CROP BREEDING AND APPLIED BIOTECHNOLOGY

ARTICLE

Statistical significance, selection accuracy, and experimental precision in plant breeding

Marcos Deon Vilela de Resende[1*] and Rodrigo Silva Alves[2]

Crop Breeding and Applied Biotechnology
22(3): e42712238, 2022
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332022v22n3a31

Abstract: *Genetic selection efficiency is measured by accuracy. Model selection relies on hypothesis testing with effectiveness given by statistical significance (p-value). Estimates of selection accuracy are based on variance parameters and precision. Model selection considers the amount of genetic variability and significance of effects. Questions arise as to which one to use: accuracy or p-value? We show there is a link between the two and both may be used. We derive equations for accuracy in multi-environment trials and determine numbers of repetitions and environments to reach accuracy. We propose a new methodology for accuracy classification based on p-values. This enables a better understanding of the level of accuracy being accepted when certain p-value is used. Accuracy of 90% is associated with p-value of 2%. Use of p-values up to 20% (accuracies above 50%) are acceptable to verify significance of genetic effects. Sample sizes for desired p-values are found via accuracy values.*

Keywords: *Enhancing breeding efficacy, experimental statistics, mixed models, number of repetitions, number of trials*

INTRODUCTION

Statistical significance, selection accuracy, and experimental precision are concepts used to assess experimental efficiency and the effectiveness of genetic selection in plant breeding (Resende and Alves 2020). The most important parameters and concepts in quantitative genetics and plant breeding are: genetic gain with selection (G_s); accuracy ($r_{\hat{g}g}$); heritability (h^2); genetic variance (σ^2_g); and genetic value (g). Genetic gain with selection ($G_s = k\, r_{\hat{g}g}\, \sigma_g$) measures the genetic and practical gains obtained with genetic improvement. Accuracy (correlation between predicted and true genetic values) and individual or plot level heritability (which is itself a component of accuracy) enables a predictive estimate of genetic gains with selection. The predicted genetic value and accuracy are essential in genetic evaluation catalogues on which selection decisions are based.

Genetic selection involves the prediction and ranking of genetic materials and is central to genetic improvement. Its efficiency is measured by selection accuracy. Tangential to genetic improvement is model selection, which is based on inference and hypothesis testing. Its effectiveness is inferred from statistical significance (p-value).

Accuracy is useful for making inferences about the quality of experiments, the reliability of predictions of genotypic values, and the statistical validity of predictive and inferential results. In practical terms, accuracy is also used to

***Corresponding author:**
E-mail: marcos.deon@ufv.br
ORCID: 0000-0002-3087-3588

Received: 06 June 2022
Accepted: 31 **August 2022**
Published: 22 September 2022

[1] Empresa Brasileira de Pesquisa Agropecuária - Embrapa Café/Current address: Universidade Federal de Viçosa, Avenida Peter Henry Rolfs, s/n, Campus Universitário, 36570-900, Viçosa, MG, Brazil
[2] Instituto Nacional de Ciência e Tecnologia do Café - INCT Café/Current address: Universidade Federal de Viçosa, Avenida Peter Henry Rolfs, s/n, Campus Universitário, 36570-900, Viçosa, MG, Brazil

compare alternative selection methods, to calculate genetic gains with selection and to plan experiments. Thus, it is one of the building blocks of statistical and genetic analyses.

In single-environment trials, the accuracy values are obtained considering the heritability (h^2) of the trait and the number of repetitions (n) of each genotype. In multi-environment trials, the accuracy is estimated considering the heritability of the trait (h^2), genotypic correlation across environments (r_{ge}), number of repetitions, and number of environments.

Conversely, a desired accuracy can be used to determine the size of experiments, inferred by choosing the number of repetitions and environments (total sample size). In this case, an optimized sample size can be obtained from the expected accuracy, heritability, and genotypic correlation across environments. In this study, we extend the work of Resende and Duarte (2007) and derived equations for the accuracy in multi-environment trials, modeling the effects of the genotype x environment (GxE) interaction using estimates of genetic parameters and Snedecor's F distribution. These equations were used to define the optimal sample sizes.

Until recently, identifying an adequate number of repetitions was mainly based on minimizing or reducing the residual variance in experimental statistics and quantitative genetics. However, this method is inefficient given the limited capacity of the coefficient of experimental variation (CV_e) to provide information about accuracy, as demonstrated by Resende and Duarte (2007). Another approach is to minimize the phenotypic variation of treatment means. This is also not entirely adequate, as a fraction of the phenotypic variance is genetic in nature. Another approach assumes the effects of genotypes as fixed and is based on maximizing the probability of detecting significant differences between treatments.

Recently, efforts have been made to determine an adequate number of repetitions (n) and environments (l) (Xu et al. 2016, Baxevanos et al. 2017a, Baxevanos et al. 2017b, George and Lundy 2019, Zhang et al. 2020, Woyann et al. 2020). Two important contributions were provided by Yan et al. (2015) and Yan (2021), who used a similar approach to the one herein but through different equations. Nevertheless, these previous studies did not contemplate a statistical way to express the equations in terms of the F test and the p-value. Storck et al. (2011), following Resende and Duarte (2007) for single-environment trials, extended the approach to determine plot size in agronomic crops. Yan's two article (2015 and 2021) were based on reliability (which is the square of the accuracy and for balanced data is equivalent to the heritability at the means level) of the prediction, called H and fixed at 0.75 as a general suitable value. Besides, they did not express the equations results in terms of the individual heritability.

Considering statistical significance, selection accuracy, and experimental precision in plant breeding, this study aims: i) to obtain accuracy estimators for multi-environment trials; ii) to obtain estimators for the number of replications and environments to maximize the selection accuracy in multi-environment trials; iii) propose a new methodology for classifying accuracy based on statistical significance via p-value. Our study extends the work of Resende and Duarte (2007) to maximize accuracy and optimize the definition of the number of replications and environments. An original approach was applied to multi-environment trials, which included deriving accuracy estimators and expressing them in terms of the F-test of the joint analysis of variance of multi-environment trials. This was then related to statistical significance via p-value. Here, quantitative genetics intersects with experimental statistics, advancing work in both areas.

ACCURACY AND ITS RELATIONSHIP WITH OTHER MEASURES OF EXPERIMENTAL QUALITY

The quality of genotypic evaluation should be inferred based on accuracy ($r_{\hat{g}g}$). In balanced experiments, Snedecor's F distribution can also be used, as $r_{\hat{g}g} = (1 - 1/F)^{1/2}$ (Resende and Duarte 2007). The mathematical expression that relates the appropriate values of F to the required accuracy is given as: $F = 1/(1 - \hat{r}^2_{\hat{g}g})$. The F statistics is the proportion between the mean square of treatments and the residuals mean square from an analysis of variance. To achieve an accuracy of 90%, an F value equal to 5.26 must be obtained. This value is independent of the species and trait evaluated and can be considered a standard value for any species and a reference value in tests of value for cultivation and use (VCU).

This statistic simultaneously contemplates the coefficient of experimental variation (CV_e), the number of replications (n), and the coefficient of genotypic variation (CV_g), as can be seen through the expression $F = 1 + (nCV^2_g/CV^2_e)$. Although traditionally used to evaluate experimental quality, the coefficient of experimental variation alone is inadequate. All three parameters are necessary because accuracy depends on them simultaneously, as shown through an alternative expression:

$$r_{\hat{g}g} = \{1/[1 + (CV_e^2/CV_g^2)/n]\}^{1/2}$$ (Resende and Alves 2020).

For the selection process in breeding programs, the aim should be to achieve accuracy values above 70% (Resende and Duarte 2007). This is equivalent to *F* values greater than 2. Therefore, *F* values less than 2 provide low accuracy (Resende and Alves 2020). Another statistic commonly calculated in the context of genetic evaluation, proposed by Vencovsky (1987), is the relative coefficient of variation ($CV_r = CV_g/CV_e$). By fixing the number (n) of repetitions or individuals per treatment, the magnitude of the relative coefficient of variation (CV_r) can be used to infer the accuracy and precision of the genetic evaluation. With $n = 2$, a $CV_r > 1$ provides high accuracy.

PROOF OF THE RELATIONSHIP BETWEEN ACCURACY AND *F* TEST

From an analysis of variance, the components of the accuracy can be expressed in terms of variance components (as used by Fisher, Kempthorne, Henderson and Robertson) or intraclass correlation coefficients (determination coefficients or proportions between variance components; as used by Lush and Wright) (Table 1).

Table 1. Illustration of the analysis of variance for random effects of genetic treatments

Source of variation	E(MS)†	E(MS)‡	F
Treatment	$\sigma_e^2 + n\sigma_g^2$	$[(1 - h^2) + nh^2]\ \sigma_y^2$	$1 + n\frac{h^2}{1-h^2}$
Error	σ_e^2	$(1 - h^2)\ \sigma_y^2$	-

†: expected mean square in terms of variance components; ‡: expected mean square in terms of intraclass correlation or coefficients of determination; σ_e^2: residual variance; σ_g^2: genotypic variance; σ_y^2: phenotypic variance; n: number of repetitions; and h^2: heritability.

At the individual (common in perennial plants) or plot (common in annual plants) levels, *F* is given as $F = 1 + n\frac{h^2}{1-h^2}$ or $F = 1 + \frac{n}{\lambda} = \frac{\lambda + n}{\lambda}$, where $\lambda = \frac{(1-h^2)}{h^2}$ is the shrinkage factor in the mixed model equations. *F* will be greater than 1 only if h^2 is greater than zero. Since $F - 1 = n\frac{h^2}{1-h^2}$, the number of repetitions is given as: $n = (F - 1)\frac{(1-h^2)}{h^2} = (F - 1)\lambda$. The significance of *F* indicates that h^2 is non-zero.

Increasing the number of repetitions (*n*), increases the value and power of the *F* test in detecting significance. It also increases the reliability or heritability at the treatment mean level, given as $h_m^2 = 1 - \frac{1}{F}$ and the accuracy given as $r_{\hat{g}g} = \sqrt{1 - \frac{1}{F}}$ (Resende and Duarte 2007). The variance components enable us to estimate heritability or coefficients of determination at the individual plot and treatment mean levels, given as: $h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma^2}$ and $r_{\hat{g}g}^2 = h_m^2 = \frac{nh^2}{1 + (n-1)h^2}$, respectively. The h^2 can also be estimated as $h^2 = \frac{CV_r^2}{CV_r^2 + 1} = \frac{1}{1 + 1/CV_r^2}$ and by $h^2 = \frac{F-1}{(F-1) + n}$, as a function of *F* and *n*.

High reliability and accuracy can be achieved using an adequate number of repetitions or individuals (*n*) per treatment. An $F = 5.26$ is reached, for example, with $n = 6.39$, for $h^2 = 0.40$. It can be inferred that $h^2 = 0.40$ and $n = 6$ provides high accuracy ($r_{\hat{g}g}^2 = 0.90$). From the desired reliability ($r_{\hat{g}g}^2$), and according to the heritability of the trait (h^2), *n* is given as:

$$n = \frac{r_{\hat{g}g}^2}{h^2}\frac{(1-h^2)}{(1-r_{\hat{g}g}^2)} = \frac{r_{\hat{g}g}^2}{(1-r_{\hat{g}g}^2)}\frac{(1-h^2)}{h^2}.$$

Yan et al. (2015) used the same approach to obtain the optimal number of repetitions (*n*) but fixed the reliability ($r_{\hat{g}g}^2$) at 0.75, which led to more restricted results. Furthermore, they did not express the equations results in terms of individual heritability.

NEW ACCURACY CLASSIFICATION BASED ON STATISTICAL SIGNIFICANCE

The quality of genetic evaluation in the context of plant breeding and experimentation is generally based on the statistical significance (p-value) of the genetic effects of the statistical model and on the accuracy of the genetic values. Initially, significance levels of 1% and 5% were considered as sufficient to statistically validate the comparison between genetic treatments (genotypes, varieties, cultivars, clones) (Fisher 1925). These cut-off points have also been used in the comparison and selection of statistical models with a hierarchical or nested structure, for example, likelihood ratio test (LRT) or deviance analysis (Resende 2007). Measures of significance associated with genetic variance (σ_g^2) and individual heritability (h_g^2) are also used, for which values must be statistically different from zero for acceptance and validity of the experiment, considering the possibility of sufficient genetic variability for genotype selection.

Geneticists also rely on the magnitude of accuracy (correlation between predicted and parametric values) to infer about the effectiveness of selection and consequent genetic improvement. Reference values were suggested by Resende and Duarte (2007), with an accuracy of ≥90% necessary for recommending cultivars, and a desirable accuracy of ≥70% for improvement in the context of recurrent selection.

In technical works and in the practice of genetic improvement, it is a common doubt to know in which situations (in terms of magnitudes of genetic variance (σ_g^2), individual heritability (h_g^2), significance of genetic effects and/or statistical differences of fit between prediction models), the selection and validation of models as well as the acceptability of the levels of genetic variability and heritability present in the breeding populations are reasonable to lead to adequate genetic gains. Which magnitudes are acceptable? For example, is a heritability value at the mean level of 70% favorable for selection? Is an individual heritability equal to 5% valid for selection? To respond these questions, information on the magnitude of the accuracy associated with these situations is necessary. Thus, the key questions are: what is the relationship between accuracy and significance (p-value) in an experiment, and which criterion should be given preference? Discussions related to such misgivings are absent from the scientific literature. This study aimed to address these issues.

Beginning from the fact that for each p-value there is a test statistic of the data distribution, some associations between the test statistic and p-value can be stipulated in experimental evaluation. The genetic values estimated from the statistical analysis can be tested against zero, using the Student's *t* test with infinite degrees of freedom (Van Vleck et al. 1987). To perform this test, a significance level (or the complement called degree of confidence) must be chosen, which is usually 5% (95% confidence) and associated with a Student's *t* test value equal to 1.96. Snedecor's *F* distribution is also used in the analysis of experiments and is asymptotically (tends to infinite degrees of freedom for the residual) equivalent to the square of the Student's *t* distribution, i.e., $F = t^2$. Asymptotic equivalence also exists between the Chi-square (χ^2) distribution with one degree of freedom and Snedecor's *F* distribution, with one degree of freedom for the numerator and infinite degrees of freedom for the residual. Resende and Duarte (2007) showed the following relationships between *F* and the square of accuracy: $r_{\hat{g}g}^2 = 1 - \frac{1}{F}$ and $F = \frac{1}{1 - r_{\hat{g}g}^2}$. Based on these relationships, knowing the value of *F* allows us to estimate the accuracy via $r_{gg} = (1 - 1/F)^{(1/2)}$. Also, $r_{gg} = (1 - 1/\chi^2)^{(1/2)}$ and $r_{gg} = (1 - 1/t^2)^{(1/2)}$. Thus, the p-value can be inferred from tables of Snedecor's *F*, Student's *t*, and Chi-square statistics, with large (tending to infinite) number of degrees of freedom for the residual, thus establishing a bridge between p-value and accuracy. A relationship also exists between *F* and the non-centrality parameter ($NCP = (z_\alpha + z_\beta)^2$) via $(F - 1) = (z_\alpha + z_\beta)^2 = NCP$, that is, $F = 1 + NCP$ (Resende and Alves 2020), discussed further below.

In Table 2, we present accuracy values in the first column and associated p-values in the second column. These two columns offer information about which p-value is being accepted when practicing a certain accuracy. For example, typical values of 0.70, 0.80 and 0.90 are associated with p-values of 16% (0.16), 10% (0.10) and 2% (0.02), respectively. An accuracy of 0.50 is associated with a p-value of 25% (0.25). Values less than 0.50 for accuracy are unacceptable as they lead to a selective coincidence of less than 50%. In this case, selection would result in more mistakes than successes. Thus, the highest acceptable p-value is less than 25%. Traditionally p-values greater than 5% are not allowed in selection. The results presented here suggest that p-values between 5% and 20% are also adequate in some situations. Accuracy values should be interpreted as: useless, leading to more wrong than correct selections (bellow 0.5); useless, leading to selection at random (equal 0.5); useful, leading to more correct than wrong selections (above 0.5).

Table 2. p-values, Snedecor's *F* or Chi-square (χ^2) statistics, Student's *t,* and reliability (r_{gg}^2) associated with different accuracy (r_{gg}) values

r_{gg}	p-value	F or χ^2	Student's t	r_{gg}^2
0.50	0.25	1.33	1.16	0.25
0.60	0.21	1.56	1.25	0.36
0.70	0.16	1.96	1.40	0.49
0.75	0.13	2.31	1.52	0.57
0.80	0.10	2.76	1.66	0.64
0.85	0.06	3.53	1.88	0.72
0.90	0.02	5.43	2.33	0.82
0.93	0.005	7.90	2.81	0.87
0.95	0.001	10.82	3.29	0.91
0.99	0.000000002	36.00	6.00	0.97

r_{gg} = 50%, p-value = 25%; r_{gg} = 70%, p-value = 16%; r_{gg} = 80%, p-value = 10%; r_{gg} = 90%, p-value = 2%; r_{gg} = 95%, p-value = 0.1%.

In Table 3, p-values are presented in the first column and associated accuracy values in the second column. These two columns show the accepted selective accuracy when

using a certain p-value. For example, typical p-values of 0.10, 0.05, and 0.01 are associated with accuracies of 79%, 86%, and 92%, respectively. These traditional 10%, 5% and 1% cut-off points for significance were recently revised, and a p-value of 0.5% (0.005) is currently widely accepted (Benjamin et al. 2018). With this p-value, the associated accuracy is 93%. Using this approach seems appropriate and can be recommended with confidence. Thus, in the final stages of breeding programs a pair p-value = 0.005 / r_{gg} = 93% is strongly indicated. The criteria presented can revive the use of the *F* distribution and its p-value in the current analytical context of genetic improvement. However, this does not mean that a traditional analysis of variance (ANOVA) is necessary, as an analysis using mixed models provides $F = \frac{\sigma_g^2}{PEV}$ (Resende and Alves 2020), where PEV is the prediction error variance associated with the empirical best linear unbiased prediction (E-BLUP) of a genetic effect. It is an element extracted from the diagonal of the generalized inverse of the coefficient matrix of the mixed model equations (Fisher information matrix).

The columns of p-values and accuracy ($r_{\hat{g}g}$) in the Table 3 were plotted in Figure 1. The curve equation describing $r_{\hat{g}g}$ as a function of the p-value is given by $r_{\hat{g}g}$ = 0.814 − 1.521 (*p value* − 0.079), where -1.521 is the regression coefficient. The fitting was good, with R^2 = 0.99 and mean absolute error of 0.01. From this function we create the estimated equivalences between p-value and accuracy (Table 4).

Table 3. Snedecor's *F* or Chi-square (χ^2) statistics, Student's *t*, reliability (r^2_{gg}), accuracy (r_{gg}), and class associated with different p-values

p-value	F or χ^2	Student's *t*	r^2_{gg}	r_{gg}	Class
0.25	1.33	1.16	0.25	0.50	Moderate
0.20	1.64	1.28	0.39	0.62	Moderate
0.15	2.07	1.44	0.52	0.72	High
0.10	2.71	1.65	0.63	0.79	High
0.05	3.84	1.96	0.74	0.86	High
0.025	5.02	2.24	0.8	0.89	High
0.010	6.63	2.57	0.85	0.92	Very high
0.005	7.88	2.81	0.87	0.93	Very high
0.001	10.83	3.29	0.91	0.95	Very high
0.0005	12.12	3.48	0.92	0.96	Very high

p-value = 10% to 2.5%, r_{gg} = 79% to 89%; p-value = 2% to 0.5%, r_{gg} = 90% to 93%; p-value = 0.1%, r_{gg} =95%.

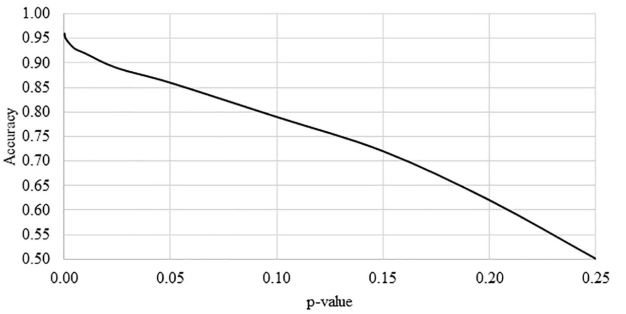

Figure 1. Accuracy as a function of the p-value, obtained from Table 3.

The strength of evidence of the p-values enables us to test a null hypothesis H_0 against an alternative hypothesis H_1 based on observed data. The p-value is defined as the probability, that a test statistic is as extreme or more extreme than its observed value, calculated considering the null hypothesis. The null hypothesis is rejected, and the finding is declared statistically significant if the p-value falls below the type I (current) error threshold: α = 0.05.

From a Bayesian perspective, a more direct measure of the strength of evidence for H_1 relative to H_0 is their odds ratio (ratio of their probabilities). By Bayes' rule, this ratio can be written as: BF x (Pr (H1) / Pr (H0)) = BF x (*a priori* probabilities), where BF is the Bayes Factor (also related to Bayesian information criteria (BIC); according to Resende and Alves 2020) that represents the evidence of the data, and the prior probabilities can be informed by the researchers' beliefs, scientific consensus, and validated evidence of similar issues in the same research field. Multiple hypothesis testing, P-hacking, and bias affect the credibility of the evidence, with some of these practices reducing the prior odds of

Table 4. Equivalences between p-value and accuracy, estimated by the regression equation: accuracy = 0.814 - 1.521 (p value - 0.079)

p-value	Accuracy	p-value	Accuracy	p-value	Accuracy	p-value	Accuracy
0.0005	0.93	0.05	0.86	0.12	0.75	0.19	0.65
0.001	0.93	0.06	0.84	0.13	0.74	0.20	0.63
0.005	0.93	0.07	0.83	0.14	0.72	0.21	0.62
0.01	0.92	0.08	0.81	0.15	0.71	0.22	0.60
0.02	0.90	0.09	0.80	0.16	0.69	0.23	0.58
0.03	0.89	0.10	0.78	0.17	0.68	0.24	0.57
0.04	0.87	0.11	0.77	0.18	0.66	0.25	0.55

H_1 over H_0. Analyses of results from reproducibility studies suggest that, for experiments in some fields, the prior odds of H_1 with respect to H_0 may only be about 1:10. Therefore, for fields where the threshold for defining statistical significance for new discoveries is a p-value < 0.05, Benjamin et al. (2018) proposed a change to a p-value < 0.005, i.e., p-value = 0.05 x 0.10 = 0.005. This simple step would immediately improve the reproducibility of scientific studies in many fields.

PAIRWISE COMPARISONS AND THE MULTIPLICITY PROBLEM IN THE VALIDITY OF THE NEW CLASSIFICATION METHOD

The relations (shown in Tables 2 and 3) between accuracy (via F values associated to 1 degree of freedom for genotypes) and p values hold, as we will show bellow. Even with treatments number higher than 2 in the trials, the precision, reliability and selection accuracy rely on precision of pairwise comparisons as the basic quantities to be averaged aiming to obtain the accuracy or its squared value (also called reliability or broad sense total heritability at mean genotype level).

We seek for a relation between reliability (squared accuracy $\hat{r}^2_{\hat{g}g}$) of the predictions and statistical significance (probability of type I error) of the difference of genotypic treatments. Such reliability can be viewed as a generalized coefficient of determination of treatments and also as a proportional reduction of errors. Piepho (2019) addressed the subject $\Omega = \hat{r}^2_{\hat{g}g}$ in statistics and gives the generalized coefficient of determination as $\Omega = 1 - \frac{\sigma^2_e}{\sigma^2_{e0}}$ From this, we can arrive at $\hat{r}^2_{\hat{g}g}$ as given by the BLUP method. F-statistic for comparing the full and reduced models can be written as a function of Ω and vice versa (Edwards et al. 2008).

In statistics, $\Omega = 1 - \frac{\sigma^2_e}{\sigma^2_{e0}}$ [1], where σ^2_e and σ^2_{e0} are the error variances of the full and of the null model, respectively (Piepho 2019). From this and in the genetics context, $\hat{r}^2_{\hat{g}g} = 1 - \frac{PEV}{\sigma^2_g}$ [2], where PEV is the genetic prediction error variance and σ^2_g is the true genotypic variance (Searle et al. 1992). According to Resende and Duarte (2007), $\hat{r}^2_{\hat{g}g} = 1 - \frac{1}{F}$ [3], where F is the calculated (from experimental data) Snedecor F statistics.

From [2] and [3] it can be perceived that $\frac{PEV}{\sigma^2_g} = \frac{1}{F}$ [4]. From [4], we have $F = \frac{\sigma^2_g}{PEV}$ [5] (Resende and Alves 2020). According to Cullis et al. (2006), the reliability is , $\hat{r}^2_{\hat{g}gC} = 1 - \frac{v_{BLUP}}{\sigma^2_g}$ [6], where v_{BLUP} is the mean (across all pair of genotypes combinations) variance of the difference of two treatments BLUP.

From [6] and [3], we have $\frac{v_{BLUP}}{2\sigma^2_g} = \frac{1}{F}$ [7], which leads to $Fv_{BLUP} = 2\sigma^2_g$ [8], and $v_{BLUP} = \frac{2\sigma^2_g}{F}$ [9]. From [8] or [9], we get $F = \frac{2\sigma^2_g}{v_{BLUP}}$ [10]. Rearranging [9], we arrive at $\frac{v_{BLUP}}{2} = \frac{\sigma^2_g}{F}$ [11], which gives $F = \frac{\sigma^2_g}{v_{BLUP}/2}$ [12]. From [12] and [5], we have $PEV = \frac{v_{BLUP}}{2}$ [13].

For n replication number of each genotype and according to Resende (2007), $v_{BLUP} = \frac{2\sigma^2}{n}$ [14], and we have PEV= $\frac{\sigma^2}{n}$ [15], as we intended to prove. The equation in [15] holds for uncorrelated fixed genetic effects (which produces BLUE of the genotypes effects) (Resende 2002, 2007).

Defining the quantity v_{BLUE} as the mean (across all pair of genotypes combinations) variance of the difference of two treatments BLUE and after noticing the similarity between v_{BLUP} and v_{BLUE}, another reliability estimation method arises, according to Piepho and Mohring (2007) and also used by Dias et al. (2020), in which the reliability is given by $\hat{r}^2_{\hat{g}gP} = 1 - \frac{v_{BLUE}}{2\sigma^2_g}$ [16], which is similar to [6].

The equivalence between the three approaches ($\hat{r}^2_{\hat{g}g}$ as given by the BLUP method, $\hat{r}^2_{\hat{g}gC}$ and $\hat{r}^2_{\hat{g}gP}$) was demonstrated above. Schmidt et al. (2019) also had an empirical evidence of this approximated equivalence. In this way, the overall squared accuracy comes from the average of all pair (two by two combinations) of comparisons based on the difference between each two genotype means. The two approaches ($\hat{r}^2_{\hat{g}gC}$ and $\hat{r}^2_{\hat{g}gP}$) showed to be equivalent as obtaining the squared accuracy from the mixed model analysis using the PEV provided by the inversion of the Fisher information matrix.

The results coming from pairwise comparisons are exact for only single comparison between two treatments. For higher number of treatments there is the multiplicity problem of all pairwise comparisons. For circumventing this, a global protection of the significance level to test the null hypothesis concerning treatments effects should be considered. Then, we adopted the Bonferroni correction and protection (Bonferroni 1936). For T treatments, this approach changes the v distribution of the p-value to that given by $v^* = v|T$ distribution. Then the significance based on v^* should be attained in order to keep the overall v-based significance. For example, for v equal to 10% and T = 10, the v^* distribution should be 1% and the F on v^* is 6.63 (and r_{gg} = 0.92), which is necessary and sufficient to keep the original v distribution on 10%. So, according to the stochastic probability laws (Papoulis and Pilla 1965, Mood et al. 1974) and mathematical logic rules (Lightstone 1978), this leads to the combined accuracy of $r_{ggc} = r_{ggv}\, r_{ggv^*} = 0.79 \times 0.92 = 0.73$. From these considerations we have the accuracies varying with the number of treatments (and so with the number of degrees of freedom for treatments), as expected in F tests obtained in practical experimentation involving diverse number of treatments. Residual degrees of freedom numbers were taken as infinity in the paper overall, as they quickly approach (according to asymptotic theory of convergence in distribution and in probability) the typically infinity value with relatively small numbers (120 for example, as is showed by Steel and Torrie (1980)). In plant breeding, given the high numbers of treatments and of replications, this assumption of approaching infinity residual degrees of freedom is easily met.

From Table 5 and from $r_{ggc} = r_{ggv}\, r_{ggv^*} = 0.79 \times 0.92 = 0.73$ as above, it can be seen that the Bonferroni correction and protection is conservative, reducing the original (unprotected) accuracy from 0.79 to the combined accuracy of 0.73, in the example of p-value of 10% and T = 10. The Bonferroni correction is also more rigorous in providing significance than is the unprotected t test. Table 5 must be read in the triples (*p-value*; T number; r_{gg} on $v|T$). The amount r_{gg} on $v|T$

Table 5. Bonferroni corrected accuracies (r_{gg}) for the number of treatments (T number), accounting for the multiplicity problem of all pairwise comparisons

p-value v	T number	F on p-value v*	rgg on v*	rgg on v	rgg on v|T	p-value v	T number	F on p-value v*	rgg on v*	rgg on v	rgg on v|T
0.25	1	1.33	0.50	0.50	0.25	0.10	1	2.71	0.79	0.79	0.62
0.25	1.25	1.64	0.62	0.50	0.31	0.10	2	3.84	0.86	0.79	0.68
0.25	1.7	2.07	0.72	0.50	0.36	0.10	4	5.02	0.89	0.79	0.70
0.25	2.5	2.71	0.79	0.50	0.40	0.10	10	6.63	0.92	0.79	0.73
0.25	5	3.84	0.86	0.50	0.43	0.10	20	7.88	0.93	0.79	0.73
0.25	10	5.02	0.89	0.50	0.45	0.10	100	10.83	0.95	0.79	0.75
0.25	25	6.63	0.92	0.50	0.46	0.10	200	12.12	0.96	0.79	0.76
0.25	50	7.88	0.93	0.50	0.47	0.05	1	3.84	0.86	0.86	0.74
0.25	250	10.83	0.95	0.50	0.48	0.05	2	5.02	0.89	0.86	0.77
0.25	500	12.12	0.96	0.50	0.48	0.05	5	6.63	0.92	0.86	0.79
0.20	1	1.64	0.62	0.62	0.38	0.05	10	7.88	0.93	0.86	0.80
0.20	1.3	2.07	0.72	0.62	0.45	0.05	50	10.83	0.95	0.86	0.82
0.20	2	2.71	0.79	0.62	0.49	0.05	100	12.12	0.96	0.86	0.83
0.20	4	3.84	0.86	0.62	0.53	0.025	1	5.02	0.89	0.89	0.79
0.20	8	5.02	0.89	0.62	0.55	0.025	2.5	6.63	0.92	0.89	0.82
0.20	20	6.63	0.92	0.62	0.57	0.025	5	7.88	0.93	0.89	0.83
0.20	40	7.88	0.93	0.62	0.58	0.025	25	10.83	0.95	0.89	0.85
0.20	200	10.83	0.95	0.62	0.59	0.025	50	12.12	0.96	0.89	0.85
0.20	400	12.12	0.96	0.62	0.60	0.010	1	6.63	0.92	0.92	0.85
0.15	1	2.07	0.72	0.72	0.52	0.010	2	7.88	0.93	0.92	0.86
0.15	1.5	2.71	0.79	0.72	0.57	0.010	10	10.83	0.95	0.92	0.87
0.15	3	3.84	0.86	0.72	0.62	0.010	20	12.12	0.96	0.92	0.88
0.15	6	5.02	0.89	0.72	0.64	0.005	1	7.88	0.93	0.93	0.86
0.15	15	6.63	0.92	0.72	0.66	0.005	5	10.83	0.95	0.93	0.88
0.15	30	7.88	0.93	0.72	0.67	0.005	10	12.12	0.96	0.93	0.89
0.15	150	10.83	0.95	0.72	0.68						
0.15	300	12.12	0.96	0.72	0.69						

T: number of treatments, v: distribution of the p-value v, v* = v|T: distribution of the p-value v*, r_{gg}: accuracy. The amount r_{gg} on v|T stands for r_{gg} on the conditional v|T, which is v given T.

stands for r_{gg} on the conditional $v|T$, which is v given T. We also have $r_{ggc} = r_{gg}$ on $v|T$. Then, in the example, it can be learnt the triple (*p-value* = 10%; T number = 10; r_{gg} on $v|T$ = 0.73).

Other relevant triples are extracted from and highlighted in the Table 5: (10%; 200; 0.76); (5%; 100; 0.83); (2.5%; 50; 0.85); (1%; 20; 0.88); (0.5%; 10; 0.89). It can be noticed that, if the T values are sufficient high, the corrected r_{gg} are not very different from that obtained in Table 3. In these bases it can be concluded that p-values equal or lower than 10% can display high and very high accuracies (Tables 3 and 5). It can also be learnt from Table 5 that the v p-values of 15% and of 20% (with $T > 4$) can display moderate accuracies and so can be considered in selection. On the other hand, p value of 25% showed to be provided low accuracies whatever the T are. So, under these circumstances p value of 25% is unsuitable for selection.

SAMPLE SIZES FOR DETECTING SIGNIFICANCE OF TREATMENT EFFECTS

Statistical reference books (Snedecor and Cochran 1967, Steel and Torrie 1980) provide the general expression to calculate the sample size (n) needed to detect significance of treatment effects, as: $[(z_\alpha + z_\beta)^2\ \sigma_D^2\ ]\delta^2$, where z_α and z_β are values of the cumulative distribution functions of Type I (α) and Type II (β) errors, under one-sided hypothesis tests; σ_D^2 is the variance of the difference between the means of two treatments; and δ is the size of the actual difference between two means that is intended to be declared significant. The quantity (1-β) is the probability (power) that the experiment presents a significant difference between the means of the treatments. In practice, powers of 80% and 90% are common and suitable. The σ_D^2 is a function of the residual variance (given as a function of $1 - h^2$) and δ^2 can be taken as the squared difference between an effect and the mass zero point (given as a function of h^2).

We then have $n = (z_\alpha + z_\beta)^2\ (1 - h^2)/h^2$. Considering $n = (F - 1)(1 - h^2)h^2$ (discussed in the previous topic), we thus have $(F - 1) = (z_\alpha + z_\beta)^2 = NCP$, which is the non-centrality parameter. The values of $(z_\alpha + z_\beta)^2$ were determined by Snedecor and Cochran (1967) as presented in Table 6. Therefore, we also have $(z_\alpha + z_\beta)^2 = \frac{n}{\lambda} = \frac{nh^2}{1-h^2}$.

From Table 6, an accuracy of 90% is associated with α equal to 10% and β equal to 80%, among other combinations of α and β. A summary of these results is presented in Table 7.

We can see that, to perform an experiment with the desired power (1-β) of 0.90 of the F test and significance of 0.05, an accuracy of 0.95 is necessary. In this case, the probability of detecting a true difference between the genotypes is 0.90 when the significance level is set at 0.05. As expected, there is a proximity between the accuracy ($r_{\hat{g}g}$) and the degree of confidence (1-α). Lee and Bjornstad (2013) demonstrated that hypothesis testing is equivalent to predicting discrete random effects, while Pawitan and Lee (2020) showed that confidence is likelihood and confidence density is, in fact, an extended likelihood.

Furthermore, a relationship between power and the coefficient of determination or square correlation ($r^2_{\hat{g}g}$) seems to exist for these high accuracy values. The coefficient of determination is also called the proportional error reduction (Linder 1951, Ceapoiu 1968) and is a measure of the proportion of coincidence, hits, correctness, or effectiveness.

Table 6. Values of $(z_\alpha + z_\beta)^2$ in one-sided test for significance levels α determined by Snedecor and Cochran (1967)

(1 − β)	$(z_\alpha + z_\beta)^2$ in one-sided test for significance levels α		
	0.01	**0.05**	**0.10**
0.80	10.0	6.2	4.5
0.90	13.0	8.6	6.6
0.95	15.8	10.8	8.6

With α = 5% and β = 90%, NCP = 8.6 and F = 9.6; thus, $r^2_{\hat{g}g} = 0.90$ and $r_{\hat{g}g} = 0.95$. With α = 5% and β = 80%, NCP = 6.2 and F = 7.2; thus, $r^2_{\hat{g}g} = 0.86$ and $r_{\hat{g}g} = 0.93$. With α = 5% and β = 80%, NCP = 4.5 and F = 5.5; thus, $r^2_{\hat{g}g} = 0.82$ and $r_{\hat{g}g} = 0.91$.

NUMBER OF REPLICATIONS PER GENETIC TREATMENT

The results of the numerical evaluations to determine the number of repetitions in experiments in a single-environment trial are presented in Table 8. Individual

Table 7. Significance level and power of the F test associated with the required accuracy levels of 0.90, 0.93, and 0.95

Accuracy ($r_{\hat{g}g}$)	$r^2_{\hat{g}g}$	Significance (Type I error: α)	Confidence (1-α)	Power (1-β)	Type II error: β	F Test
0.90	0.82	0.10	0.90	0.80	0.20	5.5
0.93	0.86	0.05	0.95	0.80	0.20	7.2
0.95	0.90	0.05	0.95	0.90	0.10	9.6

heritability (h^2) ranging from 0.05 to 0.95 were considered to achieve accuracies ($r_{\hat{g}g}$) ranging from 0.50 to 0.99. To determine the number of repetitions in experiments in a single-environment trial, the equation

$$n = \frac{r^2_{\hat{g}g}}{h^2}\frac{(1-h^2)}{(1-r^2_{\hat{g}g})} = \frac{r^2_{\hat{g}g}}{(1-r^2_{\hat{g}g})}\frac{(1-h^2)}{h^2}$$ was used (Resende et al. 2014, Resende 2015).

For traits with an individual heritability (h^2) of 0.20, an *n* equal to 17, 37, and 197 repetitions of single tree plots are required to achieve accuracies ($r_{\hat{g}g}$) equal to 0.90, 0.95, and 0.99, respectively (Table 8). As 90% is a very high accuracy (Resende and Duarte 2007, Resende and Alves 2020), 17 repetitions can be recommended. For traits with high heritability, for example 0.50, the recommended number of repetitions to achieve 90% accuracy is four (Table 8). Another way to apply these results is to use the estimated heritability of the breeding program itself in the environment in which it is conducted.

The results of the numerical evaluations to determine the number of repetitions in multi-environment trials are presented in Table 9. Individual heritability (h^2) ranging from 0.20 to 0.50 and genetic correlation across environments (r_{ge}) ranging from 0.60 to 1.00 were considered to achieve accuracies ($r_{\hat{g}g}$) of 0.70, 0.80, and 0.90. The number of repetitions

Table 8. Number of repetitions (*n*) in a single-environment trial for traits with individual heritability (h^2) ranging from 0.05 to 0.95 to reach accuracies ($r_{\hat{g}g}$) ranging from 0.50 to 0.99

Very high		**Very high**		**Very high**		**Very high**		**High**		**High**	
***$r_{\hat{g}g}$* 0.99**	**F 50.25**	***$r_{\hat{g}g}$* 0.975**	**F 20.25**	***$r_{\hat{g}g}$ 0.95***	***F 10.26***	***$r_{\hat{g}g}$ 0.90***	***F 5.26***	***$r_{\hat{g}g}$ 0.85***	***F 3.60***	***$r_{\hat{g}g}$ 0.80***	***F 2.78***
h^2_g	**N**	h^2_g	N	h^2_g	n	h^2_g	n	h^2_g	n	**h^2_g**	**n**
0.05	936	0.05	366	0.05	176	0.05	81	0.05	49	0.05	34
0.10	443	0.10	173	0.10	83	0.10	38	0.10	23	0.10	16
0.15	279	0.15	109	0.15	52	0.15	24	0.15	15	0.15	10
0.20	197	0.20	77	0.20	37	0.20	17	0.20	10	0.20	7.1
0.25	148	0.25	58	0.25	28	0.25	13	0.25	8	0.25	5
0.30	115	0.30	45	0.30	22	0.30	10	0.30	6	0.30	4.1
0.35	91	0.35	36	0.35	17	0.35	8	0.35	5	0.35	3
0.40	74	0.40	29	0.40	14	0.40	6	0.40	4	0.40	2.7
0.45	60	0.45	24	0.45	11	0.45	5	0.45	3	0.45	2
0.50	49	0.50	19	0.50	9	0.50	4	0.50	3	0.50	1.8
0.60	33	0.60	13	0.60	6	0.60	3	0.60	2	0.60	1
0.70	21	0.70	8	0.70	4	0.70	2	0.70	1	0.70	1
0.80	12	0.80	5	0.80	2	0.80	1	0.80	1	0.80	0
0.90	5	0.90	2	0.90	1	0.90	0	0.90	0	0.90	0
0.95	3	0.95	1	0.95	0	0.95	0	0.95	0	0.95	0
High		**High**		**Moderate**		**Moderate**		**Moderate**		**Moderate**	
***$r_{\hat{g}g}$* 0.75**	**F 2.29**	***$r_{\hat{g}g}$ 0.70***	**F 1.96**	***$r_{\hat{g}g}$ 0.65***	**F 1.73**	***$r_{\hat{g}g}$ 0.60***	**F 1.56**	***$r_{\hat{g}g}$ 0.55***	**F 1.43**	***$r_{\hat{g}g}$ 0.50***	**F 1.33**
h^2_g	**N**	h^2_g	n	h2g	n	h^2_g	n	h^2_g	n	**h^2_g**	**n**
0.05	24	0.05	18	0.05	14	0.05	11	0.05	8	0.05	6
0.10	12	0.10	9	0.10	7	0.10	5	0.10	4	0.10	3
0.15	7	0.15	5	0.15	4	0.15	3	0.15	2	0.15	2
0.20	5	0.20	4	0.20	3	0.20	2	0.20	2	0.20	1
0.25	4	0.25	3	0.25	2	0.25	2	0.25	1	0.25	1
0.30	3	0.30	2	0.30	2	0.30	1	0.30	1	0.30	1
0.35	2	0.35	2	0.35	1	0.35	1	0.35	1	0.35	1
0.40	2	0.40	1	0.40	1	0.40	1	0.40	1	0.40	1
0.45	2	0.45	1	0.45	1	0.45	1	0.45	1	0.45	0
0.50	1	0.50	1	0.50	1	0.50	1	0.50	0	0.50	0
0.60	1	0.60	1	0.60	0	0.60	0	0.60	0	0.60	0
0.70	1	0.70	0	0.70	0	0.70	0	0.70	0	0.70	0
0.80	0	0.80	0	0.80	0	0.80	0	0.80	0	0.80	0
0.90	0	0.90	0	0.90	0	0.90	0	0.90	0	0.90	0
0.95	0	0.95	0	0.95	0	0.95	0	0.95	0	0.95	0

Table 9. Number of repetitions (*n*) in multi-environment trials (*l* environments), for traits with individual heritability (h^2) ranging from 0.20 to 0.50, and genetic correlation across environments (r_{ge}) ranging from 0.60 to 1.00, to achieve accuracies ($r_{\hat{g}g}$) of 0.70, 0.80, and 0.90

$r_{\hat{g}g}$ = 0.90	F=5.26	l=1	l=2	l=3	l=4	l=5					
h^2_g	r_{ge}	n	n	n	n	n	n, l=1	n, l=2	n, l=3	n, l=4	n, l=5
0.2	1	17.1	8.5	5.7	4.3	3.4	17.1	17.1	17.1	17.1	17.1
0.2	0.9	-	10.9	6.6	4.7	3.7	-	19.7	19.7	18.8	18.3
0.2	0.8	-	17.1	8.3	5.4	4.1	-	24.8	24.8	21.8	20.3
0.2	0.7	-	88.0	13.0	7.0	4.8	-	38.9	38.9	28.0	24.0
0.2	0.6	-	-	-	12.3	6.6	-	-	-	49.1	32.9
0.3	1	9.9	5.0	3.3	2.5	2.0	9.9	9.9	9.9	9.9	9.9
0.3	0.9	-	6.2	3.8	2.7	2.1	-	11.3	11.3	10.7	10.5
0.3	0.8	-	9.5	4.6	3.0	2.3	-	13.8	13.8	12.1	11.3
0.3	0.7	-	47.0	6.9	3.7	2.6	-	20.8	20.8	14.9	12.8
0.3	0.6	-	-	-	6.1	3.3	-	-	-	24.5	16.5
0.4	1	6.4	3.2	2.1	1.6	1.3	6.4	6.4	6.4	6.4	6.4
0.4	0.9	-	3.9	2.3	1.7	1.3	-	7.0	7.0	6.7	6.5
0.4	0.8	-	5.7	2.8	1.8	1.4	-	8.3	8.3	7.3	6.8
0.4	0.7	-	26.4	3.9	2.1	1.4	-	11.7	11.7	8.4	7.2
0.4	0.6	-	-	-	3.1	1.6	-	-	-	12.3	8.2
0.5	1	4.3	2.1	1.4	1.1	0.9	4.3	4.3	4.3	4.3	4.3
0.5	0.9	-	2.5	1.5	1.1	0.8	-	4.5	4.5	4.3	4.2
0.5	0.8	-	3.4	1.7	1.1	0.8	-	5.0	5.0	4.4	4.1
0.5	0.7	-	14.1	2.1	1.1	0.8	-	6.2	6.2	4.5	3.8
0.5	0.6	-	-	-	1.2	0.7	-	-	-	4.9	3.3
$r_{\hat{g}g}$ = 0.80	F=2.78	l=1	l=2	l=3	l=4	l=5					
h^2_g	r_{ge}	n	n	n	n	n	n, l=1	n, l=2	n, l=3	n, l=4	n, l=5
0.2	1	7.1	3.6	2.4	1.8	1.4	7.1	7.1	7.1	7.1	7.1
0.2	0.9	-	3.8	2.5	1.8	1.4	-	7.7	7.4	7.3	7.2
0.2	0.8	-	4.3	2.6	1.9	1.5	-	8.6	7.8	7.5	7.3
0.2	0.7	-	5.1	2.8	2.0	1.5	-	10.3	8.5	7.8	7.5
0.2	0.6	-	7.3	3.3	2.1	1.6	-	14.5	9.8	8.4	7.8
0.3	1	4.1	2.1	1.4	1.0	0.8	4.1	4.1	4.1	4.1	4.1
0.3	0.9	-	2.2	1.4	1.0	0.8	-	4.4	4.2	4.2	4.1
0.3	0.8	-	2.4	1.4	1.0	0.8	-	4.8	4.3	4.2	4.1
0.3	0.7	-	2.7	1.5	1.0	0.8	-	5.5	4.5	4.2	4.0
0.3	0.6	-	3.6	1.6	1.1	0.8	-	7.3	4.9	4.2	3.9
0.4	1	2.7	1.3	0.9	0.7	0.5	2.7	2.7	2.7	2.7	2.7
0.4	0.9	-	1.4	0.9	0.6	0.5	-	2.7	2.6	2.6	2.6
0.4	0.8	-	1.4	0.9	0.6	0.5	-	2.9	2.6	2.5	2.4
0.4	0.7	-	1.5	0.9	0.6	0.4	-	3.1	2.6	2.4	2.2
0.4	0.6	-	1.8	0.8	0.5	0.4	-	3.6	2.4	2.1	1.9
0.5	1	1.8	0.9	0.6	0.4	0.4	1.8	1.8	1.8	1.8	1.8
0.5	0.9	-	0.9	0.6	0.4	0.3	-	1.8	1.7	1.7	1.6
0.5	0.8	-	0.9	0.5	0.4	0.3	-	1.7	1.6	1.5	1.5
0.5	0.7	-	0.8	0.5	0.3	0.2	-	1.6	1.4	1.3	1.2
0.5	0.6	-	0.7	0.3	0.2	0.2	-	1.5	1.0	0.8	0.8
$r_{\hat{g}g}$ = 0.70	F=1.96	l=1	l=2	l=3	l=4	l=5					
h^2_g	r_{ge}	n	n	n	n	n	n, l=1	n, l=2	n, l=3	n, l=4	n, l=5
0.2	1	3.8	1.9	1.3	1.0	0.8	3.8	3.8	3.8	3.8	3.8
0.2	0.9	-	2.0	1.3	1.0	0.8	-	3.9	3.9	3.8	3.8
0.2	0.8	-	2.0	1.3	1.0	0.8	-	4.1	3.9	3.8	3.8
0.2	0.7	-	2.2	1.3	1.0	0.7	-	4.3	4.0	3.8	3.7

0.2	0.6	-	2.4	1.4	1.0	0.7	-	4.7	4.1	3.8	3.7
0.3	1	2.2	1.1	0.7	0.6	0.4	2.2	2.2	2.2	2.2	2.2
0.3	0.9	-	1.1	0.7	0.5	0.4	-	2.3	2.2	2.2	2.2
0.3	0.8	-	1.1	0.7	0.5	0.4	-	2.3	2.2	2.1	2.1
0.3	0.7	-	1.2	0.7	0.5	0.4	-	2.3	2.1	2.0	2.0
0.3	0.6	-	1.2	0.7	0.5	0.4	-	2.4	2.0	1.9	1.8
0.4	1	1.4	0.7	0.5	0.4	0.3	1.4	1.4	1.4	1.4	1.4
0.4	0.9	-	0.7	0.5	0.3	0.3	-	1.4	1.4	1.4	1.4
0.4	0.8	-	0.7	0.4	0.3	0.3	-	1.4	1.3	1.3	1.3
0.4	0.7	-	0.6	0.4	0.3	0.2	-	1.3	1.2	1.1	1.1
0.4	0.6	-	0.6	0.3	0.2	0.2	-	1.2	1.0	1.0	0.9
0.5	1	1.0	0.5	0.3	0.2	0.2	1.0	1.0	1.0	1.0	1.0
0.5	0.9	-	0.5	0.3	0.2	0.2	-	0.9	0.9	0.9	0.9
0.5	0.8	-	0.4	0.3	0.2	0.2	-	0.8	0.8	0.8	0.8
0.5	0.7	-	0.3	0.2	0.2	0.1	-	0.7	0.6	0.6	0.6
0.5	0.6	-	0.2	0.1	0.1	0.1	-	0.5	0.4	0.4	0.4

in multi-environment trials is given as the expression $n = \frac{r^2_{\hat{g}g}(r_{ge} - h^2_g)}{(1 - r^2_{\hat{g}g})lh^2_g r_{ge} - r^2_{\hat{g}g} h^2_g(1 - r_{ge})}$), which is a function of individual heritability (h^2), genetic correlation across environments (r_{ge}), accuracy ($r_{\hat{g}g}$), and number of environments (l).

To achieve accuracies ($r_{\hat{g}g}$) equal to 0.90, 0.80, and 0.70 for traits with an individual heritability (h^2) of 0.20, genetic correlation across sites (r_{ge}) of 0.80, in three environments (l), an n equal to 8.3, 2.6, and 1.3, respectively, is required per environment. Thus, for an accuracy of 0.90, across all environments, 8.3 * 3 = 24.9 repetitions of each genetic material in the entire experimental network is required (Table 9). For traits with high heritability, for example 0.50, the recommended number of repetitions is 1.7 * 3 = 5.1 to achieve 90% accuracy (Table 9).

The values found for multiple environments (24.9 and 5.1) (Table 9) differ from the values found for single environments (17.0 and 4.0) (Table 8) as the genetic correlation across environments (r_{ge}) in Table 9 is taken as 0.80 while in Table 8 r_{ge} is implicitly equivalent to 1.00. Table 9 shows that with r_{ge} = 1, the values 17.1 and 4.3 are obtained, suggesting a coherence between the two alternative approaches to determine the number of repetitions (n).

NUMBER OF TRIALS AS A FUNCTION OF GENOTYPE X ENVIRONMENT CORRELATION

The results of the simulations to determine the number of sites (l) in multi-environment trials are presented in Table 10. We consider individual heritability (h^2) ranging from 0.20 to 0.50 and genetic correlation across environments (r_{ge}) ranging from 0.60 to 1.00 to reach accuracies ($r_{\hat{g}g}$) of 0.70, 0.80, and 0.90. The expression $l = \frac{r^2_{\hat{g}g}\left[\frac{\left(1 - h^2_g - \frac{h^2_g(1 - r_{ge})}{r_{ge}}\right)}{nh^2_g} + \frac{(1 - r_{ge})}{r_{ge}}\right]}{(1 - r^2_{\hat{g}g})}$ was used.

For traits with h^2 of 0.20, a r_{ge} equal to 0.80, and n equal to five per trial, and to achieve accuracies ($r_{\hat{g}g}$) equal to 0.90, 0.80, and 0.70, the number of trials required (l) is 4.3, 1.8, and 1.0. Thus, choosing an accuracy of 0.90 requires 4.3 * 5 = 21.3 repetitions of each genetic material in the entire experimental network (Table 10). For traits with high heritability, for example 0.50, the recommended number of repetitions is 1.7 * 5 = 8.5 to achieve 90% accuracy (Table 10).

The resulting values of 24.9 and 5.1 differ from the values 17.0 and 4.0 in Table 8 as here the genetic correlation across environments (r_{ge}) is taken as 0.80 and in Table 8 r_{ge} is implicitly equivalent to 1. Table 10 shows that with r_{ge} = 1, values of 17.1 and 4.3 are obtained, demonstrating the coherence between the two alternative approaches to determine the number of repetitions (n).

USE OF ACCURACY IN DETERMINING THE OPTIMAL PLOT SIZE

Appropriate approaches to determine the optimal plot size to evaluate p progenies, should be performed by setting

Table 10. Number of sites (*l*), conditional on the number of repetitions taken as n=2, n=3, n=4 and n=5, for traits with individual heritability (h^2) ranging from 0.20 to 0.50 and genetic correlation across environments (r_{ge}) ranging from 0.60 to 1.00, to achieve accuracies ($r_{\hat{g}g}$) of 0.70, 0.80, and 0.90

$r_{\hat{g}g}$ =0.90	F=5.02												
h^2_g	r_{ge}	n=2	n=3	n=4	n=5	l, n=2	l, n=3	l, n=4	l, n=5	n=2xl, n=2	n=3xl, n=3	n=4xl, n=4	n=5xl, n=5
0.2	1	2	3	4	5	8.5	5.7	4.3	3.4	17.1	17.1	17.1	17.1
0.2	0.9	2	3	4	5	8.8	6.0	4.6	3.8	17.5	18.0	18.5	18.9
0.2	0.8	2	3	4	*5*	9.1	6.4	5.1	*4.3*	18.1	19.2	20.3	*21.3*
0.2	0.7	2	3	4	5	9.4	6.9	5.6	4.9	18.9	20.7	22.5	24.4
0.2	0.6	2	3	4	5	9.9	7.6	6.4	5.7	19.9	22.7	25.6	28.4
0.3	1	2	3	4	5	5.0	3.3	2.5	2.0	9.9	9.9	9.9	9.9
0.3	0.9	2	3	4	5	5.2	3.6	2.8	2.4	10.4	10.9	11.4	11.8
0.3	0.8	2	3	4	5	5.5	4.0	3.3	2.8	11.0	12.1	13.1	14.2
0.3	0.7	2	3	4	5	5.9	4.5	3.9	3.5	11.8	13.6	15.4	17.3
0.3	0.6	2	3	4	5	6.4	5.2	4.6	4.3	12.8	15.6	18.5	21.3
0.4	1	2	3	4	5	3.2	2.1	1.6	1.3	6.4	6.4	6.4	6.4
0.4	0.9	2	3	4	5	3.4	2.4	2.0	1.7	6.9	7.3	7.8	8.3
0.4	0.8	2	3	4	5	3.7	2.8	2.4	2.1	7.5	8.5	9.6	10.7
0.4	0.7	2	3	4	5	4.1	3.3	3.0	2.7	8.2	10.0	11.9	13.7
0.4	0.6	2	3	4	5	4.6	4.0	3.7	3.6	9.2	12.1	14.9	17.8
0.5	1	2	3	4	5	2.1	1.4	1.1	0.9	4.3	4.3	4.3	4.3
0.5	0.9	2	3	4	5	2.4	1.7	1.4	1.2	4.7	5.2	5.7	6.2
0.5	0.8	2	3	4	*5*	2.7	2.1	1.9	*1.7*	5.3	6.4	7.5	*8.5*
0.5	0.7	2	3	4	5	3.0	2.6	2.4	2.3	6.1	7.9	9.7	11.6
0.5	0.6	2	3	4	5	3.6	3.3	3.2	3.1	7.1	9.9	12.8	15.6
$r_{\hat{g}g}$ = 0.80	F=2.78												
h^2_g	r_{ge}	n=2	n=3	n=4	n=5	l, n=2	l, n=3	l, n=4	l, n=5	n=2xl, n=2	n=3xl, n=3	n=4xl, n=4	n=5xl, n=5
0.2	1	2	3	4	5	3.6	2.4	1.8	1.4	7.1	7.1	7.1	7.1
0.2	0.9	2	3	4	5	3.7	2.5	1.9	1.6	7.3	7.5	7.7	7.9
0.2	0.8	2	3	4	5	3.8	2.7	2.1	1.8	7.6	8.0	8.4	8.9
0.2	0.7	2	3	4	5	3.9	2.9	2.3	2.0	7.9	8.6	9.4	10.2
0.2	0.6	2	3	4	5	4.1	3.2	2.7	2.4	8.3	9.5	10.7	11.9
0.3	1	2	3	4	5	2.1	1.4	1.0	0.8	4.1	4.1	4.1	4.1
0.3	0.9	2	3	4	5	2.2	1.5	1.2	1.0	4.3	4.5	4.7	4.9
0.3	0.8	2	3	4	5	2.3	1.7	1.4	1.2	4.6	5.0	5.5	5.9
0.3	0.7	2	3	4	5	2.5	1.9	1.6	1.4	4.9	5.7	6.4	7.2
0.3	0.6	2	3	4	5	2.7	2.2	1.9	1.8	5.3	6.5	7.7	8.9
0.4	1	2	3	4	5	1.3	0.9	0.7	0.5	2.7	2.7	2.7	2.7
0.4	0.9	2	3	4	5	1.4	1.0	0.8	0.7	2.9	3.1	3.3	3.5
0.4	0.8	2	3	4	5	1.6	1.2	1.0	0.9	3.1	3.6	4.0	4.4
0.4	0.7	2	3	4	5	1.7	1.4	1.2	1.1	3.4	4.2	5.0	5.7
0.4	0.6	2	3	4	5	1.9	1.7	1.6	1.5	3.9	5.0	6.2	7.4
0.5	1	2	3	4	5	0.9	0.6	0.4	0.4	1.8	1.8	1.8	1.8
0.5	0.9	2	3	4	5	1.0	0.7	0.6	0.5	2.0	2.2	2.4	2.6
0.5	0.8	2	3	4	5	1.1	0.9	0.8	0.7	2.2	2.7	3.1	3.6
0.5	0.7	2	3	4	5	1.3	1.1	1.0	1.0	2.5	3.3	4.1	4.8
0.5	0.6	2	3	4	5	1.5	1.4	1.3	1.3	3.0	4.1	5.3	6.5
$r_{\hat{g}g}$ = 0.70	F=1.96												
h^2_g	r_{ge}	n=2	n=3	n=4	n=5	l, n=2	l, n=3	l, n=4	l, n=5	n=2xl, n=2	n=3xl, n=3	n=4xl, n=4	n=5xl, n=5
0.2	1	2	3	4	5	1.9	1.3	4.0	0.8	3.8	3.8	16.0	3.8
0.2	0.9	2	3	4	5	2.0	1.4	4.0	0.9	3.9	4.1	16.0	4.3
0.2	0.8	2	3	4	5	2.0	1.4	4.0	1.0	4.1	4.3	16.0	4.8
0.2	0.7	2	3	4	5	2.1	1.6	4.0	1.1	4.3	4.7	16.0	5.5

0.2	0.6	2	3	4	5	2.2	1.7	4.0	1.3	4.5	5.1	16.0	6.4
0.3	1	2	3	4	5	1.1	0.7	4.0	0.4	2.2	2.2	16.0	2.2
0.3	0.9	2	3	4	5	1.2	0.8	4.0	0.5	2.3	2.5	16.0	2.7
0.3	0.8	2	3	4	5	1.2	0.9	4.0	0.6	2.5	2.7	16.0	3.2
0.3	0.7	2	3	4	5	1.3	1.0	4.0	0.8	2.7	3.1	16.0	3.9
0.3	0.6	2	3	4	5	1.4	1.2	4.0	1.0	2.9	3.5	16.0	4.8
0.4	1	2	3	4	5	0.7	0.5	4.0	0.3	1.4	1.4	16.0	1.4
0.4	0.9	2	3	4	5	0.8	0.6	4.0	0.4	1.5	1.7	16.0	1.9
0.4	0.8	2	3	4	5	0.8	0.6	4.0	0.5	1.7	1.9	16.0	2.4
0.4	0.7	2	3	4	5	0.9	0.8	4.0	0.6	1.9	2.3	16.0	3.1
0.4	0.6	2	3	4	5	1.0	0.9	4.0	0.8	2.1	2.7	16.0	4.0
0.5	1	2	3	4	5	0.5	0.3	4.0	0.2	1.0	1.0	16.0	1.0
0.5	0.9	2	3	4	5	0.5	0.4	4.0	0.3	1.1	1.2	16.0	1.4
0.5	0.8	2	3	4	5	0.6	0.5	4.0	0.4	1.2	1.4	16.0	1.9
0.5	0.7	2	3	4	5	0.7	0.6	4.0	0.5	1.4	1.8	16.0	2.6
0.5	0.6	2	3	4	5	0.8	0.7	4.0	0.7	1.6	2.2	16.0	3.5

the total area ($p * n * k$) of the experiment and conditioning the number of plants per plot (k) to the number of blocks (n) necessary to obtain an optimal accuracy, typically 0.90. This can be done following Storck et al. (2011) using the maximum curvature method conditioned to the desired accuracy value according to Resende and Duarte (2007). This accuracy depends on CV_g and CV_e, which provide a link between the maximum curvature (CV) and the accuracy based on the mean of the evaluated genotypes. Accuracy depends on the magnitude of the coefficient of experimental variation (CV_e), the number of repetitions (n), and the coefficient of genetic variation (CV_g), according to the alternative formula $\hat{r}_{\hat{g}g} = \{1/[1 + (CV_e^2/CV_g^2)/n]\}^{1/2}$.

On the other hand, alternative methods and applications to estimate the optimal plot size are based on the nonlinear relationship $CV(x) = A / XB$, where $CV(x)$ is the coefficient of variation for plots planned of different sizes (x), expressed as a number of base units. The maximum curvature point (Xo) of the function $CV(x) = A / XB$ is considered the optimal plot size (Meier and Lessman 1971). In this method, for values of X greater than Xo, the drop in $CV(x)$ is minimal and not efficient to reduce the experimental error. Considering that the accuracy is a function of CV_g / CV_e and n (Resende and Duarte 2007), it is possible to rewrite the function $CV(x) = A / XB$, incorporating the values of CV_e, with predefined values of n and accuracy (Storck et al. 2011). Thus, by fixing the magnitude of the selective accuracy and the number of repetitions in the design of an experiment and knowing the environmental variability (A and B) of the chosen area, we can prepare a suitable experimental plan by combining the number of repetitions and the plot size. Thus, this approach estimates the optimal plot size, relating the variability of the experimental area to the predetermined accuracy (Storck et al. 2011). Another option is to consider the desired accuracy as a function of individual heritability (h^2) and the coefficient of determination of plot effects (c^2), which measures the degree of environmental variation of the plot, indicating the appropriate values of k and n. Accuracy is given as: $r_{\hat{g}g} = \sqrt{\frac{nkh^2}{1 + (k-1)(h^2 + c^2) + (n-1)kh^2}}$ (Resende et al. 2001).

DISCUSSION

Genetic selection is the result of prediction and ranking and is central to genetic improvement programs. To measure the efficiency of such improvement, we must consider selection accuracy. Meanwhile, model selection is related to inference and hypothesis testing and is tangential to genetic improvement. Its effectiveness can be measured by the p-value, among other techniques. To estimate accuracy, models are fit via the estimation/prediction of their effects, variance parameters, and their precision. Model selection is associated with inferences about the presence of sufficient genetic variability and significance of the effects of other factors in the model, using hypothesis tests, associated with p-values or significance levels. However, questions often arise as to which one to use: accuracy or p-value? The present study shows that there is a link between the two and that both can be used simultaneously. High accuracy and effective model selection enhances the efficacy of the whole breeding program (Resende and Alves 2020).

Accuracy is one of the most important parameters in quantitative genetics and plant breeding. It is used to assess the quality of experiments and infer the reliability of predicted genotypic values and the statistical validity of the predictive and inferred results. In practical terms, accuracy is also used to compare alternative selection methods, to compute genetic gains with selection, and to plan the size of experiments. Thus, it constitutes the building blocks of statistical and genetic analyses (Resende 2002).

In a single-environment trial, accuracy values are obtained considering the heritability (h^2) and the number of repetitions (n) of each genotype. In multi-environment trials, accuracy is estimated considering the heritability (h^2), genotypic correlation across environments (r_{ge}), number of repetitions (n), and number of experimental environments (l). Conversely, an expected accuracy can be used to plan experimental size and can be inferred by choosing the number of replications (n) and experimental environments (l) (total sample size of a genotype). Selections must be based on several traits. In such a case, the most economically important and with lowest heritability is the most suitable choice for determining the replications and locations numbers.

This position paper aimed to situate and reflect on statistical significance, selection accuracy, and experimental precision in connection with the efficiency of experimentation as applied to genetic selection in plants. We derive equations for accuracy in multi-environment trials, extending the work of Resende and Duarte (2007), and develop a model with GxE interaction effects using genetic parameters and Snedecor's F statistic. Also, we consider estimators for n and l in single- and multi-environment trials. Furthermore, we propose a new methodology to classify accuracy based on statistical significance via the p-value.

CONCLUSIONS

The results referring to the number of repetitions (n) and environments (l) were given according to the coefficients of heritability (h^2) and genetic correlations across environments (r_{ge}). For traits with h^2 equal to 0.20, r_{ge} of 0.80, and l equal to three, and to achieve $r_{\hat{g}g}$ equal to 0.90, an n equal to 8.3 per environment is required. Thus, across all environments, n^*l = 8.3 * 3 = 24.9 repetitions of each genetic material is required.

The p-value can be inferred from tables of Snedecor/Fisher´s F, Student's t, and Bartlett/Pearson´s Chi-square test statistics, with large (tending to infinite) number of degrees of freedom for the residual. Therefore, a bridge between the p-value and accuracy can be established, expressing r_{gg} as a function of one of these three statistics. This link provides statisticians with information on the accuracy being accepted when practicing a certain p-value. For example, typical p-values of 0.10, 0.05, and 0.01 are associated with accuracies of 79%, 86%, and 92%, respectively. These traditional values for the 10%, 5%, and 1% cut-off points for significance were recently revised and a p-value that is now widely accepted is 0.5% (0.005) (Benjamin et al. 2018). With this p-value, the associated accuracy is 93%. This approach seems appropriate and can be recommended with confidence. Thus, for the final stages of breeding programs the pair p-value = 0.005 / $r_{\hat{g}g}$ = 93% is strongly suggested. Conversely, typical accuracy values of 50%, 70%, 80%, 90%, 93%, and 95% are associated with p-values of 25%, 16%, 10%, 2%, 0.5%, and 0.1 %, respectively. With the Bonferroni protection, p-values of up to 20% are acceptable to attest to the significance of genetic effects in models and to proceed with selection between models and between genotypes. The p-values below 20% provide r_{gg} above 50%, which are suitable to enable genetic gain.

ACKNOWLEDGMENTS

We acknowledge financial support from the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

REFERENCES

Baxevanos D, Korpetis E, Irakli M and Tsialtas IT (2017a) Evaluation of a durum wheat selection scheme under Mediterranean conditions: adjusting trial locations and replications. **Euphytica 213**: 1-14.

Baxevanos D, Tsialtas J, Vlachostergios D and Goulas C (2017b) Optimum replications and locations for cotton cultivar trials under Mediterranean conditions. **Journal of Agricultural Science 155**: 1553-1564.

Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, Bollen KA, Brembs B, Brown L, Camerer C, Cesarini D, Chambers CD, Clyde M, Cook TD, De Boeck P, Dienes Z, Dreber A, Easwaran K, Efferson C, Fehr E, Fidler F, Field AP, Forster M, George EI, Gonzalez R, Goodman S, Green E, Green DP, Greenwald AG, Hadfield JD, Hedges LV, Held L, Ho TH, Hoijtink H, Hruschka DJ, Imai K, Imbes G, Ioannidis

JPA, Jeon M, Jones JH, Kirchler M, Laibson D, List J, Little R, Lupia A, Machery E, Maxwell SE, McCarthy M, Moore DA, Morgan SL, Munafó M, Nakagawa S, Nyhan B, Parker TH, Pericchi L, Perugini M, Rouder J, Rousseau J, Savalei V, Schonbrodt FD, Sellke T, Sinclair B, Tingley D, Van Zandt T, Vazire S, Watts DJ, Winship C, Wolpert RL, Xie Y, Young C, Zinman J and Johnson VE (2018) Redefine statistical significance. **Nature Human Behaviour 2**: 6-10.

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. **Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8**: 3-62.

Ceapoiu N (1968) **Metode statistice aplicate in experientele agricole si biologice**. Agro-Silvica Bucuresti, 550p.

Cullis BR, Smith AB and Coombes NE (2006) On the design of early generation variety trials with correlated data. **Journal of Agricultural, Biological and Environmental Statistics 11**: 381-393.

Dias KOG, Piepho HP, Guimaraes LJM, Guimaraes PEO, Parentoni SN, Pinto MO, Noda RW, Magalhaes JV, Guimarães CT, Garcia AAF and Pastina MM (2020) Novel strategies for genomic prediction of untested single-cross maize hybrids using unbalanced historical data. **Theoretical and Applied Genetics 133**: 443-445.

Dickerson GE (1962) Implications of genetic-environmental interaction in animal breeding. **Animal Science 4:** 47-63.

Edwards LJ, Muller KE, Wolfinger RD, Qaqish BF and Schabenberger O (2008) An R^2 statistic for fixed effects in the linear mixed model. **Statistics in Medicine 27**: 6137-6157.

Fisher RA (1925) **Statistical methods for research workers**. Oliver and Boyd, Edinburgh and London, 239p.

George N and Lundy M (2019) Quantifying genotype x environment effects in long-term common wheat yield trials from an agroecologically diverse production region. **Crop Science 59**: 1960-1972.

Lee Y and Bjornstad JF (2013) Extended likelihood approach to large-scale multiple testing. **Journal of the Royal Statistical Society, Series B (Statistical Methodology) 75**: 553-575.

Lightstone AH (1978) **Mathematical logic: an introduction to model theory**. Springer, New York, 338p.

Linder A (1951) **Statistische methoden für naturwissenschaftler, mediziner und ingenieure**. Birkhäuser Verlag, Basel, 200p.

Meier VD and Lessman KJ (1971) Estimation of optimum field plot shape and size for testing yield in Crambe abyssinica Hochst. **Crop Science 11**: 648-650.

Mood AM, Graybill FA and Boes DC (1974) **Introduction to the theory of statistics**. McGraw-Hill, Tokyo, 564p.

Papoulis A and Pilla SU (1965) **Probability, random variables, and stochastic processes**. McGraw Hill, New York, 583p.

Pawitan Y and Lee Y (2020) Confidence as likelihood. **Statistical Science 36**: 509-517.

Piepho HP (2019) A coefficient of determination (R^2) for generalized linear mixed model. **Biometrical Journal 62**: 860-872.

Piepho HP and Mohring J (2007) Computing heritability and selection response from unbalanced plant breeding trials. **Genetics 177**: 1881-1888.

Resende MDV (1998) **Interação genótipo x ambiente e determinação do número adequado de locais de experimentação com base nas estatísticas F de Snedecor da análise de variância conjunta**. Embrapa Floresta, Colombo, p. 55-66. (Boletim de Pesquisa Florestal, 37).

Resende MDV (2002) **Genética biométrica e estatística no melhoramento de plantas perenes**. Embrapa Informação Tecnológica, Brasília, 975p.

Resende MDV (2007) **Matemática e estatística na análise de experimentos e no melhoramento genético**. Embrapa Florestas, Colombo, 362p.

Resende MDV (2015) **Genética quantitativa e de populações**. Suprema, Visconde do Rio Branco, 463p.

Resende MDV and Alves RS (2020) Linear, generalized, hierarchical, bayesian and random regression mixed models in genetics/genomics in plant breeding. **Functional Plant Breeding Journal 2**: 1-31.

Resende MDV and Duarte JB (2007) Precisão e controle de qualidade em experimentos de avaliação de cultivares. **Pesquisa Agropecuária Tropical 37**: 182-194.

Resende MDV, Furlani-Junior E, Moraes MLT and Fazuoli LC (2001) Estimativas de parâmetros genéticos e predição de valores genotípicos no melhoramento do cafeeiro pelo procedimento REML/BLUP. **Bragantia 60**:185-193.

Resende MDV, Silva FF and Azevedo CF (2014) **Estatística matemática, biométrica e computacional: modelos mistos, categóricos e generalizados (REML/BLUP), inferência Bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência**. Suprema, Visconde do Rio Branco, 881p.

Schmidt P, Hartung J, Rath J and Piepho HP (2019) Estimating broad-sense heritability with unbalanced data from agricultural cultivar trials. **Crop Science 59**: 525-536.

Searle SR, Casella G and McCulloch CE (1992) **Variance components**. Wiley, New York, 536p.

Snedecor GW and Cochran WR (1967) **Statistical methods**. Iowa State University Press, Ames, 274p.

Steel RGD and Torrie JH (1980) **Principles and procedures of statistics: a biometrical approach**. McGraw-Hill, New York, 666p.

Storck L, Lopes SJ, Lúcio ADC and Cargnelutti Filho A (2011) Optimum plot size and number of replications related to selective precision. **Ciência Rural 41**: 390-396.

Van Vleck LD, Pollak EJ, Pollak EJ and Branford EA (1987) **Genetics for animal sciences**. W.H. Freeman, San Francisco, 391p.

Vencovsky R (1987) Herança quantitativa. In Paterniani E and Viégas GP (Org) **Melhoramento e produção do milho no Brasil**. Fundação Cargil,

Campinas, p. 122-199.

Woyann LG, Zdziarski AD, Zanella R, Rosa AC, Conte J, Meira D, Storck L and Benin G (2020) Optimal number of replications and test locations for soybean yield trials in Brazil. **Euphytica 216**: 1-9.

Xu N, Jin SQ and Li J (2016) Designing the national cotton variety trials regarding the number of replicates and number of test locations in China. **Acta Agronomica Sinica 42**: 43-50.

Yan W (2021) Estimation of the optimal number of replicates in crop variety trials. **Frontiers in Plant Science 11**: 2231.

Yan W, Frégeau-Reid J, Martin R, Pageau D and Mitchell-Fetch J (2015) How many test locations and replications are needed in crop variety trials for a target region? **Euphytica 202**: 361-372.

Zhang Y, Xu NY, Guo LL, Yang ZG, Zhang XQ and Yang XN (2020) Optimization of test locations number and replication number in regional winter wheat variety trials in northern China. **Acta Agronomica Sinica 46**: 1166-1173.

APPENDIX

DERIVATION OF ACCURACY ESTIMATORS

Overall estimator for accuracy

Accuracy ($r_{\hat{g}g}$) is a correlation coefficient, given as:

$r_{\hat{g}g} = cov(\hat{g}, g)/[var(\hat{g})var(g)]^{1/2} = var(g)/[var(\hat{g})var(g)]^{1/2} = [var(g)]^{(1/2)}/[var(\hat{g})^{1/2}$ [*Equation 1*].

Expressed in terms of reliability (accuracy squared or heritability at the treatment mean level - h^2_m) of the estimation/prediction, for the case of a single-trial:

$r^2_{\hat{g}g} = \frac{var(g)}{var(\hat{g})} = h^2_m = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e/n} = \frac{nh^2}{1+(n-1)h^2}$ *[Equation 2]*;

where $h^2 = \frac{\sigma^2_g}{(\sigma^2_g + \sigma^2_e)}$ *[Equation 3]* is heritability at the plot or individual plant level and *n* is the number of replicates.

Equations for reliability and number of replicates (*n*) for single-environment trial

The reliability ($r^2_{\hat{g}g}$) is given as:

$r^2_{\hat{g}g} = h^2_m = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e/n} = \frac{nh^2}{1+(n-1)h^2}$ *[Equation 2]* or $r^2_{\hat{g}g} = \frac{nh^2}{1+(n-1)h^2} = \frac{n}{n+[(1-h^2)/h^2]} = \frac{n}{n+\lambda}$ *[Equation 4]*;

where $\lambda = [(1 - h^2)/h^2] = \sigma^2_e/\sigma^2_g$ *[Equation 5]* is shrinkage factor of mixed model equations for the best linear unbiased prediction (BLUP) of *g*.

Reliability depends on *n* and individual heritability.

Expression for the number of replications (*n*) in a single-environment trial considering heritability (h^2) and the desired selective reliability, derived from the isolation of *n* in [*Equation 2*]:

$n = \frac{r^2_{\hat{g}g}}{h^2}\frac{(1-h^2)}{(1-r^2_{\hat{g}g})} = \frac{r^2_{\hat{g}g}}{(1-r^2_{\hat{g}g})}\frac{(1-h^2)}{h^2}$ *[Equation 6]*. Or

$n = \frac{r^2_{\hat{g}g}}{(1-r^2_{\hat{g}g})}\frac{(1-h^2)}{h^2} = \frac{r^2_{\hat{g}g}}{(1-r^2_{\hat{g}g})}\lambda$ *[Equation 7]*.

The number of replicates (*n*) depends on expected/desired reliability (accuracy squared) and heritability.

Equations for reliability and number of replicates (*n*) for multi-environment trials

Expression for reliability (accuracy squared) for evaluation in multi-environment trials:

$r^2_{\hat{g}g} = h^2_m = \frac{\sigma^2_g}{\sigma^2_g + \frac{\sigma^2_{ge}}{l} + \frac{\sigma^2_e}{nl}} = \frac{h^2_g}{h^2_g + \frac{c^2_{ge}}{l} + \frac{(1-h^2_g - c^2_{ge})}{nl}}$ *[Equation 8]*.

Genetic correlation across environments (Dickerson 1962):

$r_{ge} = \frac{\sigma^2_g}{(\sigma^2_g + \sigma^2_{ge})}$ *[Equation 9]*.

Coefficient of determination of GxE interaction effects (Resende et al. 1998, Resende and Alves 2020):

$c^2_{ge} = \frac{h\sigma^2_g(1 - r_{ge})}{r_{ge}}$ *[Equation 10]*.

Coefficient of proportion between GxE interaction variance and genotypic variance (Resende 1998, Resende and Alves 2020):

$\frac{\sigma^2_{ge}}{\sigma^2_g} = \frac{(1 - r_{ge})}{r_{ge}}$ *[Equation 11]*.

Expression for the number of repetitions (*n*) in multi-environment trials given the heritability (h^2), desired reliability ($r^2_{\hat{g}g}$), and number of sites (*l*):

$n = \frac{r^2_{\hat{g}g}(1-h^2_g - c^2_{ge})}{(1 - r^2_{\hat{g}g})lh^2_g - r^2_{\hat{g}g}c^2_{ge}}$ *[Equation 12]*,

where $c^2_{ge} = \frac{h\sigma^2_g(1 - r_{ge})}{r_{ge}}$ *Equation [10]*. Thus, $n = \frac{r^2_{\hat{g}g}(r_{ge} - h^2_g)}{(1 - r^2_{\hat{g}g})lh^2_g r_{ge} - r^2_{\hat{g}g}h^2_g(1 - r_{ge})}$ *Equation [13]*.

If in *Equation 12*, $c^2_{ge} = 0$ and $l = 1$, *Equation 12* is equivalent to *Equation 6*. If in *Equation 13*, $r_{ge} = 1$ and $l = 1$, *Equation 13* is equivalent to *Equation 6*.

The number n depends on reliability (accuracy squared), c^2_g, l, and r_{ge}.

Number of sites in multi-environment trials via variety means per trial

Taking the expression for reliability (accuracy squared) in multi-environment trials (Resende 2007, Yan 2021):

$r^2_{\hat{g}g*} = \frac{h^2_g}{h^2_g + \frac{\sigma^2_{ge}}{l}} = \frac{lr_{ge})}{1+(l-1)\,r_{ge}}$ *[Equation 14]*; $l = \frac{r^2_{\hat{g}g}}{(1 - r^2_{\hat{g}g})} \frac{(1 - r_{ge})}{r_{ge}}$ *[Equation 15]*.

Number of sites (l) in multi-environment trials

Expression for reliability (accuracy squared) in multi-environment trials:

$r^2_{\hat{g}g} = \frac{h^2_g}{h^2_g + \frac{c^2_{ge}}{l} + \frac{(1-h^2_g - c^2_{ge})}{nl}}$ *[Equation 8]*.

Expression for the number of sites (l) in multi-environment trials for a trait with given heritability (h^2), desired reliability ($r^2_{\hat{g}g}$), and number of replications (n) derived from the isolation of l in *Equation 8*:

$$l = \frac{r^2_{\hat{g}g}[\frac{(1 - h^2_g - c^2_{ge})}{nh^2_g} + (c^2_{ge}/h^2_g)]}{(1 - r^2_{\hat{g}g})} = \frac{r^2_{\hat{g}g}[\frac{(\sigma^2_e)}{n(\sigma^2_g)} + \frac{(\sigma^2_{ge})}{(\sigma^2_g)}]}{(1 - r^2_{\hat{g}g})} = \frac{r^2_{\hat{g}g}[\frac{(\lambda_1)}{n} + (\lambda_2)]}{(1 - r^2_{\hat{g}g})} \quad \textit{[Equation 16]},$$

where $c^2_{ge} = \frac{h\sigma^2_g(1 - r_{ge})}{r_{ge}}$ *[Equation 10]*. Thus, l becomes:

$$l = \frac{r^2_{\hat{g}g}[\frac{(\sigma^2_e)}{n(\sigma^2_g)} + \frac{(\sigma^2_{ge})}{(\sigma^2_g)}]}{(1 - r^2_{\hat{g}g})} = \frac{r^2_{\hat{g}g}[\frac{\left(1 - h^2_g - \frac{h^2_g(1 - r_{ge})}{r_{ge}}\right)}{nh^2_g} + \frac{(1 - r_{ge})}{r_{ge}}]}{(1 - r^2_{\hat{g}g})} \quad \textit{[Equation 17]}.$$

If $n = \infty$; $l = \frac{r^2_{\hat{g}g}}{(1 - r^2_{\hat{g}g})} \frac{(1 - r_{ge})}{r_{ge}}$ *[Equation 18]*, which derives from $r^2_{\hat{g}g*} = \frac{h^2_g}{h^2_g + \frac{\sigma^2_{ge}}{l}} = \frac{lr_{ge})}{1+(l-1)\,r_{ge}}$ *[Equation 14]*.

The number of l depends on reliability (accuracy squared), h^2_g, r_{ge}, and n.

Relationships between accuracy and *F* test in a single-environment trial

Table A1. Analysis of variance for *F* test in a single-environment trial

Source of variation	E(MS)†	E(MS)‡	*F*
Treatment	$\sigma^2_e + n\sigma^2_g$	$[(1 - h^2) + nh^2]\sigma^2_y$	$1 + n\frac{h^2}{1 - h^2}$
Error	σ^2_e	$(1 - h^2)\,\sigma^2_y$	-

†: expected mean square in terms of variance components; ‡: expected mean square in terms of intraclass correlation or coefficient of determination; σ^2_e: residual variance; σ^2_g: genotypic variance; σ^2_y: phenotypic variance; and n: number of repetitions.

$h^2_g = \frac{F - 1}{(F - 1) + n}$ *[Equation 19]*.

Expression for reliability (accuracy squared):

$r^2_{\hat{g}g} = \frac{nh^2}{1 + (n - 1)h^2} = 1 - 1/F$ *[Equation 20]*.

Number of replications (n) for single-environment trials via *F* test:

$n = = \frac{r^2_{\hat{g}g}}{h^2} \frac{(1 - h^2)}{(1 - r^2_{\hat{g}g})} = \frac{r^2_{\hat{g}g}}{(1 - r^2_{\hat{g}g})} \frac{(1 - h^2)}{h^2} = \frac{(1 - 1/F)}{(1/F)} \frac{(1 - h^2)}{h^2} = \frac{(1 - 1/F)}{(1/F)}\lambda$ *[Equation 21]*.

The number of repetitions (n) depends on the F test and heritability.

Relationships between accuracy and *F* test in multi-environment trials

Equivalence between some GxE parameters and F test (Resende 1998):

$$\frac{\sigma^2_{ge}}{\sigma^2_g} = \frac{1 - \frac{1}{F^*}}{F - 1} \quad l = \frac{l}{F^*}\frac{F^* - 1}{F - 1} \quad \textit{[Equation 22]}. \quad r_{ge} = \frac{\sigma^2_g}{(\sigma^2_g + \sigma^2_{ge})} = \frac{(F-1)}{(F-1) + l\left(1 - \frac{1}{F^*}\right)} = \frac{(F-1)}{(F-1) + \left(l - \frac{l}{F^*}\right)} \quad \textit{[Equation 23]}.$$

Table A2. Analysis of variance associated with $F^{\dagger}$, F and F^* in multi-environment trials. All effects taken as random

Source of variation	MS	E(MS)	*F* Test
Environment (E)	Q_1	$\sigma^2 + n\sigma^2_{g*l} + ng\sigma^2_l$	$F^{\dagger} = Q_{1/} Q_3$
Genotype (G)	Q_2	$\sigma^2 + n\sigma^2_{g*l} + nl\sigma^2_g$	$F = Q_{2/} Q_3$
GxE interaction	Q_3	$\sigma^2 + n\sigma^2_{g*l}$	$F^* = Q_{3/} Q_4$
Error	Q_4	σ^2	

n, *g* and *l*: number of repetitions, genotypes and environments, respectively.

$$r^2_{\hat{g}g} = \frac{h^2_g}{h^2_g + \frac{c^2_{ge}}{l} + \frac{(1 - h^2_g - c^2_{ge})}{nl}} = 1 - 1/F \quad \textit{[Equation 24]}.$$

$$h^2_g = \frac{1}{1 + \frac{(1 - r_{ge})}{r_{ge}} + \frac{(1/F^*)nl}{F-1}} \quad \textit{[Equation 25]}.$$

The h^2_g depends on the F and F^* test (*Equation 23*, in terms of F) n and l.

Number of repetitions (*b*) in multi-environment trials via *F* test

Reliability (accuracy squared), heritability (h^2_g), and genetic correlation across environments (r_{ge}) estimators, via F test:

$r^2_{\hat{g}g} = 1 - 1/F$ *[Equation 24]*.

$$r_{ge} = \frac{(F-1)}{(F-1) + l\left(1 - \frac{1}{F^*}\right)} \quad \textit{[Equation 23]}.$$

$$h^2_g = \frac{1}{1 + \frac{(1 - r_{ge})}{r_{ge}} + \frac{(1/F^*)nl}{F-1}} \quad \textit{[Equation 25]}.$$

$$h^2_g = \frac{1}{1 + \frac{\left(1 - \frac{(F-1)}{(F-1) + l\left(1 - \frac{1}{F^*}\right)}\right)}{\frac{(F-1)}{(F-1) + l\left(1 - \frac{1}{F^*}\right)}} + \frac{(1/F^*)nl}{F-1}} \quad \textit{[Equation 26]}.$$