# USING PRIOR KNOWLEDGE OF THE INTRACLASS CORRELATION TO INCREASE THE POWER OF TESTS FOR TREATMENT MEANS

Marjorie E. Bond

James J. Higgins

### Recommended Citation

# USING PRIOR KNOWLEDGE OF THE INTRACLASS CORRELATION TO INCREASE THE POWER OF TESTS FOR TREATMENT MEANS

Marjorie E. Bond, Monmouth College
James J. Higgins, Kansas State University

## Abstract

It is common in agricultural research to have experimental units that consist of multiple observational units. For instance, treatments may be applied to pens of animals, pens being the experimental units, while weights are measured on individual animals, the observational units. If there are a small number of experimental units, the power of statistical tests for treatment effects can be small regardless of the number of observational units. We show that it is possible to increase the power of such statistical tests by taking advantage of prior knowledge of the intraclass correlation. Our assertion is that such prior knowledge is often available although infrequently used. We present several simple methods for taking advantage of this prior knowledge and show that the power of tests based on these methods can be substantially greater than the power of conventional tests especially when the number of experimental units is small.

## Introduction

In many areas of agricultural research, one may have a limited number of experimental units per treatment but have a large number of observational units within each experimental unit. For example, the treatments may be applied to just a few pens of animals but there are many animals per pen upon which observations are taken. The treatments may

be temperatures which are applied to growth chambers but measurements are taken on the individual plants within the chamber.

In the case of balanced experimental designs, traditional analysis of variance is equivalent to obtaining a mean for each experimental unit, then performing the analysis on means. The power of hypothesis tests for treatment effects can be low when the number of experimental units is small, and increasing the number of observational units per treatment beyond just of few units per treatment often has minimal effect on increasing the power. To overcome this problem, a researcher may be tempted to use the observational units as the units of analysis disregarding the random effects of the experimental units (e.g. disregarding pen effects as being negligible in comparison to animal to animal variability). This will increase the degrees of freedom for error but will generally result in a significant inflation in the Type I error. For instance, Blair, Higgins, Topping, Mortimer (1983) demonstrated this effect with real data in an educational setting involving classrooms as experimental units and students within classrooms as observational units. It is reasonable to conjecture that similar results would carry over to various experimental situations in agricultural.

Random variability among experimental units induces a correlation on observations within each unit called the intraclass correlation (ICC). If the value of the ICC is known, methods in Graybill (1976) can be applied to substantially improve the power of statistical tests in comparison to tests which do not take into account the value of ICC. For instance, results from Blair and Higgins (1986) confirm that using a known ICC can substantially increase the power of statistical tests for differences between means especially when the number of experimental units per treatment is small (2 or 3). The obvious problem with

their procedure is that one will rarely know the ICC. However, if a researcher has prior knowledge about the ICC, their results suggest that the power of tests for differences in treatment means can be substantially improved by incorporating this knowledge into the statistical testing procedure.

This study investigates some simple procedures that take advantage of prior knowledge about the ICC. The idea is as follows. Prior knowledge about the ICC is used to determine a "likely" value of the ICC. Both formal Bayesian and heuristic non-Bayesian methods are considered here. This value in turn is used in the test statistic proposed by Blair and Higgins (1986) as if it were the true value. Through simulations, it is shown that such test statistics hold their Type I error at near nominal levels and provide the desired increase in power.

## Model and Test Statistics

The mathematical model is

$$Y_{ijk} = \mu_i + \gamma_{ij} + \epsilon_{ijk} \qquad\qquad 1$$

where I = 1, 2, . . . , t; j = 1, 2, . . . , $m_i$; and k = 1, 2, . . . $n_{ij}$.

*Notation and Assumptions*

| | |
|---|---|
| An individual observation. | $y_{ijk}$ |
| Mean value for *i*th treatment. | $\mu_i$ |
| Random effect of *j*th experimental unit within *i*th treatment. | $\gamma_{ij}$ |
| Mean and variance of $\gamma_{ij}$. | 0, $\sigma_\gamma^2$ |
| Random variability associated with *k*th observation within *j*th experimental unit of *i*th treatment. | $\epsilon_{ijk}$ |
| Mean and variance of $\epsilon_{ijk}$. | 0, $\sigma_\epsilon^2$ |

The $\gamma_{ij}$'s are independent and have a normal distribution, likewise for the $\epsilon_{ijk}$'s. Also, the $\gamma_{ij}$'s and $\epsilon_{ijk}$'s are independent of each other. It is assumed that the $m_j$'s and $n_{ij}$'s are equal, i.e. there are m experimental units per treatment and n observations per experimental unit.

The intraclass correlation for the model is

$$\rho = \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_\epsilon^2} = \frac{\sigma_\gamma^2}{\sigma^2}$$

where $\sigma^2 = \sigma_\gamma^2 + \sigma_\epsilon^2$ .

Two treatments are considered; therefore, t = 2. The null hypothesis is $H_0$: $\mu_1 = \mu_2$, and the alternative hypothesis is $H_a$: $\mu_1 > \mu_2$ with a significance level of 0.05.

The test statistic to which the proposed statistics will be compared is the one that uses the means of the experimental units as the units of analysis. Consider the following quantities

$$x_{ij} = \overline{y}_{ij.} = \frac{1}{n} \sum_{k=1}^{n} y_{ijk} , \qquad \overline{x}_{i.} = \overline{y}_{i..} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{k=1}^{n} y_{ijk} .$$

The two-sample t-test on the means of the experimental units is given by

$$T_M = \frac{\overline{x}_{1.} - \overline{x}_{2.}}{\sqrt{s_p^2 \left( \frac{1}{m} + \frac{1}{m} \right)}} = \frac{\overline{x}_{1.} - \overline{x}_{2.}}{\sqrt{\frac{2 s_p^2}{m}}}$$

where

$$s_p^2 = \frac{1}{2}\left[ \frac{1}{m-1}\sum_{j=1}^{m}(x_{1j}-\overline{x}_{1.})^2 + \frac{1}{m-1}\sum_{j=1}^{m}(x_{2j}-\overline{x}_{2.})^2 \right] .$$

The null hypothesis is rejected if $T_M \geq t_{\alpha,df}$ where df $= 2(m - 1)$. The traditional analysis of variance F-statistic on the original observations is the square of $T_M$. Thus, the test statistic $T_M$ is the one-sided analog of the traditional analysis of variance F-test for this problem.

Based on Graybill (1976), Blair and Higgins (1986) gave the following test statistic if the ICC, $\rho$, is known:

$$T = \frac{\overline{y}_{1..} - \overline{y}_{2..}}{\sqrt{\frac{2\hat{\sigma}^2}{mn}(1+(n-1)\rho)}} \qquad\qquad 2$$

where

$$\hat{\sigma}^2 = \frac{1}{2(mn-1)}\left[ \frac{n-1}{1-\rho}\sum_{i+1}^{2}\sum_{j=1}^{m} s_{ij}^2 + \frac{n(m-1)}{1+(n-1)\rho}\sum_{i=1}^{2} b_i^2 \right] \qquad\qquad 3$$

$$s_{ij}^2 = \sum_{k=1}^{n}\frac{(y_{ijk}-\overline{y}_{ij.})^2}{n-1}, \text{ and } b_i^2 = \sum_{j=1}^{m}\frac{(\overline{y}_{ij.}-\overline{y}_{i..})^2}{m-1} .$$

The null hypothesis is rejected if $T \geq t_{\alpha,df}$ where df $= 2(mn - 1)$.

## Prior Knowledge Assumption

Our assertion is that it is often reasonable for a researcher to have prior knowledge about the ICC. Since the ICC is the fraction of the total variance due to the variability among experimental units, we believe that a researcher might very well have enough experience with similar data to be able to answer the question, "What fraction of the total variability is due to the experimental unit?" For instance, what fraction of the total variability in weights of animals is due to the pen effect, or what fraction of the total variability of dry matter of plants is due to the variability among growth chambers? Since this fraction depends on the intrinsic characteristics of the experimental units and not on scales of measurement, one might expect the ICC to be relative stable across many similar experiments. (This is what was found by Blair, et. al. (1983) in the educational setting). This prior knowledge can then be fashioned into either a prior interval for the ICC, (rl, ru), or a prior distribution for the ICC, $\pi(\rho)$.

Moreover, we speculate that in many agricultural situations the ICC is not likely to be large. For instance, in an experiment in which treatments were applied to pastures, and observations were taken on cows grazing on the pastures, we found the ICC to be less than 0.10. In an on-farm variety demonstration trial in which multiple measurements were taken on rows of corn (the rows being the experimental units), the ICC was .25. More research is needed to confirm our speculations; however, the potential gains that may accrue from knowing the ICC should encourage more researchers to publish estimated values for components of variance and the ICC.

## Procedures

A value for the ICC is determined by one of several proposed procedures. This value of the ICC is called the *assessed* value and is denoted by $\rho_1$. After $\rho_1$ is found, it replaces $\rho$ in the test statistic given in Equations 2 and 3. When $\rho_1$ is used, the test statistic and the estimate of $\sigma^2$ are denoted as $T_A$ and $\hat{\sigma}_A^2$, respectively. The degrees of freedom for the critical value is 2(mn - 1), which is based on a known value of the ICC. Satterthwaite approximations for degrees of freedom given by Milliken and Johnson (1992, pg. 281) were considered, but in our simulation study, we found little or no difference between the results with df = 2(mn - 1) and the degrees of freedom from Satterthwaite's approximation.

The procedures for determining the assessed ICC are listed below.

**Procedure 1: TRUE** As a comparison to the other procedures, the true ICC is used.

**Procedure 2: MID** The midpoint of a specified prior interval is used as the assessed value.

**Procedure 3: URL** The assessed value is the generalized maximum likelihood estimate

(GMLE) of the posterior distribution for a Uniform (rl, ru) prior distribution.

**Procedure 4: BAB** The assessed value is the GMLE of the posterior distribution for a

Beta ($\alpha$, $\beta$) prior distribution.

For comparison purposes, the parameters for the Beta distribution were chosen to match the mean and variance of the corresponding uniform prior.

**Bayes Estimation of the ICC**

The model is rewritten in vector notation as $\mathbf{Y} = X\boldsymbol{\mu} + \boldsymbol{E}$ where $\mathbf{Y}$ is a tmn x 1 vector, $\boldsymbol{\mu}$ is a t x 1 vector, $\mathbf{E}$ is a tmn x 1 vector, and X is a tmn x t design matrix. Since $E_{ijk} = \gamma_{ij} + \epsilon_{ijk}$, $\mathbf{E}$ is distributed $N_{tmn}$ ( $\mathbf{0}$, $\sigma^2 V$) where $\sigma^2 = \sigma_\gamma^2 + \sigma_\epsilon^2$, V is a tmn x tmn matrix, and

$$V = \begin{bmatrix} A & 0 & \dots & \dots & 0 \\ 0 & A & \dots & \dots & 0 \\ . & & . & & . \\ . & & & . & . \\ 0 & 0 & \dots & 0 & A \end{bmatrix}, \quad A = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix} .$$

Assume the following prior information:

a.    The prior density of $\mu$ is the noninformative constant prior over $\Re^t$.

b.    The prior density of $\sigma$ is the noninformative prior for scale parameters written as

$\pi(\sigma) \propto 1/\sigma, \sigma \geq 0.$

c.    The prior density of $\rho$ is denoted as $\pi(\rho)$, the actual density not being needed in the

derivation of the conditional posterior density.

d.    The prior parameters $\mu$, $\sigma^2$, and $\rho$ are independent random variables.

An advantage of the above prior information is that a researcher only needs to supply

prior information on the ICC since the noninformative priors for location and scale

parameters

 are used (Berger, 1985, pg. 85-86).  From the assumptions and prior information, the joint

prior density of $\mu$, $\sigma$, and $\rho$  can be written as   $\pi(\mu,\sigma,\rho) \propto \pi(\rho)/\sigma$ where $\mu \epsilon \Re^t$, $\sigma \geq 0$,

and $0 \leq \rho \leq 1$.

The steps to derive the conditional posterior density of ICC given $\mathbf{Y}$ are as follows.

(See Bond (1996) for more details).  A conditional posterior density involving all of the

parameters, $\rho$, $\sigma$, and $\mu$ given $\mathbf{Y}$ is found.  Then $\mu$ is eliminated by integrating the function

over $\Re^t$ with respect to $\mu$ and using the properties of the t-variate normal distribution. The remaining function is rewritten in terms of the precision, $\tau = 1/\sigma^2$, and $\tau$ is integrated out of the function using the properties of the gamma distribution which leaves a function which is proportional to the conditional distribution of $\rho$ given **Y**. This function is

$$\pi(\rho \mid \mathbf{Y}) \propto \frac{\pi(\rho)}{(1-\rho)^{\frac{tm(n-1)}{2}}(1+(n-1)\rho)^{\frac{tm-t}{2}}(\mathbf{Y}'V^{-1}\mathbf{Y}-\hat{\boldsymbol{\mu}}'C\hat{\boldsymbol{\mu}})^{\frac{tmn-t}{2}}}$$

where $C = X'V^{-1}X = \dfrac{mn}{1+(n-1)\rho}I_{t\times t}$, and $\hat{\boldsymbol{\mu}}$ is the least squares estimator of $\mu$ which is $\hat{\mu} = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{Y}$.

The value of $\rho$ which maximizes this function is known as the generalized maximum likelihood estimate or GMLE of the ICC (see Berger, 1985, p. 133). The GMLE is found using numerical methods.

## Type I Error and Power

A simulation study was done to investigate Type I error and power of the proposed procedures. All simulations were run in Fortran 77 using the IMSL Math/Stat Library. The simulation program selected the true ICC value, $\rho$, from a Uniform (rl, ru) distribution. By setting $\sigma_\epsilon^2$ equal to $1, \sigma_\gamma^2$ was determined. Using the variance components, a realization of the model given in Equation 1 was obtained. The test statistics, $T_M$ and $T_A$, were found with $T_A$ being calculated for the four procedures. The program counted the number of times that $H_o$ was rejected in one thousand repetitions. The parameter for these power functions is

$\delta = \mu_1 - \mu_2$; the power functions were generated from the same seed used for the Type I error,

$\delta = 0$.

Prior intervals for the Uniform(rl,ru) distribution both narrow and wide were considered and are listed below. The choice of intervals was somewhat arbitrary. The shorter intervals for the smaller values of ICC reflect our belief that smaller values of ICC tend to occur more frequently than larger values, and hence there will be a tendancy to have more precise prior information for ICC in the lower range of values. A more comprehesive study might yield different findings near the extremes for the ICC. However, for ICC near 0 or 1,

it is more likely that a satisfactory analysis of the data can be obtained by assuming "default" values of either ICC = 0 or ICC = 1, that is, no adjustment for ICC may be necessary in these cases.

| Interval # | Lower $\rho$ value | Upper $\rho$ value | Interval # | Lower $\rho$ value | Upper $\rho$ value |
|---|---|---|---|---|---|
| 1 | 0.10 | 0.15 | 7 | 0.50 | 0.60 |
| 2 | 0.10 | 0.20 | 8 | 0.50 | 0.70 |
| 3 | 0.10 | 0.30 | 9 | 0.50 | 0.80 |
| 4 | 0.20 | 0.25 | 10 | 0.70 | 0.80 |
| 5 | 0.20 | 0.30 | 11 | 0.70 | 0.90 |
| 6 | 0.20 | 0.40 | 12 | 0.70 | 0.95 |

The simulations were done for m = 2, 3, and 4 with the number of observations per experimental unit being set at n = 5, 10, 30, and 50. Therefore, for each of the four procedures a 3 x 4 x 12 study was done. To compare power functions, the maximum difference between the power of $T_A$ and $T_M$ was found for each of the procedures and each

of the 144 combinations of m, n, and prior interval selection. This is denoted as the *maximum power advantage*. The numerator mean square from the analysis of variance was obtained from the 144 combinations for each of the four procedures to see which factors (m,n, or interval) and their interactions had the greatest effect on the maximum power advantage. By far the predominant factor was m except for the BAB procedure where the prior interval effect and the prior interval by m interaction were also important.

Table 1 gives the maximum, minimum, and average of the 144 Type I error rates for each of the four procedures. On average, procedure MID holds the Type I error rate the best out of all procedures except for TRUE. The other procedures URL and BAB slightly inflate the Type I error rate. Bond (1996) showed that the Type I error rate is conservative if the assessed value of the ICC is consistently greater than the true value, while the opposite is true if the assessed value is consistently smaller than the true value. The lengths of the prior intervals 1 through 12 give an indication of what is reasonable in terms of how close one must be to the true ICC in order for the proposed test statistics to have a Type I error rate near the nominal value.

We examined the maximum power advantages of our 4 procedures for each of the 144 combinations of m, n, and prior interval selection. Table 2 contains the average, maximum, and minimum of these 144 values. Except for BAB, there is on average a 0.17-0.18 power advantage of $T_A$ over $T_M$. Table 3 gives the average maximum power advantage for the different levels of m. As m increases, the average maximum power advantage decreases. For m = 2 there is approximately a 0.30 increase in power of $T_A$ over $T_M$. As m increases to 4, this advantage in power decreases to approximately 0.10. Table 4 shows

the effects of m,n, and prior interval in more detail. With the exception of BAB we see that m is the factor that drives the maximum power advantage. BAB appears to be most adversely affected by interval 12 where perhaps the skewness of the prior distribution and the width of the prior interval diminishes its power. Figures 1 and 2 are power curves for MID, BAB, and $T_M$ for m = 2 and 5. The power of URL is similar to BAB and MID, so is not displayed. One sees clear power advantages of MID and BAB over the traditional procedure for m = 2 while the power advantage is decreased when m =4.

## Summary and Conclusions

The procedures studied here are recommended because they have desirable Type I error rates and power advantages over the traditional method. The lack of distinction among our procedures indicates that prior knowledge may be expressed by the researcher in different ways, e.g. with different priors, with improvements over the traditional method being possible. The largest improvement over the traditional method is for small values of m (2 or 3) with diminishing gains for larger values of m.

It is relatively straight forward to develop procedures for multiple treatments in balanced designs. Unequal sample sizes may present some computational difficulties in obtaining a posterior distribution for the ICC, but again, even something as simple as selecting the midpoint of a prior interval for ICC may yield improvements in power. More complicated correlation structures present additional modeling problems, of course. Again, however, correlations are something that researchers can intuitively relate to in the context of their experiments. While Bayesian procedures have not been popular in agriculture because of the subjectivity of prior information, it seems reasonable that an agricultural researcher

may be able to get a handle on correlations based on actual data obtained from prior experiments. At least this possibility should be explored.

## Bibliography

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis 2nd Edition*. New York: Springer-Verlag.

Blair, R.C., and Higgins, J.J. (1986). "Comment on 'Statistical Power with Group Mean as the Unit of Analysis'." *Journal of Educational Statistics* 11, 161-169.

Blair, R.C., Higgins, J.J., Topping, M.E.H., and Mortimer, A.L. (1983). "An Investigation of the Robustness of the t Test to Unit of Analysis Violations." *Educational and Psychological Measurement* 43, 69-80.

Bond, M.E. (1996). *Using Prior Knowledge of the Intraclass Correlation to Increase the Power of Hypothesis Tests for Treatment Means*. Diss. Kansas State U. Ann Arbor: UMI, 1997. # 9637239.

Graybill, F.A. (1976). *Theory and Application of the Linear Model*. California: Wadsworth.

Milliken, G.A., and Johnson, D.E. (1992). *Analysis of Messy Data Volume I: Designed Experiments*. New York: Van Nostrand Reinhold.

**Table 1, Type I Error Rates**

|         | TRUE  | MID   | URL   | BAB   |
|---------|-------|-------|-------|-------|
| AVG.    | 0.051 | 0.053 | 0.059 | 0.055 |
| LARGE   | 0.066 | 0.077 | 0.087 | 0.098 |
| SMALL   | 0.037 | 0.036 | 0.042 | 0.026 |

**Table 2, Maximum Power Advantage**

|         | TRUE  | MID   | URL   | BAB    |
|---------|-------|-------|-------|--------|
| AVG.    | 0.173 | 0.180 | 0.180 | 0.139  |
| LARGE   | 0.332 | 0.340 | 0.332 | 0.332  |
| SMALL   | 0.072 | 0.069 | 0.074 | -0.075 |

**Table 3, Average Maximum Power Advantage for Various Values of m**

| m | TRUE  | MID   | URL   | BAB   |
|---|-------|-------|-------|-------|
| 2 | 0.287 | 0.297 | 0.297 | 0.261 |
| 3 | 0.138 | 0.146 | 0.144 | 0.101 |
| 4 | 0.092 | 0.099 | 0.097 | 0.057 |

**Table 4, Maximum Power Advantage for Selected Values of m,n, and Prior Interval**

| m | n | interval | TRUE | MID | URL | BAB |
|---|---|---|---|---|---|---|
| 2 | 5 | 1 | .280 | .277 | .289 | .316 |
| 2 | 5 | 12 | .233 | .264 | .241 | .091 |
| 2 | 50 | 1 | .302 | .302 | .316 | .303 |
| 2 | 50 | 12 | .259 | .306 | .267 | .011 |
| 4 | 5 | 1 | .084 | .085 | .085 | .116 |
| 4 | 5 | 12 | .087 | .115 | .087 | -.012 |
| 4 | 50 | 1 | .081 | .069 | .096 | .066 |
| 4 | 50 | 12 | .094 | .145 | .090 | -.050 |