E

UPOV

TWC/28/29
ORIGINAL: English
DATE: June 22, 2010

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Twenty-Eighth Session
Angers, France, June 29 to July 2, 2010

SOME ANALYSES OF MULTINOMIAL CHARACTERISTICS THAT ARE ANALOGOUS TO COYD FOR NORMALLY DISTRIBUTED CHARACTERISTICS

Document prepared by an expert from Denmark

SOME ANALYSES OF MULTINOMIAL CHARACTERISTICS THAT ARE ANALOGOUS TO COYD FOR NORMALLY DISTRIBUTED CHARACTERISTICS

Kristian Kristensen

Department of Genetics and Biotechnology, Faculty of Agricultural sciences,
Aarhus University, Denmark

Introduction

1. Visual observations of characteristics are usually recorded as notes. The note is usually regarded as being either nominal or ordinal and, in such cases, the data may be analyzed by generalized linear mixed models, assuming the data to be distributed according to a multinomial distribution with some additional random effects. The additional random effects are included in order to take into account additional variation that is not caused by random sampling, e.g. soil fertility, climate variation and uncertainty in assigning the note. In a test with data from 2 or 3 years, it is expected that most of such additional variation can be described by a random year×variety effect – as in the COY-D method for normally distributed characteristics. Generalized linear models were presented at the twenty-seventh session of the TWC, held in Alexandria, Virginia, United States of America, June 16 to 19, 2009, together with certain other methods (see document TWC/27/14). The TWC concluded as follows:

> "56. The TWC considered document TWC/27/14, presented by Mr. Kristian Kristensen (Denmark).
>
> […]
>
> 58. The TWC agreed that it would be useful to provide an overview of VS characteristics in UPOV Test Guidelines where the method could be appropriate.
>
> 59. For possible future development, Mr. Kristensen agreed to consider the introduction of an indicator (e.g. F3) for variety to observe variation between years and to consider the possible use of a gamma distribution for the variety-by-year interaction. He also agreed to provide the method with SAS code and to consider how to deal with combining categories where zeros were present in initial categories."

2. As explained above, it was decided to produce a new paper that described the method that resembled the COY-D methods – including a test that resembled the F_3-value used in the COY-D for normally distributed data and to consider the use of gamma distribution in stead of the normal distribution for the year×variety effect. However, it has not been possible to find appropriate software that allow random gamma distributions for the models considered here, so all methods shown here assumes that the random effects are normally distributed.

Methods

3. The methods described here are based on generalized linear mixed models and separate models will be given for ordinal characteristics and nominal characteristics. For more

detailed information on generalized linear mixed models see, for example, McCulloch and Searle (2001).

Ordinal characteristics

4. Documents TWC/27/14 considered a generalized linear mixed model for analyses of ordinal characteristics using the cumulative logit as link function. The model resembled the COY-D method for continuous data by including the year×variety interaction as a random effect. The model can 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. $E_{jk} \sim N(0, \sigma^2)$

5. The parameters μ_i, δ_k 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 and test the difference between variety *j* and variety *l*. To estimate the relative number of plants for each note and variety the following formulas may be used:

First calculate: $\hat{P}_{i,j} = \mu_i + \beta_j + \frac{1}{y}\sum_{k}^{y}\delta_k$ for $i = 1, 2, ..., n-1$ for each variety, j

Then calculate $$\hat{\pi}_{i,j} = \begin{cases} e^{\hat{P}_{i,j}}/(1+e^{\hat{P}_{i,j}}) & \text{for } i = 1 \\ e^{\hat{P}_{i,j}}/(1+e^{\hat{P}_{i,j}}) - e^{\hat{P}_{i-1,j}}/(1+e^{\hat{P}_{i-1,j}}) & \text{for } i = 2, 3, ..., n-1 \\ 1 - e^{\hat{P}_{i-1,j}}/(1+e^{\hat{P}_{i-1,j}}) & \text{for } i = n \end{cases}$$

where

$\hat{\pi}_{i,j}$ is the estimated relative number of plants with note i for variety j

Other terms as defined above.

6. As pointed out in the text of COY-D, for normally distributed characteristics a large year×variety interaction for specific pair of varieties may cause that pair to be distinct because of a very large difference in only one of the years without being different in other years (TGP/8/1 Draft 13 section 3.6.3). To avoid that situation the year×variety interaction for each pair of varieties is compared to the average year×variety effects using the quotient between the Mean square for the interaction of the actual pair of varieties and the average Mean square for all varieties (the F_3 value). It is suggested that similar tests are performed here. Such

F-values can be calculated from the estimates of E_{jk} in a similar way as for the COY-D method for normally distributed characteristics. As in COY-D for normally distributed characteristics this quotient (F_3) can be tested approximately by assuming that the quotient is F-distributed with *y-1* and *(y-1)(v-1)* degrees of freedom. However, this approximate test may not work well with a low number of degrees of freedom (as the interaction for the actual pair of varieties includes both the numerator and denominator of this quotient).

7. For demonstration of the method, we have used the same data as in document TWC/27/14 (reproduced in Table 1).

Table 1: 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

8. Estimates of the relative percent of plants for each note, calculated as $\hat{\pi}_{i,j}$, are shown in Table 2.

Table 2: Estimated percent of plants for each note of each variety

Variety	Note				
	1 absent or very weak	3 weak	5 medium	7 strong	9 very strong
A	97.9	1.9	0.1	0.0	0.0
B	3.9	36.5	59.1	0.6	0.0
C	1.4	17.8	79.1	1.5	0.1
D	0.4	6.1	88.2	5.1	0.2
E	62.9	33.7	3.4	0.0	0.0
F	98.9	1.1	0.1	0.0	0.0
G	98.9	1.1	0.1	0.0	0.0
H	81.0	17.6	1.4	0.0	0.0
I	2.0	23.1	73.8	1.1	0.0
J	98.6	1.3	0.1	0.0	0.0

9. Taking variety A and B to be candidates and the remaining varieties, C, D, … , J, to be reference varieties we proceed to calculate the differences and the P-values for testing the hypothesis of no difference between candidate and reference. The differences and the P-values are shown in Table 3. The calculated F_3 values and their significance are also shown in Table 3.

10. For the data shown here candidate A could be separated from 4 of the other varieties when using a 1% level of significance while candidate B could be separated from 5 of the other varieties. The F_3 values were not significantly larger than 1 for any of the tested variety pairs shown in table 3. The largest F_3 was found for the variety pair B-C which seemed to be caused by a stronger anthocyanin coloration of varieties B and C in 2008 than in 2007. The second largest F_3 was found for the variety pair A-B and here the stronger anthocyanin coloration of variety B in 2008 seemed to be the cause.

Table 3: Differences and F_3 values together with P-values for relevant pairs of varieties

Variety	Candidate A				Candidate B			
	Difference	$P_{Difference}$	F_3	P_{F3}	Difference	$P_{Difference}$	F_3	P_{F3}
A	-	-	-	-	7.06	0.0009	2.47	0.1503
B	7.06	0.0009	2.47	0.1503	-	-	-	-
C	8.11	0.0004	0.38	0.5548	1.04	0.4648	4.78	0.0566
D	9.33	0.0001	1.42	0.2644	2.26	0.1327	0.15	0.7111
E	3.33	0.0471	0.67	0.4353	-3.73	0.0232	0.57	0.4691
F	-0.61	0.7152	1.56	0.2425	-7.68	0.0008	0.10	0.7551
G	-0.61	0.7152	1.56	0.2425	-7.68	0.0008	0.10	0.7551
H	2.41	0.1319	0.21	0.6612	-4.66	0.0079	1.25	0.2920
I	7.77	0.0005	0.03	0.8561	0.71	0.6176	1.92	0.1992
J	-0.40	0.8088	1.68	0.2273	-7.46	0.0009	0.08	0.7882

Nominal characteristics

11. Document TWC/27/14 considered a generalized linear mixed model for analyses of nominal characteristics using the generalized logit as link function. The model resembled the COY-D method for normally distributed characteristics by including the year×variety interaction as a random effect. However, for the nominal characteristics there will be a random effect for each of *n-1* notes and each of these are assumed to be normally distributed. The model can be written as:

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

$$\log\left(\frac{\pi_{ijk}}{\pi_{njk}}\right) = \mu_i + \beta_{ij} + \delta_{ik} + E_{ijk} \qquad \text{for } i = 1, 2, ..., n-1$$

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. $E_{ijk} \sim N(0, \sigma_i^2)$

12. In the formulation above it is assumed that the last note (number *n*) is taken as the reference note in the generalized logit. For improving the performance of the analyses it is recommended to ensure that the note used as the reference is the note that occurs most often (Agresti, 2002). The parameters μ_i, δ_{ik} and β_{ij} can be used to estimate the parameters of the multinomial distributions, π_{ij} and the differences between pairs of varieties can be quantified and tested by estimating the differences between β_{ij}-β_{il} for each of the *n-1* notes. The overall test will be the result of a contrast for each of those notes using a F-test with *n-1* degrees of freedom in the numerator and about *(y-1)(v-1)* to *(n-1)(y-1)(v-1)* degrees of freedom in the denominator. To estimate the relative number of plants for each note and variety the following formulas may be used:

First calculate: $\hat{P}_{ij} = \mu_i + \beta_{ij} + \frac{1}{y}\sum_{k}^{y} \delta_k$ for $i = 1, 2, ..., n-1$ for each variety, j

$$\text{Then calculate } \hat{\pi}_{ij} = \begin{cases} e^{\hat{P}_{ij}} / (1 + \sum_{l=1}^{n-1} e^{\hat{P}_{lj}}) & \text{for } i = 1, 2, ..., n-1 \\ 1 - \sum_{l=1}^{n-1} \hat{\pi}_{lj} & \text{for } i = n \end{cases}$$

where

$\hat{\pi}_{ij}$ is the estimated relative number of plants with note i for variety j

Other terms as defined above.

13. As for the ordinal characteristics, a large year×variety interaction for specific pair of varieties may cause that pair to be distinct because of a very large difference in only one of the years without being different in other years. To avoid that situation the year×variety interaction for each pair of varieties is compared to the average year×variety using the quotient between the Mean square for the interaction of the actual pair of varieties and the average. However, for the nominal characteristics, there will be a year×variety table for each of the first *n-1* notes. At least two methods of testing for a large interaction for a given variety pair may be set:

Method A: The quotient, F_3 may be based on a joint contrast for the interaction of each of the *n-1* notes. This will result in a quotient (F_3) than can be tested approximately by assuming that the quotient is F-distributed with *(n-1)(y-1)* and *(n-1)(y-1)(v-1)* degrees of freedom.

Method B: The quotient, F_3, may be based on the note that gives the largest F_3 for the given pair of varieties. This will result in a quotient (F_3) than can be tested approximately by assuming that the quotient is distributed as largest of *n-1* F-values with *(y-1)* and *(y-1)(v-1)* degrees of freedom. As an approximation the probability of getting at least one F-value to be significant was used here – calculated as $P=1-(1-P_{F3max})^{n-1}$, where P_{F3max} is the probability for the largest of the *n-1* F_3-values.

14. For both methods it was assuming that the interaction effects for different notes were independent.

15. Unfortunately neither of those two methods yields estimates that are independent of the note chosen as the reference. Using either of the two notes (3+4+5 or 7) for which there was a reasonably large number of plants did not change the conclusion for the data used, but this might happen for other datasets (see also figure 3). As for ordinal characteristics, these approximate tests may not work well for a low number of varieties (as the interaction for the actual pair of varieties includes both the numerator and denominator of this quotient).

16. For demonstration of the method, the same data as used in document TWC/27/14 are used (reproduced in table 4). Because some varieties had notes with zero plants in both years, some difficulties were found in having the algorithm meeting the convergence criteria if those varieties were used. Therefore, the varieties M, N, O, Q, R, S and V were excluded from the analyses shown here.

Table 4: Number of individual with each note for hypocotyls colours for some varieties in sugar beets

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

17. Estimates of the relative percent of plants for each note, calculated as $\hat{\pi}_{ij}$ are shown in Table 5.

Table 5: Estimated percent of plants for each note of each variety

Variety	Colour			
	1 Green	2 White	3 Pink 4 Red 5 Dark red	7 Orange
A	25.8	3.9	19.8	50.5
B	7.0	6.8	47.2	39.1
C	1.5	14.3	33.0	51.1
D	0.5	7.5	74.2	17.8
E	1.5	1.8	74.7	22.0
F	29.1	1.7	30.1	39.2
G	29.5	5.6	20.1	44.8
H	74.1	2.9	2.5	20.5
I	2.5	2.9	33.0	61.6
J	82.2	0.9	6.0	11.0
K	27.7	29.3	14.0	29.0
L	44.0	12.7	8.0	35.2
P	23.9	3.4	14.1	58.7
Q	88.0	2.5	2.0	7.5
U	41.7	12.8	3.5	42.0

18. Taking variety A and B to be candidates and the remaining varieties, C, D, … , U, to be reference varieties the F-values and the P-values for testing the hypothesis of no difference between candidate and reference were calculated. The F-values and the P-values are shown in Table 6. The two different F_3-values and their significance are also shown in Table 6.

19. Using the 1% level of significance as a decision rule for comparing the candidates, we found that candidate A was distinct from 7 of the other varieties, while candidate B was distinct from 5 of the other varieties. The largest F_3-values were found for the variety pairs B-K and A-K. This seemed to be caused mainly by variety K, which had many green and zero pink+red+dark red hypocotyls in 2007, but few green and many pink+red+dark red hypocotyls in 2008.

Table 6: Differences and F_3 values together with P-values for relevant pairs of varieties

Variety	Candidate A						Candidate B					
	F	$P_{dif.}$	F_{3A}	P_{F3A}	F_{3B}	P_{F3B}	F	$P_{dif.}$	F_{3A}	P_{F3A}	F_{3B}	P_{F3B}
A	-	-	-	-	-	-	2.34	0.1157	0.50	0.6855	0.95	0.7682
B	2.34	0.1157	0.50	0.6855	0.95	0.7682	-	-	-	-	-	-
C	5.70	0.0062	0.57	0.5829	1.04	0.6749	2.06	0.1432	0.02	0.9826	0.04	0.9883
D	6.29	0.0033	0.50	0.6485	1.49	0.5273	2.05	0.1404	0.42	0.7800	0.73	0.8003
E	5.40	0.0063	0.41	0.6601	0.91	0.6249	1.35	0.2866	0.19	0.8542	0.55	0.8371
F	0.52	0.6757	1.20	0.2671	3.18	0.1848	3.20	0.0522	0.50	0.7097	1.33	0.6461
G	0.16	0.9224	0.01	0.9976	0.02	0.9979	2.79	0.0786	0.46	0.7701	0.95	0.8215
H	6.91	0.0036	0.94	0.4998	1.69	0.4087	14.33	<.0001	0.15	0.9024	0.45	0.9199
I	5.44	0.0073	0.24	0.7018	0.36	0.7495	2.27	0.1143	0.24	0.9500	0.67	0.9558
J	10.36	0.0004	0.19	0.8365	0.36	0.9051	17.65	<.0001	0.18	0.9506	0.35	0.9603
K	2.19	0.1361	3.17	0.0405	5.07	0.1022	4.54	0.0189	4.31	0.0071	7.10	0.0540
L	2.02	0.1621	0.11	0.9719	0.21	0.9571	6.55	0.0051	0.64	0.7790	1.68	0.7654
P	0.21	0.8896	1.79	0.0934	4.80	0.0654	2.67	0.0847	0.92	0.4270	2.44	0.3158
T	13.62	<.0001	0.65	0.7695	1.47	0.8129	21.42	<.0001	0.05	0.9946	0.06	0.9932
U	2.34	0.1202	0.52	0.7387	0.81	0.6791	7.38	0.0027	1.18	0.8181	3.51	0.7334

Discussion and concluding remarks

20. The methods shown here require more computer power than the COY-D method for characteristics that can be assumed to be normally distributed. However, it seems important also to take the year×variety interaction into account when using data that are recorded on the ordinal or nominal scale as failing to do so may yield too many significant comparisons (see e.g. Tables 5 and 7 of document TWC/27/14). The year×variety interaction may be expected to take several sources of variation into account such as true year×variety interaction, variations caused by “soil fertility” and uncertainty in the recording of the notes, which the simpler χ^2-test does not. The most important benefit of taking the year×variety interaction into account is probably that this increases the probability of ensuring a decision that will be consistent over future years.

21. For nominal data, the analyses shown here requires that the expected number in plants in most cells should be at least 5 (as for the χ^2-test). Therefore, it may be necessary to merge some notes before the analyses are carried out or to run the analyses on subsets of varieties. In the analyses shown here, notes 3, 4 and 5 for pink, red and dark red were merged before the analyses were carried out and some varieties (M, N, O, Q, R, S and V) were omitted from the analyses.

22. For ordinal data, the notes with a large number of plants should occur next to each other. If they fall in groups the analyses may not be appropriate. Thus, a variety with 41, 0, 2, 32 and 25 in each of the notes 1, 3, 5, 7 and 9 should not be included in such an analysis (the reason for such an occurrence of plants could be that the variety is in fact a mixture of two genotypes – one with low anthocyanin coloration and one with medium to high anthocyanin coloration). Zeros at one or both ends of the scale should not cause problems as long as most varieties are represented in more than one note.

23. For both sets of data, the largest F_3 for both candidates was found for the same reference variety (varieties B and C for the ordinal characteristic and variety K for the nominal characteristic). This could be caused by a large variation from year-to-year for candidate C

and K. In order to investigate that, the contribution to the overall interaction from each variety was compared using F-values. The results are shown in figure 1 and 2 for the ordinals and nominal characteristic, respectively. In particular for the nominal characteristic, it is clear that variety K varies much more from year to year than the other varieties. It could be questioned weather such a variety should be included in the analyses as such a variety is expected to inflate the year×variety interaction and thus decrease the power of the tests when comparing the candidates to the other reference varieties.

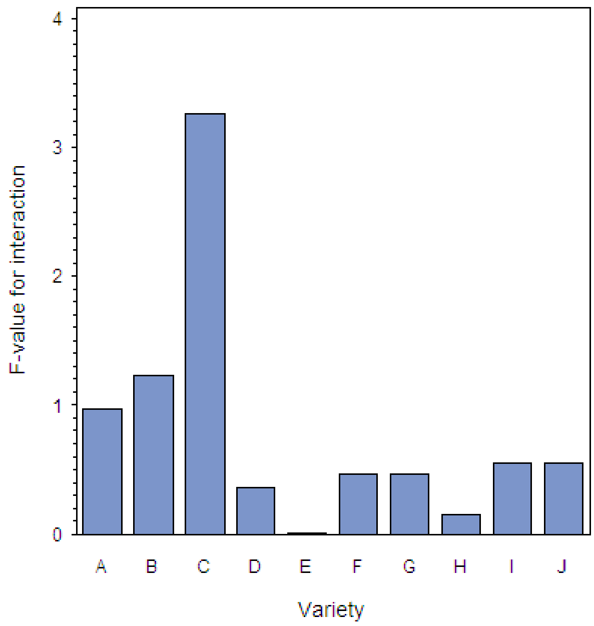

Figure 1 F-values for each varieties contribution to the interaction for ordinal characteristic

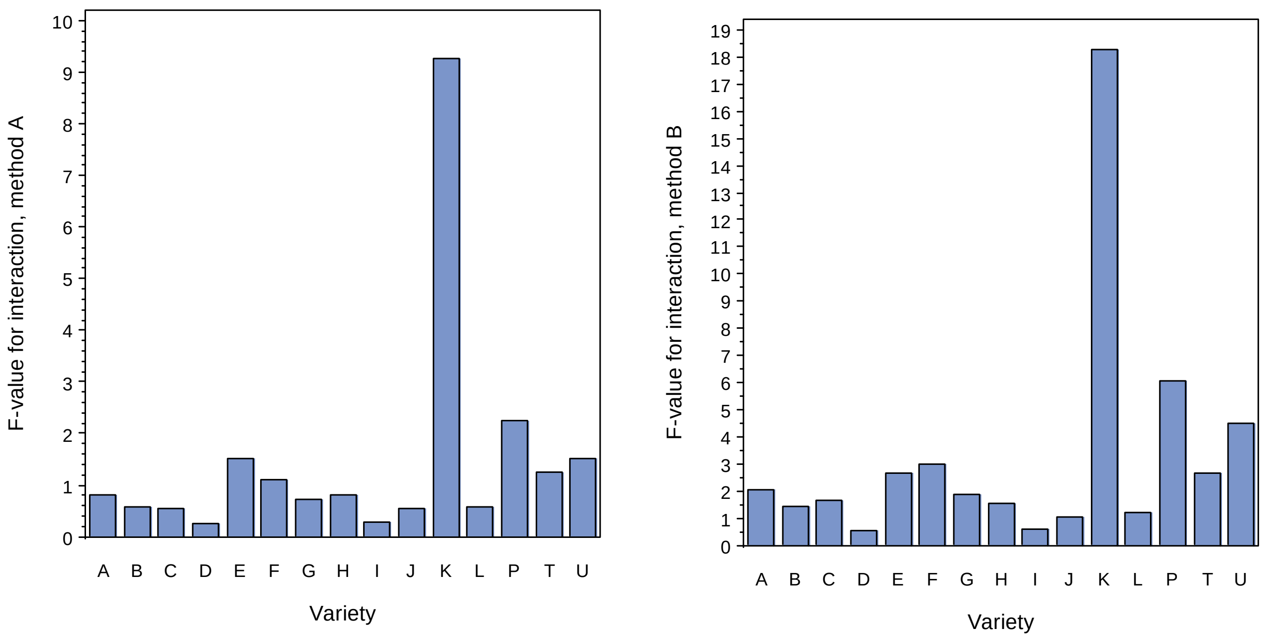

Figure 2: F-values for each varieties contribution to the interaction for nominal characteristic using method A and B

24. As mentioned above, the calculated F_3-values for nominal characteristics depend on which note is used as reference. The F_3-values for two different reference notes are shown in figure 3. The F_3-values based on method B seemed to be more dependent on the reference note than the F_3-values based on method A. The correlation between the interaction effects for different references was 0.96 and 0.94 for methods A and B, respectively.

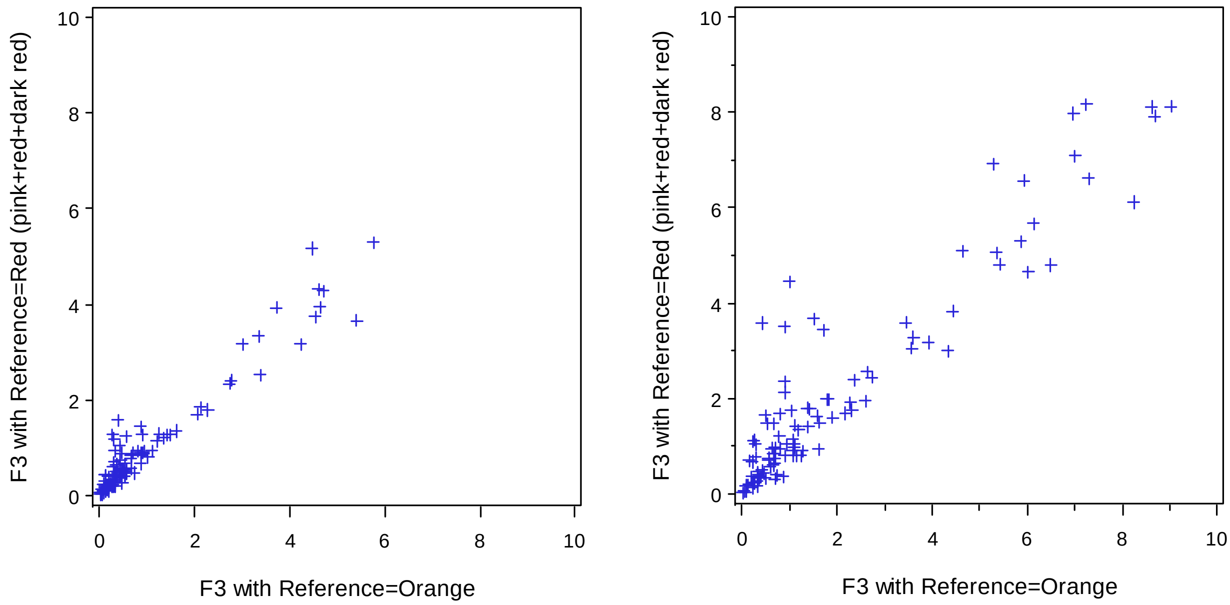

Figure 3 Plot of random effects for all varieties using either method A (left) and method B (right) for two different reference notes

Software

25. The analyses presented here were performed using the procedure GLIMMIX of SAS (SAS Institute, 2008). The models could also be analyzed using, for example, the procedure NLMIXED of SAS. Examples of how this may be done are shown in Appendix A of the book by Agresti (2002) and an updated version may be found on http://www.stat.ufl.edu/~aa/cda/software.html, but this page does not include the more recent procedure GLIMMIX, which is easier to use than the procedure NLMIXED, which Agresti (2002) describes.

26. Other software is also available. Agresti (2002) mention R, S and SPSS. The code for analysing all examples in Agresti (2002) in R or S can be found in Thompson (2009). This manual is available on the WEB (see https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf). However, recently (March 2010) the package "ordinal" has been added to R, which is not described on by Thomson (2009) but seems appropriate for the ordinal analyses described here.

References

Agresti, A., 2002. Categorical data analysis, 2nd edition. Wiley & Sons, Inc. 710 pp.

McCulloch, C.E. and Searle, S.R. 2001. Generalized, Linear, and Mixed Models. John Wiley & Sons, Inc; New York. 325 pp

Thompson, L.A., 2009. R (and S-PLUS) Manual to Accompany Agresti's *Categorical Data Analysis* (2002). 2nd edition (online access: https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf).

SAS Institute Inc. 2008. *SAS® 9.2. Procedures Guide: Statistical Procedures*. SAS Institute Inc., Cary, NC, USA. 492 pp (online access: http://support.sas.com/documentation/cdl/en/procstat/59629/PDF/default/procstat.pdf)

[End of document]