



**TWC/21/2**

**ORIGINAL:** English

**DATE:** May 15, 2003

**INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS**  
GENEVA

**TECHNICAL WORKING PARTY  
ON  
AUTOMATION AND COMPUTER PROGRAMS**

**Twenty-First Session  
Tjele, Denmark, June 10 to 13, 2003**

USE OF THE CHI SQUARE TEST IN THE ASSESSMENT OF DISTINCTNESS

*Document prepared by experts from France and the United Kingdom*

## USE OF THE CHI SQUARE TEST IN THE ASSESSMENT OF DISTINCTNESS

Authors: Vincent Gensollen (FR), Sylvain Grégoire (FR), Sally Watson (UK)

1. Within a variety, variability can occur due to phenotypic and/or environmental variation. This is the case in particular for cross-pollinated varieties, including synthetic varieties. For such varieties, the expression of a characteristic should be recorded using more than one observation. In general, records are taken from a number of individual plants (see “TGP/9”).
2. For quantitative characteristics, the statistical method recommended by UPOV for the assessment of distinctness is COYD<sup>1</sup> analysis, which takes into account variation between years (for details see TGP/9.7). UPOV has proposed several statistical methods to deal with uniformity in quantitative characteristics. One method, which takes into account variation between years is the COYU<sup>2</sup> method (for details see TGP/10.3.1).
3. For qualitative characteristics and pseudo-qualitative characteristics, two cases are possible (see Annex 1 for examples).
4. In the first case, all the plants of a given variety express the same state for a characteristic, i.e. they are homogeneous in their expression of the characteristic. For example all plants of a variety express the same state of the characteristic “Sex of plant” (e.g. dioecious female (1), dioecious male (2), monoecious unisexual (3), or monoecious hermaphrodite (4)). The ploidy level is another example of this kind of characteristic (e.g. diploid (2), tetraploid (4), or hexaploid (6)).
5. In this first case, the characteristic could be used as a grouping characteristic. In cross-pollinated varieties, including synthetic varieties, these characteristics are rather rare but very useful; UPOV recommends the assessment of uniformity by the off-type procedure in order to prevent breeders from creating varieties which are heterogeneous in these characteristics. Distinctness between two varieties can be established when they express different states in such a characteristic.
6. In the second case, plants of a given variety can have different states of expression for a characteristic, i.e. they are heterogeneous in their expression of the characteristic. These characteristics are important for distinctness purposes because the frequency with which plants expressing the different states occur in a variety can be very consistent and so helpful in determining distinctness. For example, in Lucerne the frequency of occurrence of plants with the different states of the “flower colour” characteristic ( white or yellow (1), violet (2), very dark violet (3), variegated (4)) is used to show distinctness between varieties.
7. In this second case, the Chi square test can be used to assess distinctness (SNEDECOR, G.W.; COCHRAN W. (1937)). The chi square test compares the frequencies with which plants expressing the different states of the characteristic occur in different varieties (See Annex 2). Such characteristics cannot be assessed for uniformity.

---

<sup>1</sup> COYD : Combined Over Years Distinctness

<sup>2</sup> COYU : Combined Over Years Uniformity

8. To establish distinctness in pair-wise comparisons, the differences between varieties must be significant in either 2 out of 2 or 2 out of 3 successive cycles of examination. Significance is assessed by using the Chi square test at a UPOV recommended significance probability for the crop. The differences must also be of the same sign in the 2 cycles that have significant differences, e.g. variety A must have consistently more plants with dark variegated flowers than variety B.

9. Within the second case, there is a particular situation for qualitative characteristics where plants within a variety are heterogeneous in their expression. This occurs when there are just two states (e.g. presence or absence). Examples of such characteristics include “Tendency to form inflorescences in year of sowing” or “Resistance to disease”. When recorded in the trial, these characteristics are handled as qualitative characteristics where plants within a variety express different states. However, statistically these characteristics can be analysed for distinctness either by using the Chi square test, as described above; or, alternatively, they may be quantitatively summarized for the variety as a percentage presence. In the latter approach, the COYD criterion is applied to the variety-by-years table of the percentage presence of the characteristic (for details see TGP 9.7). Such characteristics cannot be assessed for uniformity.

*Remark: When data are transformed into states of expression for the purpose of variety description, characteristics might appear to be quantitative. However, the statistical analysis is not made at this level of process. For example in the characteristic “Tendency to form inflorescences in year of sowing” the range of expressions is divided into: absent or very weak (1), weak (3), medium (5), strong (7), or very strong (9). The statistical analysis is made on the number of plants with inflorescences or without inflorescences.*

## ANNEX I

List of qualitative characteristics and descriptions of how they can be used to assess  
Distinctness and Uniformity

Name of character-istic	Distinctness			Uniformity		
	States for assess-ment	Description (states of expression)	Type of scale	Unit of assess-ment	Description (states of expression)	Type of scale
Sex of plant	1 2 3 4	dioecious female dioecious male monoecious unisexual monoecious hermaphrodite	nominally scaled qualitative data <sup>3</sup>	True-type  Off-type	Number of plants belonging to the variety Number of off-types	nominally scaled qualitative data
Ploidy	2 4 6	diploid tetraploid hexaploid	nominally scaled qualitative data <sup>3</sup>	True-type  Off-type	Number of plants belonging to the variety Number of off-types	nominally scaled qualitative data
Flower colour” for Lucerne varieties	1 2 3 4	white or yellow violet very dark violet variegated	combination of ordinal and nominal scaled qualitative data <sup>4</sup>		It is not possible to assess uniformity.	
Tendency to form inflorescences in year of sowing	1 9	absent present	nominal scaled qualitative data <sup>5</sup>		It is not possible to assess uniformity.	
Resistance to <i>Xanthomonas translucens (campestris) pv graminis</i> for Rye-grass varieties	1 9	dead plant living plant	nominal scaled qualitative data. <sup>5</sup>		It is not possible to assess uniformity.	

[Annex II follows]

<sup>3</sup> Distinctness occurs when varieties express different states of expression.

<sup>4</sup> Distinctness can be assessed by applying the Chi square test.

<sup>5</sup> Distinctness can be assessed by applying the chi-square test or the COYD criterion on percentage.

Examples of application of the Chi square test to a qualitative characteristic

1. The Chi square test can be applied to RxC tables where, for example, each of the R rows relates to one of the R different varieties which are being compared (in the case of DUS studies this is usually two, as all pair comparisons are computed), and each of the C columns relates to one of the C different states of expression of the characteristic. The value in the cell for variety *i* and state *j* is the observed frequency of plants in variety *i* which express state *j*. It should be noted that the roles of the rows and columns could equally well be reversed so that rows represented states and columns varieties.
2. The principle of the test is to compare the observed frequencies with the frequencies that would be expected if all R varieties, on average, had the same frequency of occurrence of the different states. This is the Null Hypothesis.
3. A Chi square statistic is calculated as follows:-

$$\chi^2 = \sum_{i=1}^R \sum_{j=1}^C \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where:

$O_{ij}$  denotes the observed number of plants (frequency) in variety *i* expressing state *j* of the characteristic.

$E_{ij}$  denotes the expected number of plants (frequency) in variety *i* expressing state *j*.

$E_{ij}$  is calculated as follows:-

Equation [1]

$$E_{ij} = \frac{(\text{Total number of plants for variety } i) \times (\text{Total number of plants expressing state } j)}{\text{Total number of plants for all varieties}}$$

4. The value of the Chi square statistic is compared with critical values from Chi square distribution tables on  $(R - 1)(C - 1)$  degrees of freedom.
5. For a specified significance probability, if the Chi square statistic is less than the critical value, the varieties are declared to be not distinct at that significance level for the given characteristic. If the Chi square statistic is greater than the critical value, then the varieties are declared to be distinct at that significance level for the characteristic.

**Example 1:** Use of the Chi Square test to assess distinctness between two varieties of Rye-grass for the characteristic Resistance to *Xanthomonas translucens* pv *graminis*, using Excel software:

	Variety X	Variety Y	Total number of plants
Number of dead plants	36	74	110
Number of living plants	67	29	96
Total number of plants	103	103	206

	Variety X	Variety Y	Total number of plants
Number of dead plants	55	55	110
Number of living plants	48	48	96
Total number of plants	103	103	206

Chi square statistic value = **28,17**

6. Of the 103 plants observed in variety X, 36 plants were found to be dead, whilst for variety Y, 74 plants were found to be dead out of 103 plants observed. The difference between variety X and variety Y is significant at the 0.1% level. The above table shows the observed frequencies and the Chi square statistic value. The expected frequencies are calculated using equation [1] above.

7. The table below, taken from an Excel worksheet, shows how each of the 4 cells contributes to the Chi square statistic value. The Chi square statistic value is the sum of the 4 contributions.

Example of chi square computation using Excel

	variety X	variety Y	sum
dead	36	74	110
alive	67	29	96
sum	103	103	206

	variety X	variety Y	sum
dead	55	55	110
alive	48	48	96
sum	103	103	206

Chi square statistic = sum for each cell of (observed - expected)\*(observed - expected)/ expected

Cell	formula	value
dead X	$=(36-55)*(36-55)/55$	6.56
alive X	$=(67-48)*(67-48)/48$	7.52
dead Y	$=(74-55)*(74-55)/55$	6.56
alive Y	$=(29-48)*(29-48)/48$	7.52
(sum of 4 cells) <b>Chi square</b>		<u>28.17</u>

degrees of freedom =  $(nR-1)*(nC-1) = (2-1)*(2-1) = 1$   
 $nR$  = number of rows = 2  
 $nC$  = number of columns = 2

Null Hypothesis: on average the two varieties have the same frequencies

8. The Chi square statistic is calculated as 28.17 on 1 degree of freedom. The 5%, 1% and 0.1% critical values of the Chi square distribution on 1 degree of freedom are 3.841, 6.635 and 10.83 respectively. Hence Variety X and Variety Y are distinct at the 0.1% significance level.

**Example 2:** Output from Chi square tests using SAS software to assess distinctness, between varieties of Lucerne, for the characteristic Flower Colour.

9. Trials comprise candidate and reference varieties. Each candidate variety is tested in a pair-wise comparison with all candidates and all reference varieties.

10. In the screen copy below is an example in which the following are shown:

- a simple SAS program to obtain a Chi square test from a dataset (*see editor window*)
- a partial view of the dataset, where “var” is the variety code (0104726 for Derby variety and 104730 for Europe variety) and “class” is the 4 states of the “flower colour” characteristic ( white or yellow (1), violet (2), very dark violet (3), variegated (4)). “Var” and “class” are used to obtain the rows and columns of the table of frequencies (*see Viewtable window*).
- a default output for a chi square test which shows the frequencies and the Chi square value (45.59) for a comparison of Derby and Europe on 10 replicates of at least 20 plants per replicate. In this example 201 plants have been observed for Derby and 227 plants for Europe. (*see output window*)

The screenshot shows the SAS interface with the following components:

- Output Window:** Displays the results of the FREQ procedure.
 

var(var)	class(class)				Total
	2	3	4		
0104726	196 31.78 67.66 41.09	58 13.55 28.86 82.86	7 1.64 3.48 25.93	201 46.96	
0104730	195 45.56 85.90 58.91	12 2.80 5.29 17.14	20 4.67 8.81 74.07	227 58.04	
Total	331 77.34	70 16.36	27 6.31	428 100.00	

  

Statistic	DF	Value	Prob
Chi-Square	2	45.5933	<.0001
Likelihood Ratio Chi-Square	2	48.4198	<.0001
Nominal Homogeneity Chi-Square	1	5.3428	0.0208
Phi Coefficient		0.3264	
Contingency Coefficient		0.3103	
Cramer's V		0.3264	

Sample Size = 428
- VIEWTABLE: Work Data:** Shows a partial view of the dataset with columns: var, name, rep, plant, value, class. It lists 20 rows of data for the Derby (0104730) and Europe varieties.
- Editor - Untitled1:** Contains the SAS program code:
 

```
proc freq data=data ;
table var*class /chisq;
output out=result chisq phi;
run;
```

11. In general, it is recommended that the total number of plants in each column or row in a table of frequencies to be used in a Chi Square test should have more than 5% of the total number of plants observed. Grouping of characteristic states can be done in order to avoid a situation where a column (or row) has less than 5% of the total number of plants observed. In

this case the test procedure is as before but uses a table with fewer columns or rows and consequently there are fewer degrees of freedom for the Chi square test.

12. The value of the Chi square statistic depends linearly on the number of plants observed. This means that, as is usually the case with other statistical tests, if the sample size is increased, smaller differences in percentages between 2 varieties can be shown to be “significant”.

## **REFERENCES**

SNEDECOR, G.W.; COCHRAN W. (1937); “Statistical methods”. *Iowa State University Press*.

[End of Annex II and of document]