Colloquium Biometricum 42
2012, 25–33

COMPARISON OF FOUR METHODS OF TESTING VARIETAL UNIFORMITY IN DUS TRIALS

Bogna Zawieja[1], **Wiesław Pilarczyk**[1,2]

[1]Department of Mathematical and Statistical Methods,
Poznań University of Live Sciences
Wojska Polskiego 28, 60-637 Poznań, Poland
[2]The Research Centre for Cultivar Testing,
63-022 Słupia Wielka, Poland

Summary

In the paper the four methods of testing uniformity of varieties of oil-seed rape are compared (combining over-years uniformity – COYU, the Bennett's method, the F-test and the Miller test). Partly real and partly simulated data were used. The different measures of agreement (P_z, Cohen kappa and odds ratio OR) showed high similarity of decisions between methods. Nevertheless for oil-seed rape data the most lenient (the highest number of varieties declared uniform) was the COYU method, the most restrictive was the Miller method.

Keywords and phrases: Bennett's method, coefficient of variation, COYU method, Miller method, oilseed rape, simulation, test F, variety uniformity

Classification AMS 2010: 62K99, 62P10

1. Introduction

In a paper the problem of uniformity testing of oil-seed varieties is discussed. There are two different approach possible. In the first, the standard deviations of candidate varieties are compared with average of standard deviations of reference set varieties (the set of varieties the candidate variety is compared with). This method is implemented in so-called COYU (combined over-years uniformity) method. In the second approach, the coefficients of variations of candidate varieties are compared with the average of such coefficients of reference set varieties. In this paper four method of testing uniformity (COYU, Bennett, Miller and F-test method) are compared using partly real and partly simulated data. All three methods based on testing of equality of coefficients of variations appeared to be slightly more restrictive (less candidate accepted as uniform) than COYU method.

2. Data

In DUS (distinctness, uniformity and stability) trials on oil-seed rape varieties the randomized complete block design is commonly used, and usually 30 measurements are taken from randomly chosen plants in each plot. As traditionally such trials are performed in two replicates, there is in total 60 measurements for each observed characteristic of each variety (established or candidate). In this paper two kind of data are fused and analyzed. For established set of varieties the real data taken from experiments performed by The Research Centre for Cultivar Testing, in a period from 2006-2008 are used, whereas for candidate varieties the simulated data as described by Zawieja at al.[2010] are used. Finally, there were 66 established varieties and 187 simulated (candidate) varieties in the 2006-2008 period, and there were 57 established and 272 simulated varieties in the 2007-2008 period, and 72 and 238 such varieties in the 2006-2008 period, respectively.

3. Methods

Before testing uniformity some basic statistics were calculated. So, the mean values $\bar{x}_i$ $(i = 1,2,...v)$ and the standard deviations s_i^2 $(i = 1,2,...v)$ for each variety were calculated at first, independently for each year. These statistics supplemented by numbers of measurements n_i for i^{th} variety (within years) and by number of years l are sufficient for all considered methods. Within countries associated with UPOV the COYU (combined over year uniformity) method is officially promoted for use. This method is based on comparison of (transformed) standard deviations of each candidate variety in turn with the mean value of standard deviations of the reference set varieties. The threshold for standard deviation of candidate variety is calculated as

$$UC = \bar{s}_d + t_{p;w-1}^{tab}\sqrt{s^2\left(\frac{1}{l}+\frac{1}{lw}\right)}\,, \tag{3.1}$$

where $\bar{s}_d$ is the average of (possibility adjusted using moving average method) standard deviations calculated over all varieties assigned to the reference collection (the set of varieties the new variety is compared with), s^2 is the sample variance among adjusted standard deviations of reference collection varieties after removing the effects of years. Next, l stands for the number of years (usually 2 or 3), w is the size of reference collection, t_p means the one-side t-Student's distribution critical value at probability p and degrees of freedom associated with s^2 (see Talbot 2000). Usually the value of $p = 0.001$ or $p = 0.002$ is accepted but other values are also admitted.

If (possibly adjusted) standard deviation of particular candidate variety is smaller than the UC value (threshold) for all considered characteristics, the variety is declared uniform. So, if for just one characteristic, the standard deviation is larger than the threshold, the variety is treated as non-uniform and as a consequence can not be registered.

In a Bennett's (1976) approach, the hypothesis

$$H_0:\ \zeta_1 = \ldots = \zeta_v (= \zeta, \text{say})\,, \tag{3.2}$$

is tested with use the 2Z statistic, where ζ_i denotes the coefficient of variation of i^{th} variety and v is the total number of compared varieties (one new variety and all varieties from the reference collection) and where

$$2Z = (n-v)\log\left(\frac{\sum_i y_i}{n-v}\right) - \sum_i (n_i-1)\log\left(\frac{y_i}{n_i-1}\right). \tag{3.3}$$

This statistic is approximately distributed as χ^2 with $(v-1)$ degrees of freedom. In this formula, n_i denotes the number of measurements for i-th variety, $n = \sum n_i$, y_i is calculated as

$$y_i = \frac{(n_i - 1)z_i^2}{1 + \frac{(n_i - 1)}{n} z_i^2} \tag{3.4}$$

where z_i denotes the empirical coefficient of variation and where ψ_i is the transformed value of the theoretical coefficient of variation ζ_i, namely $\psi_i = \zeta_i^2 / \left(1 + \zeta_i^2\right)$.

The Bennett test can be used for two purposes, for testing whether the varieties belonging to the reference set are uniform and for testing whether the candidate variety is sufficiently uniform (uniformity not worse then average uniformity of reference set variety). Johannes Forkman (2009) proposed to replace the Bennett's test for testing uniformity of t^{th} variety by F statistic of the form

$$F = \frac{y_t / (n_t - 1)}{\sum_i y_i / \sum_i (n_i - 1)}, \tag{3.5}$$

where summing is over all the reference set varieties. Statistic F has an approximate Fisher distribution with $n_t - 1$ and $\sum_i (n_i - 1)$ degrees of freedom. The F test is the third method considered here.

Uniformity of every "candidate" variety was tested using the three methods already described. Each variety was tested using COYU (combined over year uniformity) method, the Bennett's test and the F test. The method similar to that described by Zawieja at al. (2009) was used to compare decisions concerning uniformity. The Bennett's method can be applied when all coefficients of variation are not higher than 0.3 (Forkman 2009; Iglewicz and Meyers 1970). In our case this condition was always fulfilled.

The fourth method is called the Miller method as Miller (1991) proposed another test for hypothesis (3.2) that v coefficients of variations are homogeneous. His first test statistic was dependent on the order of tested populations, so in the following papers by Feltz and Miller (1996) and by Miller and Feltz (1997) the modified statistic D was proposed of the form

$$D = \frac{\sum_{i=1}^{v} (n_i - 1) z_i^2 - \frac{1}{n - v} \left(\sum_{i=1}^{v} (n_i - 1) z_i \right)^2}{\zeta^2 (0.5 + \zeta^2)} \tag{3.6}$$

Because the theoretical coefficient of variation ζ is not know, one must estimate it. Miller and Feltz (1997) proposed the following estimate

$$\zeta = \frac{\sum_{i=1}^{v}(n_i - 1)z_i}{n - v}. \tag{3.7}$$

The D Statistic is distributed as a central χ^2 random variable witch $v-1$ degrees of freedom. This approach was also recommended by Forkman (2006). It is worth to mention that both Forkman and Miller approximate tests are appropriate for small coefficients of variation only.

The decisions concerning uniformity of candidate varieties supported by each pair of methods are compared using two-way contingency table approach. The $n_{11} + n_{22}$ denote the number of unanimous decisions while $n_{12} + n_{21}$ denotes the number of contradictory decisions. Here the n_{11} (n_{22}) denotes the number of varieties declared as uniform (not uniform) by pair of methods. And respectively n_{12} (n_{21}) denotes the number of varieties declared as uniform by one method and as not uniform by the other.

The commonly used measure of agreement P_z between pair of methods is calculated according to formula

$$P_z = (n_{11}n_{22})/n \text{, where } n = n_{11} + n_{12} + n_{21} + n_{22} \tag{3.8}$$

If $n_{i\cdot} = n_{i1} + n_{i2}$ and $n_{\cdot j} = n_{1j} + n_{2j}$, then the Cohen (1960) coefficient of agreement between methods is defines as

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{3.9}$$

where $p_0 = \frac{n_{11} + n_{22}}{n}$ denotes the probability of unanimous decisions and $p_e = \frac{n_{1\cdot}n_{\cdot 1} + n_{2\cdot}n_{\cdot 2}}{n^2}$ means expected probability of unanimous decisions. The values of κ are from the range -1 to 1. According to Landis and Coch (1977), coefficient kappa is interpreted as follows

coefficient	degree of agreement
<0.00	lack
0.00 – 0.20	very weak
0.21 – 0.40	weak
0.41 – 0.60	medium
0.61 – 0.80	strong
0.81 – 1.00	nearly perfect

In order to check statistical significance of coefficient κ (testing of hypothesis $\kappa = 0$ against $\kappa \neq 0$) the statistic $Z = \frac{\kappa}{SE(\kappa)}$, can be used. Here $SE(\kappa) = \sqrt{\frac{p_e}{n(1-p_e)}}$. This statistic (under H_0) has a standard normal distribution.

Moreover, because in the ours previous papers, the odds ratio coefficient OR (Rudas 1998; Uebersax 2005) and its normal transformation $Z(OR)$ was used as a measure of agreement between pairs of methods, in this paper this measure of agreement is given too. This statistic tests the lack of association between methods.

In the literature (for example Wieringen and Hauvel 2005; Chmura-Kraemer et al. 2002) multi raters comparisons are proposed. In our applications these methods are of minor importance. So we focused our interest in paired comparisons of four methods.

4. Results

All considered methods were applied for three sets of generated data (data for candidate varieties). All the test were performed at the same α=0.002 level of significance (this level is recommended in the UPOV Guidelines). As already mentioned, the data for reference varieties were taken from real experiments performed at the experimental station in Słupia Wielka. The COYU analysis was performed with the use of DUST package of Weatherup (1992). The Excel spreadsheet was used for the three remaining methods. The results for two years data concerning the period 2006-2007 are given in Table 1**,** for 2007-2008 in Table 2 and for 2006-2008 in Table 3.

In rows of the tables the numbers of varieties (found in their respective categories) are shown and measures of agreement among methods are given.

Table 1. Decisions on uniformity of candidate varieties (data from the period 2006-2007), $\alpha = 0.002$

Pairs of methods	Uniformity decisions				Measures of agreements				
	UU	UN	NU	NN	P_z	OR	Z(OR)	κ	Z(κ)
Miller - COYU	159	0	28	0	0.85	-	-	0	0
Miller - Bennett	159	0	28	0	0.85	-	-	0	0
Miller - F	147	12	15	13	0.86	10.62	4.887	0.41	3.155
F- COYU	162	0	25	0	0.87	-	-	0	0
F - Bennett	162	0	25	0	0.87	-	-	0	0
COYU- Bennett	187	0	0	0	1.00	-	-	-	-

UU – uniform for both method,

UN – uniform with use first method, not uniform with use the second method,

NU – not uniform with use the first method, uniform with use the second method,

NN – not uniform according to both method.

All new methods are more restrictive than COYU as it is seen in Tables 1-3. There are – respectively 28 (in period 2006-2007), 106 (in period 2007-2008) and 88 (in period (2006-2008) less varieties declared uniform by the Miller method than by COYU method. Similarly there are 25, 60, 73 less varieties declared uniform by F-test (in successive periods) than by COYU. The most similar results are between Bennett and COYU methods (as only in period 2006-2008 the Bennett method declared uniform 16 varieties less than COYU). Comparing the other methods it can be seen that Miller method is more restrictive than Bennett method (the Miller method appeared to be the most restrictive).

Table 2. Decisions on uniformity of candidate varieties (data from the period 2007-2008), $\alpha = 0.002$.

Pairs of methods	Uniformity decisions				Measures of agreement				
	UU	UN	NU	NN	P_z	OR	Z (OR)	κ	Z(κ)
Miller - COYU	166	0	106	0	0.61	-	-	0	0
Miller - Bennett	166	0	106	0	0.51	-	-	0	0
Miller - F	158	8	54	52	0.77	19.02	7.163	0.458	6.99
F- COYU	212	0	60	0	0.78	-	-	0	0
F - Bennett	212	0	60	0	0.78	-	-	0	0
COYU- Bennett	272	0	0	0	1.00	-	-	-	-

For all pairs of methods the coefficients of agreement P_z are quite large. Odds ratio (when possible to calculate), indicates high agreement between pairs of methods. Also the Cohen coefficient indicates agreement between all pairs of methods.

It is interesting to observe so-called kappa paradox (Wieringen W. and Heuvel , 2005). Even if the coefficient of agreement P_z is relatively high (as in our case), when n_{12} or n_{21} is equal to zero, the kappa coefficient is also equal to zero falsely indicating that there is completely lack of agreement.

Table 3. Decisions on uniformity of candidate varieties (data from the period 2006-2008), α = 0.002.

Pairs of methods	Uniformity decisions				Measures of agreement				
	UU	UN	NU	NN	P_z	OR	Z(OR)	κ	Z(κ)
Miller - COYU	147	0	88	3	0.63	-	-	0.04	0.494
Miller - Bennett	147	0	72	19	0.70	-	-	0.25	3.104
Miller - F	136	11	26	65	0.85	30.91	3.805	0.66	9.353
F- COYU	162	0	73	3	0.69	-	-	0.05	0.566
F - Bennett	162	0	57	19	0.76	-	-	0.31	3.519
COYU- Bennett	218	17		2	0.92	25.65	2.595	0.16	0.766

5. Conclusions

Analysis of extent data oil-seed rape (partly real and partly simulated) allows to conclude that:

1) The Bennett's approach with replacement $2Z$ statistic by the F statistics used for testing uniformity of candidate varieties is more restrictive (less varieties accepted as uniform) than COYU;
2) The Bennett's and COYU methods were completely equivalent for two years data whereas for three years data the Bennett's method appeared to be slightly more restrictive;
3) Coefficient of agreement among remaining pairs of methods was smaller than 0.9;
4) The Miller method was the most restrictive (the smallest number of varieties declared uniform).

In majority of UPOV member countries the (freely available within the DUST package) COYU procedure is used for checking uniformity of candidate varieties. In a COYU approach the (possibly transformed) standard deviations of candidate variety and those of established varieties (reference set) are compared. In other three considered in this paper methods of testing uniformity, the equality of respective coefficients of variation is tested. In general all these methods were more restrictive (less varieties accepted as uniform) than COYU. The most similar were the results (decisions concerning uniformity) for COYU and Bennett's methods (see the largest coefficient of similarity P_z). So only the Bennett's method can potentially replace the – rather sophisticated - COYU method as it is computationally and conceptually much simpler.

References

Bennett B.M. (1976). *On an approximate test for homogeneity of coefficients of variation, in: Contributions to applied statistics* (ed. W.I. Ziegler). Birkhäuser Verlag, 169-171.

Chmura Kraemer H., Periyakoil V.S., Noda A. (2002). Kappa coefficients in medical research. Tutorial in Biostatistics 1. *Statistic In Medicine* 21, 14, 2109-2129.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1, 37–46.

Feltz, C.J., Miller, G.E. (1996). An Asymptotic test for the equality of coefficients of variation from k population. *Statistic In Medicine* 15, 647–658.

Forkman J. (2009). Estimator and Tests for Common Coefficients of Variation in Normal Distributions. *Communications in Statistics –Theory and Methods* 38, 233-251.

Iglewicz B., Meyers R. H. (1970). Comparison of approximations of the percentage points of the sample coefficient of variation. *Technometrics* 12, 166-169.

Landis J.R. Koch G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159—174.

Miller G. E. (1991). Asymptotic test statistics for coefficients of variation. *Communications in Statistics – Theory and Methods* 20, 3351–3363.

Miller G. E., Feltz, C.J. (1997). Asymptotic inference for coefficients of variation. *Communications in Statistics – Theory and Methods* 26, 715–726.

Rudas T. (1998). *Odds Ratios in the Analysis of Contingency Tables*. Thousand Oaks, CA: Sage Publ.

Talbot M. (2000). The Combined-Over-Years Distinctness and Uniformity criteria. *UPOV*, TWC/18/10, Geneva

Uebersax J. (2006). *Odds Ratio and Yule's Q*. http://www.john-uebersax.com/stat/odds.htm.

Weatherup S.T.C. (1992). *Distinctness, Uniformity and Stability trial (DUST) analysis system. User manual*. Department of Agriculture for Northern Ireland Biometrics Division, Belfast BT9 5PX.

Wieringen W. N., Heuvel E.R. (2005). A comparison of methods fort he evaluation of Binary Measurement system. *Quality Engineering* 17, 495-507.

Zawieja B., Pilarczyk W., Kowalczyk B. (2009). The comparison of uniformity decisions based on COYU and Bennett's method – oilseed rape data. *Colloquium Biometricum* 39, 170-176.

Zawieja B., Pilarczyk W., Kowalczyk B. (2010). Comparison of uniformity decisions based on COYU and Bennett's methods – simulated data. *Colloquium Biometricum* 40, 53-61.