When Good Confidence Intervals Go Bad: Predictor Sort Experiments and ANOVA

Steve VERRILL

A predictor sort experiment is one in which experimental units are allocated on the basis of the values of a predictor variable that is correlated with the response. Standard ANOVA analyses of predictor sort experiments can lead to confidence intervals whose actual coverages are poor matches to nominal coverages. Correct coverages can be obtained by adjusting confidence interval lengths by appropriate factors, or by performing analyses of covariance.

KEY WORDS: Analysis of covariance; Analysis of variance; Blocked analysis of variance: Concomitant variable; Predictor sort sampling.

1. INTRODUCTION

This article identifies errors that are likely to be made in confidence interval calculations in predictor sort ANOVA experiments. Predictor sort sampling in the context of hypothesis testing is discussed in Verrill (1993).

To perform a predictor sort, it is necessary to find a predictor that can be measured prior to the start of an experiment and is well correlated with the response being investigated. Experimental units are then sorted and allocated on the basis of this predictor. For example, in a one-way predictor sort ANOVA that compares K treatments, the specimens associated with the top K predictor values are randomly assigned to the K treatments, then the specimens associated with the next largest K predictor values are randomly assigned to the K treatments, and so on. If there are I such groups of specimens, this allocation process yields a two-way ANOVA in which each of the I blocks is composed of specimens with similar predictor values. In agricultural experimentation, typical predictors are weight and age in the case of animal subjects. Past plot yields can be used to form blocks in the case of field studies. In the behavioral sciences, predictors such as IQ, hours of training, or performance on a pre-test have been used to form blocks.

Predictor sort experiments are discussed by Cox (1958, example 3.3); Steel and Torrie (1960, sec. 8.2); Kirk (1968, sec. 5.1); Finney (1972, sec. 13.17); Ostle and Mensing (1975, example 11.3); Myers (1979, chap. 6); and Snedecor and Cochran (1989, example 6.13.1).

Steve Verrill is Mathematical Statistician, U.S. Department of Agriculture Forest Products Laboratory, Madison, WI 53705 (E-mail: steve@ws10.fpl.fs.fed.us).

In general, if an experiment could have been analyzed as an analysis of covariance, but, instead, the values of the covariate are used to define blocks, the experiment is based on predictor sort sampling.

Predictor sort experiments are quite common in wood strength research. For example, a scientist might be interested in the effects of fire retardants on wood strength. Typically the scientist obtains a random sample of lumber from a particular species of wood. Then the lumber is sorted by modulus of elasticity (MOE) which can be measured non-destructively and is known to be well correlated with modulus of rupture. If there are K fire retardants to be compared, the lumber specimens with the K largest MOE values are randomly assigned to the K retardant treatments, then the K specimens with the next largest MOE values are randomly assigned to the K treatments, and so on. Wood scientists motivate this procedure by stating that it makes pre-treatment strength distributions "reasonably equivalent" among the K groups of test specimens. Depending upon their level of statistical sophistication the scientists go on to analyze the resulting data via unblocked or blocked analyses of variance, or analyses of covariance.

As noted in Verrill (1993) (see also David and Gunnik 1997), the correlations among the order statistics of the predictor induce correlations among the responses so that the standard ANOVA assumptions are not satisfied for a predictor sort experiment. Verrill demonstrates that blocked ANOVAs are still essentially valid and that simply modified unblocked ANOVAs can also be performed on predictor sort datasets. However, one must be careful with power calculations. A program that performs predictor sort power calculations and specimen allocations can be run over the World Wide Web. See http://www1.fpl.fs.fed.us/ttweb.html.

The current article establishes that standard ANOVA analyses of predictor sort experiments can yield confidence intervals whose actual coverages are poor matches to nominal coverages. The article then discusses techniques that can correct this problem.

2. POOR CONFIDENCE INTERVAL COVERAGE

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

$$\bar{y}_{\cdot j_1 \cdot \ldots} \pm t \times s/\sqrt{I \times K_2 \times \cdots \times K_F}, \qquad (1)$$

where t is the appropriate critical value, and s is the root mean residual sum of squares from the ANOVA. Verrill (1993) established that in a predictor sort case, if the problem is treated as a $K_1 \times \cdots \times K_F$ ANOVA with I replicates per cell, the mean residual sum of squares converges in

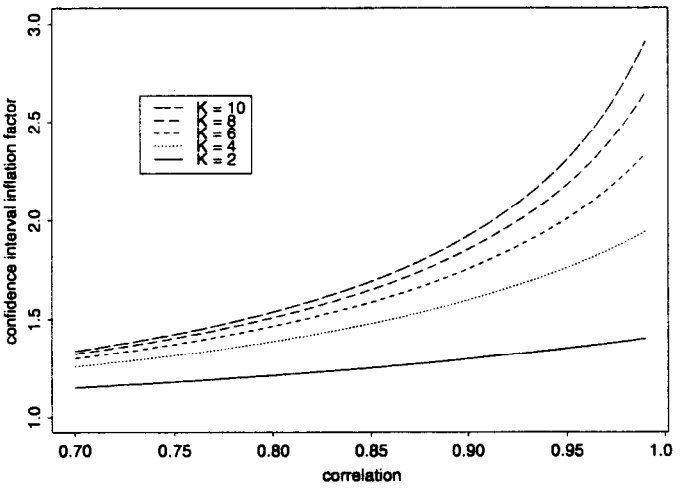

Figure 1. Unblocked ANOVA, Confidence Interval Inflation Factor.

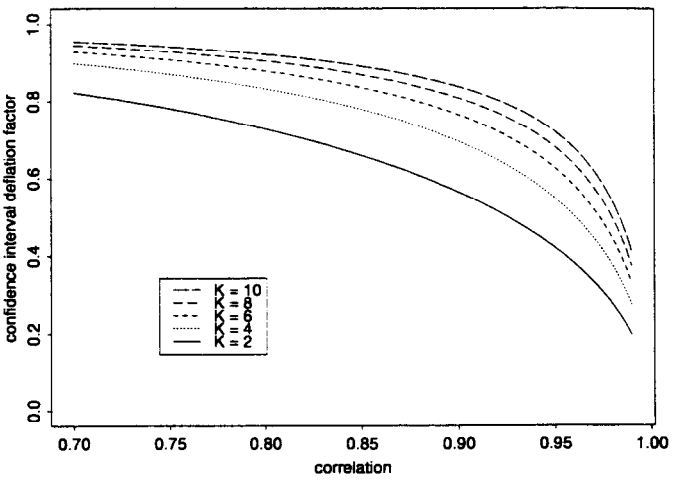

Figure 2. Blocked ANOVA, Confidence Interval Deflation Factor.

probability to σ_Y^2 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 converges in probability to $(1 - r^2)\sigma_Y^2$, where r is the correlation between the predictor used in the sort and Y.

In the theorem established in Section 4, it is shown that the appropriate large sample value for s in (1) is

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

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

Let

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

and

$$R_b(\rho, K) \equiv \left((1 - \rho^2)/(1 - \rho^2 + \rho^2/K)\right)^{1/2}.$$

In Figure 1 values of $R_{ub}(\rho, K)$ 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 2 values of $R_b(r, K)$ 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 3 values of

$$2 \times \Phi(\Phi^{-1}(.975) \times R_{ub}(r,K)) - 1$$

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

Finally, in Figure 4 values of

$$2 \times \Phi(\Phi^{-1}(.975) \times R_b(r,K)) - 1$$

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 r values, the confidence interval lengths and coverages produced by standard ANOVA analyses are unacceptable. Unfortunately, there are many situations in which the correlation between the predictor and the response can be quite high; for example, when the predictor is a measurement made on an individual before a treatment and the response is a similar measurement made on the same individual after the treatment.

3. HEURISTIC JUSTIFICATION OF THE THEOREM

To keep things relatively simple, let us focus on a one-way situation with I blocks and K treatments. We can think of a predictor sort specimen allocation in the following manner. A response value, Y, associated with a specimen is given by

$$Y = \mu_Y + \sigma_Y\left(\rho\,(X - \mu_X)/\sigma_X + \sqrt{1-\rho^2}Z\right),$$

where $(X - \mu_X)/\sigma_X$ and Z are independent N(0,1)'s. Prior to the experiment we have values for X. We rank the specimens on the basis of their associated X values and then ran-

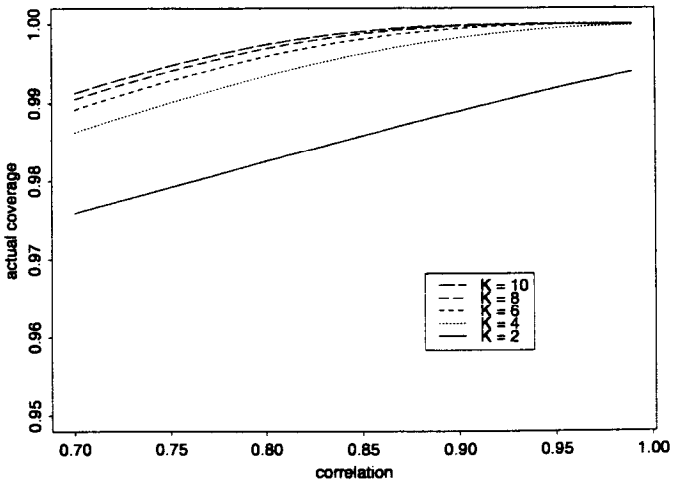

Figure 3. Unblocked ANOVA, Actual Coverage of a Nominal 95% confidence Interval.

domly allocate the top K specimens to the first block, the next K to the second block, and so on. Let $Y_{11}, \ldots, Y_{I1}$ be the responses for the specimens that receive the first treatment. We are interested in a confidence interval on $\mu_Y + \mu_1$, where μ_1 is the effect of the first treatment. We take as our estimate of this value

$$\bar{Y}_{\cdot 1} = \sum_{i=1}^{I} Y_{i1}/I = \mu_Y + \mu_1 + \sigma_Y \left(\rho \times \sum_{i=1}^{I} W_{i1}/I + \sqrt{1-\rho^2} \times \sum_{i=1}^{I} Z_{i1}/I \right),$$

where the $\mathbf{Z}$'s are iid N(0,1) and independent of the W's, and W_{i1} is randomly drawn from the ith block of $(X - \mu_X)/\sigma_X$'s. Then

$$\begin{aligned} \mathrm{var}(\bar{Y}_{\cdot 1}) &= \sigma_Y^2 \left(\rho^2 \mathrm{var} \left(\sum_{i=1}^{I} W_{i1}/I \right) + (1-\rho^2) \mathrm{var} \left(\sum_{i=1}^{I} Z_{i1}/I \right) \right) \\ &= \sigma_Y^2 \left(\rho^2 \mathrm{var} \left(\sum_{i=1}^{I} W_{i1}/I \right) + (1-\rho^2)/I \right). \end{aligned}$$

Thus, to convince ourselves that the theorem makes sense, we only need to be able to understand how

$$\mathrm{var} \left(\sum_{i=1}^{I} W_{i1}/I \right) \approx 1/IK. \tag{2}$$

We have

$$\mathrm{var} \left(\sum_{i=1}^{I} \sum_{k=1}^{K} W_{ik}/IK \right) = \mathrm{var}\,(\text{the sum of all the centered and scaled } \boldsymbol{X}\text{'s}/\boldsymbol{IK}) = 1/\boldsymbol{IK}$$

since this is just the variance of an average of IK iid N(0,1)'s. Now the claim is that, since we require one observation from each of the I blocks of adjacent X order statistics in our sum of I W's, our average of I W's is a close enough approximation to the average of all IK $(X-\mu_X)/\sigma_X$'s that result (2) holds. This intuitive closeness is established rigorously in the proof of the Theorem.

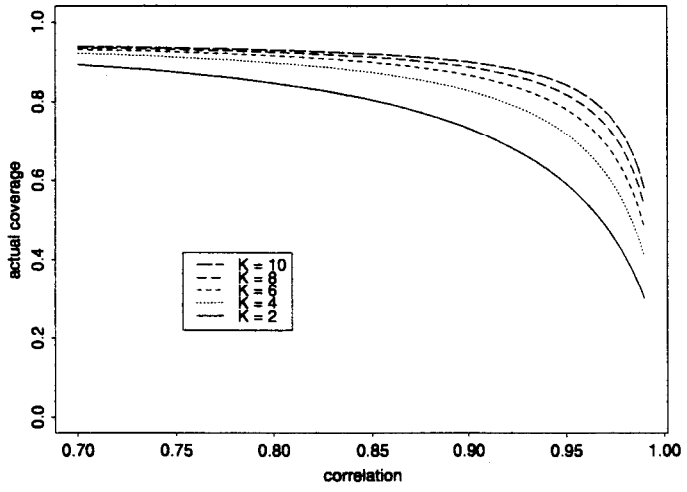

Figure 4. Blocked ANOVA, Actual Coverage of a Nominal 95% Confidence Interval.

4. THE THEOREM

Theorem 1. Assume that the predictor variable and the variable of interest, Y, have a joint bivariate normal distribution with correlation r. 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 $\bar{Y}_{\cdot j_1 \cdots \cdot}$ be the standard estimate of mean response for the j_1th level of factor 1. Then

$$\sqrt{I \times K_2 \times \cdots \times K_F} \left(\bar{Y}_{\cdot j_1 \cdots \cdot} - \mathrm{E}\left(\bar{Y}_{\cdot j_1 \cdots \cdot} \right) \right) \xrightarrow{D} \mathrm{N}\left(0, \sigma_Y^2 (1 - \rho^2 + \rho^2/K_1)\right) \tag{3}$$

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

Proof: We have

$$Y_{ij_1 \ldots j_F} = \mathrm{E}\left(Y_{ij_1 \ldots j_F}\right) + \sigma_Y \left(\rho X_{ij_1 \ldots j_F} + \sqrt{1-\rho^2} Z_{ij_1 \ldots j_F} \right),$$

where the $X_{ij_1 \ldots j_F}$'s, $j_1 \in \{1, \ldots, K_1\}$, $\ldots$, $j_F \in \{1, \ldots, K_F\}$, are a randomization of the ith group of order statistics from $I \times K_1 \times \cdots \times K_F$ iid N(0,1)'s, the $Z_{ij_1 \ldots j_F}$'s are iid N(0,1), and the X's and Z's are independent.

To establish (3), we need only show that

$$\sqrt{I \times K_2 \times \cdots \times K_F} \times \left(\sum_{i=1}^{I} \sum_{j_2=1}^{K_2} \cdots \sum_{j_F=1}^{K_F} X_{ij_1 j_2 \ldots j_F} / \left(I \times K_2 \times \cdots \times K_F\right) \right) \xrightarrow{D} \mathrm{N}(0, 1/K_1). \tag{4}$$

We have

$$\sqrt{I \times K_1 \times K_2 \times \cdots \times K_F} \times \left(\sum_{i=1}^{I} \sum_{j_1=1}^{K_1} \sum_{j_2=1}^{K_2} \cdots \sum_{j_F=1}^{K_F} X_{ij_1 j_2 \ldots j_F} / \left(I \times K_1 \times K_2 \times \cdots \times K_F\right) \right) \xrightarrow{D} \mathrm{N}(0,1)$$

so

$$\sqrt{I} \left(\sum_{i=1}^{I} \sum_{j_1=1}^{K_1} \sum_{j_2=1}^{K_2} \cdots \sum_{j_F=1}^{K_F} X_{ij_1 j_2 \ldots j_F} / \left(I \times K_1 \times K_2 \times \cdots \times K_F\right) \right) \xrightarrow{D} \mathrm{N}\left(0, 1/(K_1 \times K_2 \times \cdots \times K_F)\right). \tag{5}$$

Now by the Lemma established in the appendix of Verrill (1993),

$$\sqrt{I}\left|\sum_{i=1}^{I}\sum_{j_2=1}^{K_2}\cdots\sum_{j_F=1}^{K_F} X_{ij_1j_2\ldots j_F}/(I\times K_2\times\cdots\times K_F)\right.$$
$$\left.-\sum_{i=1}^{I}\sum_{j_1=1}^{K_1}\sum_{j_2=1}^{K_2}\cdots\sum_{j_F=1}^{K_F} X_{ij_1j_2\ldots j_F}\right.$$
$$\left./(I\times K_1\times K_2\times\cdots\times K_F)\right|$$
$$\le \sqrt{I}\sum_{i=1}^{I}\left(X_{(iK_1K_2\ldots K_F)}-X_{((i-1)K_1K_2\ldots K_F+1)}\right)/I$$
$$\le \sqrt{I}\left(X_{(IK_1K_2\ldots K_F)}-X_{(1)}\right)/I \xrightarrow{p} 0. \quad (6)$$

Here $X_{(iK_1K_2\ldots K_F)}$ is the largest value in the ith group of adjacent X order statistics, $X_{((i-1)K_1K_2\ldots K_F+1)}$ is the smallest order statistic in this group, $X_{(1)}$ is the smallest overall order statistic, and $X_{(IK_1K_2\ldots K_F)}$ is the largest. Results (5) and (6) establish (4).

5. RECOMMENDATIONS FOR PRODUCING PREDICTOR SORT CONFIDENCE INTERVALS

In the predictor sort case, there are three obvious solutions to problems of incorrect confidence interval coverages. Given the underlying relationship

$$Y=\mu_Y+\sigma_Y\left(\rho(X-\mu_X)/\sigma_X+\sqrt{1-\rho^2}Z\right), \quad (7)$$

where the predictor X is known, the best solution is to make explicit use of X in an analysis of covariance. Provided that model (7) holds, actual coverages will always equal nominal coverages in this case.

Alternatively, in the unblocked case, one could obtain an estimate of r from the data and then divide the root mean residual sum of squares by $R_{ub}(r, K)$. In the blocked case, one would divide by $R_b(r, K)$.

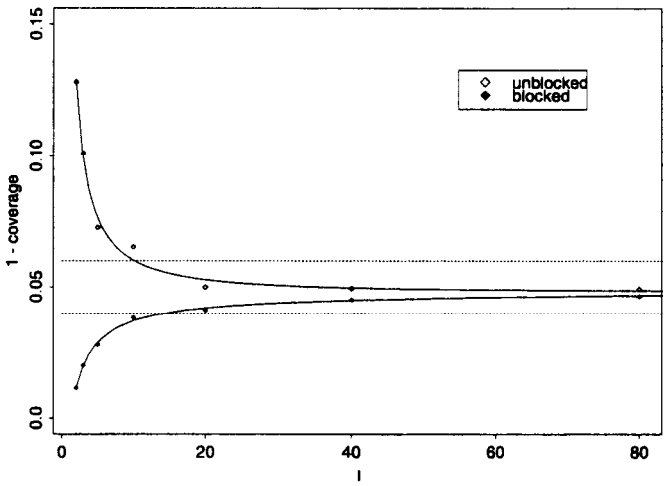

Figure 5. Simulation Results, K = 10, r = .9, Dotted Lines at 1 - coverage Equal to .04 and .06.

Simulations indicate that for I large enough any of these methods suffice. However, for small I the latter two methods can yield poor results. Tables 1 and 2 provide guidance on the number of replications needed to yield reasonable results for the latter two methods. The values in the Tables are the I values that will yield actual coverages that lie between .94 and .96 *for one-way ANOVAs.* These values were estimated from simulation runs. In these simulation runs r was estimated as the average of the cell sample correlations.

For example, to obtain the 11 value in the K = 10, r = .9 cell of Table 1, 7 simulation results were smoothed: At each of I = 2, 3, 5, 10, 20, 40, and 80, 4,000 trials were performed. This yielded the seven coverage estimates .8720, .8990, .9273, .9348, .9500, .9505, and .9507. A regression program was used to fit the model

$$\arcsin(\sqrt{1-\text{coverage}})-\arcsin(\sqrt{.05}) = c_1/I^{1/2}+c_2/I+c_3/I^{3/2}$$

and then this model was used to estimate the I at which the actual coverage first fell between .94 and .96. The data and fits for the K = 10, r = .9 case are plotted in Figure 5.

As noted previously, the values in Tables 1 and 2 are appropriate for one-way ANOVAs. For other ANOVAs they are only rough guides. The simulation program that produced the coverage estimates that were used to develop the tables can be run on additional cases over the World Wide Web at http://www1.fpl.fs.fed.us/ttconf.html. It can handle multiway ANOVAs.

Table 1. I Needed to Ensure Coverage Between .94 and .96, One-Way Unblocked ANOVAs, Dividing the Root Mean Residual Sum of Squares by $R_{ub}(r,K)$

	r				
K	.7	.8	.9	.95	.99
2	2	2	4	4	4
4	3	3	2	2	2
6	4	4	6	7	11
6	4	5	11	11	19
10	4	5	11	20	20

Table 2. I Needed to Ensure Coverage Between .94 and .96, One-Way Blocked ANOVAs, Dividing the Root Mean Residual Sum of Squares by $R_b(r,K)$

	r				
K	.7	.6	.9	.95	.99
2	12	14	29	69	>80
4	6	11	19	31	>80
6	5	11	17	35	>80
8	5	6	15	36	>80
10	3	6	15	25	>80

6. SUMMARY

It is important that statistics practictioners be able to recognize predictor sort situations. If specimens are ranked on the basis of a measured characteristic that is believed to be correlated with the response being investigated, and the specimens are placed into blocks on the basis of this rank-

ing, then the experiment needs to be treated as a predictor sort experiment rather than as a simple randomized block design.

Verrill (1993) discussed power calculations and hypothesis testing in a predictor sort context. A program that performs predictor sort power calculations and specimen allocations can be run over the World Wide Web at http://www1.fpl.fs.fed.us/ttweb.html.

Given a predictor sort experiment, if confidence intervals are of interest, then a careful analysis of covariance should be performed. Alternatively, in the unblocked case, one could obtain an estimate of r from the data and then divide the root mean residual sum of squares by $R_{ub}(r, K)$. In the blocked case, one would divide by $R_b(r, K)$. The sample sizes that are needed to justify this alternate approach in the one-way case are given in Tables 1 and 2 for a variety of r and K combinations. The nature of the coverage in a particular case can be investigated via a simulation program that can be run over the World Wide Web at http://www1.fpl.fs.fed.us/ttconf.html. This program can simulate multiway ANOVAs.

[Received March 1997. Revised January 1998.]

REFERENCES

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

David, H.A., and Gunnik, 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.

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

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

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

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.

Verrill, S.P. (1993), "Predictor Sort Sampling, Tight *T*'s, and the Analysis of Covariance," *Journal of the American Statistical Association*, 88, 119–124.