Problems with a Type of Matched Subjects Design and Analysis

Steve P. Verrill* David E. Kretschmann*

Abstract

There is a type of blocked experiment that has the potential of being poorly designed and/or analyzed. Verrill *et al.* (1993, 1999, 2004) referred to such an experiment as a "predictor sort" experiment. David and Gunnink (1997) spoke of "artificial pairing." In text books it is sometimes referred to as a "matched pair" or "matched subjects" design. The associated design process is also sometimes described as "forming blocks via a concomitant variable." Improperly designed and/or analyzed, predictor sort experiments can be associated with incorrect/inadequate power calculations and sample sizes, incorrect tests of hypotheses, and incorrect confidence intervals. In this paper we review some of the results in the literature, add a section on multiple comparisons, and present results from power and confidence interval coverage simulations that emphasize the importance of the proper design and analysis of predictor sort experiments.

Key Words: Predictor sort sampling, artificial pairing, matched pairs, matched subjects, concomitant variable, blocked ANOVA, analysis of covariance

1. Introduction

There is a type of blocked experiment that has the potential of being poorly designed and/or analyzed. Verrill *et al.* (1993, 1999, 2004) referred to such an experiment as a "predictor sort" experiment. David and Gunnink (1997) spoke of "artificial pairing." In text books it is sometimes referred to as a "matched pair" or "matched subjects" design. The associated design process is also sometimes described as "forming blocks via a concomitant variable." In a wood research context, Warren and Madsen (1977) described the specimen allocation procedure as follows:

> One can take steps, however, to ensure that the inherent [initial] strength distributions of test and control samples are reasonably equivalent. Indeed, failure to do so can only throw doubt on the results.
>
> Specifically, then, all the boards in the experiment are ordered from weakest to strongest as nearly as can be judged from their moduli of elasticity, knot size, and slope of grain. To divide the material into J equivalent groups the first J boards, after ordering, are taken and randomly allocated one to each group. This is repeated with the second, third, fourth, etc., sets of J boards. The strength distributions of the resulting groups should then be essentially the same.

Here the response is lumber strength after a treatment, and the predictor/concomitant used to form blocks (of size J) would be some combination of lumber stiffness, knot size, and slope of grain (all of which can be measured non-destructively prior to specimen allocation).

In an agricultural context, the predictor/concomitant variable might be, for example, animal age, initial animal weight, or plot fertility in a previous trial. In a behavioral or educational context, the predictor might be, for example, IQ or performance on a pre-test.

*USDA Forest Service Forest Products Laboratory

In this paper we will refer to this type of design as a "predictor sort" design (because we sort specimens on the basis of a predictor that is correlated with the response, and then form blocks via collections of specimens with adjacent predictor values). Our theory will be established for the case in which the predictor and the response have a joint bivariate normal distribution.

In his 1999 paper, Verrill cited discussions of this type of experiment in example 3.3 of Cox (1958), section 8.2 of Steel and Torrie (1960), section 5.1 of Kirk (1968), section 13.17 of Finney (1972), example 11.3 of Ostle and Mensing (1975), Chapter 6 of Myers (1979), and example 6.13.1 of Snedecor and Cochran (1989). A more recent sampling of statistical texts found such experiments discussed in Kerlinger and Lee (1999), van Zutphen *et al.* (2001), example 5.1 of Toutenburg (2002), section 4.3 of Ruxton and Colegrave (2006), problem 3.8 of Casella (2008), Cozby and Bates (2011), Tuckman and Harper (2012), and section 8.1 of Kirk (2013).

Among the variables suggested as predictors/concomitants to be used to form blocks were age, reaction time, initial weight, concentration of blood constituent, degree of disease, time since college, IQ, scores on a cognitive ability measure, grade point average, prior school performance, and pretest achievement.

Improperly designed and/or analyzed, predictor sort experiments can be associated with incorrect/inadequate power calculations and sample sizes, incorrect tests of hypotheses, and incorrect confidence intervals. Verrill (1993) and David and Gunnink (1997) focused on potential problems with hypothesis tests given a predictor sort design. Verrill (1999) focused on confidence intervals on means. Verrill *et al.* (2004) focused on confidence intervals on quantiles. Because incorrect predictor sort designs and analyses can have serious adverse effects on decision-making, and because this fact has not yet become common knowledge among statisticians (or at least among text book authors), in this paper we review some of the results in the literature, add a section on multiple comparisons, and present the results from power and confidence interval coverage simulations that emphasize the importance of the proper design and analysis of predictor sort experiments.

In section 2 we focus on hypothesis tests. In section 3 we discuss confidence intervals on means. In section 4 we discuss Scheffé and Tukey multiple comparison tests and the corresponding simultaneous confidence intervals. And in section 5, we describe web/R programs that we have written to aid in the design and analysis of predictor sort experiments.

2. Hypothesis Tests

We first set some useful notation. Here, for ease of exposition, we will restrict ourselves to the one factor case. Let Y_{ij} denote the response for the ith block, $i \in \{1, \ldots, I\}$, of the jth treatment, $j \in \{1, \ldots, J\}$. Let ρ denote the correlation between the predictor/concomitant, X, and the response, Y. We assume that X and Y have a joint bivariate normal distribution.

In a non predictor sort case, the probability model for a blocked ANOVA would be

$$Y_{ij} = \mu_{\cdot j} + \mu_{i\cdot} + \sigma_Y \times \epsilon_{ij} \tag{1}$$

where $\mu_{\cdot 1}, \ldots, \mu_{\cdot J}$ denote the treatment effects, $\mu_{1\cdot}, \ldots, \mu_{I\cdot}$ denote the block effects, and the ϵ's are i.i.d. N(0,1)'s. In a predictor sort case, we have $n = JI$ specimens. To allocate these specimens, we order the X's and randomly assign the J specimens associated with the lowest X values, to the first block, the specimens associated with the next J lowest values to the next block, and so on. In this case, the correct probability model is

$$Y_{ij} = \mu_j + \sigma_Y \left(\rho \left(X_{k(i,j),n} - \mu_X \right) / \sigma_X + \sqrt{1-\rho^2} P_{ij} \right) \tag{2}$$

where $\mu_1, \ldots, \mu_J$ denote the treatment effects, $k(i,j) \in \{(i-1)J+1, \ldots, iJ\}$, $X_{l,n}$ denotes the lth order statistic among the X's, and the P_{ij}'s are i.i.d. N(0,1)'s that are independent of the X's.

The differences between models (1) and (2) (and, in particular, the fact that $X_{k(i,j_1),n} - X_{k(i,j_2),n}$ tends to be smaller than an arbitrary $X_1 - X_2$ and yet not equal to 0) are the source of both the advantages and the problems associated with predictor sort experiments. (More detailed heuristic discussions are provided in Verrill (1993), Verrill and Green (1996), Verrill (1999), and Verrill *et al.* (2004).)

Verrill (1993) proved the following theorem on hypothesis testing following a predictor sort allocation.

Theorem 1

Assume that the predictor variable and the variable of interest have a joint bivariate normal distribution with correlation ρ. Let the allocation of samples be as described in Section 1. (For the multi-factor case, enough adjacent (in predictor values) experimental units are chosen at a time to provide one additional observation for each cell.) Then, for a factor with J levels, for $0 \leq \rho < 1$, the asymptotic distribution of the ANOVA test statistic that treats the groups of adjacent (in predictor values) experimental units as a block is $\chi^2_{J-1}/(J-1)$. The asymptotic distribution of the ANOVA test statistic that ignores the block structure generated by these groups is $(1-\rho^2)\chi^2_{J-1}/(J-1)$.

Proof

See Verrill (1993) for a 1-factor proof, and Appendix G of Verrill and Kretschmann (2015) for a multi-factor proof.

Because of this asymptotic behavior, if we analyze a predictor sort experiment as a blocked ANOVA (where the blocks are formed of specimens with similar predictor/concomitant values — "matched subjects") and I is sufficiently large, the nominal size of the test will be approximately equal to the true size. (Actually, for very large ρ values, the true size is reduced from the nominal size even for fairly large samples. See the web version of table 1 referenced below.) However, if we ignore the blocks in our analysis, the actual size can be *much* lower than nominal size and power will suffer significantly. (We essentially end up comparing $(1-\rho^2)\chi^2_{J-1}$ random variables with χ^2_{J-1} critical values.)

To explore these effects we have performed a large power simulation of the 1-factor case. For all combinations of X,Y correlations 0.0, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99; number of treatments, J, equal to 2, 3, 5, 7, 9, 11, and 20; sample sizes, I, equal to 3, 5, 10, 20, and 40; and 21 noncentrality parameters, we performed 40,000 trials. We created two versions of each of the resulting data sets. One version was created by allocating the specimens in a data set to the J treatment conditions via a standard randomization. The second version was created by allocating the specimens in a data set to the J treatment conditions via a predictor sort.

We then performed performed seven hypothesis tests on the data sets, and two theoretical power calculations:

1. A standard 1-way analysis of variance on the non predictor sort version of the data set.

2. A standard (and thus incorrect) 1-way analysis of variance on the predictor sort version of the data set.

3. A *corrected* 1-way analysis of variance on the predictor sort version of the data set. The corrected F statistic is the standard 1-way statistic divided by $1-\hat{\rho}^2$.

4. A second *corrected* 1-way analysis of variance on the predictor sort version of the data set. The corrected F statistic is the standard 1-way statistic divided by $1-\rho^2_{\text{true}}$.

5. A 2-way analysis of variance on the predictor sort version of the data set. (The blocks are formed by specimens with adjacent [randomized within the block] values of the predictor.)

6. An analysis of covariance on the non predictor sort version of the data set.

7. An analysis of covariance on the predictor sort version of the data set.

8. The "theoretical" power for a a corrected 1-way analysis of variance on a predictor sort version of the data set:

$$\text{Prob}\left(\text{NCF}_{J-1,J(I-1)}(\gamma) > F^{-1}_{J-1,J(I-1)}(1-\alpha)\right) \tag{3}$$

   where NCF denotes a non-central F distribution function, $\gamma = \sum_{J=1}^{J} I(\mu_j - \bar{\mu}_.)^2/(\sigma^2_Y(1-\rho^2))$ is the non-centrality parameter of the noncentral F, and F^{-1} denotes the inverse of a central F distribution function.

9. The "theoretical" power for a 2-way anova on a predictor sort version of the data set:

$$\text{Prob}\left(\text{NCF}_{J-1,(J-1)(I-1)}(\gamma) > F^{-1}_{J-1,(J-1)(I-1)}(1-\alpha)\right) \tag{4}$$

   (The same non-centrality parameter is used in both (3) and (4).)

The results of these simulations for the $J = 2, 5$; $I = 3, 5, 10, 20, 40$; $\rho = 0.5, 0.7, 0.9$ cases are presented in table 1 of Verrill and Kretschmann (2015). The results for the remaining cases can be found at http://www1.fpl.fs.fed.us/ps15_table1.html. Plots that present a portion of the results of these simulations ($J = 2, 5$; $I = 10, 20$; $\rho = 0.0, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99$) appear in Verrill and Kretschmann (2015). The plot for the $J = 2$, $I = 10$, $\rho = 0.9$ case appears in Figure 1. In these plots, the "noncentrality parameter index" is the m in column 4 of the corresponding power table. See Appendix A of Verrill and Kretschmann (2015) for a discussion of this index. "th" is the theoretical power calculated by (3) and presented in column 12 of the power tables. "ps, ancov" is the power of an analysis of covariance after a predictor sort allocation (column 11 of the power tables). "ps, two-way" is the power of a blocked analysis of variance after a predictor sort allocation (column 9 of the power tables). "no ps, 1-way" is the power of a 1-way analysis of variance after a standard (non predictor sort) random allocation of specimens (column 5 of the power tables). "ps, 1-way, no rho" is the power of an uncorrected 1-way analysis of variance after a predictor sort allocation (column 6 of the power tables).

An analysis of these tables and plots yields the following conclusions:

1. Large increases in statistical power and/or sample size reductions can be gained by performing a predictor sort allocation and analysis. These improvements become larger as the correlation, ρ, between the predictor and the response increases. Specifically, if n samples are needed to achieve a given power when predictor sort allocation is not used, approximately $(1-\rho^2)n$ samples are needed to achieve the same power when predictor sort allocation is used. Thus, for example, a 0.7 correlation yields, roughly, a halving of necessary sample size.

2. It is a statistical *blunder* to perform a predictor sort allocation and then follow the allocation with a standard non predictor sort analysis of variance. Such an approach can considerably *reduce* power.

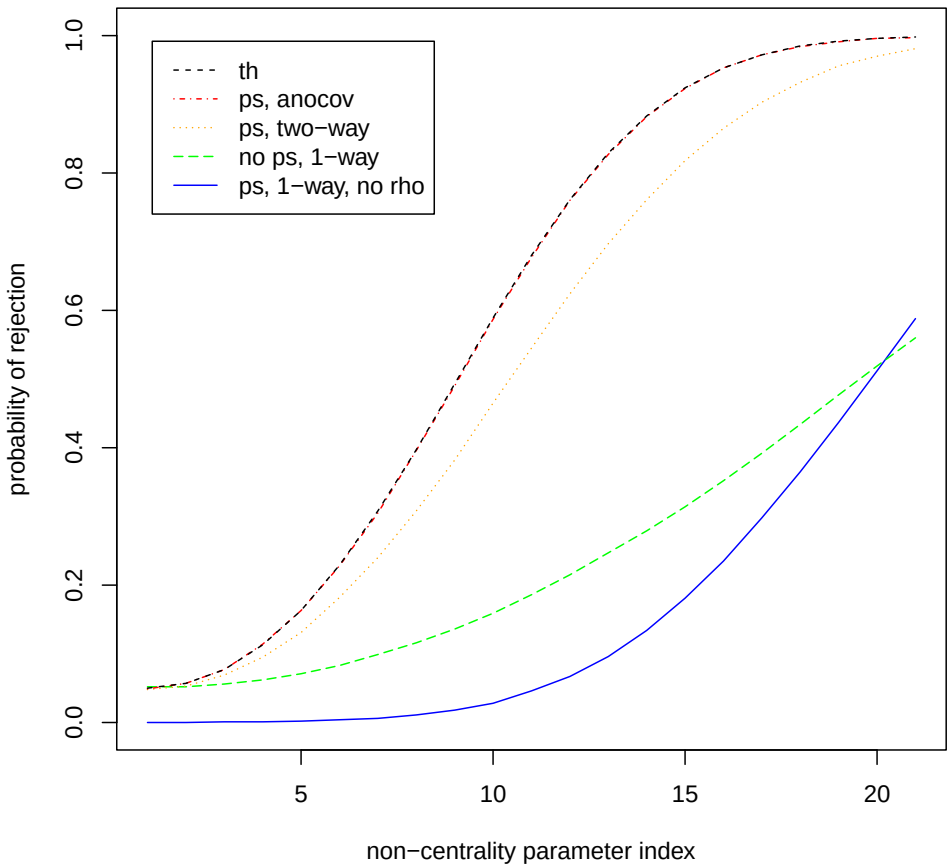

Figure 1: Power plot, $\rho = 0.9, J = 2, I = 10$.

3. After a predictor sort allocation has been performed, either an analysis of covariance or a blocked analysis of variance should be performed. For $\rho \leq 0.8$ and $I \geq 10$, the blocked analysis of variance performs almost as well as the analyis of covariance. For higher ρ and/or smaller I, the analysis of covariance performs better.

4. For $\rho \leq 0.8$ and $I \geq 10$, the theoretical power of a blocked anova can be well approximated by (4). For higher ρ, (4) overestimates the power available from a blocked anova (especially for lower I).

5. It is well known (also see result (47) in Appendix B of Verrill *et al.* (2015)) that the power of a 1-factor analysis of covariance for testing the hypothesis $\mu_1 = \ldots = \mu_J$ is given by

$$\text{Prob}(\text{NCF}_{J-1,IJ-(J+1)}(\gamma) > F^{-1}_{J-1,IJ-(J+1)}(1-\alpha))$$

where

$$\begin{aligned}
\sigma^2_{\text{anocov}} \times \gamma &= \sum_{j=1}^{J} I(\mu_j - \bar{\mu}_.)^2 - \left(\sum_{j=1}^{J} I(\mu_j - \bar{\mu}_.)(\bar{x}_{.j} - \bar{x}_{..}) \right)^2 / \sum_{j=1}^{J} \sum_{i=1}^{I} (x_{ij} - \bar{x}_{..})^2 \\
&\geq \sum_{j=1}^{J} I(\mu_j - \bar{\mu}_.)^2 - \left(\sum_{j=1}^{J} I(\mu_j - \bar{\mu}_.)^2 \sum_{j=1}^{J} I(\bar{x}_{.j} - \bar{x}_{..})^2 \right) / \sum_{j=1}^{J} \sum_{i=1}^{I} (x_{ij} - \bar{x}_{..})^2 \\
&= \sum_{j=1}^{J} I(\mu_j - \bar{\mu}_.)^2 \left(1 - \sum_{j=1}^{J} I(\bar{x}_{.j} - \bar{x}_{..})^2 / \sum_{j=1}^{J} \sum_{i=1}^{I} (x_{ij} - \bar{x}_{..})^2 \right) \qquad (5)
\end{aligned}$$

and γ is the non-centrality parameter of the noncentral F. (The inequality in (5) is due to the Cauchy-Schwarz theorem.)

The simulations established that for $I \geq 5$, this is well approximated by

$$\text{Prob}(\text{NCF}_{J-1,IJ-(J+1)}(\hat{\gamma}) > F^{-1}_{J-1,IJ-(J+1)}(1-\alpha))$$

where $\hat{\gamma} = \sum_{J=1}^{J} I(\mu_j - \bar{\mu}_.)^2 / (\sigma^2_Y(1-\rho^2))$ is an approximation to the non-centrality parameter of the noncentral F. (Recall that in our case $\sigma^2_{\text{anocov}} = \sigma^2_Y(1-\rho^2)$.)

A listing of the simulation program that produced the power estimates can be obtained at http://www1.fpl.fs.fed.us/ps15_powersim_code.html. A web-based simulation program that can be run on additional cases (including multi-factor cases) can be found at http://www1.fpl.fs.fed.us/pspower.html.

It can be argued that in a predictor sort situation a statistician would undoubtedly perform a blocked analysis (or an analysis of covariance using the predictor/concomitant as the covariate). However, some authors of statistical texts for non-statisticians (see, for example, some of the texts listed in Section 1) appear to treat "matched subject" allocations as good experimental practice regardless of the subsequent analyses. (For example, one of the texts discussed matching, t tests, and ANOVAs, but not paired t-tests, blocked ANOVAS, or analyses of covariance.) Given that very poor power can result if a predictor sort allocation is analyzed via an unblocked ANOVA, authors of (at least) statistical texts for non-statisticians need to make this clear. This is especially true for fields in which concomitants might be highly correlated with responses.

3. Confidence Intervals on Means

Verrill (1999) established the following theorem.

Theorem 2

Assume that the predictor variable and the variable of interest, Y, have a joint bivariate normal distribution with correlation ρ. Denote the variance of Y by σ^2_Y. Suppose that there are I blocks and F factors with $K_1, \ldots, K_F$ levels. Let the allocation of samples be as described in Section 1. (For a multiple factor case, enough adjacent experimental units would be chosen at a time to provide one additional observation for each cell.) Let $\bar{Y}_{\cdot j_1 \cdot \ldots \cdot}$ be the standard estimate of the mean response for the j_1th level of factor 1. Then

$$\sqrt{I \times K_2 \times \ldots \times K_F} \left(\bar{Y}_{\cdot j_1 \cdot \ldots \cdot} - \text{E}\left(\bar{Y}_{\cdot j_1 \cdot \ldots \cdot} \right) \right) \xrightarrow{D} \text{N}\left(0, \sigma^2_Y(1 - \rho^2 + \rho^2/K_1)\right)$$

as $I \to \infty$. The analogous results hold for factors $2, \ldots, F$.

Why can this result lead to problems?

If the predictor sort nature of an experiment is neglected, then the confidence interval that is constructed for the mean response associated with level j_1 of factor 1 is

$$\bar{y}_{\cdot j_1 \cdot \ldots \cdot} \pm t \times s / \sqrt{I \times K_2 \times \ldots \times K_F} \tag{6}$$

where t is the appropriate critical value, and s is the root mean residual sum of squares from the ANOVA. Verrill (1993) established (see the appendix of the 1993 paper, or Appendix G of Verrill *et al.* (2015)) that in a predictor sort case, if the problem is treated as a $K_1 \times \ldots \times K_F$ ANOVA with I replicates per cell, the mean residual sum of squares, $s^2_{\text{unblocked}}$, satisfies

$$s^2_{\text{unblocked}} \xrightarrow{p} \sigma^2_Y \tag{7}$$

as I increases to infinity. If the problem is treated as as one involving I blocks with 1 replicate per cell, the mean residual sum of squares, s^2_{blocked}, satisfies

$$s^2_{\text{blocked}} \xrightarrow{p} (1-\rho^2)\sigma^2_Y \tag{8}$$

where ρ is the correlation between the predictor used in the sort and Y.

But by Theorem 2 above, the appropriate large sample value for s in (6) is

$$\sigma_Y \sqrt{1-\rho^2+\rho^2/K_1}$$

rather than σ_Y or $\sigma_Y\sqrt{1-\rho^2}$. This discrepancy is the source of the coverage problems.

Let

$$R_{\text{ub}}(\rho, J) \equiv 1/\left((1-\rho^2+\rho^2/J)\right)^{1/2}$$

and

$$R_{\text{b}}(\rho, J) \equiv \left((1-\rho^2)/(1-\rho^2+\rho^2/J)\right)^{1/2}.$$

(Notice that we are switching from the "K" treatments notation of the 1999 paper to a "J" treatments notation here.)

In figure 33 of Verrill *et al.* (2015) values of $R_{\text{ub}}(\rho, J)$ are plotted. These R values approximate the factor by which confidence interval sizes are incorrectly inflated when a standard unblocked ANOVA is performed in a predictor sort case.

In figure 34 of Verrill *et al.* (2015) values of $R_{\text{b}}(\rho, J)$ are plotted. These values approximate the factor by which confidence interval sizes are incorrectly deflated when a standard blocked ANOVA is performed in a predictor sort case.

In figure 35 of Verrill *et al.* (2015) values of

$$2 \times \Phi\left(\Phi^{-1}(.975) \times R_{\text{ub}}(\rho, J)\right) - 1$$

are plotted where Φ denotes the cumulative distribution function of a N(0,1). These values approximate the actual confidence levels that are associated with nominal 95% confidence intervals in the unblocked case.

Finally, in figure 36 of Verrill *et al.* (2015), reproduced here as Figure 2, values of

$$2 \times \Phi\left(\Phi^{-1}(.975) \times R_{\text{b}}(\rho, J)\right) - 1$$

are plotted. These values approximate the actual confidence levels that are associated with nominal 95% confidence intervals in the blocked case.

From these plots it is clear that, given a predictor sort design, for higher ρ values, the confidence interval lengths and coverages produced by standard ANOVA analyses are unacceptable.

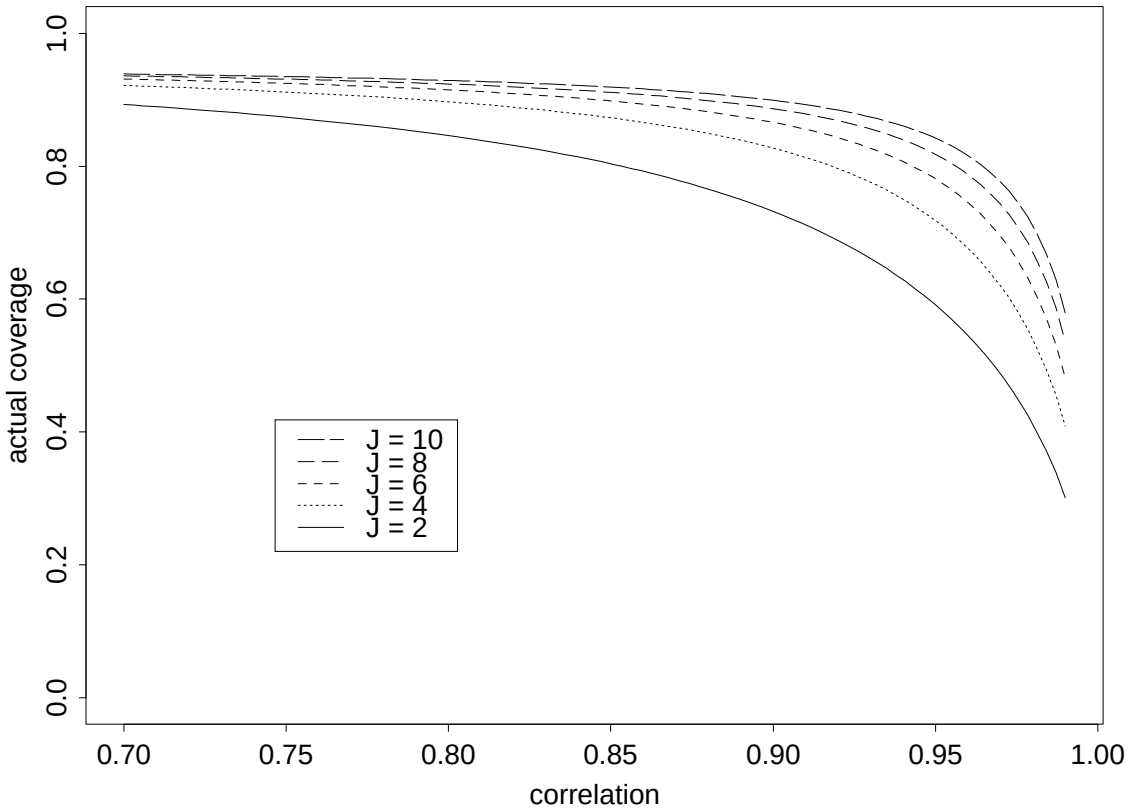

Figure 2: Blocked ANOVA. Actual coverage of a nominal 95% confidence interval.

Verrill (1999) suggested two possible fixes to incorrect confidence intervals on the μ_j's (1-factor treatment means) in the predictor sort case. First, he noted that the s in (6) could be "corrected" by multiplying it by an estimate of $\sqrt{1-\rho^2+\rho^2/J}$ in the unblocked case or by an estimate of $\sqrt{1-\rho^2+\rho^2/J}/\sqrt{1-\rho^2}$ in the blocked case. He then performed simulations that indicated the number of replications that would be needed to ensure that these asymptotically correct adjustments would yield good confidence interval coverages. These numbers were reported in his tables 1 and 2. These tables indicated that for larger ρ's, fairly large sample sizes would be needed to ensure good μ_j confidence interval coverages. This problem appears to be driven by the sensitivity of the corrections to $\hat{\rho}$.

We have since realized that it is possible to avoid this problem by making use of results (7) and (8). Together, they imply that

$$s^2_{\text{unblocked}} - s^2_{\text{blocked}} \xrightarrow{p} \rho^2\sigma^2_Y$$

and

$$s^2_{\text{blocked}} + \left(s^2_{\text{unblocked}} - s^2_{\text{blocked}}\right)/J \xrightarrow{p} (1-\rho^2+\rho^2/J)\sigma^2_Y$$

so, in the 1-factor case, we can take the corrected "anova z" confidence interval on μ_j to be

$$\bar{y}_{\cdot j} \pm z\left(\sqrt{s^2_{\text{blocked}} + \left(s^2_{\text{unblocked}} - s^2_{\text{blocked}}\right)/J}\right)/\sqrt{I} \tag{9}$$

where $z = \Phi^{-1}(1-\alpha/2)$ for a $1-\alpha$ confidence level, and the corrected "anova t" confi-

dence interval on μ_j to be

$$\bar{y}_{\cdot j} \pm t\left(\sqrt{s^2_{\text{blocked}} + \left(s^2_{\text{unblocked}} - s^2_{\text{blocked}}\right)/J}\right)/\sqrt{I} \tag{10}$$

where $t = T^{-1}_{J(I-1)}(1-\alpha/2)$ for a $1-\alpha$ confidence level, and $T_{J(I-1)}$ denotes the cumulative distribution function of a t distribution with $J(I-1)$ degrees of freedom (this is an ad hoc choice for degrees of freedom).

The second solution that was blithely and incorrectly proposed by Verrill (1999) was an analysis of covariance. Recall (equation (2)) that in a 1-factor predictor sort case, we have

$$Y_{ij} = \mu_j + \sigma_Y\left(\rho\left(X_{k(i,j),n} - \mu_X\right)/\sigma_X + \sqrt{1-\rho^2}P_{ij}\right)$$

or

$$\begin{aligned} Y_{ij} &= \mu_j - \rho\,\sigma_Y\,\mu_X/\sigma_X + \rho\,\sigma_Y\,X_{k(i,j),n}/\sigma_X + \sqrt{1-\rho^2}\,\sigma_Y\,P_{ij} \\ &= a_j + b\,X_{k(i,j),n} + \sqrt{1-\rho^2}\,\sigma_Y\,P_{ij} \end{aligned} \tag{11}$$

where

$$a_j = \mu_j - \rho\,\sigma_Y\,\mu_X/\sigma_X$$

and

$$b = \rho\,\sigma_Y/\sigma_X$$

Now

$$a_j + b\,\bar{x}_{\cdot\cdot} = \mu_j - \rho(\sigma_Y/\sigma_X)\mu_X + \rho(\sigma_Y/\sigma_X)\bar{x}_{\cdot\cdot}$$

is an approximation to μ_j and, *given* the x's and model (11) , it is well known that $\hat{a}_j + \hat{b}\bar{x}_{\cdot\cdot}$ has variance

$$\left(1/I + (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2 / \left(\sum_{k=1}^{J}\sum_{i=1}^{I}(x_{ik} - \bar{x}_{\cdot k})^2\right)\right)(1-\rho^2)\sigma_Y^2 \tag{12}$$

For large I this is of the order

$$(1-\rho^2)\sigma_Y^2/I$$

But as noted above, this is an underestimate of the variance of the estimator. This is essentially due to the fact that we are treating $\bar{x}_{\cdot\cdot}$ as a constant when, instead,

$$\begin{aligned} \text{Var}(a_j + b \times \bar{x}_{\cdot\cdot}) &= b^2\sigma_X^2/(IJ) \\ &= \rho^2\sigma_Y^2/\sigma_X^2 \times \sigma_X^2/(IJ) \\ &= \sigma_Y^2\,\rho^2/(IJ) \end{aligned}$$

In Appendix C of Verrill *et al.* (2015), we show that under a standard (non predictor sort) allocation,

$$\hat{a}_j + \hat{b}\bar{x}_{\cdot\cdot} = \bar{y}_{\cdot j} - \hat{b}(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) = \bar{y}_{\cdot j} - \hat{\rho}(\hat{\sigma}_Y/\hat{\sigma}_X)(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) = \hat{\mu}_j \tag{13}$$

where $\hat{a}_j$ and $\hat{b}$ are the standard analyis of covariance estimators, and $\hat{\rho}$, $\hat{\sigma}_Y$, $\hat{\sigma}_X$, and $\hat{\mu}_j$ are maximum likelihood estimators. In Appendix D of Verrill *et al.* (2015), we show that under maximum likelihood regularity conditions (which we do not verify at this point)

$$\sqrt{I}(\hat{\mu}_j - \mu_j) \xrightarrow{\text{D}} \text{N}(0, \sigma_Y^2(1-\rho^2+\rho^2/J)) \tag{14}$$

From results (13) and (14) we have, under a standard (non predictor sort) allocation,

$$\sqrt{I}(\bar{y}_{\cdot j} - \hat{b}(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) - \mu_j) \xrightarrow{\mathrm{D}} \mathrm{N}(0, \sigma_Y^2(1 - \rho^2 + \rho^2/J)) \tag{15}$$

In Appendix E of Verrill *et al.* (2015), we show that under a predictor sort allocation, result (15) continues to hold.

Thus, in the 1-factor case, we can take the corrected "anocov z" confidence interval on μ_j to be

$$\bar{y}_{\cdot j} - \hat{b}(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) \pm z \left(\sqrt{s^2_{\text{blocked}} + \left(s^2_{\text{unblocked}} - s^2_{\text{blocked}}\right)/J} \right) / \sqrt{I} \tag{16}$$

where $z = \Phi^{-1}(1 - \alpha/2)$ for a $1 - \alpha$ confidence level, and the corrected "anocov t" confidence interval on μ_j to be

$$\bar{y}_{\cdot j} - \hat{b}(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) \pm t \left(\sqrt{s^2_{\text{blocked}} + \left(s^2_{\text{unblocked}} - s^2_{\text{blocked}}\right)/J} \right) / \sqrt{I} \tag{17}$$

where $t = T^{-1}_{JI-(J+1)}(1 - \alpha/2)$ for a $1 - \alpha$ confidence level (again, the degrees of freedom are ad hoc).

We have performed simulations on 1-factor predictor sort anovas and anocovs to evaluate the resulting confidence interval coverages. For all combinations of X,Y correlations 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99; number of factor levels, J, equal to 2, 3, 5, 7, 9, 11, and 20; sample sizes, I, equal to 3, 5, 10, 20, 40, and 80; and confidence levels 90% and 95%, we performed 10,000 trials. For each of the resulting data sets we calculated standard unblocked and blocked anova confidence intervals on μ_1, corrected z and t anova confidence intervals on μ_1, a "standard" anocov confidence interval on μ_1, corrected z and t anocov confidence intervals on μ_1, and the maximum likelihood confidence interval on μ_1.

In accord with the reasoning associated with (12), we calculated the "standard" anocov confidence interval as

$$\bar{y}_{\cdot j} - \hat{b}(\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot}) \pm t s_{\text{anocov}} \sqrt{1/I + (\bar{x}_{\cdot j} - \bar{x}_{\cdot\cdot})^2 / \left(\sum_{k=1}^{J} \sum_{i=1}^{I} (x_{ik} - \bar{x}_{\cdot k})^2 \right)}$$

The results of the 95% confidence interval simulations are presented in table 2 of Verrill *et al.* (2015). The results of both the 90% and 95% confidence interval simulations are available at
http://www1.fpl.fs.fed.us/ps15_table2.html.

For the 95% confidence interval simulations, we fit the model

$$\text{coverage} - .95 = c_1/I^{1/2} + c_2/I + c_3/I^{3/2}$$

to the tabled coverages, and then used the resulting fits to estimate the I's at which the actual coverages would first fall between .94 and .96. As an illustration, the data and fits for the $J = 3$, $\rho = 0.8$ case are plotted in figures 37 (anova) and 38 (anocov) of Verrill *et al.* (2015). The estimated needed I's are provided in table 3 of Verrill *et al.* (2015). For blocked anovas, these I's are much improved over the comparable values reported in Verrill's (1999) table 2. For unblocked anovas and $J \geq 5$, these I's are much improved over the comparable values reported in Verrill's (1999) table 1.

It is clear from the 95% and 90% confidence interval simulation tables that

1. The confidence interval coverage simulation results are in accord with the large sample results expressed in figures 35 and 36 of Verrill *et al.* (2015). That is, for higher ρs, an uncorrected 1-factor/unblocked ANOVA will lead to confidence interval coverages that are larger than the nominal coverages and a 1-factor/blocked ANOVA will lead to coverages that are lower than nominal coverages. Also, as expected from the discussion in connection with result (12), uncorrected anocov analyses will lead to actual coverages smaller than nominal coverages.

2. Corrected anova's and anocov's yield good coverages for reasonable sample sizes.

3. For $\rho \leq 0.80$, good coverages are obtained most quickly/for the smallest samples sizes by taking a "uses t" approach. That is, we use the appropriate t critical value rather than the appropriate z critical value. For $\rho \geq 0.90$, corrected anova's yield correct coverages most quickly if a "uses t" aproach is taken for $J = 2$ and a "uses z" approach is taken otherwise. For $\rho \geq 0.90$, corrected anocov's yield correct coverages most quickly if a "uses t" aproach is taken for $J = 2, 3$ and a "uses z" approach is taken otherwise.

4. The maximum likelihood actual coverage is slow to converge to the nominal coverage.

A listing of the simulation program that produced the table 2 coverage estimates in Verrill *et al.* (2015) can be obtained at http://www1.fpl.fs.fed.us/ps15_confsim_code.html. A web-based simulation program that can be run on additional cases (including multi-factor cases) can be found at http://www1.fpl.fs.fed.us/psconf.html.

4. Scheffé and Tukey multiple comparison procedures after a predictor sort allocation

As one would expect given the hypothesis test results established in Verrill (1993), for large sample sizes, suitably altered versions of the Scheffé and Tukey multiple comparison procedures are valid after a predictor sort allocation. In this section, for the Tukey case, we describe the alterations and establish the needed asymptotic results. The Scheffé case is considered in section 5 of Verrill *et al.* (2015).

We first introduce the notation that we will use in this section. Assume that we have F factors, K_j levels for the jth factor, and I blocks (formed by specimens with adjacent [randomized within a block] order statistics of the predictor). Let $n \equiv IK_1 \ldots K_F$, and $\{X_i, i = 1, \ldots, n\}$, $\{Z_i, i = 1, \ldots, n\}$ be i.i.d. N(0,1) random variables. Define $W_i \equiv \sigma_Y \left(\rho X_i + \sqrt{1-\rho^2}\, Z_i\right)$. Then the W_i's are i.i.d. $N(0, \sigma^2)$ and

$$\text{corr}(X_i, W_j) = \begin{cases} \rho & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

We model predictor sort allocation by ordering the X's (the predictors), and randomly dividing the W's that correspond to $X_{(i-1)K_1 \ldots K_F+1,n}, \ldots, X_{iK_1 \ldots K_F,n}$ among the $K_1 \times K_2 \times \ldots \times K_F$ treatments. (Here, $X_{l,n}$ is the lth order statistic among the X's.)

Let $Y_{ij_1 \ldots J_F}$ denote $\mu_{j_1 \cdot \ldots \cdot} + \ldots + \mu_{\cdot \ldots \cdot j_F}$ plus the ith W that is assigned to treatment $j_1 \ldots j_F$, where, for example, $\mu_{j_1 \cdot \ldots \cdot}$ denotes the effect associated with the j_1th level of factor 1, and $\mu_{\cdot \ldots \cdot j_F}$ denotes the effect associated with the j_Fth level of factor F. Then

$$Y_{ij_1 \ldots j_F} = \mu_{j_1 \cdot \ldots \cdot} + \ldots + \mu_{\cdot \ldots \cdot j_F} + \sigma_Y \left(\rho X_{k(ij_1 \ldots j_F),n} + \sqrt{1-\rho^2}\, P_{ij_1 \ldots j_F}\right) \qquad (18)$$

where $k(ij_1 \ldots j_F) \in \{(i-1)K_1 \ldots K_F + 1, \ldots, iK_1 \ldots K_F\}$, and the $P_{ij_1 \ldots j_F}$ are i.i.d. N(0,1) and are independent of the X's.

(Note that there is some ugly notation here. The "$ij_1 \ldots j_F$" in $k(ij_1 \ldots j_F)$ is not a product and $k(ij_1 \ldots j_F)$ should really be written as $k(i, j_1, \ldots, j_F)$, but for simplicity, we omit the commas. On the other hand, "$iK_1 \ldots K_F$" actually is a product.)

4.1 Tukey's multiple comparison test/procedure

Theorem 4

Assume that the predictor variable and the variable of interest, Y, have a joint bivariate normal distribution with correlation ρ. Denote the variance of Y by σ_Y^2. Suppose that there are I blocks and F factors with $K_1, \ldots, K_F$ levels. Let the allocation of samples be as described in Section 1. (For a multiple factor case, enough adjacent experimental units would be chosen at a time to provide one additional observation for each cell.) Let $\bar{Y}_{\cdot j_1 \cdot \ldots \cdot}$ be the standard estimate of the mean response for the j_1th level of factor 1. For comparisons of the factor 1 levels, let the numerator of the test statistic be given by

$$Q_I \equiv \max_{l_1, l_2 \in \{1, \ldots, K_1\}} \sqrt{IK_2 \ldots K_F} |\bar{Y}_{\cdot l_1 \cdot \ldots \cdot} - \bar{Y}_{\cdot l_2 \cdot \ldots \cdot}|$$

Let s_{ub}^2 denote the estimate of σ^2 in the unblocked case. That is,

$$s_{\mathrm{ub}}^2 \equiv \mathrm{SS}_{\mathrm{den,unbl}} / (IK_1 \ldots K_F - (K_1 + K_2 - 1 + \ldots + K_F - 1)) \tag{19}$$

where

$$\mathrm{SS}_{\mathrm{den,unbl}} = \sum_{i=1}^{I} \sum_{j_1=1}^{K_1} \ldots \sum_{j_F=1}^{K_F} (y_{ij_1 \ldots j_F} - (\bar{y}_{\cdot \ldots \cdot} + (\bar{y}_{\cdot j_1 \cdot \ldots \cdot} - \bar{y}_{\cdot \ldots \cdot}) + \ldots + (\bar{y}_{\cdot \ldots \cdot j_F} - \bar{y}_{\cdot \ldots \cdot})))^2$$

Let s_{b}^2 denote the estimate of σ^2 in the blocked case. That is,

$$s_{\mathrm{b}}^2 \equiv \mathrm{SS}_{\mathrm{den,bl}} / (IK_1 \ldots K_F - (I + K_1 - 1 + \ldots + K_F - 1)) \tag{20}$$

where

$$\mathrm{SS}_{\mathrm{den,bl}} = \sum_{i=1}^{I} \sum_{j_1=1}^{K_1} \ldots \sum_{j_F=1}^{K_F} (y_{ij_1 \ldots j_F} - (\bar{y}_{\cdot \ldots \cdot} + (\bar{y}_{i \cdot \ldots \cdot} - \bar{y}_{\cdot \ldots \cdot}) + \ldots + (\bar{y}_{\cdot \ldots \cdot j_F} - \bar{y}_{\cdot \ldots \cdot})))^2$$

Let $F_{\mathrm{R(K_1)}}$ denote the distribution of the range of a sample of K_1 independent N(0,1)'s. Then, under the null hypothesis that $\mu_{1 \cdot \ldots \cdot} = \ldots = \mu_{K_1 \cdot \ldots \cdot}$,

$$Q_I / \left(s_{\mathrm{ub}} \sqrt{1 - \rho^2} \right) \xrightarrow{\mathrm{D}} F_{\mathrm{R(K_1)}} \tag{21}$$

and

$$Q_I / s_{\mathrm{b}} \xrightarrow{\mathrm{D}} F_{\mathrm{R(K_1)}} \tag{22}$$

Similar results hold for factors 2 through F.

Proof

The proof appears in Appendix F of Verrill *et al.* (2015).

We have performed simulations on 1-factor predictor sort anovas to evaluate the resulting sizes of Tukey tests. For all combinations of X,Y correlations 0.0, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99, number of treatments, J, equal to 3, 5, 7, 9, 11, and 20, and sample

sizes, I, equal to 3, 5, 10, 20, and 40, we performed 100,000 trials. For each of the resulting data sets we calculated a standard unblocked Tukey test statistic on the equivalence of the J treatment means, an unblocked Tukey test statistic that has been corrected by an estimated $\sqrt{1-\rho^2}$ factor, an unblocked Tukey test statistic that has been corrected by the true $\sqrt{1-\rho^2}$ factor, and a blocked Tukey test statistic. In table 4 of Verrill *et al.* (2015), we report the resulting actual test sizes when the nominal size is 0.05.

We can conclude from this table that

1. For the non-zero ρ's considered, a standard unblocked Tukey test (one that uses an s_{ub} denominator with no ρ correction) yields test sizes that can be much less than the nominal test size.

2. For lower I, an unblocked Tukey test that has been corrected via an estimated ρ yields test sizes that can be much more than the nominal test size.

3. For lower I, and $\rho = 0.5, 0.6$, an unblocked Tukey test that has been corrected via the true ρ yields test sizes that can be lower than the nominal test size. For lower I, and $\rho = 0.8, 0.9, 0.95, 0.99$, an unblocked Tukey test that has been corrected via the true ρ yields test sizes that can be much higher than the nominal test size.

4. In general, a blocked Tukey test (one that uses an s_{b} denominator) yields actual test sizes that closely match the nominal 0.05 test size. The blocked Tukey test does perform somewhat poorly in the $\rho = 0.95, .99$ cases.

A listing of the simulation program that produced the size estimates can be found at http://www1.fpl.fs.fed.us/ps15_tukey_size_sim_code.html.

Simultaneous confidence intervals based on Tukey's multiple comparison test are described in section 5.4 of Verrill *et al.* (2015).

5. Web Programs/R Programs

Forest Products Laboratory scientists have produced predictor sort web programs that help researchers

1. choose sample sizes (perform power calculations) for predictor sort hypothesis tests

2. allocate specimens via a predictor sort

3. perform hypothesis tests for a simple 1 factor, 2 levels predictor sort experiment

4. perform simulations to estimate the coverage of predictor sort confidence intervals on treatment means

These programs can be accessed at http://www1.fpl.fs.fed.us/predsort.html. This web page also contains a link to R code that helps users calculate predictor sort confidence intervals on treatment means.

We have also developed interactive Java code that permits a user to obtain small sample confidence intervals on quantiles in the predictor sort case. This work will appear in a separate Forest Products Laboratory technical report.

6. Summary

We have reminded readers that, properly designed and analyzed, predictor sort experiments (experiments in which the predictor variable is used to form blocks) permit scientists to achieve considerable increases in statistical power and/or reductions in sample sizes (and thus reductions in experimental costs). For analysis of variance tests of hypotheses, sample sizes can be reduced from roughly n to $(1-\rho^2)n$ where ρ is the correlation between the predictor/concomitant and the variable of interest. For confidence intervals on quantiles, approximate sample size reductions are illustrated in figure 40 of Verrill *et al.* (2015). They can amount to 30% for ρ equal to 0.70, and increase as ρ increases.

Our studies also indicate that the $1-\rho^2$ factor is only approximate, especially for blocked anovas (as opposed to analyses of covariance). Thus, we have provided a web based simulation program that yields estimates of actual powers (in addition to estimates based on large sample theory).

We have demonstrated that if a scientist performs a predictor sort allocation, but then analyzes the experiment as an unblocked analysis of variance, their experiment can have extremely low statistical power (an inability to detect actual differences). This amounts to a serious scientific blunder.

We have demonstrated theoretically that given a predictor sort allocation, unmodified analyses of variance (blocked or unblocked) and analyses of covariance yield incorrect confidence intervals on treatment means. (The confidence intervals are too wide in the unblocked anova case and too narrow in the blocked anova and analysis of covariance cases.) We have provided a web based simulation program that estimates the coverages of incorrect (unmodified) anova confidence intervals on treatment means, and the coverages of corrected anova and anocov confidence intervals. We have also provided an R function that helps a scientist calculate corrected confidence intervals on treatment means estimated from a predictor sort experiment.

Finally, we have developed Scheffé and Tukey multiple comparison tests and associated simultaneous confidence intervals that are appropriate in the predictor sort case.

All of our results have been established under an assumption of a joint bivariate normal relationship between the predictor and the response.

As noted earlier, it can be argued that in a predictor sort situation a professional statistician would undoubtedly perform a blocked analysis or an analysis of covariance using the predictor/concomitant as the covariate, and thus, we need not exercise special care in identifying, designing, and analyzing predictor sort experiments. We have a five-fold response. First, identifying a design as a predictor-sort design permits a scientist to perform correct power calculations. Second, blocked anova hypothesis tests can perform poorly as ρ becomes sufficiently large. Third, although unmodified blocked anovas (for lower ρ's) and analyses of covariance yield essentially correct hypothesis tests, they yield incorrect confidence intervals on treatment means. Fourth, blocked anovas and analyses of covariance do not help us in the quantile estimation case. Finally, as noted at the end of Section 2, we have seen introductory texts that treat predictor sort allocation as a good experimental practice independent of the method of analysis. (For example, one of the texts we sampled discussed matching, t tests, and ANOVAs, but not paired t-tests, blocked ANOVAS, or analyses of covariance.) Given the large decrease in power (especially for larger ρ's) that can occur if a predictor sort experiment is analyzed via an unblocked anova, this pedagogy must be corrected.

REFERENCES

ASTM. (2003), Standard Practice for Establishing Allowable Properties for Visually-Graded Dimension Lumber from In-Grade Tests of Full-Size Specimens. ASTM D1990-00. *Annual Book of ASTM Standards*, Volume 04.10, American Society for Testing and Materials, West Conshohocken, Pennsylvania.

ASTM. (2003), Standard Practice for Evaluating Allowable Properties for Grades of Structural Lumber. ASTM D2915-03. *Annual Book of ASTM Standards*, Volume 04.10, American Society for Testing and Materials, West Conshohocken, Pennsylvania.

Casella, G. (2008), *Statistical Design*, Berlin: Springer.

Cox, D.R. (1958), *Planning of Experiments*, New York: John Wiley.

Cozby, P. and Bates, S. (2011), *Methods in Behavioral Research, 11th Edition*, New York: McGraw-Hill.

David, H.A. (1981), *Order Statistics*, New York: John Wiley & Sons.

David, H.A. and Gunnink, J.L. (1997), "The Paired *t* Test Under Artificial Pairing," *The American Statistician*, 51, 9–12.

Finney, D.J. (1972), *An Introduction to Statistical Science in Agriculture*, New York: John Wiley.

Gibbons, R.D. (1994), *Statistical Methods for Groundwater Monitoring*, New York: John Wiley and Sons.

Guttman, I. (1970), *Statistical Tolerance Regions: Classical and Bayesian*, Hafner Publishing Company, Darien, Connecticut.

Kerlinger, F.N. and Lee, H.B. (1999), *Foundations of Behavioral Research, 4th Edition*, Cengage Learning: Boston.

Kirk, R.E. (1968), *Experimental Design: Procedures for the Behavioral Sciences*, Belmont, CA: Brooks/Cole.

Kirk, R.E. (2013), *Experimental Design: Procedures for the Behavioral Sciences*, 4th Edition, Los Angeles, CA: SAGE Publications.

Myers, J.L. (1979), *Fundamentals of Experimental Design*, Boston: Allyn and Bacon.

Lehmann, E.L. (1983), *Theory of Point Estimation*, New York: John Wiley & Sons.

Michigan DEQ (1994), *Guidance Document, Verification of Soil Remediation*, Hazardous Waste Program Section, Waste Management Division, Michigan Department of Environmental Quality, Lansing, Michigan.

MIL-HDBK-17-1 (2003), *Guidelines for Characterization of Structural Materials*, Department of Defense Single Stock Point, Philadelphia, Pennsylvania, http://www.dodssp.daps.mil/.

Ostle, B., and Mensing, R.W. (1975), *Statistics in Research*, Ames, IA: The Iowa State University Press.

Owen, D.B. (1963), *Factors for one-sided tolerance limits and for variable sampling plans*, Sandia Corporation Monograph No. SCR-607 (19th edn).

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons.

Ruxton, G.D. and Colgrave, N. (2006), *Experimental Design for the Life Sciences, 2nd Edition*, Oxford University Press.

Scheffé, H. (1959), *The Analysis of Variance*, New York: John Wiley & Sons.

Snedecor, G.W., and Cochran, W.G. (1989), *Statistical Methods*, Ames, IA: The Iowa State University Press.

Steel, R., and Torrie, J. (1960), *Principles and Procedures of Statistics*, New York: McGraw-Hill.

Toutenburg, H. (2002), *Statistical Analysis of Designed Experiments*, New York: Springer.

Tuckman, B.W., and Harper, B.E. (2012), *Conducting Educational Research, 6th Edition*, Lanham, Maryland: Rowman and Littlefield Publishers.

van Zutphen, L.F., Baumans, V., and Beynen, A.C. (2001), *Principles of Laboratory Animal Science, Revised Edition: A contribution to the humane use and care of animals and to the quality of experimental results*, Amsterdam: Elsevier.

Verrill, S.P. (1993), "Predictor Sort Sampling, Tight *T*'s, and the Analysis of Covariance," *Journal of the American Statistical Association*, 88, 119–124. http://www1.fpl.fs.fed.us/jasa1993.pdf

Verrill, S.P., and Green, D.W. (1996), " Predictor Sort Sampling, Tight t's, and the Analysis of Covariance: Theory, Tables, and Examples," Research Paper FPL-RP-558, Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. 105 pages. http://www1.fpl.fs.fed.us/fpl_rp558.pdf

Verrill, S.P. (1999), "When Good Confidence Intervals Go Bad: Predictor Sort Experiments and ANOVA," *The American Statistician*, 53(1), 38–42. http://www1.fpl.fs.fed.us/amstat1999.pdf

Verrill, S.P., Herian, V.L., and Green, D.W. (2004), "Predictor Sort Sampling and Confidence Bounds on Quantiles I: Asymptotic Theory," Research Paper FPL-RP-623, Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. 67 pages. http://www1.fpl.fs.fed.us/fpl_rp623.pdf

Verrill, S.P. and Kretschmann, D.E. (2015), "A Reminder about Potentially Serious Problems with a Type of Blocked ANOVA Analysis," draft research paper, Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. http://www1.fpl.fs.fed.us/ps15.pdf

Warren, W.G., and Madsen, B. (1977), "Computer-Assisted Experimental Design in Forest Products Research: A Case Study Based on Testing the Duration-of-Load Effect," *Forest Products Journal*, 27, 45–50.

Verrill, Kretschmann, (2015) “Problems with a Type of Matched Subjects Design and Analysis,” In: Proceedings of the 2015 American Statistical Association Joint Statistical Meetings, Section on Physical and Engineering Sciences, pages 3491-3505; 2015