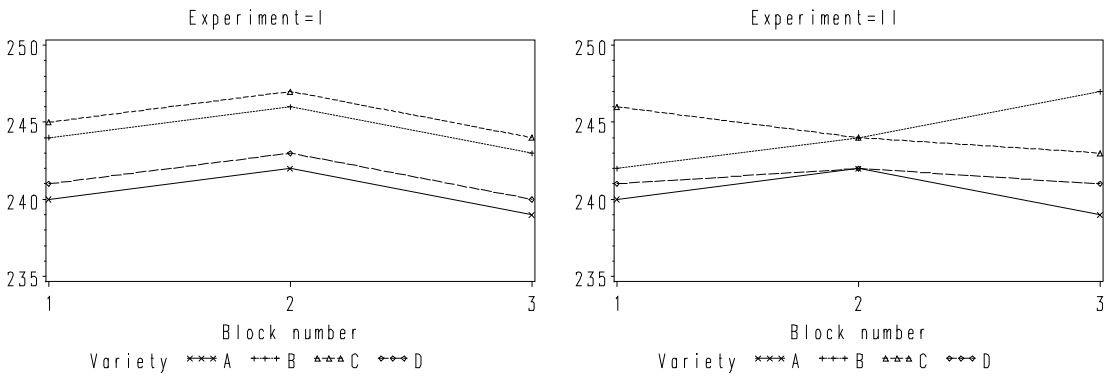


Figure 3. Artificial plot means from two experiments showing additive block and variety effects (left) and non-additive block and variety effects (right) using same data as in table 3.

4.3.1.5.2 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 produces 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 may be masked by random variation.

4.3.2 Validation of assumptions for statistical analysis

4.3.2.1 Introduction

4.3.2.1.1 The main purpose of validation is to check that the assumptions underlying the statistical analyses are fulfilled. However, it also serves as a secondary check that the data are without mistakes.

4.3.2.1.2 There are different methods to use when validating the assumptions. Some of these are:

- look through the data to verify the assumptions
- produce plots or figures to verify the assumptions
- make formal statistical tests for the different types of assumptions. In the literature several methods to test for outliers, variance homogeneity, additivity 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 statistically significant in large datasets)

4.3.2.2 Looking through the data

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 tedious 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.

4.3.2.3 Using figures

4.3.2.3.1 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).

4.3.2.3.2 The plot of the residuals versus the predicted values may be used to judge the dependence of the variance on the mean. If there is no dependence, 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 mistakes, 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 in figure 4, 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 a mistake and also that a mistake will not necessarily show up as an outlier.

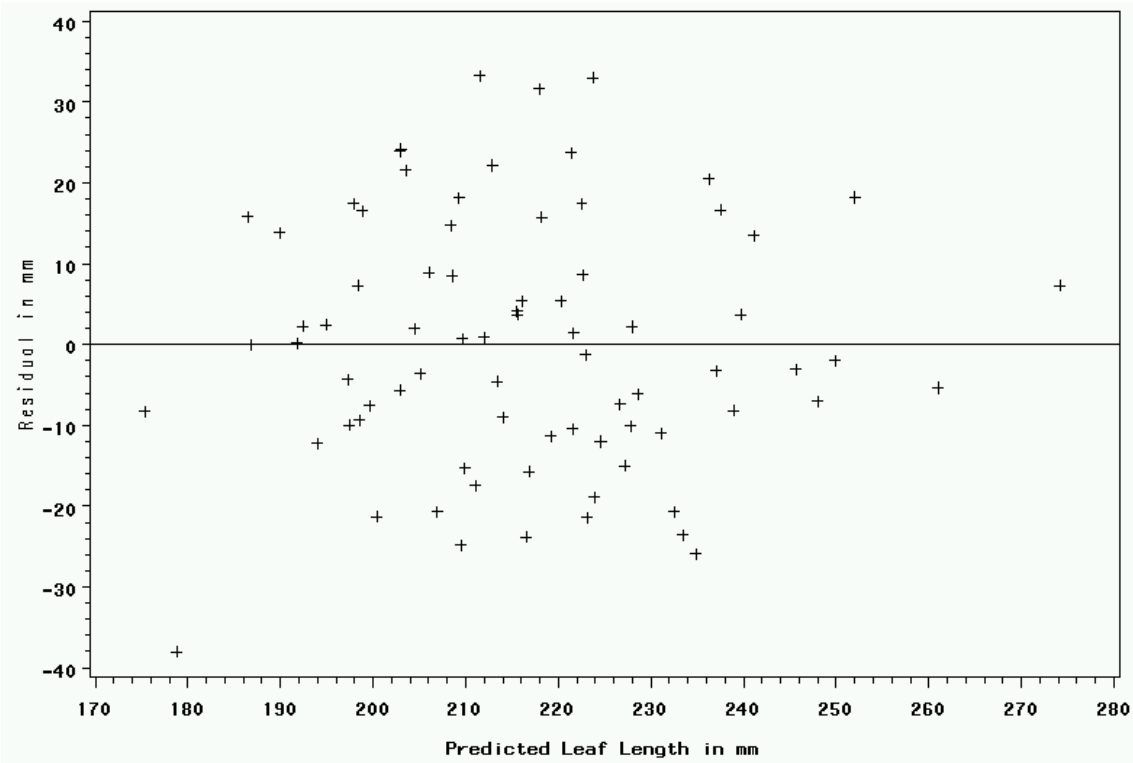


Figure 4. Plot of residuals versus plot predicted values for leaf length in 26 oil seed rape varieties in 3 blocks

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

4.3.2.3.4 The range (maximum value minus minimum 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 varieties with an extremely large variation (all plots of the variety with a large value) or plots where the variation is extremely large (maybe caused by a single plant). It is clearly seen that the range for one of variety 13's plots is much higher than in the other two plots. Also the range in one of variety 3's plots seems to be relatively large.

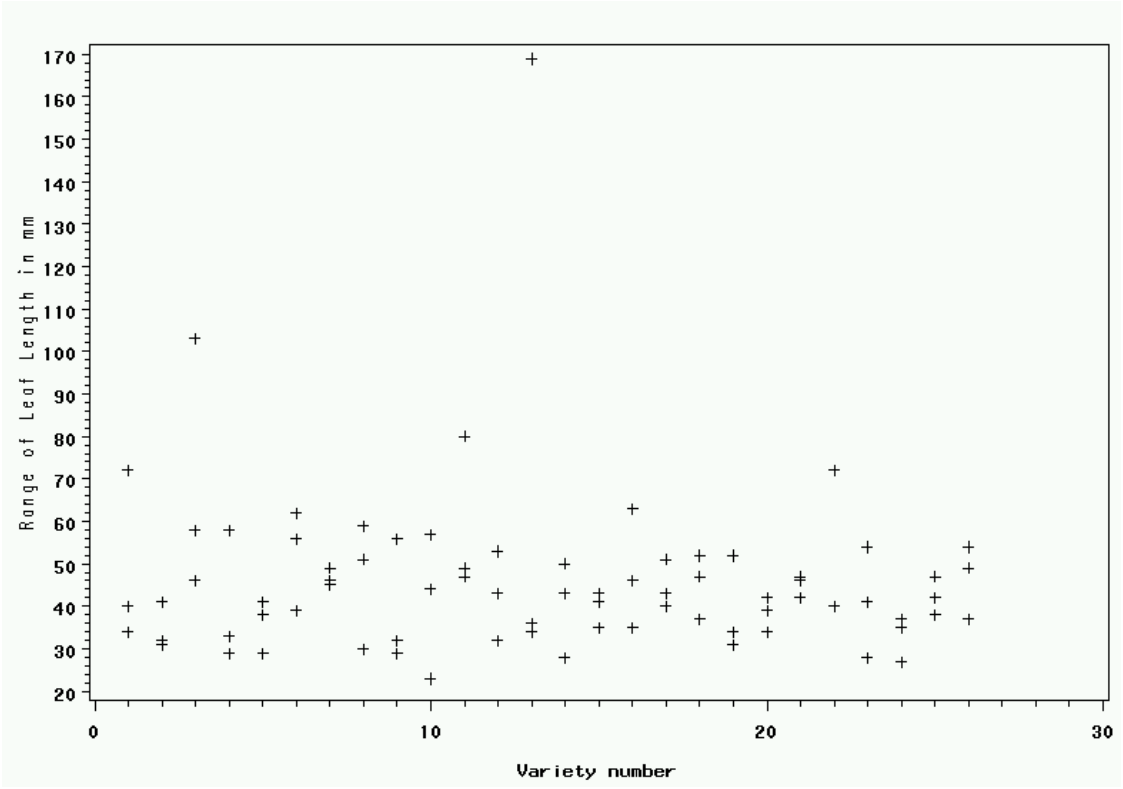


Figure 5. Differences between minimum and maximum of 20 leaf lengths for 3 plots versus oil seed rape variety number

4.3.2.3.5 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 are numbered such that the numbers indicate the relative location. In the example shown in Figure 6, 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).

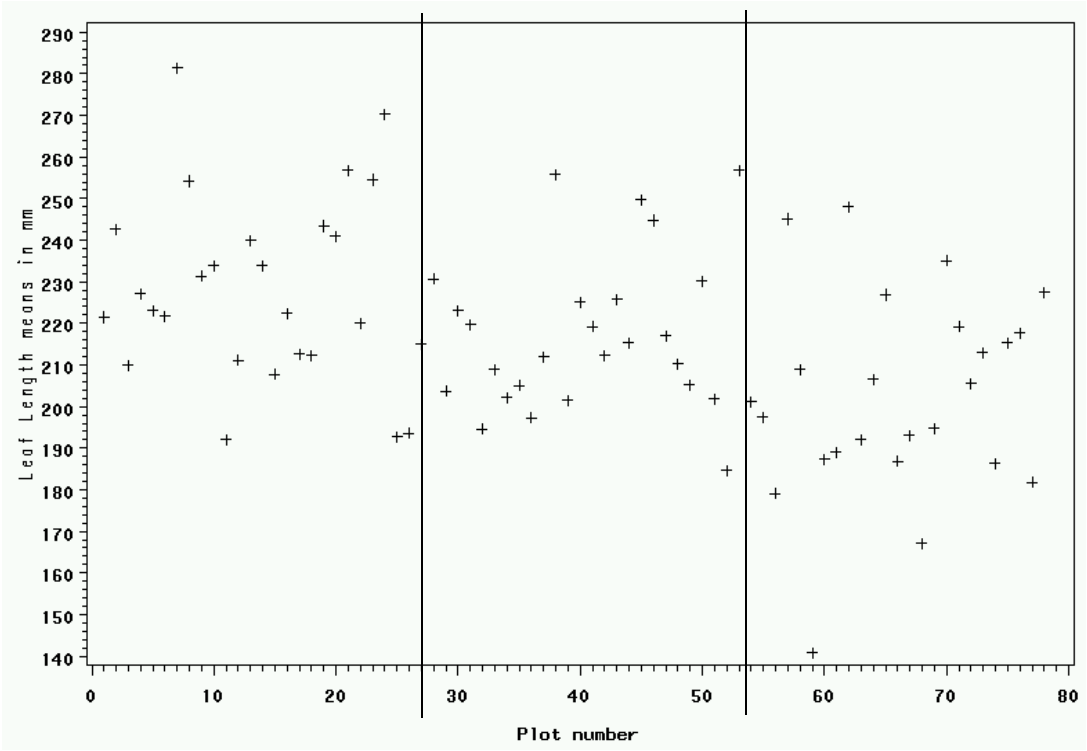


Figure 6. Plot means of 20 leaf lengths versus plot numbers

4.3.2.3.6 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).

4.3.2.3.7 Normal Probability Plots (Figure 7). This type of graph is used to evaluate to what extent the distribution of the variable follows the normal distribution. The selected variable will be plotted in a scatter plot against the values “expected from the normal distribution.” The standard normal probability plot is constructed as follows. First, the residuals (deviations from the predictions) 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.

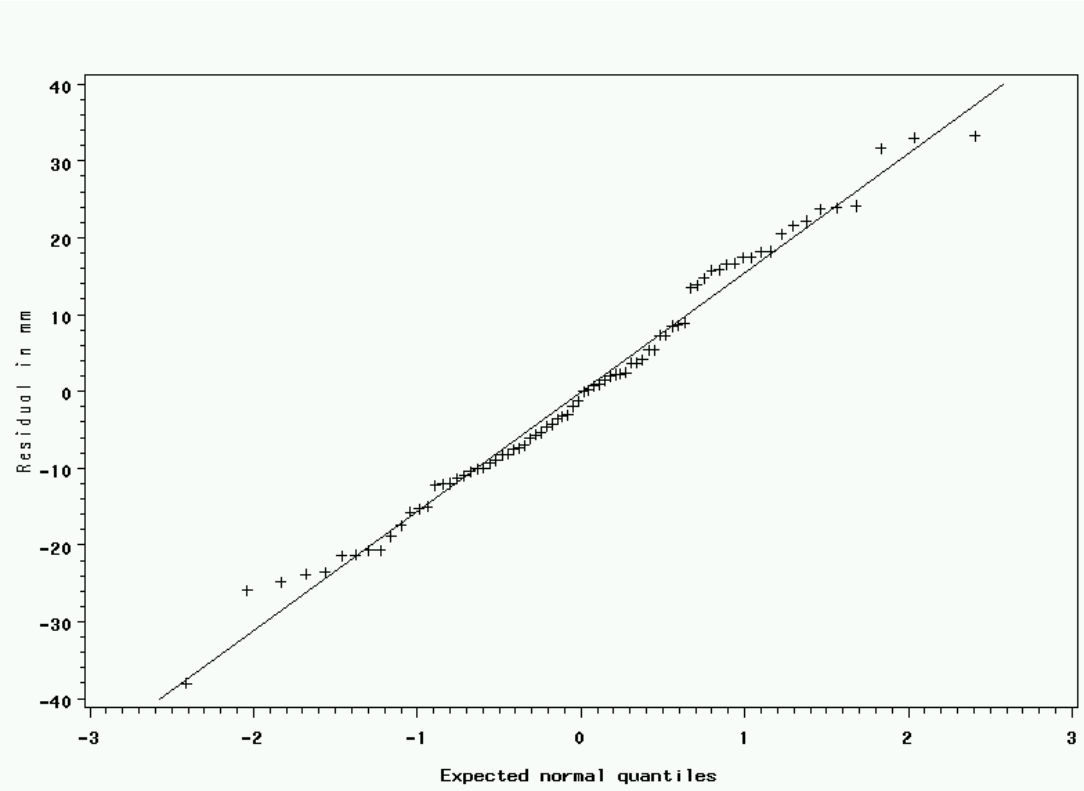


Figure 7. Normal probability plot for the residuals of leaf length in 26 oil seed rape varieties in 3 blocks

5. CHOICE OF STATISTICAL METHODS FOR EXAMINING DISTINCTNESS

5.1 Introduction

5.1.1 This section addresses some general considerations when choosing suitable statistical methods for the assessment of distinctness. It contains a discussion of factors influencing the choice of method and, as the statistical test used by each method is an essential part of that method, it includes a brief discussion of statistical tests, factors influencing their selection and some comments on their usefulness in particular situations.

5.1.2 Statistical methods are most commonly used for the assessment of distinctness of measured quantitative characteristics for cross-pollinated varieties when the data from the growing trial for a variety are subject to variation. Because of this variation, distinctness criteria based on statistical methods are needed in order to separate genuine varietal differences from chance variation and so make decisions about whether the candidate variety is distinct with a certain level of confidence that the decision is the correct one.

5.1.3 The variation may occur for example from plant to plant, from plot to plot and from year to year. Whether a single growing cycle or more than a single growing cycle is needed to provide assurance that the differences observed between varieties are sufficiently consistent will depend on the levels or amounts of variation from these different sources that are observed in a species. Section 1.2 of PART I of this document provides information on growing cycles.

5.2 Statistical methods for use with two or more independent growing cycles

5.2.1 Introduction

5.2.1.1 A number of different statistical methods have been developed to assess distinctness when there are at least two independent growing cycles. The choice of which method to use depends partly on the species and partly on whether the trial and data requirements for the different statistical methods are met. Where those requirements are not met, such as where only one, or very few, known varieties exist for a taxon, and so a large trial is not possible, then other suitable approaches might be used.

5.2.1.2 The principles common to suitable statistical methods used to assess distinctness when there are at least two independent growing cycles include:

- statistical tests of the differences between variety means are used to determine whether the differences between varieties in the expression of their characteristics are significant.
- a requirement for the differences to be consistent across the different growing cycles. This requirement may be part of the statistical test as in the COYD method, or not part of the statistical test as in the 2x1% and Match methods.

For the sake of brevity, in the following, the term ‘year’ is used, though for these purposes it is interchangeable with the term ‘independent growing cycle’.

5.2.1.3 Examples of suitable statistical methods include:

(a) The COYD and long-term COYD methods to assess distinctness, which have been developed by UPOV to analyze data from two or more years of growing trials where there are either at least a certain minimum number of varieties in trial or data from sufficient trials in earlier years. Whether differences are sufficiently consistent is assessed using a statistical test based on a two-tailed LSD to assess whether differences in over-year variety means are significant. Details of the COYD and long-term COYD methods and the requirements for their use are given in document TGP/8 Part II section 3.

(b) The 2x1% method to assess distinctness, which has also been developed by UPOV to analyze data from two or more years of growing trials. Differences are assessed in each year using a statistical test based on a two-tailed LSD to compare the within-year variety means. Whether differences are sufficiently consistent is determined by the requirement that two varieties are significantly different in the same direction at the 1% level in both years, or, where trials are conducted in three years, in at least two out of three years. Details of the 2x1% method and how it compares with the COYD method are given in document TGP/8 Part II section 4.

(c) The Match approach to assess distinctness was developed to analyze data from more than one growing cycle. Trials are conducted by the breeder in the first growing cycle and examined by the testing authority in the second growing cycle (see document TGP/6 “Arrangements for DUS Testing”, Section 2 “Examples of Arrangements for DUS Testing”). Whether differences are sufficiently consistent is assessed using a statistical test (e.g. LSD, Multiple Range Test (MRT), Chi-Square or Fischer’s Exact) to gauge whether the differences in the second growing cycle are significant and agree with the “direction of the differences” declared by the breeders in the first growing cycle. The choice of statistical test depends on the type of expression of the characteristic concerned. Details of the Match method approach are given in document TGP/8 Part II, Section 7.

5.2.1.4 In the context of consistency and harmonization, it should be noted that different statistical methods may produce different results.

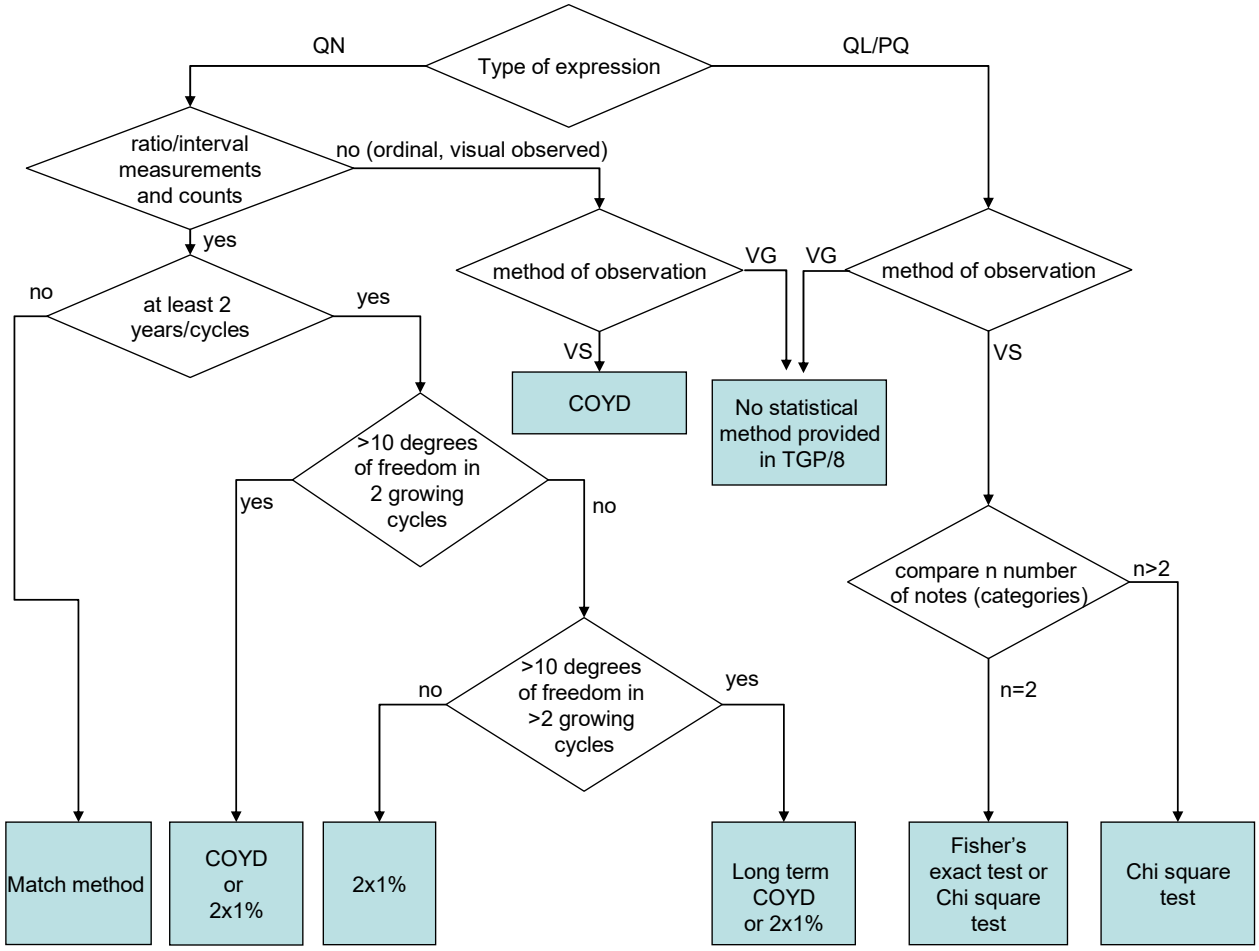
5.3 Summary of selected statistical methods for examining distinctness

5.3.1 Selected methods used in DUS examination

		Method of observation	Minimum Number of years/growing cycles	Minimum Degrees of freedom	Hypothesis to be tested	Type of characteristic	Other
COYD*		MS/VS	2	20 in two years/ growing cycles	D/non-D for variety means	QN	-
Long Term COYD*		MS/VS	2	20 (using data from more than 2 years/growing cycles)	D/non-D for variety means	QN	-
2x1 %		MS	2		D/non-D for variety means	QN	-
Match methods	Chi square	VS	-	-	Hypothesis for D based on previously known facts or principles	PQ/QN	2 or more varieties compared by one characteristic Expressions allocated to two or more categories Value of each category is more than five
Match methods	Fisher's exact test	VS	-	-	Hypothesis for D based on previously known facts or principles*	PQ/QN	2 varieties compare by one characteristic Expressions allocated to two categories Value of each category is less than 10

***** Under certain circumstances *MG and VG may also be applicable

5.4 Requirements for statistical methods for distinctness assessment



There may be other statistical methods suitable for the assessment of distinctness which are not included in the above diagram.

6. CYCLIC PLANTING OF VARIETIES FROM THE VARIETY COLLECTION TO REDUCE TRIAL SIZE

6.1 Summary of requirements for application of method

Cyclic planting of varieties from the variety collection (established varieties) to reduce trial size is appropriate for use in trials where:

- distinctness is determined by COYD;
- the number of established varieties is excessive for cost or for practical reasons;
- there should be at least 20 degrees of freedom for the MJRA-adjusted varieties-by-years mean square in the adapted COYD analysis of variance. If there are not, then cyclic planting of established varieties should not be used.
- three independent growing cycles are normally grown. The guidance below is for this case. However, it may also be adapted for crops where two independent growing cycles are normally grown.

6.2 Summary

Cyclic planting of the established varieties in trial and analysis by compensated data is a system to reduce DUS trial sizes while maintaining testing stringency. It may be used in trials where distinctness is determined by COYD.

The system comprises allocating each of the established varieties in trial to one of three series~~,~~ with one series omitted in turn from trial each year[5]. Candidate varieties are included in trial for the three years of their test period plus a fourth year. If, after DUS testing, a variety is added to the variety collection it is allocated to a series and is cyclically omitted from the trial every third year.

Distinctness is assessed by applying an adaptation of COYD to the incomplete table of variety characteristic means (candidate and established varieties) in the three year test period. Where data is missing for a variety, it is compensated for by use of two years' data from before the test period. If uniformity is determined by COYU, it may be applied to the incomplete table of variety characteristic standard deviations (candidate and established varieties) in the three year test period. Prior to its adoption, historical data should be used to compare the DUS decisions made based on the cyclic planting system with those based on the existing system.

6.3 Cyclic Planting of Established Varieties in Trial

Established varieties in trial are allocated to one of three series One series is omitted cyclically from trial each year (Fig. 1). Thus varieties belonging to Series 1 in Fig. 1 will not be planted in 2010, 2013 or 2016, whereas those in Series 3 will not be planted in 2012, 2015 or 2018. This will result in a smaller trial size as one third of the established varieties are omitted from the trial each year. Each candidate variety is planted in trial and has data recorded on it in each year of a three year test period (2014 to 2016 in Fig. 1 below), after which a DUS decision is taken. Because of a possible lag between final DUS testing and the decision on the application, candidate varieties are kept in trial for a fourth year after the three-year test period. If a positive decision is taken, they will become an established variety and will enter the cyclic planting system. Thus all newly accepted varieties are initially present in trial for four consecutive years, and all varieties entering trial in the same year follow the same cycle of omissions in future years. Hence candidate varieties that had their final year of DUS testing in 2012 in Fig. 1 are in trial for a fourth year in 2013 and so join the Series 2 established varieties. Candidate varieties final DUS tested in 2013, 2014 and 2015, would join Series 3, 1 and 2 respectively.

Established varieties are initially allocated to series in a manner to minimize the risk of bias. Other than the initial allocation, the choice of established varieties following each series is determined by the candidate varieties entered for trial in earlier years and by which established varieties the applicants choose to withdraw.

---

[5] For the purpose of this document, “year” means a “growing cycle”.

Although an exactly equal number of established varieties belonging to each series is not essential, it is likely to be beneficial to balance the numbers in each series in the future. This should be done by transferring established varieties between the series by planting them in years when they should be omitted.

Figure 1. **Data patterns and usage for the test period 2014 to 2016**

					TEST PERIOD				
TRIAL YEARS	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**
Candidate Varieties					X	X	X	*	
Established Varieties									
Series 1		X	X		X	X		*	*
Series 2	O		X	X		X	X		*
Series 3	O	X		X	X		X	*	
New Established Varieties – Assimilation into matrix									
Final DUS tested in 2012 (Series 2)	O	O	X^F	X		X	X		*
Final DUS tested in 2013 (Series 3)		O	X	X^F	X		X	*	
Final DUS tested in 2014 (Series 1)			X	X	X^F	X		*	*
Final DUS tested in 2015 (Series 2)				O	X	X^F	X		*

X Indicates data retrieved using maximum of 4 years for distinctness testing and within the (boxed) test period for uniformity testing

O Indicates data present but not retrieved

F Indicates final DUS test year of new established varieties

* Indicates future inclusion in trial

☐ (within box) Indicates the data used for uniformity testing

6.3.1 The assessment of distinctness by data compensation

Conventionally, when using COYD to assess distinctness, it is applied to a complete variety (candidate and established) by test period years matrix of characteristic means. With cyclic planting, this matrix is incomplete for the established varieties. For the assessment of distinctness, where data on an established variety is missing, data held in computer files from earlier years are used to compensate for the loss of data. Due to lack of overlap years with the candidates, the value of back-data is not as high as data from the test period. In the crops to which cyclic planting has been applied to date, to maintain stringency of testing, two years of past data must be included when one year of current data is missing for an established variety. Thus for the 2014 to 2016 test period illustrated in Fig. 1, established varieties in Series 1 would have data from 2011 and 2012 retrieved, those in Series 2 data from 2012 and 2013 and those in Series 3 data from 2011 and 2013. Even where more years of past data are available (marked by an O in Fig. 1), to avoid reducing the stringency of the distinctness test, only the two most recent years are used to compensate for the missing current year. Hence, while data from 2010 and before are available for varieties in Series 2 and 3, such data are not retrieved for the 2014 to 2016 test period.

Sometimes data on an established variety will be available for a year when its series suggests it would not be present in the trial. Such cases are the fourth year after the three year test period where a candidate variety has become an established variety in trial, or where an established variety is needed for a special test with a problem variety. In this case the established variety would have full data available during the test period and so no historical data would be retrieved for the distinctness testing. Thus for the test period of 2014 to 2016, successful candidate varieties final DUS tested in 2015 would have no historical data retrieved, whereas successful candidate varieties final DUS tested in 2012, 2013 and 2014 would have historical data retrieved.

6.3.2 Method of analysis for distinctness assessment

Distinctness is assessed by applying an adaptation of COYD with Modified Joint Regression Analysis (MJRA) applied to data comprising the incomplete table of variety (candidate and established) characteristic means in the three year test period together with the compensating back-data for established varieties missing during the test period. Details of the method of analysis and an example are given in section 1.7.

6.3.3 The assessment of uniformity

Conventionally, when using COYU to assess uniformity, it is applied to a complete variety (candidate and established) by test period years matrix of within variety standard deviations. With cyclic planting, as may be seen from the boxed year by variety combinations in Fig. 1, this matrix is incomplete for the established varieties. COYU is applied to this matrix and no attempt is made to compensate for the incomplete data. This is because COYU consists of pooling over years the within variety standard deviations for all available established varieties while taking into account any relationship between variety means and the standard deviations. This is done to provide a uniformity standard against which to compare the standard deviations of the candidate varieties. Consequently, it is not possible to make a correction for standard deviations from years outside the test period. As a result, only uniformity data from the established varieties within the test period are used to set the uniformity standard for the candidates.

6.4 Comparison of the cyclic planting system with the existing system

Prior to adoption of the system of cyclic planting, historical data should be used to compare the DUS decisions made based on the cyclic planting system with those based on the existing system. Providing all established varieties were planted with the existing system, the cyclic planting system can be simulated by allocating established varieties to the series, replacing their data with missing data symbols in the computer files where appropriate, and including the previous years' files from which data are to be retrieved to compensate for this 'missing' data. The distinctness and uniformity decisions that would have been made based on the cyclic planting system can then be compared with those that would have been made based on the existing system. This approach also permits assessment of the number of years of back-data that should be included to compensate for when one year of data in the test period is missing for an established variety.

Note: if the DUSTNT software is used, a variety can be made to appear missing simply by removal of the variety from the "E file". In United Kingdom DUS Herbage trials, when compared with the previous system, the cyclic planting system was found to be slightly less stringent in distinctness testing and slightly more stringent in uniformity testing, with a minimal overall effect on the DUS variety pass rate.

6.5 Cyclic planting system software

The DUST program CYCL, has been developed to enable the compensated data to be retrieved, statistically analyzed using MJRA, and the results presented in reports suitable for the assessment of distinctness. Uniformity assessment is based on the data within the test period and uses the DUST program COYU. Both programs are available as part of the DUST9 (MSDOS based) and DUSTNT (Windows NT and 95) versions of the DUST software.

6.6 Additional technical detail and example of analysis for distinctness assessment

Distinctness is assessed by applying an adaptation of COYD to n data values comprising the incomplete table of variety (candidate and established) characteristic means in the three year test period together with the compensating back-data for established varieties missing during the test period. Characteristics are all analyzed by Modified Joint Regression Analysis (MJRA). This scales all the variety effects in a year up or down depending on the year by multiplying the variety effects by a sensitivity for the year.

The MJRA model for the cyclic planting data with n_v varieties in n_y years is as follows:

$$c_{ij} = \mu + y_j + \beta_j v_i + \varepsilon_{ij}$$

where c_{ij} is the value on a characteristic for variety i in year j, $i = 1,\ldots,n_v$ and $j = 1,\ldots, n_y$
μ is the overall mean
v_i is the effect of the ith variety with $\Sigma v_i = 0$
y_j is the effect of the jth year with $\Sigma y_j = 0$
β_j is the sensitivity of year j.
ε_{ij} is a random error associated with variety i in year j

This model is an adaptation of one proposed by Digby (1979) where year effects are scaled for a variety by multiplying them by a variety sensitivity. As the model is non-linear, it cannot be fitted directly to the data, but

must be fitted iteratively to obtain estimates of the variety means and least significant differences (LSD's), which are based on the MJRA-adjusted varieties-by-years mean square and are used to compare the variety means and determine distinctness. The LSD's and the MJRA-adjusted varieties-by-years mean square are on (n - 1 - 2(n_y - 1) - (n_v - 1)) degrees of freedom, which should be at least 20 degrees of freedom.

6.6.1 Example of distinctness assessment

Consider the following matrix of n within year variety means c_{ij}. Variety A represents candidate varieties and varieties B, C and D represent the three series of established varieties. The test period is years 4 to 6.

Example data

	Year					
Variety	1	2	3	4	5	6
A	-	-	-	6	2	3
B	-	6	4	-	6	7
C	7	10	-	8	11	-
D	11	-	14	10	-	17

Model fitting provides final estimates of $\hat{\mu},(\hat{y}_1,...,\hat{y}_6),(\hat{\beta}_1,...,\hat{\beta}_6),(\hat{v}_1,...,\hat{v}_4)$ as 7.862, (-2.12, 0.55, -1.20, -0.12, 1.16, 1.73), (0.91, 1.14, 1.26, 0.36, 1.39, 1.28), (-5.09, -2.12, 1.38, 5.81), from which the following table of means is derived:

	Year						
Variety	1	2	3	4	5	6	Means
A	-	-	-	6	2	3	2.78 = 7.86 + -5.09
B	-	6	4	-	6	7	5.76
C	7	10	-	8	11	-	9.24
D	11	-	14	10	-	17	13.67
Means	5.74	8.42	6.66	7.75	8.92	9.03	
Sensitivities	0.91	1.14	1.26	0.36	1.37	1.39	

The model fitting also provides standard errors for the means on 1 degree of freedom, which together with the two-tailed 1% critical t-value on 1 degree of freedom, gives the following table of 1% LSD values between all variety pairs:

Variety	A	B	C
B	15.75		
C	18.00	15.64	
D	18.39	15.64	18.83

Comparison of the 1% LSD between varieties A and D (18.39) with the difference in their means of 10.89 indicates these varieties are not significantly different at the 1% level. Further details of the analysis and the worked example are given in Camlin *et al* (2001).

Note: the above example serves to illustrate the method, but is on an artificially small dataset. It results in LSD's and the MJRA-adjusted varieties-by-years mean square on 1 degree of freedom. The recommended minimum for use of the method in practice is 20 degrees of freedom.

6.7 References

Camlin, M.S., Watson, S., Waters, B.G. and Weatherup, S.T.C. (2001). The potential for management of reference collections in herbage variety registration trials using a cyclic planting system for reference varieties. Plant Varieties and Seeds, 14:1-14.

Digby, P. (1979) Modified joint regression for incomplete variety x environment data. Journal of Agricultural Science 93, Cambridge, 81-86.

[Part II follows]

PART II: SELECTED TECHNIQUES USED IN DUS EXAMINATION

1. THE GAIA METHODOLOGY

The GAIA method has been developed to optimize trials, by avoiding the growing of some of the varieties in the variety collection. The principle is to compute a phenotypic distance between each pair of varieties, this distance being a sum of distances on each individual observed characteristic. The background of the method relies on the possibility given to the crop expert to express his confidence on the differences observed, by giving weights to the difference for each observed characteristic.

The GAIA methodology is mainly used after a first growing cycle to identify those varieties in the variety collection which can be excluded from the subsequent growing cycle(s) because they are “Distinct Plus” (see Part II section 1.3.2.1) from all the candidate varieties. GAIA can also identify similar varieties, on which the DUS examiner will need to focus attention in the subsequent growing cycle

1.1 Some reasons to sum and weight observed differences

1.1.1 When assessing distinctness, a DUS examiner first observes a variety characteristic-by-characteristic. In the case of similar varieties, the DUS examiner also considers all observed differences as a whole. The GAIA software helps the DUS examiner to assess differences characteristic-by-characteristic and for all characteristics together.

1.1.2 A DUS examiner may see that two varieties are so distinct after the first growing cycle that it is not necessary to repeat the comparison. Those two varieties, which are “distinct plus” (see Part II section 1.3.2.1), are obviously distinct.

1.1.3 A DUS examiner may have a situation where two varieties receive a different notes, but the two varieties are considered by the examiner to be similar. The difference could be due to the fact that the varieties were not grown very close each other (i.e. had different environmental conditions), or to variability of the observer when assessing the notes, etc.

1.1.4 Characteristics vary in their susceptibility to environmental conditions and the precision with which they are observed (i.e. visual observation/measurement). For characteristics which are susceptible to environmental conditions and which are not assessed very precisely, the examiner requires a large difference between Variety A and Variety B to be confident that the observed difference indicates distinctness.

1.1.5 For characteristics which are independent of environmental conditions and which are assessed precisely, the examiner can be confident in a smaller difference between Variety A and Variety B.

1.1.6 In the GAIA method, the examiner decides the appropriate weights for the observed differences for each observed characteristic. The software computes the sum of the weightings and indicates to the crop examiner which pairs of varieties are “distinct plus” and which are not. The examiner can then decide which of the varieties of in the variety collection can be excluded from the subsequent growing cycle(s), because they are already obviously distinct from all candidate varieties.

1.2 Computing GAIA phenotypic distance

1.2.1 The principle of the GAIA method is to compute a phenotypic distance between two varieties, being the total distance between a pair of varieties resulting from the addition of the weightings of all characteristics. Thus, the GAIA phenotypic distance is:

$$dist(i, j) = \sum_{k=1,nchar} W_k(i, j)$$

where:

$dist(i, j)$ is the computed distance between variety i and variety j.

k is the kth characteristic, from the *nchar* characteristics selected for computation.

$W_k(i, j)$ is the weighting of characteristics k, which is a function of the difference observed between variety *i* and variety *j* for that characteristic k.

$W_k(i, j) = f\left(\left|OV_{ki} - OV_{kj}\right|\right)$

where OV_{ki} is the observed value on characteristic *k* for variety *i*.

1.2.2 Detailed information on e is provided in section 1.3.

1.3 Detailed information on the GAIA methodology

1.3.1 Weighting of characteristics

1.3.1.1 It is important to take account of the correlation between characteristics when weighting. If two characteristics are linked (e.g. plant height including panicle; plant height excluding panicle), it is advisable to use only one of them in GAIA, to avoid double weight. For example, assuming that panicle length is used as a characteristic, it would be advisable to use only plant height including panicle, or plant height excluding panicle.

1.3.1.2 Weighting is defined as the contribution in a given characteristic to the total distance between a pair of varieties. For each species, this system must be calibrated to determine the weight which can be given to each difference and to evaluate the reliability of each characteristic in a given environment and for the genetic variability concerned. For that reason the role of the crop expert is essential.

1.3.1.3 Weighting depends on the size of the difference and on the individual characteristic. The weightings are defined by the crop expert on the basis of his expertise in the crop and on a “try-and-check” (see Diagram 3 at the end of this annex) learning process. The expert can give zero weighting to small differences, thus, even if two varieties have different observed values in many characteristics, the overall distance might be zero. For a given difference, the same weighting is attributed to any pair of varieties for a given characteristic.

1.3.1.4 The weighting should be simple and consistent. For instance the crop expert can base the weights for a characteristic only with integer values, i.e. 0, 1, 2, 3, (or more).

If so,

- a weight of 0 is given to observed differences which for this characteristic are considered by the crop expert as possibly caused by environment effects or lack of precision in measure.
- a weight of 1 is the minimum weight which can contribute as a non zero distance
- a weight of 3 is considered to be about 3 times greater in term of confidence or distance than a weight of 1.

1.3.1.5 The distinctness plus threshold will be defined as a value for which the sum of the differences with a non zero weight is great enough to ensure a reliable obvious distinction.

1.3.1.6 Diagram 3 is a flowchart which describes how an iterative “try and learn” process can be used to obtain step by step a satisfactory set of weights for a given crop.

1.3.1.7 The following simple example on *Zea mays* shows the computation of the distance between two varieties:

Example: taking the characteristic “Weighting matrix shape of ear”, observed on a 1 to 3 scale, the crop expert has attributed weighting to differences which they consider significant:

Shape of ear:
1 = conical
2 = conico-cylindrical
3 = cylindrical

Comparison between difference in notes and weighting		
	Different in notes	Weighting
conical (1) vs. conical (1)	0	0
conical (1) vs. conico-cylindrical (2)	1	2
conical (1) vs. cylindrical (3)	2	6
conico-cylindrical (2) vs. conico-cylindrical (2)	0	0
conico-cylindrical (2) vs. cylindrical (3)	1	2
cylindrical (3) vs. cylindrical (3)	0	0

When the crop expert compares a variety 'i' with conical ear (note 1) to a variety 'j' with cylindrical ear (note 3), he attributes a weighting of 6 etc. The weightings are summarized in the form of a weighting matrix:

Weighting matrix 'i'				
		Variety 'i'		
Variety 'j'		**1**	**2**	**3**
	1	0	2	6
	2		0	2
	3			0

When the crop expert compares a variety i with conical ear (note 1) to a variety j with cylindrical ear (note 3), he attributes a weighting of 6.

1.3.2 Examples of use

1.3.2.1 Determining "Distinctness Plus"

1.3.2.1.1 The threshold for the phenotypic distance used to eliminate varieties from the growing trial is called "Distinctness Plus" and is settled by the crop expert at a level which is higher than the difference needed to establish distinctness. This ensures that all pairs of varieties having a distance equal or greater than the threshold (Distinctness Plus) would be distinct if they were grown in another trial.

1.3.2.1.2 The Distinctness Plus threshold must be based on experience gained with the varieties in the variety collection and must minimize the risk of excluding in a next growing trial a pair of varieties which should need to be further compared in the field.

1.3.2.2 Other examples of use

Using phenotypic distance in the first growing cycle

1.3.2.2.1 A crop that has a large variety collection and uses only characteristics on a 1 to 9 scale; GAIA methodology allows the selection of varieties to be included in the growing trial. This can be used to plan the first growing cycle trials as well as the subsequent growing cycles.

1.3.2.2.2 In crops with relatively few candidates and a small variety collection, which enables the crop expert to sow all candidates (e.g. an agricultural crop), and the appropriate varieties in the variety collection, in two or three successive growing cycles. The same varieties are sown in growing cycles 1, 2 and 3, in a randomized layout. The software will help to identify the pairs with a small distance, to enable the expert to focus his attention on these particular cases when visiting the field.

Using phenotypic distance after the first growing trial

1.3.2.2.3 After one growing cycle (e.g. in the examination of an ornamental crop), the absolute data and distance computations are an objective way to secure the decision of the expert, because the quality of the observation and reliability of differences observed have been taken into account in the weighting system. If more growing cycles are necessary before a decision is taken, the software helps to identify on which cases the expert will need to focus.

1.3.2.2.4 In cases where there are many candidate varieties and many varieties in the variety collection and there is a wide variability in the species (e.g. a vegetable crop such as *Capsicum*); on the one hand there are already obvious differences after only one cycle, but on the other hand some varieties are very similar. In order to be more efficient in his checks, the crop expert wishes to grow “similar” varieties close to each other. The raw results and distances will help to select the “similar” varieties and decide on the layout of the trial for the next growing cycle.

1.3.2.2.5 In crops in which there are many similar varieties, for which it is a common practice to make side-by-side comparisons, GAIA can be used to identify the similar varieties after the first cycle, in particular, when the number of varieties in a trial increases, making it less easy to identify all the problem situations. The software can help to “not miss” the less obvious cases.

1.3.2.2.6 In vegetatively propagated ornamental varieties, the examination lasts for one or two growing cycles: after the first growing cycle, some varieties of the variety collection in the trial are obviously different from all candidates, and their inclusion in the second growing cycle is not necessary. When the number of varieties is large, the raw data and distance(s) can help the expert to detect varieties in the variety collection for which the second growing cycle is unnecessary.

1.3.3 Computing GAIA phenotypic distance

The principle is to compute a phenotypic distance between two varieties, which is the sum of weightings given by the crop expert to the differences he observed.

GAIA phenotypic distance is:

$$dist(i, j) = \sum_{k=1,nchar} W_k(i, j)$$

where:

$dist(i, j)$ is the computed distance between variety *i* and variety *j*.

k is the kth characteristic, from the *nchar* characteristics selected for computation.

$W_k(i, j)$ is the weighting of characteristics k, which is a function of the difference observed between variety *i* and variety *j* for that characteristic k.

$$W_k(i, j) = f\left(\left|OV_{ki} - OV_{kj}\right|\right)$$

where OV_{ki} is the observed value on characteristic *k* for variety *i*.

This phenotypic distance computation allows to:

- compare two varieties,
- compare a given variety to all other varieties,
- compare all candidate varieties to all [candidate + varieties in the variety collection included in the growing trial] observed varieties,
- compare all possible pair combinations.

1.3.4 GAIA software

1.3.4.1 GAIA software allows the computation of the phenotypic distance using UPOV characteristics of the test guidelines, which can be used alone or in combination. The user can decide on the type of data and the way it is used. He can select all the available characteristics, or different subsets of characteristics.

1.3.4.2 The main use of GAIA is to define a “distinct plus” threshold which corresponds to a reliable and obvious distinction.

1.3.4.3 Remember that all differences with a zero weight do not contribute at all to the distance. Two varieties can have different notes in a number of observed characteristics, and end with a zero distance.

1.3.4.4 Non zero weights are summed in the distance. If the distance is smaller than the distinct plus threshold, even if there are a number of clear differences in notes or measures, the varieties will not be suggested as reliably and obviously distinct. If the distance is greater than the distinct plus threshold set by the crop expert, this shall correspond to a case where a pair comparison in a further growing trial is unnecessary.

1.3.4.5 GAIA enables the crop expert to use the threshold parameter in two other ways for practical means other than distinctness plus:

- a low threshold helps to find the more difficult cases (to identify similar varieties or close varieties) on which the expert will have to focus his attention in the next cycle
- a very big threshold allows all available raw data and the weightings for each characteristic to be seen on screens and printouts

1.3.4.6 In practice different thresholds can be used according to different needs. They can easily be selected before running a comparison. Different comparisons can be computed, stored and recalled from the database with their appropriate threshold, set of characteristics, set of varieties, etc.

1.3.4.7 The software provides a comprehensive report for each pair-wise comparison and a classification of all pair-wise comparisons, from the more distinct to the more similar. The software computes an overall distance, but also provides all the individual absolute values and the distance contribution of each characteristic.

1.3.4.8 In order to minimize computation time, as soon as the threshold is achieved for a comparison between two given varieties, the software proceeds to the next pair of varieties. Remaining characteristics and their raw values will not be shown in the summary output, and will not contribute to the distance.

1.3.4.9 Section 1.3.5 provides a screen copy of a display tree which shows how the expert can navigate and visualise the results of computations.

1.3.4.10 GAIA software has been developed with WINDEV. The general information (species, characteristics, weighting, etc.), the data collected on the varieties and the results of computations are stored in an integrated database. Import and export facilities allow for other information systems to be used in connection with the GAIA software. ODBC allows access to the GAIA database and to other databases simultaneously.

1.3.4.11 1 or 2 notes per variety can be used. 1 note occurs when one cycle is available. Two notes are present for instance when two trials are made in different locations in a given year, or if 2 cycles are obtained in the same location. For electrophoresis data, only one description can be entered per variety. For measurements, at least 2 values (different trials, repeats, etc.) are necessary and the user can select which to use in the computation.

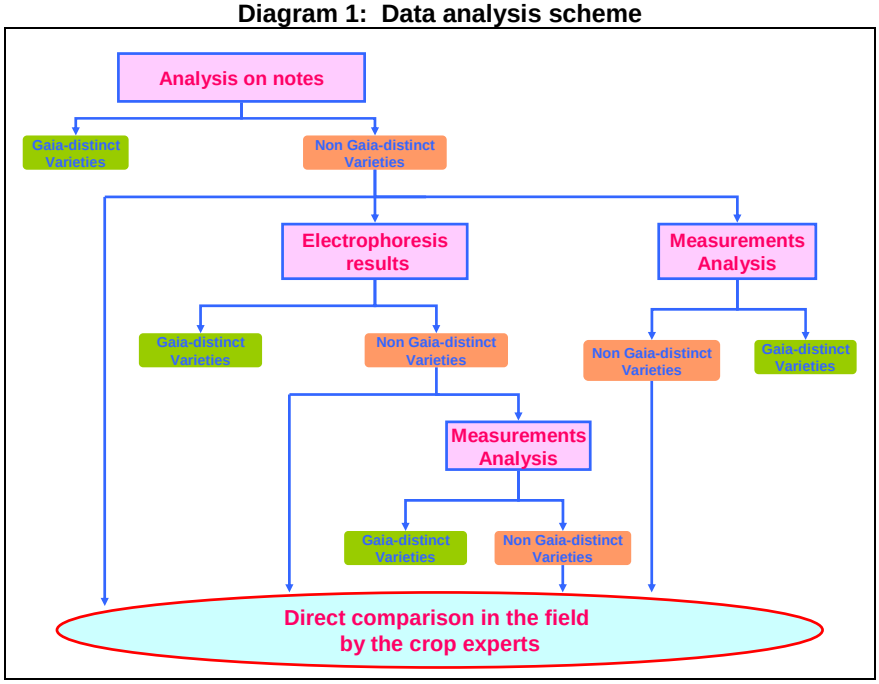
1.3.4.12 GAIA is most suitable for self-pollinated and vegetatively propagated varieties, but can also be used for other types of varieties.

1.3.5 Example with Zea mays data

1.3.5.1 Introduction

The software can use notes, measurements and/or electrophoresis results. These types of data can be used alone or in combination, as shown in Diagram 1.

Diagram 1: Data analysis scheme



In this example, it is assumed that the crop expert has decided to use a Distinctness Plus threshold $\mathbf{S_{dist}}$ of 10.

1.3.5.2 Analysis of notes

1.3.5.2.1 In qualitative analysis notes (1 to 9) are used. Notes can come from qualitative, quantitative and pseudo-quantitative characteristics.

1.3.5.2.2 For each characteristic, weightings according to differences between levels of expression are pre-defined in a matrix of distances.

1.3.5.2.3 "Shape of ear": observed on a 1 to 3 scale, the crop expert has attributed weightings greater than zero to differences which they consider significant:

1 = conical
2 = conico-cylindrical
3 = cylindrical

		Variety 'i'		
		1	2	3
Variety 'j'	1	0	2	6
	2		0	2
	3			0

1.3.5.2.4 When the crop expert compares a variety 'i' with conical ear (note 1) to a variety 'j' with cylindrical ear (note 3), they attribute a weighting of 6.

1.3.5.2.5 “Length of husks”, observed on a 1 to 9 scale, the crop expert has defined the following weighting matrix:

1 = very short
2 = very short to short
3 = short
4 = short to medium
5 = medium
6 = medium to long
7 = long
8 = long to very long
9 = very long

		Variety ‘i’								
		1	2	3	4	5	6	7	8	9
Variety ‘j’	1	0	0	0	2	2	2	2	2	2
	2		0	0	0	2	2	2	2	2
	3			0	0	0	2	2	2	2
	4				0	0	0	2	2	2
	5					0	0	0	2	2
	6						0	0	0	2
	7							0	0	0
	8								0	0
	9									0

1.3.5.2.6 The weighting between a variety ‘i’ with very short husks (note 1) and a variety ‘j’ with short husks (note 3) is 0. The expert considers a difference of 3 notes is the minimum difference in order to recognise a non-zero distance between two varieties. Even if the difference in notes is greater than 3, the expert keeps the distance weight to 2 while in very reliable characteristics a difference of 1 is given a weight of 6.

1.3.5.2.7 The reason for using a lower weighting for some characteristics compared to others can be that they are less “reliable” or “consistent” (e.g. more subject to the effect of the environment); and/or they are considered to indicate a lower distance between varieties.

1.3.5.2.8 The matrix for a qualitative analysis for 5 characteristics for varieties A and B:

	Ear shape	Husk length	Type of grain	Number of rows of grain	Ear diameter	
Notes for variety A (1 to 9 scale)	1	1	4	6	5	
Notes for variety B (1 to 9 scale)	3	3	4	4	6	
Difference observed	2	2	0	2	1	
Weighting according to the crop expert	*6*	*0*	*0*	*2*	*0*	$D_{qual} = 8$

In this example $\mathbf{D_{qual} = 8 < 10}$ **(**$\mathbf{S_{dist}}$ =10 in this example) varieties A and B are declared “GAIA NON-distinct” on the basis of these 5 characteristics.

1.3.5.3 Electrophoresis analysis

1.3.5.3.1 In some UPOV Test Guidelines electrophoresis results can be used, as in *Zea mays*. The software does not allow the use of heterozygous loci, but only the use of homozygous loci, in conformity with the Test Guidelines. Results used are 0 (absent) and 1 (present), and the knowledge of chromosome number.

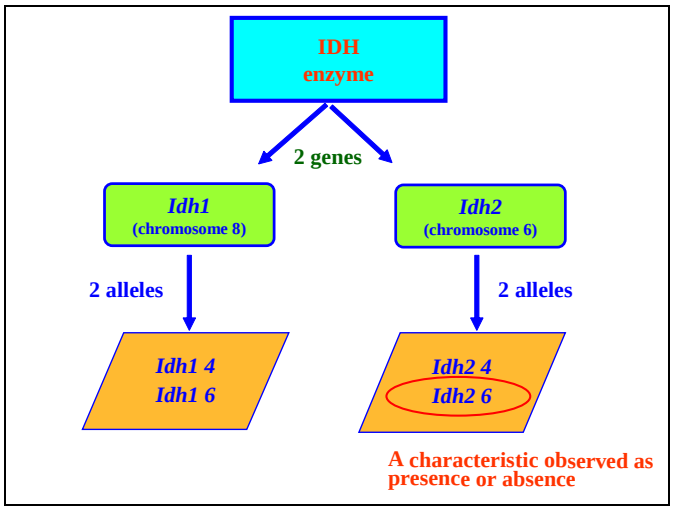


Diagram 2: The Isocitrate Dehydrogenase (IDH) enzyme has two genes (*Idh1* and *Idh2*) located on two different chromosomes. Each of them has two alleles which are observed as 1 (presence) or 0 (absence).

1.3.5.3.2 Electrophoresis results are noted as 0 or 1 (absence or presence). The decision rule, used to give a weighting to two varieties, is the addition of the weighting number of differences observed and the weighting number of chromosomes related to these differences (see example below):

	Chromosome 8		Chromosome 6	
	Idh1 4	*Idh1 6*	*Idh2 4*	*Idh2 6*
Variety A	0	1	1	0
Variety B	0	1	0	1
Difference	0	0	1	1

1.3.5.3.3 In this example, varieties A and B are described for 4 electrophoresis results:

Idh1 4, *Idh1 6*, *Idh2 4* and *Idh2 6*. The software looks at differences and gives the phenotypic distance using the following computation:

$$D_{elec} = 2 \times 0{,}25 + 1 \times 1 = 1{,}5$$

2 is the number of differences observed

0,25 is the weighting attributed by experts to the number of differences

1 is the number of chromosomes on which differences are observed

1 is the weighting associated by experts to chromosome

1.3.5.3.4 This formula, which might be difficult to understand, was established by the crop expert in collaboration with biochemical experts. Both the *number of differences* and the *number of chromosomes on which differences are observed* are used. Thus, less importance is attached to differences when these occur on the same chromosome, than when they occur on different chromosomes.

1.3.5.3.5 After qualitative and electrophoretic analysis, the phenotypic distance between varieties A and B is equal to:

$$D = D_{qual} + D_{elec} = 8 + 1.5 = 9.5$$

1.3.5.3.6 The phenotypic distance is *lower than* $\mathbf{S_{dist}}$ ($\mathbf{S_{dist}}$=10 in this example) *therefore varieties A and B are considered "GAIA NON-distinct".*

1.3.5.3.7 The crop expert can decide if he does not want to establish distinctness solely on the basis of electrophoresis analysis. It is necessary to have a minimal phenotypic distance in qualitative analysis in order to take into account the electrophoresis results. This minimal phenotypic distance must also be defined by the crop expert.

1.3.5.4 Analysis of measurements

1.3.5.4.1 Analysis of measurements computes differences on observed or computed measurements, counts are handled as measurements

1.3.5.4.2 For each measured characteristic, the comparison of two varieties is made by looking for consistent differences in at least two different experimental units. Experimental units are defined by the user depending on data present in the database. It can, for example, be the data from two geographical locations of the first growing cycle, or 2 or 3 replications from the same trial in the case of a single geographical location, or data from 2 cycles in the same location.

1.3.5.4.3 For a comparison to be made, the two varieties must be present in the same experimental units. The differences observed must be greater than one of the two threshold values (or minimal distances), fixed by the crop expert.

- $\mathbf{D_{min\text{-}inf}}$ is the lower value from which a weighting is attributed,
- $\mathbf{D_{min\text{-}sup}}$ is the higher minimal distance. These values could be chosen arbitrarily or calculated (15% and 20% of the mean for the trial, or LSD at 1% and 5%, etc.)

For each minimal distance a weighting is attributed:

- $\mathbf{D_{min\text{-}inf}}$ a weighting P_{min} is attributed;
- $\mathbf{D_{min\text{-}sup}}$ a weighting P_{max} is attributed;
- the observed difference is lower than $\mathbf{D_{min\text{-}inf}}$ a zero weighting is associated.

1.3.5.4.4 Varieties A and B have been measured for characteristics "Width of blade" and "Length of plant" in two trials.

For each trial, and each characteristic, the crop expert has decided to define $\mathbf{D_{min\text{-}inf}}$ and $\mathbf{D_{min\text{-}sup}}$ by calculating respectively the 15% and 20% of the mean for the trial:

	Width of blade		Length of plant	
	Trial 1	Trial 2	Trial 1	Trial 2
$\mathbf{D_{min\text{-}inf}}$ = 15% of the trial mean	1.2 cm	1.4 cm	28 cm	24 cm
$\mathbf{D_{min\text{-}sup}}$ = 20% of the trial mean	1.6 cm	1.9 cm	37 cm	32 cm

For each characteristic, the crop expert has attributed the following weighting:

A weighting $\mathbf{P_{min}}$ = 3 is attributed when the difference is greater than $\mathbf{D_{min\text{-}inf}}$.

A weighting $\mathbf{P_{max}}$ = 6 is attributed when the difference is greater than $\mathbf{D_{min\text{-}sup}}$.

	Width of blade		Length of plant		
	Trial 1	Trial 2	Trial 1	Trial 2	
Variety A	9.9 cm	9.8 cm	176 cm	190 cm	
Variety B	9.6 cm	8.7cm	140 cm	152 cm	
Difference	0.3 cm	1.1 cm	36 cm	38 cm	
Weighting according to the crop expert	0	0	3	6	D_{quan} =?

1.3.5.4.5 In this example, for the characteristic "Width of blade", the differences observed are lower than $\mathbf{D_{min\text{-}inf}}$, so no weighting is associated. On the other hand, for the characteristic "Length of plant" one difference is greater than the $\mathbf{D_{min\text{-}inf}}$ value and the other is greater than the $\mathbf{D_{min\text{-}sup}}$ value. These two differences are attributed different weightings.

1.3.5.4.6 The user must decide which weighting will be used for the analysis:

- the weighting chosen is that attributed to the lowest difference (minimalist option);
- the weighting chosen is that attributed to the highest difference (maximalist option);
- mean option: the weighting chosen is the mean of the others (mean option).

1.3.5.4.7 In this example, the crop expert has decided to choose the lowest of the two weightings, so the phenotypic distance based on measurements is $\mathbf{D_{quan} = 3}$.

1.3.5.4.8 In summary, at the end of all analysis, the phenotypic distance between varieties A and B is:

$$\mathbf{D = D_{qual} + D_{elec} + D_{quan} = 8 + 1.5 + 3 = 12.5 > S_{dist}}$$

1.3.5.4.9 The phenotypic distance is greater than the distinction threshold $\mathbf{S_{dist}}$, fixed by the crop expert at 10, so varieties A and B are declared "GAIA-distinct".

1.3.5.4.10 In this example, the use of electrophoresis data "confirms" a distance between the two varieties; but on the basis of qualitative and quantitative data alone, the threshold is exceeded (8 + 3 = 11 is greater than 10).

1.3.5.4.11 If the threshold had been set at 6, the difference on the characteristic ear shape would have been sufficient, as variety A is conical and variety B is cylindrical, which is already a clear difference.

1 = conical
2 = conico-cylindrical
3 = cylindrical

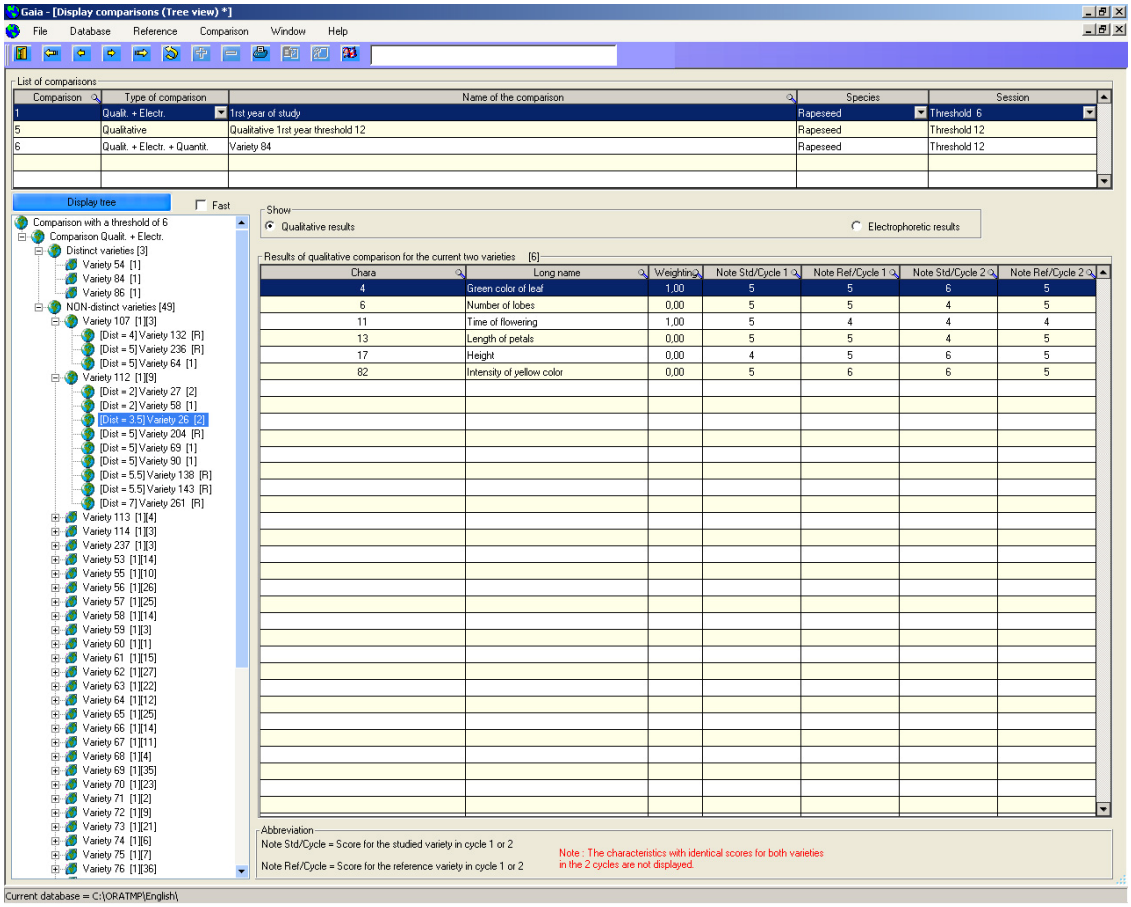
Variety i			
	1	2	3
1	0	2	6
2		0	2
3			0

1.3.5.5 Measurements and 1 to 9 scale on the same characteristic

1.3.5.5.1 For some crops, it is common practice to produce values on a 1 to 9 scale from measurements. Sometimes the transformation process is very simple, sometimes it is complex.

1.3.5.5.2 GAIA can include both as two separate characteristics: the original measurements and the 1 to 9 scale. They are associated in the description of the characteristics. Using the knowledge of this association, when both are present, only one of them is kept, in order to avoid the information being used twice for weighting.

1.3.6 Example of GAIA screen copy



1.3.6.1 The upper part "List of comparisons" shows 3 different computations which have been kept in the database. Comparison 1 is highlighted (selected) and shown on the display tree.

1.3.6.2 The "Display tree" on the left shows results for a [qualitative + electrophoresis at threshold of 6] computation.

1.3.6.3 *Distinct varieties [3]* indicates that 3 varieties were found distinct from all others. There was a total of 52 (49 + 3) varieties in the computation.

1.3.6.4 The display tree is used to navigate through all possible pairs.

1.3.6.5 The user can expand or reduce the branches of the tree according to his needs.

1.3.6.6 *NON-distinct varieties [49].* Forty-nine varieties were found "not distinct from all others" with a threshold of 6.

1.3.6.7 The first variety, *Variety 107*, has only 3 close varieties, whereas the second, *Variety 112*, has 9 close varieties, the third, *Variety 113*, 4 close varieties, etc.

1.3.6.8 *Variety 112 [1][9]* indicates variety 112 is in the first year of examination *[1]*; and has 9 close varieties according to the threshold of 6 *[9].*

1.3.6.9 *[dist=3.5] Variety 26 [2]* indicates variety 26 (comparison highlighted=selected) has a GAIA distance of 3.5 from variety 112, which is in second year of examination.

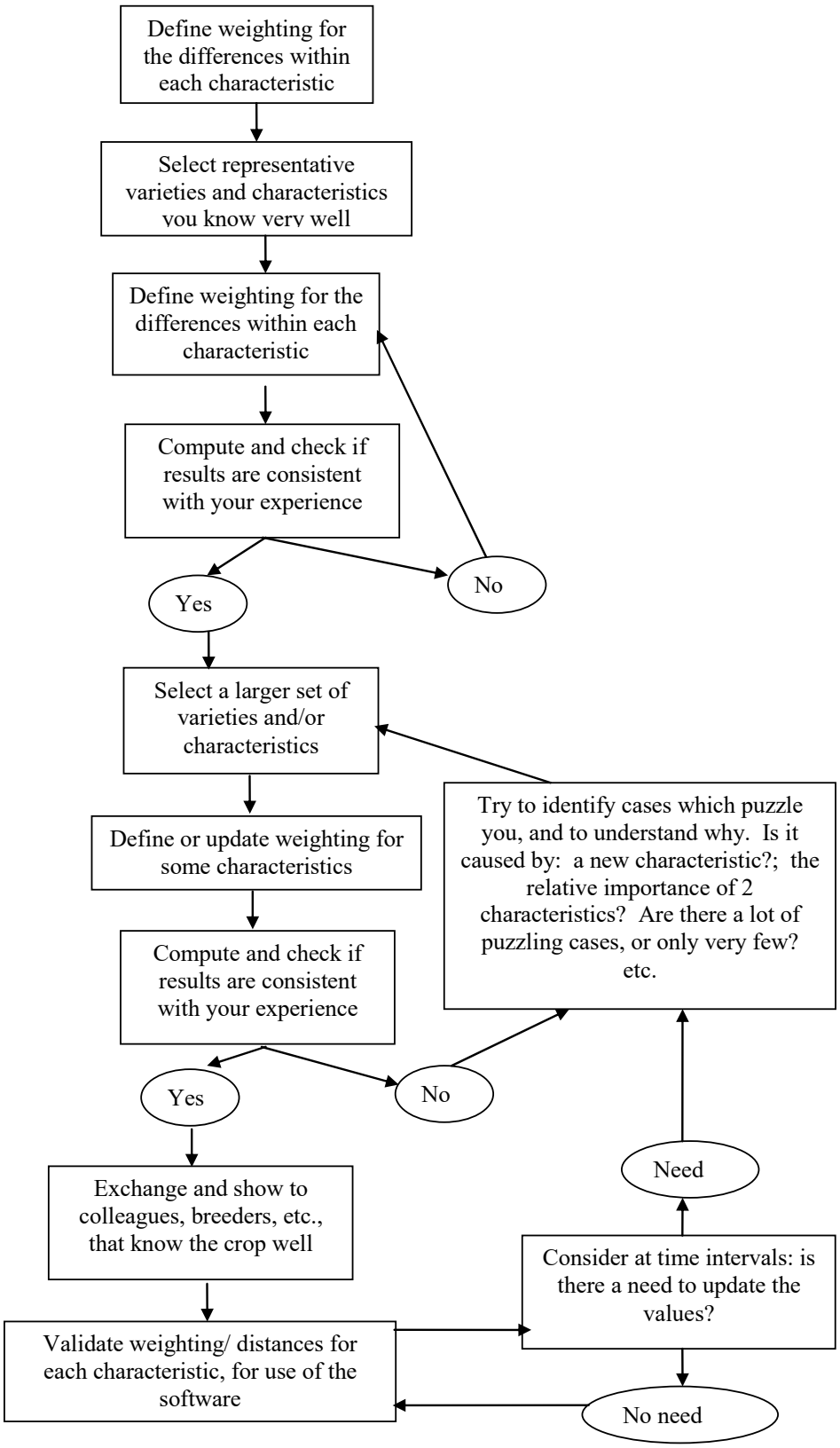
1.3.6.10 On the right of the Display tree, the raw data for *Variety 112* and *Variety 26* are visible for the 6 qualitative characteristics observed on both varieties (two cycles).

1.3.6.11 The third column “weighting" is the weighting according to the pre-defined matrices. The notes for both varieties are displayed for the two available cycles (Std stands for “studied” which are the candidate varieties).

1.3.6.12 As noted in red, if two varieties have the same description on a given characteristic, this characteristic is not displayed.

1.3.6.13 In this screen copy the varieties have been numbered for sake of confidentiality, the crop expert can name the varieties according to their need (lot or application number, name, etc.).

Diagram 3: "Try-and-check" process to define and revise the weightings for a crop



2. PARENT FORMULA OF HYBRID VARIETIES

2.1 Introduction

2.1.1 When examining distinctness of hybrid varieties, authorities may consider the possibility of using the parental formula approach described in this section. In cases where it is considered that the use of the parental formula might be appropriate, this possibility is mentioned in the Test Guidelines.

2.1.2 The use of the parental formula requires that the difference between parent lines is sufficient to ensure that the hybrid obtained from those parents is distinct. The method is based on the following steps:

(i) description of parent lines according to the Test Guidelines;

(ii) checking the originality of those parent lines in comparison with the variety collection, based on the table of characteristics in the Test Guidelines, in order to identify similar parent lines;

(iii) checking the originality of the hybrid formula in relation to the hybrids in the variety collection, taking into account the most similar parent lines; and

(iv) assessment of distinctness at the hybrid level for varieties with a similar formula.

2.2 Requirements of the method

The application of the method requires:

(i) a declaration of the formula and submission of plant material of the parent lines of hybrid varieties;

(ii) inclusion in the variety collection of the parent lines used as parents in the hybrid varieties of the variety collection (for guidance on the constitution of a variety collection see document TGP/4 section 1) and a list of the formulae of the hybrid varieties;

(iii) application of the method to all varieties in the variety collection. This condition is important to obtain the full benefit; and

(iv) a rigorous approach to assess the originality of any new parent line in order to be confident on the distinctness of the hybrid variety based on that parent line.

2.3 Assessing the originality of a new parent line

2.3.1 The originality of a parental line is assessed using the characteristics included in the relevant Test Guidelines.

2.3.2 The difference between parent lines must be sufficient to be sure that hybrids produced using different parent lines will be distinct. For example:

Characteristic 1: a characteristic having two states of expression (absent/present), which are determined by two alleles of a single gene, with one dominant allele (+) for the expression “present” and one recessive allele (-) for the expression “absent”.

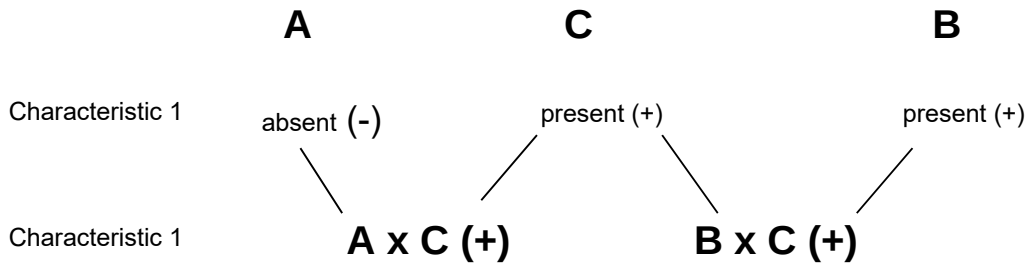
Three parent lines:

A: with the recessive allele (-) with expression “absent”
B: with the dominant allele (+) with expression “present”
C: with the dominant allele (+) with expression “present”

Crossing the above-mentioned parent lines to obtain the following F1 hybrids:

(A x C): having expression “present” for Characteristic 1
(B x C): having expression “present” for Characteristic 1

The following diagram shows the ways the two different crossings result in the same expression of Characteristic 1 (i.e. “present” in both hybrids), although parent line A(-) and parent line B(+) have different expressions.



2.3.3 Although the parent lines A and B are clearly different for characteristic 1, the two hybrid varieties A x C and B x C have the same expression. Thus, a difference between A and B for Characteristic 1 is not sufficient.

2.3.4 With a more complex genetic control involving several genes, not precisely described, the interaction between the different alleles of each gene and between genes might also lead to similar expression at the level of the hybrid varieties. In such cases, a larger difference is appropriate to establish distinctness between two parent lines.

2.3.5 Determining the difference required is mainly based on a good knowledge of the species, of the characteristics and, when available, on their genetic control.

2.4 Verification of the formula

2.4.1 The aim of verifying the formula is to check if the candidate hybrid variety has been produced by crossing the parent lines declared and submitted by the applicant.

2.4.2 Different characteristics can be used to perform this check when the genetic pattern of each parent can be identified in the hybrid. Generally, characteristics based on polymorphism of enzymes or of some storage proteins can be used.

2.4.3 If no suitable characteristics are available, the only possibility is to cross the parent lines using the plant material submitted by the applicant and to compare the hybrid variety seedlots (the sample submitted by the applicant and the sample harvested after the cross).

2.5 Uniformity and stability of parent lines

2.5.1 The uniformity and stability of the parent lines should be assessed according to the appropriate recommendations for the variety concerned. The uniformity and stability of the parent lines are important for the stability of the hybrid. Another requirement for the stability of the hybrid is the use of the same formula for each cycle of the hybrid seed production.

2.5.2 A check of the uniformity on the hybrid should also be done, even if distinctness of the hybrid has been established on the basis of the parent lines.

2.6 Description of the hybrid

A description of the hybrid variety should be established, even where the distinctness of the hybrid has been established on the basis of the parent formula.

3. THE COMBINED OVER-YEARS CRITERIA FOR DISTINCTNESS (COYD)

3.1 Summary of requirements for application of method

COYD is an appropriate method for assessing the distinctness of varieties where:

- the characteristic is quantitative;
- there are some differences between plants (or plots) of a variety;
- observations are made on a plant (or plot) basis over at least two years or growing cycles, and these should be carried out at a single location;
- there should be at least 10, and preferably at least 20 degrees of freedom for the varieties-by-years mean square in the COYD analysis of variance, if there are not, then in some circumstances Long-Term COYD can be used whereby additional data from other varieties and earlier years are used and the degrees of freedom for the varieties-by-years mean square is increased correspondingly (see 3.6.2).

3.2 Summary

3.2.1 Document TGP/9/1, section 5.2.4.5.1.1 explains that “To assess distinctness for varieties on the basis of a quantitative characteristic it is possible to calculate a minimum distance between varieties such that, when the distance calculated between a pair of varieties is greater than this minimum distance, they may be considered as ‘distinct’ in respect of that characteristic. Amongst the possible ways of establishing minimum distances is the method known as the "Combined-Over-Years Distinctness (COYD)”. The COYD analysis takes into account variation between years. Its main use is for cross-pollinated, including synthetic, varieties but, if desired, it can also be used for self-pollinated and vegetatively propagated varieties in certain circumstances. This method requires the size of the differences to be sufficiently consistent over the years and takes into account the variation between years.

3.2.2 The COYD method involves:

- for each characteristic, taking the variety means from the two or three years of trials for candidates and established varieties and producing over-year means for the varieties;
- calculate a least significant difference (LSD), based on variety-by-years variation, for comparing variety means;
- if the over-years mean difference between two varieties is greater than or equal to the LSD then the varieties are said to be distinct in respect of that characteristic.

3.2.3 The main advantages of the COYD method are:

- it combines information from several seasons into a single criterion (the “COYD criterion”) in a simple and straightforward way;
- it ensures that judgements about distinctness will be reproducible in other seasons; in other words, the same genetic material should give similar results, within reasonable limits, from season-to-season;
- the risks of making a wrong judgement about distinctness are constant for all characteristics.

3.3 Introduction

The following sections describe:

- the principles underlying the COYD method;
- UPOV recommendations on the application of COYD to individual species;

– details of ways in which the procedure can be adapted to deal with special circumstances. This includes when there are small numbers of varieties in trial;

– the computer software which is available to apply the procedure.

3.4 The COYD method

3.4.1 The COYD method aims to establish for each characteristic a minimum difference, or distance, which, if achieved by two varieties in trials over a period of two or three years, would indicate that those varieties are distinct with a specified degree of confidence.

3.4.2 The method uses variation in variety expression of a characteristic from year-to-year to establish the minimum distance. Thus, characteristics which show consistency in variety ranking between years will have smaller minimum distances than those with marked changes in ranking.

3.4.3 Calculation of the COYD criterion involves analysing the variety-by-year table of means for each characteristic to get an estimate of the varieties-by-years variation, which is used in the next step: to calculate an LSD. Usually data for all candidate and established varieties which appeared in trials over the two or three test years are included in the table, the analysis is by analysis of variance, the varieties-by-years mean square is used as the estimate of the varieties-by-years variation, and the resulting LSD is known as the COYD LSD. However, where there are small numbers of varieties in trial, the approach is different.

3.4.4 Where there are small numbers of varieties in trial, the table used to calculate of the COYD criterion is expanded with means from other varieties and earlier years, a different method of analysis is used to get a varieties-by-years mean square to estimate the varieties-by-years variation, and the resulting LSD is known as the Long-Term LSD. This is discussed later.

3.4.5 Equation [1]

$$LSD_p = t_p \times \sqrt{2} \times SE(\bar{x})$$

where $SE(\bar{x})$ is the standard error of a variety's over-year mean calculated as:

$$SE(\bar{x}) = \sqrt{\frac{\text{varieties - by - years mean square}}{\text{number of test years}}}$$

and t_p is the value in Student's t table appropriate for a two-tailed test with probability p and with degrees of freedom associated with the variety-by-years mean square. The probability level p that is appropriate for individual species is discussed under UPOV RECOMMENDATIONS ON COYD below.

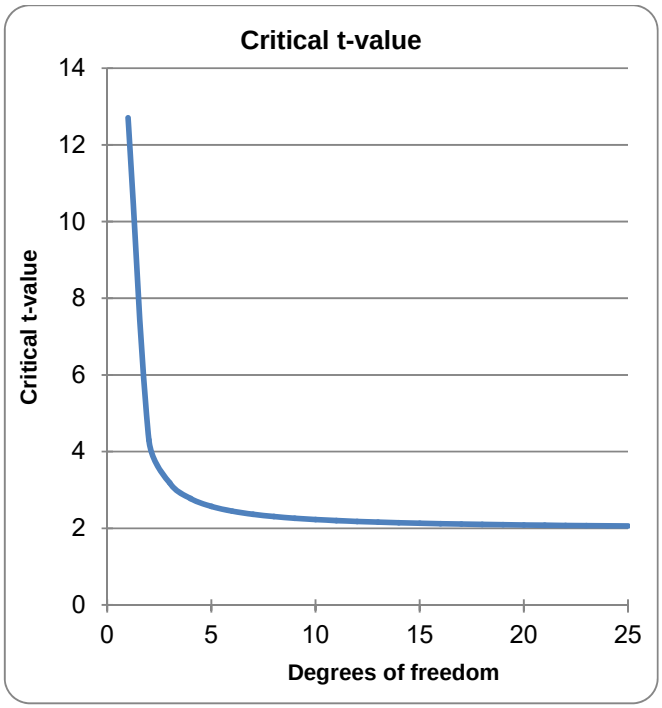
3.4.6 An example of the application of COYD to a small data set is given in Figure 1. Statistical details of the method are in Part II section 3.9. Further information about the COYD criterion can be found in Patterson and Weatherup (1984).

3.5 Use of COYD

3.5.1 COYD is an appropriate method for assessing the distinctness of varieties where:

- the characteristic is quantitative;
- there are some differences between plants (or plots) of a variety;
- observations are made on a plant (or plot) basis over two or more years;
- there should be at least 10, and preferably at least 20 degrees of freedom for the varieties-by-years mean square in the COYD analysis of variance, if there are not, then in some circumstances Long-Term COYD can be used whereby additional data from other varieties and earlier years are used and the degrees of freedom for the varieties-by-years mean square is increased correspondingly (see section 3.6.2).

The reason for this recommendation is to ensure that the varieties-by-years mean square is based on sufficient data to be a reliable estimate of the varieties-by-years variation in the LSD. The fewer the data, the fewer the degrees of freedom for the varieties-by-years mean square, and the less reliable the estimate of the varieties-by-years variation used in the LSD. This is compensated for by use of a larger critical t-value, t_p, in the LSD. The result is a less powerful test, which means that there is a reduced chance of declaring varieties as being distinct. From the graph below, it can be seen that the power of the test is good with 20 or more degrees of freedom for the varieties-by-years mean square, that it is still reasonably powerful if the degrees of freedom drop to 10, though more is preferable.



Twenty degrees of freedom corresponds to 11 varieties common in three years of trials, or 21 varieties common in two years, whereas, ten degrees of freedom corresponds to 6 varieties common in three years of trials, or 11 varieties common in two years. Trials with fewer varieties in common over years are considered to have small numbers of varieties in trial.

3.5.2 A pair of varieties is considered to be distinct if their over-years means differ by at least the COYD LSD in one or more characteristics.

3.5.3 The UPOV recommended probability level p for the t_p value used to calculate the COYD LSD differs depending on the crop and for some crops depends on whether the test is over two or three years. The testing schemes that usually arise in distinctness testing are described in Part II section 3.11.

3.6 Adapting COYD to special circumstances

3.6.1 Differences between years in the range of expression of a characteristic

Occasionally, marked differences between years in the range of expression of a characteristic can occur. For example, in a late spring, the heading dates of grass varieties can converge. To take account of this effect it is possible to fit extra terms, one for each year, in the analysis of variance. Each term represents the linear regression of the observations for the year against the variety means over all years. The method is known as modified joint regression analysis (MJRA) and is recommended in situations where there is a statistically significant ($p \leq 1\%$) contribution from the regression terms in the analysis of variance. Statistical details, and a computer program to implement the procedure, are described in Part II sections 3.9 and 3.10.

3.6.2 Small numbers of varieties in trials: Long-Term COYD

3.6.2.1 It is recommended that there should be at least 20 degrees of freedom for the varieties-by-years mean square in the COYD analysis of variance. This is in order to ensure that the varieties-by-years mean square is based on sufficient data to be a reliable estimate of the varieties-by-years variation for the LSD. Twenty degrees of freedom corresponds to 11 varieties common in three years of trials, or 21 varieties common in two years. Trials with fewer varieties in common over years are considered to have small numbers of varieties in trial.

3.6.2.2 In trials with small numbers of varieties the variety-by-year tables of means can be expanded to include means for earlier years, and if necessary, other established varieties. As not all varieties are present in all years, the resulting tables of variety-by-year means are not balanced. Consequently, each table is analysed by the least squares method of fitted constants (FITCON) or by REML, which produces an alternative varieties-by-years mean square as a long-term estimate of variety-by-years variation. This estimate has more degrees of freedom as it is based on more years and varieties.

$$\text{degrees of freedom} = \begin{pmatrix} \text{No. values in expanded} \\ \text{variety-by-year table} \end{pmatrix} - (\text{No. varieties}) - (\text{No. years}) + 1$$

3.6.2.3 The alternative varieties-by-years mean square is used in equation [1] above to calculate an LSD. This LSD is known as a “Long-Term LSD” to distinguish it from COYD LSD based on just the test years and varieties. The Long-Term LSD is used in the same way as the COYD LSD is used to assess the distinctness of varieties by comparing their over-year (the test years) means. The act of comparing the means of varieties using a “Long-Term LSD” is known as “Long-Term COYD”.

3.6.2.4 Long-Term COYD should only be applied to those characteristics lacking the recommended minimum degrees of freedom. However, when there is evidence that a characteristic’s LSD fluctuates markedly across years, it may be necessary to base the LSD for that characteristic on the current two or three-years of data, even though it has few degrees of freedom.

3.6.2.5 Figure 2 gives an example of the application of Long-Term COYD to the Italian ryegrass characteristic “Growth habit in spring”. A flow diagram of the stages and DUST modules used to produce Long-Term LSD’s and perform Long-Term COYD is given in Figure B2 in Part II: section 3.10.

3.6.2.6 Marked year-to-year changes in an individual variety’s characteristic

Occasionally, a pair of varieties may be declared distinct on the basis of a t-test which is significant solely due to a very large difference between the varieties in a single year. To monitor such situations a check statistic is calculated, called F_3, which is the variety-by-years mean square for the particular variety pair expressed as a ratio of the overall variety-by-years mean square. This statistic should be compared with F-distribution tables with 1 and *g*, or 2 and *g*, degrees of freedom, for tests with two or three years of data respectively where *g* is the degrees of freedom for the variety-by-years mean square. If the calculated F_3 value exceeds the tabulated F value at the 1% level then an explanation for the unusual result should be sought before making a decision on distinctness.

3.6.3 Crops with grouping characteristics

3.6.3.1 In some crops, it is possible to use grouping characteristics to define groups of varieties such that all the varieties within a group will be distinct from all the varieties of any other group (“distinct groups”). This grouping may be preserved in trial layouts so that, within a replicate, varieties in the same group are in the same vicinity. (See TG/1/3, section 4.8 “Functional Categorization of Characteristics).

3.6.3.2 When grouping is possible, such that all the varieties within a group will be distinct from all varieties of any other group, comparisons are only necessary between varieties in the same group. Since varieties within groups tend to be more similar to each other, it is possible to tailor the COYD method by accounting for the groups. If there is a sufficient number of varieties in each group, COYD can be applied separately for each group. However, in practice some groups will generally have too few varieties. In such cases, the over-years analysis of variance (COYD) can be adjusted to take into account the grouping. This method is known as COYD for groups (COYDG).

3.6.3.3 Whereas the standard COYD analysis of variance has terms for 'year' and 'variety', COYDG has terms for 'year', 'group', 'variety-within-group' and 'group-by-year'. The LSD is then calculated for comparisons between pairs of varieties within the same group. It is assumed that the same standard error is applicable within all groups. Note that a larger LSD will apply for comparisons between pairs of varieties from different groups.

3.6.3.4 So the LSD for COYDG is given by $LSD_p = t_p \times SED_G$

where SED_G is the standard error for the difference between two varieties within the same group and calculated as:

$$SED_G = \frac{\sqrt{2 \times \text{varieties} - \text{within} - \text{group - by- years mean square}}}{\text{number of test years}}$$

Note that the varieties-within group-by-years mean square is the same as the residual mean square from the COYDG analysis of variance.

3.6.3.5 The COYDG LSD is used in place of the COYD LSD as a distinctness criterion. Usually it should be smaller. However it is sensible to verify whether this is true on historical data sets.

3.6.3.6 The COYDG method can be applied using GTVRP module of the DUST package for the statistical analysis of DUS data, which is available from Dr. Sally Watson (Email: info@afbini.gov.uk) or from http://www.afbini.gov.uk/dustnt.htm.

3.7 Implementing COYD

COYD is an appropriate method for assessing the distinctness of varieties where:

- the characteristic is quantitative;
- there are some differences between plants (or plots) of a variety;
- observations are made on a plant (or plot) basis over two or more years;
- there should be at least 10, and preferably at least 20 degrees of freedom for the varieties-by-years mean square in the COYD analysis of variance, or if there are not, then Long-Term COYD can be used (see 3.6.2).

The COYD method can be applied using TVRP module of the DUST package for the statistical analysis of DUS data, which is available from Dr. Sally Watson (Email: info@afbini.gov.uk) or from http://www.afbini.gov.uk/dustnt.htm. Sample outputs are given in Part II section 3.10.

3.8 References

DIGBY, P.G.N. (1979). Modified joint regression analysis for incomplete variety x environment data. J. Agric. Sci. Camb. 93, 81-86.

PATTERSON, H.D. & WEATHERUP, S.T.C. (1984). Statistical criteria for distinctness between varieties of herbage crops. J. Agric. Sci. Camb. 102, 59-68.

TALBOT, M. (1990). Statistical aspects of minimum distances between varieties. UPOV TWC Paper TWC/VIII/9, UPOV, Geneva.

Figure 1: Illustrating the calculation of the COYD criterion

Characteristic: Days to ear emergence in perennial ryegrass varieties

Varieties	Years 1	Years 2	Years 3	Over Year Means	*Difference (Varieties compared to C2)*	
Reference		Means				
R1	38	41	35	38	*35*	*D*
R2	63	68	61	64	*9*	*D*
R3	69	71	64	68	*5*	*D*
R4	71	75	67	71	*2*	
R5	69	78	69	72	*1*	
R6	74	77	71	74	*-1*	
R7	76	79	70	75	*-2*	
R8	75	80	73	76	*-3*	
R9	78	81	75	78	*-5*	*D*
R10	79	80	75	78	*-5*	*D*
R11	76	85	79	80	*-7*	*D*
Candidate						
C1	52	56	48	52	*21*	*D*
C2	72	79	68	73	*0*	-
C3	85	88	85	86	*-13*	*D*

ANALYSIS OF VARIANCE

Source	df	Mean square
Years	2	174.93
Variety	13	452.59
Variety-by-years	26	2.54

$LSD_p = t_p * \sqrt{2} * SE(\overline{X})$

$LSD_{0.01} = 2.779 * 1.414 * \sqrt{\ } (2.54/3) = 3.6$

Where t_p is taken from Student's t table with p = 0.01 (two-tailed) and 26 degrees of freedom.

To assess the distinctness of a candidate, the difference in the means between the candidate and all other varieties is computed. In practice a column of differences is calculated for each candidate. In this case, varieties with mean differences greater than, or equal to, 3.6 are regarded as distinct (marked *D* above).

Figure 2: Illustrating the application of Long-Term COYD

Characteristic: Growth habit in spring in Italian ryegrass varieties

Varieties	Years 1	2	3*	4*	5*	Mean over test years	*Difference (Varieties compared to C2)*	
Reference			Means					
R1	43	42	41	44				
R2		39	45					
R3	43	38	41	45	40	42	*6*	*D*
R4	44	40	42	48	44	44.7	*3.3*	*D*
R5	46	43	48	49	45	47.3	*0.7*	
R6	51	48	52	53	51	52	*-4*	*D*
Candidate								
C1			43	45	44	44	*4*	*D*
C2			49	50	45	48	*0*	
C3			48	53	47	49.3	*-1.3*	

* indicates a test year

The aim is to assess the distinctness of the candidate varieties C1, C2 & C3 grown in the test years 3, 4 & 5.

The trial has a small number of varieties in trial because there are just seven varieties in common over the test years 3, 4 & 5 (data marked by a black border).

FITCON analysis of the variety-by-years table of means expanded to nine varieties in five years gives: varieties-by-years mean square = 1.924, on 22 degrees of freedom

Long-term $LSD_p = t_p * \sqrt{2} * SE(\overline{X})$

Long-term $LSD_{0.01} = 2.819 * 1.414 * \sqrt{(1.924/3)} = 3.19$

Where t_p is taken from Student's t table with $p = 0.01$ (two-tailed) and 22 degrees of freedom

To assess the distinctness of a candidate, the difference in the means between the candidate and all other varieties is computed. In practice a column of differences is calculated for each candidate. In the case of variety C2, varieties with mean differences greater than, or equal to 3.19 are regarded as distinct (marked *D* above).

3.9 COYD statistical methods

3.9.1 Analysis of variance

The standard errors used in the COYD criterion are based on an analysis of variance of the variety-by-years table of a characteristic's means. For *m* years and *n* varieties this analysis of variance breaks down the available degrees of freedom as follows:

Source	Df
Years	*m*-1
Varieties	*n*-1
Varieties-by-years	(*m*-1)(*n*-1)

3.9.2 Modified joint regression analysis (MJRA)

3.9.2.1 As noted above, the COYD criterion bases the standard error of a variety mean on the varieties-by-years variation as estimated by the varieties-by-years mean square. Systematic variation can sometimes be identified as well as non-systematic variation. This systematic effect causes the occurrence of different slopes of the regression lines relating variety means in individual years to the average variety means over all years. Such an effect can be noted for the heading date characteristic in a year with a late spring: the range of heading dates can be compressed compared with the normal. This leads to a reduction in the slope of the regression line for variety means in that year relative to average variety means. Non-systematic variation is represented by the variation about these regression lines. Where only non-systematic varieties-by-years variation occurs, the slope of the regression lines have the constant value 1.0 in all years. However, when systematic variation is present, slopes differing from 1.0 occur but with an average of 1.0. When MJRA is used, the standard error of a variety mean is based on the non-systematic part of the varieties-by-year variation.

3.9.2.2 The difference between the total varieties-by-years variation and the varieties-by-years variation adjusted by MJRA is illustrated in Figure B1, where variety means in each of three years are plotted against average variety means over all years. The variation about three parallel lines fitted to the data, one for each year, provides the total varieties-by-years variation as used in the COYD criterion described above. These regression lines have the common slope 1.0. This variation may be reduced by fitting separate regression lines to the data, one for each year. The resultant residual variation about the individual regression lines provides the MJRA-adjusted varieties-by-years mean square, on which the standard error for a variety mean may be based. It can be seen that the MJRA adjustment is only effective where the slopes of the variety regression lines differ between years, such as can occur in heading dates.

3.9.2.3 The use of this method in assessing distinctness has been included as an option in the computer program which applies the COYD criterion in the DUST package. It is recommended that it is only applied where the slopes of the variety regression lines are significantly different between years at the 1% significance level. This level can be specified in the computer program.

3.9.2.4 To calculate the adjusted variety means and regression line slopes the following model is assumed.

$$y_{ij} = u_j + b_j v_i + e_{ij}$$

where y_{ij} is the value for the ith variety in the jth year.

u_j is the mean of year j (j = 1, ..., m)

b_j is the regression slope for year j

v_i is the effect of variety i (i = 1, ..., n)

e_{ij} is an error term.

3.9.2.5 From equations (6) and (7) of Digby (1979), with the meaning of years and varieties reversed, the following equations relating these terms are derived for the situation where data are complete:

$$\sum_{i=j}^{n} \mathrm{v}_i y_{ij} = \mathrm{b}_j \sum_{i=1}^{n} v_i^2$$

$$\sum_{j=1}^{m} b_j y_{ij} = v_i \sum_{j=1}^{m} b_j^2$$

3.9.2.6 These equations are solved iteratively. All b_j values are taken to be 1.0 as a starting point in order to provide values for the v_i's. The MJRA residual sum of squares is then calculated as:

$$\sum_{j=1}^{m} \sum_{i-1}^{n} \left(y_{ij} - u_j - b_j v_i \right)^2$$

3.9.2.7 This sum of squares is used to calculate the MJRA-adjusted varieties-by-years mean square on $(\mathrm{m}-1)(\mathrm{n}-1)-\mathrm{m}+1$ degrees of freedom.

3.9.3 Comparison of COYD with other criteria

It can be shown that, for a three-year test, the COYD criterion applied at the 1% probability level is of approximately the same stringency as the 2x1% criterion for a characteristic where the square root of the ratio of the variety-by-years mean square to the variety-by-replicates-within-trials mean square (λ) has a value of 1.7. The COYD criterion applied at the 1% level is less stringent than the 2x1% criterion if $\lambda < 1.7$, and more stringent if $\lambda > 1.7$.

3.10 COYD software

3.10.1 An example of the output from the computer program in the DUST package which applies the COYD criterion is given in Tables B 1 to 3. It is taken from a perennial ryegrass (diploid) trial involving 40 varieties selected from the variety collection (R1 to R40) and 9 candidate varieties (C1 to C9) in 6 replicates on which 8 characteristics were measured over the years 1988, 1989 and 1990.

3.10.2 Each of the 8 characteristics is analysed by analysis of variance. As this analysis is of the variety-by-year-by-replicate data, the mean squares are 6 (= number of replicates) times the size of the mean squares of the analysis of variance of the variety-by-year data referred to in the main body of this paper. The results are given in Table B 1. Apart from the over-year variety means there are also presented:

YEAR MS:	the mean square term for years
VARIETY MS:	the mean square term for varieties
VAR.YEAR MS:	the mean square for varieties-by-years interaction
F1 RATIO:	ratio of VARIETY MS to VAR.YEAR MS (a measure of the discriminating power of the characteristic - large values indicate high discriminating power)
VAR.REP MS:	average of the variety-by-replicate mean squares from each year
LAMBDA VALUE (λ):	square root of the ratio of VAR.YEAR MS to VAR.REP MS
BETWEEN SE:	standard error of variety means over trials on a plot basis i.e. the square root of the VAR.YEAR MS divided by 18 (3 years x 6 replicates)
WITHIN SE:	the standard error of variety means within a trial on a plot basis i.e. the square root of the VAR.REP MS divided by 18
DF:	the degrees of freedom for varieties-by-years
MJRA SLOPE:	the slope of the regression of a single year's variety means on the means over the three years
REGR F VALUE:	the mean square due to MJRA regression as a ratio of the mean square about regression
REGR PROB:	the statistical significance of the REGR F VALUE
TEST:	indicates whether MJRA adjustment was applied (REG) or not (COY).

3.10.3 Each candidate variety is compared with every other candidate variety and every other variety in the trial selected from the variety collection. The mean differences between pairs of varieties are compared with the LSD for the characteristic. The results for the variety pair R1 and C1 are given in Table B 2. The individual

within year t-values are listed to provide information on the separate years. Varieties R1 and C1 are considered distinct since, for at least one characteristic, a mean difference is COYD significant at the 1% level. If the F_3 ratio for characteristic 8 had been significant at the 1% level rather than the 5% level, the data for characteristic 8 would have been investigated, and because the differences in the three years are not all in the same direction, the COYD significance for characteristic 8 would not have counted towards distinctness.

3.10.4 The outcome in terms of the tests for distinctness of each candidate variety from all other varieties is given in Table B 3, where D indicates “distinct” and ND denotes “not distinct.”

Table B 1: An example of the output from the COYD program showing variety means and analysis of variance of characteristics

PRG (DIPLOID) EARLY N.I. UPOV 1988-90

VARIETY MEANS OVER YEARS

	5 SP.HT	60 NSPHT	8 DEEE	10 H.EE	11 WEE	14 LFL	15 WFL	24 LEAR
1 R1	45.27	34.60	67.87	45.20	70.05	20.39	6.85	24.54
2 R2	42.63	31.84	73.85	41.96	74.98	19.68	6.67	24.44
3 R3	41.57	27.40	38.47	27.14	57.60	17.12	6.85	22.57
4 R4	33.35	21.80	77.78	30.77	78.04	18.25	6.40	21.09
5 R5	37.81	25.86	50.14	27.24	62.64	16.41	6.41	16.97
6 R6	33.90	21.07	78.73	32.84	79.15	19.44	6.46	21.79
7 R7	41.30	31.37	73.19	41.35	71.87	20.98	6.92	24.31
8 R8	24.48	19.94	74.83	32.10	62.38	15.22	6.36	19.46
9 R9	46.68	36.69	63.99	44.84	68.62	18.11	7.02	22.58
10 R10	25.60	20.96	75.64	32.31	57.20	14.68	5.51	20.13
11 R11	41.70	30.31	74.60	40.17	76.15	19.45	6.79	22.72
12 R12	28.95	21.56	66.12	27.96	59.56	14.83	5.53	20.55
13 R13	40.67	29.47	70.63	36.81	74.12	19.97	7.04	24.05
14 R14	26.68	20.53	75.84	34.14	63.29	15.21	6.37	20.37
15 R15	26.78	20.18	75.54	30.39	66.41	16.34	6.01	20.94
16 R16	42.44	27.01	59.03	30.39	72.71	17.29	6.47	22.48
17 R17	27.94	21.58	76.13	32.53	68.37	16.72	6.11	22.03
18 R18	41.34	30.85	69.80	37.28	69.52	20.68	7.09	25.40
19 R19	33.54	23.43	73.65	30.35	75.54	18.97	6.37	22.43
20 R20	44.14	34.48	68.74	42.60	64.17	18.63	6.56	22.02
21 R21	27.77	21.53	80.52	31.59	69.41	16.81	5.81	22.35
22 R22	38.90	27.83	75.68	43.25	75.08	19.63	7.46	23.99
23 R23	42.43	31.80	72.40	42.07	74.77	20.99	6.78	23.57
24 R24	38.50	27.73	73.19	37.12	75.76	19.28	6.91	22.77
25 R25	43.84	29.60	68.82	39.79	74.83	20.63	7.08	22.65
26 R26	49.48	36.53	63.45	42.01	70.46	22.14	7.84	25.91
27 R27	25.61	19.25	78.78	29.81	56.81	15.81	5.07	18.94
28 R28	26.70	20.31	79.41	32.75	66.54	16.92	6.00	21.91
29 R29	27.90	20.94	72.66	29.85	67.14	16.85	6.28	21.79
30 R30	43.07	30.34	70.53	40.51	73.23	19.49	7.28	23.70
31 R31	38.18	25.47	74.23	36.88	80.23	20.40	7.09	25.21
32 R32	35.15	27.56	71.49	37.26	63.10	18.18	6.80	23.13
33 R33	42.71	31.09	67.58	39.14	70.36	19.85	7.12	23.35
34 R34	23.14	18.05	72.09	24.29	59.37	13.98	5.63	18.91
35 R35	32.75	25.41	77.22	38.90	67.07	17.16	6.42	21.49
36 R36	41.71	31.94	77.98	44.33	73.00	19.72	7.09	23.45
37 R37	44.06	32.99	74.38	45.77	71.59	20.88	7.40	24.06
38 R38	42.65	32.97	74.76	44.42	74.13	20.29	7.38	24.32
39 R39	28.79	22.41	76.83	35.91	64.52	16.85	6.34	22.24
40 R40	44.31	31.38	72.24	43.83	74.73	21.53	7.60	25.46
41 C1	42.42	31.68	64.03	40.22	67.02	20.73	6.90	26.16
42 C2	41.77	32.35	86.11	46.03	75.35	20.40	6.96	22.99
43 C3	41.94	31.09	82.04	43.17	74.04	19.06	6.26	23.44
44 C4	39.03	28.71	78.63	45.97	70.49	21.27	6.67	23.37
45 C5	43.97	30.95	72.99	39.14	77.89	19.88	6.68	25.44
46 C6	37.56	27.14	83.29	39.16	81.18	19.47	6.97	25.25
47 C7	38.41	28.58	83.90	42.53	76.44	19.28	6.00	23.47
48 C8	40.08	27.25	83.50	43.33	80.16	22.77	7.92	26.81
49 C9	46.77	34.87	51.89	37.68	61.16	19.25	6.92	24.82
YEAR MS	1279.09	3398.82	3026.80	2278.15	8449.20	672.15	3.36	51.32
VARIETY MS	909.21	476.72	1376.10	635.27	762.41	80.21	6.44	74.17
VAR.YEAR MS	23.16	18.86	14.12	23.16	46.58	4.76	0.28	2.73
F1 RATIO	39.26	25.27	97.43	27.43	16.37	16.84	22.83	27.16
VAR.REP MS	8.83	8.19	4.59	11.95	23.23	1.52	0.15	1.70
LAMBDA VALUE	1.62	1.52	1.75	1.39	1.42	1.77	1.37	1.27
BETWEEN SE	1.13	1.02	0.89	1.13	1.61	0.51	0.13	0.39
WITHIN SE	0.70	0.67	0.50	0.81	1.14	0.29	0.09	0.31
DF	96	94	96	96	96	96	96	96
MJRA SLOPE 88	0.90	0.86	0.99	0.91	0.99	1.09	0.97	0.95
MJRA SLOPE 89	1.05	1.08	1.01	0.99	1.06	0.97	1.02	0.98
MJRA SLOPE 90	1.05	1.06	1.00	1.10	0.95	0.94	1.01	1.07
REGR F VAL	4.66	6.17	0.06	4.48	0.76	1.62	0.29	1.91
REGR PROB	1.17	0.30	93.82	1.39	47.08	20.27	74.68	15.38
TEST	COY	REG	COY	COY	COY	COY	COY	COY

Table B 2: An example of the output from the COYD program showing a comparison of varieties R1 and C1

PRG (DIPLOID) EARLY N.I. UPOV 1988-90

41 C1 VERSUS 1 R1 *** USING REGR WHERE SIG ***

(T VALUES + VE IF 41 C1 > 1 R1)

		SIG LEVELS YEARS				COYD			T VALUES YEARS					
		88	89	90		T	PROB%	SIG	88	89	90	TSCORE	F3	
5	SP.HGHT	-	-	-1	ND	-1.78	7.88	NS	-1.05	-1.34	-2.64	-2.64	0.23	NS
60	NATSPHT	-	-1	-	ND	-2.02	4.61	*	-1.58	-2.61	-1.17	-2.61	0.22	NS
8	DATEEE	-1	-1	+	D	-3.06	0.29	**	-4.14	-6.33	0.80	-6.74	3.99	*
10	HGHT.EE	-1	-1	-5	D	-3.11	0.25	**	-2.79	-2.69	-2.06	-7.55	0.06	NS
11	WIDTHEE	-	-	-	ND	-1.33	18.58	NS	-1.47	-1.80	-0.21	0.00	0.32	NS
14	LGTHFL	+	+	-	ND	0.47	63.61	NS	0.17	1.83	-0.67	0.00	0.56	NS
15	WIDTHFL	+	-	+	ND	0.27	78.83	NS	0.31	-0.41	0.67	0.00	0.17	NS
24	EARLGTH	5	1	+	ND	2.93	0.42	**	2.10	3.33	1.01	5.43	0.84	NS

Notes

1. The three “COYD” columns headed, T PROB% and SIG give the COYD t value, its significance probability and significance level. The t value is the test statistic formed by dividing the mean difference between two varieties by the standard error of that difference. The t value can be tested for significance by comparing it with appropriate values from Students t-table. Calculating and testing a t value in this manner is equivalent to deriving an LSD and checking to see if the mean difference between the two varieties is greater than the LSD.

2. The two right-hand “F3” columns give the F_3 variance ratio statistic and its significance level. The F_3 statistic is defined in Part II section 3.6.2.

3. The sections in boxes refer to earlier distinctness criteria. The three “T VALUES, YEARS” columns headed 88, 89 and 90 are the individual within year t-test values (the Student’s two-tailed t test of the variety means with standard errors estimated using the plot residual mean square), and the three “SIG LEVELS, YEARS” columns headed 88, 89 and 90 give their direction and significance levels. The column containing D and ND gives the distinctness status of the two varieties by the 2 x 1% method criterion described in Part II: Section 4. The column headed T SCORE gives the obsolete T Score statistic and should be ignored.

Table B 3: An example of the output from the COYD program showing the distinctness status of the candidate varieties

PRG (DIPLOID) EARLY N.I. UPOV 1988-90

SUMMARY FOR COYD AT 1.0% LEVEL *** USING REGR ADJ WHEN SIG ***

CANDIDATE VARIETIES		C1	C2	C3	C4	C5	C6	C7	C8	C9
1	R1	D	D	D	D	D	D	D	D	D
2	R2	D	D	D	D	ND	D	D	D	D
3	R3	D	D	D	D	D	D	D	D	D
4	R4	D	D	D	D	D	D	D	D	D
5	R5	D	D	D	D	D	D	D	D	D
6	R6	D	D	D	D	D	D	D	D	D
7	R7	D	D	D	D	D	D	D	D	D
8	R8	D	D	D	D	D	D	D	D	D
9	R9	D	D	D	D	D	D	D	D	D
10	R10	D	D	D	D	D	D	D	D	D
11	R11	D	D	D	D	D	D	D	D	D
12	R1	D	D	D	D	D	D	D	D	D
13	R13	D	D	D	D	ND	D	D	D	D
14	R14	D	D	D	D	D	D	D	D	D
15	R15	D	D	D	D	D	D	D	D	D
16	R16	D	D	D	D	D	D	D	D	D
17	R17	D	D	D	D	D	D	D	D	D
18	R18	D	D	D	D	D	D	D	D	D
19	R19	D	D	D	D	D	D	D	D	D
20	R20	D	D	D	D	D	D	D	D	D
21	R21	D	D	D	D	D	D	D	D	D
22	R22	D	D	D	D	D	D	D	D	D
23	R23	D	D	D	D	D	D	D	D	D
24	R24	D	D	D	D	D	D	D	D	D
25	R25	D	D	D	D	D	D	D	D	D
26	R26	D	D	D	D	D	D	D	D	D
27	R27	D	D	D	D	D	D	D	D	D
28	R28	D	D	D	D	D	D	D	D	D
29	R29	D	D	D	D	D	D	D	D	D
30	R30	D	D	D	D	D	D	D	D	D
31	R31	D	D	D	D	D	D	D	D	D
32	R32	D	D	D	D	D	D	D	D	D
33	R33	D	D	D	D	D	D	D	D	D
34	R34	D	D	D	D	D	D	D	D	D
35	R35	D	D	D	D	D	D	D	D	D
36	R36	D	D	D	ND	D	D	D	D	D
37	R37	D	D	D	D	D	D	D	D	D
38	R38	D	D	D	D	D	D	D	D	D
39	R39	D	D	D	D	D	D	D	D	D
40	R40	D	D	D	D	D	D	D	D	D
41	C1	-	D	D	D	D	D	D	D	D
42	C2	D	-	D	D	D	D	D	D	D
43	C3	D	D	-	D	D	D	ND	D	D
44	C4	D	D	D	-	D	D	D	D	D
45	C5	D	D	D	D	-	D	D	D	D
46	C6	D	D	D	D	D	-	D	D	D
47	C7	D	D	ND	D	D	D	-	D	D
48	C8	D	D	D	D	D	D	D	-	D
49	C9	D	D	D	D	D	D	D	D	-
NO OF ND VARS		0	0	1	1	2	0	1	0	0
DISTINCTNESS		D	D	ND	ND	ND	D	ND	D	D
CANDIDATE VAR		C1	C2	C3	C4	C5	C6	C7	C8	C9

Figure B1. Heading date yearly variety means against over-year variety means

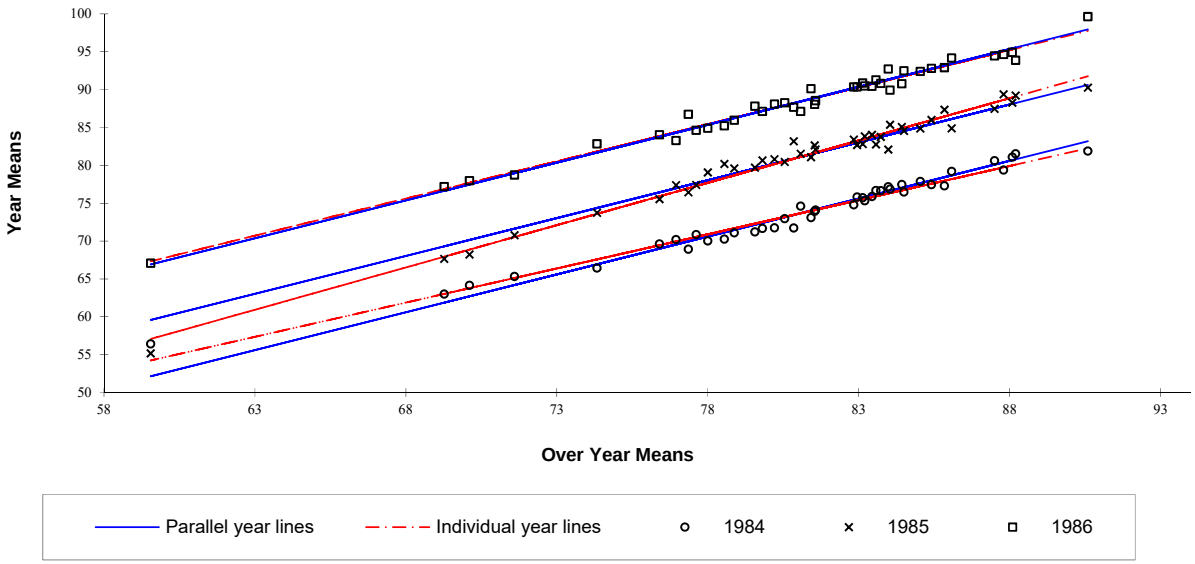
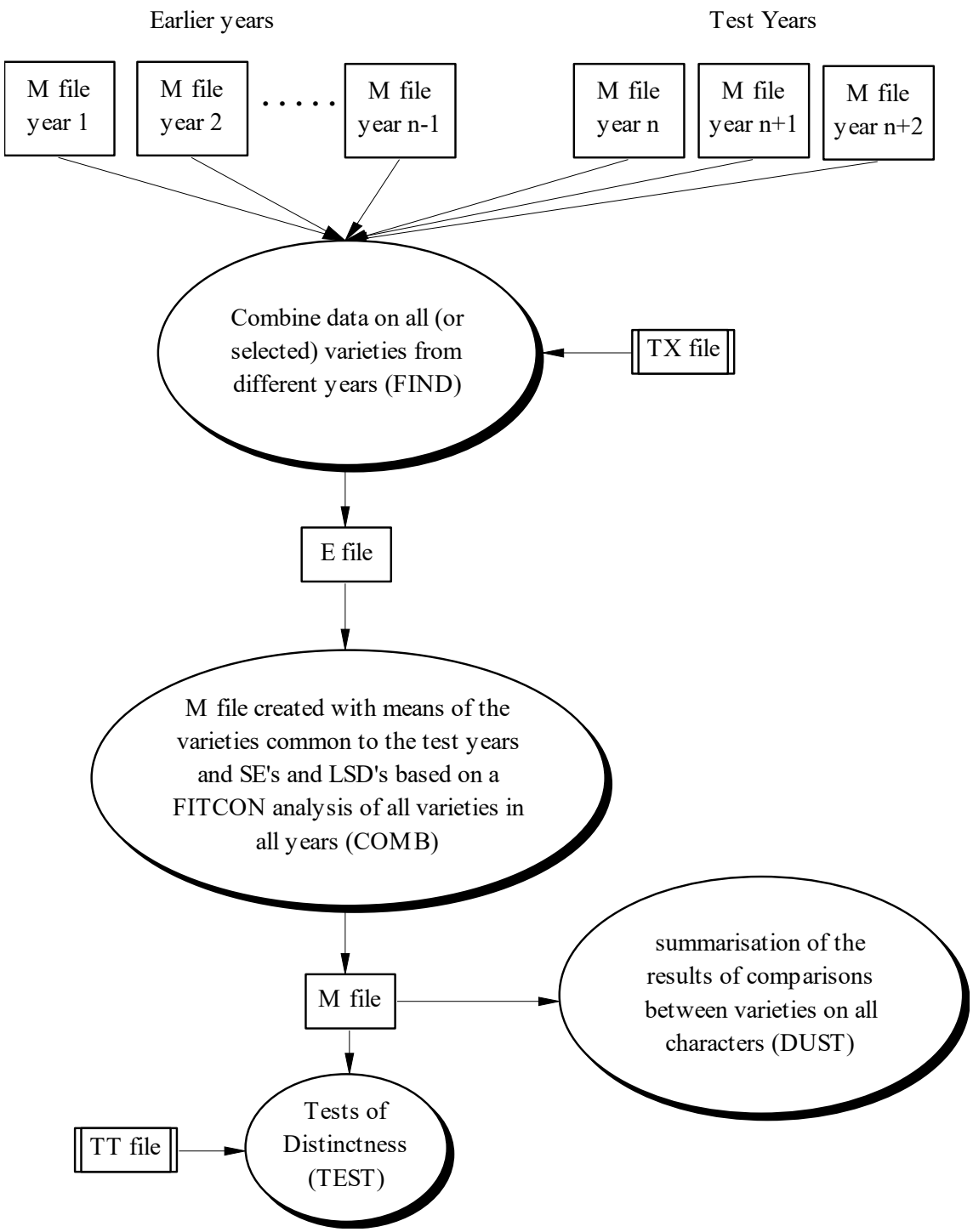


Figure B2. Flow Diagram of the stages and DUST modules used to produce long-term LSD's and perform long-term COYD



3.11 Schemes used for the application of COYD

3.11.1 The following four cases are those which, in general, represent the different situations which may arise where COYD is used in DUS testing:

Scheme A: Test is conducted over 2 independent growing cycles and decisions made after 2 growing cycles (a growing cycle could be a year and is further on denoted by cycle)

Scheme B: Test is conducted over 3 independent growing cycles and decisions made after 3 cycles

Scheme C: Test is conducted over 3 independent growing cycles and decisions made after 3 cycles, but a variety may be accepted after 2 cycles

Scheme D: Test is conducted over 3 independent growing cycles and decisions made after 3 cycles, but a variety may be accepted or rejected after 2 cycles

3.11.2 The stages at which the decisions are made in Cases A to D are illustrated in figures 1 to 4 respectively. These also illustrate the various standard probability levels (p_{d2}, p_{nd2}, p_{d3}, p_{u2}, p_{nu2} and p_{u3}) which are needed to calculate the COYD criteria depending on the case. These are defined as follows:

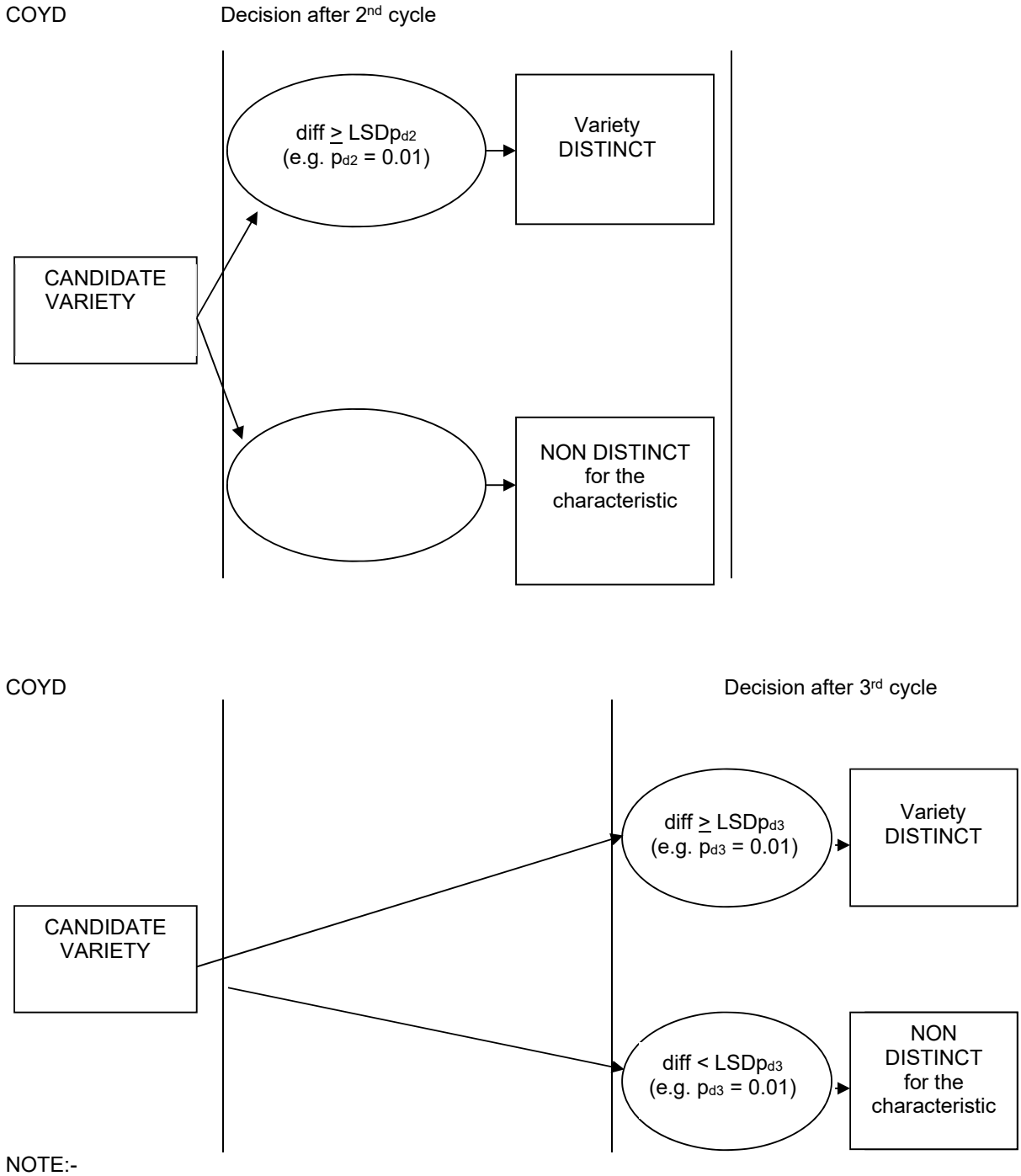
Probability Level	**Used to decide whether a variety is :-**
p_{d2}	distinct after 2 cycles
p_{nd2}	non-distinct in a characteristic after 2 cycles
p_{d3}	distinct after 3 cycles

3.11.3 In Figures 1 to 4 the COYD criterion calculated using say the probability level p_{d2} is denoted by $LSDp_{d2}$ etc. The term “diff” represents the difference between the means of a candidate variety and another variety for a characteristic.

3.11.4 Table 1 summarizes the various standard probability levels needed to calculate the COYD criteria in each of Cases A to D. For example, in Case B only one probability level is needed (p_{d3}), whereas Case C requires two (p_{d2}, p_{d3}).

Table 1	COYD		
CASE	p_{d2}	p_{nd2}	p_{d3}
A			
B			
C			
D			

Figure 1. COYD decisions and standard probability levels (p_i) in Case A



NOTE:-
"diff" is the difference between the means of the candidate variety and another variety for the characteristic.
LSDp is the COYD criterion calculated at probability level p.

Figure 3. COYD decisions and standard probability levels (p_i) in Case C

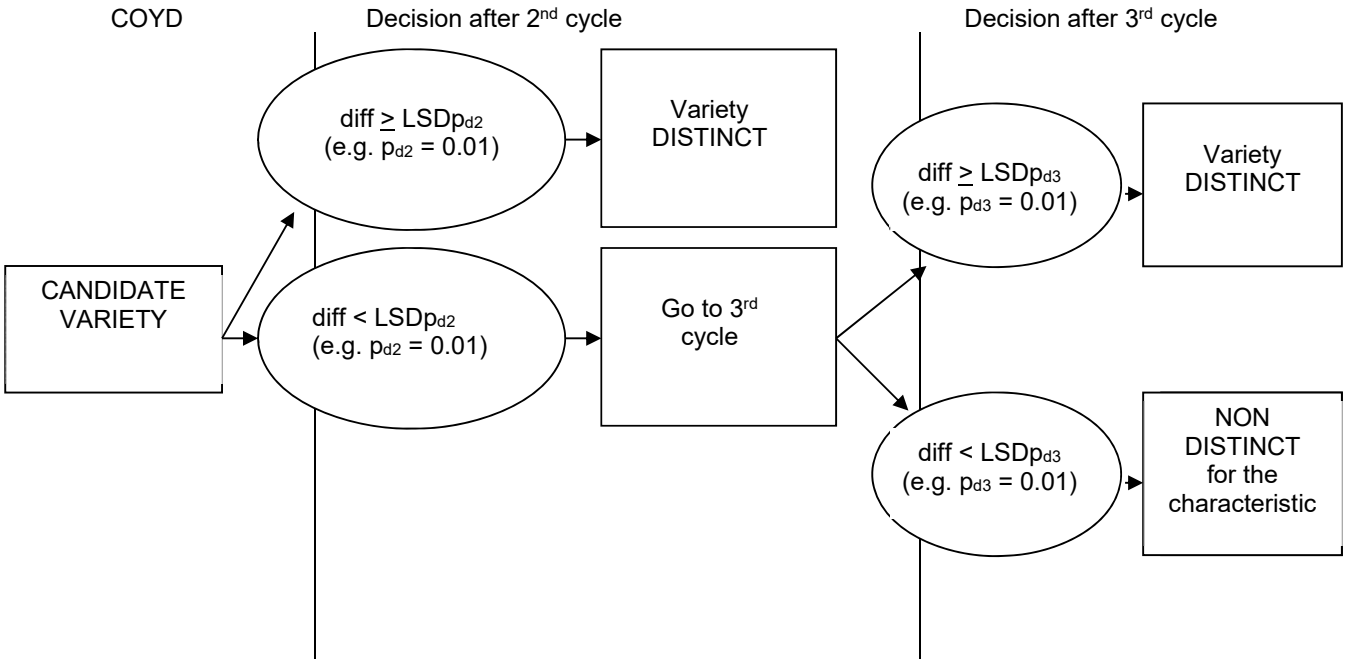
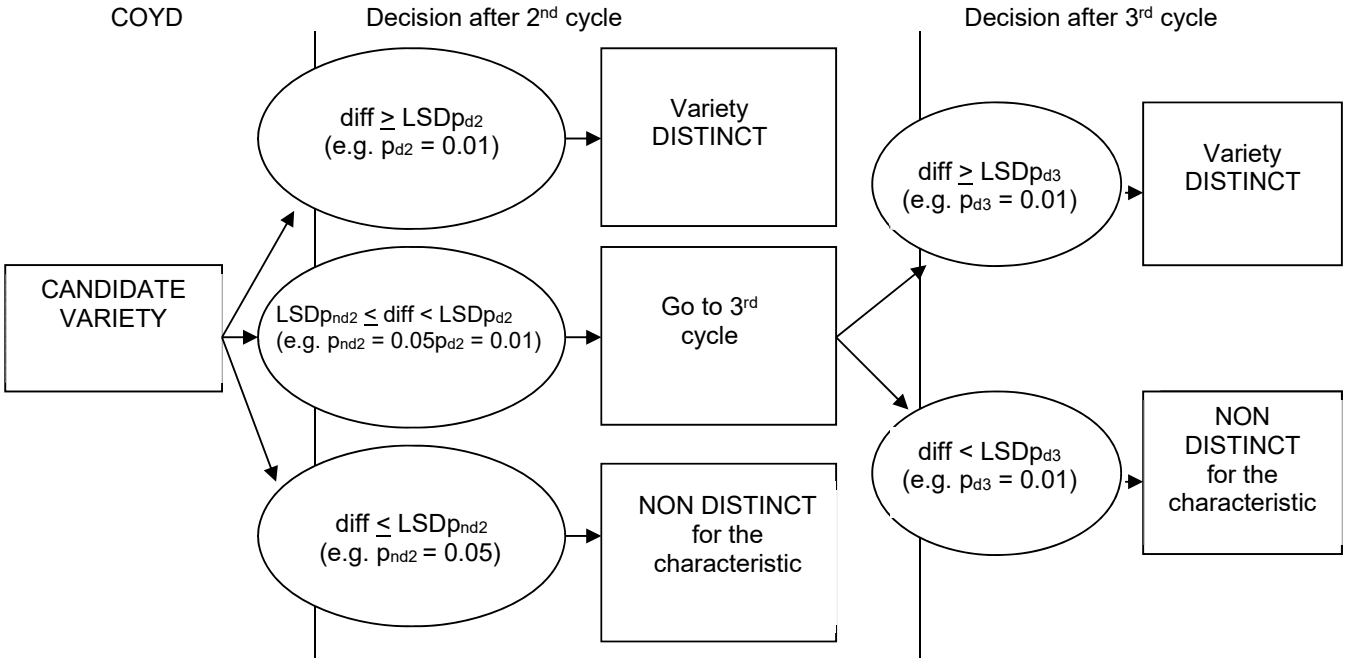


Figure 4. COYD decisions and standard probability levels (p_i) in Case D



NOTE:-
"diff" is the difference between the means of the candidate variety and another variety for the characteristic.
LSDp is the COYD criterion calculated at probability level p.

4. 2X1% METHOD

4.1 Requirements for application of method

4.1.1 The 2x1% Criterion is an appropriate method for assessing the distinctness of varieties where:

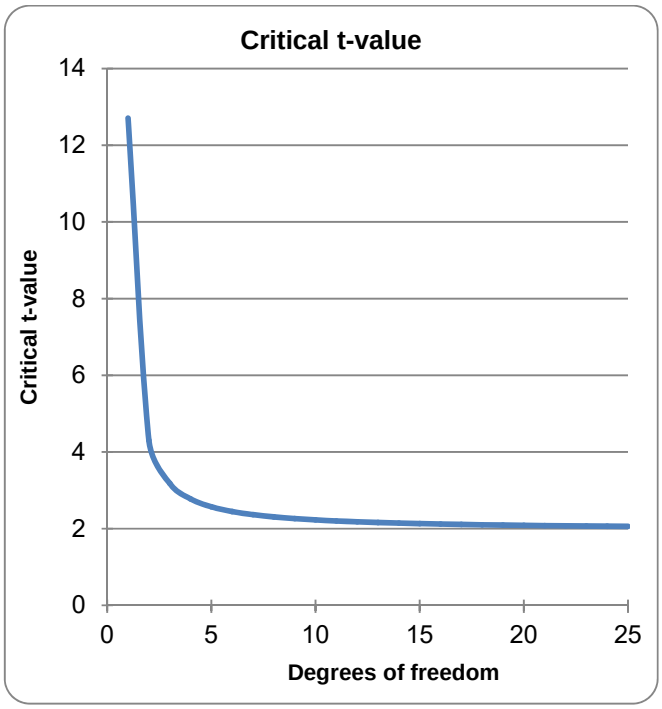
- the characteristic is quantitative;
- there are some differences between plants (or plots) of a variety;
- observations are made on a plant (or plot) basis over two or more years;
- there are at least 10, and preferably at least 20, degrees of freedom for the residual mean square used to estimate the standard error in the t-test in each year;
- to have replicated plots.

4.2 The 2x1% Criterion (Method)

4.2.1 For two varieties to be distinct using the 2x1% criterion, the varieties need to be significantly different in the same direction at the 1% level in at least two out of three years in one or more measured characteristics. The tests in each year are based on Student's two-tailed t-test of the differences between variety means with standard errors estimated using the residual mean square from the analysis of the variety x replicate plot means.

4.2.2 With respect to the 2x1% criterion, compared to COYD, it is important to note that:

- Information is lost because the criterion is based on the accumulated decisions arising from the results of t-tests made in each of the test years. Thus, a difference which is not quite significant at the 1% level contributes no more to the separation of a variety pair than a zero difference or a difference in the opposite direction. For example, three differences in the same direction, one of which is significant at the 1% level and the others at the 5% level would not be regarded as distinct.
- Some characteristics are more consistent over years than others in their expression of differences between varieties. However, beyond requiring differences to be in the same direction in order to count towards distinctness, the 2x1% criterion takes no account of consistency in the size of the differences from year to year.
- It is recommended that there should be at least 10, and preferably at least 20, degrees of freedom for the residual mean square used to estimate the standard error in the t-test in each year. This is to ensure that the residual mean square is based on sufficient data to be a reliable estimate of the varieties-by-replicates variation used in the standard error in the t-test. The fewer the data, the fewer the degrees of freedom for the residual mean square, and the less reliable the estimate of the standard error in the t-test. This is compensated for by use of a larger critical t-value in the t-test. The result is a less powerful test, which means that there is a reduced chance of declaring varieties as being distinct. From the graph below, it can be seen that the power of the test is good with 20 or more degrees of freedom for the residual mean square, that it is still reasonably powerful if the degrees of freedom drop to 10, though more is preferable.



Assuming replicates are arranged in blocks, 20 degrees of freedom corresponds to 11 varieties in three replicates, or 5 varieties in six replicates, whereas, ten degrees of freedom corresponds to 6 varieties in three replicates, or 3 varieties in six replicates.

5. PEARSON'S CHI-SQUARE TEST APPLIED TO CONTINGENCY TABLES

5.1 A contingency table is a table showing the responses of subjects to one factor as a function of another factor. In DUS testing it is generally used for categorical data where individuals of a variety can be allocated to discrete states of expression for a characteristic. Various statistical tests can be used to analyze the data in contingency tables depending on the particular circumstances. For example, Pearson's Chi-square test, as applied to contingency tables, is useful where:

- observations on a characteristic are allocated to two or more categories (classes) and are recorded in a contingency table

- there are some differences between plants (or plots) of a variety;

- the only source of variation should be caused by random sampling, e.g. there should be no variation due to soil conditions, etc.

- the minimum expected value in each category should be five

5.2 In some cases, distinctness may be established by classifying individual varieties into broad groups and demonstrating statistically different grouping patterns for different varieties. Examples include counts based on broad flower color groups - such as dark blue violet versus not dark blue violet and the disease/pest/nematode infection classes. Data based on counts of individuals in a sample/population belonging to each of several classes require statistical analysis capable of dealing with categorical data.

5.3 To use the Chi-square analysis for plant breeder rights' (PBR) purposes, we should consider how we are going to arrive at certain conclusions about distinctness by formulating certain hypotheses using the classification data.

The standard formula for the chi-square statistic used in such analysis is:

$$\chi^2 = \Sigma \frac{(\text{Observed value of a class} - \text{Expected value of a clase})^2}{\text{Expected value}}$$

5.4 Hence, the Chi-square distribution is a continuous distribution based upon an underlying normal distribution.

5.5 The following precautions are to be considered before using the chi-square test.

(1) Selection of the hypothesis to be tested should be based on previously known facts or principles

(2) Given the hypothesis, you should be able to assign expected values for each class correctly. Avoid using the chi-square test if the smallest expected class is less than five. By increasing the sample size the size of the smallest expected class can be made larger. Alternatively, if some classes have a size less than five, either pool those adjacent classes to bring the size of the pooled class to five or more than five, or use an exact test.

(3) Degrees of freedom is defined as the number of classes that are independent to be assigned an arbitrary value. For example, if we have two classes the degrees of freedom is 2-1 = 1. Hence, in using this method to test a hypothesis, the degrees of freedom for the chi-square test is one less than the number of classes.

(4) Avoid using two class situations which follow more like the binomial distribution, with np or nq less than 5. If you encounter such situations, calculate expected values using formulae based on the binomial distribution. In a two class situation, np is the size of one of the classes determined by the number of events (n) times the probability of falling into that class (p). Similarly the size of the other class (nq) is determined by n times the probability (q) of falling into that class. So in a situation where the probability of falling into either class is equal (p=q=0.5) and the sample size is 10 (n) the number expected in each class is 5. Always use Yates Correction for determining the chi-square test with only one degree of freedom.

5.6 Let us examine the following data on the disease scoring of a Lucerne candidate variety and four varieties of common knowledge. The disease scored was *Colletotrichum trifolii* (Characteristic 19, TG/6/5, Lucerne). The scoring was on a 5 class scale, with class 1 (note 9) being resistant and class 5 (note 1) being susceptible.

Contingency table of number of plants counted in different classes in each variety after 7-10 days of inoculation

Note(Class)	Candidate		Variety 1	Variety 2	Variety 3	Variety 4
9(1)	34		12	6	1	7
7(2)	4		7	6	5	10
5(3)	1		9	5	5	5
3(4)	1		7	9	8	7
1(5)	6		9	19	9	15
Total	46		44	45	28	44

5.7 It can be seen from the table that the candidate variety has more plants in the resistant category than the four varieties of common knowledge. However, to statistically test the significance of the difference, we need to formulate a hypothesis:

(1) Whether the four varieties of common knowledge differ significantly or not from the candidate in the distribution of scores i.e. by testing the null hypothesis. The null hypothesis in this case is all the varieties show similar reaction to the Colletotrichum crown rot. This can be done by testing the "distinctness χ^2".

5.8 Pooling of classes to form a new intermediary pooled class is necessary to meet the minimum expected value requirement for the use of the chi square test.

Now the observed data is reduced to:

Class/Score	Candidate		Variety 1	Variety 2	Variety 3	Variety 4
1	*34*		12	6	1	7
2	*6*		23	20	18	22
3	*6*		9	19	9	15
Total	*46*		44	45	28	44

5.9 For each comparison of the candidate with each variety of common knowledge, a two-way table of observed values is formed. The expected values are calculated as the product of the row and column totals divided by the grand total, and the chi square statistic is calculated. The distributions of expected values for different varieties are as follows:

Observed for Variety 1

Class/Score	**Candidate**	**Variety 1**	**Total**
1	34	12	46
2	6	23	29
3	6	9	15
Total	**46**	**44**	**90**

Expected for Variety 1

Class/Score	**Candidate**	**Variety 1**	**Total**
1	23.5=46x46/90	22.5=46x44/90	46
2	14.8=29x46/90	14.2=29x44/90	29
3	7.7=15x46/90	7.3=15x44/90	15
Total	46	44	90

Similarly, using the table of observed data in 5.3.8, the expected values for varieties 2, 3 and 4 are;

Class/Score	Candidate	Variety 2	Total
1	20.2	19.8	40
2	13.1	12.9	26
3	12.6	12.4	25
Total	46	45	91

Class/Score	Candidate	Variety 3	Total
1	21.8	13.2	35
2	14.9	9.1	24
3	9.3	5.7	15
Total	46	28	74

Class/Score	Candidate	Variety 4	Total
1	21.0	20.0	41
2	14.3	13.7	28
3	10.7	10.3	21
Total	46	44	90

5.10 For calculating the "distinctness χ^2" for variety 1

χ^2 = (34-23.5)2/23.5 + (12-22.5)2/22.5 + (6-14.8)2/14.8 + (23-14.2)2/14.2 + (6-7.7)2/7.7 + (9-7.3)2/7.3 = 21.1 on (No rows – 1)(No cols – 1) = 2 df

5.11 The number of degrees of freedom for looking up the χ^2 table is one less than the number of rows multiplied by one less than the number of columns i.e., (3 – 1) x (2-1) =2.

5.12 At P = 0.01, for 2 df, the tabular value is 9.21. The calculated distinctness χ^2 is more than the tabulated χ^2 value. Therefore, we reject the null hypothesis that variety 1 has a similar reaction to the disease as the candidate variety.

5.13 Similarly the calculated "distinctness χ^2" for variety-2, variety-3 and variety-4 are 33.9, 35.4 and 30.8, respectively, which are all greater than the tabulated χ^2 value of 9.21 at 2 df.

5.14 Hence, the four varieties of common knowledge are significantly different from the candidate variety in reaction to Colletotrichum crown rot.

6. FISHER'S EXACT TEST

Fisher's Exact Test is a statistical test used in the analysis of categorical (qualitative) data where the number of samples (i.e. sample size) is small. Fisher's Exact test applied to 2 x 2 contingency tables is useful where;

- observations on a characteristic are allocated to two or more categories (classes)
- the only source of variation should be caused by random sampling, e.g. there should be no variation due to soil conditions, etc.
- the expected values in each category are less than 10

6.1 Assessment of Distinctness

6.1.1 Fisher's Exact Test is used to determine if there are non-random associations between two categorical variables in a 2 x 2 contingency table[6] and can be used when the sample number for one or more categories for each variety is less than 10 (see bold framed cells in Table 1) or when the table is very unbalanced. Where there is a larger number of samples (i.e. 10 or more), a chi-square test is often preferred.

6.1.2 This test only applies to the analysis of categorical data. The following hypothetical examples illustrate this method:

Example 1

6.1.3 In the following example, the frequency of dark blue flowers is used as a relevant characteristic in the DUS trial. In this example of a DUS trial with two varieties, plants are scored as having dark blue flowers or not having dark blue flowers.

6.1.4 Assume that the two varieties (Variety 1 and Variety 2) have some observed differences in the proportion of dark blue flowers. Examiners need to be able to reliably determine whether these differences can be accepted as clearly distinguishable and Fisher's Exact Test method provides an accepted method to test the hypothesis that the observed differences are statistically significant. Hypothetical data from a total of 24 plants is presented in Table 1.

Table 1: A 2 x 2 Contingency Table - Number of plants with not dark blue and dark blue flowers observed in Variety 1 and Variety 2

	Variety 1	Variety 2	Total
Not dark blue	**4**	**9**	13
Dark blue	**8**	**3**	11
Total	12	12	24

In a 2 x 2 contingency table, the number of degrees of freedom is always 1.

6.1.5 What is the probability that Variety 1 is distinct from Variety 2 on the basis of this characteristic, knowing that 11 of these 24 flowers are dark blue and 8 of these are from Variety 1 and 3 of them are from Variety 2? Or, in other words, is the observed difference in flower color associated with the varietal differences, or is it likely to have arisen through chance sampling? Fisher's method calculates the exact probability of a non-random association, from a 2 x 2 contingency table, using a hypergeometric distribution[7]. In this case, the probability is calculated as the sum of the probabilities for each possible event that is as larger or larger than the observed. Consequently, in addition to the observed, the number of dark blue flowers that would give a successful outcome would be 9,10 or 11 for Variety 1 and 2, 1 or 0 for Variety 2.

6.1.6 Representing the above cells with algebraic notation, the general formula for calculating the probability of the observed numbers is found (Table 2).

Table 2: Algebraic notation for Fisher's Exact Test

	Variety 1	Variety 2	Total

[6] A contingency table is used to record and analyze the relationship between two or more variables, most usually categorical variables.

[7] A hypergeometric distribution is a discrete probability distribution that describes the number of successes in a sequence of n draws from a finite population without replacement.

Not dark blue	a	b	a + b
Dark blue	c	d	c + d
Total	a + c	b + d	n

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!(b+d)!}{n!a!b!c!d!}$$

6.1.7 Where p is the Fisher's Exact probability of finding a non-random distribution between the varieties and the characteristics. (! is the symbol for factorial).

6.1.8 When the algebraic notations in Table 2 are replaced with the observed numbers from Table 1:

$$p = \frac{(13)!\,(11)!\,(12)!(12)!}{24!4!9!8!3!}$$

After solving the factorials:

$$p = 0.05$$

6.1.9 Interpreting the p value calculated by Fisher's Exact Test is straight forward. In the example above, p = 0.05 meaning that there is a 5% chance that, given the sample size and distribution in Table 1, observed differences are due to sampling alone. Given the small sample size, and the need for varieties to be clearly distinguishable from each other, it is open to examination authorities to choose p = 0.01 as the upper cut off significance acceptability level of our null hypothesis. That being so, an examination authority would conclude from this example that the observed difference in the dark blue vs. not dark blue characteristic is not significantly different and the two varieties (Variety 1 and Variety 2) are not distinct on that basis.

Example 2

6.1.10 Observations for Variety 3 and Variety 4 for the same characteristic and observations are given in Table 3:

Table 3: Number of plants with Not dark blue and Dark blue flowers observed in Variety 3 and Variety 4

	Variety 3	Variety 4	Total
Not dark blue	1	9	10
Dark blue	11	3	14
Total	12	12	24

Putting the above values in Fisher's hypergeometric distribution:

$$p = \frac{(10!)\,(14!)(12)!(12)!}{24!1!9!11!3!}$$

After solving the factorials the Fisher's probability value is calculated as:

$$p = 0.001$$

6.1.11 In this particular case, the null hypothesis (that the varieties are similar on the basis of dark blue vs. not dark blue characteristic) is rejected because the calculated Fisher's probability is much lower than the acceptable level of significance (p = 0.01). Accordingly the two varieties (Variety 3 and Variety 4) should be declared as distinct.

7. MATCH APPROACH

7.1 Requirements for application of method

7.1.1 The match method is appropriate for assessing distinctness of varieties where:

- data from more than one year are analyzed,
- observations made on a plant (or plot) in the second year are compared to observations made by the breeder in the first year,
- there are claimed differences between plants (or plots) of a variety based on information from the first year trial,
- the requirements of the method depend on the particular statistical test that is used (e.g. LSD, Multiple Range Tests (MRT), Chi-square or Fisher's Exact).

7.2 Match Method

7.2.1 The Match method to assess distinctness was developed for use where the trials are conducted by the breeder in the first year and examined by the testing authority in the second year (see document TGP/6 section 2.1). Whether differences are sufficiently consistent is assessed using a statistical test (e.g. LSD, MRT, Chi-Square or Fisher's Exact) to gauge whether the differences in the second year are significant and agree with the "direction of the differences" declared by the breeders in the first year. The choice of statistical test depends on the type of expression of the characteristic concerned. For two varieties to be distinct using the Match method, the varieties need to be significantly different in the same direction claimed by the breeder in the first year.

7.2.2 The requirements of the method depend on the particular statistical test that is used (e.g. LSD, MRT, Chi-Square or Fisher's Exact). For quantitative characteristics the statistical test may for example be based on a one-tailed LSD, if there is one candidate, or on a one-tailed MRT, if there is more than one candidate included in the growing trial. A Chi-square test or Fisher's exact test may be used for pseudo-qualitative or qualitative characteristics where the requirements for these tests are met. Although these tests are most useful in trials of cross-pollinated varieties, they can be similarly applied to trials of self-pollinated and vegetatively propagated varieties provided the relevant requirements are met

7.2.3 The Match method typically involves relatively small scale trials where the number of varieties in the trials is limited to the candidate varieties and the most similar varieties of common knowledge.

8. THE METHOD OF UNIFORMITY ASSESSMENT ON THE BASIS OF OFF-TYPES

8.1 Fixed Population Standard

8.1.1 Introduction

Document TGP/10 section 4 provides guidance on when it would be appropriate to use the approach of uniformity assessment on the basis of off-types, using a fixed population standard. It also provides guidance on the determination of crop dependent details such as sample size and the acceptable number of off-types. This section describes the off-type approach from the following perspectives:

- Use of the off-type approach to assess uniformity in a crop.

- The issues to be considered when deciding on the crop dependent details for assessing the uniformity of a crop by the method of off-types. These details include the sample size, the acceptable number of off-types, whether to test in more than one year, and whether to use sequential testing.

8.1.2 Using the approach to assess uniformity in a crop

8.1.2.1 To use the approach to assess uniformity in a crop, the following crop dependent details are either obtained from the UPOV Test Guidelines or decided on the basis of experience, in particular with reference to other UPOV Test Guidelines for comparable types of variety:

- a sample size, e.g. 100 plants
- a maximum number of off-types to be allowed in the sample, e.g. 3
- a fixed population standard, e.g. 1%
- and an acceptance probability, e.g. at least 95%

8.1.2.2 Next, a sample of the correct size of candidate variety plants is taken and the number of off-types counted. If this number is less than or equal to the maximum allowed, the variety is accepted as uniform, otherwise it is rejected as non-uniform. In making these decisions there are two statistical errors that could be made. The risks of making these errors are controlled by the choice of sample size and the maximum allowed number of off-types.

8.1.2.3 The fixed population standard, or “population standard”, is the maximum percentage of off-types that would be permitted if all individuals of the variety could be examined. In the example above it is 1%. Varieties with less than the population standard of off-types are uniform, and those with more than the population standard are non-uniform. However, not all individuals of the variety can be examined, and a sample must be examined instead.

8.1.2.4 Consider a variety which, if all individuals of the variety were examined, would have no more than the population standard of off-types. In taking a sample there are two possible outcomes. Either the sample contains no more than the maximum allowed number of off-types, in which case the variety is accepted as uniform, or the sample contains more than the maximum allowed number of off-types and the variety is rejected. In the latter case a statistical error known as a “Type I error” would have been made. The probability of accepting this variety and the probability making a Type I error are linked as follows:

“probability accept” + “probability make a Type I error” = 100%

8.1.2.5 The chances of accepting or rejecting a variety on the basis of a sample depend on the sample size, the maximum allowed number of off-types, and the percentage of off-types that would be found if all individuals of the variety were examined. The sample size and maximum allowed number of off-types are chosen so as to satisfy the “acceptance probability”, which is the minimum probability of accepting a variety with the population standard of off-types. Thus for the example above, the sample size and maximum number of off-types have been chosen to give at least a 95% chance of accepting a variety which, if all individuals of the variety were examined, would have 1% off-types.

8.1.2.6 To verify the sample size and maximum number of off-types in the example above, the reader should refer to Table A, which lists table 5 and figure 5 as relevant for a population standard of 1% and an acceptance probability of ≥95%. Turning to Table 5, the reader will see that a sample size of 100 (between 83 and 137) and a maximum number of off-types of 3 will give an acceptance probability of >95% for a population standard

of 1%. Figure 5 gives more detail: the lowest of the four traces gives the probability of a Type I error for the different sample sizes and maximum numbers of off-types listed in Table 5. Thus for a population standard of 1%, a sample size of 100, and allowing up to 3 off-types, the probability of a Type I error is 2%, so the probability of accepting on the basis of such a sample a variety with the population standard, i.e. 1%, of off-types is 100% - 2% = 98%, which is greater than the "acceptance probability" (95%) as required.

8.1.2.7 It can be seen from figure 5 that as the sample size increases, the probability of a Type I error increases and the probability of accepting a variety with the population standard, i.e. 1%, of off-types decreases, until this probability becomes too low to satisfy the "acceptance probability", and it becomes necessary to increase the maximum number of off-types in accordance with table 5.

8.1.2.8 Just as a variety with the population standard or fewer off-types can be either accepted or rejected (Type I error) on the basis of a sample, so can a variety with more than the population standard of off-types be either accepted or rejected. To accept on the basis of a sample a variety with more than the population standard of off-types is known as a "Type II error". The probability of a Type II error depends on how non-uniform the variety is. The three upper traces in figure 5 give the probabilities of Type II errors for three degrees of non-uniformity for the different sample sizes and maximum numbers of off-types listed in table 5. The three degrees of non-uniformity are 2, 5 and 10 times the population standard. They are represented by the top, middle and bottom of the three upper traces respectively. Thus for a sample size of 100, and allowing up to 3 off-types, the probability of accepting a variety with 2% off-types is 86%, that of accepting a variety with 5% off-types is 26%, and that of accepting a variety with 10% off-types is 1%. In general:

- The greater the non-uniformity, the smaller the probability of a Type II error.
- For a given maximum number of off-types, as the sample size increases the probability of a Type II error decreases.
- The probability of a Type II error increases as the maximum number of off-types increases.

8.1.3 Issues to be considered when deciding on the use of the method

8.1.3.1 In the preceding section it has been seen that the probability of accepting a variety with the population standard or fewer off-types, or rejecting it (Type I error), and the probability of accepting a variety with more than the population standard of off-types (Type II error) or rejecting it all depend on the choice of sample size and maximum allowed number of off-types. The remainder of this chapter is a discussion of how these choices can be used to balance the risks of Type I and Type II errors. This will be illustrated through a series of examples. The discussion is extended to include the situation where the test is carried out over more than one year, including the possibility of using sequential testing to minimise sampling effort. The reader is provided with tables and figures from which to obtain the Type I and Type II error probabilities for different combinations of population standard and acceptance probability. The reader is also given details of how to calculate the probabilities directly, both for single year tests and for two or more year tests, including two-stage testing.

8.1.3.2 The two types of error described above can be summarized in the following table:

	Decision based on number of off-types in a sample	
Decision that would be made if all plants of a variety could be examined	Variety is accepted as uniform	Variety is rejected as non-uniform
Variety is Uniform	Same decision	Different decision, Type I error
Variety is not uniform	Different decision, Type II error	Same decision

8.1.3.3 The probability of Type II error depends on "how non-uniform" the candidate variety is. If it is much more non-uniform than the population standard then the probability of Type II error will be small and there will be a small probability of accepting such a variety. If, on the other hand, the candidate variety is only slightly more non-uniform than the standard, there is a large probability of Type II error. The probability of acceptance will approach the acceptance probability for a variety with a level of uniformity near to the population standard.

8.1.3.4 Because the probability of Type II error is not fixed but depends on "how non-uniform" the candidate variety is, this probability can be calculated for different degrees of non-uniformity. As mentioned above, this document gives probabilities of Type II error for three degrees of non-uniformity: 2, 5 and 10 times the population standard.

8.1.3.5 In general, the probability of making errors will be decreased by increasing the sample size and increased by decreasing the sample size.

8.1.3.6 For a given sample size, the balance between the probabilities of making Type I and Type II errors may be altered by changing the number of off-types allowed.

8.1.3.7 If the number of off-types allowed is increased, the probability of Type I error is decreased but the probability of Type II error is increased. On the other hand, if the number of off-types allowed is decreased, the probability of Type I errors is increased while the probability of Type II errors is decreased.

8.1.3.8 By allowing a very high number of off-types it will be possible to make the probability of Type I errors very low (or almost zero). However, the probability of making Type II errors will now become (unacceptably) high. If only a very small number of off-types is allowed, the result will be a small probability of Type II errors and an (unacceptably) high probability of Type I errors. The process of balancing the Type I and Type II errors by choice of sample size and number of off-types allowed will now be illustrated by examples.

8.1.4 Examples

Example 1

8.1.4.1 From experience, a reasonable standard for the crop in question is found to be 1%. So the population standard is 1%. Assume that a single test with a maximum of 60 plants is used. From relevant tables (chosen to give a range of target acceptance probabilities), the following schemes can be applied:

Scheme	Sample size	Target acceptance probability*	Maximum number of off-types
a	60	90%	2
b	53	90%	1
c	60	95%	2
d	60	99%	3

8.1.4.2 The following probabilities are obtained for the Type I error and Type II error for different percentages of off-types (denoted by P_2, P_5 and P_{10} for 2, 5 and 10 times the population standard).

Scheme	Sample size	Maximum number of off-types	Probabilities of error (%)			
			Type I	Type II		
				P_2 = 2%	P_5 = 5%	P_{10} = 10%
a	60	2	2	88	42	5
b	53	1	10	71	25	3
c	60	2	2	88	42	5
d	60	3	0.3	97	65	14

* See section 8.1.9.

8.1.4.3 The table lists four different schemes and they should be examined to see if one of them is appropriate to use. (Schemes a and c are identical since there is no scheme for a sample size of 60 with a probability of Type I error between 5 and 10%). If it is decided to ensure that the probability of a Type I error should be very small (scheme d) then the probability of the Type II error becomes very large (97, 65 and 14%) for a variety with 2, 5 and 10% of off-types, respectively. The best balance between the probabilities of making the two types of error seems to be obtained by allowing one off-type in a sample of 53 plants (scheme b).

Example 2

8.1.4.4 In this example, a crop is considered where the population standard is set to 2% and the number of plants available for examination is only 6.

8.1.4.5 Using the relevant tables, the following schemes a-d can be applied:

Scheme	Sample size	Acceptance probability	Maximum number of off-types	Probability of error (%)			
				Type I	Type II		
					P_2 = 4%	P_5 = 10%	P_{10} = 20%
a	6	90	1	0.6	98	89	66
b	5	90	0	10	82	59	33
c	6	95	1	0.6	98	89	66
d	6	99	1	0.6	98	89	66
e	6		0	11	78	53	26

8.1.4.6 Scheme e of the table is found by applying the formulas (1) and (2) shown later in this document.

8.1.4.7 This example illustrates the difficulties encountered when the sample size is very low. The probability of erroneously accepting a non-uniform variety (a Type II error) is large for all the possible situations. Even when all five plants must be uniform for a variety to be accepted (scheme b), the probability of accepting a variety with 20% of off-types is still 33%.

8.1.4.8 It should be noted that a scheme where all six plants must be uniform (scheme e) gives slightly smaller probabilities of Type II errors, but now the probability of the Type I error has increased to 11%.

8.1.4.9 However, scheme e may be considered the best option when only six plants are available in a single test for a crop where the population standard has been set to 2%.

Example 3

8.1.4.10 In this example we reconsider the situation in example 1 but assume that data are available for two years. So the population standard is 1% and the sample size is 120 plants (60 plants in each of two years).

8.1.4.11 The following schemes and probabilities are obtained from relevant tables:

Scheme	Sample size	Acceptance probability	Maximum number of off-types	Probability of error (%)			
				Type I	Type II		
					P_2 = 2%	P_5 = 5%	P_{10} = 10%
a	120	90	3	3	78	15	<0.1
b	110	90	2	10	62	8	<0.1
c	120	95	3	3	78	15	<0.1
d	120	99	4	0.7	91	28	1

8.1.4.12 Here the best balance between the probabilities of making the two types of error is obtained by scheme c, i.e. to accept after two years a total of three off-types among the 120 plants examined.

8.1.4.13 Alternatively a two-stage sequential testing procedure may be set up. Such a procedure can be found for this case by using formulae (3) and (4) later in this document.

8.1.4.14 The following schemes can be obtained:

Scheme	Sample size	Acceptance probability	Largest number for acceptance after year 1	Largest number before reject in year 1	Largest number to accept after 2 years
e	60	90	can never accept	2	3
f	60	95	can never accept	2	3
g	60	99	can never accept	3	4
h	58	90	1	2	2

8.1.4.15 Using the formulas (3), (4) and (5) the following probabilities of errors are obtained:

Scheme	Probability of error (%)				Probability of testing in a second year
	Type I	Type II			
		P_2 = 2%	P_5 = 5%	P_{10} = 10%	
e	4	75	13	0.1	100
f	4	75	13	0.1	100
g	1	90	27	0.5	100
h	10	62	9	0.3	36

8.1.4.16 Schemes e and f both result in a probability of 4% for rejecting a uniform variety (Type I error) and a probability of 13% for accepting a variety with 5% off-types (Type II error). The decision is:

* Never accept the variety after 1 year
* More than 2 off-types in year 1: reject the variety and stop testing
* Between and including 0 and 2 off-types in year 1: do a second year test
* At most 3 off-types after 2 years: accept the variety
* More than 3 off-types after 2 years: reject the variety

8.1.4.17 Alternatively, one of schemes a and h may be chosen. However, scheme g seems to have a too large probability of Type II errors compared with the probability of Type I error. For example, there is a 1% probability of rejecting a uniform variety (Type I error) and a 27% probability of accepting a variety with 5% off-types (Type II error).

8.1.4.18 Scheme h has the advantage of often allowing a final decision to be taken after the first test (year) but, as a consequence, there is a higher probability of a Type I error. In this case, there is a 10% probability of rejecting a uniform variety (Type I error) and a 9% probability of accepting a variety with 5% off-types (Type II error).

Example 4

8.1.4.19 In this example, we assume that the population standard is 3% and that we have 8 plants available in each of two years.

8.1.4.20 From relevant tables, we have:

Scheme	Sample size	Acceptance probability	Maximum number of off-types	Probability of error (%)			
				Type I	Type II		
					P_2 = 6%	P_5 = 15%	P_{10} = 30%
a	16	90	1	8	78	28	3
b	16	95	2	1	93	56	10
c	16	99	3	0.1	99	79	25

8.1.4.21 Here the best balance between the probabilities of making the two types of error is obtained by scheme a.

8.1.4.22 The International Seed Testing Association (ISTA) “seedcalc” method can be used for calculating Type I and Type II errors. “Seedcalc” is available at the following website address: http://www.seedtest.org/en/stats_tool_box_content---1--1143.html.

8.1.5 Introduction to the tables and figures

8.1.5.1 In the TABLES AND FIGURES section (Part II section 8.1.10), there are 7 table and figure pairs corresponding to different combinations of population standard and acceptance probability. These are designs to be applied to a single off-type test. An overview of the tables and the figures are given in table A.

8.1.5.2 Each table shows the maximum numbers of off-types (k) with the corresponding ranges in sample sizes (n) for the given population standard and acceptance probability. For example, in table 1 (population standard 10%, acceptance probability ≥ 95%), for a maximum set at 2 off-types, the corresponding sample size (n) is in the range from 11 to 22. Likewise, if the maximum number of off-types (k) is 10, the corresponding sample size (n) to be used should be in the range 126 to 141.

8.1.5.3 For small sample sizes, the same information is shown graphically in the corresponding figures (figures (1 to 7). These show the actual risk of rejecting a uniform variety and the probability of accepting a variety with a true proportion of off-types 2 times (2P), 5 times (5P) and 10 times (10P) greater than the population standard. (To ease the reading of the figure, lines connect the risks for the individual sample sizes, although the probability can only be calculated for each individual sample size).

Table A. Overview of Table and Figure 1 to 7.

Population standard %	Acceptance probability %	See table and figure no.
10	>95	1
5	>95	2
3	>95	3
2	>95	4
1	>95	5
0.5	>95	6
0.1	>95	7

8.1.5.4 When using the tables the following procedure is suggested:

(a) Choose the relevant population standard.

(b) Choose the decision scheme with the best balance between the probabilities of errors.

8.1.5.5 The use of the tables and figures is illustrated in the example section.

8.1.6 Method for one single test

The mathematical calculations are based on the binomial distribution and it is common to use the following terms:

(a) The percentage of off-types to be accepted in a particular case is called the "population standard" and symbolized by the letter P.

(b) The "acceptance probability" is the probability of accepting a variety with P% of off-types. However, because the number of off-types is discrete, the actual probability of accepting a uniform variety varies with sample size but will always be greater than or equal to the "acceptance probability." The acceptance probability is usually denoted by 100 - α, where α is the percent probability of rejecting a variety with P% of off-types (i.e. Type I error probability). In practice, many varieties will have less than P% off-types and hence the Type I error will in fact be less than α for such varieties.

(c) The number of plants examined in a random sample is called the sample size and denoted by n.

(d) The maximum number of off-types tolerated in a random sample of size n is denoted by k.

(e) The probability of accepting a variety with more than P% off-types, say P_q% of off-types, is denoted by the letter β or by β_q.

(f) The mathematical formulae for calculating the probabilities are:

$$\alpha = 100 - 100 \sum_{i=0}^{k} \binom{n}{i} P^i (1-P)^{n-i} \qquad (1)$$

$$\beta_q = 100 \sum_{i=0}^{k} \binom{n}{i} P_q^i (1-P_q)^{n-i} \qquad (2)$$

P and P_q are expressed here as proportions, i.e. percents divided by 100.

8.1.7 Method for more than one single test (year)

8.1.7.1 Guidance on assessing uniformity by off-types on the basis of more than one growing cycle and on the basis of sub-samples within a single test/trial is provided in document TGP/10 “Assessing Uniformity”.

8.1.8 Note on balancing the Type I and Type II errors

8.1.8.1 We cannot in general obtain Type I-errors that are nice pre-selected values because the number of off-types is discrete. The scheme a of example 2 with 6 plants above showed that we could not obtain an α of 10% - our actual α became 0.6%. Changing the sample size will result in varying α and β values. Figure 3 - as an example - shows that α gets closer to its nominal values at certain sample sizes and that this is also the sample size where β is relatively small.

8.1.8.2 Larger sample sizes are generally beneficial. With same acceptance probability, a larger sample will tend to have proportionally less probability of Type II errors. Small sample sizes result in high probabilities of accepting non-uniform varieties. The sample size should therefore be chosen to give an acceptably low level of Type II errors. However small increases in the sample size may not always be advantageous. For instance, a sample size of five gives α = 10% and β_2 = 82% whereas a sample size of six gives α = 0.6% and β_2 = 98%. It appears that the sample sizes, which give α-values in close agreement with the acceptance probability are the largest in the range of sample sizes with a specified maximum number of off-types. Thus, the largest sample sizes in the range of sample sizes with a given maximum number of off-types should be used.

8.1.9 Definition of statistical terms and symbols

The statistical terms and symbols used have the following definitions:

Population standard. The percentage of off-types to be accepted if all the individuals of a variety could be examined. The population standard is fixed for the crop in question and is based on experience.

Acceptance probability. The probability of accepting a uniform variety with P% of off-types. Here P is population standard. However, note that the actual probability of accepting a uniform variety will always be greater than or equal to the acceptance probability in the heading of the table and figures. The probability of accepting a uniform variety and the probability of a Type I error sum to 100%. For example, if the Type I error probability is 4%, then the probability of accepting a uniform variety is 100 – 4 = 96%, see e.g. figure 1 for n=50). The Type I error is indicated on the graph in the figures by the sawtooth peaks between 0 and the upper limit of Type I error (for instance 10 on figure 1). The decision schemes are defined so that the actual probability of accepting a uniform variety is always greater than or equal to the acceptance probability in the heading of the table.

Type I error: The error of rejecting a uniform variety.

Type II error: The error of accepting a variety that is too non-uniform.

P Population standard

P_q The assumed true percentage of off-types in a non-uniform variety. $P_q = q\,P$.
In the present document q is equal to 2, 5 or 10. These are only 3 examples to help the visualization of Type II errors. The actual percentage of off-types in a variety may take any value. For instance we may examine different varieties which in fact may have respectively 1.6%, 3.8%, 0.2%, … of off-types.

n Sample size
k Maximum number of off-types allowed
α Probability of Type I error
β Probability of Type II error

8.1.10 Tables and figures

Table and figure 1: **Population Standard = 10%**
Acceptance Probability ≥ 95%
n=sample size, k=maximum number of off-types

n			k
1	to	3	1
4	to	8	2
9	to	14	3
15	to	20	4
21	to	27	5
28	to	34	6
35	to	41	7
42	to	48	8
49	to	56	9
57	to	63	10
64	to	71	11
72	to	79	12
80	to	86	13
87	to	94	14
95	to	102	15
103	to	110	16
111	to	119	17
120	to	127	18
128	to	135	19
136	to	143	20
144	to	152	21
153	to	160	22
161	to	168	23
169	to	177	24
178	to	185	25
186	to	194	26
195	to	200	27

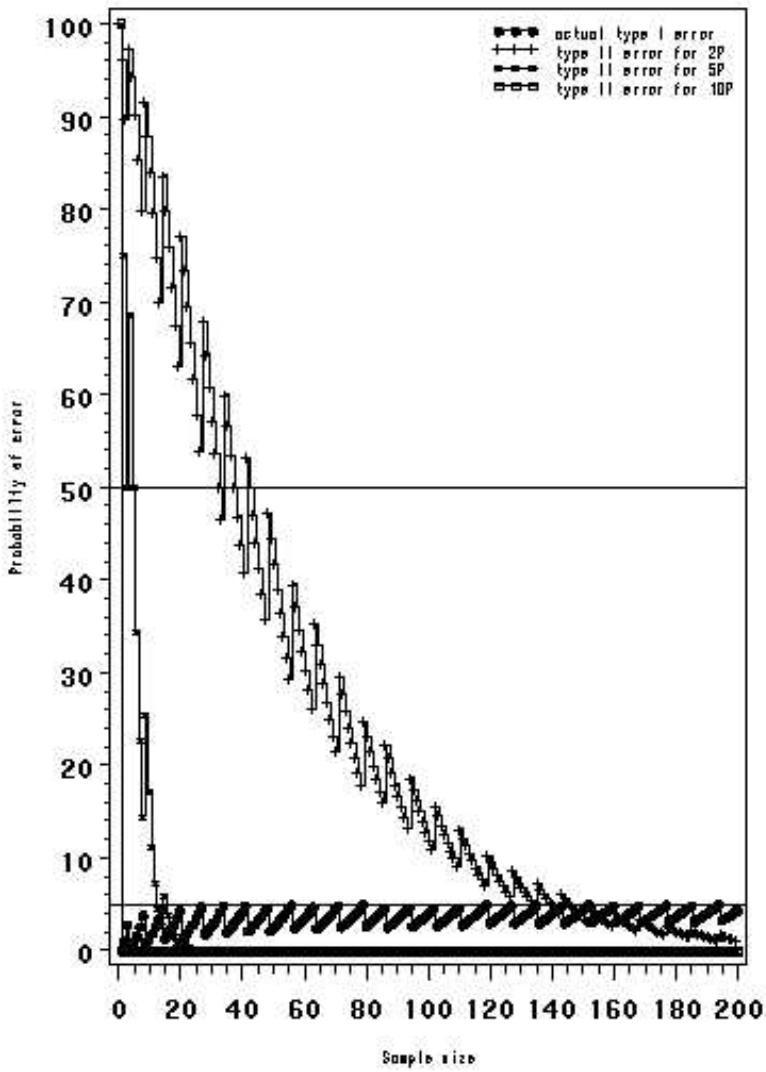


Table and figure 2: **Population Standard = 5%**
Acceptance Probability ≥95%
n=sample size, k=maximum number of off-types

n			k
1	to	1	0
2	to	7	1
8	to	16	2
17	to	28	3
29	to	40	4
41	to	53	5
54	to	67	6
68	to	81	7
82	to	95	8
96	to	110	9
111	to	125	10
126	to	140	11
141	to	155	12
156	to	171	13
172	to	187	14
188	to	203	15
204	to	219	16
220	to	235	17
236	to	251	18
252	to	268	19
269	to	284	20
285	to	300	21
301	to	317	22
318	to	334	23
335	to	351	24
352	to	367	25
368	to	384	26
385	to	401	27
402	to	418	28
419	to	435	29
436	to	452	30
453	to	469	31
470	to	487	32
488	to	504	33
505	to	521	34
522	to	538	35
539	to	556	36
557	to	573	37
574	to	590	38
591	to	608	39
609	to	625	40
626	to	643	41
644	to	660	42
661	to	678	43
679	to	696	44
697	to	713	45
714	to	731	46
732	to	748	47
749	to	766	48
767	to	784	49
785	to	802	50
803	to	819	51
820	to	837	52
838	to	855	53
856	to	873	54
874	to	891	55
892	to	909	56
910	to	926	57
927	to	944	58
945	to	962	59
963	to	980	60
981	to	998	61

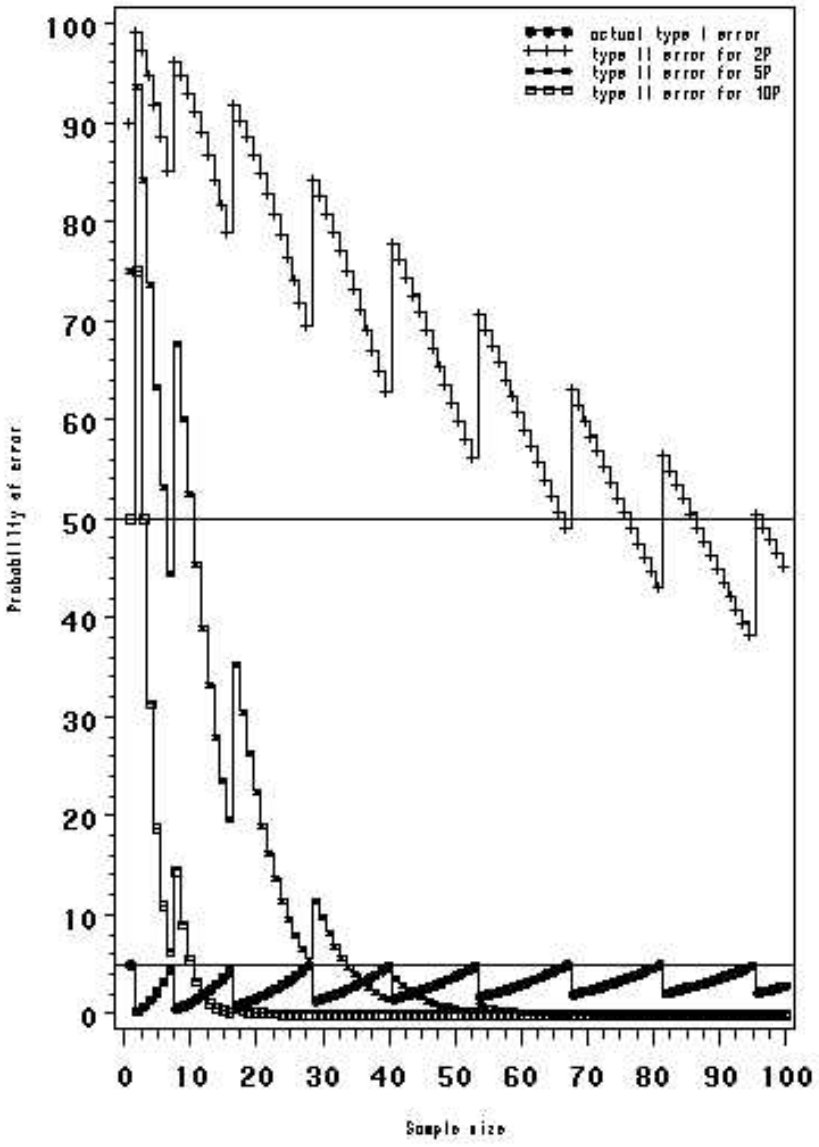


Table and figure 3: **Population Standard = 3%**
Acceptance Probability ≥95%
n=sample size, k=maximum number of off-types

n			k
1	to	1	0
2	to	12	1
13	to	27	2
28	to	46	3
47	to	66	4
67	to	88	5
89	to	110	6
111	to	134	7
135	to	158	8
159	to	182	9
183	to	207	10
208	to	232	11
233	to	258	12
259	to	284	13
285	to	310	14
311	to	337	15
338	to	363	16
364	to	390	17
391	to	417	18
418	to	444	19
445	to	472	20
473	to	499	21
500	to	527	22
528	to	554	23
555	to	582	24
583	to	610	25
611	to	638	26
639	to	666	27
667	to	695	28
696	to	723	29
724	to	751	30
752	to	780	31
781	to	809	32
810	to	837	33
838	to	866	34
867	to	895	35
896	to	924	36
925	to	952	37
953	to	981	38
982	to	1010	39
1011	to	1040	40
1041	to	1069	41
1070	to	1098	42
1099	to	1127	43
1128	to	1156	44
1157	to	1186	45
1187	to	1215	46
1216	to	1244	47
1245	to	1274	48
1275	to	1303	49
1304	to	1333	50
1334	to	1362	51
1363	to	1392	52
1393	to	1422	53
1423	to	1451	54
1452	to	1481	55
1482	to	1511	56
1512	to	1541	57
1542	to	1570	58
1571	to	1600	59
1601	to	1630	60
1631	to	1660	61
1661	to	1690	62
1691	to	1720	63
1721	to	1750	64
1751	to	1780	65
1781	to	1810	66
1811	to	1840	67
1841	to	1870	68
1871	to	1900	69

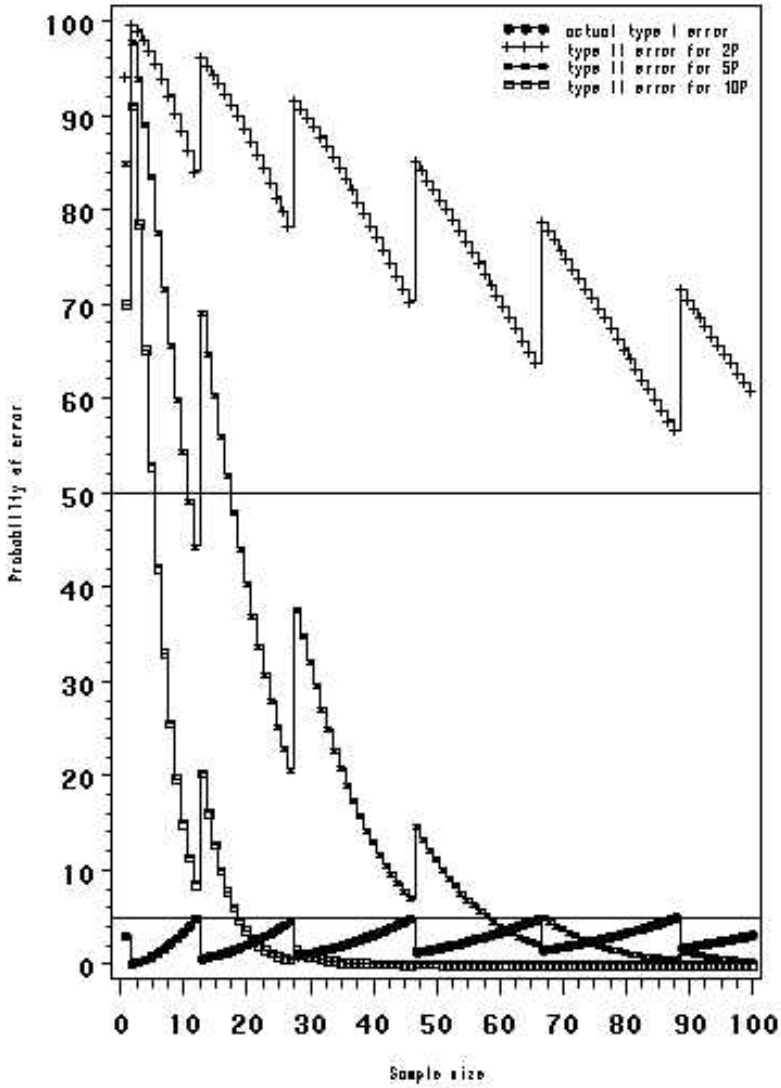


Table and figure 4: Population Standard = 2%
Acceptance Probability ≥95%
n=sample size, k=maximum number of off-types

n			k
1	to	2	0
3	to	18	1
19	to	41	2
42	to	69	3
70	to	99	4
100	to	131	5
132	to	165	6
166	to	200	7
201	to	236	8
237	to	273	9
274	to	310	10
311	to	348	11
349	to	386	12
387	to	425	13
426	to	464	14
465	to	504	15
505	to	544	16
545	to	584	17
585	to	624	18
625	to	665	19
666	to	706	20
707	to	747	21
748	to	789	22
790	to	830	23
831	to	872	24
873	to	914	25
915	to	956	26
957	to	998	27
999	to	1040	28
1041	to	1083	29
1084	to	1126	30
1127	to	1168	31
1169	to	1211	32
1212	to	1254	33
1255	to	1297	34
1298	to	1340	35
1341	to	1383	36
1384	to	1427	37
1428	to	1470	38
1471	to	1514	39
1515	to	1557	40
1558	to	1601	41
1602	to	1645	42
1646	to	1689	43
1690	to	1732	44
1733	to	1776	45
1777	to	1820	46
1821	to	1864	47
1865	to	1909	48
1910	to	1953	49
1954	to	1997	50
1998	to	2000	51

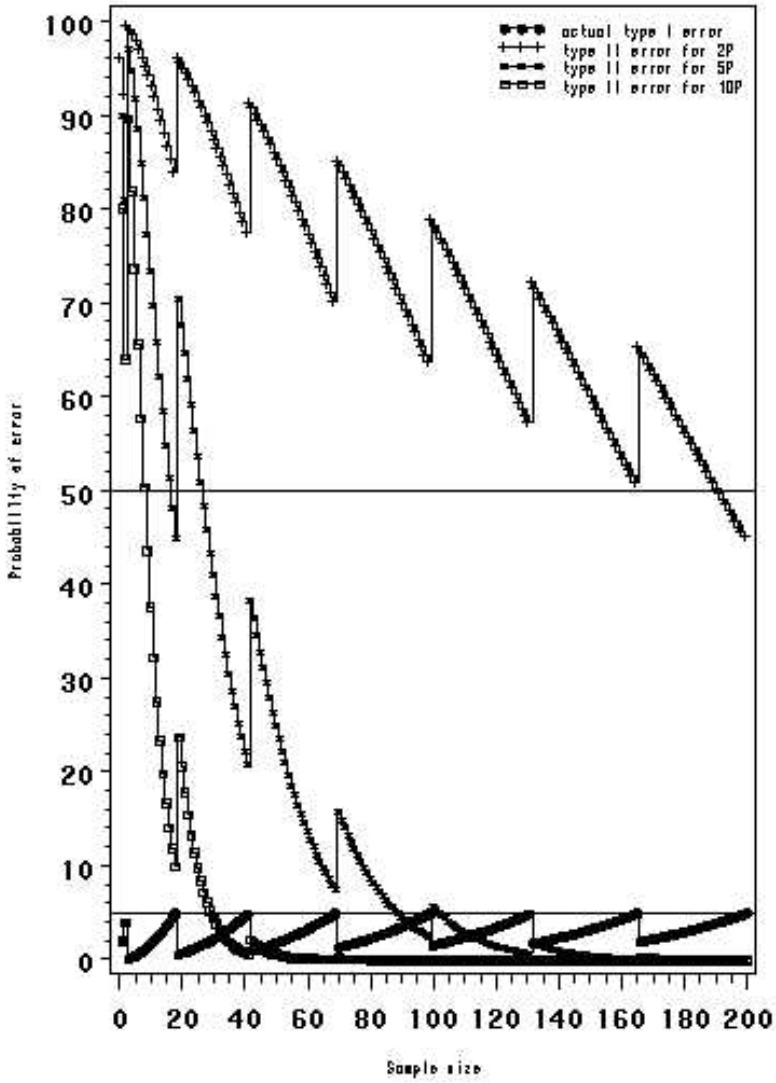


Table and figure 5: Population Standard = 1%
Acceptance Probability ≥95%
n=sample size, k=maximum number of off-types

n			k
1	to	5	0
6	to	35	1
36	to	82	2
83	to	137	3
138	to	198	4
199	to	262	5
263	to	329	6
330	to	399	7
400	to	471	8
472	to	544	9
545	to	618	10
619	to	694	11
695	to	771	12
772	to	848	13
849	to	927	14
928	to	1006	15
1007	to	1085	16
1086	to	1166	17
1167	to	1246	18
1247	to	1328	19
1329	to	1410	20
1411	to	1492	21
1493	to	1575	22
1576	to	1658	23
1659	to	1741	24
1742	to	1825	25
1826	to	1909	26
1910	to	1993	27
1994	to	2078	28
2079	to	2163	29
2164	to	2248	30
2249	to	2333	31
2334	to	2419	32
2420	to	2505	33
2506	to	2591	34
2592	to	2677	35
2678	to	2763	36
2764	to	2850	37
2851	to	2937	38
2938	to	3000	39

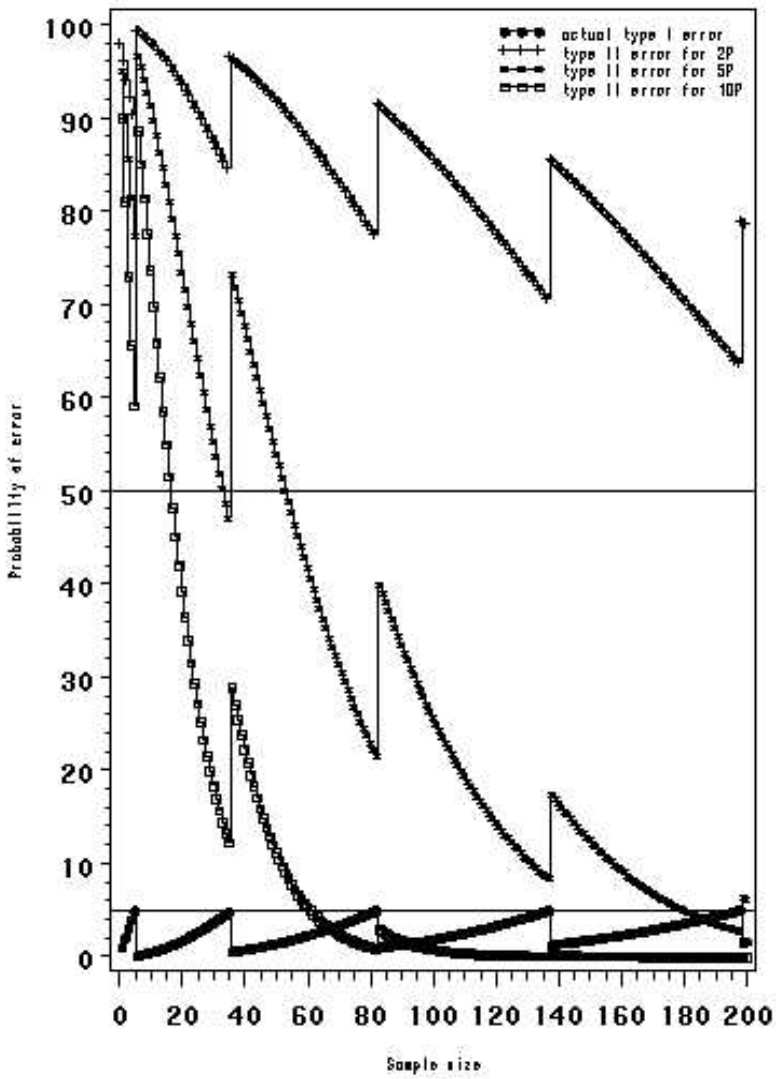


Table and figure 6: Population Standard = .5%
Acceptance Probability ≥95%
n=sample size, k=maximum number of off-types

n			k
1	to	10	0
11	to	71	1
72	to	164	2
165	to	274	3
275	to	395	4
396	to	523	5
524	to	658	6
659	to	797	7
798	to	940	8
941	to	1086	9
1087	to	1235	10
1236	to	1386	11
1387	to	1540	12
1541	to	1695	13
1696	to	1851	14
1852	to	2009	15
2010	to	2169	16
2170	to	2329	17
2330	to	2491	18
2492	to	2653	19
2654	to	2817	20
2818	to	2981	21
2982	to	3000	22

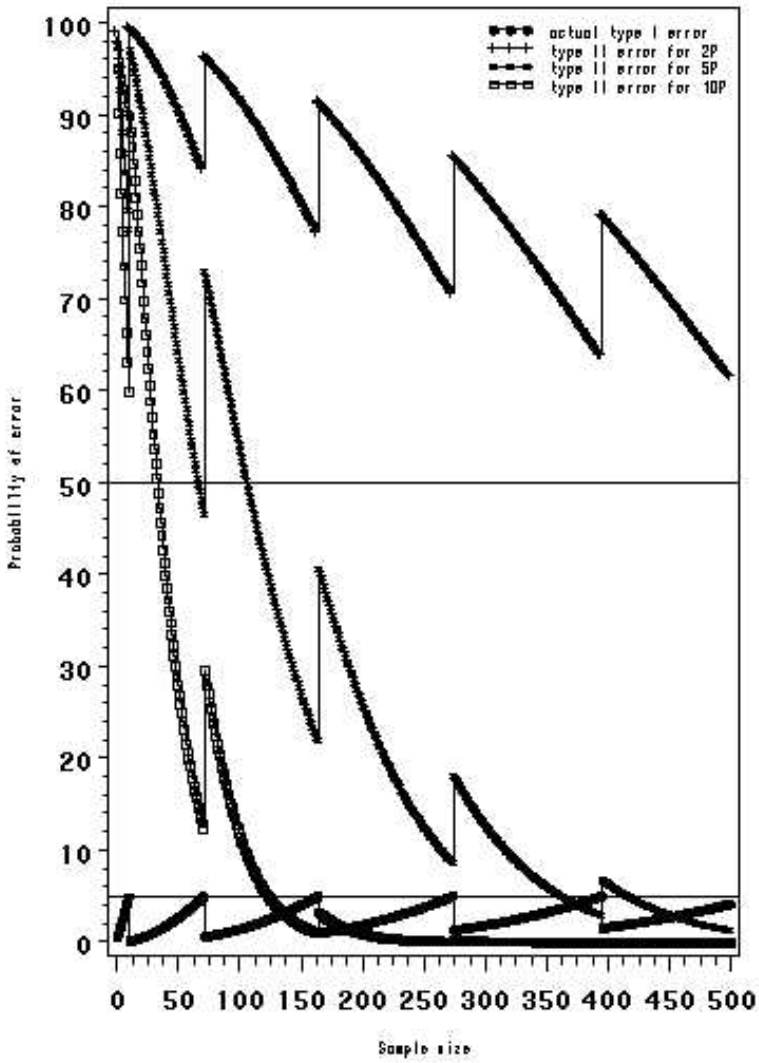
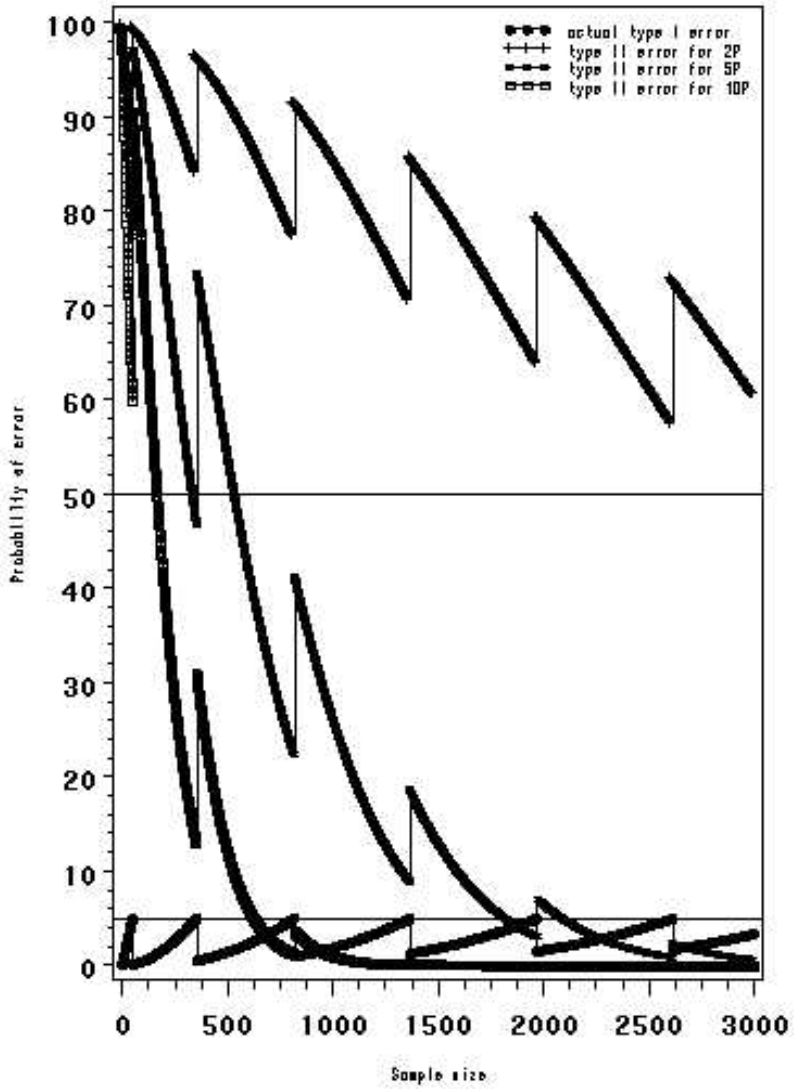


Table and figure 7: Population Standard = .1%
Acceptance Probability ≥95%
n=sample size, k=maximum number off-types

n			k
1	to	51	0
52	to	355	1
356	to	818	2
819	to	1367	3
1368	to	1971	4
1972	to	2614	5
2615	to	3000	6



9. THE COMBINED-OVER-YEARS UNIFORMITY CRITERION (COYU)

9.1 Summary of requirements for application of method

- For quantitative characteristics.
- When observations are made on a plant basis over two or more years.
- When there are some differences between plants of a variety, representing quantitative variation rather than presence of off-types.
- It is recommended that there should be at least 20 degrees of freedom for the estimate of variance for the comparable varieties formed in the COYU analysis.

Comparable varieties are varieties of the same type within the same or a closely related species that have been previously examined and considered to be sufficiently uniform (see document TGP/10, Section 5.2 “Determining acceptable level of variation”).

9.2 Summary

9.2.1 Document TGP/10 explains that when the off-type approach for the assessment of uniformity is not appropriate for the assessment of uniformity, the standard deviation approach can be used. It further states the following with respect to determination of the acceptable level of variation.

> “5.2 Determining the acceptable level of variation
>
> “5.2.1 The comparison between a candidate variety and comparable varieties is carried out on the basis of standard deviations, calculated from individual plant observations. UPOV has proposed several statistical methods for dealing with uniformity in measured quantitative characteristics. One method, which takes into account variations between years, is the Combined Over Years Uniformity (COYU) method. The comparison between a candidate variety and comparable varieties is carried out on the basis of standard deviations, calculated from individual plant observations. This COYU procedure calculates a tolerance limit on the basis of comparable varieties already known i.e. uniformity is assessed using a relative tolerance limit based on varieties within the same trial with comparable expression of characteristics.”

9.2.2 Uniformity is often related to the expression of a characteristic. For example, in some species, varieties with larger plants tend to be less uniform in size than those with smaller plants. If the same standard is applied to all varieties then it is possible that some may have to meet very strict criteria while others face standards that are easy to satisfy. COYU addresses this problem by adjusting for any relationship that exists between uniformity, as measured by the plant-to-plant SD, and the expression of the characteristic, as measured by the variety mean, before setting a standard.

9.2.3 The method involves ranking comparable and candidate varieties by the mean value of the characteristic. Each variety’s SD is taken and the mean SD of the most similar varieties is subtracted. This procedure gives, for each variety, a measure of its uniformity expressed relative to that of similar varieties. The term comparable varieties here refers to established varieties which have been included in the growing trial and which have comparable expression of the characteristics under investigation.

9.2.4 The results for each year are combined in a variety-by-years table of adjusted SDs and analysis of variance is applied. The mean adjusted SD for the candidate is compared with the mean for the comparable varieties using a standard t-test.

9.2.5 COYU, in effect, compares the uniformity of a candidate with that of the comparable varieties most similar in relation to the characteristic being assessed. The main advantages of COYU are that all varieties can be compared on the same basis and that information from several years of testing may be combined into a single criterion.

9.3 Introduction

9.3.1 Uniformity is sometimes assessed by measuring individual characteristics and calculating the standard deviation (SD) of the measurements on individual plants within a plot. The SDs are averaged over all replicates to provide a single measure of uniformity for each variety in a trial.

9.3.2 This section outlines a procedure known as the combined-over-years uniformity (COYU) criterion. COYU assesses the uniformity of a variety relative to comparable varieties based on SDs from trials over several years. A feature of the method is that it takes account of possible relationships between the expression of a characteristic and uniformity.

9.3.3 This section describes:

- The principles underlying the COYU method.
- UPOV recommendations on the application of COYU to individual species.
- Mathematical details of the method with an example of its application.
- The computer software that is available to apply the procedure.

9.4 The COYU Criterion

9.4.1 The application of the COYU criterion involves a number of steps as listed below. These are applied to each characteristic in turn. Details are given under Part II section 9.6.

- Calculation of within-plot SDs for each variety in each year.
- Transformation of SDs by adding 1 and converting to natural logarithms.
- Estimation of the relationship between the SD and mean in each year. The method used is based on moving averages of the log SDs of comparable varieties ordered by their means.
- Adjustments of log SDs of candidate and comparable varieties based on the estimated relationships between SD and mean in each year.
- Averaging of adjusted log SDs over years.
- Calculation of the maximum allowable SD (the uniformity criterion). This uses an estimate of the variability in the uniformity of comparable varieties derived from analysis of variance of the variety-by-year table of adjusted log SDs.
- Comparison of the adjusted log SDs of candidate varieties with the maximum allowable SD.

9.4.2 The advantages of the COYU criterion are:

- It provides a method for assessing uniformity that is largely independent of the varieties that are under test.
- The method combines information from several trials to form a single criterion for uniformity.
- Decisions based on the method are likely to be stable over time.
- The statistical model on which it is based reflects the main sources of variation that influence uniformity.
- Standards are based on the uniformity of comparable varieties.

9.5 Use of COYU

9.5.1 COYU is recommended for use in assessing the uniformity of varieties

- For quantitative characteristics.
- When observations are made on a plant basis over two or more years.

- When there are some differences between plants of a variety, representing quantitative variation rather than presence of off-types.

9.5.2 A variety is considered to be uniform for a characteristic if its mean adjusted log SD does not exceed the uniformity criterion.

9.5.3 The probability level "p" used to determine the uniformity criterion depends on the crop. Recommended probability levels are given in section 9.11.

9.5.4 The uniformity test may be made over two or three years. If the test is normally applied over three years, it is possible to choose to make an early acceptance or rejection of a variety using an appropriate selection of probability values.

9.5.5 It is recommended that there should be at least 20 degrees of freedom for the estimate of variance for the comparable varieties formed in the COYU analysis. This corresponds to 11 comparable varieties for a COYU test based on two years of trials and 8 comparable varieties for three years. In some situations, there may not be enough comparable varieties to give the recommended minimum degrees of freedom. Advice is being developed for such cases.

9.6 Mathematical details

Step 1: Derivation of the within-plot standard deviation

9.6.1 Within-plot standard deviations for each variety in each year are calculated by averaging the plot between-plant standard deviations, SD_j, over replicates:

$$SD_j = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{ij} - \mathbf{y}_j\right)^2}{(n-1)}}$$

$$SD = \frac{\sum_{j=1}^{r} SD_j}{r}$$

where y_{ij} is the observation on the ith plant in the jth plot, $\mathbf{y}_j$ is the mean of the observations from the jth plot, n is the number of plants measured in each plot and r is the number of replicates.

Step 2: Transformation of the SDs

9.6.2 Transformation of SDs by adding 1 and converting to natural logarithms. The purpose of this transformation is to make the SDs more amenable to statistical analysis.

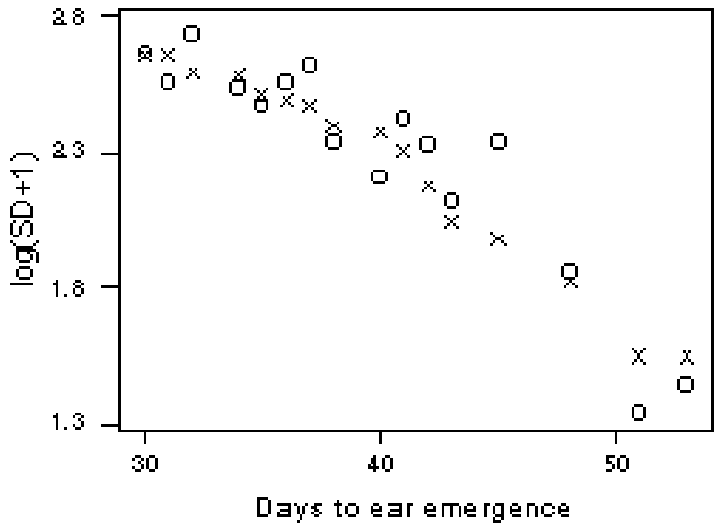
Step 3: Estimation of the relationship between the SD and mean in each year

9.6.3 For each year separately, the form of the average relationship between SD and characteristic mean is estimated for the comparable varieties. The method of estimation is a 9-point moving average. The log SDs (the Y variate) and the means (the X variate) for each variety are first ranked according to the values of the mean. For each point (X_i, Y_i) take the trend value T_i to be the mean of the values Y_{i-4}, Y_{i-3}, ... , Y_{i+4} where i represents the rank of the X value and Y_i is the corresponding Y value. For X values ranked 1st and 2nd the trend value is taken to be the mean of the first three values. In the case of the X value ranked 3rd the mean of the first five values are taken and for the X value ranked 4th the mean of the first seven values are used. A similar procedure operates for the four highest-ranked X values.

9.6.4 A simple example in Figure 1 illustrates this procedure for 16 varieties. The points marked "0" in Figure 1a represent the log SDs and the corresponding means of 16 varieties. The points marked "X" are the 9-point moving-averages, which are calculated by taking, for each variety, the average of the log SDs of the

variety and the four varieties on either side. At the extremities the moving average is based on the mean of 3, 5, or 7 values.

Figure 1: Association between SD and mean – days to ear emergence in cocksfoot varieties (*symbol O is for observed SD, symbol X is for moving average SD*)



Step 4: Adjustment of transformed SD values based on estimated SD-mean relationship

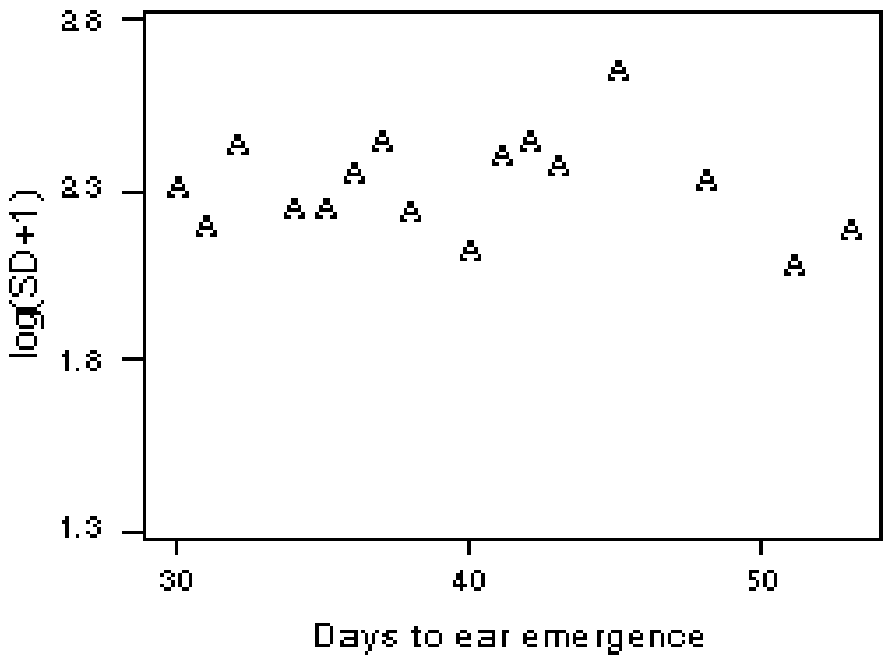
9.6.5 Once the trend values for the comparable varieties have been determined, the trend values for candidates are estimated using linear interpolation between the trend values of the nearest two comparable varieties as defined by their means for the characteristic. Thus if the trend values for the two comparable varieties on either side of the candidate are T_i and T_{i+1} and the observed value for the candidate is X_c, where $X_i \leq X_c \leq X_{i+1}$, then the trend value T_c for the candidate is given by

$$T_c = \frac{(X_C - X_i)T_{i+1} + (X_{i+1} - X_C)T_i}{X_{i+1} - X_i}$$

9.6.6 To adjust the SDs for their relationship with the characteristic mean the estimated trend values are subtracted from the transformed SDs and the grand mean is added back.

9.6.7 The results for the simple example with 16 varieties are illustrated in Figure 2.

Figure 2: Adjusting for association between SD and mean – days to ear emergence in cocksfoot varieties (*symbol A is for adjusted SD*)



Step 5: Calculation of the uniformity criterion

9.6.8 An estimate of the variability in the uniformity of the comparable varieties is derived by applying a one-way analysis of variance to the adjusted log SDs, i.e. with years as the classifying factor. The variability (V) is estimated from the residual term in this analysis of variance.

9.6.9 The maximum allowable standard deviation (the uniformity criterion), based on k years of trials, is

$$UC_p = SD_r + t_p\sqrt{V\left(\frac{1}{k} + \frac{1}{Rk}\right)}$$

where SD_r is the mean of adjusted log SDs for the comparable varieties, V is the variance of the adjusted log SDs after removing year effects, t_p is the one-tailed t-value for probability p with degrees of freedom as for V, k is the number of years and R is the number of comparable varieties.

9.7 Early decisions for a three-year test

9.7.1 Decisions on uniformity may be made after two or three years depending on the crop. If COYU is normally applied over three years, it is possible to make an early acceptance or rejection of a candidate variety using an appropriate selection of probability values.

9.7.2 The probability level for early rejection of a candidate variety after two years should be the same as that for the full three-year test. For example, if the three-year COYU test is applied using a probability level of 0.2%, a candidate variety can be rejected after two years if its uniformity exceeds the COYU criterion with probability level 0.2%.

9.7.3 The probability level for early acceptance of a candidate variety after two years should be larger than that for the full three-year test. As an example, if the three-year COYU test is applied using a probability level of 0.2%, a candidate variety can be accepted after two years if its uniformity does not exceed the COYU criterion with probability level 2%.

9.7.4 Some varieties may fail to be rejected or accepted after two years. In the example set out in section 9.8, a variety might have a uniformity that exceeds the COYU criterion with probability level 2% but not the criterion with probability level 0.2%. In this case, such varieties should be re-assessed after three years.

9.8 Example of COYU calculations

9.8.1 An example of the application of COYU is given here to illustrate the calculations involved. The example consists of days to ear emergence scores for perennial ryegrass over three years for 11 comparable varieties (R1 to R11) and one candidate (C1). The data is tabulated in Table 1.

Table 1: Example data-set – days to ear emergence in perennial ryegrass

	Character Means			Within Plot SD			Log (SD+1)		
Variety	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3	Year 1	Year 2	Year 3
R1	38	41	35	8.5	8.8	9.4	2.25	2.28	2.34
R2	63	68	61	8.1	7.6	6.7	2.21	2.15	2.04
R3	69	71	64	9.9	7.6	5.9	2.39	2.15	1.93
R4	71	75	67	10.2	6.6	6.5	2.42	2.03	2.01
R5	69	78	69	11.2	7.5	5.9	2.50	2.14	1.93
R6	74	77	71	9.8	5.4	7.4	2.38	1.86	2.13
R7	76	79	70	10.7	7.6	4.8	2.46	2.15	1.76
R8	75	80	73	10.9	4.1	5.7	2.48	1.63	1.90
R9	78	81	75	11.6	7.4	9.1	2.53	2.13	2.31
R10	79	80	75	9.4	7.6	8.5	2.34	2.15	2.25
R11	76	85	79	9.2	4.8	7.4	2.32	1.76	2.13
C1	52	56	48	8.2	8.4	8.1	2.22	2.24	2.21

9.8.2 The calculations for adjusting the SDs in year 1 are given in Table 2. The trend value for candidate C1 is obtained by interpolation between values for varieties R1 and R2, since the characteristic mean for C1 (i.e. 52) lies between the means for R1 and R2 (i.e. 38 and 63). That is

$$T_c = \frac{(X_C - X_i)T_{i+1} + (X_{i+1} - X_C)T_i}{X_{i+1} - X_i} = \frac{(52-38)x2.28 + (63-52)x2.28}{63-38} = 2.28$$

Table 2: Example data-set – calculating adjusted log(SD+1) for year 1

Variety	Ranked mean (X)	Log (SD+1) (Y)	Trend Value T	Adj. Log (SD+1)
R1	38	2.25	(2.25 + 2.21 + 2.39)/3 = 2.28	2.25 - 2.28 + 2.39 = 2.36
R2	63	2.21	(2.25 + 2.21 + 2.39)/3 = 2.28	2.21 - 2.28 + 2.39 = 2.32
R3	69	2.39	(2.25 + . . . + 2.42)/5 = 2.35	2.39 - 2.35 + 2.39 = 2.42
R5	69	2.50	(2.25 + . . . + 2.48)/7 = 2.38	2.50 - 2.38 + 2.39 = 2.52
R4	71	2.42	(2.25 + . . . + 2.32)/9 = 2.38	2.42 - 2.38 + 2.39 = 2.43
R6	74	2.38	(2.21 + . . . + 2.53)/9 = 2.41	2.38 - 2.41 + 2.39 = 2.36
R8	75	2.48	(2.39 + . . . + 2.34)/9 = 2.42	2.48 - 2.42 + 2.39 = 2.44
R7	76	2.46	(2.42 + . . . + 2.34)/7 = 2.42	2.46 - 2.42 + 2.39 = 2.43
R11	76	2.32	(2.48 + . . . + 2.34)/5 = 2.43	2.32 - 2.43 + 2.39 = 2.28
R9	78	2.53	(2.32 + 2.53 + 2.34)/3 = 2.40	2.53 - 2.40 + 2.39 = 2.52
R10	79	2.34	(2.32 + 2.53 + 2.34)/3 = 2.40	2.34 - 2.40 + 2.39 = 2.33
Mean	70	2.39		
C1	52	2.22	2.28	2.22 – 2.28 + 2.39 = 2.32

9.8.3 The results of adjusting for all three years are shown in Table 3.

Table 3: Example data-set – adjusted log(SD+1) for all three years with over-year means

	Over-Year Means		Adj. Log (SD+1)		
Variety	Char. mean	Adj. Log (SD+1)	Year 1	Year 2	Year 3
R1	38	2.26	2.36	2.13	2.30
R2	64	2.10	2.32	2.00	2.00
R3	68	2.16	2.42	2.10	1.95
R4	71	2.15	2.43	1.96	2.06
R5	72	2.20	2.52	2.14	1.96
R6	74	2.12	2.36	1.84	2.16
R7	75	2.14	2.43	2.19	1.80
R8	76	2.02	2.44	1.70	1.91
R9	78	2.30	2.52	2.16	2.24
R10	78	2.22	2.33	2.23	2.09
R11	80	2.01	2.28	1.78	1.96
Mean	70	2.15	2.40	2.02	2.04
C1	52	2.19	2.32	2.08	2.17

9.8.4 The analysis of variance table for the adjusted log SDs is given in Table 4 (based on comparable varieties only). The variability in the uniformity of comparable varieties is estimated from this (V=0.0202).

Table 4: Example data set – analysis of variance table for adjusted log (SD+1)

Source	Degrees of freedom	Sums of squares	Mean squares
Year	2	1.0196	0.5098
Varieties within years (=residual)	30	0.6060	**0.0202**
Total	32	1.6256	

9.8.5 The uniformity criterion for a probability level of 0.2% is calculated thus:

$$UC_p = SD_r + t_p\sqrt{V\left(\frac{1}{k}+\frac{1}{Rk}\right)} = 2.15 + 3.118x\sqrt{0.0202x\left(\frac{1}{3}+\frac{1}{3x11}\right)} = 2.42$$

where t_p is taken from Student's t table with p=0.002 (one-tailed) and 30 degrees of freedom.

9.8.6 Varieties with mean adjusted log (SD + 1) less than, or equal to, 2.42 can be regarded as uniform for this characteristic. The candidate variety C1 satisfies this criterion.

9.9 Implementing COYU

The COYU criterion can be applied using COYU module of the DUST software package for the statistical analysis of DUS data. This is available from Dr. Sally Watson, (Email: info@afbini.gov.uk) or from http://www.afbini.gov.uk/dustnt.htm.

9.10 COYU software

9.10.1 DUST computer program

9.10.1.1 The main output from the DUST COYU program is illustrated in Table A1. This summarises the results of analyses of within-plot SDs for 49 perennial ryegrass varieties assessed over a three-year period. Supplementary output is given in Table A2 where details of the analysis of a single characteristic, date of ear emergence, are presented. Note that the analysis of variance table given has an additional source of variation; the variance, V, of the adjusted log SDs is calculated by combining the variation for the variety and residual sources.

9.10.1.2 In Table A1, the adjusted SD for each variety is expressed as a percent of the mean SD for all comparable varieties. A figure of 100 indicates a variety of average uniformity; a variety with a value less than 100 shows good uniformity; a variety with a value much greater than 100 suggests poor uniformity in that characteristic. Lack of uniformity in one characteristic is often supported by evidence of poor uniformity in related characteristics.

9.10.1.3 The symbols "*" and "+" to the right of percentages identify varieties whose SDs exceed the COYU criterion after 3 and 2 years respectively. The symbol ":" indicates that after two years uniformity is not yet acceptable and the variety should be considered for testing for a further year. Note that for this example a probability level of 0.2% is used for the three-year test. For early decisions at two years, probability levels of 2% and 0.2% are used to accept and reject varieties respectively. All of the candidates had acceptable uniformity for the 8 characters using the COYU criterion.

9.10.1.4 The numbers to the right of percentages refer to the number of years that a within-year uniformity criterion is exceeded. This criterion has now been superseded by COYU.

9.10.1.5 The program will operate with a complete set of data or will accept some missing values, e.g. when a variety is not present in a year.

Table A1: Example of summary output from COYU program

**** OVER-YEARS UNIFORMITY ANALYSIS SUMMARY ****

WITHIN-PLOT STANDARD DEVIATIONS AS % MEAN OF REFERENCE VARIETY SDS

	CHARACTERISTIC NUMBER				
	5	60	8	10	11
R1	100	100	95 1	100	97	97
R2	105	106	98	99	104	101
R3	97	103	92 1	103	96	98
R4	102	99	118 2	105	101	101
R5	102	99	116 3	95	104	110
R6	103	102	101	99	97	104
R7	100	95	118 2	102 1	98	99
R8	97	98	84	95	97	93
R9	97	105	87	99	101	99
R10	104	100	96	105 1	96	102
R11	99	96	112	99	101	98
R12	100	97	99 1	103	105	106
R13	95	96	101	100	96	101
R14	105	103	90	97	101	97
R15	102	100 1	89	105	105 1	101
R16	99	98	92 1	98	102	98
R17	97	101	98	101	101	95
R18	99	97	96	96	102	99
R19	103	101	105	102	100	98
R20	104	99	93	91	100	102
R21	97	94	103	97	100	102
R22	101	110*1	112	107 1	103 1	101
R23	94	101	107	99	104	97
R24	99	97	95	99	100	103
R25	104 1	103	93 1	99	101	96
R26	98	97	111 2	96	102 1	106
R27	102	99	106 1	99	103	107
R28	101	106	90	95	101	101
R29	101	105	83	102	94	93
R30	99	96	97	99	95	100
R31	99	102	107	107 1	102	99
R32	98	93	111 2	102	98	103
R33	104	102 1	107 1	103	100	97
R34	95	94	82	95	97	96
R35	100	102	95	100	99	94
R36	99	98	111 1	99	100	103
R37	100	107 1	107	101	100	107
R38	95	97	102	107 1	97	101
R39	99	99	90	98	101	100
R40	104	102	112 1	100	101	97
C1	100 1	106	113 2	104 1	106 1	106
C2	103	101	98	97	101	109
C3	97	93	118 2	98	99	109
C4	102	101	106	103	99	101
C5	100	104	99	103	100	107
C6	101	102	103	100	103	107
C7	96	98	106	97	102	103
C8	101	105 1	116 2	103	103	93
C9	99	99	90 2	91	97	98

CHARACTERISTIC KEY

5	SPRING	60	NATURAL SPRIN
8	DATE OF EAR	10	HEIGHT AT EAR
11	WIDTH AT EAR	14	LENGTH OF FLA
15	WIDTH OF FLAG	24	EAR LENGTH

SYMBOLS

* - SD EXCEEDS OVER-YEARS CRITERION AFTER
\+ - SD EXCEEDS OVER-YEARS CRITERION AFTER
: - SD NOT YET ACCEPTABLE AFTER 2 YEARS W

1,2,3 - THE NUMBER OF OCCASIONS THE WITHIN-YE

Table A2: Example of supplementary DUST output for date of ear emergency (char.8)

**** UNIFORMITY ANALYSIS OF BETWEEN-PLANT STANDARD DEVIATIONS (SD) ****

	OVER-YEARS			INDIVIDUAL YEARS								
VARIETY	CHAR. MEAN	ADJ. LOG SD	UNADJ LOG SD	CHAR. MEAN 88	CHAR. MEAN 89	CHAR. MEAN 90	LOG (SD+1) 88	LOG (SD+1) 89	LOG (SD+1) 90	ADJ LOG(SD+1) 88	ADJ LOG(SD+1) 89	ADJ LOG(SD+1) 90
REFERENCE												
R3	38.47	1.823	2.179	39.07	41.21	35.12	2.02	2.18	2.34X	1.73	1.78	1.96
R5	50.14	2.315	2.671	48.19	53.69	48.54	2.52X	2.74X	2.76X	2.23	2.33	2.39
R16	59.03	1.833	2.179	57.25	63.33	56.50	2.28X	2.24	2.01	1.96	1.73	1.81
R26	63.44	2.206	2.460	61.00	66.53	62.81	2.50X	2.75X	2.13	2.18	2.33	2.11
R9	63.99	1.739	1.994	62.92	68.32	60.72	2.21	2.03	1.74	1.96	1.64	1.62
R12	66.12	1.964	2.086	67.89	65.35	65.12	2.07	2.58X	1.60	1.97	2.14	1.78
R33	67.58	2.124	2.254	66.66	71.54	64.53	2.55X	2.26	1.95	2.32	1.92	2.12
R1	67.87	1.880	1.989	69.07	70.64	63.90	1.60	2.45X	1.93	1.60	2.08	1.96
R20	68.74	1.853	1.893	67.17	74.31	64.74	2.05	1.95	1.68	1.92	1.75	1.89
R25	68.82	1.853	1.905	68.28	72.38	65.81	1.83	2.39X	1.49	1.75	2.09	1.72
R18	69.80	1.899	1.853	68.61	75.22	65.58	1.88	1.84	1.84	1.82	1.80	2.08
R30	70.53	1.919	1.864	70.36	75.08	66.15	2.04	1.84	1.71	2.00	1.78	1.98
R13	70.63	2.005	2.000	70.23	75.00	66.66	1.97	2.03	2.01	1.91	1.86	2.24
R32	71.49	2.197	2.238	70.03	74.98	69.44	2.32X	2.45X	1.94	2.31	2.27	2.01
R34	72.09	1.630	1.545	71.32	77.35	67.59	1.57	1.49	1.58	1.54	1.58	1.78
R40	72.24	2.222	2.178	72.71	75.07	68.95	2.25X	2.26	2.03	2.29	2.16	2.22
R23	72.40	2.122	2.058	69.72	78.39	69.10	2.11	2.14	1.93	2.16	2.14	2.06
R29	72.66	1.657	1.580	73.13	75.80	69.04	1.46	1.63	1.65	1.47	1.69	1.81
R7	73.19	2.341	2.342	72.23	75.80	71.52	2.62X	2.30X	2.10	2.61	2.30	2.11
R24	73.19	1.888	1.796	74.00	76.37	69.20	1.62	1.84	1.93	1.71	1.91	2.04
R19	73.65	2.083	2.049	73.32	76.06	71.57	1.96	2.05	2.14	1.96	2.13	2.16
R2	73.85	1.946	1.897	72.98	78.16	70.42	1.76	1.96	1.97	1.79	2.02	2.03
R31	74.23	2.119	2.012	73.73	78.23	70.71	2.05	1.86	2.13	2.25	1.94	2.17
R37	74.38	2.132	2.020	74.87	76.95	71.32	1.97	2.04	2.04	2.23	2.11	2.06
R11	74.60	2.224	2.150	73.87	78.07	71.87	2.21	2.08	2.16	2.36	2.10	2.21
R38	74.76	2.029	1.916	76.11	78.24	69.93	1.84	2.15	1.75	1.98	2.24	1.87
R8	74.83	1.677	1.593	74.27	78.77	71.45	1.62	1.55	1.61	1.75	1.64	1.64
R15	75.54	1.760	1.682	75.72	78.68	72.22	1.53	1.79	1.73	1.64	1.84	1.80
R10	75.64	1.915	1.847	73.47	79.24	74.23	1.87	1.66	2.00	1.99	1.78	1.98
R22	75.68	2.228	2.133	74.57	79.17	73.32	2.18	2.21	2.01	2.40	2.26	2.03
R14	75.84	1.797	1.688	74.53	79.56	73.43	1.54	1.63	1.90	1.70	1.76	1.93
R17	76.13	1.942	1.832	75.34	79.09	73.96	1.65	2.04	1.81	1.90	2.10	1.83
R39	76.83	1.781	1.676	75.49	80.50	74.50	1.56	1.51	1.96	1.72	1.70	1.92
R35	77.22	1.886	1.773	76.67	80.85	74.15	1.73	1.67	1.92	1.88	1.85	1.93
R4	77.78	2.349	2.268	76.80	81.22	75.33	2.36X	2.13	2.31X	2.52	2.33	2.20
R36	77.98	2.209	2.173	78.97	79.85	75.11	2.13	2.15	2.25X	2.24	2.21	2.18
R6	78.73	2.009	1.935	77.53	82.88	75.78	2.00	1.75	2.06	2.03	2.09	1.91
R27	78.78	2.116	2.098	77.61	80.03	78.69	1.80	2.25	2.24X	1.87	2.39	2.09
R28	79.41	1.785	1.722	78.28	81.99	77.97	1.68	1.43	2.05	1.79	1.67	1.89
R21	80.52	2.045	1.950	77.43	85.02	79.11	1.98	1.75	2.13	2.07	2.09	1.98
CANDIDATE												
C1	64.03	2.252	2.438	63.85	63.33	64.92	2.49X	2.81X	2.02	2.25	2.29	2.21
C2	86.11	1.940	1.837	84.83	88.63	84.85	1.79	1.71	2.01	1.90	2.05	1.87
C3	82.04	2.349	2.248	82.26	87.45	76.40	2.37X	2.03	2.35X	2.48	2.37	2.20
C4	78.63	2.104	2.033	78.01	82.17	75.72	2.05	2.01	2.04	2.15	2.27	1.90
C5	72.99	1.973	1.869	71.98	79.40	67.59	1.95	1.78	1.88	1.93	1.90	2.08
C6	83.29	2.050	1.947	84.10	85.57	80.21	2.05	1.69	2.10	2.16	2.03	1.96
C7	83.90	2.100	1.997	84.12	87.99	79.60	1.93	1.95	2.11	2.04	2.29	1.97
C8	83.50	2.304	2.201	82.43	85.98	82.08	2.27X	2.00	2.34X	2.38	2.33	2.20
C9	51.89	1.788	2.157	52.35	55.77	47.56	1.83	2.34X	2.31X	1.52	1.91	1.93
MEAN OF REFERENCE	71.47	1.988		70.78	74.97	68.65	1.97	2.03	1.96	1.99	1.99	1.99

UNIFORMITY CRITERION

		PROB. LEVEL
3-YEAR REJECTION	2.383	0.002
2-YEAR REJECTION	2.471	0.002
2-YEAR ACCEPTANCE	2.329	0.020

**** ANALYSIS OF VARIANCE OF ADJUSTED LOG(SD+1) *** *

	DF	MS	F RATIO
YEARS	2	0.06239	
VARIETIES	39	0.11440	5.1
RESIDUAL	78	0.02226	
TOTAL	119	0.05313	

SYMBOLS

* - SD EXCEEDS OVER-YEARS UNIFORMITY CRITERION AFTER 3 YEARS.
\+ - SD EXCEEDS OVER-YEARS UNIFORMITY CRITERION AFTER 2 YEARS.
: - SD NOT YET ACCEPTABLE ON OVER-YEARS CRITERION AFTER 2 YEARS.
X - SD EXCEEDS 1.265 TIMES MEAN OF REFERENCE VARIETIES

9.11 Schemes used for the application of COYU

The following four cases are those which, in general, represent the different situations which may arise where COYU is used in DUS testing:

Scheme A: Test is conducted over 2 independent growing cycles and decisions made after 2 growing cycles (a growing cycle could be a year and is further on denoted by cycle)

Scheme B: Test is conducted over 3 independent growing cycles and decisions made after 3 cycles

Scheme C: Test is conducted over 3 independent growing cycles and decisions made after 3 cycles, but a variety may be accepted after 2 cycles

Scheme D: Test is conducted over 3 independent growing cycles and decisions made after 3 cycles, but a variety may be accepted or rejected after 2 cycles

The stages at which the decisions are made in Cases A to D are illustrated in figures 1 to 4 respectively. These also illustrate the various standard probability levels (p_{u2}, p_{nu2} and p_{u3}) which are needed to calculate the COYU criteria depending on the case. These are defined as follows:

Probability Level	**Used to decide whether a variety is :-**
p_{u2}	uniform in a characteristic after 2 cycles
p_{nu2}	non-uniform after 2 cycles
p_{u3}	uniform in a characteristic after 3 cycles

In Figures 1 to 4 the COYU criterion calculated using say the probability level p_{u2} is denoted by UCp_{u2} etc. The term “U” represents the mean adjusted log(SD+1) of a variety for a characteristic.

Table 1 summarizes the various standard probability levels needed to calculate the COYD and COYU criteria in each of Cases A to D. For example, in Case B only one probability level is needed (p_{u3}), whereas Case C requires two (p_{u2} and p_{u3}).

Table 1	COYU		
CASE	p_{u2}	p_{nu2}	p_{u3}
A		■	■
B	■	■	
C		■	
D			

Figure 1. COYU decisions and standard probability levels (p_i) in Case A

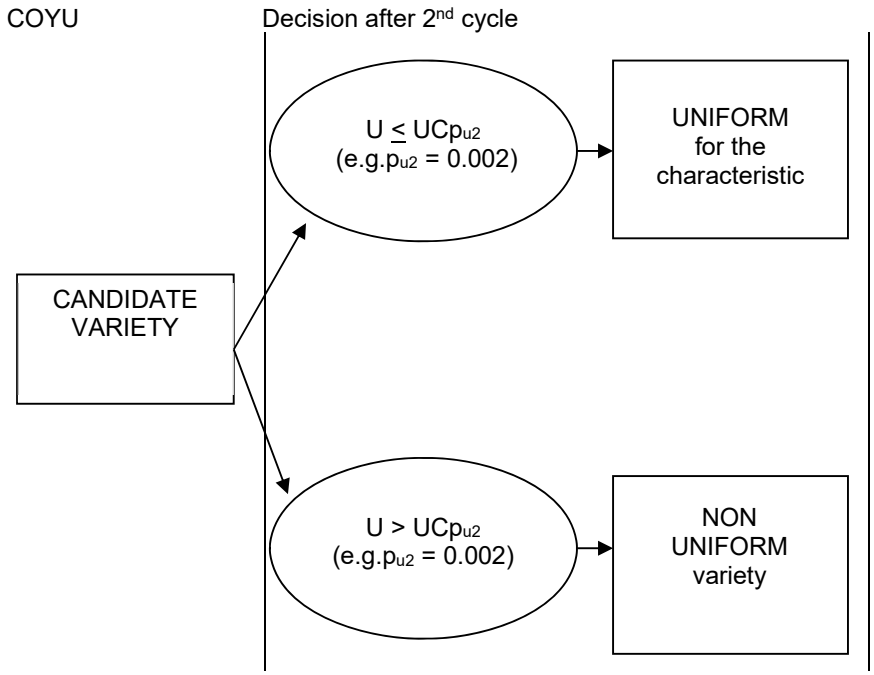
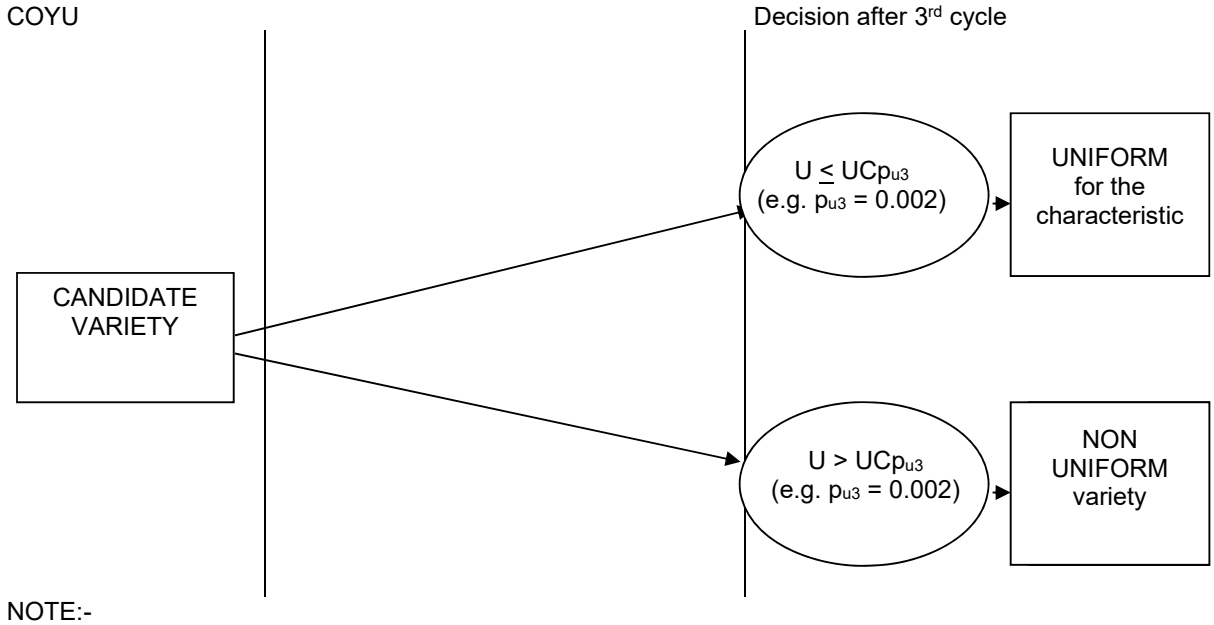


Figure 2. COYD and COYU decisions and standard probability levels (p_i) in Case B



NOTE:-
"U" is the mean adjusted log(SD+1) of the candidate variety for the characteristic.
UCp is the COYU criterion calculated at probability level p.

Figure 3. COYU decisions and standard probability levels (p_i) in Case C

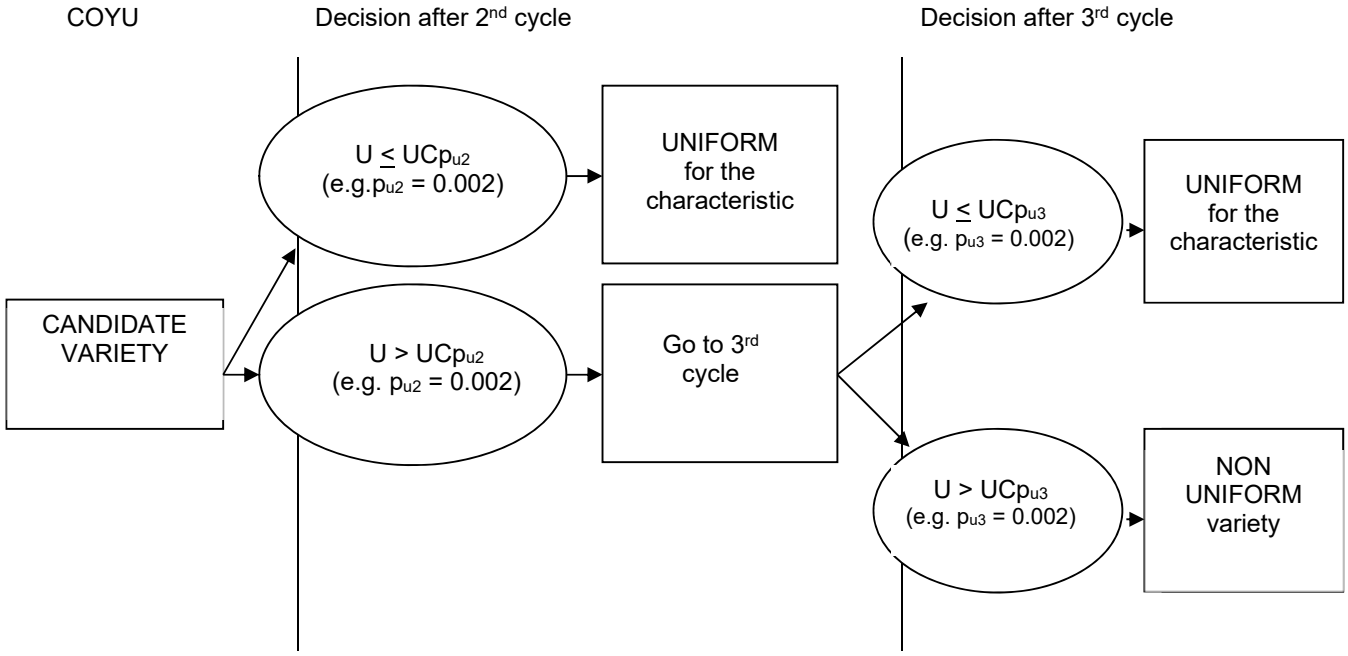
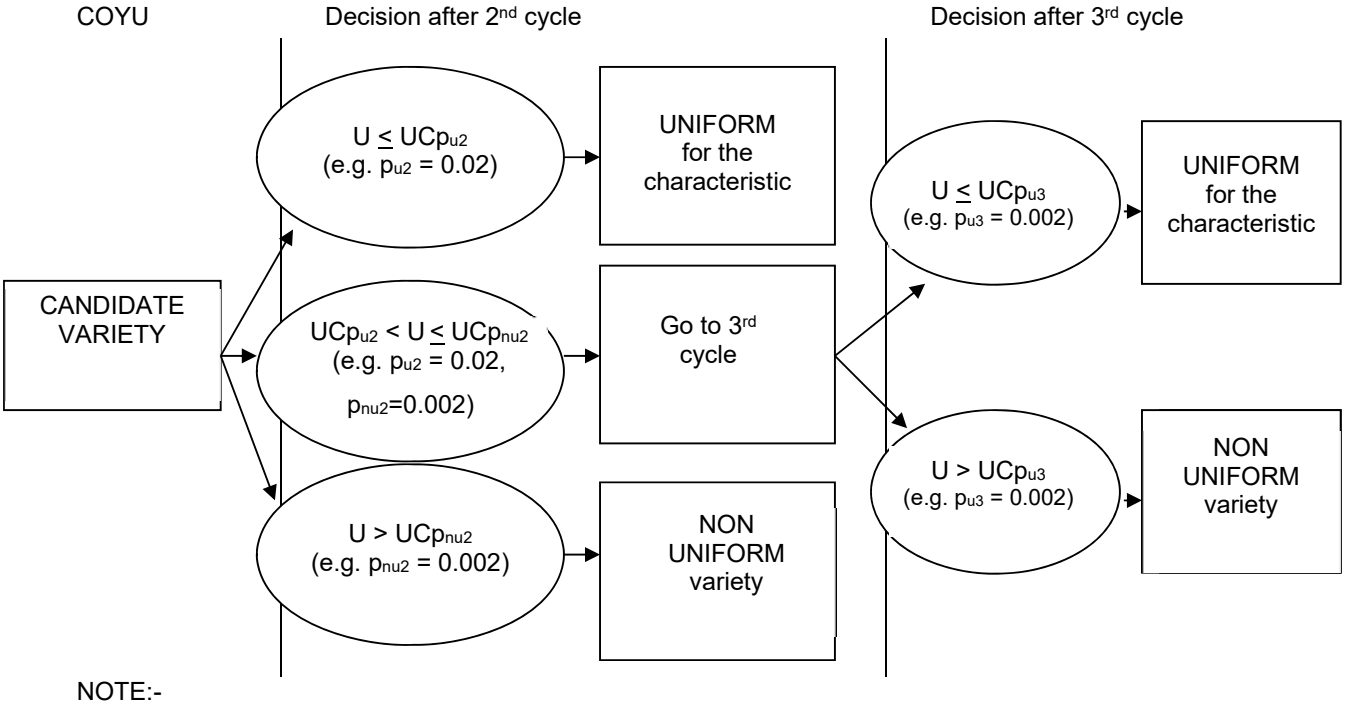


Figure 4. COYD and COYU decisions and standard probability levels (p_i) in Case D



NOTE:-
"U" is the mean adjusted log(SD+1) of the candidate variety for the characteristic
UCp is the COYU criterion calculated at probability level p

10. UNIFORMITY ASSESSMENT ON THE BASIS OF THE RELATIVE VARIANCE METHOD

10.1 Use of the relative variance method

The relative variance for a particular characteristic refers to the variance of the candidate divided by the average of the variance of the comparable varieties (i.e. Relative variance = variance of the candidate/average variance of the comparable varieties). The data should be normally distributed. The relative variance method may be applied to any measured characteristic that is a continuous variable, irrespective of the method of propagation of the variety. Comparable varieties are varieties of the same type within the same or a closely related species that have been previously examined and considered to be sufficiently uniform (see document TGP/10, Section 5.2 "Determining acceptable level of variation").

In cross-pollinated varieties, a common recommendation in the UPOV Test Guidelines is to take 60 measurements per characteristic per variety. In essence, the variance ratio equates to the F statistic, and the tabulated value of F at P = 0.01 under df_1 =60 (degrees of freedom of candidate) and $df_2 = \infty$ (degrees of freedom of comparable variety(ies)) is 1.47. $df_2 = \infty$ is chosen as a conservative estimate, as it is assumed that comparable varieties accurately represent the infinite number of possible comparable varieties for the species as a whole. Therefore, 1.47 is the threshold for cross-pollinated species with 60 measurements per characteristics per variety. For different sample sizes, a different F statistic should be used for the df_1, although the df_2 should remain at ∞.

10.2 Thresholds for different sample sizes

10.2.1 Different thresholds of F (at P = 0.01) should be applied for different sample sizes of the candidate variety. The df_1 will vary according to different sample sizes of the candidate variety. However, in all cases the df_2 will be considered to be ∞, to cover the whole range of possible comparable varieties within a species - thus providing a conservative estimate of the threshold. Under these conditions and taking the relevant values from the F table, Table 1 shows the thresholds that would apply for different sample sizes of the candidate varieties. In the case of different sample sizes than those included in Table 1, the correct threshold should be used for the exact sample size.

Table 1: Thresholds for relative variance for some different sample sizes

Sample size of candidate	Thresholds for relative variance
30	1.70
40	1.59
50	1.53
60	1.47
80	1.41
100	1.36
150	1.29
200	1.25

Source: Table of F published in 'Tables for Statisticians' Barnes &Noble, Inc. New York

10.2.2 For a given sample size, if the relative variance exceeds the threshold, the candidate variety will be deemed to be non-uniform for that characteristic.

10.3 The relative variance test in practice

10.3.1 When the calculated relative variance is lower than the tabulated value of F statistic presented in Table 1, for the relevant sample size, then it is reasonable to assume that the variances are equal and the candidate variety is uniform in that particular characteristic. If the calculated relative variance is higher than the tabulated value of F, then the null hypothesis, that the varieties have equal variances, is rejected. The candidate variety would then be deemed to have a higher variance than the comparable varieties for that particular characteristic and, therefore, would not meet the uniformity criteria.

10.4 Example of relative variance method

Example

10.4.1 In a DUS trial, a cross-pollinated candidate variety is grown together with a number of varieties representing the required level of uniformity for all relevant characteristics. In order to illustrate the calculation of the relative variance, an example with 4 comparable varieties is given. The variance data on plant height measurements for the five varieties are presented in Table 2. For each variety, 60 plants were measured for plant height:

Table 2: variances of candidate and comparable varieties for plant height data

Candidate	Comparable variety 1	Comparable variety 2	Comparable variety 3	Comparable variety 4
5.6	7.8	4.5	3.2	5.8

10.4.2 The number of observations per variety is the same (n=60); therefore, we can take the average variance of the comparable varieties as their pooled variance.

10.4.3 The average variance for comparable varieties is (7.8 + 4.5 + 3.2 + 5.8)/4 = 5.32

10.4.4 The relative variance for a particular characteristic refers to the variance of the candidate divided by the average of the variance of the comparable varieties.

Relative variance = variance of the candidate/average variance of the comparable varieties

= 5.6/5.32 = 1.05

10.4.5 Now, in Table 1, for a sample size of 60, the threshold is 1.47; therefore, we can conclude that the candidate variety is sufficiently uniform for that characteristic.

10.5 Relationship between relative variance and relative standard deviation

10.5.1 Sometimes in DUS trials, the uniformity data is presented in terms of standard deviations, not as variances. Mathematically there is a simple relationship between variance and standard deviation, as follows:

Standard deviation = square root of Variance

10.5.2 Therefore, when dealing with relative standard deviations, Table 1 needs to be modified to include the square roots of the threshold, which is presented in Table 4.

Table 4: Thresholds for relative standard deviations for some different sample sizes

Sample size of candidate	**Thresholds for relative standard deviations**
30	1.30
40	1.26
50	1.24
60	1.21
80	1.19
100	1.17
150	1.14
200	1.12

10.5.3 When making a decision on uniformity based on relative standard deviations, the examiner needs to use Table 4, instead of Table 1, to get the appropriate threshold. The same principle for acceptance or rejection applies for relative standard deviation; only the thresholds are lower due to the square root of appropriate values. For example, for 60 samples the relative variance threshold is 1.47; however, for relative standard deviation the threshold is 1.21, which is the square root of 1.47.

11. EXAMINING CHARACTERISTICS USING IMAGE ANALYSIS

11.1 Introduction

Characteristics which may be examined by image analysis should also be able to be examined by visual observation and/or manual measurement, as appropriate. Explanations for observing such characteristics, including where appropriate explanations in Test Guidelines, should ensure that the characteristic is explained in terms which would enable the characteristic to be understood and examined by all DUS experts.

11.2 Combined characteristics

11.2.1 The General Introduction (document TG/1/3, Chapter 4, Section 4) states that:

> "4.6.3 Combined Characteristics
>
> "4.6.3.1 A combined characteristic is a simple combination of a small number of characteristics. Provided the combination is biologically meaningful, characteristics that are assessed separately may subsequently be combined, for example the ratio of length to width, to produce such a combined characteristic. Combined characteristics must be examined for distinctness, uniformity and stability to the same extent as other characteristics. In some cases, these combined characteristics are examined by means of techniques, such as Image Analysis. In these cases, the methods for appropriate examination of DUS are specified in document TGP/12, 'Special Characteristics'."

11.2.2 Thus, the General Introduction clarifies that the use of image analysis is one possible method for examining characteristics which fulfill the basic requirements for use in DUS testing (see document TG/1/3, Chapter 4.2), which includes the need for the uniformity and stability of such characteristics to be examined. With regard to combined characteristics, the General Introduction also explains that such characteristics should be biologically meaningful.

11.2.3 Image analysis is the extraction of information (e.g. plant measurements) from (digital) images by means of a computer. Image analysis is used in plant variety testing to help in the assessment of plant characteristics. It can be regarded as an intelligent measurement device (advanced ruler). This document aims to give guidance when using image analysis in plant variety testing.

11.2.4 Image analysis can be used in a fully automated or semi-automated way. When fully automated, the expert just records images of plant parts with a camera or scanner and the computer automatically calculates relevant characteristics without human interference. In a semi-automated way, the computer shows the images on a screen and a user can interact with the software to measure specific plant parts, e.g. by clicking with a mouse.

11.3 Image recording: calibration and standardization

11.3.1 An important aspect to consider when recording and analyzing digital images is standardization and calibration in cases where image analysis is automated. Standardization is done by using as much as possible the same setup (illumination, camera, camera-settings, lens, perspective, and object-camera distance) for every recording. It is important to check that the recordings are done according to a prescribed protocol, as the software may depend on it. For example, pods may have to be orientated horizontally in the images, with the beaks pointing to the left. Calibration of the system is needed to make the recording as much as possible independent of any varying conditions by correcting for the variations, e.g. in size or color.

11.3.2 Size calibration is necessary. Since the measure unit in pictures is the pixel, a relation needs to be established between the pixels on the image and millimeters. A standard way to perform this calibration is to include a ruler in every recorded image, at the same distance from the camera as the plant part being recorded. In that case the user can relate the size of the ruler to the number of pixels, and make the calibration manually. A preferred way is to use an object of standard dimensions, e.g. a coin, which can automatically be analyzed with the software and then used for an implicit size calibration. A coin also allows checking if pixels are square (i.e. if the aspect ratio of every pixel is 1:1). In all cases, the object should be sufficiently close to the calibration object and sufficiently far from the camera, to minimize the effect of varying magnification with distance. Alternatively a telecentric lens could be used to minimize this effect.

11.3.3 Illumination calibration is also necessary: an object has to be segmented from the background in the image. An often used and very simple way to do this, is to use thresholding: a pixel with a (grey) value above a certain threshold is considered an object pixel and below the threshold a background pixel (or vice versa). If the illumination is not constant, it may occur that the segmentation is not optimal for every image and that part of the pixels are assigned to the wrong class (object/background), even if the threshold value is determined automatically. This may result in erroneous measurements. It is therefore advisable to check the segmentation results by having a quick look at the segmented binary images.

11.3.4 In many situations only a silhouette/contour of the plant material is necessary, e.g. for size and shape. In these cases it is often advisable to use a background illumination, e.g. a light box. This will increase the contrast between the background and the object, and make the segmentation result much less dependent on the threshold value.

11.3.5 It should be ensured that the lighting is homogenously distributed over the image. Darker parts in the image may result in a wrong segmentation and hence lead to incorrect and incomparable measures, especially when multiple objects are recorded in the same image.

11.3.6 For colors and (variegation or blush) patterns on the plant part, it is essential that the illumination is done correctly and checked regularly, preferably for every image. In that case illumination calibration can be done by recording (part of) a standard color chart in the image. Special algorithms are available to correct for color changes due to differing illumination conditions, but in many situations this correction causes some loss of precision.

11.3.7 The light source is of large influence on the observed color in the image. Especially for color, the type of light source is important. In many cases, lamp color and intensity change during warming up of the lamps which should consequently sufficiently be warmed up before starting the recordings. If fluorescent tubes are used, it should regularly be verified that they have more or less the same intensity/color, as they may change rather rapidly with age. Calibration charts can be used to this purpose.

11.3.8 Especially when recording shiny objects like apples or certain flowers, specular reflection needs to be taken into account. Objects with specular spots cannot be measured reliably. In such cases, attention should be paid to uniform and indirect illumination, using special light tents.

11.3.9 Both (color) cameras and scanners can be used for image recording. The choice is dependent on the application and the preference of the user. Other more advanced systems, such as 3D cameras or hyperspectral cameras are not yet used in standard plant variety testing.

11.3.10 In general image analysis is used to automate the measurement of characteristics described in the guidelines of UPOV. In that case the aim is to replace a hand measurement by a computer measurement. This requires an additional calibration in addition to the image recording calibration. The measurements can then be checked with manual measurements for consistency, e.g. by a scatterplot of hand versus computer measurement with a regression line and the line y=x.

11.3.11 In some cases, image analysis requires a more precise and mathematical definition of the characteristic than is required for human experts. E.g. the length of the pod can be redefined as the length of the medial axis of the pod, excluding the stem. In such cases, there is a special need to check for differences in behavior for different genotypes (bias). The measurement for some genotypes may be exactly the same, whereas for others a systematic difference may be present. A nice example is for determining the bulb height in onions (van der Heijden, Vossepoel and Polder, 1996), where the top of the bulb was defined as the bending point of the shoulder. As long as such a change or refinement of the definition of a characteristic is known and accounted for, this is not a problem. In general, it is advisable to consult the crop experts for redefining a characteristic and check if a minor modification of the guideline might be necessary.

11.3.12 In some cases the object consists of different parts which have to be measured separately, e.g. the pod, beak and stem of a pod of French bean. This requires a special algorithm to separate the different parts (distinguish stem and beak from the pod) and this has to be tested extensively on a large number of genotypes in the reference collection, to be sure that the implementation is robust over the entire range of expressions.

11.3.13 Shape characteristics can also be measured with image analysis, but in general it will be restricted to characteristics already in the guideline, e.g. by defining the shape as the ratio between length and width.

11.3.14 Although color is a standard UPOV characteristic, and could be measured by image analysis, it is not used often. In most cases, crop experts still rely on visual observation with RHS colour charts.

11.4 Conclusions

11.4.1 Image analysis is used for measurements and to automate, at least partially, the assessment of characteristics. It requires a good and precise definition of the characteristic, computerization using existing or in-house software, a good preparation of samples, checking with existing procedures, careful calibration and standardization. It often necessitates therefore an investment which can only be profitable versus hand assessment of characteristics if it concerns a significant number of measurements or measurements which are difficult and time consuming to assess by the examiner. In case of organs of a small size, seed size for example, image analysis will be more precise and more reliable.

11.4.2 Image analysis offers the possibility to store information: images can be recorded and analyzed at a later stage in order to avoid peaks of work and they can be retrieved at a later stage to compare varieties for example in case of doubt.

11.4.3 Today it is mainly used for size and shape features but with the development of techniques, it will be possible to use it for a wider range of standard UPOV characteristics in future.

11.5 References

Van der Heijden, G., A. M. Vossepoel & G. Polder (1996) Measuring onion cultivars with image analysis using inflection points. *Euphytica*, 87, 19-31.

12. EXAMINING CHARACTERISTICS ON THE BASIS OF BULK SAMPLES

The following criteria should be observed when examining characteristics on the basis of bulk samples:

(a) the characteristic should fulfill the requirements of a characteristic, as set out in the "General Introduction to the Examination of Distinctness, Uniformity and Stability and the Development of Harmonized Descriptions of new Varieties of Plants" (see document TG/1/3, Section 4.2.1);

(b) there should be knowledge of the genetic control of the characteristic;

(c) the suitability of the characteristic should be validated through an initial assessment of uniformity on individual plants;

(d) information on plant-by-plant variation and differences between growing cycles should be provided (data from routine measurement of the characteristic from different years);

(e) a full description of the method of assessment should be provided;

(f) states of expression should be based on existing variation between varieties considering environmental influence.

[End of document]