Theor Appl Genet (1989) 77:473–481

Accuracy and selection success in yield trial analyses *

H.G. Gauch, Jr. and R.W. Zobel

Department of Agronomy and USDA-ARS, Cornell University, Ithaca, NY 14853, USA

Received May 10, 1988; Accepted November 18, 1988
Communicated by A.R. Hallauer

Summary. Yield trials serve research purposes of estimation and selection. Order statistics are used here to quantify the successes or problems to be expected in selection tasks commonly encountered in breeding and agronomy. Greater accuracy of yield estimates implies greater selection success. A New York soybean yield trial serves as a specific example. The Additive Main effects and Multiplicative Interaction (AMMI) statistical model is used to increase the accuracy of these soybean yield estimates, thereby increasing the probability of successfully selecting, on the basis of the empirical yield data, that genotype which has the maximum true mean. The statistical strategy for increasing accuracy is extremely cost effective relative to the alternative strategy of increasing the number of replications. Better selections increase the speed and effectiveness of breeding programs, and increase the reliability of variety recommendations. Selection tasks are frequently more difficult than may be suspected.

Key words: AMMI – Order statistics – Selection – Soybean – Yield trials

Introduction

Yield trial data serve two basic research purposes which may be distinguished as estimation and selection. This paper's focus is on the latter purpose, with particular emphasis on the influence of accuracy upon selection success. That is, having first distinguished estimation from selection, the next matter is to quantify the relationship between these two tasks.

Statistical procedures familiar to agronomists and breeders are mostly aimed at the problem of yield estimation. They answer representative questions such as: What is the yield estimate for a particular trial and what is the standard error of this estimate? What yield interval gives a 95% probability of including the true mean? Does one yield trial mean differ from another mean at the 5% level of significance?

By contrast, the field of order statistics explores the properties of means which have been ranked (ordered), and is aimed at the problem of selection. Representative questions include: What is the probability of a successful selection, that is, what is the probability that the genotype selected on the basis of having the largest empirical mean is also the genotype which actually has the largest true mean? What is the probability that a subset, of a given size, of the best genotypes will include that genotype which actually has the largest true mean? Given N realizations from a normal distribution, what is the expected value of the largest of these realizations?

The problems of estimation and selection thus ask different questions. Both sets of questions are, however, relevant to the objectives of agronomists and breeders. Nevertheless, the answers to these two problems require two different statistical methodologies, and neither methodology constitutes a valid substitute for the other.

Were the accuracy of yield estimates increased, by whatever refinements, then intuition correctly suggests that the probability of making successful selections will also increase. However, intuition cannot supply a quantitative assessment of the exact magnitude of this increase.

When a quantitative understanding of this relationship between the accuracy of yield estimation and the selection success is lacking, it becomes impossible to

* This research was supported by the Rhizobotany Project of the USDA-ARS

know whether the costs associated with achieving a given increase in accuracy are justifiable in terms of the economic benefits associated with realizing a quantitatively predictable increase in the probability of making successful selections. This paper attempts to clarify this relationship between accuracy and selection success, thus placing such choices of research strategy upon a firmer foundation. This clarification is especially important given the current trends toward using fewer replications (Bradley et al. 1988), and also given the possibility of using the Additive Main effects and Multiplicative Interaction (AMMI) statistical analysis to achieve greater yield estimation accuracy even with fewer replications (Gauch and Zobel 1988).

This paper has three objectives. First, some basic results from order statistics are recapitulated. Second, simulation studies explore selection systems having two or three cycles (years) of selection. Third, selection success is examined for a New York soybean [*Glycine max* L. (Merr.)] yield trial, comparing selections based on yield estimates derived from simple treatment means and from an AMMI model.

The results presented here may provide agronomists and breeders with a more realistic concept of the successes or problems to be expected in their selection programs. It is suggested that selection tasks are often considerably more difficult than an uninformed intuition may suspect.

Order statistics

General discussions of order statistics are given by Gibbons et al. (1977) and Gupta and Panchapakesan (1979). The first of these texts is suitable for non-specialists, whereas the second includes a Bayesian approach to selection. Gupta and Berger (1988a, b) provide recent Bayesian results.

First, some terminology may be clarified. The standard normal distribution is a normal distribution with a mean of 0 and variance and standard deviation of 1. A realization is simply a particular instance or draw from a distribution. The expected value of a distribution is the average value of an infinite number of realizations. Attention here will focus on the maximum value encountered in a set of *N* realizations. A probability density function specifies the probability for an entire distribution, such as the familiar bell-shaped curve for the normal distribution, and is scaled to give unit area (probability).

Table 1 shows the average or expected value for the maximum of *N* normal realizations from a standard normal distribution. These maxima are surprisingly large for even rather small *N*. For example, when *N* is only 4, the maximum of these 4 normal realizations has an expected value of 1.03 standard deviations above the mean of 0. Note that equivalent increases in the maximum require progressively greater increases in *N*.

Table 1. Average value of the maximum of *N* normal realizations

N	Maximum
2	0.56
4	1.03
10	1.53
30	2.04
100	2.51
500	3.04
3,000	3.54
20,000	4.02

The conventional perspective in yield trials is to consider the genotypes and the environments (site-year combinations) to be factors of interest. Often yield trials are replicated, that is, they contain two or more yield plots for a given genotype and environment combination. For a replicated trial, the standard error of mean yields, averaged over replications, is equal to the standard deviation of individual yield plots (as estimated by the square root of the error mean square) divided by the square root of the number of replications. When yield estimates are based upon means over replications, then the standard error is used in the place of the standard deviation in order statistics calculations.

The expected ranges of *N* normal realizations are simply twice the values for the maxima shown in Table 1, because the expected minimum and maximum values are the same except for opposite signs. For example, given a replicated yield trial with 10 genotypes all having identical true means and standard errors, the range in empirical means is expected to be $2 \cdot 1.53 = 3.06$ standard errors. Likewise the expected separation between empirical means for 2 genotypes with identical true means is $2 \cdot 0.56 = 1.12$ standard errors.

Consider a replicated yield trial with 10 genotypes, a grand mean of 2,000 kg/ha, and a standard error of 200 kg/ha (that is, 10% of the grand mean). If there are no real yield differences between these genotypes, then mere statistical fluctuations may be expected to generate a best trial with $(2{,}000 + 200 \cdot 1.53) = 2{,}306$ kg/ha, a worst trial with 1,694 kg/ha, and a yield range of 612 kg/ha. Likewise were there 100 entries instead of 10, yields may be expected from 1,498 to 2,502 kg/ha, which is a range of over half of the magnitude of the grand mean.

On the other hand, even if statistically significant and agronomically important differences do exist among the true means of these 10 genotypes, nevertheless it may be difficult to correctly identify the superior genotypes. Since a best trial of 2,306 kg/ha, or 15% better than the grand mean, is expected merely because of statistical fluctuations, one may react by moderating excitement over any smaller yields. Patterson et al. (1977) and Talbot (1984) provide useful tables for interpreting yield trial results with caution. Simply ignoring genotypes with yields below 2,306 kg/ha may not be entirely satisfactory, however, because an increase in a true mean of considerably less than 15% could be of substantial agronomic and economic importance. Indeed, if a farmer's profit margin is about 10%, then a real yield increase of 15% would more than double the farmer's profits.

For some purposes it is inadequate to know merely the expected value of the maximum of N normal realizations, as in Table 1, but rather the entire distribution function is required. Figure 1 shows the probability density function for the maximum of N normal realizations for the cases of N equalling 1, 5, 10, 100, and 1,000. The case for $N=1$ is, of course, simply the normal curve itself, with a mean of 0 and a standard deviation of 1. The other curves may look much like normal curves, but in fact they are not symmetric and are skewed to the right. As N increases, these curves move to the right and become narrower. Each curve is scaled to give unit area under the curve, so narrower curves are also taller.

For a given curve in Fig. 1, values of X may be weighted by the probability value Y, using numerical integration, in order to obtain the expected or average value of X. Such values were given in Table 1. For example, the mean of 10 normal realizations is 1.53 (from Table 1). However, the curve for $N=10$ in Fig. 1 makes it clear that a maximum as large as 2 or even 2.5, or as small as 1 or even 0.5, would not be rare. Recall from the above example of a yield trial with 10 entries that its expected maximum yield is 2,306 kg/ha. Although this is the average maximum value, it would not be rare in a particular instance to obtain a maximum empirical mean of 2,400 or even 2,500 kg/ha, despite the true mean being only 2,000 kg/ha.

The method for calculating the curves in Fig. 1 may be described briefly. These technicalities are not essential for understanding this paper, but they may be of interest to some of its readers. Let $C(x)$ be the cumulative density function of the normal distribution, integrated from $-\infty$ to x. Then $C(x)$ is the probability that a realization will be smaller than x. The probability that all N realizations will be smaller than x is simply $C(x)^N$, and consequently the probability that one or more realizations will exceed x is the one-complement, namely $1-C(x)^N$. This latter quantity provides the cumulative density for the desired curves in Fig. 1, which thus are obtained by numerical differentiation followed by standardization to unit area. Numerical integration of the x values of such curves as those in Fig. 1, weighting by the relative frequencies on the Y-axis, provides the average values given in Table 1.

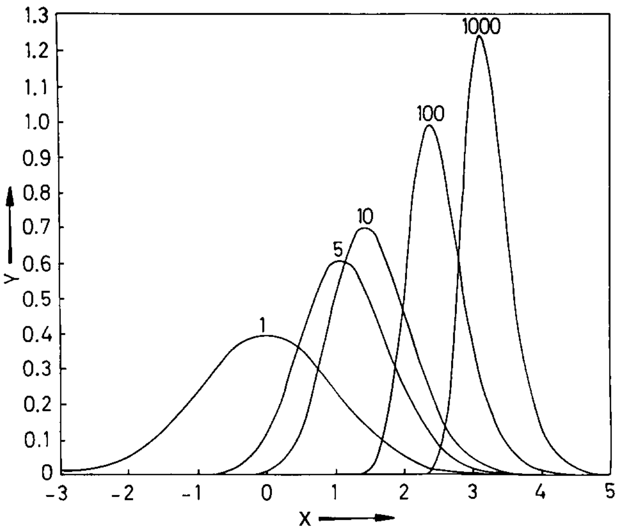

Fig. 1. The distributions for the maximum of 1, 5, 10, 100, and 1,000 realizations from a standard normal distribution with mean 0 and standard deviation 1. The X-axis shows standard deviation units centered at 0 for the standard normal curve, and the Y-axis is scaled to give unit probability under each curve

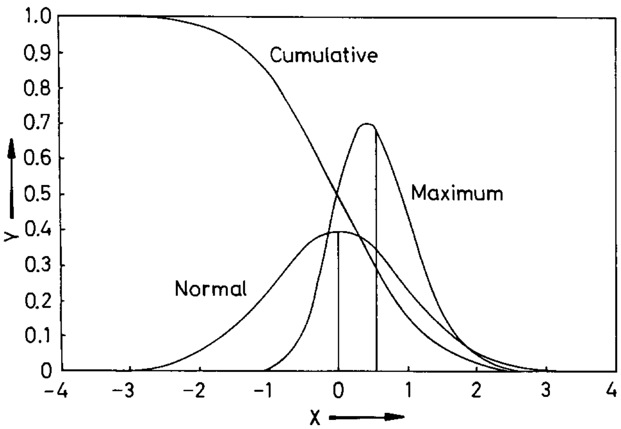

Fig. 2. A standard normal curve with mean 0 and standard deviation 1, its cumulative curve integrating from x to ∞, and the distribution for the maximum of 10 realizations from a normal distributions with mean -1 and standard deviation 1. The X-axis shows standard deviation units centered at 0 for the standard normal curve, and the Y-axis is scaled to give unit probability under the normal and maximum curves and a maximum probability of 1 for the cumulative curve

The distributions shown in Fig. 1 may be used to calculate the probability of a successful selection by the method depicted in Fig. 2. This figure represents a selection task having 10 inferior genotypes and 1 superior genotype. Selection success consists of selecting the superior genotype on the basis of its having the largest empirical mean. The superior genotype is represented by a standard normal distribution with a mean of 0, whereas the 10 inferior genotypes all have a mean of -1 standard deviation unit (and a normal distribution, not shown, to avoid clutter in Fig. 2). However, since the expected value of the maximum of 10 normal realizations is 1.53

Table 2. Probability of success in selecting the superior entry

R	No. of standard entries										
	5	10	20	30	50	100	200	500	1,000	2,000	5,000
0.2	21.3	12.3	6.9	4.8	3.0	1.6	0.9	0.4	0.2	0.1	0.0
0.4	26.5	16.3	9.6	6.9	4.6	2.5	1.4	0.6	0.3	0.2	0.1
0.6	32.2	21.0	13.0	9.7	6.6	3.9	2.2	1.1	0.6	0.3	0.2
0.8	38.5	26.4	17.2	13.2	9.4	5.7	3.5	1.7	1.0	0.6	0.3
1.0	45.0	32.4	22.2	17.5	12.8	8.2	5.2	2.7	1.7	1.0	0.5
1.2	51.6	38.8	27.9	22.6	17.1	11.4	7.5	4.2	2.6	1.7	0.9
1.4	58.2	45.6	34.2	28.4	22.2	15.4	10.5	6.2	4.0	2.6	1.5
1.6	64.5	52.5	41.0	34.8	28.0	20.3	14.4	8.8	6.0	4.0	2.3
1.8	70.5	59.3	48.0	41.7	34.4	25.9	19.0	12.3	8.6	6.0	3.6
2.0	75.9	65.8	55.1	48.8	41.4	32.3	24.5	16.6	12.1	8.6	5.4
2.2	80.7	71.9	61.9	56.0	48.6	39.1	30.8	21.7	16.3	12.1	7.9
2.4	84.9	77.3	68.4	62.9	55.8	46.4	37.6	27.7	21.4	16.3	11.1
2.6	88.4	82.1	74.4	69.4	62.8	53.7	44.9	34.3	27.4	21.5	15.2
2.8	91.2	86.2	79.7	75.3	69.4	60.9	52.3	41.5	34.0	27.5	20.2
3.0	93.5	89.6	84.3	80.6	75.4	67.7	59.6	48.9	41.2	34.2	26.0
3.2	95.3	92.3	88.1	85.1	80.7	74.0	66.6	56.3	48.7	41.4	32.6
3.4	96.7	94.5	91.2	88.8	85.2	79.5	73.0	63.6	56.2	48.9	39.7
3.6	97.7	96.1	93.6	91.8	89.0	84.3	78.7	70.3	63.5	56.4	47.3
3.8	98.5	97.3	95.5	94.1	92.0	88.3	83.7	76.4	70.3	63.7	54.9
4.0	99.0	98.2	96.9	95.9	94.3	91.4	87.8	81.8	76.4	70.6	62.3
4.2	99.3	98.8	97.9	97.2	96.0	93.9	91.1	86.3	81.8	76.7	69.3
4.4	99.6	99.2	98.6	98.1	97.3	95.8	93.7	89.9	86.3	82.1	75.6
4.6	99.7	99.5	99.1	98.8	98.2	97.2	95.6	92.8	90.0	86.6	81.1
4.8	99.8	99.7	99.5	99.2	98.9	98.1	97.1	95.0	92.9	90.2	85.8
5.0	99.9	99.8	99.7	99.5	99.3	98.8	98.1	96.6	95.1	93.1	89.6

standard deviations above their mean, in this case −1, the curve for the maximum of these *inferior* genotypes has its mean at 0.53. This is above the mean of 0 for the single *superior* genotype. Unfortunately this arrangement portends trouble for the selection task.

Consider a given yield, x, in Fig. 2. The probability that the superior genotype's empirical yield will exceed x is simply the cumulative of the normal distribution, integrating from x to $+\infty$ (which is the one-complement of the more customary cumulative from $-\infty$ to x). Weighting this cumulative distribution by the maximum distribution, using numerical integration, produces the desired probability of selection success. In this case, the result is a 32.4% probability of a successful selection.

The configuration of true means underlying Fig. 2, in which 1 mean is largest and N means are all inferior by the same amount R (expressed in units of standard deviations), is termed the Generalized Least Favorable (GLF) configuration (Gibbons et al. 1977:21–22). Given an interest in a genotype superior by the amount R, but indifference regarding lesser superiorities in yield, the GLF configuration poses the most difficult selection problem.

A GLF selection problem can be characterized by two quantities: the number N of inferior genotypes, and the separation R between the inferior and superior means scaled in units of standard deviations. For example, given a superior mean of 2,000 kg/ha, inferior means of 1,800 kg/ha, and a standard deviation of 200 kg/ha, then $R = (2{,}000 - 1{,}800)/200 = 1$.

Table 2 gives the GLF probability of selection success for various values of N and R. Note that this probability decreases as N becomes larger and as R becomes smaller. This can be understood readily by considering Fig. 2. Increasing N moves the maximum curve to the right (as in Fig. 1), hence more heavily weighting small values of the cumulative curve. Likewise, decreasing R amounts to shifting the maximum curve to the right relative to the cumulative curve, again more heavily weighting small values of the cumulative curve.

Typical selection problems in agronomy and breeding involve 10–100 entries and R values in the range of 0.4–2. Many of the values in this portion of Table 2 may be considerably smaller than most researchers' intuitions might suggest.

For example, the probability of successfully selecting the superior genotype from $N = 10$ competitors inferior by $R = 1$ standard deviation is, as noted earlier, only 32.4%. It appears that most agronomists and breeders would expect this probability to be considerably larger. This example with only 11 entries is, after all, a fairly easy selection problem when compared with many or most actual selection problems in agronomy and breeding. Furthermore, since there are only 11 entries, even without conducting a yield trial and without having any data or information whatsoever, one would still have a 9.1% probability of selecting the superior genotype just by chance alone. Consequently, conducting this yield trial experiment increases the probability of a successful selection by only 23.3%. Again, few agronomists or breeders would suspect such a small increase. Frequently selection tasks are considerably more difficult than may be recognized.

AMMI Analysis

The Additive Main effects and Multiplicative Interaction (AMMI) model, sometimes called the "biplot" model, is described by Bradu and Gabriel (1978), Kempton (1984), Gauch (1988), and Zobel et al. (1988). The AMMI model equation is:

$$Y_{ge}=\mu+\alpha_g+\beta_e+\sum_{n=1}^{N}\lambda_n\,\gamma_{gn}\,\delta_{en}+\varrho_{ge}$$

where Y_{ge} is the yield of genotype g in environment e,
μ is the grand mean,
α_g are the genotype mean deviations (the genotype means minus the grand mean),
β_e are the environment mean deviations,
λ_n is the eigenvalue of principal components analysis (PCA) axis n,
γ_{gn} and δ_{en} are the genotype and environment PCA scores for PCA axis n,
N is the number of PCA axes retained in the model,
and ϱ_{ge} is the residual.

Ordinarily N is chosen to retain only a few PCA axes in the model, leaving the higher axes in the residual. The eigenvectors γ and δ may be scaled as unit vectors, but another convenient scaling is to multiply both eigenvectors by the square root of the eigenvalue λ so that products of genotype and environment scores ($\lambda^{0.5}\gamma$ and $\lambda^{0.5}\delta$) estimate interaction effects directly without requiring another multiplication by λ. If the experiment is replicated, an error term ε_{ger}, which is the difference between the single observation Y_{ger} for replicate r and the Y_{ge} mean over replicates, may be added to the model. The AMMI model is regarded here as a fixed effects model because all inferences pertain to specific genotypes and specific environments, rather than to some population of genotypes or environments.

The least-squares fit to this model for balanced data (equal replication and no missing observations) is obtained in two steps. First, the additive main effects (μ, α_g, and β_e) are fitted using the ordinary calculations for a two-way analysis of variance (ANOVA; as in Chap. 14 of Snedecor and Cochran 1980). The variance not captured by this additive model remains in its residual, namely the interaction. Second, the multiplicative interaction effects (γ_{gn} and δ_{en}) are fitted using principal components analysis (PCA). Note that PCA is applied here to the interaction, that is to the residual from the additive ANOVA model, rather than to the original data. Since only the first few eigenvalues and eigenvectors are desired, an efficient direct iteration algorithm is applicable (Acton 1970; Gauch 1988). This algorithm achieves a linear workload (computing time proportional to the amount of data, that is to GE), so there is no computational difficulty in applying AMMI to large data sets. The AMMI calculations were performed by the FORTRAN77 program MATMODEL (Gauch 1987).

AMMI is usually the model of choice for data having significant main effects and significant interaction. This constitutes the main case in yield trial research. Even when a different model or a subcase of AMMI is best (such as only the additive part or only the multiplicative part of AMMI), this best model is routinely most easily diagnosed by first inspecting the results from an AMMI analysis (Bradu and Gabriel 1978). AMMI largely integrates and subsumes the several statistical models customarily applied to yield trial data, including the additive ANOVA, multiplicative PCA, and Finlay-Wilkinson linear regression models (Zobel et al. 1988).

AMMI has proven effective for a variety of interrelated agricultural research purposes. (1) Better understanding of GE interactions is facilitated by the genotype and environment interaction PCA scores, particularly as presented graphically in a biplot (Kempton 1984; Zobel et al. 1988). (2) More predictively accurate yield estimates result from discarding a noise-rich residual (Gauch 1988; Gauch and Zobel 1988). (3) Greater accuracy translates into new options to create experimental designs with fewer replications or with more treatments (genotypes, fertilizers, or whatever). (4) Greater accuracy also improves success in selecting truly superior material. In a typical plant breeding scenario, the selection gain requiring 3 years without AMMI could be achieved in only 2 years with AMMI. (5) Examination of the residuals from the AMMI model with respect to the physical layout of the yield plots can reveal spatial heterogeneity in experimental fields. (6) Ultimately the better understanding of interactions and the greater accuracy of yield estimates makes possible more reliable variety recommendations and more rapid advancement in breeding programs.

The focus of this paper is on the fourth purpose above. Greater accuracy from AMMI implies better selections, as considered next.

Selection systems

Simulation studies were used to investigate the efficiency of two selection systems for a substantial selection problem. There were 1,000 simulated genotypes with a uniform distribution of true means (yields) of 2,000, 1,999, 1,998, . . ., 1,002, and 1,001. Hence the worst genotype was about half as productive as the best genotype, and the grand mean was approximately 1,500. Note that this was not the GLF configuration. No phenotypic trait other than yield is considered here. The empirical or measured means resulting from a yield trial were simulated, for each genotype, by a normal distribution around its true mean. The standard deviation for these normal distributions was 450, corresponding to a standard deviation for individual plots of 30% of the grand mean. For those selection systems considered below involving replication, these normal distributions were based upon the standard error. For example, 4 replications resulted in a standard error of $450/\sqrt{4}$ or 225, which was 15% of the grand mean. This would represent a typical noise level in the New York soybean trials considered here.

The selection task was to select that genotype with the highest yield. In this study, success was measured not in terms of the probability of selecting the genotype with a true mean of 2,000, but rather in terms of the average true mean of the selected genotype, based on 1,000 simulations. The setting was to conduct a yield trial in several environments so that AMMI was applicable to the resulting multivariate data, but the following discussion focuses upon the selection task in any one particular environment.

Two selection systems were compared. (1) A traditional program tested 1,000 entries in year 1 using 2 replications, advancing the best 100 entries to year 2 using 4 replications, and finally advancing the best 10 entries to year 3 using 4 replications as a basis for the final selection of the best entry. (2) An alternative selection system relied on AMMI analysis to give an increase in the accuracy of yield estimates comparable to that expected from doubling the number of replicates (which was a modest increase given actual gain factors from AMMI ranging from 2.5 to 4.3 in Gauch and Zobel 1988). This was implemented by replacing the original standard deviation of 450 by an improved standard deviation of $450/\sqrt{2}$ or 318. Of course, equivalent results would arise from literally doubling the number of replications, rather than using AMMI, although this is a costly alternative. The AMMI selection system tested the 1,000 entries in year 1 using only 1 replication (with the improved standard deviation of 318), advancing the best 100 entries to year 2, using 4 replications (with a standard error of $318/\sqrt{4}$ or 159) as a basis for the final selection of the best entry.

Comparing the resources required by these two selection systems, the traditional program used 2,440 plots and 3 years, whereas the AMMI program used 1,400 plots and 2 years. The question was whether the AMMI selection system could produce equivalent or even superior results, even though it required less resources and less time.

The average true mean for the AMMI selection was 1,950, or 97.5% of the potential yield of 2,000. The average for the traditional system was only 1,942.

Instead of using an individual plot standard deviation of 450, the simulation was also repeated using values of 150, 300, 600, and 750. Over this broad range in noisiness of the data, the AMMI system was always the winner. Hence this result is not peculiar to the particular noise level chosen.

The results of this particular simulation study may be generalized somewhat as follows: when AMMI effectively increases the level of replication by a factor of 2–5 for a given data set, this corresponds to increasing R in Table 2 by a factor of the square root of these numbers, namely 1.41 to 2.24. Inspection of representative values in Table 2 shows that such an increase in R usually results in at least a factor of 3 increase in the number of entries N which can be handled while still maintaining the same level of successful selections. For example, a yield trial with $N=10$ inferior entries and $R=1$ gives P(success) = 32.4%, but a representative AMMI factor of 2.5 increases R to about 1.6 and P(success) to 52.5%, and at this level of R somewhat over 30 entries can be accommodated before P(success) drops back to its original value of 32.4%.

In other words, AMMI commonly allows about 3 times as many entries to be handled without reducing the probability of successful selections. Consequently, a three-year selection system assessing 1,000, 100, and then 10 entries can be replaced by an equally successful two-year AMMI selection system assessing 1,000 and then 30 entires.

New York soybean yield trial

A New York soybean yield trial is analyzed as an example. The yield data and details of the field methods are available in reports from the Department of Agronomy, Cornell University. Dr. Madison Wright kindly made available the original data on individual replicates. The subset of these data analyzed here has 7 genotypes grown in 40 environments (certain combinations of nine sites and 9 years). Yields are expressed in kg/ha at 13% moisture content. Most trials had 4 replicates, but due to occasional problems some had only 2 or 3; of the 1,120 plots planted, 1,044 plots were harvested.

Table 3 shows the ranked yield data for a representative one of the 40 environments analyzed by AMMI (Gauch and Zobel 1988), namely Aurora in 1979. General results for all 40 environments will follow later. Chippewa 64 at 2,491 kg/ha had the lowest yield, and SRF 200 at 3,432 kg/ha had the greatest yield. The trial had 4 replications and a standard error of 162 kg/ha. AMMI gave a statistical gain factor of 2.5 for this data set.

Assuming for the moment that the yield data and the standard error in Table 3 are in fact the true values, what is the probability that SRF 200 would again win this trial were it duplicated? It may be difficult to give a realistic estimate of this probability by inspecting Table 3. A better grasp on this problem may result from presenting these same data in graphical form.

Figure 3 repeats the information in Table 3, depicting empirical measurements of the 7 genotypes' yields as standard normal curves. The curve for SRF 200 is centered at 0 standard deviation units. The curve for Chippewa 64, e.g., is centered at (2,491−3,432)/163 = −5.81 standard deviation units. Inspection of this graph shows that realizations of SRF 200 above its mean are quite likely to win the trial, but that the other half of its realizations face significant competition, especially from Hodgson and Evans.

The exact probability of SRF 200 winning can be calculated by numerical integration. The calculation follows the preceding explanation for the GLF configuration, with the minor elaboration that the cumulative for each inferior variety must be computed individually and these values multiplied together (instead of computing a single value and raising it to the power N). The probability of success is 78.9%.

Were the curves in Fig. 3 broader, the probability of success would drop. Conversely, narrower curves would imply an increased probability of success. This difference corresponds to having less, or more, accurate data. This could arise, for example, from having fewer, or alternatively more, replications.

Figure 4 shows the percent probability of success for this soybean selection problem given 0–16 replications. Note that without conducting a yield trial (0 replications), the probability of success in selecting the truly superior genotype is 1/7 or 14.3% merely due to chance, since there are 7 entries in this competition. In order to achieve a 95% probability of success, 12 replications would be needed.

Agronomists and breeders can rarely afford planting 12 replications. However, the AMMI gain factor of 2.5 implies equivalent results with only 12/2.5 or 5 replicates. This yield trial's actual 4 replicates with AMMI analysis equates to 10 replicates in Fig. 4, resulting in a probability of success of 93.2%.

This probability of a successful selection may be partitioned into three sources. A 14.3% probability is free, due to mere chance. An increase of 64.6% is attributable to the yield trial experiment with its 4 replications, resulting in a 78.9% probability of success. An additional increase of 14.3% is attributable to the AMMI statistical analysis, resulting in the final probability of 93.2%. (It is merely coincidental that this value of 14.3% is the same

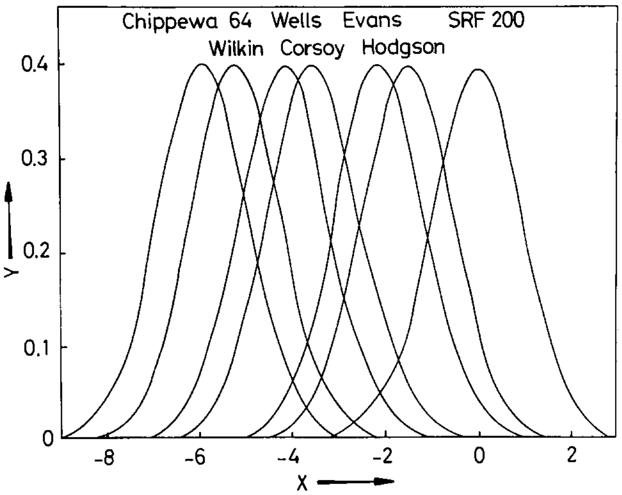

Fig. 3. Normal distibutions expected for empirical measurements of the yields of seven soybean varieties, expressed in units of standard deviations centered at the largest mean yield for variety SRF 200. The X-axis shows standard deviation units centered at 0 for SRF 200, and the Y-axis is scaled to give unit probability under each curve

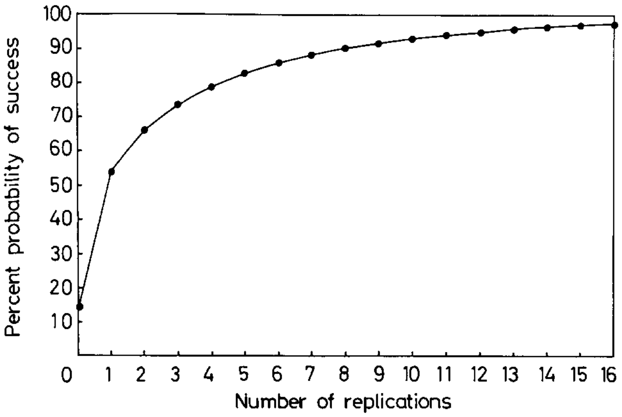

Fig. 4. The percent probability of successfully selecting SRF 200 on the basis of empirical data for 0–16 replications

Table 3. Data and AMMI yield estimates in kg/ha for Aurora in 1979 with rankings (from lowest to highest yields). The root error mean square is 324 kg/ha, and the standard error for a mean of 4 replicates is 162 kg/ha. The AMMI efficiency factor of 2.5 gives an adjusted standard error of 102 kg/ha. The yield trial's grand mean is 2,518 kg/ha

Genotype	Data	AMMI
Chippewa 64	2,491 [1]	2,701 [2]
Wilkin	2,621 [2]	2,627 [1]
Wells	2,788 [3]	2,861 [3]
Corsoy	2,877 [4]	3,132 [6]
Evans	3,127 [5]	3,006 [4]
Hodgson	3,203 [6]	3,197 [7]
SRF 200	3,432 [7]	3,015 [5]

as that due to chance.) Hence the AMMI analysis improves the probability of a successful selection about a fifth or a quarter as much as does conducting the yield trial itself. This is a substantial accomplishment, however, because successive increases in the probability of success are increasingly difficult to achieve.

Because the cost of the statistical analysis is trivial relative to the cost of the yield trial, this analysis represents an extremely cost effective means for improving the probability of making a successful selection. The alternative of planting 6 more replications (for a total of 10 replications) should produce equivalent selection success, but the cost of this alternative is considerably greater. This study involved 7 genotypes in 40 environments, so an additional 6 replications amounts to an additional 1,680 yield plots.

The above considerations are based on the simplifying assumption that the measured yields are the true yields. Some important observations may be offered regarding the more realistic assumption that the empirical means are merely a particular statistical realization of the actual (but unavailable) true means. Even if the simplistic outlook were true, there would still be only a 78.9% probability (without AMMI) of correctly identifying the best genotype. Obviously a more realistic perspective, as follows, will give an even less optimistic picture.

The AMMI model partitions this data set's 279 df for treatments (genotype and environment combinations) with a sum of squares (SS) of 165694661 into two basic sources: the AMMI model (including the environment and genotype main effects and one interaction PCA axis) with 89 df and a SS of 157825432, and a discarded residual with 190 df and a SS of 7869229 (Gauch and Zobel 1988). Because the AMMI model is a reduced model, its yield estimates differ from the actual data. In particular, since the SS of the AMMI model is less than the SS of the data (because a residual has been discarded), AMMI estimates tend to be less extreme than the original data. Furthermore, AMMI estimates for a given environment may give different genotype rankings than do the corresponding original data.

Table 3 also gives the AMMI model estimates and their rankings for Aurora in 1979. Again these estimates are based on calculations using all the data for all 40 environments. AMMI indicates that Hodgson, not SRF 200, is the superior genotype in this trial, and that Corsoy also outyields SRF 200.

If the original data are more to be trusted, then planting Hodgson instead of SRF 200 would result in a yield loss of 229 kg/ha, or 6.7% below SRF 200's yield. On the other hand, if AMMI estimates are more to be trusted, then the result is a yield gain of 182 kg/ha, or 5.7% over SRF 200's yield. Differences of this magnitude would be of economic significance.

For the 40 environments in this soybean yield trial, data rankings and AMMI rankings pick different winners in slightly over half of the cases. A similar result was obtained for an international corn (*Zea mays* L.) yield trial (unpublished data). These results should not be too surprising. Selection tasks are difficult and when accuracy increases, rankings and winners often change.

The calculation of selection success may be repeated for this soybean yield trial using the AMMI yield estimates in Table 3 instead of the original data. The AMMI estimated means are closer together, making the resulting selection problem more difficult, although the smaller adjusted standard error of 102 kg/ha offsets some of this difficulty. The probability of success for this selection task is less optimistic, being only 61.7%. On the other hand, if the 2% yield difference between Hodgson and Corsoy were declared indifferent, so that selection of either Hodgson or Corsoy would be regarded as a success, then the probability of success would increase considerably.

Discussion

The purposes and applications of agricultural research generate demands for accurate yield estimates and for reliable selections of superior entries. The contexts of breeding and agronomy may be considered more specifically.

Because breeding programs typically involve several cycles (years) of selections, an increase in genetic gain of only 1% or 2% per cycle could accumulate to a considerable amount. Much accuracy is needed to successfully select such marginally superior material from among a large number of slightly inferior genotypes. Furthermore, genotype-environment interaction must be well understood in order to make selections appropriate for a given growing region and to avoid unpleasant surprises (Kempton 1984; Bradley et al. 1988; Gauch and Zobel 1988). Finally, a distinctively predictive perspective is needed which supports inferences to new sites and new years as well as possible (Gauch 1988; Gauch and Zobel 1988).

Agronomic recommendations also require great accuracy, but for different reasons than breeding programs. An increased yield of 2%, for example, may seem small. However, if a farmer's income is distributed into 90% against costs and 10% profit, then a yield increase of merely 2% would provide a 20% profit increase. This is economically important, regardless of whether the farming situation be that of intense competition within a prospering country or be that of intense need within an agriculturally deficient country.

The problem is that such accuracy is rarely achieved. The required level of replication is ordinarily simply not feasible.

The problem of inadequate experimental accuracy is mitigated by several conventional means. A breeding system may provide covariates useable for refining the data. Block designs may remove some variation due to heterogeneous field conditions. Formal or informal judgments based on many years or many sites may be more reliable than judgments based on a single site-year environment. Nevertheless, after all is said and done, practical decisions must often be made about which genotypes to advance in a breeding program or which varieties to recommend in an agronomy program. Often the data and its analysis do not guide these decisions as reliably as one might think or intend.

Agronomists and breeders are accustomed to using adjusted means from various experimental designs, such as incomplete block designs. Likewise, AMMI may be used to produce adjusted means which can be demonstrated empirically to have greater predictive accuracy, and hence greater value for making selections, than do the unadjusted treatment means. The statistical strategies underlying these blocking and AMMI methods for adjusting the means are, however, quite different (Gauch and Zobel 1988). The customary methods partition blocks (or whatever) from the error degrees of freedom (df), producing sources for blocks and for pure error. The AMMI method partitions a model from the treatment df, leaving a discarded residual. Given a balanced design, treatments and error are orthogonal, so it is possible to use both methods. However, the adjustments from AMMI are typically several times as large and influential as are those from blocking (Gauch and Zobel 1988). It should not be surprising that more is to be gained by careful analysis of the treatment df than from careful analysis of the error df because the treatments are the factors of principal interest, whereas replication is of peripheral interest basically for error control. The significance of adjusting the means in the present context is that in general such adjustments will alter genotype rankings and hence may change selections.

Two recommendations seem worthwhile.

First, order statistics calculations may be done to quantify the magnitude of selection problems in a given research program. These results may confirm the present research strategy, or they may indicate the need for some adjustments. Quantification and evaluation of selection problems is particularly crucial in the current setting of rapidly changing experimental designs (Bradley et al. 1988).

Second, most alternatives for refining estimation and improving selection simply cannot and will not be done because of limitations of money, time, or other resources. Order statistics may show that a 95% probability of successful selections requires 12 replications, but no solution is forthcoming if the research budget limits experimental designs to 3 replications. Consequently special attention should be given to cost effective possibilities, such as using AMMI analysis in applicable instances in order to achieve greater accuracy and better selections without requiring better or more data.

Even if resources or circumstances do not allow for a refinement of experimental designs or analyses, it may still be worthwhile to analyze a research project using order statistics in order to adjust researchers' expectations and claims to a known and realistic level. More hopefully, however, this exercise will define new possibilities for greater research efficiency, greater yield estimation accuracy, and more reliable selections of superior genotypes.

Acknowledgements. We appreciate helpful suggestions from M. Smith, M. Sorrells, D. Wallace, and the reviewers.

References

Acton FS (1970) Numerical methods that work. Harper and Row, New York

Bradley JP, Knittle KH, Troyer AF (1988) Statistical methods in seed corn product selection. J Prod Agric 1:34–38

Bradu D, Gabriel KR (1978) The biplot as a diagnostic tool for models of two-way tables. Technometrics 20:47–68

Gauch HG (1987) MATMODEL. Microcomputer Power, Ithaca, New York

Gauch HG (1988) Model selection and validation for yield trials with interaction. Biometrics 44:705–715

Gauch HG, Zobel RW (1988) Predictive and postdictive success of statistical analyses of yield trials. Theor Appl Genet 76:1–10

Gibbons JD, Olkin I, Sobel M (1977) Selecting and ordering populations: A new statistical methodology. Wiley, New York

Gupta SS, Berger JO (1988a) Statistical decision theory and related topics IV, vol 1. Springer, New York Heidelberg Berlin

Gupta SS, Berger JO (1988b) Statistical decision theory and related topics IV, vol 2. Springer, New York Heidelberg Berlin

Gupta SS, Panchapakesan S (1979) Multiple decision procedures: Theory and methodology of selecting and ranking populations. Wiley, New York

Kempton RA (1984) The use of biplots in interpreting variety by environment interactions. J Agric Sci 103:123–135

Patterson HD, Silvey V, Talbot M, Weatherup STC (1977) Variability of yields of cereal varieties in UK trials. J Agric Sci 89:239–245

Snedecor GW, Cochran WG (1980) Statistical methods, 7th edn. Iowa State University Press, Ames/IA

Talbot M (1984) Yield variability of crop varieties in the UK. J Agric Sci 102:315–321

Zobel RW, Wright MJ, Gauch HG (1988) Statistical analysis of a yield trial. Agron J 80:388–393