Journal of Applied Horticulture, 11(1): 10-16, January-June, 2009

Extracting within-experiment precision of horticultural experiments useful for meta-analysis

Guido Knapp[1], Bimal K. Sinha[2] and Dihua Xu[2]

[1]Department of Statistics, TU Dortmund University, 44221 Dortmund, Germany, [2]Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA

Abstract

For combining results from independent experiments, it is essential that information about the precision of the estimates of treatment effects is available. In publications of horticultural experiments, the results of multiple comparisons tests are often reported without sufficient information about the precision of the experiments. Based on limited information of the precision of an experiment such as treatments with the same letter are not significantly different, we develop a method for extracting a possible range of the precision of the experiment which can then be used for meta-analysis. The procedure is demonstrated using a real data example where alternatives to methyl bromide are studied in pre-plant soil fumigation. We also provide an R program which computes the possible range of the precision.

Key words: Duncan's multiple range test, Student-Newman-Keuls multiple range test, Fisher's LSD test, standardized mean differences, ratio of means, random effects meta-analysis.

Introduction

Meta-analysis (MA) has become a common and widely accepted tool in different spheres of sciences including horticulture for combining tests of significance as well as methods for comparing differences between treatments. Olkin and Shaw (1995) presented methods for the latter, the standardized difference of two normal means is the effect size of interest and they also presented the statistical meta-analysis in the fixed effects model, that is, they assumed homogeneous effect sizes in all the experiments eligible for meta-analysis. A major obstacle to conducting meta-analysis is the non-availability of appropriate experimental results. As Olkin and Shaw (1995) said "The minimum information required for quantitative research synthesis includes means, sample sizes, and either standard errors or standard deviations."

Shaw and Larson (1999) conducted a meta-analysis of strawberry yield response to preplant soil fumigation with combinations of methyl bromide-cloropicrin and several alternative systems. They followed the lines of Olkin and Shaw (1995) and an inclusion criterion was that means, sample sizes, and either standard errors or standard deviations of the treatments had to be reported in the published articles. Since the published results of four studies, basically eligible for that meta-analysis, lacked these necessary parameters for inclusion, these four studies were omitted.

The quality of reporting results in the published articles determines the quality of the meta-analysis. Unfortunately, Olkin and Shaw's summation about the minimum information required for meta-analysis is still often ignored by presenting results of horticultural experiments in publications nowadays. The lack of information about the precisions of the estimated means in most published articles was a major obstacle for the meta-analysis of Porter *et al.* (2006) on validating the yield performance of alternatives to methyl bromide for preplant fumigation. To circumvent the problem of missing within-experiment precision, Porter *et al.* (2006) simply assumed that the between-experiment variability is so large compared to within-experiment variability that the latter one can be ignored in the analysis. Obviously, this assumption is a very critical one and gives all the different experiments the same weight in the meta-analysis.

But sometimes the results of multiple testing procedures for comparing the treatment means are reported using the presentation that groups of means with the same letter are not significantly different. For instance, Student-Newman-Keuls' or Duncan's multiple range test are applied in the analysis of horticultural experiments. In this paper we present a method to demonstrate how to extract information on the within-experiment precision when only the results of a multiple range test are reported besides the estimated means of the treatments.

The paper is organized as follows: In Section 2 we briefly summarize the basic ideas of multiple range tests. In Section 3 we present the extraction method using a simulated data set. Section 4 contains the results of the extraction method using the reported results from two published articles. Since the sample sizes (number of observations or replicates) are often not available from the published articles, we present a meta-analytical approach in Section 5 using the ratio of means as effect size of interest. In Section 6, we give some concluding remarks and show how to extract the precision when simultaneous test procedures like Fisher LSD, Scheffe, or Tukey test were used in the statistical analysis of the horticultural experiments. In the Appendix, an R code is given to extract the precision from experiments when the results of Duncan's multiple range test or Fisher LSD are known.

Multiple Range Tests

Let Y_{ij}, $i=1, \ldots, r$, $j=1, \ldots, n$, be r independent samples of n independently, normally distributed random variables with a common variance σ^2 and expectations $E(Y_{ij}) = \mu_i$, $i=1, \ldots r$, $j=1, \ldots, n$.

Following Miller (1981), the basic credo of multiple range tests is: the difference between any two means in a set of r means is significant provided that the range of each and every subset which contains the given means is significant according to an α_p-level studentized range test where p is the number of means in the subset concerned.

Let S^2 be the error mean sum of squares from one-way analysis of variance and assume S^2 is a multiple of a χ^2-random variate with $r(n-1)$ degrees of freedom. The α_p-level studentized range test is then conducted by comparing the range (divided by $S/\sqrt{n}$) of the p means involved with the critical value $q_{p,r(n-1),\alpha_p}$ of the studentized range distribution.

In Student-Newman-Keuls multiple range test, the α_p-levels are chosen as

$$\alpha_p = \alpha, \quad p = 2,\ldots,r, \tag{1}$$

whereas Duncan's multiple range test uses, see Miller (1981),

$$\alpha_p = 1-(1-\alpha)^{p-1}, \quad p = 2,\ldots,r. \tag{2}$$

Given a subset of p means, if $\overline{Y}^{(p)}_{\max}$ denotes the largest mean in this subset and $\overline{Y}^{(p)}_{\min}$ the smallest one, the corresponding null hypothesis is then rejected at level α_p when

$$\frac{\overline{Y}^{(p)}_{\max}-\overline{Y}^{(p)}_{\min}}{S/\sqrt{n}} > q_{p,r(n-1),\alpha_p}. \tag{3}$$

Extracting method using a simulated data set

Let us consider a simulated data set with five treatments each with ten replications. The standard deviation is $\sigma = 10$ and we choose the following true means: $\mu_1 = \mu_2 = 510$ and $\mu_3 = \mu_4 = \mu_5 = 490$.

Then, starting with the global null hypothesis, that is, $p = 5$, the following hypotheses are sequentially tested:

$p = 5$ $H_0^{5}: \mu_1=\mu_2=\mu_3=\mu_4=\mu_5$

$p = 4$ $H_0^{4,1}: \mu_1=\mu_2=\mu_3=\mu_4$ $H_0^{4,2}: \mu_2=\mu_3=\mu_4=\mu_5$

$p = 3$ $H_0^{3,1}: \mu_1=\mu_2=\mu_3$ $H_0^{3,2}: \mu_2=\mu_3=\mu_4$ $H_0^{3,3}: \mu_3=\mu_4=\mu_5$

$p = 2$ $H_0^{2,1}: \mu_1=\mu_2$ $H_0^{2,2}: \mu_2=\mu_3$ $H_0^{2,3}: \mu_3=\mu_4$ $H_0^{2,4}: \mu_4=\mu_5$

In case, for instance, the null hypothesis $H_0^{3,1}$ cannot be rejected at level α_3, the hypotheses $H_0^{2,1}$ and $H_0^{2,2}$ are also not rejected without further testing.

We used the function RAND('NORMAL') in SAS 9.1.3 and the ROUND function to obtain the data given in Table 1.

Using SAS PROC GLM, we obtain an error mean square of $S^2 = 85.5244$. For Student-Newman-Keuls (SNK) and Duncan's multiple range test, we get the same grouping, see Table 2. The grouping in Table 2 means that treatments with the same letter are not significantly different. Thus we can conclude that we cannot reject $H_0^{2,1}: \mu_1=\mu_2$ and $H_0^{3,3}: \mu_3=\mu_4=\mu_5$ Consequently, we also accept $H_0^{2,3}: \mu_3=\mu_4$ and $H_0^{2,4}: \mu_4=\mu_5$ without further testing. All the other null hypotheses are rejected.

Table 2. Grouping for Student-Newman-Keuls (SNK) and Duncan's multiple range test

Grouping	Mean	Treatment
A	512.2	1
A	511.4	2
B	492.2	5
B	490.7	4
B	483.9	3

Now, let us assume that only the results from Table 2 are available. The question is: can we extract any information about the within-experiment precision, that is, $S/\sqrt{n}$, from this table? Note that it holds in the above example

$$\frac{S}{\sqrt{n}} = \frac{\sqrt{85.5244}}{\sqrt{10}} = 2.924456.$$

Recall that we reject the null hypothesis at level α_p if

$$\frac{\overline{Y}^{(p)}_{\max}-\overline{Y}^{(p)}_{\min}}{S/\sqrt{n}} > q_{p,r(n-1),\alpha_p}, \tag{4}$$

that is, if

$$\overline{Y}^{(p)}_{\max}-\overline{Y}^{(p)}_{\min} > q_{p,r(n-1),\alpha_p} S/\sqrt{n} \tag{5}$$

or if

$$\frac{\overline{Y}^{(p)}_{\max}-\overline{Y}^{(p)}_{\min}}{q_{p,r(n-1),\alpha_p}} > S/\sqrt{n}. \tag{6}$$

Conversely, we do not reject the null hypothesis at level α_p if

$$\frac{\overline{Y}^{(p)}_{\max}-\overline{Y}^{(p)}_{\min}}{S/\sqrt{n}} \le q_{p,r(n-1),\alpha_p}, \tag{7}$$

that is, if

$$\overline{Y}^{(p)}_{\max}-\overline{Y}^{(p)}_{\min} \le q_{p,r(n-1),\alpha_p} S/\sqrt{n} \tag{8}$$

or if

$$\frac{\overline{Y}^{(p)}_{\max}-\overline{Y}^{(p)}_{\min}}{q_{p,r(n-1),\alpha_p}} \le S/\sqrt{n}. \tag{9}$$

It is obvious from (5) and (8) that for each subset of p means the critical range is identical, namely

$$q_{p,r(n-1),\alpha_p} S/\sqrt{n}.$$

Furthermore, if a hypothesis is rejected at level α_p, then we know from (6) that the standard error $S/\sqrt{n}$ is **at most**

$$\frac{\overline{Y}^{(p)}_{\max}-\overline{Y}^{(p)}_{\min}}{q_{p,r(n-1),\alpha_p}},$$

that is, each rejected hypothesis gives an information about an upper bound of the standard error. Calculating the upper bounds from all rejected hypotheses and taking the minimum of all possible upper bounds, we obtain a sharp upper bound of the within-experiment standard error. To facilitate the computation, it is sufficient to consider only the rejected hypothesis with the smallest range of means for subsets of magnitude p, as this range provides the smallest upper bound for the within-experiment standard error.

Table 1. Simulated data set for five treatments

Treatment	Observations									
1	506	509	502	514	504	513	514	538	506	516
2	501	513	518	505	524	498	516	512	512	515
3	479	488	482	491	487	479	479	504	482	468
4	489	485	490	498	500	487	496	479	501	482
5	473	496	495	498	498	482	501	491	482	507

Conversely, if a hypothesis is not rejected at level α_p, then we know from (8) that the standard error $S/\sqrt{n}$ is **at least**

$$\frac{\overline{Y}^{(p)}_{\max} - \overline{Y}^{(p)}_{\min}}{q_{p,r(n-1),\alpha_p}},$$

that is, each non-rejected hypothesis gives an information about a lower bound of the standard error. Calculating the lower bounds from all non-rejected hypotheses and taking the maximum of all possible lower bounds, we obtain a sharp lower bound of the within-experiment standard error. To facilitate the computation, it is again sufficient to consider only the non-rejected hypothesis with the largest range of means for subsets of magnitude *p*, as this range provides the largest lower bound for the within-experiment standard error.

Let us now apply this method to the results from Table 2. The critical values of Student-Newman-Keuls multiple range test are

$q_{5,45,.05}$=4.0184, $q_{4,45,.05}$=3.7727, $q_{3,45,.05}$=3.4275 and $q_{2,45,.05}$=2.8484

and the critical values of Duncan's multiple range test are

$q_{5,45,.1855}$=3.1617, $q_{4,45,.1426}$=3.0919, $q_{3,45,.0975}$=2.9954 and $q_{2,45,.05}$=2.8484

In our example only the hypotheses $H_0^{2,1}: \mu_1=\mu_2$ and $H_0^{3,3}: \mu_3=\mu_4=\mu_5$ were tested and not rejected. Consequently, we can extract the following possible lower bounds:

H_0	Range	SNK	Duncan
$H_0^{3,3}$	$\overline{Y}_5-\overline{Y}_3$	2.4216	2.7709
$H_0^{2,1}$	$\overline{Y}_1-\overline{Y}_2$	0.2809	0.2809

For the determination of the upper bound, we consider, for each *p*, the smallest observed range of the rejected null hypotheses and calculate the bounds using the critical values of both tests. This results in

p	Range	SNK	Duncan
5	$\overline{Y}_1-\overline{Y}_3$	7.0426	8.9509
4	$\overline{Y}_1-\overline{Y}_4$	5.6989	6.9536
3	$\overline{Y}_1-\overline{Y}_5$	5.8351	6.6768
2	$\overline{Y}_2-\overline{Y}_5$	6.7407	6.7407

The resulting range for possible values of the standard error is [2.4216,5.6989] using Student-Newman-Keuls multiple range test and [2.7709,6.6768] using Duncan's test.

Note that for *p > 2*, Student-Newman-Keuls multiple range test is more informative for the upper bound of the standard error than Duncan's multiple range test.

In the above calculation we have used the information that the number of replications is ten for each treatment. Sometimes, this information is not available from the published articles and, thus, no information about the error degrees of freedom can be deduced.

In Table 3, critical values of Student-Newman-Keuls multiple range test are presented for *p=2(1)10* and several error degrees of freedom (*ddf*). The corresponding critical values of Duncan's multiple range test are given in Table 4. For given *p* and increasing *ddf*, the critical values decrease but converge to a certain value. In case no information on the *ddf* are available, a possible choice would be using the limiting value, that is, $C_{p,\infty,\alpha}$ which can be, for instance, determined with the SAS function PROBMC [*e.g.*, x = PROBMC('RANGE', ., 1- alpha , ., p) with alpha = $1-(1-\alpha)^{p-1}$ or alpha = α].

Applying the limiting values in the above calculations we obtain the following possible lower bounds :

H_0	Range	SNK	Duncan
$H_0^{3,3}$	$\overline{Y}_5-\overline{Y}_3$	2.5042	2.8440
$H_0^{2,1}$	$\overline{Y}_1-\overline{Y}_2$	0.2886	0.2886

The possible upper bounds are given as

p	Range	SNK	Duncan
5	$\overline{Y}_1-\overline{Y}_3$	7.3361	9.1606
4	$\overline{Y}_1-\overline{Y}_3$	5.9177	7.1271
3	$\overline{Y}_1-\overline{Y}_4$	6.0341	6.8530
2	$\overline{Y}_3-\overline{Y}_5$	6.9269	6.9269

The resulting range for possible values of the standard error is [2.5004,5.9177] using Student-Newman-Keuls multiple range test, and [2.8440,6.8530] using Duncan's test. With respect to the upper bound, the use of the limiting critical value is a conservative choice.

We have explicitly demonstrated the idea of the extraction method in this section, but our example is restricted to the situation that the means can be divided into disjunct subgroups of non-significant means (or treatments). In practice, however,

Table 3. Critical values $C_{p,\ ddf,\ 0.05}$ of Student-Newman-Keuls multiple range test

ddf	Number of means *p*								
	2	3	4	5	6	7	8	9	10
10	3.1511	3.8768	4.3266	4.6543	4.9120	5.1242	5.3042	5.4605	5.5984
20	2.9500	3.5779	3.9583	4.2319	4.4452	4.6199	4.7676	4.8954	5.0079
50	2.8405	3.4159	3.7584	4.0020	4.1904	4.3437	4.4727	4.5839	4.6814
100	2.8058	3.3646	3.6950	3.9289	4.1093	4.2557	4.3785	4.4842	4.5768
1000	2.7752	3.3194	3.6393	3.8647	4.0379	4.1781	4.2954	4.3962	4.4843
10000	2.7721	3.3150	3.6338	3.8584	4.0309	4.1704	4.2872	4.3875	4.4751
∞	2.7718	3.3145	3.6332	3.8577	4.0301	4.1696	4.2863	4.3865	4.4741

Table 4. Critical values $C_{p,\ ddf,\ 0.05}$ of Duncan's multiple range test

ddf	Number of means *p*								
	2	3	4	5	6	7	8	9	10
10	3.1511	3.2928	3.3763	3.4297	3.4652	3.4891	3.5052	3.5156	3.5218
20	2.9500	3.0965	3.1896	3.2546	3.3026	3.3392	3.3678	3.3905	3.4086
50	2.8405	2.9876	3.0843	3.1544	3.2082	3.2511	3.2862	3.3155	3.3403
100	2.8058	2.9527	3.0505	3.1217	3.1771	3.2219	3.2590	3.2904	3.3173
1000	2.7752	2.9218	3.0200	3.0925	3.1494	3.1957	3.2345	3.2676	3.2964
10000	2.7721	2.9188	3.0170	3.0896	3.1466	3.1931	3.2320	3.2653	3.2943
∞	2.7718	2.9184	3.0167	3.0893	3.1463	3.1928	3.2317	3.2651	3.2941

the subgroups of non-significant means (or treatments) usually overlap. In the Appendix a computer program is given written in R (R Development Core Team, 2008) which can be used to extract the possible range of standard errors in the general scenario of overlapping subgroups.

Application to real data examples

We consider some results from Bartual *et al.* (2002) for demonstrating the extraction method in a real data scenario. Bartual *et al.* (2002) studied alternatives to methyl bromide in preplant soil fumigation. Seven treatments were investigated, where treatment 1 is a non treated control, treatment 2 is the standard methyl bromide treatment, and treatments 3 to 7 are alternative treatments. We refer to Bartual *et al.* (2002) for details.

The published article contains the information that the experimental design consisted of two years crop with a complete randomized block with three replications in the first year. The treatments were repeated on the same plot for a second year but with only two replicates. Duncan's multiple range tests were done for statistical comparison among treatments.

Several outcomes were measured like marketable yield, first quality fruit yield, first quality fruit size, and percentage of second quality fruit yield. The observed outcomes for marketable yield and first quality fruit size along with the results of Duncan's multiple range tests are reproduced in Table 5.

With seven treatments and three replicates in a completely randomized block design, the denominator degrees of freedom are *ddf = 12* for the first year. Only two replicates in the second year provide *ddf = 6*.

Table 5. Marketable yield and first quality fruit size in the first and second year of planting

Treatment	Marketable yield		First quality fruit size	
	First year	Second year	First year	Second year
1	319 C	392 C	17.6 C	17.3 B
2	544 A	738 A	19.4 A	19.5 A
3	513 A	683 AB	19.6 A	20.1 A
4	562 A	579 AB	18.7 B	18.2 B
5	554 A	542 BC	18.6 B	18.2 B
6	427 B	410 C	18.6 B	17.7 B
7	284 C	193 D	18.4 B	16.1 C

Applying the extraction method from the previous section, we obtain the standard errors of estimated marketable yield in both years as [14.7927,27.910] (first year) and [44.3336,47.1219] (second year)

Replacing the denominator degrees of freedom by infinity (∞) yields [16.2432,31.0267] (first year) and [54.4815,57.9080] (second year)

The interval for the second year is relatively tight, because we know that the critical range of two means for *p = 3* is between 159 and 169.

For the first quality fruit size, the limits of the standard errors are [0.0906,0.2272] (first year) and [0.2466,0.3468] (second year)

Replacing the denominator degrees of freedom by infinity (∞) yields [0.0994,0.2525] (first year) and [0.2983,0.4329] (second year)

Since there is one replicate less in the second year, one expects that the result of the first year is more precise. This is reflected in the deduced ranges of the standard error.

Note that we can easily extract the error mean sum of squares if the number of replicates (blocks) in a completely randomized block design is known.

A small meta-analysis

Olkin and Shaw (1995) used the standardized mean difference as the effect size for combining differences of treatments from several independent experiments. In our scenario, this combination method is possible if all the experiments provide the number of replicates besides the treatment means and the results from the multiple range test. In case the number of replicates (blocks) is unknown, we can use the extracted information of the standard errors to combine the results of several experiments using the ratio of means as the effect size for comparing treatments. Combining results using the standardized mean difference or the ratio of means will be demonstrated in the next two sections. We will consider only the combination of two independent experiments. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be generally two effect size estimates of a common effect size, say θ, with estimated variances $\widehat{\mathrm{Var}}(\hat{\theta}_1)$ and $\widehat{\mathrm{Var}}(\hat{\theta}_2)$, respectively. Then, the combined estimator of θ is given as

$$\hat{\theta}=\hat{w}_1\hat{\theta}_1 + \hat{w}_2\hat{\theta}_2$$

with $\hat{w}_1=\dfrac{[\widehat{\mathrm{Var}}(\hat{\theta}_1)]^{-1}}{[\widehat{\mathrm{Var}}(\hat{\theta}_1)]^{-1}+[\widehat{\mathrm{Var}}(\hat{\theta}_2)]^{-1}}$ and $\hat{w}_2 = 1-\hat{w}_1$.

Since the extraction method provides a range of possible values of the standard errors, we only will consider point estimates of the common effect θ in the present paper. For further statistical inference on θ, we refer to the textbooks of Hedges and Olkin (1985) and Hartung, Knapp, and Sinha (2008), or to the overview of Olkin and Shaw (1995).

Standardized mean difference

Recall that the standardized mean difference is defined as

$$\delta=\frac{\mu_1-\mu_2}{\sigma},$$

where μ_1 is the expected value in the first group (treatment), μ_2 the expected value in the second group (treatment), and σ the common standard deviation of both groups (treatments). An estimator of δ, called Hedges's g, is given as (Hartung *et al.* (2008))

$$g=\frac{\bar{Y}_1-\bar{Y}_2}{S},$$

where $\bar{Y}_1$ is the sample mean in the first group (treatment), $\bar{Y}_2$ the sample mean in the second group (treatment), and S the pooled standard deviation of both groups (treatments) or the root error mean sum of squares in case of more than two treatments considered in an experiment. Since g is biased for δ, an approximately unbiased estimator of δ is given as (Hartung *et al.* (2008))

$$g^{*}=\left(1-\frac{3}{4N-9}\right)g$$

with $N = n_1 + n_2$ and n_i, i=1,2, is the number of replicates in the ith group (treatment).

For the variance of g, it holds that (Hartung *et al.* (2008))

$$\mathrm{Var}\,(g) \approx \frac{1}{\tilde{n}} + \frac{\theta^2}{2(N-3.94)},$$

which can be easily estimated by

$$\widehat{\mathrm{Var}}\,(g) \approx \frac{1}{\tilde{n}} + \frac{g^2}{2(N-3.94)},$$

where, $\tilde{n} = n_1 n_2 / N$.

Let us now consider the estimated marketable yield in the first and second year from the real data example of Section 4. We first want to determine the effect of the (active) standard treatment (No. 2) compared to the control treatment (No. 1). Since the ranges of the standard errors are [14.7927,27.9101] in the first year and [44.3336,47.1219] in the second year with knowing the number of replicates, the ranges for the root error mean sum of squares are [25.6217,48.3417] (first year) and [62.6972,66.6404] (second year). Consequently, the true observed standardized mean differences (Hedges's g) lies in the range [4.6544,8.7816] (first year) and [5.1920,5.5186] (second year).

Applying the bias correction to Hedges's g, see Hedges and Olkin (1985) or Hartung, Knapp, and Sinha (2008), we obtain the possible range of observed values for g^* as [3.7235,7.0253] (first year) and [2.9669,31535] (second year)

Note that the larger the estimated standardized mean difference the larger the estimated variance given a fixed sample size. The estimated variances for Hedges's g are [5.9247,19.3844] (first year) and [225.6405,254.792] (second year)

For the bias corrected estimates g^* we obtain the following estimated variances [4.0318,12.6460] (first year) and [74.3514,83.871] (second year)

Let us combine the results for the marketable yield of the two years considering the extreme cases, that is, using the limits of the extracted intervals. Combining the results using Hedges's g, we obtain the following range of possible values of the common effect size: [4.6682,8.5509]. Using the bias corrected g^*'s yields the range: [3.6846,6.5180].

Ratio of means: Let $\rho = \frac{\mu_1}{\mu_2}$ be the ratio of two (normal) means and

$$\xi = \ln(\mu_1) - \ln(\mu_2)$$

be the logarithm of ρ. An estimate of ξ is readily given as

$$\hat{\xi} = \ln(\bar{Y}_1) - \ln(\bar{Y}_2)$$

Using the delta method, an estimate of the variance of p can be deduced as

$$\widehat{\mathrm{Var}}(\hat{\xi}) = S^2\left(\frac{1}{n_1\bar{Y}_1^2} + \frac{1}{n_2\bar{Y}_2^2}\right)$$

where S^2 is the pooled sample variance of both groups (treatments) or the error mean sum of squares in case of more than two treatments considered in an experiment. Note that for identical replications, that is, $n_1 = n_2 = n$, the variance estimate reads

$$\widehat{\mathrm{Var}}\,(\hat{\xi}) = \frac{S^2}{n}\left(\frac{1}{\bar{Y}_1^2} + \frac{1}{\bar{Y}_2^2}\right)$$

and the knowledge of the standard error besides the treatment means is sufficient to calculate this variance estimate.

Let us combine the results for the first quality fruit size of the two years for the ratio of means of treatment 2 and 1 using the extracted standard errors with infinite degrees of freedom, that is, assuming that the number of replicates is unknown. Recall that the extracted standard errors are [0.0994,0.2525] (first year) and [0.2983,0.4389] (second year)

Since the effect size here depends only on means, we obtain, in contrast to the standardized mean difference, exactly one estimate of the effect size from each experiment. But we get a range of possible values of the estimated variances for estimated effect sizes. For the first year of the first quality fruit size, we get

$$\hat{\xi}_1 = \ln(19.4) - \ln(17.6) = 0.0974$$

with

$$\widehat{\mathrm{Var}}(\hat{\xi}_1) \in [0.00006, 0.00038],$$

and for the second year

$$\hat{\xi}_1 = \ln(19.5) - \ln(17.3) = 0.1197$$

with

$$\widehat{\mathrm{Var}}(\hat{\xi}_2) \in [0.00053, 0.00112].$$

Clearly, the weighted average lies between 0.0974 and 0.1197. We obtain the minimum value of all possible weighted averages by using the smallest possible value of $\widehat{\mathrm{Var}}(\hat{\xi}_1)$ and the largest possible value of $\widehat{\mathrm{Var}}(\hat{\xi}_2)$. Conversely, the maximum value of all possible weighted averages is given by using the largest possible value of $\widehat{\mathrm{Var}}(\hat{\xi}_1)$ and the smallest possible value of $\widehat{\mathrm{Var}}(\hat{\xi}_1)$. This leads to the following range of possible estimates of the common effect size: [0.0985,0.1067].

Backtransforming the results to the original scale we obtain the range of estimates of ρ as

[exp(0.0985), exp(0.1067)]=[1.1035,1.1126]

Concluding Remarks

In this paper we have demonstrated how information about the within-experiment variability can be extracted when several treatments are compared and only the treatment means and the results of a multiple range test, either Student-Newman-Keuls or Duncan, are reported. Based on the results of the multiple range test we can deduce a lower and an upper bound of possible values of the root error mean sum of squares.

The extracted within-experiment variability can be used in meta-analysis, when the results of several independent experiments should be combined. Possible effect sizes are the standardized mean difference and the ratio of means. For using the standardized mean difference, the number of replicates has to be additionally known to determine the common variance estimate or the error mean sum of squares. Moreover, we can only calculate a range of possible effect size estimates leading also to a possible range of variance estimates. Fortunately, there is a one-to-one relationship between effect size estimate and variance estimate. Using the ratio of means, we always get one estimate of the effect size per experiment but a range of possible variance estimates.

The extraction method described in this paper is a little bit elaborate as the critical ranges of the multiple range tests vary with the number of means. Since Tukey's and Scheffe's multiple comparisons or Fisher's LSD tests are simultaneous comparisons for all treatments and consequently have only one critical range, the interval of the standard error can be easily determined using the largest range of two means which do not lead to rejection of the null hypothesis and the smallest range of two means which lead to a rejection of the null hypothesis.

To demonstrate the less elaborate extraction method for the simultaneous multiple comparison methods, let us consider Fisher's LSD test. Let $\bar{Y}_i$ and $\bar{Y}_j$, $i \neq j$, be two sample means. Then Fisher's LSD test rejects the null hypothesis of equal means if

$$\left|\bar{Y}_i - \bar{Y}_j\right| \geq t_{v,\alpha/2} S\sqrt{2/n} = LSD$$

and $t_{v,\alpha/2}$ is the upper critical value of t_v.

If *LSD* is explicitly given, then it holds

$$\frac{S}{\sqrt{n}} = \frac{LSD}{t_{n,a/2}\sqrt{2}}.$$

Since the error degrees of freedom *v* are unknown, we approximate the standard error by

$$\frac{S}{\sqrt{n}} \approx \frac{LSD}{z_{a/2}\sqrt{2}}.$$

If *LSD* is not explicitly given but only the information *non-significant* then a lower bound of the standard error is

$$\frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{t_{n,a/2}\sqrt{2}} \left(\leq \frac{S}{\sqrt{n}}\right)$$

which, when *v* is unknown, can again be approximated by

$$\frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{z_{a/2}\sqrt{2}}.$$

If *LSD* is not explicitly given but it is known which differences of means are significant and which ones are not, then we proceed as follows. Consider the largest range of two means which do not lead to rejection of the null hypothesis, say $\left|\bar{Y}_i - \bar{Y}_j\right|$, $i \neq j$, and the smallest range of two means which lead to a rejection of the null hypothesis, say $\left|\bar{Y}_{(l)} - \bar{Y}_{(k)}\right|$, $l \neq k$. Then we have

$$\left|\bar{Y}_{(i)} - \bar{Y}_{(j)}\right| < LSD \quad \text{and} \quad \left|\bar{Y}_{(l)} - \bar{Y}_{(k)}\right| > LSD.$$

Consequently, $$\frac{\left|\bar{Y}_{(i)} - \bar{Y}_{(j)}\right|}{t_{v,a/2}\sqrt{2}} < \frac{S}{\sqrt{n}} \quad \text{and} \quad \frac{\left|\bar{Y}_{(l)} - \bar{Y}_{(k)}\right|}{t_{v,a/2}\sqrt{2}} > \frac{S}{\sqrt{n}},$$

and we can approximate the sought-after interval by

$$\left[\frac{\left|\bar{Y}_{(i)} - \bar{Y}_{(j)}\right|}{z_{a/2}\sqrt{2}}, \frac{\left|\bar{Y}_{(l)} - \bar{Y}_{(k)}\right|}{z_{a/2}\sqrt{2}}\right]$$

References

Bartual, R., V. Cebolla, J. Bustos, A.Giner and J.M. Lopez-Aranda, 2002. The Spanish project in alternatives to methyl bromide(2): The case of strawberry in the area of Valencia. *Acta Horticulturae*, 567: 431-434.

Hartung, J., G. Knapp and B.K. Sinha, 2008. *Statistical Meta-Analysis with Applications*. Wiley, New York.

Hedges, L. and I. Olkin, 1985. *Statistical Methods for Meta-Analysis*. Academic Press, Orlando, Fl.

Miller, R.G.J. 1981. *Simultaneous Statistical Inference*. Springer, New York.

Olkin, I. and D.V. Shaw, 1995. Meta-analysis and its applications in horticultural science. *HortScience,* 30: 1343-1348.

Porter, Ian J., L. Trinder, D. Partington, J. Banks, S. Smith, M. Hannah and N. Karavarsamis, 2006. *Validating the Yield Performance of Alternatives to Methyl Bromide for Pre-Plant Fumigation*. UNEP, Nairobi, Kenya.

R Development Core Team 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Shaw, D.V. and K.D. Larson, 1999. A meta-analysis of strawberry yield response to preplant soil fumigation with combinations of methyl bromide-chloropicrin and four alternative systems. *HortScience*, 34: 830-845.

Appendix: Computer code in R

Extracting the Standard Error from Multiple Range Test

Description: `multi` is used to extract the standard error from Duncan's Multiple Range Test or Fisher's LSD Test.

Idea: For one study, we compare every possible pair of treatments to find out if they have at least one letter in common. If this happens, we do not reject the null hypothesis for that pair of treatments, thus enabling us to get a lower bound for the standard error. On the other hand, if they have no letter in common, we get an upper bound.

Usage: multi(y.mean, y.mrt, method, method=c("Duncan", "Fisher"), alpha=.05)

Arguments:

`y.mean:` a vector of mean values reported in the study.

`y.mrt:` a vector of statistical comparison results reported in the study.

`method:` method used to report comparison results in the study, either Duncan's Multiple Range Test or Fisher's LSD Test. For Fisher's Test, we only consider the situation that neither LSD nor "not significant" is reported.

`alpha:` level for the reported statistical comparison results, default is .05

Details:

```
# Function for comparing two strings
# if they have at least one letter in common return "L", else return "U"
charcomp <- function(x, y) {
 nsep <- nchar(x) + nchar(y)
    # summing up number of letters in x, and number of letters in y
 xy <- c( strsplit(x,split=character(0))[[1]], strsplit(y,split=character(0
))[[1]])
    # combine x and y into a new vector, letters by letters
    # for example, x = "abc", y="a", then xy = c("a","b","c","a")
 nbind <- length(unique(xy))
    # report number of unique elements in xy
    # for previous example, will be "a", "b", "c", then return 3.
 if(nsep==nbind) {return("U")}
    # the two values are equal means x and y have no letter in common
 else {return("L")}
}
# Function for extracting the standard error
# from Duncan's Multiple Range Test or Fisher's LSD Test
multi <- function(y.mean, y.mrt, method=c("Duncan","Fisher"),
alpha=.05) {
 n <- length(y.mean)
n.pair <- choose(n, 2)  # number of possible pairs
c.pair <- combn(n, 2)  # all possible combinations of n choose 2
 bound <- decision <- rep(NA, n.pair)
 if(method=="Duncan") {
 y.rank <- 13- rank(y.mean, ties.method="min")
 for (i in 1:n.pair) {
  j <- c.pair[1, i]
  k <- c.pair[2, i]
  mean.diff <- abs(y.mean[j]- y.mean[k])
  rank.diff <- abs(y.rank[j]- y.rank[k])
  alpha.p <- 1- (1- alpha)^rank.diff
  bound[i] <- mean.diff / qtukey(1-alpha.p, rank.diff+1, 100000)
  decision[i] <- charcomp(y.mrt[j], y.mrt[k])
} }
 if(method==”Fisher”) {
for (i in 1:n.pair) {
  j <- c.pair[1, i]
  k <- c.pair[2, i]
  mean.diff <- abs(y.mean[j]- y.mean[k])
  bound[i] <- mean.diff /(qnorm(1-alpha / 2) * sqrt(2))
  decision[i] <- charcomp(y.mrt[j], y.mrt[k])
} }
 c(max(bound[decision==”L”], na.rm=T), min(bound[decision==”U”], na.rm=T))
}
```

Value: The lower bound and upper bound of the standard error will be returned.

Example:

```
yield <- c(392, 738, 683, 579, 542, 410, 193)
dmrt <- c(“c”, “a”, “ab”, “ab”, “bc”, “c”, “d”)
multi(yield, dmrt, method=“Duncan“, alpha=0.05)
[1] 54.48149 57.90800
```