E

TWC/22/12
ORIGINAL: English
DATE: June 7, 2004

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Second Session Tsukuba, Japan, June 14 to 17, 2004

COMMENTS ON STATISTICAL METHODS USED FOR COMPARISON OF VARIETY DESCRIPTIONS

Document prepared by experts from Germany

COMMENTS ON STATISTICAL METHODS USED FOR COMPARISON OF VARIETY DESCRIPTIONS

Uwe Meyer and Beate Rücker
Bundessortenamt, Hannover, Germany

1. Introduction

During its fortieth session, held in Geneva, from March 29 to 31, 2004, the Technical Committee discussed the developments in the project for the publication of variety descriptions. The TC agreed that the Chairman of the TWC should, after consultation with the members of the TWC, develop guidance on how to present the variation in the states of expression between different descriptions of the same variety and communicate this guidance to the coordinators of the model studies via the Office.

The aim of this paper is to discuss the application of standard deviations on variety descriptions and to develop guidance on how to present the variation in the states of expressions between different descriptions of the same variety.

It is obvious to use standard deviation (STD) to estimate the variation between different variety descriptions. The known formula is:

$$STD = \sqrt{\frac{1}{n-1} * \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

x_i note for a characteristic for the ith country,
n number of countries,
i varies from 1 to n and
$\overline{x}$ arithmetic mean of this characteristic over all countries

From the statistical point of view the application of standard deviation on notes requires the fulfilment of particular conditions.

Table 1: Examples for a Barley variety (fictitious):

Characteristic		Scale	Notes	STD
number	name		Country 1 2 3 4 5 6	
16	Ear: length	1 – 9	5 . 5 7 5 7	1.10
29	Seasonal type	1, 2, 3	1 . 3 1 1 3	1.10
XX	Leaf: color	1, 2, 3	1 . 3 1 1 3	1.10

For characteristic 16 (Ear: length) the note 5 is ‘medium’ and the note 7 is ‘long’. For characteristic 29 (Seasonal type) note 1 is ‘‘winter type’ and note 3 is ‘spring type’.

Characteristic XX (Leaf: color) is fictitious to demonstrate a special type of characteristic which does not exist in Barley. For this characteristic note 1 is ‘green’ and note 3 is ‘‘red’.

2. Type of characteristics and scale level of data

The characteristic ‘Ear: length’ is quantitative and the notes for variety description are ordinal scaled data.

The characteristic ‘Seasonal type’ is quantitative. The note 2 is the ‘alternative type’ and lies between note 1 and 3.

The characteristic ‘Leaf: color’, in this theoretical case, is qualitative and the notes for variety description are nominal scaled data (see TGP/8.3 draft 3). In practice it is unlikely to observe this characteristic with note 1 in one country and with note 3 in another. The example has been chosen specifically just to illustrate the statistical consequences.

3. Relation between scale level of data and conditions for application of standard deviations

To apply standard deviation on a set of data it is necessary to have interval scaled data or ratio scaled data. The requirement is the existence of equal distances between the expressions (here notes). That means the intervals between all pairs of notes have the same length. This condition is not fulfilled for nominal data. For ordinal data, the fulfilment of this condition varies from characteristic to characteristic. Sometimes the condition is fulfilled. In this case the scale level lies on the border between ordinal and interval scaled data. In other cases the scale level tends from ordinal towards nominal data.

3.1. Nominal scaled data:

In table 1 the example for characteristic XX (Leaf: color) is given. Although the expressions are well-defined in the Test Guidelines, the order of the expressions may not be clearly specified, so that we get a different definition described in table 2. This new definition is only used to describe examples and not to propose any change in the current guidelines.

Table 2: Reorder of expressions of characteristic XX

Definition	Notes		
	1	2	3
TG/YY/ZZ	green	yellow	red
new proposal	red	green	yellow

The example for characteristic XX from table 1 can be transferred into the new notes by using the new proposal from table 2. The result is given and compared in table 3.

Table 3: Comparison of two definitions of expressions of characteristic XX (Leaf: color)

Definition	Notes per Country 1 2 3 4 5 6	STD
TG/YY/ZZ	1 . 3 1 1 3	1.10
new proposal	2 . 1 2 2 1	0.55

Thus, if we assign expressions to different notes for nominal scaled data we get different standard deviations for the same example. This is the reason to declare that for nominal scaled data calculation of standard deviation is not allowed.

From a statistical point of view there is no measurement for deviations for nominal scaled data.

3.1.1 Guidance for nominal scaled data:

To compare variety descriptions from different countries the structure of table 1 is very useful but the standard deviation must be eliminated from the table. Additionally another structure could provide a new view of the data. Therefore the frequencies of all expressions should be computed as shown in table 4.

Table 4: Frequency table for a variety (fictitious)

Characteristic		Notes		
number	name	1	2	3
XX	Leaf: color	3	0	2

This kind of frequency table is easy to produce and appropriate to show the variation of the characteristic. The same information can be illustrated graphically by using a histogram.

Some examples are given in table 5 in order to show variation for nominal scaled data for different varieties.

Table 5: Frequency table for different varieties (fictitious) for characteristic Leaf: color (1 – green, 2 – yellow, 3 – red)

Variety	Notes			Variation class in description		STD
	1	2	3			
A	5	0	0	absent	1	0.00
B	0	5	0	absent	1	0.00
C	0	0	5	absent	1	0.00
D	4	1	0	very small	2	0.45
E	4	0	1	very small	2	**0.89**
F	0	4	1	very small	2	0.45
G	0	1	4	very small	2	0.45
H	1	4	0	very small	2	0.45
I	1	0	4	very small	2	0.89
J	3	0	2	small	3	1.10
K	3	2	0	small	3	0.55
L	0	3	2	small	3	0.55
M	0	2	3	small	3	0.55
N	2	3	0	small	3	0.55
O	2	0	3	small	3	1.10
P	3	1	1	medium	4	0.89
R	1	3	1	medium	4	0.71
S	1	1	3	medium	4	0.89
T	2	2	1	high	5	**0.84**
U	2	1	2	high	5	1.00
V	1	2	2	high	5	0.84

Examples given in table 5 show the differences between the classification from expert's point of view in classes from 1 to 5 (in this example) and the classical standard deviations (STD). Here it is very important to understand the formation of different classes. For nominal scaled data the notes themselves are not allowed to be used for estimation of variation. Variation can be described only by frequencies.

It is obvious that varieties A, B and C are in the same class without any variation. Varieties D to I fall into class 2 (very small variation) because only one country gave another note than all the other and so on. The maximum variation is when the variety has the same or nearly the same frequency in each note and, where the frequency for no note is equal to zero.

It is important that the standard deviation varies not continuously from class 1 to 5. Variety E has a standard deviation of 0.89 and it is part of variation class 2. Variety T has a standard deviation of 0.84 and it is part of class 5. Thus, the application of standard deviation for nominal scaled data leads to incorrect results.

The definition of variation classes must be done by crop experts. In the example above it is possible to decide that variation class 3 and 4 have the same variation. The decision depends on the characteristics.

The best way to analyse the variation of nominal scaled data is the use of frequency tables and classifications of variation made by the crop expert. The use of standard deviation is not appropriate.

3.1.2. Dichotomous or binary data:

Dichotomous or binary data are special cases of nominal data with two classes (absent and present for example). For this kind of data it has been agreed to use note 1 for absent and note 9 for present. From the statistical point of view there are also other possibilities such as 1 and 2 or 0 and 1.

In table 6 a fictitious example for varieties are given by using three different types of combination of notes and the consequences to variation classes and the standard deviation as parameter of variation.

Table 6: Frequency table for a fictitious variety with different combination of notes

Variety	Combination	Notes				Variation	class	STD
		0	1	2	9			
A	1 and 9		5		0	absent	1	0.00
A	1 and 2		5	0		absent	1	0.00
A	0 and 1	5	0			absent	1	0.00
B	1 and 9		4		1	very small	2	**3.58**
B	1 and 2		4	1		very small	2	0.44
B	0 and 1	4	1			very small	2	0.44
C	1 and 9		3		2	small	3	**4.38**
C	1 and 2		3	2		small	3	0.55
C	0 and 2	3	2			small	3	0.55

There are three variation classes for this example in table 6. The symmetrical cases (0 5; 1 4 and 2 3) fall into variation class 1, 2 or 3 respectively. It is obvious that the standard deviation depends on the choice of notes. The pair (1 9) produces more variability than the other (see variation classes 2 and 3 in table 6).

The definition of the notes of a characteristic is given by the Test Guidelines and should be obligatory for all crop experts. So there is no problem by using the standard deviation in this case.

3.2. Ordinal scaled data

In table 1 the example for characteristic 16 (Ear: length) is given. The expressions are well defined in the Test Guidelines (TG/19/10). The example given in table 1 (Characteristic

16 or Ear: length) was used and modified in order to show the influence of the variation of ordinal scaled data on statistical parameters like minimum, maximum, range and standard deviation (STD). The range is the difference from maximum to minimum. Results are shown in table 7.

Table 7: Statistical parameters of ordinal scaled data by using of fictitious examples (Ear: length)

Variety	Notes									Min	Max	Range	STD
	1	2	3	4	5	6	7	8	9				
A									5	9	9	0	0.00
B					5					5	5	0	0.00
C	5									1	1	0	0.00
D					2	1	2			5	7	2	1.00
E					3		2			5	7	2	1.10
F					2		3			5	7	2	1.10
G				1	2		2			4	7	3	1.34
H			1		2		2			3	7	4	1.67
I			1		1		2		1	3	9	6	2.28
J			1		1		1		2	3	9	6	2.61
K			2		1		1		1	3	9	6	2.61
L	2		1		1		1			1	7	6	2.61
M	1		1		1		1		1	1	9	8	3.16
N	1								4	1	9	8	3.58
O	2								3	1	9	8	4.38
P	3								2	1	9	8	4.38

It is obvious that the range correlates with the standard deviation in this example. From the statistical point of view the range is the best parameter to estimate variation within ordinal scaled data. The conditions to apply the standard deviation are not fulfilled. But the correlation between both measures shows that the error is very small by using the standard deviation in this case.

3.2.1 Guidance for ordinal scaled data:

To compare different countries the structure of table 1 is very useful. The standard deviation as measure for variation correlates with the range (difference between maximum and minimum). The range is the best parameter to estimate variation in notes (from the statistical point of view).

The calculation of frequencies of all expressions can provide additional information as shown in table 8.

Table 8: Frequency table for a Barley variety (fictitious)

Characteristic		Notes									Range
number	name	1	2	3	4	5	6	7	8	9	
16	Ear: length					3		2			2

This kind of frequency table is easy to produce and appropriate to show the variation of the characteristic.

4. Comparison of a set of characteristics for a chosen variety and other comparisons:

To compare varieties regarding to a chosen ordinal scaled characteristic variation parameters like standard deviation or range are useful. The comparison of two or more characteristic for a given variety is easy for characteristics of the same type of scale. But in the case of different scale levels of characteristics (nominal and ordinal or ordinal with 1 to 9 and 1 to 3 for example) the comparison of variation is not so easy as demonstrated in TWA/30/16. In this document there are also tables to compare a set of characteristics for a chosen variety (table 3 on page 7) and a table to show the variation for each variety regarding to all characteristics by using intervals of standard deviations (table 5 on page 9). On page 11 of document TWA/30/16 the variation for each characteristic is shown regarding to all varieties (table 6). In all cases the standard deviation is the used measure.

For the cases of comparison of varieties regarding to a chosen characteristic (nominal or ordinal) the guidance is given above. For other kind of comparisons there is the problem that only in the case of ordinal scaled data the standard deviation is comparable. For nominal scaled data it is not allowed to compute the standard deviation.

Proposal for a guidance:

Standard deviation should be accepted as measure of variation for ordinal scaled data but not for nominal scaled data. Most of the grouping characteristics are nominal scaled and the expected variation is zero. Thus, there is no need for a parameter of variation. For the other nominal scaled characteristics crop experts could define variation classes (see chapter 3.1.1). The variation classes of two or more characteristics with nominal scaled data can be compared by the number of classes and by the frequencies.

Nominal scaled data with two categories (dichotomous or binary data) can be handled as nominal or ordinal.

5. Discussion

The participants of the twenty-second TWC meeting are invited to discuss the use of variation parameter standard deviation, range and variation class for characteristics with nominal or ordinal scaled data.

A proposal for the Coordinators of the Crop Subgroups for the Publication .of Variety Descriptions to present and analyze data is presented in Annex I to this document.

[Annex follows]

ANNEX

Proposal for Recommendations for Coordinators for the Ad Hoc Crop subgroups for the Publication of Variety Descriptions

Following the request made by the Technical Committee during its fortieth session, held in Geneva, from March 29 to 31, 2004, the TWC recommends that the coordinators of the Crop Subgroups for the Publication of Variety descriptions use the following tables to present and analyze the data.

Table 1 Qualitative Characteristics (QL) (e.g. Ploidy type)

VARIETY	Nº of descriptions	NOTES		
		2	4	6
A	5	4	1	0
B	4	0	4	0
……				
……				
……				

Table 2 Pseudo Qualitative Characteristics (PQ) (e.g. Flower color)

VARIETY	Nº of description	Notes					
		White 1	Yellow 2	Green 3	Blue 4	Red 5	Purple 5
A	5	4	1				
B	4				3		1
C	5		1	4			
……	……						
……	……						

Table 3 Quantitative Characteristics (QN) (e.g. Leaf length)

Variety	Nº of description	NOTES									Range	STD deviation
		1	2	3	4	5	6	7	8	9		
A	5					2	1	2			2	1.00
B	5				1	2		2			3	1.34
C	5	1								4	8	3.58
…..	……										…..	…..
								Average			X	Y