DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 5 | Issue 2 Article 17

11-1-2005

A Method for Analyzing Unreplicated Experiments Using Information on the Intraclass Correlation Coefficient

Jamis J. Perrett
University of Northern Colorado, jamis.perrett@unco.edu

Recommended Citation

Perrett, Jamis J. (2005) "A Method for Analyzing Unreplicated Experiments Using Information on the Intraclass Correlation Coefficient," *Journal of Modern Applied Statistical Methods*: Vol. 5 : Iss. 2 , Article 17.
DOI: 10.22237/jmasm/1162354560
Available at: http://digitalcommons.wayne.edu/jmasm/vol5/iss2/17

Journal of Modern Applied Statistical Methods
November, 2006, Vol. 5, No. 2, 432- 442

1538 – 9472/06/$95.00

A Method for Analyzing Unreplicated Experiments Using Information on the Intraclass Correlation Coefficient

Jamis J. Perrett
Department of Applied Statistics and Research Methods
University of Northern Colorado

Many studies are performed on units that cannot be replicated; however, there is often an abundance of subsampling. By placing a reasonable upper bound on the intraclass correlation coefficient (ICC), it is possible to carry out classical tests of significance that have conservative levels of significance.

Key words: Intraclass correlation coefficient, unreplicated experiment, subsamples

Introduction

A researcher, wishing to compare two different teaching methods, teaches two classes: one with method 1 and one with method 2. The grade of each student in the two classes is recorded with the purpose of comparing the average grade for the students taught by method 1 to the average grade for the students taught by method 2. The within class variation is the variability from student to student. The between class variation is due to such factors as time of day, difference in classroom setting, etc. One would expect the variation from class to class to be small relative to the within class variation, regardless of whether the students are being taught mathematics, creative writing, etc. The majority of the total variability will be explained by the difference in performance of the students within a class, and that should be fairly similar for one subject as it is another. Thus, the intraclass correlation coefficient (ICC) should be consistent in studies of similar types, and it will tend to be small (In an education example such as this, it would not be unusual for the ICC to be less than 0.1) in many situations.

An unreplicated experiment is one in which a treatment of interest is applied to only one unit. Some experiments logistically cannot be replicated. Circumstances that might prevent replication are cost in time or money or both, scarcity of experimental units, destructive experimentation, among other things. Some farmers just don't have an extra plot of land to experiment on. Consideration is given for what can be done in such cases.

The Model

Let y_{ij} be the measurement taken on the j^{th} student given the i^{th} treatment, μ_i is the fixed effect of treatment *i*, δ_i is the random effect of class *i*, and ε_{ij} is the random effect of student *j* given treatment *i*, $i = 1, 2, \ldots, t$; $j = 1, 2, \ldots, n_i$. Let $\delta_i \sim n\left(0, \sigma_\delta^2\right)$, where σ_δ^2 represents the between class variability; let $\varepsilon_{ij} \sim n\left(0, \sigma_\varepsilon^2\right)$, where σ_ε^2 represents the between student within class variability. It is assumed that δ_i and ε_{ij} are independent. A model is written for the experiment as follows:

Jamis Perrett is Assistant professor of Applied Statistics at the University of Northern Colorado. His research interests include the analysis of unreplicated experiments, computational statistics, and repeated measures analysis. E-mail: jamis.perrett@unco.edu

$$y_{ij} = \mu_i + \delta_i + \varepsilon_{ij} \qquad (1)$$

This type of model is a single factor completely randomized design with subsampling, where classes are the experimental units for each treatment level, and the students within each class are the subsamples, or observational units. A researcher, who uses the students as the experimental units, ignores the variability that can exist between different classes receiving the same treatment. Such an assumption is to claim that σ_δ^2=0. If the researcher correctly uses classes as the experimental units, there is only one unit per treatment level and zero degrees of freedom available for testing the difference between these treatment means (Barcikowski, 1981, Blair, 1986).

The Intraclass Correlation Coefficient and the Independence Assumption

The intraclass correlation coefficient (ICC) is defined as the correlation between y_{ij} and $y_{ij'}$ (two subsample units within one experimental unit). In this study, ρ refers to the true value of the ICC and ρ_0 refers to a best guess value, chosen by the researcher, to plug into formulas in place of the ICC in the analysis. The ICC for the model in Equation 1 can be obtained using the following formula:

$$\rho = \frac{\operatorname{cov}(y_{ij}, y_{ij'})}{\sqrt{\operatorname{var}(y_{ij})\operatorname{var}(y_{ij'})}} = \frac{\sigma_\delta^2}{\sigma_\delta^2 + \sigma_\varepsilon^2} \qquad (2)$$

Thus, if σ_δ^2=0, the ICC is also zero. The result is independent samples assuming normality of error terms. To illustrate the ideas, consider the two-treatment case in which $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$ is tested. The variance of the difference between the two sample means is given by

$$\operatorname{var}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = 2\sigma_\delta^2 + \sigma_\varepsilon^2\left[\frac{n_1 + n_2}{n_1 n_2}\right]$$

$$= \sigma_\varepsilon^2\left\{2\left(\frac{\rho}{1-\rho}\right) + \left[\frac{n_1 + n_2}{n_1 n_2}\right]\right\} \qquad (3)$$

using the substitution $\sigma_\delta^2 = \sigma_\varepsilon^2 \dfrac{\rho}{1-\rho}$.

$$Z = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{\sqrt{\sigma_\varepsilon^2\left\{2\left(\frac{\rho}{1-\rho}\right) + \left[\frac{n_1 + n_2}{n_1 n_2}\right]\right\}}} \qquad (4)$$

where Z has a standard normal distribution. If ρ is incorrectly assumed to be zero, then

$$Z = \frac{\bar{y}_{1\cdot} - \bar{y}_{2\cdot}}{\sqrt{\sigma_\varepsilon^2\left(\frac{n_1 + n_2}{n_1 n_2}\right)}}. \qquad (5)$$

A test statistic for a hypothesis test based on the incorrect assumption that observations are independent will be too large and consequently inflate the associated Type 1 error.

Bounding the ICC

In practice, σ_ε^2 may be estimated from the pooled variance of observational units within the experimental unit. With many subsamples, σ_ε^2 can be estimated quite accurately. However, in an unreplicated experiment there is no way to estimate σ_δ^2 and consequently ρ.

Although it may not be possible to know the value of the ICC, in many cases it may be reasonable to make assumptions about its upper bound. To do so, one must consider the relative size of the between unit variability to the within unit variability. In the example considered in this study, the classes were similar, so it is reasonable to assume the component of the variance due to classes is relatively small. On the other hand, the component of variance due to differences among students within a class tends to be relatively large due to the inherent differences in students: maturity, study habits, initial understanding, intelligence, etc. Thus, it would seem to be reasonable to place a bound on the ICC that is less than .5 and possibly quite a bit smaller than this. Data discussed in this study indicate that a bound of $\rho < .15$ is reasonable for this example.

Other examples of this also are common in agriculture. Consider for instance feeding treatments applied to pens with measurements made on individual animals within pens. For many measurements such as weight gain, body condition scores, and various blood parameters, the greatest source of variability is among the animals within the pens. The component of variance due to pens, while not negligible, is often just of fraction of the component of variance due to the animals. In such cases it is quite reasonable to assume that ρ is small. The upper bound will be denoted as $\rho_{\max}$.

The importance of a small value of ρ can be seen in Equation 3 which shows that the variance of the difference of sample means gets smaller as ρ gets smaller. In the limit, the variance is that of the difference of means of independent observations. Intuitively, the closer ρ is to zero, the more the observations behave as if they were independent. Moreover, the analysis using a known ICC becomes more powerful as ρ gets closer to zero with the limiting power being that obtained when the observations are independent.

Methodology

The Testing Strategy

Although ρ is not known, if prior experience allows for an upper bound to be placed on it, then it is possible to carry out statistical tests for unreplicated experiments. Consider a test of the hypotheses H$_0$: $\mu_1 = \mu_2$ vs. H$_A$: $\mu_1 \neq \mu_2$. Let ρ_0 denote a value of ρ that the researcher assumes to be reasonable based on prior experience. Let $\mu_{01} - \mu_{02}$ represent the hypothesized difference of the mean for treatment 1 and the mean for treatment 2 respectively. The test statistic is

$$T = \frac{(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) - (\mu_{01} - \mu_{02})}{\sqrt{\hat{\sigma}_\varepsilon^2 \left\{ 2\left(\frac{\rho_0}{1-\rho_0} \right) + \left[\frac{n_1 + n_2}{n_1 n_2} \right] \right\}}} \qquad (6)$$

where $\hat{\sigma}_\varepsilon^2$ represents a pooled estimate of the within class variance. Let

$$s_i^2 = \frac{\sum_{k=1}^{n_i} (y_{ik} - \bar{y}_{i\cdot})^2}{n_i - 1}, \text{ i=1, 2} \qquad (7)$$

be the variance of the measurements under treatment *i*. Then

$$\hat{\sigma}_\varepsilon^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \qquad (8)$$

Let α_0 denote the nominal level of the test and α the true level of significance. Depending on assumptions, α_0 may or may not equal α.

Balancing simplicity and desirable properties, it is suggested that tests be based on p-values obtained when $\rho_0 = \rho_{\max}$. It is also recommended that confidence intervals and multiple comparisons be carried out using $\rho_0 = \rho_{\max}$.

Properties of the test statistic

For simplicity, assume the number of subsamples per class is the same for all classes, i.e. $n = n_1 = n_2$. The sample size is one–one class per treatment. Let $\mu_1 - \mu_2$ be the true value of the difference between the treatment means to be compared in the hypothesis test. Let υ=2(n-1), the degrees of freedom for the test. Let $t_{0.05,\upsilon}$ denote the upper tail 0.05 value of the t-distribution with υ degrees of freedom. The probability of rejecting H$_0$ for an upper-tail test at $\alpha_0 = 0.05$ can be determined using the following steps:

$$P\left[\frac{(\bar{y}_1-\bar{y}_2)-(\mu_{01}-\mu_{02})}{\sqrt{2\hat{\sigma}_\varepsilon^2\left(\frac{\rho_0}{1-\rho_0}+\frac{1}{n}\right)}}>t_{0.05,\upsilon}\right] \quad (9)$$

If the following is defined as

$$Z=\frac{(\bar{y}_1-\bar{y}_2)-(\mu_1-\mu_2)}{\sqrt{2\sigma_\varepsilon^2\left(\frac{\rho}{1-\rho}+\frac{1}{n}\right)}} \quad (10)$$

and

$$\lambda=\frac{(\mu_1-\mu_2)-(\mu_{01}-\mu_{02})}{\sqrt{2\sigma_\varepsilon^2\left(\frac{\rho}{1-\rho}+\frac{1}{n}\right)}} \quad (11)$$

then Equation 9 can be expressed as

$$P\left[\frac{Z+\lambda}{\sqrt{U/\upsilon}}>(t_{0.05,\upsilon})\frac{\sqrt{\frac{\rho_0}{1-\rho_0}+\frac{1}{n}}}{\sqrt{\frac{\rho}{1-\rho}+\frac{1}{n}}}\right] \quad (12)$$

where Z~n(0,1), λ is a constant, and U~χ_υ^2. To the left of the inequality is a random variable with a non-central t-distribution with non-centrality parameter λ and degrees of freedom, υ.

Evaluation

The plug-in method involves choosing a value ρ_0 to plug into Equation 6. The method is evaluated by determining significance levels and power curves for the tests, for different values of ρ_0 and ρ. In particular, the method is considered useful if it can maintain a Type 1 error level close to the nominal level (α_0) while providing power to detect differences in treatment means for a reasonable range of ρ_0 near ρ.

Equation 12 may be used to evaluate the probability of rejection for tests of the hypothesis that the two means are equal. These probabilities depend on the values for ρ, ρ_0, n, and the non-centrality parameter, λ. In order to measure the deviation between the two means, a standardized difference is defined as StDiff = $\frac{\mu_1-\mu_2}{\sigma_\varepsilon}$. Probabilities will depend on StDiff through λ. Using Equation 12, probabilities were generated using the following values as indices:

StDiff = 0, 1
n = ∞
ρ = 0 through 0.99 in increments of 0.01
ρ_0 = 0 through 0.99 in increments of 0.01

Figure 1 is a result of the generated probabilities. The data used for this plot were created by evaluating Equation 12. For the plot, let $n \to \infty$ resulting in $\lim_{n\to\infty}\frac{1}{n}=0$ being used in Equation 12. The nominal significance level $\alpha_0=0.05$ is used for the plot. Also, neither the power nor the significance level is defined at ρ=1 or ρ_0=1. This plot depicts the case with a theoretically infinite number of students per class. A similar graph results from using 10 students per class (omitted). This power plot is overlaid with the corresponding plot of two-tail significance levels. The red, green, and blue areas denote power in the ranges ≤0.2, 0.2 to 0.5, and 0.5 to 0.9 respectively. The lines represent the two-tail probabilities of rejecting the null hypothesis. If $\rho=\rho_0$,

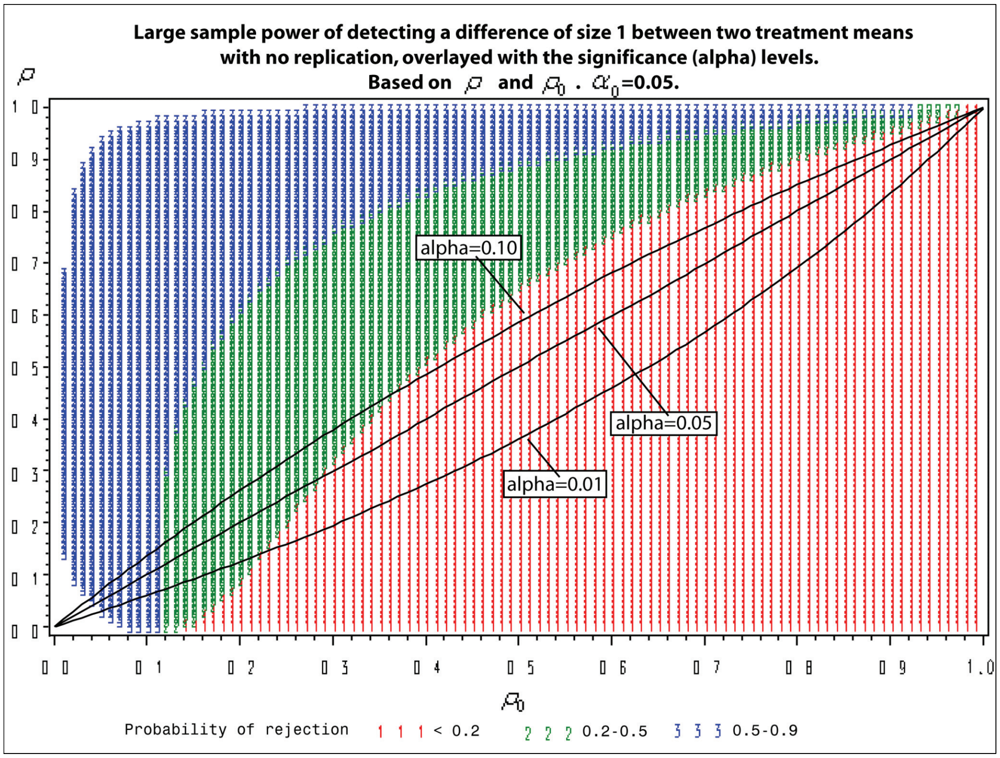

Figure 1. Large (sub)sample power of detecting a difference of size 1 between two treatment means with no replication, overlaid with large sample significance (alpha) levels.

then α=0.05. If the plug-in value ρ_0 is less than the true value ρ ($\rho \geq \rho_0$), the Type 1 error is inflated ($\alpha \geq 0.05$). For example, if a value of ρ_0=0.2 is used when ρ=0.4, alpha will be 0.2253. If the plug-in value ρ_0 is greater than the true value ρ ($\rho \leq \rho_0$), the probability of a Type 1 error is smaller than α_0=0.05.

Because the p-value increases as ρ_0 increases, the p-value obtained when ρ_0 = $\rho_{\max}$ is greater than or equal to the true p-value, so tests using this methodology are conservative. On the other hand, if $\rho_{\max}$ is set too small by mistake, Type 1 error will be inflated. In terms of power and length of confidence intervals, a smaller $\rho_{\max}$ is better than a larger one.

This study does not suggest that problems with lack of replication magically disappear with this methodology. Even if ρ is known, ρ>0 presents problems. In the two sample case, for instance, the variance of the difference between two means when the number of subsamples n_1 and n_2 approaches infinity becomes

$$= 2\sigma_{\varepsilon}^{2}\frac{\rho}{1-\rho} \qquad (13)$$

This would be the formula for the variance of the difference between two sample means if each treatment had n = $\frac{1-\rho}{\rho}$ replications. Thus, $\frac{1-\rho}{\rho}$ can be thought of as the number of pseudo replications. For instance if ρ=.2, the number of pseudo replications is 4. This study simply suggests the proposed methodology gives researchers a way to analyze data when conventional analysis of variance could not be carried out due to lack of replication.

It is also useful to examine the maximum power attainable using a plug-in value for ρ in hypothesis testing. Figure 2 demonstrates the maximum attainable power under the most ideal conditions: namely $\rho = \rho_0$ and $n = \infty$.

As seen in Figure 2, if the true value ρ is 0.5, the power is low. For smaller values of ρ the power increases considerably. For instance, for a standardized difference of 1.0 and ρ=0.1, the power is 0.564. Thus, using a plug-in value for ρ is only going to be effective for smaller values of ρ.

What the Researcher May Know About ρ

It is possible for the researcher to obtain information about ρ based on prior experiments of a similar nature, or from knowledge about the behavior of ρ for a current experiment.

Distributional Information

If a large amount of distributional information is available from prior studies of a nature similar to that of the current study, the researcher may be able to put a prior distribution or empirical distribution on ρ (not considered in this study).

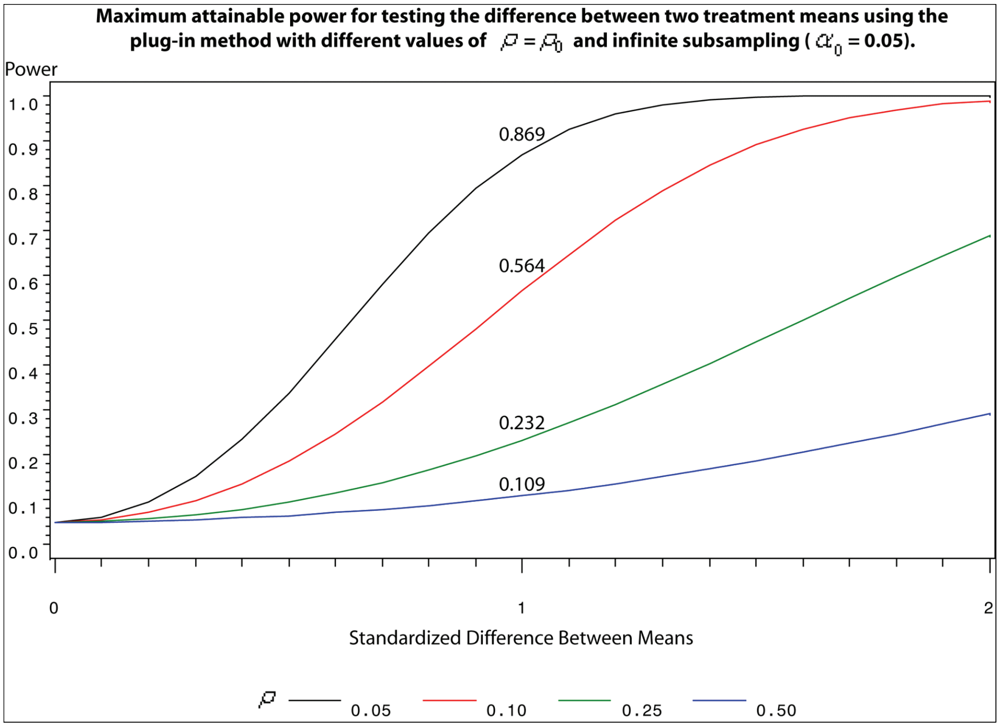

Figure 2. Maximum attainable power for testing the difference between two treatment means using the plug-in method with different values of $\rho = \rho_0$ and infinite subsampling.

Point or Interval Information

The researcher may not have extensive distributional information about ρ, but may have an indication of a reasonable minimum or maximum value of ρ.

The Plug-in Value

In the case of no replication, zero degrees of freedom exist for conducting a hypothesis test comparing means. So, the test cannot be performed using traditional methods. With this procedure a value, chosen by the researcher, is used as if it were the true value, ρ. This value, ρ_0, called a plug-in value, can be used in hypothesis testing and in producing confidence intervals of differences of treatment means. The strategies for choosing a value for ρ_0 given in this section are proposed to researchers who have an unreplicated experiment and have a reasonable idea of the actual value, ρ. Other methods for dealing with unreplicated experiments have been investigated with similar results to the plug-in methods.

Proposed Strategies

Two strategies that make use of prior information about ρ to test hypotheses about treatment means in an unreplicated, two-treatment experiment will be presented in this study.

Strategy 1: Plot of the Conditional P-value Given ρ_0

Description

For a given set of data, the p-value may be obtained for various assumed values of ρ. It is computed as the probability that a t-distributed random variable with $n_1 + n_2 - 2$ degrees of freedom is more extreme than the observed value of the statistic defined by Equation 10. A conditional p-value plot plots p-values for testing a certain hypothesis over the range of possible values of ρ.

Properties

The three situations a researcher may encounter with a conditional p-value plot are 1) all p-values are above α_0 for reasonable values of ρ, 2) some p-values are below α_0 for reasonable values of ρ and some are above α_0 for reasonable values of ρ, and 3) all p-values are below α_0 for reasonable values of ρ. When the first situation occurs, the result of the test is to fail to reject the null hypothesis at level α_0. When the third situation occurs, the result of the test is to reject the null hypothesis at level α_0. When the second situation occurs, it is less obvious what the results of the test of hypothesis should be. The researcher is advised to refrain from making a decision about the acceptance of the null hypothesis if the second situation is observed. The accuracy of the results will depend on the accuracy of the choice of the likely range of ρ.

Implementation

The researcher determines the range of reasonable values for the ICC based on information obtained from prior studies. Then the researcher creates the conditional p-value plot given the possible values of ρ. The decision to reject or fail to reject the null hypothesis, or abstain from judgment on the acceptance of the null hypothesis is made based on the observed p-values within the range of reasonable values of ρ.

Strategy 2: Maximum Rho

Description

The Maximum Rho procedure simply involves choosing $\rho_0 = \rho_{\max}$. That value, ρ_0, is then incorporated into the test statistic (Equation 6). The test rejects the null hypothesis if the p-value is less than α_0.

Properties

The Maximum Rho procedure assures the true significance level, α, is less than or equal to the nominal value α_0, with equality when $\rho_0 = \rho$. The closer ρ_0 is to ρ, the higher the power of the test.

Implementation

To implement the Maximum Rho procedure, the researcher simply computes a p-value for the test based on using $\rho_0 = \rho_{\max}$ in Equation 6.

Example

The class data consists of final course grades for two classes of introductory statistics taught by two different methods. The sample means for the two classes were 2.83 and 3.37 with sample standard deviations of 1.04 and 0.84 respectively. A researcher would like to see if there is a difference between the average grade received by students taught by the two different methods. Only one class was observed for each of the two methods making this an unreplicated experiment with class as the unit of study, student as the subsampling unit. Let $H_0 : \mu_1 = \mu_2$ and $H_A : \mu_1 \neq \mu_2$. Let $\alpha_0 = 0.05$. It is assumed that replication would have yielded small class-to-class variability relative to the total variability. To determine if that could in fact be the case, a study was conducted on prior undergraduate courses taught to multiple sections over multiple semester.

College Course Grades

The value ρ was estimated from a variety of courses offered at Kansas State University using the SAS® MIXED procedure. The components of variance consist of variability of scores due to section, σ_δ^2; and the variability of scores due to students within sections, σ_ε^2. Fourteen different undergraduate courses were selected (CHM 111, 210, 230; CIS 101; ENGL 100, 125; MATH 010, 100; MUSIC 250, 255; PSYCH 110, 202, 350; SPAN 161) each with multiple sections, covering both Fall and Spring semesters over the years 2001-2003, for a total of 43 course-semester combinations. These values are graphed in a frequency histogram in Figure 3.

All 43 estimated ICC values are at or below 0.33. The majority, 95%, is at or below 0.2–90% are below 0.15. Only one value is at 0.33. That value is for a honors English course (ENGL 125). Other courses include undergraduate courses in chemistry, English, music, CIS, math, psychology, and Spanish. Based on this, it would be reasonable to use the range of ρ from 0.02 to 0.15, with an assumed maximum at $\rho_0 = \rho_{\max}$=0.15 for the plug-in methods involving grades for this study.

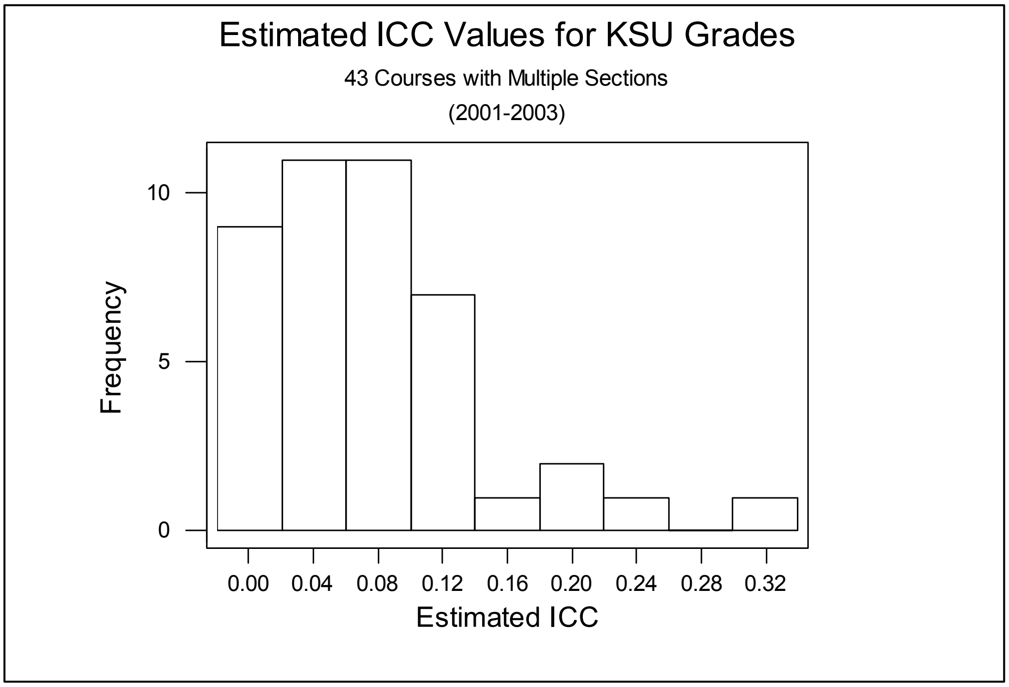

Figure 3. Estimated ICC values for KSU grades for 43 courses with multiple sections (2001-2003).

Results

Plot of the Conditional P-value Given ρ_0

Figure 4 plots the p-value as a function of ρ for the class data. It can be seen that p-values are only significant if ρ is less than 0.013. Because the likely value of ρ is greater than 0.013, the result of the test is to fail to reject the null hypothesis at $\alpha_0 = 0.05$.

Maximum Rho

Using ρ_0=0.15 and Equation 6, the test statistic is found to be t=-0.894 with a p-value of 0.374. So the result of the test is to fail to reject the null hypothesis in favor of the alternative (at $\alpha_0 = 0.05$).

Based on both the conditional p-value plot and the results of the Maximum Rho procedure, the conclusion is that the difference between the mean grades for the two classes is not significant. Assuming $\rho_0 \geq \rho$, the probability of a Type 1 error for the test in this example is at most 0.05. If, in fact, $\rho = \rho_0$=0.15, the power for detecting a difference of one grade point is approx. 0.3373. However, if ρ<0.15, the power will be less. A significant difference would have been detected using a classical t-test to test the same hypothesis under the assumption of independent samples. However, the probability of a type 1 error would be inflated under that assumption.

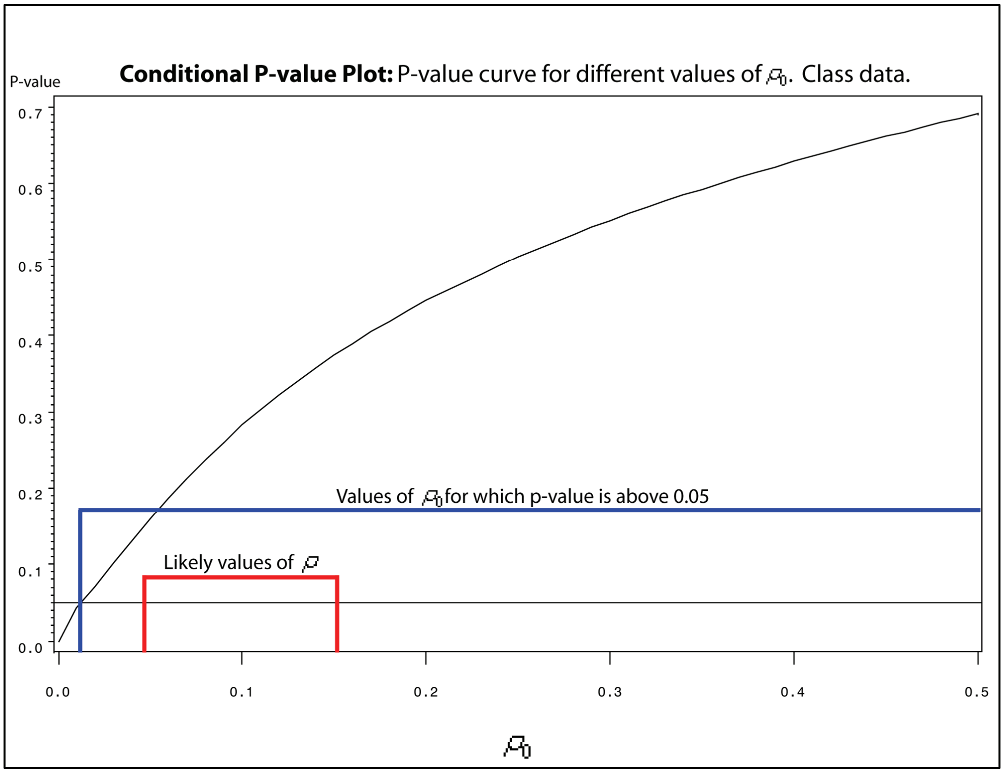

Figure 4. Conditional P-value Plot. P-value curve for different values of ρ_0 using the Class data.

Conclusion

If replication is feasible, a replicated experiment is always preferred over an unreplicated experiment. However, many studies are performed on units that cannot be replicated. The method described in this paper makes it possible to accurately analyze unreplicated experiments in which the intraclass correlation coefficient (ICC) is small and relatively stable. By placing a reasonable upper bound on the ICC, it is possible to carry out classical tests of significance that have conservative levels of significance. This methodology has wide applicability for analyzing unreplicated experiments in many fields of research and its simple computations will surely appeal to the applied researcher.

References

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6,* 267-285.

Blair, R C., et. al. (1983). An investigation of the robustness of the t-test to unit of analysis violations. *Educational and Psychological Measurement, 43,* 69-80.

Blair, R. C., & Higgins, J. J. (1986). Comment on statistical power with group mean as the unit of analysis. *Journal of Educational Statistics Summer, 2,* 161-169.

Casella, G., & Berger, R. L. (2002). *Statistical inference (2nd ed.)*. Pacific Grove, CA: Duxbury.

Graybill, F. A. (1976). *Theory and application of the linear model*. Pacific Grove, CA: Wadsworth.

Little, R. C., et. al. (1996). *SAS system for mixed models.* Cary, NC: SAS Institute Inc.

Milliken, G. A., & Johnson, D. E. (1992). *Analysis of messy data, Volume 1: Experimental design*. London, UK: Chapman & Hall.

SAS Institute Inc. (1999). *SAS/STAT user's guide, version 8.* Cary, NC: Author.

SAS is a registered trademark of SAS Institute Inc., Cary, N.C.