Significance Testing For Two Cluster Samples with Identical Clusters and Different Units

Pedro J. Saavedra[1], R. Lee Harding[1] and Ronaldo Iachan[1]
[1]ICF International, 11785 Beltsville Drive, Suite 300, Calverton, MD 20705

Abstract

In comparing groups using data from a cluster sample, the analyst needs to take into account the design effect in order to calculate statistical significance. There are several ways of handling statistically a situation where the same clusters are drawn but different units are selected for two surveys. The simplest seems to be to use a jackknife, creating one set of replicate weights. Then one can produce an estimate of the difference between means or proportions, and use the jackknife to obtain the variance of that estimate. Unlike separate jackknife estimates, this approach can use information about the common sample clusters. This research used simulated data to compare several methods of testing for significant differences in proportions. The methods included jackknifes or Taylor Series that take into account the common clusters in the two samples.

Key words: Jackknife, probabilities proportional to size, Taylor Series

1. Introduction

A common problem for analysts is comparing estimates from two populations or estimates from the same populations across years or surveys. If there is a difference between the two estimates, is the difference significant? One usual way of comparing the two survey estimates is to estimate the confidence interval for a variable of interest, independently for each survey. This is what would have been done if the two sampling procedures had been independent. Typically an assumption of independence is made when comparing estimates from two different surveys. Often the analysts are working with two separate reports or datasets with limited information regarding the sampling procedures. Under these conditions, the assumption of independence seems reasonable.

This research examines the situation were the independence assumptions may not hold up. It explores the implications of not accounting for the lack of independence between the two samples. There are situations where samples of the same population for multiple surveys are "piggy backed" for operational efficiencies. The assumption of independence does not hold in such situations as when, for example, two surveys of high school students are administered at the same time within the same sample of schools.

Some surveys are designed so that the same PSUs are used in consecutive years. Sometimes this is done as a matter of expediency, and sometimes it is done precisely in order to facilitate year-to-year comparisons. But not every comparison takes advantage of the sample design or the common PSUs for the two years.

This research compares the results when the independence is assumed and not assumed. For this study, data from the 2009 and 2010 Florida Youth Tobacco Surveys (FYTS) and the 2009 Florida Youth Substance Abuse Survey (FYSAS) were used.

1.1.Sample Design and Weighting

In 2009, Florida fielded three different surveys to high schools students using the same sampled schools. The desired number of completed surveys in each study was 5,000 students. After inflating for student nonresponse the total sample size was 27,000 targeted students for all three surveys combined. A sample of 80 schools was drawn from all high schools in Florida. Within each sampled school, classrooms were sampled, and students were sampled from each classroom. Each survey was administered to one third of the students in each sampled class. For this research only the FYTS and FYSAS were used. Both the FYTS and the FYSAS used their own edit rules to eliminate cases with too many missing items.

The remaining survey respondents were then weighted. Each survey dataset was weighted independently of the other. The weighting procedure for each dataset was identical. Nonresponse adjustments were conducted at the school, class and student level. The dataset was then poststratified by grade, race and gender to population totals obtained from the Florida Department of Education.

1.2 Common Questions

While the FYTS and FYSAS dataset were chosen primarily because of the sampling method, they were also chosen because they both target risk behaviors among the same population. In particular, they both target high school students in Florida and they both ask questions about tobacco use habits. Most of the questions were very different between the two surveys, but there were a few tobacco questions that could be considered equivalent.

The question was asked as to whether equivalent questions were really equivalent. In other words, whether the estimates derived from the two surveys would yield similar results. One concern was with the slight variations in the wording of questions between the two surveys. A second concern dealt with the editing criteria. A difference may exist due to the removal of students with too many missing items. On one of the surveys, these students were considered nonrespondents.

Two tobacco use variables were chosen from each of the surveys for comparison. One of the two variables yielded large differences, the other did not. For comparison purposes a third variable was created from the variable which yielded marginal differences. The variable was created through simulation, by changing the answer through a random process. The purpose was to illustrate a situation where one approach would yield significant differences and the other would not.

2. Analysis

When comparing the point estimates and testing for significance, the assumption of independence or lack of independence will affect the variance estimate and therefore affect the significance test. In order to obtain variance estimates for either survey, the schools should be seen as primary sampling units. A variance approximation method is necessary due to the complex nature of the sample design. Variances can be calculated

with a jackknife approach that drops one primary sampling unit (PSU) at a time to create replicate weights or the Taylor Series linearization method that uses the PSUs as clusters. Both variance approximation methods produce similar results.

These alternate approaches can be used in a situation where the same clusters are drawn, but different units are selected for two surveys. In estimating the difference between means or proportions, or testing for their significance, the simplest approach seems to be to use a jackknife estimate for the variance, creating one set of replicate weights. This has An advantage that this approach offers over separate jackknife estimates is that the analysis can use information about the common sample clusters. While this approach has similarities to the use of blocking in Analysis of Variance, it permits taking into account the weighting process.

2.1 Comparing Across Surveys

These analyses will use both variance approximation methods. Under the Taylor Series approximation, each school (PSU) is defined as a cluster. Under the Jackknife method each school is defined as a replicate group. We used three methods to create the clusters. The first method (M1) assumes complete knowledge of the sample design. In method 1, each school is defined as a cluster ignoring the survey source. Under the assumptions in method 1, there were 80 clusters defined with students from both surveys. All the students were guaranteed to be from the same school.

The second method (M2) for defining clusters assumes knowledge of the sample design but an inability to match schools between the two surveys. Schools are randomly sorted 1 to 80 on each survey and then each school is assigned a cluster from 1 to 80. Through random chance, a school on one survey may have been combined with the same school in the other survey. Under the assumptions in method 2, there were also 80 clusters, but students in each cluster may or may not have been from the same school.

The third method (M3) for defining clusters assumes complete independence. Each survey-school combination was defined as a cluster. This assumed the clusters were sampled independently, and that there was no link between a cluster in the FYTS and the same cluster in the FYSAS. Method 3 calculated separate variances for each survey to determine if there were significant differences. This method is likely to overestimate the variance of the difference. This phenomenon may not be apparent if one is not aware of the relationship between the surveys. Under the assumptions of method 3, there were 160 clusters.

Using PROC SURVEYFREQ in SAS, a Rao-Scott Chi-Square was computed to test for a significant difference between the three variables chosen for analysis. Only significant results are presented. Table1 below presents the analysis for the variable found to have a highly significant difference between the two surveys. Notice that the difference is significant regardless of the independence assumptions made. While method 1 produced the largest Chi-Square all three methods showed significant results.

Table 1: Rao-Scott Chi-Square Test Results for a Highly Significant Difference

Variance Approximation Method	*PSU Methods*	*Rao-Scott Chi-Square*	*P* <
Taylor Series	Matched by School (M1)	40.3219	.0001
	Random match (M2)	16.4338	.0001
	Independent (M3)	16.8318	.0001
Jackknife	Matched by School (M1)	41.2228	.0001
	Random match (M2)	19.9722	.0001
	Independent (M3)	22.9791	.0001

Table2 presents the analysis for the created variable. The difference between the two surveys was found to be less significant. When the difference is smaller, the assumption of independence does not detect the significance.

Table 2: Rao-Scott Chi-Square Test Results for the Created Variable with a Marginal Difference

Variance Approximation Method	*PSU Methods*	*Rao-Scott Chi-Square*	*P* <
Taylor Series	Matched by School (M1)	5.2448	.0220
	Random match (M2)	2.2990	.1285
	Independent (M3)	2.2863	.1305
Jackknife	Matched by School (M1)	5.2652	.0244
	Random match (M2)	2.7964	.0945
	Independent (M3)	3.0306	.0836

A third set of analyses were conducted using a variable for which there does not seem to be significant differences between the two surveys. None of the Rao-Scott Chi-Square tests were significant for this variable.

2.2 Comparing Two Cycles

Similar to the case where two surveys were administered to the same sample of schools, a single survey could be administered to the same sampled schools at two different points in time. Once again, the assumption of independence would not hold. We used data from two surveys to illustrate this point, the 2009 FYTS, a State level survey, and the 2010 FYTS, a County level survey. The two survey samples were drawn separately but there was some overlap between the two years, 62 of the 80 high schools in 2009 were in the 2010 survey. Treating the two years as part of the same survey with some PSUs overlapping and some not was beyond the scope of this study.

The 62 common schools were selected with a probability equal to the product of the probability in each cycle. Weights for each cycle were divided by the probability of selection in the other and were then adjusted to add to population totals for the cycle, by

grade, race and gender. The Jackknife variance approximation method was chosen for the remainder of the analysis. Since methods 2 and 3 for creating the replicate groups produced similar results in the earlier analysis only M1 and M2 were used in creating the Jackknife replicate groups for the FYTS 2009/2010 dataset. Table 3 below presents the results of the comparison of two variables in the combined 2009/2010 dataset. Again, the analysis included a variable created by simulation to be less significant.

Table3: Matched and Independent Analyses

Variable	*Method*	*Year 09*	*Year 10*	*Difference*	*Std Error*	*Lower 95%*	*Upper 95%*
Variable 1	Matched by School (M1)	0.503	0.391	0.112	0.012	0.089	0.136
	Independent (M3)	0.503	0.391	0.112	0.015	0.082	0.143
Variable 2 simulated	Matched by School (M1)	0.503	0.476	0.027	0.013	.0014	0.054
	Independent (M3)	0.503	0.476	0.027	0.015	- .0032	0.058

The standard errors are consistently higher if the schools are not matched. For a marginal result, one might detect significant differences using the combined database which one might miss using the reported means and standard deviations.

3. Conclusions

The analysis indicates that the incorporation of the common sampling design and common PSUs in a comparison between two surveys can detect significant differences which might otherwise go undetected, but that this is an issue only when the differences are marginally significant. In practice, treatment of the samples as independent will lead to conservative results, and may be associated with a Type II error.

This approach is particularly useful when two school surveys use the same schools, but different students. In Our analyses did not consider the situation where the school samples overlap partially, but not completely, When the sample design makes use of the same PSUs, however, it makes sense to use this information if available.