

Submission and review of manuscripts submitted to the variety testing and evaluation section of HortTechnology: Where do we go from here?

In the instruction for authors included in every January issue of HortTechnology, the category 'Variety trials' is defined as the repository of 'articles reporting the results of studies in which varieties or species are evaluated for comparative performance. Manuscripts should be oriented toward testing and differentiating varieties using traits of interest to growers and other professional horticulturists.' Manuscripts submitted to the 'variety testing and evaluation' section of HortTechnology are peer-reviewed in order to 'assure readers that the published papers have been found acceptable by competent, independent professionals'. As a refereed publication, variety trial manuscripts are expected to follow the criteria of excellence set forth by the ASHS publications committee (see instruction for authors). The strength of a refereed publication comes from (1) the trust in the quality of the work described and in the relevance of the methodology used, and (2) the scope of the inference that can be made from the experimental results. When applied to variety testing, these concepts become (1) issues related to data collection, and (2) statistical methods available to compare varieties.

Traditionally, results of variety trials have been published in non-refereed publications such as bulletins, station reports, regional reports, news letters, or other clientele-oriented publications. The articles that have been published in the 'variety testing and evaluation' section of HortTechnology as of 2001 well illustrate the similitudes and differences between applied research and variety trials. The review process itself revealed challenges in reaching a

consensus between authors and reviewers on what constitutes an acceptable manuscript. Therefore, authors and reviewers needed to be able to refer to guidelines on what may be 'acceptable' in submitting variety trial results. The objective of this article is to list and discuss some essential and desirable characteristics of manuscripts submitted to the 'variety testing and evaluation' section of HortTechnology. This list (Table 1) is intended to be a start, and many may not find it exhaustive. Therefore, this article is intended to serve a flexible guide for authors and reviewers, not a rigid one. Hopefully, those who feel the need to modify, expand or update this list will be stimulated to develop the missing research data, or provide an in-depth summary of existing literature. For clarity of presentation and because issues discussed here have left potential authors and reviewers wondering, issues are presented as answers to typical questions.

1. Questions related to methodology and data collection:

What are the objectives of cultivar testing?

The strength of a research manuscript comes in part from the coherence between the title, the abstract, the statement of objective, and the conclusion. The message conveyed by all these sections should be similar. In particular, a manuscript should have a clear goal. Possible goals for cultivar trials are 'identify poor-performing cultivars', 'select best-performing cultivars', or 'update recommendations by comparing the performance of new cultivars and/or advanced breeding lines to those of the industry standards'(Table 2). These goals are clear because they will allow a specific message in the conclusion. The work will therefore make a significant contribution to the

literature. Also, these clear goals contain the blueprints of a statistical model since they suggest a comparisons of means. On the other hand, goals such as 'evaluate varieties', 'grow new genotypes', or 'try new plant materials' are not action goals; they do not suggest a comparison of means. From past experience, manuscripts without a clearly stated goal seldom generate specific information. The reader is left wondering 'and so what?'

Should we include quality rating for the trials?

Several variety trial reports (Simonne, 1999b; 2000) have standardized the description of growing conditions of cultivar trials, and proposed ratings of weather conditions, fertilization, irrigation, pest pressure, and overall (Table 3). This approach was adopted by several authors contributing the South-eastern Regional Bulletins, who found these ratings practical and simple. In addition, these ratings may be used as quality control when, for example, trials with at least one rating below 3 are not reported.

When preparing a manuscript for HortTechnology, this approach should be considered as an addition to the materials and methods section, and not a substitute. It is still essential to reference precisely the fertilization and irrigation programs used during the trial, as they may influence the result or help explain year-to-year or across-location differences. Typically, cultivar trials should be conducted following current recommendations or industry practices. However, in situations where cultural practices used do not follow current recommendations, the practices used should be described and their selection justified.

To what geographical area do trial results apply?

Traditionally, recommendations are made for each state. With decreasing resources allocated to variety testing, multi-state or regional variety trial programs have emerged such as the one in the Southeast (Simonne, 2000) or the one in the mid-west (Morales and Maynard, 2000). Regional vegetable production recommendations (including cultivars) have been developed (Sanders, 2000), but they are still presented by state. Similarly, most articles currently published in the variety testing and evaluation section of HortTechnology include a state name in the title. By encouraging authors to identify the geographical zone similar to that of the evaluation, progress could be made toward regional recommendations based on growing conditions (soil, climate, production system) rather than state line. Instead of describing an area, authors could include a map with an area of 'potential application of the results presented in this work based on soil type, weather conditions, or planting seasons'. In any cases, inference for cultivar performance cannot be legitimately extended beyond the region in which the experiment was conducted.

How many entries should there be in a trial?

There is no simple answer to this fundamental question. Cultivar trials should contain a reference variety (current industry standard(s) or well-known variety) together with new varieties and advanced breeding lines. When no single standard exist, this should be explained. Practically, the two main factors that determine the number of entries are crop type, and resources needed to perform the trial. For some crops such as watermelon, sweet corn, or tomato, large number of cultivars are introduced each year.

For these crops, cultivar trials may include up to thirty varieties. In contrast, few new releases occur each year for crops such as strawberries. For these crops, it is possible to have a valid cultivar trial with only two entries - the industry standard and the new introduction.

The other factor determining the number of entries in a trial is the resources - mainly in terms of space, labor, and cost- needed to perform the trial. For example, hand-harvest takes an estimated 150 men-hour/acre for cucumber or eggplant (multiple harvests), but it takes 30 and 25 and for sweet corn, and watermelon (once-over harvest). For comparison, it only takes 20 men-hour/acre to harvest potato mechanically (Brown et al., 1983). Increasing the number of entries without increasing the number of replications has statistical consequences which will be discussed in section 2.

What are the typical plot sizes used in cultivar evaluation?

Guidelines have been proposed (Maynard, 1987) regarding plot size and number of plants per plot for the main vegetable crops grown in the United States (Table 4). While these guidelines have been largely adopted in cultivar evaluation, the coefficients of variation observed for total marketable yield were often above the 20

In general, CV are even higher for weights within each grade because grade weights are fractions of total marketable yields. Yet, when market prices are much higher at the beginning of the season, growers may make higher profits with a small percentage of their total production. These results suggest two points. First, because of the inherent variability in plant yields, plot sizes commonly used may not allow to keep CVs in vegetable trials down

to the accepted levels in cultivar trials of other commodities. The statistical implication of high CV in multiple comparison tests will be discussed later. Second, as commodities are grown for profit, early yields could be better compared using market values.

Can yield data be published alone?

The most common goal of cultivar testing is to make a recommendation. Hence, any attribute useful to distinguish or compare cultivars should be measured. Typical data collection include yield, grade distribution, and horticultural attributes (Maynard, 2001). Some data commonly collected are crop-specific (Table 4). Yet, photosynthetic response (Bhagsari, 1990), plant nutritional characteristics (Quintana et al., 1996; Southwick et al., 1999), vitamin content (Wang and Goldman, 1996; Simonne et al., 1997), chemical composition (Kalt and McDonald, 1996), cooking tests (Paull et al., 2000), consumer acceptance (Frank et al., 2001), taste tests (Brittain and McDonald, 1987; Simonne et al., 1999), disease reaction (Schultheis and Waters, 1998; Southwick et al., 1999), or post harvest behavior (Liang and Harbaugh, 2001) also provided useful information in assessing cultivar performance. It is therefore unlikely that yield alone be sufficient to make a recommendation, and therefore should not be the sole data reported in manuscripts submitted to the variety trial and evaluation section of HortTechnology.

Should data be corrected for stand? When?

When assessing yield, it makes sense not to correct for stand since stand itself provides information on yield. Analysis of the stand rate itself may be of interest, but inflating yield by correction for stand would seem to be an

inaccurate measurement.

What units should be used to report data?

If data do not approximately conform to Gaussian distributional assumptions underlying ANOVA then transformations such as logarithmic, exponential or power transformations may cure the problem. When transformations of data fail, ranks of responses can be used in nonparametric analyses. Nonparametric methods require weaker assumptions than parametric ANOVA and are generally less efficient than ANOVA for normally distributed data, so that ranks should only be used when necessary.

Occasionally, analyses for incomplete and unbalanced designs can be simplified by using differences from or ratios to some control or standard measurement. Drawbacks to these techniques of measurement include a general loss of information and loss of degrees of freedom in ANOVA. When the design is complete and independent measurements can be made for each response, then there is no need for differences or ratios. In other cases, it may only be possible for responses for treatments to be made relative to the control.

How can global indices be used to establish overall comparisons?

Data collected on cultivar trials are usually analyzed using univariate statistical procedures (analysis of variance, means comparison tests, non parametric tests). Yet, the recommendation of a cultivar is based on a global judgement that includes several attributes. For example, yield, ear characteristics, and eating quality all contribute to the quality of sweet corn (Simonne et al., 1999). Stem length, diameter and vase life described the

performance of Trachelium cultivars and were used together to identify best overall cultivar (Liang and Harbaugh, 2001). Overall evaluation of lemon cultivars included fruit yield, juice yield and chemical composition, and tree survival rate (Fallahi et al., 1990). In many published articles, the overall evaluation is part of the discussion. It is subjective, based on the individual measurements (Fallahi et al., 1990). In some articles where ranking procedures have been used, overall performance was based on a rank sum index (Simonne et al., 1999; Liang and Harbaugh, 2001). Global indices base on rank sums can be used to cumulate the partial contribution of each attribute to the overall evaluation of that cultivar, thereby reducing the subjectivity of the evaluation.

Global indices are easy to define. Yet, some basic rules must be followed so that they are used correctly. First, each variety has to be ranked for each attribute. Usually, 1 is assigned to the variety with the highest mean, and N is assigned to the one with the lowest (when the trial had N entries). It is essential that all rankings are oriented the same way in regard to desirability. Some variables such as yield represent desirable attributes. In this case, the higher the value, the better. Other variables describe adverse or undesirable attributes. This may be levels of bitterness in lettuce, or cull weights. In this case, the higher the value, the least desirable the attribute.

When results from trials from different years/location are used to define a global index, it is not uncommon for the number of ranks to be different (because of a different number of entries at each year/location). In this case, attention has to be paid to the way the ranks are assigned so that no bias is

introduced.

When two or more means are numerically the same, then a tie in rank occurs. The proper way to handle two-way ties at rank p , is to assign twice the rank $p + 1/2$. The following rank is then $p + 2$. This procedure allows the sum of the ranks to be constant, despite the presence of ties. The three sums (1) $p + (p+1) + (p + 2)$, (2) $(p+1/2) + (p+1/2) + (p+2)$, and (3) $(p+1) + (p+1) + (p+1)$ are identical ($3p+3$).

Inspection of results using these indices can be informative, but statistical inference for multiple comparisons among them is decidedly incomplete. Because of their dependence on one another, ranks or rank-sums and hence the RSI may behave differently in repeated sampling than do raw measurements. The statistical literature is particularly undeveloped for multiple comparisons using ranks. For discussion, see Conover and Iman (1981) and Hsu (1996).

2. Issues related to statistics and experimental design

How much detail is required to describe the statistical methodology used?

Guidelines for improved presentation of ANOVA and regression results are provided in clear and persuasive article by Wehner and Shaw (1994) and should be adopted for the scholarly publication of variety trials as well. In particular, these authors applaud the inclusion of ANOVA tables. This should be standard practice. Without a complete ANOVA table, it is difficult to understand all of the sources of variability (block, treatment, time, etc.) in an experiment and to what degree they explain variation in the response variables. In particular, the *mean square for error* (MSE) term should always be reported, as it estimates error for replications within a block x cultivar

combination and is central to all subsequent inference, including tests and confidence intervals, and can provide a reference point for precision in comparison of multiple studies. The experimental design should also be clearly stated in the methods section.

Standard errors for cultivar means, perhaps in parentheses next to the reported means would be extremely informative. Even better would be simultaneous 95% confidence intervals for all pairwise differences. If there are k cultivar means, then a $k \times k$ table of intervals can provide much more information than the traditional table of means accompanied by letters indicating whether or not two means differ significantly. Indeed, knowing whether or not mean differences are statistically significant alone does not provide growers with all the information they need when selecting cultivars. A table of intervals all of which cover the true differences with high probability would seem to be much more informative.

Some mention of whether or not the data conform, at least approximately to the mathematical assumptions underlying ANOVA techniques ought to be included to validate the statistical methodology. Most statistics texts advocate inspection of residual plots or goodness-of-fit statistics. Inclusion of these plots may not be appropriate for articles here, but some mention that they were inspected would be reassuring to readers. It is straightforward to generate residuals from any model and use them in diagnostic plots. For example, the output statement in the SAS code below creates a temporary SAS dataset containing the original data along with the fitted values \hat{y} for a response variable y and the residuals, $e = y - \hat{y}$. These residuals can then

be plotted against the fitted values to check for homogeneity of variance. Normal plots for the residuals or goodness-of-fit statistics can then be used to assess the assumption of normality.

To the greatest extent possible, data and analyses should be clarified for ease of understanding and for possible use in future work. There may be considerable overlap among multiple experiments or publications and all-important statistical power for finding variety differences could be gained by pooling results. Ensuring high standards in presentation and publication will do much towards this end. Even better would be inclusion of the steps followed when software is used. An example of inclusion of SAS code would be the following:

```
proc glm;
  class cultivar block;
  model yield=cultivar block;
  lsmeans cultivar/cl pdiff=all adjust=tukey;
  output out=resdata r=residuals p=fitted;
run;

proc plot;
  plot residuals*fitted;
run;

proc univariate normal plot;
  var residuals;
run;
```

Another possibility is to include a URL pointing readers to locations on the internet where data and SAS or other software code can be found.

Are single year/location trials as well as non-replicated data publishable?

Inference is the legitimate claim that the results observed on a sample

also apply to the population from which that sample was drawn. In order to control its level of statistical risk, inference is based on an assessment of experimental repeatability in time and space. Multiple locations allow the estimation of the cultivar x environment (location) interaction. In most trials, this interaction is significant, thereby indicating that the performance of cultivars differ from location to location (Poysa et al., 1986; Hodges et al., 1995). When non-replicated data are collected, a broad inference cannot be made as no estimate of variance is available. If block effects (such as fertility or irrigation gradients in soil) are strong and can conceivably affect different varieties differently, then replication is needed to estimate these interactions. If it is reasonable to believe that all cultivars benefit to the same degree from block effects, then non-replicated data are acceptable. If interactions exist, then the non-replicated RBD model is underspecified, and estimates for treatment effects will be biased.

It should be noted that replicated data may be collected from single-plot trials only when intensive variables are measured. Intensive variables such as growth habit, disease resistance, fruits type and shape, may be collected more than once on a single plot. Except in this situation, it is highly unlikely that non-replicated data will be acceptable in manuscripts submitted to the variety testing and evaluation section of HortTechnology. In some limited cases, single-location trials may be acceptable, especially when the environmental conditions are relatively controlled such as in greenhouse studies.

Reporting raw data, ranks or percentage of check within rep (like herbicides): what are the pros and cons?

If data do not approximately conform to Gaussian distributional assumptions underlying ANOVA then transformations such as logarithmic, exponential or power transformations may cure the problem. When transforming data fails, ranks of responses can be used in nonparametric analyses for comparisons of cultivar medians. Nonparametric methods are generally less efficient than ANOVA for normally distributed data, so that ranks should only be used when necessary. Drawbacks to subtracting off means, or using ratios of measurements to some standard include a loss of degrees of freedom, but in some cases, the difference from some reference measurement point really is an appropriate response to analyze.

What are the most appropriate mean separation techniques for mean comparisons in cultivar trials?

A survey of the voluminous literature on the topic indicates that the multiple comparisons issue is a controversial one. The collection of papers defies enumeration. Fortunately, many of these are insightful, informative and some humorous. See e.g. Chew (1973), Little (1978), Swallow (1984), Saville (1990), Gates (1990), Mihail and Niblack (1991) and Tukey (1991). There have been a number of simulation studies of *multiple comparison procedures* (MCPs) or *mean separation procedures*. See e.g. Carmer and Swanson (1973) or Einot and Israel (1975). Finally, most textbooks on statistical methods address the issue and Hochberg and Tamhane (1987) and Hsu (1996) are devoted entirely to it. Here the problem will be reviewed and insight into pertinence in variety trials will be discussed.

As mentioned at the outset, the quality of a publication depends upon

the scope and strength of the inference that can be made from the analysis of the experiment. Hsu (1996) classifies the strength of inference of any MCP as falling in one of several categories (in order of increasing strength): 1) Individual: without any adjustment for multiplicity 2) Inhomogeneity: means are different 3) Tests of Inequalities: which sample means differ significantly 4) Simultaneous confidence intervals for differences of means. A type I error occurs when a difference between sample means from two equivalent cultivars is found to be statistically significant. A type II error occurs when two sample means for two non-equivalent cultivars is not big enough to be declared statistically significant. MCPs have been developed for experiments that attempt to answer many questions at once. Without adjustment for multiplicity, the expected number of type I errors increases quadratically with the number of entries in a variety trial, as shown in Table 7. For some this is unacceptable, for others, it is an affordable price to pay.

Both types of errors are unavoidable, but MCPs have been developed to control the type I error rate while accounting for multiplicity. When making many comparisons among cultivars, the *experimentwise error rate* is defined as the average proportion of experiments in which at least one type I error is committed. The *comparisonwise error rate* is the average proportion of comparisons in which a type I error is committed. Some MCPs, such as Tukey's procedure, sometimes called the honestly significant difference, or the Tukey-Welsch procedure (denoted REGWQ in SAS) are constructed to control for the experimentwise error rate, others, such as the least significant difference (LSD), the protected LSD, or Duncan's Multiple Range test are

not.

Persuasive arguments are made in Saville (1990) that only individual comparisons are needed in experiments such as variety trials, that stronger forms of inference are too complex for interpretation, can lead to inconsistencies in declaring differences significant and suffer from a high type II error rate. So, Hsu's weakest form of inference can be achieved simply by comparing any observed difference to a least significant difference or LSD based on the t -distribution of any estimated, standardized difference of sample means. If it is acceptable to accept a comparisonwise error rate of $\alpha = 0.05$, without undue concern for experimentwise error rates, then the LSD procedure is a reasonable recommendation. A scan of current articles indicates that this is a popular technique and easy to explain and report in tables of means. Confidence intervals can also be constructed without adjustment for multiplicity. Their interpretation is that 95% of them will "cover" the true cultivar differences. If this is an acceptable level of confidence, then no adjustment for multiplicity is warranted, so long as the limitation of the strength of the inference is mentioned in the analysis.

Note the following table, which gives the number of pairwise comparisons needed in a variety trial with between 5 and 40 entries and the average number of confidence intervals which will "miss" the true differences:

Entries	Comparisons	Avg. # missed
5	10	0.5
10	45	2.3
20	190	9.5
30	435	21.8
40	780	39

One well-known study (Carmer and Swanson, 1973) reported the following Monte Carlo estimates of experimentwise error rates for randomized block designs with $k = 5, 10$ or 20 equal treatments and varying numbers of replications $n = 3, 4, 6, 8$. The Monte Carlo standard errors for the estimated experimentwise error rates appear in parentheses and are determined from the fact that 4000 simulations were used for each treatment configuration.

Procedure	k		
	5	10	20
Tukey	0.05	0.048	0.047 (0.003)
Duncan	0.182	0.373	0.626 (0.008)
LSD	0.256	0.584	0.895 (0.008)

This indicates that in trials with $k = 20$ equal treatments if Duncan’s procedure is used with comparisonwise error rate $\alpha = 0.05$, there will be false discoveries of differences among varieties in about 61 – 64% of these types of experiments. The example is slightly pathological, as few would attach much credence to an omnibus equality of all cultivars in many variety trials in the first place, but the same thing happens for many configurations in which there are some equalities among cultivar means.

If stronger inference is desired, then a simple and highly informative analysis of variety trials which address the multiplicity issue is to obtain simultaneous confidence intervals for all pairwise differences with the property that the chance that they all “cover” the true mean differences is 95%. This method is easy to implement using SAS or many other packages. The sample code provided earlier will report the intervals for all pairwise comparisons. Space considerations would make it difficult to report all of these in an article, particularly when there are many attributes in addition to yield under

consideration. Only those intervals which are of interest to growers or researchers need to be reported. Consider an example taken from a variety trial in Michigan to evaluate 6 varieties of early season strawberry. Assume the actual means (in $100\text{lbs}/\text{acre}$) which would be achieved with infinite sample size are about

$$\mu_1 = 40, \mu_2 = 40, \mu_3 = 50, \mu_4 = 90, \mu_5 = 90, \mu_6 = 110.$$

Assume further that a complete randomized block design is used, with $n = 3$ replications. The block effects were assumed to be modest: with 2 blocks yielding on average 2000 more lbs/acre and 2 blocks yielding 2000 less and 2 blocks yielding the average. Data simulated under this model assuming that the standard deviation for replications of each variety is about $\sigma = 20$ appear in the table below:

	block					
cultivar	1	2	3	4	5	6
1	46	32	38	54	96	40
2	6	12	54	26	46	16
3	40	14	50	48	70	78
4	110	64	114	104	102	102
5	98	82	96	98	72	90
6	108	54	80	104	138	132

The SAS code used to generate these data appears below:

```

data one;
  array alpha{6} (-30,-30,-20,20,20,40);
  array beta{6} (-20,-20,0,0,20,20);
  do block=1 to 6;
    do cultivar=1 to 6;
      y=70+alpha{i}+beta{j}+20*normal(2);
      y=round(y,2);
      output;
    end;
  end;
run;

```

The ANOVA table and all differences are reported below. The output was produced by SAS and cut and pasted into this document. The only differences in varieties illuminated by the analysis are those involving pairs (1, 4), (1, 5), (1, 6), (2, 4), (2, 5), (2, 6), (3, 4), (3, 5), (3, 6) though the (1,3) and (1,5) differences contain plausible differences as small as 700 or 800 *lbs/acre*. Since the true cultivar means are known, the error rate from this particular simulated experiment is known. No type I errors occurred. The differences involving pairs (1, 3), (2, 3), (4, 6), (5, 6) were all missed, so that 4 type II errors resulted. Note that the magnitude of these 4 errors is very small relative to other differences among the varieties, so that there may be smaller cost associated with missing them. Listing the simultaneous confidence intervals provides more information about the precision with which the effects can be estimated. An interval estimate for the difference between 2 and 6 is (4400, 10800) in *lbs/acre*. Reporting the interesting confidence intervals may be more informative than tables with the requisite "means with the same letter do not differ significantly" message that Little (1978) belittles.

The SAS System
The GLM Procedure

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
cultivar	5	29639.66667	5927.93333	18.86	<.0001
block	5	6497.00000	1299.40000	4.13	0.0071
Error	25	7858.33333	314.33333		
Corrected Total	35	43995.00000			

Least Squares Means
Adjustment for Multiple Comparisons: Tukey

cultivar	y LSMEAN	LSMEAN Number
1	51.000000	1
2	26.666667	2
3	50.000000	3
4	99.333333	4
5	89.333333	5
6	102.666667	6

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	24.333333	-7.212162	55.878828
1	3	1.000000	-30.545495	32.545495
1	4	-48.333333	-79.878828	-16.787838
1	5	-38.333333	-69.878828	-6.787838
1	6	-51.666667	-83.212162	-20.121172
2	3	-23.333333	-54.878828	8.212162
2	4	-72.666667	-104.212162	-41.121172
2	5	-62.666667	-94.212162	-31.121172
2	6	-76.000000	-107.545495	-44.454505
3	4	-49.333333	-80.878828	-17.787838
3	5	-39.333333	-70.878828	-7.787838
3	6	-52.666667	-84.212162	-21.121172
4	5	10.000000	-21.545495	41.545495
4	6	-3.333333	-34.878828	28.212162
5	6	-13.333333	-44.878828	18.212162

These same data could also be analyzed without regard to multiplicity ($6(6 - 1)/2 = 15$ comparisons) using LSD. The output was again produced by SAS and cut and pasted into this document.

The GLM Procedure
Least Squares Means

i	j	Difference	95% Confidence Limits for	
		Between Means	LSMean(i)-LSMean(j)	
1	2	24.333333	3.251687	45.414979
1	3	1.000000	-20.081646	22.081646
1	4	-48.333333	-69.414979	-27.251687
1	5	-38.333333	-59.414979	-17.251687
1	6	-51.666667	-72.748313	-30.585021
2	3	-23.333333	-44.414979	-2.251687
2	4	-72.666667	-93.748313	-51.585021
2	5	-62.666667	-83.748313	-41.585021
2	6	-76.000000	-97.081646	-54.918354
3	4	-49.333333	-70.414979	-28.251687
3	5	-39.333333	-60.414979	-18.251687
3	6	-52.666667	-73.748313	-31.585021
4	5	10.000000	-11.081646	31.081646
4	6	-3.333333	-24.414979	17.748313
5	6	-13.333333	-34.414979	7.748313

NOTE: To ensure overall protection level, only probabilities associated with pre-planned comparisons should be used.

Note the warning. Note also the type I error that occurred when comparing cultivars 1 and 2. The observed comparison wise error rate here is then $1/15 = 6\%$ while the observed experimentwise error rate is 100%.

Hayter and Hsu (1989) show that Tukey's procedure applies to randomized block designs and to balanced designs, but for unbalanced incomplete

block designs, it is currently only conjecture that the Tukey MCP preserves the experimentwise error rate for more than $k = 3$ cultivars. In this last case, a Bonferroni adjustment or simulation approach is recommended. Both are easily provided for in SAS using `adjust=bon` or `adjust=simulate` in the LSMEANS statement. For theoretical details, see Westfall et al (199?).

Controlling experimentwise error rate comes at a price. There is generally a type I for type II error trade-off. This alone is perhaps a strong argument not to adjust for multiplicity. Without attaching some loss or cost to the two types of errors or incorporating production costs associated with actual cultivars which could be used in conjunction with interval estimates to attempt to make decisions based on profitability, there is no immediate reason why controlling for type I error rate is more important than minimizing type II error rate. This is perhaps unusual in that many areas where multiple comparisons might be used, in medical or pharmaceutical applications for example, type I errors are often more grievous than type II errors. The response to the next question in the list attempts to address power considerations. For now, some power can be gained while controlling for experimentwise error rate when it is not necessary to make all pairwise comparisons. Sometimes inference can be restricted to questions concerning a control or identification of a “best” variety and a search for inferior cultivars relative to this unknown best. MCPs that are useful in variety trials can be classified as one of the following types

1. all-pairwise comparisons (MCA)
2. multiple comparisons with a control (MCC).
3. multiple comparisons with the best (MCB).

Dunnett's test for comparisons with a reference mean If there exists a control against which it is of interest to compare varieties in a trial, then Dunnett's procedure can be used and gives substantial gains in statistical power over other all-pairwise-comparison MCPs. Dunnett's stepdown procedure, easily invoked using SAS

Multiple comparisons with a control (MCC) can be achieved while preserving an experimentwise error rate using Dunnett's procedure. SAS has implemented this procedure in the GLM procedure. MCA was discussed previously. Dunnett's procedure can be used for MCC problems. It preserves the experimentwise error, is more powerful than MCA procedures and is easy to implement using SAS. In the strawberry example, suppose that interest lies in comparisons involving cultivar 4 so that it is a standard or reference cultivar or control. The following LSMEANS statement within PROC GLM will get SAS to carry out Dunnett's MCP: `lsmeans cultivar/pdiff=control('4') adjust=dunnett;` The output appears below:

The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Dunnett

cultivar	y LSMEAN	H0:LSMean= Control Pr > t
1	51.000000	0.0004
2	26.666667	<.0001
3	50.000000	0.0003
4	99.333333	
5	89.333333	0.7889
6	102.666667	0.9973

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	4	-48.333333	-75.844227	-20.822439
2	4	-72.666667	-100.177561	-45.155773
3	4	-49.333333	-76.844227	-21.822439
5	4	-10.000000	-37.510894	17.510894
6	4	3.333333	-24.177561	30.844227

Inspection reveals that only one nonzero difference (4,6) was missed, and precision is greater. These intervals are more narrow than the ones necessary for the MCA problem. From this analysis, the standard appears to produce significantly more than cultivars 1,2 and 3 and substantially more than cultivars 1 and 2 while cultivar 6 may produce as high as *5200lbs/acre* more than the standard.

Lastly inference may be only for identification of a best cultivar with respect to a single attribute at a time, such as yield. Hsu (1996) has developed an MCB procedure which has not been enabled yet in PROC GLM, but is

available in SAS/INSIGHT. This procedure is not as commonly used yet, but seems appropriate for variety trials.

Is there a method to determine the adequate number of replications needed? and should we report power?

When designing the experiment, both types of error rates should be considered: false discovery or accidental declaration of equivalent cultivars to be different (type I) and failure to declare sample differences among substantially different cultivars to be significant (type II). Researchers have rightly criticized undue emphasis on type I error rates. Indeed, p-values and hypothesis tests are formulated to control the probability of false discovery, but many researchers point out that false discovery is no more severe an error as failure to discover. Power has been insufficiently addressed in many fields, perhaps because it is inconvenient to calculate. Software which will compute power is much less prevalent and less well-known and few practitioners are familiar with appropriate packages. Power considerations also require the practitioner to be more specific and to elaborate about the questions being investigated. As is often the case when a client asks a question of a statistician, the client gets several questions back in return. How much of a difference between cultivars is meaningful or is it possible to hypothesize a meaningful configuration of cultivar means *a priori*? What procedure will be used to address the problem of multiplicity of comparisons? How high must the probability be to detect a given cultivar difference or what proportion of actual differences declared significant is acceptable?

Sample size computations are straightforward for the F-test in one-way

ANOVA. See, Ch. 9.8 Rao (1998) for example. Such a computation requires only the specification of an effect size and a guess at the experimental standard deviation of a response. This approach provides a start for a variety trial. As an example, suppose that 6 varieties of early season strawberry are to be evaluated in a variety trial and the actual means in *100lbs/acre* which would be achieved with infinite sample size for the 6 varieties are about

$$\mu_1 = 40, \mu_2 = 40, \mu_3 = 50, \mu_4 = 90, \mu_5 = 90, \mu_6 = 110.$$

The effects, or differences from average, of the 6 varieties are then

$$\alpha_1 = -30, \alpha_2 = -30, \alpha_3 = -20, \alpha_4 = 20, \alpha_5 = 20, \alpha_6 = 40.$$

Suppose that the standard deviation for replications of each variety is about $\sigma = 20$. Suppose $n = 3$ replications will be made in a completely randomized design. Following Rao (1998), the F -ratio will have a noncentral F distribution with $\nu_1 = 5$ and $\nu_2 = 12$ respective numerator and denominator degrees of freedom and noncentrality parameter $2\lambda = n \sum \alpha_i^2 / \sigma^2 = 34.5$. The area to the right of the $\alpha = 0.05$ critical value $F(0.95, 5, 12) = 3.1059$ under this noncentral F distribution is 0.9680. The SAS code below computes this power:

```
data one;
    lambda=3*(30**2+30**2+20**2+20**2+20**2+40**2)/400;
    fstar=finv(0.95,5,12);
    power=1-probf(fstar,5,12,lambda);
run;
```

Beyond this simple computation, another power consideration involves the proportion of the real differences that are expected to be detected. In the example above, of the 15 pairwise cultivar differences, 13 are nonzero. If a MCA procedure is used, what proportion of these 13 nonzero differences are expected to be found? One approach to answering this question is by simulation. A Monte Carlo estimate of the proportional power can be constructed from averaging the proportion of 13 differences detected over many simulations. The more simulated datasets, the more accurate the estimate of power. The SAS macro %SimPower developed by Westfall, et al (1999) accomplishes this easily and allows for each of MCA MCC and MCB types of comparisons using Tukey, Dunnett or the Tukey-Welsch procedures. The macro is currently available on the internet at <http://ftp.sas.com/samples/A56648>. Once the macro has been compiled, the following statement is all that is needed:

```
%SimPower(method=tukey,n=4,s=20,truemeans=(40,40,50,90,90,110),
nrep=100,seed=123);
```

The output from this invocation of the macro appears below:

Method=TUKEY, Nominal FWE=0.05, nrep=100, Seed=123	1	
True means = (40,40,50,90,90,110), n=4, s=20		
Quantity	Estimate	---95% CI----
Complete Power	0.00000	(0.000,0.000)
Minimal Power	0.98000	(0.953,1.000)
Proportional Power	0.49308	(0.460,0.526)

So that with this design and this *a priori* specification of cultivar effects and error, about half of the real differences will be detected, while the chance of rejecting the overall hypothesis of no cultivar effects is estimated as 0.98. (The widths of the confidence intervals can be decreased by increasing nrep when the macro is called.) In publishing the findings from variety trials, computations like this would illuminate the possible limitations of the experiment to find all differences.

The shortcoming of this software is lack of functionality beyond one-way models. Most variety trials are block designs and require two factor ANOVA. In the absence of block x cultivar interaction, many designs, including balanced randomized block designs have what is called a one-way structure and theory for MCPs carries over from one-way models to general linear models. So, the MCPs and power simulators apply to some more general models. For models that do not have a simple structure, such as unbalanced incomplete block designs, special provisions should be made to enable computation of power under meaningful alternatives.

Literature cited

Bhagsari, A.S. 1990. Photosynthetic evaluation of sweetpotato germplasm. *J. Amer. Soc. Hort. Sci.* 115(4):634-639.

Brittain, M.J. and N.A. McDonald. 1987. Techniques used in performance trials of celery, leeks, parsnips and sweet corn. *J. Natl. Inst. Agric. Bot.* 17:345-352.

Brown, G.K., D.E. Marshall, B.R. Tennes, D.E. Booster, P. Chen, R.E. Garrett, M.O. O'Brien, H.E. Studer, R.A. Kepner, S.L. Hedden, C.E. Wood, D.H. Lenker, W.F. Miller, G.E. Rehkugler, D.L. Peterson, and L.N. Shaw. 1983. Status of harvest mechanization of horticultural crops. *Amer Soc. Agric. Eng., St. Joseph, Mich.*

Cramer, C.S. 2001. Comparison of open-pollinated and hybrid onion varieties for New Mexico. *HortTechnology* 11(1):119-123.

Crossa, J. 1990. Statistical analysis of multilocation trials. *Adv. Agron.* 44:55-85.

Einot, I. and K.R. Gabriel. 1975. A study of the powers of several methods of multiple comparisons. *JASA*. 70(351):574-583.

Eskridge, K.M and R.F. Mumm. 1992. Choosing plant cultivars based on the probability of outperforming a check. *Theor. Appl. Genet.* 84:494-500.

Fernandez, G.C.J. 1991. Analysis of genotype x environment interaction by stability estimates. *HortScience* 26(8):947-950.

Frank, C.A., R.G. Nelson, E.H. Simonne, B.K. Behe, and A.H. Simonne. 2001. Consumer preference for color, price, and vitamin C content of bell peppers. *HortScience* 36(4):795-800.

Gates, C.E. 1991. A User's Guide to Misanalyzing Planned Experiments. *Hortscience* 26(10):1262-1265.

Hochberg, Y. and A.C. Tamhane. 1987. *Multiple Comparison Procedures*. New York: John Wiley & Sons.

Hodges, L, D.C. Sanders, K.B. Perry, K.M. Eskridge, K.M. Batal, D.M. Granberry, W.J. McLaurin, D. Decoteau, J. Dufault, J.T. Garrett, and R. Nagata. 1995. Adaptability and reliability of four bell pepper cultivars across three southeastern states. *HortScience* 30(6):1205-1210.

Hsu, J.C. 1996. *Multiple Comparisons: Theory and Methods*. London: Chapman and Hall.

Kalt, W. and J.E. McDonald. 1996. Chemical composition of lowbush blueberry cultivars. *J. Amer. Soc. Hort. Sci.* 121(1):142-146.

Kemble, J.M. (Ed.). 2000. Spring 2000 commercial vegetable variety trials. *Regional Bul. 5*, Auburn Univ., AL.

Kemble, J.M. (Ed.). 2001. Fall 2000 commercial vegetable variety trials.

Regional Bul. 6, Auburn Univ., AL.

Kessler, J.R., Jr., J.L. Sibbly, B.K. Behe, D.M. Quinn, and J.S. Bannon. 2000. Herbaceous perennial trials in central Alabama. *HortTechnology* 10(1):222-228.

Liang, R. and B.K. Harbaugh. 2001. Evaluation of *Trachelium* cultivars as cut flowers. *HortTechnology* 11(2):316-318.

Little, T.M. 1978. If Galileo published in *HortScience*. *HortScience* 13(5):504-506.

Maynard, D.N. 1987. Vegetable variety evaluation demonstrations: A manual for county Extension faculty. Fla. Coop. Ext. Ser. Circ. 762, UF/IFAS, Gainesville, FL.

Maynard, D.M.N. 2001. Variety selection, p.15. In: Maynard, D. and S.M. Olson (eds.) *Vegetable production guide for Florida*, UF/IFAS, Gainesville, Fla.

Mihail, J.D. and T.L. Black. 1991. Comparison of Treatment Means: A Statistical Fantasy. *Journal of Nematology* 23(4S):557-563.

Mullins, C.A. and R.A. Straw. 2001 Performance of filet-type snap bean cultivar in Tennessee. *HortTechnology* 11(1):124-127.

Morales, M.R. and L. Maynard. 2000. Midwestern vegetable variety trial report for 2000. Bul. 798, Purdue Univ., West Lafayette, Ind.

Mullins, C.A., R.A. Straw, B. Pitt, Jr., D.O. Onks, M.D. Mulles, J. Reynolds, and M. Kirchner. 1999. Response of selected sweet corn cultivars to nitrogen fertilization. *HortTechnology* 9(1):32-35.

Orzolek, M.D., W.J. Lamont, and L. Otjen. 2000. 1997 spring and fall

cabbage cultivar trials in Pennsylvania. HortTechnology 10(1):218-221.

Paull, R.E., G. Uruu, and A. Arakaki. 2000. Variation in the cooked and chipping quality of taro. HortTechnology. 10 (4):823-829.

Poysa, V.W., R. Garton, W.H. Courtney, J.G. Metcalf, and J. Muehmer. 1986. Genotype-environment interactions in processing tomatoes in Ontario. J. Amer. Soc. Hort. Sci. 111(2):293-297.

Quintana, J.M., H.C. Harrison, J. Nienhius, J.P. Palta, and M.A. Grusak. 1996. Variation in calcium concentration among sixty Si families and four cultivars of snap bean (*Phaseolus vulgaris* L.). J. Amer. Soc. Hort. Sci. 121(5):789-793.

Sanders, D.C. (Ed.). 2001. 2001-2002 vegetable crops guidelines for the Southeastern U.S. 192 pp, Vance Pub., Lincolnshire, IL.

Saville, D.J. 1990. Multiple Comparison Procedures: The Practical Solution. The American Statistician. 44(2):174-180.

Schultheis, J.R. and S.A. Waters. 1998. Yield and virus resistance of summer squash cultivars and breeding lines in North Carolina. HortTechnology 8(1):31-39.

Simonne, E.H. (Ed.). 1996a. Fall 1995 commercial vegetable variety trials. Ala. Ag. Exp. Sta. Prog. Rpt. 129, Auburn Univ., Ala.

Simonne, E.H. (Ed.). 1996b. Spring 1996 commercial vegetable variety trials. Ala. Ag. Exp. Sta. Prog. Rept. 130, Auburn Univ., Ala.

Simonne, A.H., E.H. Simonne, R.R. Eitenmiller, H.A. Mills and N.R. Green. 1997. Ascorbic acid and provitamin A contents in unusually colored bell peppers (*Capsicum annuum* L.) J. Food Comp. Anal. 10(4):299-311.

Simonne, E.H. (Ed.). 1997a. Fall 1996 commercial vegetable variety trials. Ala. Ag. Exp. Sta. Prog. Rept. 131, Auburn Univ., Ala.

Simonne, E.H. (Ed.). 1997b. Spring 1997 commercial vegetable variety trials. Ala. Ag. Exp. Sta. Prog. Rept. 132, Auburn Univ., Ala.

Simonne, E.H. (Ed.). 1998a. Fall 1997 commercial vegetable variety trials. Ala. Ag. Exp. Sta. Prog. Rept. 133, Auburn Univ., Ala.

Simonne, E.H. (Ed.). 1998b. Spring 1999 commercial vegetable variety trials. Regional Bul. 1, Auburn Univ., Ala.

Simonne, E.H., A.H. Simonne and R. Boozer. 1999. Yield, ear characteristics and consumer acceptance of selected white sweet corn varieties in the Southeast. HortTechnology 9(2):289-293.

Simonne, E.H. (Ed.). 1999a. Fall 1998 commercial vegetable variety trials. Regional Bul. 2, Auburn Univ., Ala.

Simonne, E.H. (Ed.). 1999b. Spring 1999 commercial vegetable variety trials. Regional Bul. 3, Auburn Univ., Ala.

Simonne, E.H. (Ed.). 2000. Fall 1999 commercial vegetable variety trials. Regional Bul. 4, Auburn Univ., AL.

Southwick, S.M., J.T. Yeager, J. Osgood, W. Olson, M. Norton, and R. Buchner. 1999. Performance of New Marianna rootstocks in California for 'French' prune. HortTechnology 9(3):498-505.

Swallow, W.H. 1984. Those overworked and oft-misused mean separation procedures-Duncan's, LSD, etc. Plant Disease 68:919-921.

Tukey, J.W. 1991. The Philosophy of Multiple Comparisons. Statistical Science 6(1):100-116.

Wang, M and I.L. Goldman. 1996. Phenotypic variation in free folic acid content among F1 hybrids an open-pollinated cultivars of red beet. *J. Amer. Soc. Hort. Sci.* 121(6):1040-1042.

Westfall, P.H., R.D. Tobias, D. Rom, R.D. Wolfinger and Y. Hochberg. 1999. Multiple Comparisons and Multiple Tests Using the SAS System. SAS Institute Inc., Cary, NC.

Method=TUKEY, Nominal FWE=0.05, nrep=100, Seed=12345 1
 True means = (5600,8400,3000,9600,4400,2800), n=3, s=2000

Quantity	Estimate	---95% CI----
Complete Power	0.00000	(0.000,0.000)
Minimal Power	0.95000	(0.907,0.993)
Proportional Power	0.26067	(0.231,0.290)
Directional FWE	0.00000	(0.000,0.000)

Method=TUKEY, Nominal FWE=0.05, nrep=100, Seed=12345 2
 True means = (5600,8400,3000,9600,4400,2800,5600,8400,3000,9600,4400,2800),
 n=3,s=2000

Quantity	Estimate	---95% CI----
Complete Power	0.00000	(0.000,0.000)
Minimal Power	0.99000	(0.970,1.000)
Proportional Power	0.20283	(0.181,0.225)
True FWE	0.00000	(0.000,0.000)
Directional FWE	0.00000	(0.000,0.000)

Table 1: Table 1. List of some essential and desirable traits of manuscripts submitted to the “Variety Testing and Evaluation” section of HortTechnology.

Trait	Necessary	Desirable
Clear objectives	****	
Quality rating for the trials		*
Definition and discussion of the area where trial results may apply		*
Follow standard production practices		***
Reference variety		***
Multiple seed source		**
Replication	****	
Multi-year/location		***
Power calculation		**
Non parametric statistics		**
Clear conclusion	****	