E

TWC/26/11 Rev.
ORIGINAL: English
DATE: August 22, 2008

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Sixth Session
Jeju, Republic of Korea, September 2 to 5, 2008

DOCUMENT TGP/8: NON PARAMETRIC STATISTICAL METHODS

Document prepared by experts from Australia

1. At its forty-fourth session, held in Geneva from April 7 to 9, 2008, the Technical Committee (TC) considered the proposed structure and content of document TGP/8/1 Draft 9 and agreed, with regard to Part II to structure the section into separate sections on parametric and non-parametric methods and to include further methods for non-parametric methods, to be drafted by Australia.

2. The following proposed text has been prepared by experts from Australia for the section on non-parametric methods of TGP/8 Part II:

Introduction

3. Nonparametric methods are useful tools for DUS testing particularly when either:

- Observations are made using qualitative scales where the intervals between states of expression are not known or not necessarily equal (e.g. ordinal or nominal scales, see TGP/8 Part I section 2.5.4.2); or
- The underlying statistical assumptions needed by the parametric methods are not met or are untested.

4. Ordinal and nominal scaled data contain less information than interval or ratio data, and their analysis is by definition, less sensitive. This leads to the conclusion that nonparametric

methods are less powerful because, for the same sample size, they are less likely to confirm small differences between varieties. However where properly used, this may be an acceptable outcome which contributes to the maintenance of minimum distance and assists determination of "clearly distinct" as compared with "distinct by the smallest of differences".

5. Nonparametric methods are well suited to the analysis of characteristics assessed by "notes" such as for pseudo-qualitative and qualitative data and in situations where objective rigor is required in the development of national descriptors.

6. While nonparametric methods are usually applied to the analysis of ordinal and nominal scaled data, they can also be used to analyze interval or ratio data. Nominal scaled data can only be analyzed using nonparametric methods.

7. Where sample size is small, (say less than 6 observations), there is no alternative to using nonparametric methods unless the distribution of the states of expression of the candidate variety are known exactly (a rare circumstance for DUS testing authorities).

Role of non-parametric analysis for analyzing quantitative data

8. Generally, for quantitative measured data, such as plant length in centimeters or number of stamens (cross ref TGP8 Part I, 2.5.4), parametric statistical methods are preferred. The use of parametric methods relies on underlying assumptions of the population distribution. They are usually robust and powerful even if there is moderate departure from the statistical assumptions (such as departure from a normal distribution). If assumptions are badly violated, nonparametric tests could be employed, however, before doing so, it is necessary to first investigate whether experimental error is the cause (cross ref TGP 8 Part I section 4.2) or establish that the type of data collected does not fit the parametric assumptions. There are many nonparametric tests (e.g. Kruzkal-Wallis one way analysis of variance and Mann-Whitney U test) that could be used and these are well documented and described. The use of nonparametric statistics for quantitative measured data from DUS trials is the exception rather than the rule and it is not necessary to describe these further here. Instead it is sufficient to note that these methods are documented in statistical literature and can be considered if necessary.

Role of non-parametric analysis for analyzing qualitative data

9. Some characteristics routinely used in DUS testing do not usually satisfy the assumptions required for parametric methods. Qualitatively scaled data are usually obtained from visually assessed characteristics using ordinal or nominal scales (cross ref TGP/8 Part 1 2.5.4.2). For example, where individual plants are scored on 1 to 10 scale of increasing resistance to a particular disease, the position within the scale is important (i.e. it is an ordinal scale). If one plant is assessed as having a higher level of resistance than another then it is scored with a higher number on the scale. However, it is usually difficult to precisely identify the limit of each interval of the scale. Consequently, the exact interval size is unknown and is likely to vary. For this reason the scores cannot be treated as quantitative data with an assumed normal distribution which would allow the use of parametric methods. Instead it is appropriate to use nonparametric methods that do not rely on equally spaced intervals. Another example is scoring of results from an iodine starch test in assessing the maturity of apples using an ordinal scale.

10. Sometimes individual plants can be placed in "categories" where the order does not matter (i.e. a nominal scale) e.g. scoring plants as shattering or non-shattering in *Phalaris*.

11. Where all or most plants of a variety fall into one category it is unnecessary to apply a statistical method to decide on distinctness. However, in some cases, particularly for cross pollinated varieties, the allocation to categories is not absolute and there will be a certain amount of heterogeneity in the population due to the breeding system of the species. The consequence is that large numbers of plants of the variety may be allocated to different categories. This is acceptable provided the degree of heterogeneity is within that for comparable varieties of the species. A decision has to be made as to whether there is sufficient separation to establish distinctness between varieties.

12. In these cases, nonparametric statistical methods can be used as they do not rely on assumptions about the underlying population distribution of the data.

13. Whilst there are many nonparametric methods that can be used for qualitative data, two methods commonly used in plant variety testing are the Chi-square (χ^2) and Fishers Exact Test. For convenience these are briefly described below.

Chi square test

14. The Chi-square test is useful where observations on a characteristic are allocated to two or more categories (classes). Each category should have a minimum of five counts.

15. In DUS trials, many of the characteristics are observed by measurements such as plant height, leaf length, leaf width, flower diameter etc. These are continuous variables and are expected to follow normal distribution with μ mean and σ^2 variance. These can be in general, statistically analyzed using 'Student t criterion' or F test. However, in some cases, distinctness may be established by classifying individual varieties into broad groups and demonstrating statistically different grouping patterns for different varieties. Such examples include counts based on the flower color groups - red, pink or white etc. and the disease/pest/nematode infection classes. Data based on counts of individuals in a sample/population belonging to each of several classes require a different kind of statistical analysis. A method commonly used for analyzing such enumeration data is called the *Chi-square* (χ^2).

16. 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 and stability by formulating certain hypotheses using the classification data.

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

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

17. This factor by definition is *the sum of squares of independent, normally distributed variables with zero mean and unit variance.* Hence, the Chi-square distribution is a continuous distribution based upon an underlying normal distribution.

Note: 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 value can be made larger. Alternatively, if some classes have a size less than five, pool those classes to bring the size of the pooled class to five or more than five.

(3) The number of degrees freedom to look up on the chi-square table is not always obvious. *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 testing any 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. If you encounter such situations, calculate expected values using formulae based on the binomial distribution. Always use Yates Correction for determining the chi-square test with only one degree of freedom.

18. Let us examine the following data on the disease scoring of two generations of a lucerne candidate variety and its four comparator varieties. The disease scored was Coletotrichum crown rot in lucerne. The scoring was on 1-5 scale, note 1 being resistant and note 5 being susceptible.

Number of plants counted in different classes in each variety after 7-10 days of inoculation

Class/Score	Candidate Generation 1	Candidate Generation 2	Comparator 1	Comparator 2	Comparator 3	Comparator 4
1	34	32	12	6	1	7
2	4	3	7	6	5	10
3	1	3	9	5	5	5
4	1	2	7	9	8	7
5	6	4	9	19	9	15
Total	46	44	44	45	28	44

19. It can be seen from the table that the two generations of the candidate variety have more plants in the resistant category than the comparators. However, to statistically test the significance of these differences, we need to formulate two hypotheses:

(1) Whether the comparator varieties differ significantly or not from the generation 1 of 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 Coletotrichum crown rot. This can be done by testing the "distinctness χ^2".

(2) If the two generations of the candidate differ from one another in the distribution of scores. This can be approached by testing another null hypothesis that the two generations behave similarly to the inoculation of Colectrichum crown rot. This can be done by testing "stability χ^2".

20. The generation 1 of the candidate variety is considered as a reference variety for PBR comparisons. Hence, the distribution of scores in different classes observed for this reference variety is considered to be the expected distribution. The expected values of classes 2, 3 and 4 for generation 1 of the candidate are less than 5 and it would be appropriate to pool all the values in those classes to form a new intermediary pooled class for all the varieties under consideration.

Now the observed data is reduced to:

Class/Score	Candidate Generation 1	Candidate Generation 2	Comparator 1	Comparator 2	Comparator 3	Comparator 4
1	*34*	32	12	6	1	7
2	*6*	8	23	20	18	22
3	*6*	4	9	19	9	15
Total	*46*	44	44	45	28	44

21. The distribution of expected values for different varieties are as using the distribution of the scores for the reference variety (0.74 (34/46) for class 1, 0.13 (6/46) for class 2 and 3 respectively) is as follows:

Class/Score	Candidate Generation 1	Candidate Generation 2	Comparator 1	Comparator 2	Comparator 3	Comparator 4
1	*34*	32.52	32.52	33.26	20.70	32.52
2	*6*	5.74	5.74	5.87	3.65	5.74
3	*6*	5.74	5.74	5.87	3.65	5.74
Total	*46*	44	44	45	28	44

The total χ^2 for the whole set of data is as follows:
χ^2 = (34 -34)^2/34 +... (32-32.52)^2/32.52 +... (12 - 32.52)^2 /32.52+...(6 - 33.27)^2/33.27 + (1 20.70)^2/20.70 +...(7-32.52)^2/32.52+... (15-5.74)^2/5.74 = 317.87

22. At v(n-1) degrees of freedom i.e., 6(2) = 12 df the table χ^2 value is 26.22 at P = 0.01. The calculated value is more than the table value and hence there are significant differences among varieties for Coletotrichum crown rot (CCR). Hence, the null hypothesis that there are no significant differences in reaction to CCR among the varieties is rejected.

23. For calculating the "distinctness χ^2" for comparator 1

χ^2 = (12 -32.52)^2/32.52 + (23 -5.74)^2/5.74 + (9 - 5.74)^2/5.74
= 35.1 + 12.95 + 1.18
= 49.23

24. The number of degrees of freedom for looking up the χ^2 table is one less than the number of classes i.e., 3 - 1 =2.

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

26. Similarly the calculated "distinctness χ^2" for comparator-2, comparator-3 and comparator-4 are 142.92, 402.53 and 110.79, respectively, which are all greater than the table χ^2 value of 9.21 at 2 df.

27. Hence, all the comparator varieties are significantly different from the generation 1 of the candidate variety in reaction to Coletotrichum crown rot.

28. Similarly, for calculating the "stability χ^2" the observed and expected values of generation 2 of the candidate variety are to be used.

29. Thus, "Stability χ^2" is

$$\chi^2 = (32 - 32.52)^2/32.52 + (8 - 5.74)^2/5.74 + (4 - 5.74)^2/5.74$$
$$= 0.01 + 0.64 + 0.76$$
$$= 1.41$$

30. This should be tested again at 2 df and it turns out to be non-significant. Hence, the null hypothesis is accepted and it is concluded that the two generations of the candidate show similar reaction to Coletotrichum crown rot.

31. Thus, χ^2 analysis is a useful analytical tool to analyze such categorical data for PBR.

Fisher's Exact Test

32. 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 and is named after its inventor, R.A. Fisher.

33. 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[1] 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 - as it is usually quicker to calculate.

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

Example 1

35. In a self-pollinated species, seed retention in the inflorescence is accepted as a relevant characteristic in the DUS trial. In this example of a DUS trial with two varieties, the seed

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

retention characteristic has two states of expression (i) shattering and (ii) non-shattering inflorescence.

36. Assume that the two varieties (Variety 1 and Variety 2) have some observed differences in the proportion of non-shattering inflorescences. Examiners need to be able to reliably determine whether these differences can be accepted as clearly distinct 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 shattering and non shattering plants observed in Variety 1 and Variety 2

	Variety 1	Variety 2	Total
Shattering	**4**	**9**	13
Non-shattering	**8**	**3**	11
Total	12	12	24

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

37. What is the probability that Variety 1 is distinct from Variety 2 on the basis of this characteristic, knowing that 11 of these 24 plants are non-shattering 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 seed retention 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[2].

38. 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
Shattering	a	b	a + b
Non-shattering	c	d	c + d
Total	a + c	b + d	N

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

39. 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).

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

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

[2] 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.

After solving the factorials:

$$p = 0.04$$

41. Interpreting the p value calculated by Fisher's Exact Test is straight forward. In the example above, p = 0.04 meaning that there is a 4% 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 distinct 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 non-shattering vs. shattering characteristic is not significantly different and the two varieties (Variety 1 and Variety 2) are not distinct on that basis.

Example 2

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

Table 3: Number of shattering and non shattering plants observed in Variety 3 and Variety 4

	Variety 3	Variety 4	Total
Shattering	1	9	10
Non-shattering	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:

$$p = 0.001$$

43. In this particular case, the null hypothesis (that the varieties are similar on the basis of non-shattering vs. shattering 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.

Uniformity

44. Uniformity for this characteristic could be assessed if the trial is replicated. Assuming that the trial used in example 2 has two more replicates. The data for the candidate variety (Variety 3) from all three replicates are compared in Tables 4, 5 and 6.

Table 4: Number of shattering and non shattering plants observed in Variety 3 (Rep 1 and Rep 2)

	Variety 3 (rep1)	Variety 3 (rep2)	Total
Shattering	1	2	3
Non-shattering	11	10	21
Total	12	12	24

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

p = 0.39

Table 5: Number of shattering and non shattering plants observed in Variety 3 (Rep 1 and Rep 3)

	Variety 3(rep 1)	Variety 3 (rep3)	Total
Shattering	1	3	4
Non-shattering	11	9	20
Total	12	12	24

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

p = 0.24

Table 6: Number of shattering and non shattering plants observed in Variety 3 (Rep 2 and Rep 3)

	Variety 3(rep 2)	Variety 3 (rep3)	Total
Shattering	2	3	5
Non-shattering	10	9	19
Total	12	12	24

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

p = 0.34

45. In the comparisons above, the calculated p values are much higher than the threshold limit (p=0.01) for rejecting the null hypothesis that the candidate variety is same in all three replicates. Therefore, we accept the null hypothesis and conclude that the candidate variety is sufficiently uniform for this characteristic.

[End of document]