DOI: 10.1515/bile-2015-0012

Biometrical Letters
Vol. 52 (2015), No. 2, 131 - 137

A remark on genotype selection in plant breeding projects

Ewa Bakinowska[1], Andrzej Bichoński[2], Radosław Kala[3], Bogna Zawieja[3]

[1]Institute of Mathematics, Poznan University of Technology, Piotrowo 3A, 60-965 Poznań,
Poland, e-mail: ewa.bakinowska@put.poznan.pl
[2]Malopolska Breeding Company Polanowice, Zbożowa 4, 30-002 Kraków, Poland,
e-mail: andrzejb@hbp.pl
[3]Department of Mathematical and Statistical Methods, Poznan University of Life Sciences,
Wojska Polskiego 28, 60-637 Poznań, Poland, e-mail: kalar@up.poznan.pl,
bogna13@up.poznan.pl

SUMMARY

One of the main problems in plant breeding is the selection of the best genotypes. Most often the selection is made using yield as a main trait of continuous type, and ignoring the other traits of discrete type. Here, a simple procedure is proposed for dealing with the selection problem using not only the yield, but also an auxiliary discrete trait. The method is based on transforming the continuous variable into a discrete one and testing the dependence of variables with the use of contingency tables. The procedure is illustrated by a real unreplicated experiment with winter wheat.

Key words: continuous trait, discrete trait, contingency table, chi-squared test

1. Introduction

In one of the stages of plant breeding projects, a breeder deals with the problem of selecting the most elite lines from a large number of genotypes. Usually in such research there are insufficient seeds. Consequently, the plant breeder is forced to employ so-called unreplicated designs, i.e. experiments in which each genotype (line) appears only once. Usually the whole experimental field in such an unreplicated experiment is divided into small plots, and each line is sown only on one plot. If the whole experimental field is uniform there is no need to control the fertility, and the plots may be organized mainly with regard to the simplicity of the experiment. Sometimes, when the fertility of the experimental field varies in a systematic manner in one direction, the plots are arranged in several rows orthogonal to the fertility trend, each row consisting of sequences of narrow plots. To control the fertility trend, some plots, systematically distributed along each row, receive the check cultivar (see e.g. Kempton, 1984; Dobek and Kala, 1995). In both cases the main issue is to choose from the whole collection of lines or genotypes those which are the most promising for further yield trials (see e.g. Ambrożу et al. 2008a, 2008b). The decisions are taken based on the observations of traits of various types. Usually, the main trait is the yield, being of continuous type, accompanied by some other characteristics of discrete type. The aim of this note is to propose a simple procedure for dealing with the selection problem using not only the main continuous trait but also an auxiliary discrete one.

2. The procedure

The difference in the types of the observed traits (variables) causes some difficulties in performing simultaneous analysis of the experimental data. One approach is to transform the discrete variable into a continuous one. However, the lack of replications, which is standard in unreplicated experiments, means that there is no guarantee of success. We propose the opposite approach, namely transforming the continuous variable into a discrete one. This is an easy task; the only difficulty concerns the question of how to divide the range of the continuous variable into several disjoint classes. Such transformation leads to the loss of some information, but in experiments in which the main objective is to distinguish between good, average and poor genotypes, the loss is not significant. After such a transformation, methods appropriate for discrete variables, in particular correlation analysis with the use of contingency tables, can be applied (Yates, 1934; Agresti, 1984 p. 5). Evaluation of the independence of discrete variables is usually carried out using the well-known chi-squared test. The essential limitations are then the frequencies of the subclasses, which should not be too small or empty (Agresti, 2002 pp. 395–398).

The conversion of a continuous variable to a discrete one requires first of all the determination of the number of classes (disjoint intervals). This is arbitrary, but if it is large, then a small number of observations will be classified in separate intervals. The division of the range may be uniform, but a better solution is to base the division on selected statistics of location and dispersion. In the simplest case the range of the variable can be divided into two classes: observations smaller and larger than the average (or median). A division into three classes (low, medium or high yield) is described, for example, by the intervals:

$$[\, x_{\min},\ \text{mean} - s/2\,), \quad [\, \text{mean} - s/2,\ \text{mean} + s/2\,), \quad [\, \text{mean} + s/2,\ x_{\max}],$$

where *s* is the sample standard deviation.

For illustration, let us consider a simple case involving two traits, one continuous and one discrete with three classes. Moreover, let the continuous variable be transformed to a discrete one with three classes also. As a result we have a three-by-three contingency table. If the analysis of such a table indicates a significant correlation between variables, the selection should be focused on the genotypes classified only in one subclass. If there is no correlation, the selection of lines may be based only on the main variable.

3. A real experiment

To illustrate the procedure, we use a winter wheat trial conducted in 2012 at the experimental station in Polanowice. In the experiment, 51 breeding lines (genotypes) of wheat and three control cultivars (Tonacja, Ozon and Patras) were analyzed. The genotypes were marked with consecutive numbers. In the experiment two levels of fertilizer were applied: A1 (a small dose of N-P-K) and A2 (a larger dose of N-P-K). As a result, the study included 108 combinations of genotypes and fertilization levels.

Table 1. Yield [t/ha] and lodging of winter wheat

Fertilization A1						Fertilization A2					
Line	L[a]	Yield	Line	L[a]	Yield	Line	L[a]	Yield	Line	L[a]	Yield
44	3	6.207	30	9	8.690	26	3	7.279	8	7	10.635
29	4	6.417	6	8	8.901	50	2	8.497	17	7	10.704
1	3	6.462	33	8	8.959	9	5	8.530	41	7	10.787
20	4	7.558	27	9	8.993	20	4	8.701	54	7	10.962
26	5	7.564	23	8	9.234	1	3	9.477	32	6	11.082
50	2	7.581	8	9	9.353	7	2	9.900	39	6	11.191
35	5	8.590	4	8	9.554	35	3	10.393	21	7	11.285
51	5	9.154	39	9	9.617	23	5	10.662	13	7	11.292
7	3	9.223	13	9	9.700	5	2	10.854	37	7	11.399
48	3	9.475	18	8	9.803	12	4	10.930	4	6	11.431
28	4	9.777	38	9	9.863	48	4	10.942	45	7	11.433
5	4	10.849	36	8	9.940	44	3	10.995	49	7	11.683
32	6	7.609	19	8	9.940	36	3	11.413	52	7	11.757
40	7	8.510	21	8	9.972	10	5	11.460	53	7	12.216
54	7	8.629	25	8	10.010	28	3	11.540	11	8	9.314
34	7	8.800	31	9	10.048	3	5	11.635	43	8	9.603
12	7	9.242	46	8	10.056	22	4	11.898	25	8	10.214
3	7	9.265	15	9	10.128	29	7	6.153	46	8	10.611
14	7	9.455	37	8	10.172	51	6	8.487	30	8	11.203
42	7	9.455	47	9	10.368	33	7	9.299	15	9	11.212
41	6	10.260	49	9	10.552	40	6	9.430	27	8	11.223
16	6	10.260	24	8	10.760	42	6	9.455	38	8	11.251
17	7	10.595	2	9	10.775	14	7	9.614	18	8	11.736
45	7	10.800	10	9	10.785	34	7	10.203	47	9	11.864
9	9	6.800	53	9	11.006	6	7	10.231	16	8	12.242
43	8	7.520	22	9	11.132	2	6	10.327	24	9	12.473
11	9	8.550	52	8	11.487	19	7	10.454	31	9	12.564

[a] L – lodging

The experiment was performed on a uniform field divided into two parts (A1, A2). Each part comprised single plots, each with a size of 10 m^2. During the study several traits were observed. The yield and the weight of 1000 grains represented the continuous variables. The other traits, for example the degree of lodging and brown rust resistance, were of discrete type. The discrete traits were measured on nine-point scales. The yield was converted to the common moisture content of 15%, while the nine degrees of lodging were reduced to three classes: strong lodging (1–5), medium lodging (6–7) and low lodging (8–9).

The observations obtained from the experiment are presented in Table 1. For each level of fertilization and in each *yield*×*lodging* subclass, the observations are ordered according to the yield. The lodging classes are separated by longer horizontal lines, while the yield classes are separated by shorter lines.

4. Results

The basic descriptive statistics of yield are contained in Table 2. These were required to determine the division of yield variability into disjoint classes. Because during the plants' growth two different fertilization levels were applied, the division of the yield into classes was performed separately for fertilization levels A1 and A2.

Table 2. Descriptive statistics of yield

	A1	A2
Mean	9.34	10.59
Standard deviation (*s*)	1.26	1.28
Range	5.28	6.41
$x_{\min}$	6.21	6.15
$x_{\max}$	11.49	12.56

Due to the relatively small number of studied genotypes, the yield range was divided into three classes. They were determined in accordance with the formulae given in section 2:

A1: low [6.21, 8.71), medium [8.71, 9.97), high [9.97, 11.49],

A2: low [6.15, 9.96), medium [9.96, 11.23), high [11.23, 12.56].

These classes are marked in Table 1 by shorter horizontal lines. Similarly, three lodging classes, as mentioned at the end of section 4, were also established. These are marked in Table 1 by longer horizontal lines. The results of classification of the data into a 3×3 contingency table, together with the chi-square statistic, are presented in Table 3.

Because the null hypothesis of independence of the yields and lodging may be rejected, the selection should be limited to the 20 lines classified in the subclass *high yield*×(*8–9*), i.e. the lines of high yield and resistance to lodging. Among them are six lines grown in favorable conditions (A2) and fourteen lines grown under less favorable conditions (A1). The first group contains the lines 38, 18, 47, 16, 24 and 31. Three of them, lines 47, 24 and 31, were resistant to lodging and of the highest yield under both fertilization levels. Other lines of this group, 38 and 18, yielded less at the lower fertilization level A1, while line 16 produced a high yield but was not very resistant to lodging in the case of fertilization A1.

Table 3. 3×3 table for yield and lodging

Yield	Lodging			χ^2 statistic	p-value
	(1–5)	(6–7)	(8–9)		
Low	7 + 6 = 13	3 + 6 = 9	4 + 2 = 6	9.9412 *	0.041
Medium	4 + 6 = 10	5 + 10 = 15	12 + 5 = 17		
High	1 + 5 = 6	4 + 8 = 12	14 + 6 = 20		

* $p < 0.05$

The check cultivars Patras (52) and Ozon (53) produced high yields at both levels of fertilization; however, with the higher fertilization A2 they were not so resistant to lodging. In turn, the variety Tonacja (54) gave low yields at both levels of fertilization and was not resistant to lodging.

5. Conclusions

A method has been proposed which can be used in the selection of genotypes in unreplicated experiments. It has been shown how a selection can be made using simultaneous continuous and discrete variables, after transforming the continuous variable to a discrete one and performing selection using a contingency table. The procedure is illustrated by a real experiment with winter wheat, which besides the yield takes into account the degree of lodging. The procedure is effective when the analyzed traits are significantly correlated. It can be used even if yields are improved due to environmental conditions, and also when more traits are analyzed. However, because of the requirements of correlation analysis, the number of searched lines (genotypes) must not be too small.

REFERENCES

Agresti, A. (1984): Analysis of Ordinal Categorical Data. New York: Wiley.

Agresti, A. (2002): Categorical Data Analysis, 2nd edn. NewYork:Wiley.

Ambroży K., Bakinowska E., Bocianowski J., Budka A., Pilarczyk W., Zawieja B. (2008a): Statistical support of selection decisions at early stage of cereal breeding. Part I. Methods of estimation of treatment effects. Biuletyn Instytutu Hodowli i Aklimatyzacji Roślin (Bulletin of the Institute of Plant Breeding and Acclimatization) 250: 21-28.

Ambroży K., Bakinowska E., Bocianowski J., Budka A., Pilarczyk W., Zawieja B. (2008b): Statistical support of selection decisions at early stage of cereal breeding. Part II. Empirical comparison of treatment effects estimation methods. Biuletyn Instytutu Hodowli i Aklimatyzacji Roślin (Bulletin of the Institute of Plant Breeding and Acclimatization) 250: 29-39.

Dobek A., Kala R. (1995): On the analysis of experiments with unreplicated varieties. Biuletyn Oceny Odmian 26-27: 73-82.

Kempton R. (1984): The design and analysis of unreplicated field experiments. Vtr. Pfanzenzchtg. 7: 219-242.

Yates, F. (1934): Contingency tables involving small numbers and the χ^2 test. J. Royal Statist. Soc., Suppl. 1: 217-235.