Published in final edited form as:

J Prim Prev. 2009 September ; 30(5): 497–512. doi:10.1007/s10935-009-0191-y.

Using a Nonparametric Bootstrap to Obtain a Confidence Interval for Pearson's r with Cluster Randomized Data: A Case Study

David A. Wagstaff,

College of Health and Human Development, The Pennsylvania State University, 153 Henderson Building, University Park, PA 16802, USA

Elvira Elek,

RTI International, Washington, DC, USA

Stephen Kulis, and

Arizona State University, Tempe, AZ, USA

Flavio Marsiglia

Arizona State University, Tempe, AZ, USA

David A. Wagstaff: daw22@psu.edu; Elvira Elek: ; Stephen Kulis: ; Flavio Marsiglia:

Abstract

A nonparametric bootstrap was used to obtain an interval estimate of Pearson's r , and test the null hypothesis that there was no association between 5th grade students' positive substance use expectancies and their intentions to not use substances. The students were participating in a substance use prevention program in which the unit of randomization was a public middle school. The bootstrap estimate indicated that expectancies explained 21% of the variability in students' intentions ($r = 0.46$, 95% CI = [0.40, 0.50]). This case study illustrates the use of a nonparametric bootstrap with cluster randomized data and the danger posed if outliers are not identified and addressed. *Editors' Strategic Implications:* Prevention researchers will benefit from the authors' detailed description of this nonparametric bootstrap approach for cluster randomized data and their thoughtful discussion of the potential impact of cluster sizes and outliers.

Keywords

Cluster randomization; Nonparametric bootstrap; Confidence interval; Pearson's r

Introduction

When data have been obtained from individuals who constitute intact social groups (e.g., members of the same family, students attending the same neighborhood school, or patients seen by the same physician), the individuals' responses are not statistically independent. Indeed, as a result of prior selection, common exposure, and/or mutual influence, the responses of any two randomly selected group members tend to be more similar to one another than the responses of any two randomly selected individuals from the parent population (Donner 1998; Donner and Klar 1999; Kish 1957; Murray 1998; Myers et al. 1981). Failure to account for these dependencies during a study's planning phase and when the data are analyzed can result in a seriously underpowered study, underestimated SEs, inflated test statistics, and actual

Type I error rates that are many times larger than the nominal Type I error rates. These consequences were first identified by Walsh (1947) and later stated with highly memorable phrasing by Cornfield (1978). The latter and other researchers have shown that estimation and inference are affected whether the data have been collected from (a) intact groups that have been randomly sampled (Kish and Frankel 1974; Korn and Graubard 1999; Levy and Lemeshow 1999); (b) intact groups that have been randomly assigned to study conditions (Bieler and Williams 1995; Localio et al. 1995); or (c) consist of repeated observations on individuals grouped by one or more characteristics (LaVange et al. 1994).

Finding appropriate interval estimates for means, variances, and regression coefficients when observations are not independently and identically distributed has received greater attention in the last two decades because of the increased use of cluster randomization by education, health, and prevention researchers (see Altman 2000; Bryk and Raudenbush 1992; Donner and Klar 1999; Goldstein 1995; Murray 1998). To date, this multidisciplinary literature and, more importantly, the commercial software programs that have been developed to analyze cluster randomized data, do not provide measures like Pearson's r to assess the strength of association. This oversight is surprising given the ubiquitous role that Pearson's r plays in studies where the data were collected with a simple random sampling plan.

Past Research

There are two areas where one might expect to find researchers using Pearson's r with observations obtained from intact groups: the analysis of familial data and the analysis of complex sample survey data.

Familial Data

Statisticians, geneticists, and family researchers have developed a rich literature that addresses the extent to which observations taken on members from two different classes (e.g., parent and child) or from the same class (e.g., siblings) are similar to one another. Unfortunately, the theoretical arguments that have been used to derive the respective interclass (different classes) and intraclass (same class) correlation coefficients have assumed that siblings' responses are statistically independent or they have not addressed the dependencies. Researchers have accomplished the latter by taking the mean of the siblings' responses or randomly selecting one of the sibs' responses and analyzing the resulting data set (Keen and Elston 2003; Rosner et al. 1977). The literature on the analysis of familial data is relevant to the present study because it demonstrates that researchers have not been able to derive suitable closed form expressions for the maximum likelihood estimator of an interclass or intraclass correlation coefficient when the number of elements in each cluster varies (Rosner et al. 1977). Specific examples from this literature include the number of family members and the number of children in each family. Examples from prevention literature include the number of students within each randomized school and the number of patients seen by a participating physician who has been randomized to a study condition. Finally, the literature on the analysis of familial data has shown that some iterative or computer-intensive approach will be required if a prevention researcher wants to obtain point and interval estimates of a correlation coefficient when the data were obtained with a cluster randomized design and cluster sizes vary.

Complex Sample Survey Data

In the 1950s, design-based survey statisticians developed statistical theory and methods that addressed the analytic challenges they confronted when their random sampling plan involved stratification, clustering, and the selection of units with unequal probability at one or more stages (Levy and Lemeshow 1999). With this design-based approach, a researcher did not have to specify a distributional model in order to obtain consistent estimates of population regression

parameters. Today, a number of commercial programs (e.g., Stata, SUDAAN) can fit a variety of regression models to complex sample survey data. More importantly, any researcher who uses these programs can calculate point and interval estimates for an odds ratio, the preeminent measure of association used by epidemiologists and other health researchers. However, no commercial package calculates interval estimates for Pearson's r when observations are not independently distributed.

Sribney (2001) noted that the expressions for the population regression coefficient, β , and population correlation coefficient, ρ , took their familiar forms with complex sample survey data and, further, that the same relation between the two population parameters held (i.e., $\rho = \beta\sigma_X/\sigma_Y$) whether the data were obtained from a simple random sample or complex sample. However, Sribney noted that the equivalence between the t -statistic used to test the null hypothesis that the population regression coefficient equaled zero, $H_0 : \beta = 0$, and the t -statistic used to test the null hypothesis that the population correlation coefficient equaled zero, $H_0 : \rho = 0$, did not hold for complex sample survey data (e.g., a cluster sample).

Although Sribney (2001) did not indicate why researchers could use the standard expression to obtain a point estimate of Pearson's r with complex sample survey data, Kish (1957) had considered point estimation and the construction of confidence intervals for Pearson's r with clustered samples. Kish showed that the familiar expression for Pearson's r often provided a good point estimate of the population correlation coefficient when observations were not independently and identically distributed. However, he also showed that the simple random sampling estimator of the SE of r would exhibit greater variability with clustered samples because of the dependencies among units from the same cluster.

When Kish and Frankel (1974) investigated the sampling behavior of simple, partial, and multiple correlation coefficients in a large empirical study of complex sample survey data, they found that the three methods commonly used to estimate population parameters (specifically, Taylor linearization, balanced repeated replication, and jackknife repeated replication) yielded point estimates that exhibited little bias and, further, that the proportion of test statistics that actually fell in the 95% confidence interval based on the appropriate t distribution showed good agreement with the proportion of test statistics that were expected to fall between the lower and upper limits. Of particular relevance to the present study, Kish and Frankel found that a resampling method performed better than Taylor linearization, which is an analytic method based on approximating the value of a function with a Taylor series. In addition, they found that this performance difference was particularly noticeable for simple correlation coefficients.

The Nonparametric Bootstrap

The present study takes its inspiration from Kish and Frankel's (1974) empirical study and Efron's early work on the bootstrap (Efron 1979; Efron and Gong 1983; Efron and Tibshirani 1986). Efron had proposed the bootstrap as a general method for determining the SE of any estimator. The bootstrap is a computer-intensive method that draws independent samples from the data and calculates the target statistic on each draw. It then uses the resulting empirical distribution to obtain an estimate of the target statistic's SE. The bootstrap's promise of better performance than standard methods was based in part on inconsistencies between the process that generated the data and the assumptions and analytic approximations (e.g., bivariate normality, asymptotic theory) that were used to derive the SE (Davison and Hinkley 1997).

In the 1983 and 1986 papers that introduced the method, Efron used a nonparametric bootstrap to estimate the SE of Pearson's r with a simple random sample of 15 observations (with a nonparametric bootstrap as opposed to a parametric bootstrap, the researcher does not specify a distribution for the data or the quantity of interest). In a subsequent paper (DiCiccio and Efron 1996), Efron discussed the problem of estimating a 90% confidence interval for Pearson's r in

a simple random sample of 20 observations. Recently, Shao (2003) reviewed the impact that survey statisticians working in the 1980s and 1990s have had on the development of the bootstrap. To date, no researcher has published a paper in which a nonparametric bootstrap was used with cluster randomized data to obtain an interval estimate for Pearson's r and test the null hypothesis that the population correlation coefficient equals zero.

A Nonparametric Bootstrap with Cluster Randomized Data

To obtain the bootstrap distribution for Pearson's r in a cluster randomized design with m clusters, the researcher (1) samples m clusters with replacement from the original sample of N values that are nested within the m clusters; (2) calculates r for each bootstrap sample; and (3) repeats this process B times. Within Stata, all of the necessary steps are executed with the following three lines.

```
set seed 7593

corr expect intent

bootstrap "corr expect intent" r = r(rho), cluster(schl1) reps(10000) /* */ saving("e:
\data\bsout10000.dta") replace bca nowarn
```

The first command, **set seed**, sets the seed for the random number generator. This step is recommended because it permits the researcher to reproduce results if the need should arise. The second command, **corr expect intent**, calculates Pearson's r using the data stored in memory. The third statement directs Stata to take 10,000 bootstrap samples of the data, use the correlation command to calculate Pearson's r for the specified variables (positive drug use expectancies and substance use intentions) and write the bootstrap estimate of Pearson's r to the designated file. More specifically, the line instructs Stata to sample clusters with replacement, use the data pairs in the sampled clusters to calculate Pearson's r for the specified variables, and save the returned value of Pearson's r to the file, bsout10000.dta. The existing variable that identifies each cluster is schl1 (i.e., school).

After calculating Pearson's r using Stata's correlation command ("corr"), the estimate is temporarily stored in memory as $r(\rho)$. The Stata user does not have to write much code because Stata supports bootstrapping. For example, Stata determines when the estimates are written to the output file, bsout10000.dta, and handles other bookkeeping tasks. Because Stata handles so much of the heavy lifting, the Stata user only needs to keep in mind the key assumption behind the bootstrap: the obtained sampling distribution has to provide a reasonable picture of what one would observe by repeatedly sampling with replacement from the population and constructing an empirical distribution of $\hat{\rho}$.

Bootstrap Confidence Intervals

As research on the bootstrap accelerated, Efron and other researchers (see Efron 1987) have found it necessary to propose several methods to obtain bootstrap confidence intervals: the standard or normal bootstrap confidence interval, the simple percentile method, the bias-corrected percentile method, and an accelerated bias-corrected percentile method. With the standard or normal bootstrap confidence interval for Pearson's r , the SE is estimated by the standard deviation of the B bootstrapped estimates of $\hat{r}$ (i.e., $\hat{\sigma}$); and the 95% standard or normal bootstrap confidence interval is calculated as $\hat{r} \pm z_{\alpha/2} \hat{\sigma}$ or $\hat{r} \pm 1.96 \hat{\sigma}$. For the normal bootstrap confidence interval to provide accurate coverage, the distribution of the B bootstrapped values of $\hat{r}$ has to be approximately normal; the mean of the B values of $\hat{r}$ has to be ρ (i.e., $\hat{r}$ has to be an unbiased estimator of ρ); and the bootstrap distribution has to provide a good estimate of the standard deviation of the population sampling distribution.

With Efron's simple percentile method of constructing a bootstrap confidence interval, the lower limit is simply the value of $\hat{r}$ that is exceeded by 97.5% of the B bootstrapped estimates; and the upper limit is simply the value of $\hat{r}$ that is only exceeded by 2.5% of the B bootstrapped estimates (one avoids having to do any interpolation to calculate the limits if B is an odd number like 101, 201, 5,001, and so on). If the mean of the B bootstrapped estimates is not close to the population parameter or the bootstrapped distribution is skewed, the actual coverage provided by the simple percentile confidence interval can be quite poor (Efron 1987; Hall 1988; Manly 1997).

To address the poor performance of the simple percentile method when the bootstrapped distribution was skewed, Efron (1987) proposed a bias-corrected percentile method and an accelerated bias-corrected percentile method. With the bias-corrected percentile method, the degree of bias is estimated by taking the difference between the mean of the B bootstrapped estimates and the estimate obtained with the original sample (this simple difference provides an estimate of the bias). Then, the bootstrap distribution is used to determine the confidence interval's lower and upper limits. With the accelerated bias-corrected percentile method, the confidence interval's lower and upper limits are found after correcting for bias and skew. Research has shown that the bias-corrected and accelerated bias-corrected methods often perform better than the simple percentile method. However, many more bootstrap samples are needed to obtain accurate estimates of the lower and upper limits of a $100(1 - \alpha)\%$ confidence interval when either of these two methods is used (Efron 1987).

Number of Bootstrap Samples

Several researchers have addressed the question, "how many bootstrap samples are needed to construct 90% or 95% confidence intervals"? The number appears to have increased over time. Efron and Tibshirani (1986) suggested that the number should lie between 1,000 and 2,000. About 10 years later, DiCiccio and Efron (1996) suggested that 2,000 replications would not be too many if one wanted to estimate confidence intervals. Later, Ukoumunne et al. (2003) cited Efron and Tibshirani (1993) and claimed that 2,000 bootstrap samples would be sufficient to estimate coverage probabilities for 95% confidence intervals with a SE of just under 0.5 percent. However, when Buckland (1984) studied percentile confidence intervals obtained through Monte Carlo simulation, he showed that the actual 95% confidence interval would lie within 93.6–94.6% when B was 1,000, and that it would lie within 94.6–95.4% when B was 10,000. Clearly, 1,000 samples would not be enough if one wanted the lower and upper limits of the confidence interval to be close to the corresponding limits that one would obtain with an infinite number of bootstrap samples (c.f. Hall 1986). In the present study, we let the number of bootstrap samples vary from 100 to 10,000.

Why the Bootstrap Can Perform Poorly

As noted previously, the bootstrap can perform poorly when the resampled distribution fails to provide a reasonable approximation to empirical sampling distribution (c.f. Carpenter and Bithell 2000; Efron and Gong 1983; Manly 1997). However, it can also perform poorly when resampling observations results in bootstrap samples that contain more outliers than there were in the original sample. Moreover, the bootstrap may perform poorly when the target population parameter is on the boundary of the parameter space (e.g., estimation of the mean when the true population mean is 0 and values are restricted to be nonnegative). Andrews (2000) discusses the latter example and provides numerous references to other examples where the nonparametric bootstrap has been shown to perform poorly.

Past Implementations of the Bootstrap with Cluster Randomized Data

Other researchers have used the bootstrap to estimate quantities when the data were not independently and identically distributed. For example, Ren et al. (2006) analyzed data

observed on l regions, m_i families within the i th region, and n_{ij} members from the j th family within the i th region. To assess the prevalence rate for a hepatitis epidemic in China in 1992 as well as estimate intraclass correlation coefficients for family and for region, Ren, Yang, and Lai used a parametric bootstrap with 1,000 bootstrap samples. For their parametric bootstrap, Ren et al. resampled the residuals from the fit of an appropriate multilevel model.

Ukoununne et al. (2003) used simulation to compare different methods for calculating bootstrap confidence intervals for an intraclass correlation coefficient when the observations were drawn from a balanced design. Specifically, the authors simulated multilevel data structures by varying the numbers of clusters (10, 30, 50), the intraclass correlation coefficient (0.001, 0.01, 0.05, 0.3) and the outcome distribution (normal, non-normal continuous). They then compared the performance of five bootstrap confidence intervals as well as bootstrap confidence intervals applied to observations that had been transformed in order to stabilize the variance of the intraclass correlation coefficient. Each bootstrap distribution was based on 2,000 bootstrap samples where clusters were sampled with replacement. Ukoununne et al. found that the standard bootstrap methods only provided coverage levels for 95% confidence intervals that were close to the nominal level when there were 50 clusters. More importantly, they found that application of a bootstrap after applying a variance-stabilizing transformation to the intraclass correlation coefficient improved the performance of the standard bootstrap methods and provided coverage that was close to nominal.

To assess the relation between school achievement and class size, Carpenter et al. (2003) used a nonparametric residual bootstrap. As Carpenter et al. noted, there was no well-established, nonparametric bootstrap for their multilevel structure: The participating students were nested within classrooms that were nested within schools that were nested within educational authorities. Because there were few participating educational authorities, resampling at the highest level of their hierarchy was not feasible. Carpenter et al. fit a multilevel regression model to students' mathematics scores and randomly drew residuals with replacement from a rescaled and centered set of empirical residuals. This approach assumed that their regression model for students' mathematics scores was specified correctly and, moreover, that the conditional variances were homogeneous (the latter assumption ensured that the resampled residuals are exchangeable).

Finally, Field and Welsh (2007) discussed some of the issues that researchers have encountered when they have bootstrapped multilevel data. These issues reflect the data's often complex structure, the different ways that multilevel data can be bootstrapped, and the different ways that the researchers' particular bootstrap can be evaluated. Using simulation, Field and Welsh examined the performance of six different bootstraps. With the cluster bootstrap, they sampled entire clusters with replacement and fit a model to the resulting sample. Field and Welsh concluded that the cluster bootstrap provided a simple resampling scheme that resulted in consistent estimates as the number of clusters increased. However, they cautioned that the cluster bootstrap may not be easily generalized for use with complicated random and mixed effects models.

Study Design

We chose to study the relation between students' substance use expectancies and intentions because prior research with individuals who were older than those participating in the present study had demonstrated that alcohol use expectancies predicted alcohol use behavior (see Stacy et al. 1990 and the many references therein). Indeed, we could have fit a mixed effects regression model, a random coefficients regression model, or a hierarchical linear regression/multilevel model to our cluster randomized data had we simply wanted to determine if there was a linear relation between preadolescent students' positive substance use expectancies and

their intentions to not use substances. However, only by using a nonparametric bootstrap could we obtain a point and an interval estimate of the strength of the association in a metric so well known to prevention researchers. More importantly, as many have noted (c.f. Kelly and Maxwell 2003; Chan and Chan 2004; Kelly 2005; Maxwell 2004), prevention researchers have been asked to pay more attention to providing estimates of an effect's magnitude as well as an estimate of the uncertainty associated with those estimates. Because Pearson's r provides a common metric that facilitates comparisons among effect sizes, it has played an important role in meta analyses (c.f. Rosenthal 1991; Derzon 2007).

Methods

Participants

The data were collected during the first half of the 2004–2005 school year from fifth grade students who were participating in the baseline assessment of an on-going, NIDA-funded, substance use prevention program. The students were attending 29 public schools in Phoenix, Arizona that were participating in the parent study (the participating schools were from seven of ten school districts in the study area. Approximately half of the schools in the districts agreed to participate in the study). The graduate research assistants who were responsible for coordinating the parent study in a specific school initiated the consent process and student recruitment a few weeks before the classroom teachers gave their students consent forms to take home to their parents. Students received “kooky eggs” as an incentive to encourage them to return the consent forms to their classroom teachers in a timely manner. The classroom teachers collected and returned the completed forms to the research assistants. After the research assistants received the completed consent forms, they administered the larger study's baseline assessment and obtained the data that are described in the present study. Approximately 84% of the students in the 29 study schools received their parents' consent, and 96% ($n = 1,934$) of these students provided data.

We used Stata's bootstrap command (Stata Corporation 2005) to resample baseline data from 28 intact groups of fifth grade students; execute Stata's correlation command to obtain an estimate of the correlation between students' positive substance use expectancies and their substance use intentions; and calculate normal, simple percentile, and accelerated bias-corrected 95% confidence intervals that were then used to test the hypothesis that correlation between the two variables equaled zero. We initially varied the number of bootstrap samples in a straightforward manner, obtaining 100, 200, 500, 1,000, 2,000, 5,000, and then 10,000 bootstrap samples. Prior to each run, we set a different random number seed. At the end of each run, we used Stata to obtain a report on our point estimate; an estimate of the SE and the bias; the three bootstrap confidence intervals; and graph our data. Following our first run, when our graphs and tabled data indicated that we had a problem with one or more outliers, we inspected our data, identified the observations from the problem school, and sought to determine if we were successful. To accomplish the latter, we conducted four runs for each of the previously mentioned number of bootstrap samples (i.e., 100, 200, 500, 1,000, 2,000, 5,000, and 10,000). Again, at the end of each of the 28 runs, we used Stata to obtain a report.

Measures

The data were collected with a 104-item questionnaire administered during a 45-minute classroom session. Students could complete the scannable questionnaires in Spanish or English, which was printed on the back of the page. Approximately 9.6% of the students completed the Spanish language version of the questionnaire.

Positive Substance Use Expectancies—Positive substance use expectancies were assessed with three Likert items measured with the same 4-point scale (1 = “Strongly agree,”

2 = “Agree,” 3 = “Disagree,” and 4 = “Strongly disagree”). The items were: “Drinking alcohol makes parties more fun,” “Smoking cigarettes makes people less nervous,” and “Smoking marijuana makes it easier to be part of a group.” Scale scores were calculated by taking the mean of the item scores with increasing values indicating less positive drug use expectancies. Cronbach’s alpha was 0.80.

Intentions to Use Substances—Intentions to use substances were assessed with 3 Likert items measured with the same 4-point scale (1 = “Definitely yes,” 2 = “Yes,” 3 = “No,” and 4 = “Definitely no”). The question stem was “If you had a chance *this weekend*, would you use ____.” The question was completed by asking each student to report their intention to use “... alcohol?” “... cigarettes?” and “... marijuana?” Scale scores were calculated by taking the mean of the item scores with increasing values indicating a stronger intention to not use substances this weekend if one had the chance. Cronbach’s alpha was 0.92.

Results

Figure 1 displays the point estimates and the normal, simple percentile, and accelerated bias-corrected 95% confidence intervals for 1,698 participating fifth grade students’ positive substance use expectancies and substance use intentions when the data include the school that was subsequently identified as an outlier. The figure is remarkable for two reasons. First, at each of the seven bootstrap sample sizes, both the lower and upper limits of the normal 95% confidence interval were slightly larger than the corresponding limits of the simple percentile or accelerated bias-corrected 95% confidence intervals. However, neither the simple percentile nor the accelerated bias-corrected confidence interval was consistently shorter or larger than the normal theory confidence interval despite the fact that the bootstrap distributions had noticeable departures from the symmetric, bell-shaped normal distribution. Second, although the means of the bootstrap estimates for Pearson’s r vary with the bootstrap sample sizes, the means are relatively stable among the sets of confidence intervals. In addition to the information provided by Fig. 1, Table 1 displays estimates of the bias and SEs for the various bootstrap sample sizes.

The departures from the symmetric, bell-shaped normal distribution displayed by the various density plots and the pattern displayed by the means of the bootstrap estimates for Pearson’s r in Fig. 1 suggested that one or more outliers might be present in the data. Figure 2 displays a scatterplot of expectancies and intentions at the school level. It suggests that a number of points may be outliers. We used Stata to fit an ordinary least squares regression to the school-level data and calculate Cook’s distance (Cook 1977). Values of Cook’s D summarize the influence observations have on the slope of the regression line as well as how far observations are from the regression line. Stata’s robust regression command flags observations when Cook’s D exceeds 1.0 (specifically, it assigns a weight of 0 to each extreme point). In the aggregated sample, the three largest values of Cook’s D were 1.444, 0.196, and 0.068. Using the information provided by the graphs and Cook’s D , we dropped the data reported by the 44 students attending school 706 and resampled the remaining observations. With this pass, we took four runs at each of the seven levels (i.e., 100, 200, 500, 1,000, 2,000, 5,000, and 10,000) and used Stata to calculate and report a point estimate for Pearson’s r and a bootstrap confidence interval based on normal theory, the simple percentile method, and the accelerated bias-corrected method. A different random number seed was set for each run.

Figure 3 displays the findings from the 28 runs. In marked contrast to Figs. 2 and 3 shows that there was remarkable stability of the means of the bootstrap estimates for Pearson’s r among the three methods (normal, simple percentile, and accelerated bias-corrected confidence intervals) and across the seven bootstrap sample sizes. Table 2 displays the point estimate, SE,

bias, and the lower and upper limits of the 95% confidence interval at the seven bootstrap sample sizes.

Discussion

The bootstrap is a computer-intensive method that can permit prevention researchers to address questions for which analytic answers may be difficult to obtain. Like traditional methods, it relies on an assumption: the obtained sampling distribution has to provide a reasonable picture of what one would observe by repeatedly sampling with replacement from the population and constructing an empirical distribution. However, unlike traditional methods, the bootstrap does not rely on asymptotic theory.

In the present study, we used a nonparametric bootstrap to obtain an interval estimate of the correlation between fifth grade students positive substance use expectancies and their intentions to not use substances. The issue was not whether these two variables were associated. Prior experimental and observational research has demonstrated that these variables are associated in populations whose fifth grade experiences were far behind them. Then, a fit of any number of competing regression models (aka hierarchical linear models, multilevel regression models, mixed effects regression models, random coefficient models) to the cluster randomized data would have demonstrated that there was a statistically significant linear relation among students' positive substance use expectancies and their intentions to not use substances. However, by using a nonparametric bootstrap, we were able to determine that reasonable point and interval estimates of Pearson's r were 0.46 and [0.40, 0.50] respectively, and that expectancies explained 21% of the variability in students' intentions to use substances.

Contributions of the Present Study

A significant contribution of the present study is that it describes in an accessible manner how practitioners and prevention researchers can use a nonparametric bootstrap to address a difficult analytic problem and obtain an interval estimate for Pearson's r when the data have been obtained with a cluster randomized design. More importantly, the present study illustrates (a) how a nonparametric bootstrap can fail when the number of clusters (schools) is small and an outlier is present among the resampled clusters, and (b) how one can identify when an outlier is present and address the problem.

Conclusions

With the present case study, we sought to demonstrate how straightforward it would be to use a nonparametric bootstrap to obtain an interval estimate of Pearson's r with cluster randomized data and test the null hypothesis that there was no association between 5th grade students' positive substance use expectancies and their intention to use substances in the coming weekend. We had the good fortune to have reasonably well-behaved data. Indeed, the one outlying cluster (school) that was present was not masked by any other data points and, therefore, relatively easy to detect. However, the present study illustrates that even one outlier can be problematic when prevention researchers use a nonparametric bootstrap with cluster randomized data and there are relatively few clusters.

The number of cluster randomized studies has increased at an accelerating rate with each passing decade. Because the audience for an interim or final report may range in statistical sophistication, it would be desirable to provide a familiar summary statistic like Pearson's r to express the magnitude of an association. Until a useful analytic expression is provided for the SE of Pearson's r when the data have been collected with a cluster randomized design, prevention researchers can use a nonparametric bootstrap to obtain satisfactory interval

estimates. However, as simple as it might be to employ a nonparametric bootstrap with cluster randomized data, prevention researchers will have to use the method in a thoughtful manner until further research is conducted that varies the number of clusters, the intraclass correlation, and the heterogeneity among cluster sizes.

Acknowledgments

The project described was supported by Grant Number DA005629 awarded by the National Institute On Drug Abuse to The Pennsylvania State University (Grant Recipient), Michael Hecht, Principal Investigator, with Arizona State University as the collaborating subcontractor. The data used in the present study would not have been available had it not been for the dedication of the Drug Resistance Strategies Project team members in Phoenix, Arizona. These researchers are led by Drs. Flavio Marsiglia, Stephen Kulis, and Patricia Dustman. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health. Finally, we would like to thank Drs. Eric Loken and Michael Rovine for helpful comments and suggestions on the preparation of this article.

References

- Altman DG. Statistics in medical journals: Some recent trends. *Statistics in Medicine* 2000;19:3275–3289. [PubMed: 11113959]
- Andrews DWK. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* 2000;68:399–405.
- Bieler GS, Williams RL. Cluster sampling techniques in quantal response teratology and developmental toxicity studies. *Biometrics* 1995;51:764–776. [PubMed: 7662858]
- Bryk, AS.; Raudenbush, SW. *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage; 1992.
- Buckland ST. Monte Carlo confidence intervals. *Biometrics* 1984;40:811–817.
- Carpenter J, Bithell J. Bootstrap confidence intervals: When? Which? What? A practical guide to medical statisticians. *Statistics in Medicine* 2000;19:1141–1164. [PubMed: 10797513]
- Carpenter JR, Goldstein H, Rasbash J. A novel bootstrap procedure for assessing the relationship between class size and achievement. *Applied Statistics* 2003;52:431–443.
- Chan W, Chan DWL. Bootstrap standard error and confidence interval for the correlation corrected for range restriction: A simulation study. *Psychological Methods* 2004;9:369–385. [PubMed: 15355154]
- Cook RD. Detection of influential observations in linear regression. *Technometrics* 1977;19:15–18.
- Cornfield J. Randomization by group: A formal analysis. *American Journal of Epidemiology* 1978;108:100–102. [PubMed: 707470]
- Davison, AC.; Hinkley, DV. *Bootstrap methods and their application*. New York: Cambridge University Press; 1997.
- Derzon J. Using correlational evidence to select youth for prevention programming. *The Journal of Primary Prevention* 2007;28:421–447. [PubMed: 17990119]
- DiCiccio TJ, Efron B. Bootstrap confidence intervals. *Statistical Science* 1996;13:189–228.
- Donner A. Some aspects of the design and analysis of cluster randomization trials. *Applied Statistics* 1998;47:95–113.
- Donner, A.; Klar, N. *Design and analysis of cluster randomization trials in health research*. New York: Oxford University Press; 1999.
- Efron B. Bootstrap methods: Another look at the Jackknife. *Annals of Statistics* 1979;7:1–26.
- Efron B. Better bootstrap confidence intervals. *Journal of the American Statistical Association* 1987;82:171–185.
- Efron B, Gong G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician* 1983;37:36–48.
- Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science* 1986;1:54–77.
- Efron, B.; Tibshirani, R. *An introduction to the bootstrap*. New York: Chapman & Hall; 1993.

- Field CA, Welsh AH. Bootstrapping clustered data. *Journal of the Royal Statistical Society, Series B* 2007;69:369–390.
- Goldstein, H. Multilevel statistical models. Vol. 2. London: Edward Arnold; 1995.
- Hall P. On the number of bootstrap simulations required to construct a confidence interval. *Annals of Statistics* 1986;14:1453–1462.
- Hall P. Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics* 1988;16:927–953.
- Keen K, Elston RC. Robust asymptotic theory for correlations in pedigrees. *Statistics in Medicine* 2003;22:3229–3247. [PubMed: 14518025]
- Kelly K. The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement* 2005;65:51–69.
- Kelly K, Maxwell SE. Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods* 2003;8:305–321. [PubMed: 14596493]
- Kish L. Confidence intervals for clustered samples. *American Sociological Review* 1957;22:154–165.
- Kish L, Frankel MR. Inference from complex samples. *Journal of the Royal Statistical Society, Series B* 1974;36:1–37.
- Korn, EL.; Graubard, BI. Analysis of health surveys. New York: Wiley; 1999.
- LaVange LM, Keys LL, Koch GG, Margolis PA. Application of sample dose-response modeling ratios to incidence densities. *Statistics in Medicine* 1994;13:343–355. [PubMed: 8177985]
- Levy, PS.; Lemeshow, S. Sampling of populations: Methods and applications. Vol. 3. New York: Wiley; 1999.
- Localio, AR.; Sharp, TJ.; Landis, JR. Analysis of clustered categorical data in an experimental design: Sample survey methods compared to alternatives. *Proceedings of the Biometrics Section, American Statistical Association*; 1995. p. 71-76.
- Manly, B.F.J. Randomization, bootstrap and Monte Carlo methods in biology. Vol. 2. London: Chapman & Hall; 1997.
- Maxwell SE. The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods* 2004;9:147–163. [PubMed: 15137886]
- Murray, DM. Design and analysis of group-randomized trials. New York: Oxford University Press; 1998.
- Myers JL, DiCecco JV, Lorch RF Jr. Group dynamics and individual differences: Pseudogroup and quasi-F analyses. *Journal of Personality and Social Psychology* 1981;40:86–98.
- Ren S, Yang S, Lai S. Intraclass correlation coefficients and bootstrap methods of hierarchical binary outcomes. *Statistics in Medicine* 2006;25:3576–3588. [PubMed: 16416402]
- Rosenthal, R. Meta-analytic procedures for social research. Newbury Park, CA: Sage; 1991. revised edition
- Rosner B, Donner A, Hennekens CH. Estimation of interclass correlation from familial data. *Applied Statistics* 1977;26:179–187.
- Shao J. Impact of the bootstrap on sample surveys. *Statistical Science* 2003;18:191–198.
- Sribney, B. How can I estimate correlations and their level of significance with survey data?. 2001. Retrieved March 06, 2007 from <http://www.stata.com/support/faqs/stat/survey.html>
- Stacy AW, Widaman KF, MarLatt GA. Expectancy models of alcohol use. *Journal of Personality and Social Psychology* 1990;58:918–928. [PubMed: 2348377]
- Stata Corporation. Stata statistical software: Release 9.0. College Station, TX: Author; 2005.
- Ukoununne OC, Davison AC, Gulliford MC, Chinn S. Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Statistics in Medicine* 2003;22:3805–3821. [PubMed: 14673940]
- Walsh JE. Concerning the effect of intraclass correlation on certain significance tests. *Annals of Mathematical Statistics* 1947;18:88–96.

Fig. 1. Normal (*N*), simple percentile (*P*), and accelerated bias-corrected (*B*) 95% bootstrap confidence intervals for 5th grade students positive substance use expectancies and substance use intentions (Outlier School Present)

Fig. 2.
School level scatterplot of positive substance use expectancies and substance use intentions.
Note: The (y_i, x_i) value for School 706 was (1.67, 1.11)

Fig. 3.

Normal, simple percentile, and accelerated bias-corrected 95% bootstrap confidence intervals for 5th grade students positive substance use expectancies and substance use intentions (Outlier School Omitted)

Table 1

Nonparametric bootstrap point and interval estimates of Pearson's r

Number of samples	Point estimate	BiasSE	Normal theory		Simple percentile		BCa	
			LLUL		LLUL		LLUL	
100	0.39	0.010.11	0.170.61	0.140.58	0.140.58	0.140.58	0.140.58	0.140.58
200	0.21	0.000.18	-0.140.57	-0.180.56	-0.180.56	-0.170.55	-0.170.55	-0.170.55
500	0.60	0.000.10	0.400.80	0.380.78	0.380.78	0.330.77	0.330.77	0.330.77
1,000	0.40	-0.010.17	0.070.74	0.010.70	0.010.70	0.020.70	0.020.70	0.020.70
2,000	0.50	0.000.12	0.270.73	0.260.70	0.260.70	0.260.70	0.260.70	0.260.70
5,000	0.51	-0.020.17	0.180.85	0.100.75	0.100.75	0.110.76	0.110.76	0.110.76
10,000	0.54	-0.010.15	0.250.83	0.170.76	0.170.76	0.150.75	0.150.75	0.150.75

LL lower limit, UL upper limit, Bca accelerated bias-corrected confidence interval

Table 2

Nonparametric bootstrap point and interval estimates of Pearson's r with the Outlier School Deleted

Number of samples	Estimate	BiasSE	Normal		Simple percentile		Bca	
			LL	UL	LL	UL	LL	UL
100	0.455	-0.0030.024	0.41	0.50	0.41	0.50	0.41	0.50
100	0.455	-0.0030.029	0.40	0.51	0.40	0.51	0.40	0.51
100	0.455	-0.0020.022	0.41	0.50	0.41	0.50	0.41	0.50
100	0.455	-0.0060.027	0.40	0.51	0.40	0.51	0.40	0.51
200	0.455	-0.0040.026	0.40	0.51	0.40	0.51	0.40	0.51
200	0.455	-0.0010.026	0.40	0.51	0.40	0.51	0.40	0.51
200	0.455	-0.0030.027	0.40	0.51	0.40	0.51	0.40	0.51
200	0.455	-0.0020.025	0.41	0.50	0.41	0.50	0.41	0.51
500	0.455	-0.0020.025	0.41	0.50	0.40	0.50	0.40	0.50
500	0.455	-0.0020.026	0.40	0.51	0.40	0.51	0.40	0.50
500	0.455	-0.0030.026	0.40	0.51	0.40	0.51	0.41	0.51
500	0.455	-0.0030.027	0.40	0.51	0.40	0.51	0.41	0.51
1,000	0.455	-0.0010.026	0.40	0.51	0.40	0.51	0.40	0.50
1,000	0.455	-0.0010.026	0.40	0.51	0.40	0.51	0.40	0.51
1,000	0.455	0.0000.026	0.40	0.51	0.40	0.51	0.40	0.50
1,000	0.455	0.0000.027	0.40	0.51	0.40	0.51	0.40	0.50
2,000	0.455	-0.0010.026	0.40	0.51	0.40	0.51	0.40	0.50
2,000	0.455	-0.0010.026	0.40	0.51	0.40	0.51	0.40	0.50
2,000	0.455	-0.0020.025	0.41	0.51	0.40	0.50	0.40	0.50
2,000	0.455	0.0000.025	0.41	0.51	0.40	0.50	0.40	0.50
5,000	0.455	-0.0010.026	0.40	0.51	0.40	0.50	0.40	0.50
5,000	0.455	-0.0010.026	0.40	0.51	0.40	0.50	0.40	0.50
5,000	0.455	-0.0010.026	0.40	0.51	0.40	0.50	0.40	0.50
5,000	0.455	-0.0010.026	0.40	0.51	0.40	0.50	0.40	0.50
10,000	0.455	-0.0010.026	0.40	0.51	0.40	0.50	0.40	0.50
10,000	0.455	-0.0010.026	0.40	0.51	0.40	0.50	0.40	0.50
10,000	0.455	0.0000.026	0.40	0.51	0.40	0.50	0.40	0.50

LL lower limit, UL upper limit, Bca accelerated bias-corrected confidence interval