E

TWC/27/14
ORIGINAL: English
DATE: June 4, 2009

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Seventh Session
Alexandria, Virginia, United States of America
June 16 to 19, 2009

STATISTICAL METHODS FOR VISUALLY OBSERVED CHARACTERISTICS
(WITH EMPHASIS ON GENERALIZED LINEAR MODELS)

Document prepared by an expert from Denmark

STATISTICAL METHODS FOR VISUALLY OBSERVED CHARACTERISTICS (WITH EMPHASIS ON GENERALIZED LINEAR MODELS)

Kristian Kristensen
Department of Genetics and Biotechnology, Faculty of Agricultural Sciences,
University of Aarhus, Denmark

Introduction

1. Visual observation of characteristics usually results in a note (often on a 1-9 scale). The note is usually regarded as being either nominal or ordinal. In some cases, an interval scale may be assumed and, only in that case, it could be reasonable to analyze the mean characteristic as a continuous variable. When the scale is either nominal or ordinal, other methods are needed. This paper describes some methods that may be appropriate for such data.

Methods

2. One of the simplest methods is to analyze the data as a contingency table formed by variety and notes (such as table 1) and then test for independence in that table using the following formula:

Method A

$$\chi^2 = \sum_{i}^{n}\sum_{j}^{v} \frac{(Y_{ij} - E_{ij})^2}{E_{ij}} \text{ with } E_{ij} = \frac{Y_{i.}Y_{.j}}{Y_{..}},\ Y_{i.} = \sum_{j}^{v} Y_{ij},\ Y_{.j} = \sum_{i}^{n} Y_{ij},\ Y_{..} = \sum_{i}^{n}\sum_{j}^{v} Y_{ij}$$

where

n is the number of notes and v is the number of varieties

Y_{ij} is the number of plants for variety j with note i

χ^2 is a test statisctis that is $\chi^2 -$ distrbuted with $(n-1)(v-1)$ degrees of freedom if the notes are distributed independently of variety.

3. This is an acceptable method for analyzing both nominal and ordinal data *if the only source of variation is sampling error*. However, some drawbacks may be mentioned:

(a) More efficient methods are available, especially for characteristics on the ordinal scale;
(b) The method cannot take into account other sources of variation, such as variations caused by “soil fertility”, variation from year- to-year and uncertainty in the recording of a note (e.g. when the characteristics of a plant is somewhere between two notes); and
(c) Pair wise comparisons require the above formula to be evaluated for all pairs of interest

Method B

4. The test above can also be formulated using a generalized linear model where it is assumed that the logarithm of the expected value can be formulated as a linear model including the effects of notes, variety and interaction between varieties and notes. The test for

significance can then be done by testing for interactions between varieties (or pairs of varieties) and notes. The model may be written as:

Y_{ij} are Poisson-distributed with mean value λ_{ij}

$\log(\lambda_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$

where

Y_{ij} is the number of plants for variety j with note i

μ is the average level of $\log(\lambda_{ij})$

α_i is the effect of notes i

β_j is the effect of variety j

γ_{ij} is the interaction between variety j and note i

Method C

5. For nominal characteristics, the model can be reformulated as a model where we instead treat the data as if they are multinomial-distributed with a probabilities π_{ij} (the probability of variety j having note i). In this model we assume that a generalized logit of the probabilities (also called log odds ratios) can be regarded as a linear model. In the example here we take the last note as the basic note. Other notes could be used as basic model, but this will not change the comparisons of variety pairs (but will change the interpretation of the estimated parameters). The model can be formulated as:

$(Y_{1j}, Y_{2j}, Y_{3j}, ..., Y_{nj})$ are multinomial distributed with parameters $(\pi_{1j}, \pi_{2j}, \pi_{3j}, ..., \pi_{nj})$

$$\log\left(\frac{\pi_{ij}}{\pi_{nj}}\right) = \mu_i + \beta_{ij}$$

where

Y_{ij} is the number of plants for variety j with note i

μ_i is the effect of note i $(i = 1, 2, 3, ..., n-1)$

β_{ij} is the effect of variety j for note i $(i = 1, 2, 3, ..., n-1,\ j = 1, 2, 3, ..., v)$

6. The parameters μ and β_{ij} can be used to estimate the parameters of the multinomial distributions, π_{ij} and the differences β_{ij}-β_{il} can be used to quantify the difference between variety j and variety l.

7. The assumptions for validity of the tests are as for the two previous methods: The χ^2 are usually good if no more than 20% of the expected number of plants for each combination of note and variety are less than 5 and none are below 1.

Method D

8. For ordinal characteristics, the model can usually be modified and simplified. Here this is done by using cumulative logit instead of the generalized logit in method C and using a common effect of all notes for each variety instead of one for each note. Then the model may be written as:

$(Y_{1j}, Y_{2j}, Y_{3j}, ..., Y_{nj})$ are multinomial distributed with parameters $(\pi_{1j}, \pi_{2j}, \pi_{3j}, ..., \pi_{nj})$

$$\log\left(\frac{\sum_{l=1}^{i}\pi_{ij}}{\sum_{l=i+1}^{n}\pi_{ij}}\right) = \mu_i + \beta_j \quad \left(\text{for each combination of note } i = 1,2,3,...n-1, \text{ and vareity } j\right)$$

where

Y_{ij} is the number of plants for variety j with note i

μ_i is the effect of note i $(i = 1,2,3,...,n-1)$

β_j is the effect of variety j $(j = 1,2,3,...,v)$

9. The parameters μ and β_j can be used to estimate the parameters of the multinomial distributions, π_{ij} and the differences β_j-β_l can be used to quantify the difference between variety j and variety l.

Analyses taking into account sources of variation other than sampling

10. If more than one set of results is recorded for each variety, the multinomial model for both the nominal-scaled characteristics and the ordinal-scaled characteristics can be extended to take into account additional variation such as “soil fertility”, variation from year-to-year and uncertainty in the recording. This means that, if we have repeated counts for each variety, such as counts from each of a number of replicates in a given trial or counts for each of at least two years, we can take such additional variation into account. Here the model for the situation with recordings in more than one year will be described (see examples of such data in table 4 and 6).

Method C_{COYD}

11. Model C (for nominal scale) may be extended by adding two additional effects: the fixed effect of year and the random effect of year-by-variety (for each of the n-1 levels of the note). The random effect is assumed to be normally distributed. Such a model will then be analogical to the COYD method for continuous (normally distributed) data. The modified model C may then be written as:

$(Y_{1jk}, Y_{2jk}, Y_{3jk}, ..., Y_{njk})$ are multinomial distributed with parameters $(\pi_{1jk}, \pi_{2jk}, \pi_{3jk}, ..., \pi_{njk})$

$$\log\left(\frac{\pi_{ijk}}{\pi_{njk}}\right) = \mu_i + \beta_{ij} + \delta_{ik} + E_{ijk}$$

where

Y_{ijk} is the number of plants for variety j in year k with note i

μ_i is the effect of note i $(i = 1, 2, 3, ..., n-1)$

β_{ij} is the effect of variety j for note i $(i = 1, 2, 3, ..., n-1,\ j = 1, 2, 3, ..., v)$

δ_{ik} is the effect of year k for note i $(i = 1, 2, 3, ..., n-1,\ k = 1, ..., y)$

E_{ijk} is the random effect of vareity j in year k for note i $(i = 1, 2, 3, ..., n-1,\ j = 1, 2, 3, ..., v, k = 1, ..., y)$

E_{ijk} is assumed to be normally distributed with mean zero and a constant variance for each of the $n-1$ levels of the note, i.e. $E_{ijk} \sim N(0, \sigma_i^2)$

Method D_{COYD}

12. Similarly, model D (for ordinal scale) may be extended by adding two effects: fixed effect of year and the random effect of year-by-variety. The random effects are also here assumed to be normally distributed. Such a model will similarly be analogical to the COYD method for continuous (normally distributed) data. The modified model D may then be written as:

$(Y_{1jk}, Y_{2jk}, Y_{3jk}, ..., Y_{njk})$ are multinomial distributed with parameters $(\pi_{1jk}, \pi_{2jk}, \pi_{3jk}, ..., \pi_{njk})$

$$\log\left(\frac{\sum_{L=1}^{i} \pi_{ijk}}{\sum_{l=i+1}^{n} \pi_{ijk}}\right) = \mu_i + \beta_j + \delta_k + E_{jk}$$

where

Y_{ijk} is the number of plants for variety j in year k with note i

μ_i is the effect of note i $(i = 1, 2, 3, ..., n-1)$

β_j is the effect of variety j $(j = 1, 2, 3, ..., v)$

δ_k is the effect of year k $(k = 1, ..., y)$

E_{jk} is the random effect of vareity j in year k. E_{jk} is asumed to be normally distributed with zero mean and constant variance, i.e. $E_{jk} \sim N(0, \sigma^2)$

Examples

13. The first data to look at are those presented at the twenty-sixth session of the TWC, held in Jeju, Republic of Korea, September 2 to 5, 2008, (see document TWC/26/11) and then later we look at two other datasets, where some varieties were recorded in two years.

Analyses of data on Coletotrichum crown rot in lucerne

14. The data presented in document TWC/26/11 were collected in Australia for two generations of a candidate variety and 4 reference varieties. Each variety/generation was scored on a 1-5 scale with note 1 being resistant and note 5 being susceptible. The number of individuals in each combination of variety/generation and note are shown in table 1.

Table 1: Number of individual with each note for Coletotrium crown rot in 6 varieties/generations

Note	Candidate Generation 1	Candidate Generation 2	Reference 1	Reference 2	Reference 3	Reference 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

15. We first analyze the data using the method based on contingency tables. Doing so, we get a χ^2 test statistic of 96.7 with 20 degrees of freedom for an overall test of independent distribution for all varieties (table 2, top-left value), which very clearly rejects the null-hypothesis of independent distributions. Then, as examples, 3 pair-wise tests were performed by analyzing only the actual pair of varieties to be compared. Firstly, the distributions for the two generations of the candidate were compared. This resulted in a χ^2 test statistic of 1.9 with 4 degrees of freedom, which was far from being significant (P=0.7554), so we accept the hypothesis of same distribution for those two generations. Next, we compared the first generation of the candidate with reference variety number 1 and finally we compared the average distribution of two generations of the candidate with reference variety number 1 (by summing the numbers for the two generations of the candidate and comparing the distribution of these sums with those of reference variety 1). Both those tests were highly significant (two bottom rows of method A in table 2), so we conclude that the distribution for the candidate is different for that of reference variety no 1.

16. For methods B, C and D all data in table 1 were analyzed jointly. The pair-wise tests in those analyses were performed by setting up contrasts for each of the three comparisons. Methods B and C gave slightly different results to method A, because the χ^2 test statistic in method A is based on Pearsons χ^2 test statistic, whereas that of methods B, C (and D) is based on likelihood ration χ^2 test statistic. However, the conclusions are the same. Method D, which is the most appropriate method here, gives the same conclusions as the other tests, but the significance of the pair-wise tests between the candidate and reference variety number 1 was somewhat stronger – showing that this method is more powerful than the others (because the information about the ordering of the notes is built in).

Table 2: Chi-square and probability of rejecting some 0-hypotheses using different statistical methods

Comparison	Method*							
	A		B		C		D	
	$\chi^2_{(DF)}$	$P(\chi^2>c)$	$\chi^2_{(DF)}$	$P(\chi^2>c)$	$\chi^2_{(DF)}$	$P(\chi^2>c)$	$\chi^2_{(DF)}$	$P(\chi^2>c)$
Varieties	$96.7_{(20)}$	<.0001	$74.2_{(20)}$	<.0001	$74.2_{(20)}$	<.0001	$68.9_{(5)}$	<.0001
CG1 vs. CG2	$1.9_{(4)}$	0.7554	$1.8_{(4)}$	0.7799	$1.8_{(4)}$	0.7795	$0.9_{(1)}$	0.9094
CG1 vs. Ref1	$22.8_{(4)}$	0.0001	$18.2_{(4)}$	0.0011	$18.2_{(4)}$	0.0011	$18.0_{(1)}$	<.0001
CG. vs. Ref1	$28.5_{(4)}$	<.0001	$24.7_{(4)}$	<.0001	$24.7_{(4)}$	<.0001	$25.0_{(1)}$	<.0001

*) A=Contingency table: testing for independent distribution
B=Poisson model with main effects of notes and variety/generation and interaction: testing for interaction
C=Multinomial model based on odds ratio assuming nominal notes
D=Multinomial model based on odds ratio assuming ordinal notes

17. In method D the model was simplified by assuming that the variety effect on the log odds ratios was the same for all notes. To see if this was reasonable, some measures of model fit was calculated. Akaike's information criterion, AIC was calculated (table 3). As the value for method D is less than that for method C it can be concluded that model D fits the data at least as well as model C.

Table 3: Goodness of fit statistics of the models for analyzing Coletotrichum crown root in Lucerne using methods C and D

Comparison	Method*	
	C	D
No of parameters	1+24	1+9
AIC	701	692

*) C=Multinomial model based on odds ratio assuming nominal notes
D=Multinomial model based on odds ratio assuming ordinal notes

Analyses of data on colors of hypocotyls in sugar beets

18. In order to demonstrate the method of C_{COYD} it was necessary to use a dataset where the same information was collected in more years.

Table 4: Number of individual plants with each note for hypocotyls colors for varieties of sugar beet

Year	Variety	Colour				
		1 Green	2 White	3 Pink 4 Red 5 Dark red	7 Orange	Total
2007	A	30	9	15	46	100
	B	5	9	48	38	100
	C	0	17	31	52	100
	D	1	7	71	21	100
	E	0	5	80	20	105
	F	30	0	30	40	100
	G	33	12	16	39	100
	H	72	2	3	23	100
	I	3	4	37	56	100
	J	82	2	7	9	100
	K	52	16	0	32	100
	L	50	17	5	28	100
	M	0	12	58	30	100
	N	0	9	74	17	100
	O	0	12	58	30	100
	P	25	0	17	58	100
	Q	0	0	65	35	100
	R	0	0	75	25	100
	S	0	6	53	41	100
	T	83	5	3	9	100
	U	54	12	3	31	100
	V	0	6	71	23	100
2008	A	21	1	25	53	100
	B	9	5	46	40	100
	C	3	12	35	50	100
	D	0	8	77	15	100
	E	3	0	72	25	100
	F	28	4	30	38	100
	G	25	2	24	49	100
	H	76	4	2	18	100
	I	2	2	29	67	100
	J	82	0	5	13	100
	K	7	33	44	16	100
	L	37	9	12	42	100
	M	0	2	56	42	100
	N	0	8	69	23	100
	O	0	10	65	25	100
	P	22	10	11	57	100
	Q	0	10	64	26	100
	R	0	0	55	45	100
	S	0	1	61	38	100
	T	92	1	1	6	100
	U	30	13	4	53	100
	V	0	18	63	19	100

19. The data was collected in Denmark for a number of candidate variety and reference varieties in two years. In each year, approximately 50 hypocotyls were selected from each of two replicates and their color was recorded using 7 notes (1 green, 2 white, 3 pink, 4 red, 5 dark red, 6 yellow and 7 orange). Among the varieties present in both years, 22 varieties were selected for demonstration purpose. As the number of hypocotyls recorded as yellow were zero for all varieties, this note were left out. The number of hypocotyls recorded as red and dark red was very few, so they were merged with pink and treated as one note. The numbers in each of the two replicates were summed in order to form a year-by-variety table. The number of individuals in each combination of year, variety and note are shown in table 4.

20. Those data were analyzed by two methods; method A, which does not take into account any additional variation from year-to-year (or variation from replicate-to-replicate); and method $C_{COYD,}$ which does that by including a random effect for the interaction year-by-variety for each of the n-1 first levels of the characteristic. Variety A and B are treated as candidates, while the others are treated as reference varieties in all analyses and the pair-wise tests between the two candidates and two reference varieties (C and G) are shown as examples. The results are summarized in table 5. For the method C_{COYD}, the tests were performed as F-tests because the test included the variety-by-year effects, which was estimated and based on a limited number of degrees of freedom (as in the COYD tests for continuous variables). The overall test for differences between varieties was highly significant for both methods, whereas for the pair-wise tests the C_{COYD} methods gave less significance and, in some cases, different conclusions. The reason for these differences was that method A did not take into account the variation caused by other sources than random sampling. The additional variation from year-to-year is illustrated in figure 1. For example, variety K had about 50% of the recorded plants as note 1 in 2007, but only about 10% of the recorded plants had this note in 2008.

Figure 1: Percent hypocotyls in each note for each of 22 varieties in 2007 and 2008

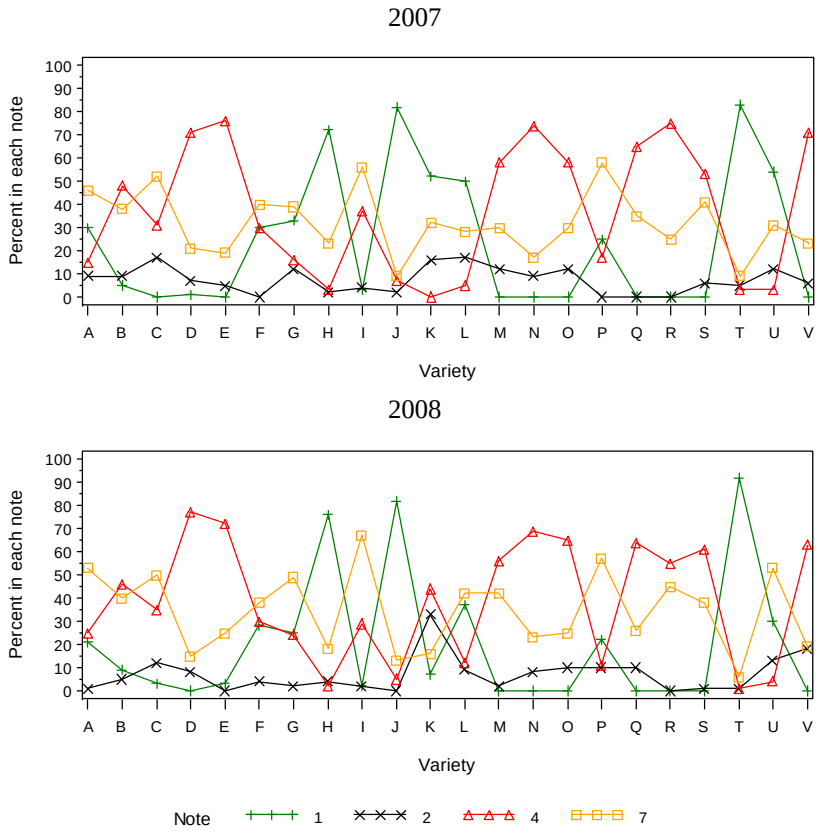


Table 5: Chi-square and probability of rejecting some 0-hypotheses using different statistical methods

Comparison	Method*			
	A		C_{COY-D}	
	$\chi^2_{(DF)}$	$P(\chi^2>c)$	$F_{(ndf,ddf)}$	P(F>c)
Varieties	$2785_{(63)}$	<.0001	$21.76_{(63,22)}$	<.0001
C_A vs. R_C	$58_{(3)}$	<.0001	$5.72_{(3,27)}$	0.0036
C_A vs. R_G	$1.8_{(3)}$	0.6230	$0.15_{(3,16)}$	0.9259
C_B vs. R_C	$20_{(3)}$	<.0001	$2.17_{(3,25)}$	0.1165
C_B vs. R_G	$49_{(3)}$	<.0001	$3.14_{(3,19)}$	0.0492

*) A=Contingency table: testing for independent distribution
C_{COY-D}=Multinomial model based on odds ratio assuming nominal notes

Analyses of data on intensity of anthocyanin coloration on coleoptiles in winter wheat

21. In order to demonstrate the method of D_{COYD} it was necessary to use a dataset where the same information was collected in more years using an ordinal scale. As an example, data collected in Denmark for a number of candidate variety and reference varieties in two years were used. In each year approximately 100 coleoptiles were selected and their anthocyanin coloration was recorded using 5 notes (1 absent or very weak, 3 weak, 5 medium, 7 strong and 9 very strong). Ten varieties present in both years were used. The number of individuals in each combination of year, variety and note are shown in table 6.

Table 6: Number of individual plants with each note for anthocyanin coloration on coleoptiles for some varieties in winter wheat

Year	Variety	Note					
		1 absent or very weak	3 weak	5 medium	7 strong	9 very strong	Total
2007	A	98	1	0	0	0	99
	B	4	14	178	0	0	196
	C	6	32	56	0	0	94
	D	1	5	75	17	1	99
	E	84	106	3	0	0	193
	F	96	4	0	0	0	100
	G	96	4	0	0	0	100
	H	77	23	0	0	0	100
	I	8	15	55	4	0	82
	J	95	3	2	0	0	100
2008	A	86	3	0	0	0	89
	B	14	65	20	0	0	99
	C	0	6	83	4	0	93
	D	4	13	82	1	0	100
	E	62	19	0	0	0	81
	F	100	0	0	0	0	100
	G	100	0	0	0	0	100
	H	84	16	0	0	0	100
	I	4	16	69	1	0	90
	J	93	0	0	0	0	93

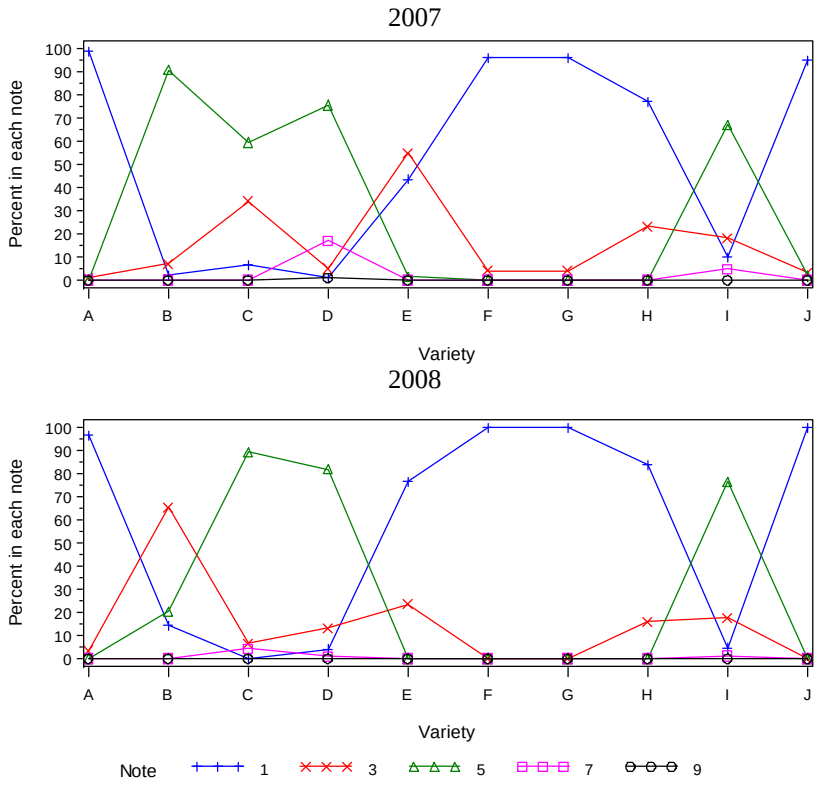
22. The data were analyzed by method A and D_{COYD}. In both analyses, varieties A and B were treated as candidates, while the others were treated as reference varieties. Also here, large differences were found between the two methods (table 7). The largest difference was found for the comparison between candidate variety B and reference variety D, where method A yields a very strong significance (<0.0001) while method D_{COYD} concluded that this pair was not significantly different (P≈13%). Again, the reason was that method A only took into account the variation caused by random sampling. There was clearly some additional variations from year-to-year (figure 2) where, for example variety C, about 35% of the recorded plants had note 3 in 2007, but only about 6% of the recorded plants had this note in 2008.

Table 7: Chi-square and probability of rejecting some 0-hypotheses using different statistical methods

Comparison	Method*			
	A		D_{COY-D}	
	$\chi^2_{(DF)}$	$P(\chi^2>c)$	$F_{(ndf,ddf)}$	P(F>c)
Varieties	$2000_{(36)}$	<.0001	$13.86_{(9,9)}$	0.0003
C_A vs. R_C	$337_{(3)}$	<.0001	$30.68_{(1,9)}$	0.0004
C_A vs. R_D	$354_{(4)}$	<.0001	$40.37_{(1,9)}$	0.0001
C_B vs. R_C	$11_{(3)}$	0.0114	$0.58_{(1,9)}$	0.4648
C_B vs. R_D	$53_{(4)}$	<.0001	$2.73_{(1,9)}$	0.1327

*) A=Contingency table: testing for independent distribution
D_{COY-D}=Multinomial model based on odds ratio assuming ordinal notes

Figure 2: Percent coleoptiles in each note for each of 10 varieties in two years



Discussion

23. The previous sections describe some methods that are mainly based on generalized mixed models. All methods based on generalized mixed models are based on maximum-likelihood or pseudo-likelihood estimation. The methods are iterative because no explicit equations can be set up for estimating the parameters. Because of the iterative method the analyses take considerably more time than traditional analyses of variance methods and problems in having the algorithms to converge may occur (see below). More details on the methods may be found in statistical text books such as McCullagh and Nelder (1989). The analyses shown here were performed using the procedure Glimmix of SAS (SAS, 2008), but similar methods are available in other statistical packages such as GENSTAT and R.

24. In some cases there may be difficulties in applying the methods. Varieties which show no variation (all plants having the same note) may make such analyses impossible because then log odds ratio for all notes will approach either $+\infty$ or $-\infty$ ($\pm$infinity), and such varieties may have to be left out of the analyses and compared to the candidates (or references) using other methods. Also, other types of extremes, such as notes that are present in only a few plants, may make it difficult to apply the methods (especially method C and C_{COYD}) and here some *ad hoc* modifications or alternative methods may be needed.

25. Other methods exist that could be used for analyzing such data. Those could be methods based on ranks, as described by Van der Laan and Verdoren (1987). They show some non-parametric methods based on ranks that are analogical to classical tests and analyses of variance. However, none of their methods are counterparts of mixed model type

analyses, so care has to be taken if they are used for analyzing data with variation at different levels.

26. The models that include year-by-variety interactions (methods C_{COYD} and D_{COYD} above) gave less significant results than the contingency table method (method A above). The reason for this was that method C_{COYD} and D_{COYD} took into account all types of variation present in the data, whereas method A only takes into account the variation causes by random sampling. If a decision that is consistent over years has to be taken, it is important to take into account all types of variation present in the data. As an example, take variety K (table 4 or figure 1): if the distribution of the notes for this variety in 2007 and 2008 were compared using method A, then hypotheses of independent distribution would have been rejected at the 0.01% level of significance. The differences found here between C_{COYD}/D_{COYD} and A are analogical to those that must be expected if the COYD method for continuous variables were compared with a method using the within-plot variation as error.

References

McCullagh, P., Nelder, J. A. 1989. Generalized Linear models. Second edition. Chapman and Hall. 511 pp

SAS Institute Inc. 2008. SAS/STAT® 9.2. Users Guide. SAS Institute Inc. Cary, NC, USA. 7880 pp (online access: http://support.sas.com/documentation/cdl/en/statug/59654/PDF/default/statug.pdf)

Van der Laan, P.; Verdoren, L.R. 1987. Classical Analysis of Variance Methods and Nonparametric Counterparts. Biom. J. 29, 635-665.

[End of document]