TWC/24/11
ORIGINAL: English
DATE: June 13, 2006

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Fourth Session
Nairobi, June 19 to 22, 2006

COMPARISON OF DISTRIBUTION OF COLOURS OF COTYLEDONS IN SUGAR BEETS FOR DISTINCTNESS PURPOSES

Document prepared by experts from Denmark

Introduction

1. In beets, one of the characteristics used in Denmark for description and distinctness purposes is the frequency of different colours on the cotyledons. The criteria for distinctness here is whether the distribution of the colours differs significantly from the other varieties. The test for that can be done in different ways:

- The frequency of each colour can be analysed separately and the varieties can be considered different if at least the frequency is significantly different for at least one colour. One problem with this method may be that a single character gets more than one chance to become significantly distinct. This may easily result in liberal test, so that it may be expected to yield some false significant distinctness (too many varieties becomes distinct).

- All colours of the characteristics are tested simultaneously by testing whether the number of observations in the different colours can be considered independent. This test cannot be carried out in the traditional COY-D analysis. However, the hypothesis of independence can be tested using the simple χ^2-test for independence. By doing that, the additional variability caused by year-to-year and plot-to-plot variation are ignored, which means that this method may also be expected to yield false significantly distinct pairs.

2. The present paper tries to document that and to suggest an alternative method to be used.

Data

3. The data used for this investigation are from the Danish DUS trial with sugar beets in 2001, 2002, 2003 and 2004. From each of 2 replicates, approximately 50 plants were selected randomly. The cotyledon of each plant was then visually accessed and assigned to one of the following 7 colours: green, white, pink, red, dark red, yellow or orange, and the number of plants with each colour were then counted.

Statistical methods for analysing these colours

4. The simplest method for analysing the data for independence is to calculate Persons χ^2 using the following formula:

$$\chi^2 = \sum_{v=1}^{2}\sum_{c=1}^{p}\frac{(O_{vc} - E_{vc})^2}{E_{vc}}$$

where

O_{vc} is the observed number of cotyledons with colour c for variety v

E_{vc} is the expected number of cotyledons with colour c for variety v if the 2 varieties has the same distribution of colours

5. The calculations are demonstrated using the following example:

Observed distribution of colours in the 2 varieties

Variety	Green	White	Pink	Orange	Total
A	0	18	103	163	284
B	166	9	31	92	298
Total	166	27	134	255	582

Expected distribution of colours in the 2 varieties if they had same distribution of coloured cotyledons

Variety	Green	White	Pink	Orange	Total
A	81.0	13.2	65.4	124.4	284
B	85.0	13.8	68.6	130.6	298
Total	166	27	134	255	582

6. The expected values are the values that can be expected if the proportion of observations for each colour is the same for both varieties: as an example, 28.52% of the total number of observations are green and therefore we expect 28.52% of the total number of observations for variety A to be green, i.e. 28.52% of 284=81.0, and 28.52% of the total number of observations for variety B to be green, i.e. 28.52% of 298=85.0.

7. Based on those two tables we calculate:

$\chi^2 = \frac{(0-81.0)^2}{81.0} + \frac{(18-13.2)^2}{13.2} + ... + \frac{(92-130.6)^2}{130.6} = 227.2$ with 3 degrees of freedom which means that χ^2 is much larger that what can be expected to be caused by random sampling

8. This can be shown to give approximately the same results as if the data were analysed using the following generalised linear model:

Y_{vc} = Poisson distributed with parameter λ_{vc}
where
$\log(\lambda_{vc}) = \mu + \alpha_v + \beta_c$
and then use the deviance from that model as a χ^2 statistics.

9. This is so because non-additivity on a log-scale is the same as deviation from a proportional model on the untransformed scale. For the example above, the generalised linear model gave a χ^2 of 293.7. For the exact method of calculation, see a book on generalised linear models.

10. However, these methods only take into account the variability caused by random sampling among a homogeneous sample, i.e. they do not take into account the variability caused by varying growing conditions from plot to plot or from year to year. To demonstrate this, the above test was carried out by testing if the distribution of colours in the 2 replicates of the same variety could be the same. In the data above, there were in total 444 combinations of years and varieties. For each of those 444 combinations the above χ^2-tests were performed. The results are shown in Table 2. If the only variation were random sampling and a 0.1% test were performed one should expect that 0 or 1 (or may be 2-3) of those 444 tests were found to be significant. However, 17 (≈3.8%) were found to be

significant at the 0.1% level of significance when the simple method (Pearson chi-square test) above is used. So this test does clearly give more significant results than expected. If the tests are performed at the 1% or 5% level of significance similar results are found.

Table 2: Empirical percent of significant times the distribution in the 2 replicates of the same variety were significantly different. Calculated for 4 different nominal levels for all 444 combinations of varieties and years

Nominal level	Empirical level
	Pearson chi-square
0.1	3.8
1.0	10.1
5.0	21.8
10.0	31.5

Alternative methods

11. The analyses of the frequency for a given colour can be carried out using the standard COY-D method (after an appropriate transformation) but then each pair of varieties is given several chances to be significant for the same characteristic.

12. A possible way to do the analyses would be to generalise the COY-D methods to handle proportions in a way where the variability from plot to plot and from year to year are also taken into account. Here it is suggested to use the following generalised linear mixed model:

Y_{yvc} = Poisson distributed with parameter λ_{yvc}

where

$$\log(\lambda_{yvc}) = \mu + \alpha_v + \beta_c + (\alpha\beta)_{vc} + B_y + C_{yv} + D_{yc} + E_{yvc}$$

where

Y_{yvc} is the sum total number of observations with colour c for variety v in year y

μ, α_v, β_c and $(\alpha\beta)_{vc}$ are fixed effect of mean, variety, colour and interaction between variety and colour

$(\alpha\beta)_{vc}$ is the effect of interest that has to be tested for the actual pair of varieties

B_y, C_{yv}, D_{yc} and E_{yvc} are random effects of year, variety $\times$ year, colour $\times$ year and variety $\times$ colour $\times$ year

B_y, C_{yv}, D_{yc} and E_{yvc} are assumed to be i.i.d. normal distributed variables with means zero and variances σ_{B}^2, σ_{C}^2, σ_D^2 and σ_E^2, respectively

13. One problem with this model is that not all colours are present in all varieties. Therefore, zero expectations will occur, which gives problems when we model the logarithms of the expectations. One solution would be to set up the above model for each pair of varieties to be tested using only the colours present for that pair, but this would then give a poor power of the tests because only few degrees of freedom will be available for estimating the variance components.

Methods and results

14. Applying the simple method and the alternative method to all pairs of varieties that were present in at least 2 years (2, 3 or 4 years) yielded the results shown in the following tables. The results obtained by using the simple method are shown in Table 3 and the results obtained by using the generalised linear mixed model are shown in Table 4. By comparing those two methods it is clear that the simple method yields much more significant results than then generalised linear mixed model.

15. For the simple method almost the same number of significant results was obtained for pairs present in 2, 3 or 4 years. For the method based on the generalised linear mixed model, the percent of distinct pairs was clearly higher for the pairs that were present in 4 years than for the pairs that were present in only 2 years. This may be caused partly by the fact that the number of degrees of freedom in the denumerator of the tests was much lower for tests based on 2 years (about 4-6) than for tests based on 3 years (about 8-12) and 4 years (about 12-18) and partly because the total number of observations increases with the number of years.

Table 3: Empirical percent of significant pairs of varieties for different number of years in test and 4 different nominal levels using a simple method

	Nominal level				Number of pairs
Number of years	0.1%	1%	5%	10%	
2	93.8	96.2	97.7	98.3	2365
3	95.1	96.7	97.9	98.4	822
4	95.1	96.7	97.9	98.4	1078

Table 4: Empirical percent of significant pairs of varieties for different number of years in test and 4 different nominal levels when the data were analysed using a generalized linear mixed model. The last 2 columns shows the variance component for interaction with year and the residual (an over dispersion factor).

	Nominal level				Year×Colour ×Variety	Residual
Number of years	0.1%	1%	5%	10%		
2	23.1	52.6	73.0	79.7	0.014	1.31
3	51.8	68.6	80.1	86.0	0.061	1.16
4	67.3	77.8	86.1	88.9	0.035	1.32

Concluding Remarks

16. The comparison of distribution of the colours in two replicates of the 444 combination of years and varieties clearly shows that the simple method based Pearson χ^2 yields too many significant results. It is suggested that a generalised linear model can be used for comparing the distributions of e.g. colours instead of the method based Pearson χ^2. The generalised linear mixed model is to some extent similar to the well known COY-D method for characteristics.

17. However, it might be expected that the power of the tests will be rather low if the analyses are carried out using only the data for each pair of varieties in each analysis. Therefore, it is necessary to find a method to include several (all) varieties in the analyses in order to obtain a reasonably high number of degrees of freedom in the denominator of the F-tests. The problem here is that not all colours are present in all varieties.

18. It is suggested that the same method can be used for other segregating characteristics.

[End of document]