E

TWC/19/7
ORIGINAL: English
DATE: May 11, 2001

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

TECHNICAL WORKING PARTY ON AUTOMATION AND COMPUTER PROGRAMS

Nineteenth Session
Prague, June 4 to 7, 2001

METHODS FOR TESTING UNIFORMITY ON CHARACTERS WHERE THE DATA HAS BEEN BULKED

Document prepared by the experts from Denmark

METHODS FOR TESTING UNIFORMITY ON CHARACTERS WHERE THE DATA HAS BEEN BULKED

Kristian Kristensen

Biometry Research Unit, Danish Institute of Agricultural Sciences,
Post-box 50, DK-8830 Tjele, Denmark,
E-mail: Kristian.kristensen@agrsci.dk

May 2001

Background

In some crops, individual samples are bulked before the measurements are taken. In these cases characters are at present not tested for uniformity. The following paper briefly describes some possible methods to use in order to carry out a test for uniformity in those variables.

Method

At present, the within plot standard deviation between individual plants is used as a measurement of uniformity. The log of those standard deviations plus one are analysed in an over years analysis. This can be formulated as follows:

$Z_{vy} = \log(s_{vy} + 1)$

which are then analysed in the following model

$Z_{vy} = \mu + \alpha_v + \beta_y + E_{vy}$

where

E_{vy} can be assumed to consist of 2 sources of variation, one caused by the random variability when estimating a variance and one caused by the random change of varaibility due to the environment, i.e. (1.1)

$E_{vy} = C_{vy} + D_{vy}$

If the recorded variable is normally distributed then the approximate variance of C_{vy}, D_{vy} and E_{vy} may be written as

$$Var(C_{vy}) = \sigma_C^2 \approx \frac{1}{2\nu}\left(\frac{\sigma_{vy}}{\sigma_{vy}+1}\right)^2 \approx \frac{1}{2\nu}\left(\frac{\sigma}{\sigma+1}\right)^2 \quad \text{(if } \sigma_{vy} \text{ is not too variable)}$$

$Var(D_{vy}) = \sigma_D^2$, say (1.2)

$$Var(E_{vy}) \approx \frac{1}{2\nu}\left(\frac{\sigma}{\sigma+1}\right)^2 + \sigma_D^2$$

where σ_{vy} is the standard deviation of variety *v* in year *y*; s_{vy} is an estimate of σ_{vy} with ν degrees of freedom; σ is the average standard deviation of all varieties

μ, α_v and β_y are fixed effects for the general mean, variety and year, respectively.

The thresholds for rejection and acceptance are calculated as:

$$T = R + LSD$$

$$LSD = t\ s_d = t\ \hat{\sigma}_E \sqrt{\frac{1}{N_y} + \frac{1}{N_y N_r}} \qquad (1.3)$$

where

T is the threshold

R is the mean of the $\log(s_{vy})$ for all reference varieties

t is a t-value taken at the appropriate probability level and degrees of freedom

N_y is the number of years

N_r is the number of reference varieties

The variance components may be calculated and then used to show how much the LSD may be expected to change if the number of degrees of freedom is changed.

The estimates of standard deviations, s_{vy}, in a single year may be formulated as if the following statistical model were used for the recorded character

$$X_{vbp} = \mu_v + B_b + C_{vb} + E_{vbp} \qquad (1.4)$$

where

$v = 1, 2..., N_v; b = 1, 2, ...N_b;\ p = 1, 2, ...N_p$

μ_v is the mean of the character in that experiment for variety v

B_b is the random effect of block b

C_{vb} is the random effect of the plot with variety v in block b

E_{vbp} is the random effect of plant p of variety v in block b

$B_b \square N(0, \sigma_B^2)$

$C_{vb} \square N(0, \sigma_C^2)$

$E_{vbp} \square N(0, \sigma_v^2)$

$B_{b,}$ σ_B^2 and σ_C^2 are assumed to be common to all varieties in the experiment

σ_v^2 depends on the variety

(Editing note: in the electronic version, the boxes in equation (1.4) should be replaced by a tilde.)

If just a single variety is analysed then this model results in the following analysis of variance table:

Source	Df	E(MS)
Block+ Plot	(N_b-1)	$\sigma_v^2+N_p(\sigma_B^2+\sigma_C^2)$
Residual=Plant	$N_b(N_p-1)$	σ_v^2

Now if all plants are bulked without noticing from which block and plot they originated then the analysis of variance table for a single variety becomes:

Source	Df	E(MS)
Residual=Mix of all effects	N_bN_p-1	σ_v^2+(N_b -1)/(N_b-1/N_p)(σ_B^2+σ_C^2)

If *k* random sub-samples, each of *l* plants, are taken from such a bulk then the variance of the mean of these will be:

$$Var\ \bar{X}_{vs} = \frac{1}{l}\sigma_v^2 + \frac{1}{l}\frac{N_b - 1}{N_b - 1/N_p}(\sigma_B^2 + \sigma_C^2) \qquad (1.5)$$

Collecting all these variances from two or three years we may set up an additive model for the variances

$$Z_{vy} = \mu + \alpha_v + \beta_y + E_{vy}$$

where

$Z_{vy} = Var(\bar{X}_{vs})$ is the variance of the k subsample means for variety v in year y (1.6)

μ, α_v and β_y parameters describing the effects of the general mean, varieties and years

E_{vy} is the random error term that may be expected to be approximately gamma distributed

Note that the variance components for blocks and plots are assumed to be identical for all varieties in a certain year. Therefore, the effect of these variance components will be absorbed by and confounded with the year effect. The additive model assumes that the size, *l*, of the sub-samples is the same in all years.

It may be suggested to analyse such variances in a generalised linear model (for more details on generalised linear models see e.g. McCullagh, P. & Nelder, J. A. 1989). It is suggested to use the gamma distribution in combination with the identity link. Compared to a situation where all individual plants are recorded there will be some loss of information:

- When bulking is done, fewer degrees of freedom are available for estimation of the variance.

- When bulking is done without knowing from which plot the individuals come, then the noise may be expected to be increased because the ratio of plants from different plots may vary from sub-sample to sub-sample.

This means that a candidate variety needs to be more variable before the new candidate becomes significantly larger than the mean of the reference varieties.

If the sub-samples can be formed by plants from the same plot then the results become much simpler. The variance between the means of *k* sub-samples then becomes

$$Var(\bar{X}_{vs}) = \frac{1}{l}\sigma_v^2 \qquad (1.7)$$

and the variances can be analysed exactly as for characters where all plants are recorded individually. The only difference is that the variance is scaled by a factor that will be absorbed by and confounded with the year effect when the logarithms are analysed, and that the number of degrees of freedom behind the variance is reduced (and therefore, less precision must be expected).

The statistical calculations shown in the following sections were carried out using the procedures Mixed and Genmod of SAS (SAS Institute, 1997). The data generated for simulations was also generated using SAS.

Results

The model and method described in equation (1.1) and (1.2) was used to judge the effect of bulking several plants in each plot before a mean response was recorded. The following data were used:

Table 1. Summary of data and variables.

Crop	N_v	N_b	N_p	Variable	Variable description
Peas	141	2	10	12	Stem: length
				31	Stipule: length
				32	Stipule: width
				48	Pod: length
				49	Pod: width
Ryegrass, 4n	41	3	20	10	Flag leaf: length
				11	Flag leaf: width
				12	Stem: length
				13	Inflorescences: length
Sugarbeets, 2n	70	4	15	1	Leaf: length incl. petiole
				2	Leaf: length
				5	Petiole: width
				12	Root: length
Sugarbeets, 3n	85	4	15	1	Leaf: length incl. petiole
				2	Leaf: length
				5	Petiole: width
				12	Root: length

Table 2. Effects of bulking the the samples down to 10, 6 or 2 degrees of freedom per variety

Crop	Vari-able	Mean of log(s+1)	Residual Variance	Var. components		Actual std_{diff}	Relative std. on difference			
				Df	V×Y		act.df	10 df	6 df	2 df
Peas	12	1.937	0.03313	0.03313	0.01278	0.129	1.00	1.22	1.49	2.43
	31	1.895	0.02631	0.02631	0.00626	0.115	1.00	1.27	1.59	2.66
	32	1.557	0.02370	0.02370	0.00640	0.109	1.00	1.26	1.57	2.62
	48	1.557	0.01672	0.01672	-0.00058	0.092	1.02	1.36	1.76	3.05
	49	0.465	0.00592	0.00592	0.00208	0.055	1.00	1.23	1.52	2.49
Rye-grass, 4n	10	3.549	0.01200	0.01200	0.00373	0.078	1.00	2.06	2.62	4.47
	11	0.736	0.00423	0.00423	0.00185	0.047	1.00	1.91	2.41	4.06
	12	2.253	0.00957	0.00957	0.00254	0.070	1.00	2.11	2.69	4.60
	13	3.644	0.00792	0.00792	-0.00040	0.064	1.02	2.45	3.16	5.47
Sugar-beets 2n	1	3.616	0.00844	0.00844	-0.00001	0.065	1.00	2.37	3.06	5.30
	2	3.482	0.01240	0.01240	0.00401	0.079	1.00	2.03	2.58	4.39
	5	1.214	0.00694	0.00694	0.00253	0.059	1.00	1.98	2.51	4.26
	12	3.556	0.00746	0.00746	-0.00097	0.062	1.06	2.51	3.25	5.62
Sugar-beets 3n	1	3.564	0.01428	0.01428	0.00585	0.085	1.00	1.93	2.43	4.12
	2	3.447	0.01041	0.01041	0.00204	0.073	1.00	2.17	2.77	4.77
	5	1.258	0.00565	0.00565	0.00108	0.053	1.00	2.17	2.78	4.78
	12	3.603	0.00724	0.00724	-0.00121	0.061	1.08	2.56	3.30	5.72

The results are shown in Table 2 and show that the variability caused by the estimation of the standard deviation was relatively large for all examined characters whereas the random variability caused by the environment seem smaller. Therefore, the effect of bulking is relatively large for these characters. When having only 2 sub-samples in each of 2 plots the standard deviation on the difference between the mean of the reference varieties and a new candidate variety may increase by a factor 5 or more.

In order to check how much the power may decrease if the records of sub-samples from a bulked sample are used, instead of making records on each individual plant, a small simulation study was carried out. Data were generated using the model in formula (1.4) for each of 2 years. The data were assumed to consist of 30 varieties (of which 29 were reference varieties) grown in trials with 3 randomised blocks and with 20 plants in each plot. The standard deviation of all reference varieties was assumed to be 1 and that of the candidate, σ_n, was varied from 1 to 10. The simulations were carried out for different values of the other two variance components and of the number of sub-samples. For each case 1000 independent sets of data were generated and the percent of sets were the candidate variety were rejected at the 5%, 1% and 0.1% level were recorded.

When only 2 or 4 sub-samples vas taken the estimated variances had only 1 or 3 degrees of freedom. In those cases the software used had difficulties in converging. When the variances were estimated with only 1 degree of freedom then the computations converged in less than 10% of the sets. There was a clear tendency that the percent of sets where the calculations did not converge increased as the variance of the new candidate increased.

Therefore the results should be taken as only provisional when the degree of freedom low. The results are shown in Table 3.

It should be noted that if no bulking is done, similar simulations demonstrated that the probability of rejecting the candidate variety was 100% at all three levels in all sets when σ_n was greater than or equal to 2. When σ_n was equal to 1 and no bulking was done then the probability of rejecting the candidate variety was 4.9%, 1.0% and 0.1% at the three levels (using 4000 simulations).

Table 3. Percent of sets where the candidate variety was rejected at the 5%, 1% or 0.1% level using the gamma distribution in a generalized linear model. The results are given for different combinations of variance components, number of sub-samples and degree of uniformity (the standard deviation of the reference varieties is 1 while that of the new candidate is given by σ_n).

σ_B	σ_C	σ_n	Number of sub-samples, *k*, and number of plants per bulk, *l*											
			10 á 6, ν=9			6 á 10, ν=5			4 á 15, ν=3*			2 á 30, ν=1*		
			5%	1%	.1%	5%	1%	.1%	5%	1%	.1%	5%	1%	.1%
0	0	1	5	2	0	4	1	0	3	0	0	0	0	0
0	0	2	97	89	76	82	67	47	63	44	21	6	0	0
0	0	3	100	100	99	99	96	91	90	80	57	22	8	0
0	0	5	100	100	100	100	99	98	96	91	76	53	19	0
0	0	10	100	100	100	100	100	100	100	100	99	61	44	0
0	0.5	2	94	84	65	77	62	37	62	42	18	10	0	0
0.5	0	2	94	86	68	77	61	40	57	39	18	12	0	0
0.5	0.5	2	88	74	50	68	49	26	53	35	15	8	3	1

*) problems with convergence (see text)

All sets were also analysed using the present method, COY-U. The results are shown in Table 4. When 6 or 10 sub-samples were taken then the results found by the generalised linear model and the COY-U method are very much the same. For smaller numbers of sub-samples, more sets where the candidate variety are shown to have a larger variance are found by the present method than by the method using the generalized linear model. This may be partly or fully explained by the above-mentioned problems with convergence. For the present simulations where the number of plants, *l*, in each sub-sample and the two variance components for block and plot (σ_B and σ_C) are the same in both years, both these methods should also be applicable. However, if *l*, σ_B or σ_C vary from year to year the present method may not be applicable - unless σ_B and σ_C are both zero or the sub-samples are formed within plots.

Table 4 Percent of sets where the candidate variety was rejected at the 5%, 1% or 0.1% levels using the COY-U method. The results are given for different combinations of variance components, number of sub-samples and degree of uniformity (the standard deviation of the reference varieties is 1 while that of the new candidate is given by σ_n).

σ_B	σ_C	σ_n	Number if bulk samples, *k*, and number of plants per bulk, *l*											
			10 á 6, ν=9			6 á 10, ν=5			4 á 15, ν=3			2 á 30, ν=1		
			5%	1%	.1%	5%	1%	.1%	5%	1%	.1%	5%	1%	.1%
0	0	1	5	1	0	5	1	0	5	1	0	4	1	0
0	0	2	98	92	79	84	72	52	70	52	32	39	25	15
0	0	3	100	100	100	99	98	93	93	88	76	61	49	33
0	0	5	100	100	100	100	100	99	99	97	92	76	67	52
0	0	10	100	100	100	100	100	100	100	100	100	96	93	88
0	0.5	2	95	86	68	81	65	43	65	48	27	35	21	10
0.5	0	2	95	87	69	79	64	44	61	44	26	32	19	11
0.5	0.5	2	90	77	54	72	53	32	57	39	22	28	18	8

Example

In peas, the kernel weight is recorded for each variety in two sub-samples from each of two plots. Each sample consists of 200 seeds and the weight of these - multiplied by 5 - gives the weight of 1000 kernels. In each plot a standard deviation with one degree of freedom is calculated, which is then pooled over blocks to give one standard deviation with 2 degree of freedom for each variety. These standard deviations for all varieties recorded in both 1999 and 2000 were then analysed both by the present method and the method using a generalized linear model. A total of 148 varieties were present in both years (140 reference varieties and 8 new candidate varieties).

The overall test for differences among varieties was not significant (P=0.21) when using the COY-U method, but significant (P<0.0001) when using the generalised linear model. This very different result seems to be caused by just one variety, where the standard deviations in the two years were 2.0 and 13.5. When this variety was left out both methods yielded similar results (the P-values were 0.34 and 0.44, respectively).

None of the 8 candidate varieties were rejected because of non-uniformity (not even at the 10% level). The results are shown in Table 5.

Table 5 Estimates of uniformity for 8 candidate varieties on weight of 1000 kernels when compared to the mean of 139 reference varieties.

Candidate variety	Estimate of uniformity for candidate varieties	
	Present Method, COY-U, log(sd+1)	Generalized linear model, sd^2
H	0.64	1.27
D	0.67	0.95
C	0.72	1.18
G	0.74	1.20
A	0.82	1.63
B	0.94	2.45
E	0.95	2.73
F	1.10	4.05
References	0.93	2.35

Discussions and Conclusions

There are two important differences to note when using the generalised linear model instead of the present COY-U method.

- The reference varieties are weighted differently when calculating the standard to compare the candidates against. In the COY-U method, the mean of the reference varieties are based on the *log(sd+1)* which has approximately equal variances, whereas in the generalised linear model the mean of the reference varieties are based on sd^2 which do not have equal variances. This means that a few reference varieties with large variances may increase the mean of these more when using the generalised linear model than when using the present method.

- In the calculations shown, there is no adjustment for dependence between *sd* and the mean when using the generalized method. However, it may be possible to modify the method to take such possible dependence into account.

When using the generalised linear model it is possible to test for uniformity in bulked samples under some assumed conditions. The more important of these are that the observations are normally distributed and that the change of the within plot variances from year to year are additive. In theory it is possible to test for uniformity if at least 2 sub-samples are drawn from the bulk and recorded. However, the simulations show that with only 2 sub-samples it may be difficult to get the algorithm to converge and that the degree of non-uniformity must be large in order to be reasonable sure to detect non-uniformity. In all cases some loss of power may be expected when testing for uniformity in sub-samples from bulked samples.

If the sub-samples are formed within plots then the present COY-U method can still be used.

References

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

SAS. 1997. SAS/STAT Software: Changes and Enhancements through Release 6.12. SAS Institute Inc. NC, USA. 1162 pp.

[End of document]