NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

02-23

Inference for Pairwise Comparisons of Random Variety Effects.

Alison Smith and Brian Cullis

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia. Phone +61 2 4221 5076, Fax +61 2 4221 4998. Email: karink@uow.edu.au

Inference for pairwise comparisons of random variety effects.

Alison Smith and Brian Cullis[1]

[1]*Mixed Models and Experimental Design Lab, National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia. Email: alismith@uow.edu.au*

October 30, 2023

In comparative experiments it is often of interest to make formal statistical inference for pairwise comparisons of treatments. In linear mixed model (LMM) analyses in which treatment effects are regarded as fixed effects, we can proceed using hypothesis tests. If we let τ_i and τ_j be the fixed effects for treatments i and j respectively we could test the null hypothesis of $H_0 : \tau_i = \tau_j$ against the alternative hypothesis of $H_a : \tau_i \neq \tau_j$. A test statistic and associated $p-$value is calculated. A formal test of significance involves stipulation of a critical value, α, such that we reject H_0 if the $p-$value is less than α. This means that, for this single comparison, the probability of falsely rejecting H_0 is less than α.

However, it is typically of interest to conduct multiple pairwise comparisons. As the number of comparisons increases, so too does the likelihood of falsely rejecting H_0. To address this problem, Benjamini and Hochberg (1995) developed an approach to control the "False Discovery Rate" (FDR) which is the expected proportion of false "discoveries". In the fixed effects setting, a false discovery may be interpreted as rejection of H_0 when H_0 is actually true.

In the context of plant variety trials, the treatments are the varieties and the aim is to select the best subset of varieties. In a LMM analysis, variety effects are therefore regarded as random effects. It is still possible to make inferences about random effects, but the framework of hypothesis testing for equality is no longer appropriate. Instead, we can make probability statements concerning the ranking or superiority of one variety over another. For example, we can determine the probability that, given the data, the true (genetic) effect for variety i is higher than that for variety j. This is clearly of interest if the aim is to make comparisons of individual varieties against a control, where interest lies in determining whether varieties are superior to the control so in the fixed effects setting we would be conducting a one-sided test. If the aim is to make "generic" pairwise comparisons of varieties, akin to two-sided tests in the fixed effects setting, it is sensible in the random effects setting to consider each comparison as a selection problem. Thus for the pair comprising varieties i and j we wish to "select" the superior variety, that is, establish the ranking (first=top and second=bottom) of the varieties.

Selection is based on the data, or more specifically, on the best linear unbiased predictions (BLUPs) of the variety effects from the LMM analysis of the data. If we denote the BLUPs for varieties i and j as $\tilde{u}_i$ and $\tilde{u}_j$, then the selection process for this pair of varieties proceeds by ordering them on the basis of their BLUPs and selecting the top ranked variety. We then let $\tilde{u}_{(1)}$ be the BLUP for the top ranked variety and let $\tilde{u}_{(2)}$ be the BLUP for the other. Thus $\tilde{u}_{(1)} = \max(\tilde{u}_i, \tilde{u}_j)$ and $\tilde{u}_{(1)} - \tilde{u}_{(2)} > 0$. We then denote the true genetic effect for the selected variety as $u_{(1)}$ and the true effect for the unselected variety as $u_{(2)}$. The probability of a selection error for this pair of varieties is then given by $\Pr\left(u_{(1)} < u_{(2)}|\boldsymbol{y_2}\right)$ where $\boldsymbol{y_2}$ is the transformed vector of data used for residual maximum likelihood (REML) estimation. If we let b be the number of varieties in the experiment and $\boldsymbol{\Omega}$ be the $b \times b$ prediction error variance matrix for the variety effects then we can write

$$\begin{aligned}
\Pr\left(u_{(1)} < u_{(2)}|\boldsymbol{y_2}\right) &= \Pr\left(u_{(1)} - u_{(2)} < 0|\boldsymbol{y_2}\right) \\
&= \Pr\left((u_{(1)} - \tilde{u}_{(1)}) - (u_{(2)} - \tilde{u}_{(2)}) < -(\tilde{u}_{(1)} - \tilde{u}_{(2)})|\boldsymbol{y_2}\right) \\
&= \Pr\left(\frac{(u_{(1)} - \tilde{u}_{(1)}) - (u_{(2)} - \tilde{u}_{(2)})}{\sqrt{\boldsymbol{d}^\top \boldsymbol{\Omega} \boldsymbol{d}}} < -\frac{(\tilde{u}_{(1)} - \tilde{u}_{(2)})}{\sqrt{\boldsymbol{d}^\top \boldsymbol{\Omega} \boldsymbol{d}}}|\boldsymbol{y_2}\right) \\
&= 1 - \Phi\left(\frac{(\tilde{u}_{(1)} - \tilde{u}_{(2)})}{\sqrt{\boldsymbol{d}^\top \boldsymbol{\Omega} \boldsymbol{d}}}\right)
\end{aligned}$$

where $\boldsymbol{d}$ is a vector of length b containing all zeros apart from a "1" in position i and "-1" in position j and Φ is the standard normal cumulative distribution function.

We note that Bueno Filho and Gilmour (2007) use this framework in the context of model-based experimental design for variety trials. They formulate an optimality criterion that seeks to minimise the probability of selection errors.

In a similar manner to hypothesis testing we may wish to stipulate a critical value, α, such that we proceed with selecting the top ranked variety in the pair if the probability of a selection error (the $p-$value) is less than α. The issue of multiple comparisons can be addressed using the Benjamini and Hochberg (1995) approach for controlling the FDR. In this context a "false discovery" is synonymous with a "selection error", that is, an incorrect ranking of the two varieties. This proceeds as follows. We compute the p-values (probability of selection errors) as above for each of t pairwise comparisons. We denote these as $P_1, P_2, \ldots, P_t$. We let $P_{(1)} < P_{(2)} < \ldots < P_{(t)}$ be the ordered $p-$values with the smallest value (lowest probability of a selection error) appearing first. We let $S_{(s)}$ be the selection decision (pairwise comparison) associated with $P_{(s)}$ (for $s = 1, \ldots, t$). Then to control the proportion of false discoveries to a level q^*, we let k be the largest value of s for which $P_{(s)} < \frac{s}{t}q^*$ and only proceed with selections $S_{(s)}$ for $s = 1, \ldots, k$. Thus the new "critical value" is effectively $\frac{k}{t}q^*$.

We illustrate this approach using data from a multi-environment trial (MET). Variety predictions for individual trials were obtained from a factor analytic linear mixed model and there was particular interest in the predictions for 41 key varieties. Here we consider a single trial in which 36 of these varieties were grown so we wish to conduct pairwise comparisons amongst this subset, making a total of $t = 630$ comparisons. We set the FDR level to $q^* = 0.05$. Figure 1 shows a plot of the ordered $p-$values ($P_{(s)}$) graphed against the (ordered) comparison number (s). Also shown is the line with slope given by q^*/t. This

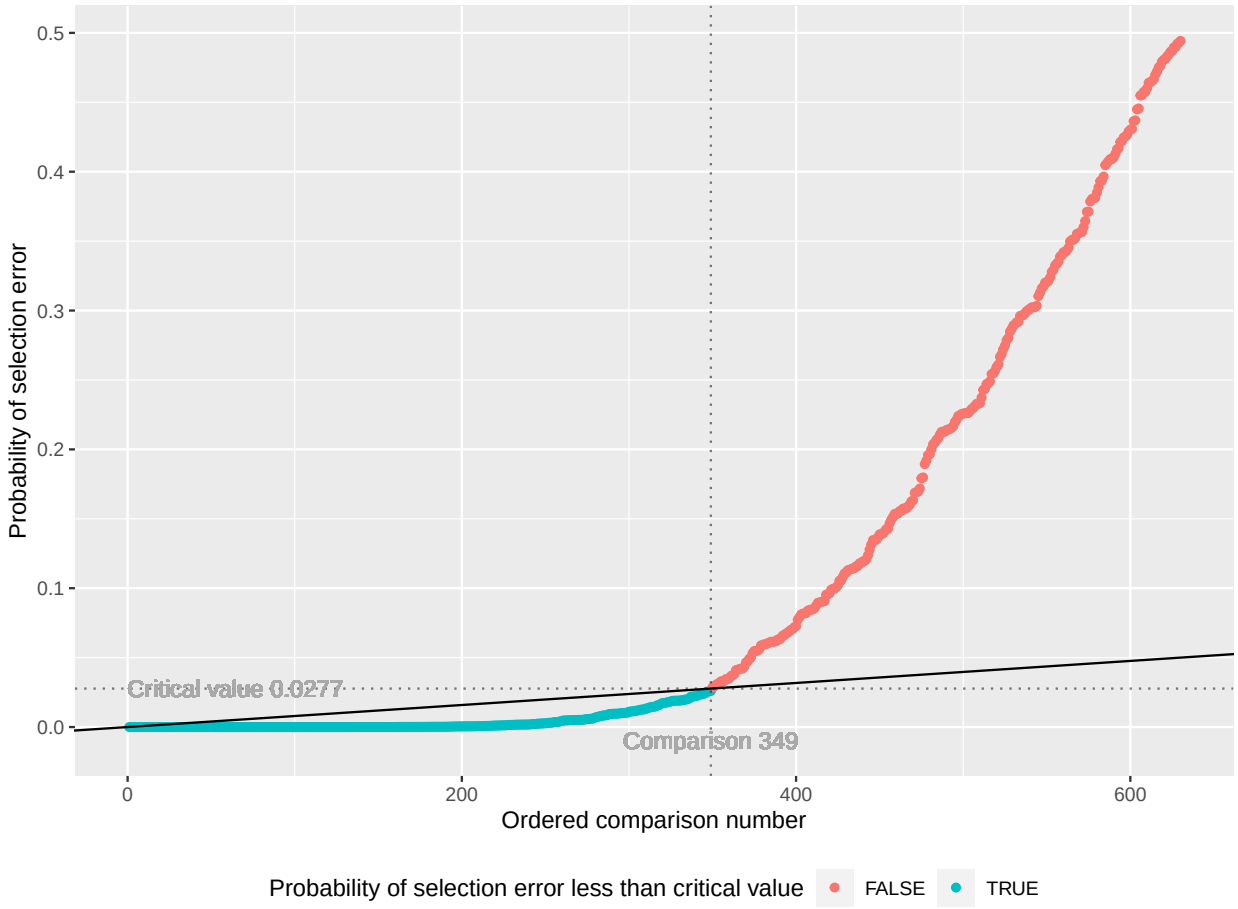

Figure 1: Probabilities of selection errors for 630 pairwise comparisons plotted against ordered comparison number, s (ordering is from smallest to largest probability). The plot has been supplemented with a straight line with slope $q^*/t = 0.05/630$ so that the intersection point provides the critical value that controls the FDR at $q^* = 0.05$.

line cuts the $p-$values at comparison number $k = 349$ where $kq^*/t = 0.02769841$ so that every comparison with a $p-$value smaller than kq^*/t is deemed to be a sufficiently small probability of a selection error.

We may display the results of all pairwise comparisons using a heatmap where cells are coloured differently according to whether the $p-$value for the comparison is less than or greater than kq^*/t. For ease of interpretation the heatmap can be ordered (top to bottom and left to right) in decreasing order of the BLUPs of the variety effects. The heatmap for the example is shown in Figure 2. The cells coloured green in the heatmap relate to those where the $p-$value of the comparison was less than kq^*/t. So these comparisons are the ones with small probabilities of selection errors. It may be easier to interpret these in terms of the converse probability, namely the probability of *not* making a selection error. Given the use of $q^* = 0.05$ we may then think of the green cells as those where we are at least 95% certain that the ranking for the pair of varieties is correct and therefore we are at least 95% certain we would be selecting the best variety. On the other hand, the red cells are those

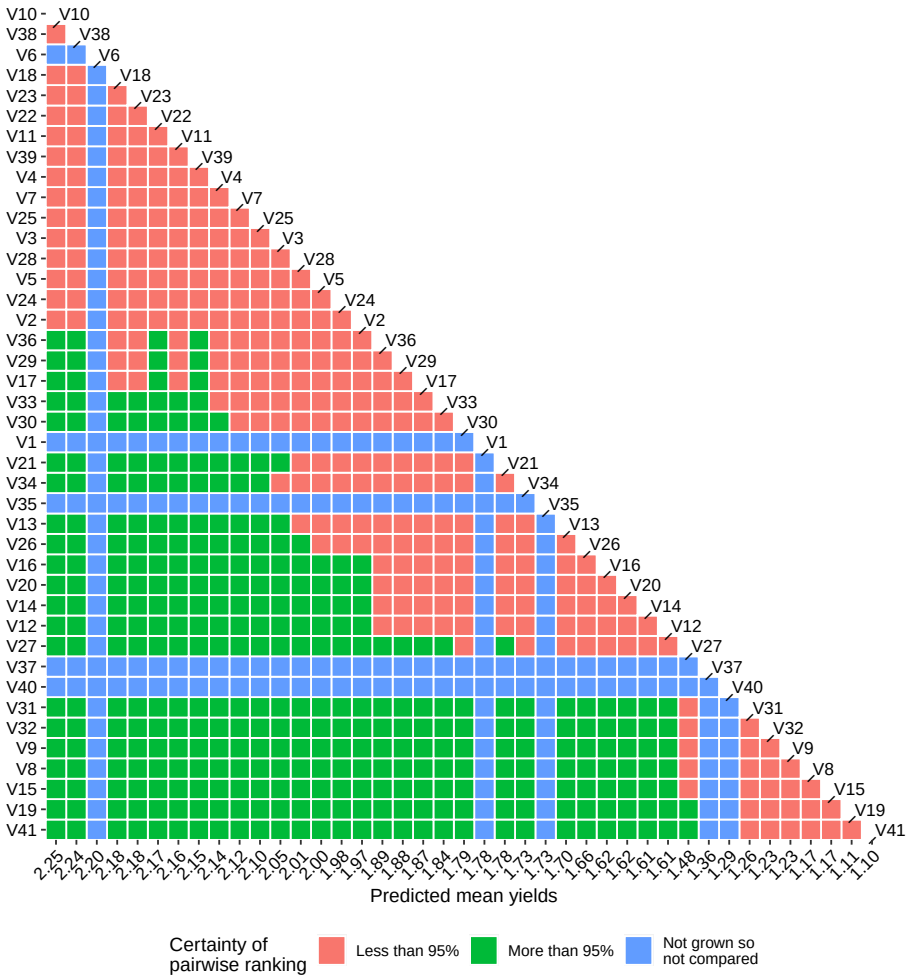

Figure 2: Heatmap of pairwise variety comparisons for a single trial from a MET data-set. 41 varieties were of interest across the entire data-set but 5 were not grown in this trial so the probability statements are restricted to the 36 that were grown.

where we are less than 95% certain that the ranking is correct. Given the ordering of the heatmap, we can examine individual varieties by first scanning left to right. The green cells will then relate to varieties where we are at least 95% certain that they are superior to the variety of interest. The red cells relate to varieties where we are less than 95% certain that they are superior to the variety of interest. Once the diagonal of the heatmap for this variety has been reached we then scan top to bottom. Now the red cells relate to varieties where we are less than 95% certain that they are inferior to the variety of interest. And the green cells relate to varieties where we are at least 95% certain that they are inferior to the variety of interest. The cells coloured blue relate to the additional 5 varieties that were of interest across the entire MET data-set but which were not grown in this particular trial.

References

Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B.*, 57:289–300, 1995.

J. S. de S. Bueno Filho and S. G. Gilmour. Block designs for random treatment effects. *Journal of Statistical Planning and Inference*, 137:1446–1451, 2007.