Estimating Degrees of Freedom for Data from Complex Surveys

Allison M. Burns, Robert J. Morris, Jun Liu, Margaret Z. Byron
RTI International, Research Triangle Park, NC 27709

Abstract

Estimating the appropriate degrees of freedom when conducting tests using data from a complex survey is not trivial. The conventional method of the number of PSUs minus the number of strata may substantially overestimate the degrees of freedom when the domain of interest is rare. We discuss an alternative approach and compare it to the conventional method using simulations.

Introduction

In the analysis of complex survey data, comparisons usually involve t-tests. One basic issue involved in these tests is the calculation of degrees of freedom. A typical two-stage complex survey design has primary sampling units (PSUs) selected with probability proportional to size (PPS) independently within first-stage strata and secondary sampling units (SSUs) selected independently within PSUs using simple random sampling (SRS). For a stratified selection of PSUs, the conventional method of calculating degrees of freedom is to use the number of sampled PSUs minus the number of strata (Korn and Graubard 1999). We call this method the "fixed PSU method".

There is, however, a potential problem with this method of calculation of degrees of freedom. Of concern is that this method could overestimate the degrees of freedom when the domain of interest is rare. For example, a sampled PSU may have no observations in the domain of interest yet that PSU is being taken into account in the calculation of degrees of freedom. Due to this, for a study with a small sample size, the degrees of freedom could be larger than the sample size because there are PSUs that have no observations for a certain domain. In this paper, we present an alternative method for calculating the degrees of freedom for a t-test for complex survey data and compare this to the conventional fixed PSU method, using a series of simulations.

Calculations

An alternative method for calculating degrees of freedom is what we call the "variable PSU method". It is a modification of the fixed PSU method recommended by Korn and Graubard. This calculation is the number of sampled PSUs with sampled observations in the domain minus the number of strata with sampled observations in the domain (Korn and Graubard 1999). The adjustment of using only sampled PSUs and strata with sampled observations in the domain accounts for situations when some sampled PSUs have no observations in the domain. Therefore, this method is more appropriate for studies with rare domains; it should produce the same number of degrees of freedom as the fixed PSU method when the domain is common. Additionally, this method will vary across domains whereas the fixed PSU method will not.

Methods

To compare the two methods of calculating degrees of freedom empirically, we used simulations. Our simulations mimicked the sampling procedures of the 2002 Community Health Center (CHC) User and Visit Study. This study was funded by the Health and Human Resources and Services Administration (HRSA) of the U.S. Department of Health and Human Services. Primary sampling units were grantees which were selected using PPS sampling. Grantees are the administrative entities by which a community health center receives Section 330 funding from the Bureau of Primary Health Care (BPHC) to provide primary medical care in the United States or its territories (Carcagno et al 1997). Secondary sampling units were users within each chosen grantee selected using SRS. Samples of 25 grantees with 30 users per grantee were chosen in each iteration of our simulation.

To create a sampling frame from which to take these simulated samples, we generated a dataset with a structure and characteristics that were similar to that of the Uniform Data System (UDS) database maintained by the BPHC. The UDS database has characteristics of each grantee including total number of users and percentage of users by gender, race, income, and age. Once we had our samples, we used the UDS frame information to randomly classify each user into demographic categories with probabilities based on the demographic distribution in their grantee. Each user was assigned an indicator of certain health conditions using a logistic model based on the 1995 Community Health Center (CHC) User and Visit Study.

In each of 10,000 iterations, the null hypothesis was that the simulated sample mean from that iteration would equal the estimated population mean, which was simulated in an earlier, independent set of 10,000 iterations. For each simulated sample, within each domain and for each health indicator, we calculated a t-statistic as:

$$T = (\bar{x} - \mu) / \text{s.e.}(\bar{x}),$$

where $\bar{x}$ was the sample mean, μ was the estimated population mean, and s.e. ($\bar{x}$) was the standard error of the sample mean. The standard error of the sample mean was calculated using SUDAAN, which accounts for the stratification, clustering, and unequal weighting introduced by complex surveys (Research Triangle Institute 2001). We considered three domains of interest: people at less than one hundred percent federal poverty level, Asian/Pacific Islanders, and Hispanics over the age of 65. We were interested in the poverty domain because we would expect users in this group to be evenly distributed across grantees, with observations in each grantee. Since this domain is common, we would expect the fixed PSU method of calculating degrees of freedom to produce the same number of degrees of freedom as the variable PSU method. Of greater interest were the Asian/Pacific Islander and Hispanic domains because these are considered rare domains; for these domains there may not be sampled observations in each grantee. We would expect the fixed PSU method to overestimate degrees of freedom for these domains.

For each t-test, we calculated degrees of freedom using two methods: fixed PSU and variable PSU. In both cases, we enforced a lower bound of 1 on the degrees of freedom. For each method, we identified the region from the appropriate t-distribution where the null hypothesis would be rejected at the 5% level. We then compared each calculated T to these rejection regions to determine what percentage of them were rejected. Since the null hypothesis is true in expectation, it should be rejected 5% of the time after many iterations. If it is rejected more than 5% of the time, we may conclude that the degrees of freedom was overestimated.

Results

Table 1 shows the mean degrees of freedom for each health indicator, by domain. The degrees of freedom for the fixed PSU method is the same for all domains and health indicators, which must be true by definition. The degrees of freedom for the variable PSU method is constant across health indicators within a domain, but, unlike the fixed PSU method, varies across domains. The variable PSU method is similar to the fixed PSU method for the common domain, and produces much smaller degrees of freedom for the rare domains.

Table 2 shows the percentage of time we rejected the null hypothesis using each method of calculating degrees of freedom, by domain and health indicator. Recall that we expected the fixed PSU method to overestimate the degrees of freedom for the rare domains (Asian/Pacific Islanders and older Hispanics). From Table 2, we see that the rejection rate for the fixed PSU method is greater than 5% for all domains, implying that it is overestimating the degrees of freedom in all cases. However, the level of overestimation is very high for the rare domains, while it is only slight for the common domain. Comparing our alternate method to the fixed PSU method, we see that both methods are equivalent for the common domain, each producing rejection rates slightly higher than 5%. For the rare domains, the fixed PSU method has very large rejection rates, while the variable PSU method has much lower rejection rates. For example, for Asian/Pacific Islanders with coronary heart disease, the rejection rate for the variable PSU method was 4.5% compared with 14.4% for the fixed PSU method. The rejection rates for the rare domains for the variable PSU method were closer to 5%, ranging from 3.3% to 11.0%. By comparison, the rejection rates of the fixed PSU method ranged from 14.4 to 53.0%.

Conclusions

As expected, for common domains, the t-test results are quite similar for both methods. Therefore, there appears to be no advantage in using one method over the other when testing means for domains found commonly in the population. However, for rare domains, the degrees of freedom may be small enough that changing it could significantly alter the t-test results. In our results for the rare domains, the fixed PSU method produced larger degrees of freedom than the variable PSU method. This resulted in t-test rejection rates suggesting that the fixed PSU method overestimates the degrees of freedom. Therefore, for domains that are rare in the population, it appears that using the variable PSU method is most appropriate because it may reduce the risk of overestimating the degrees of freedom. More research is needed to study whether using different domains or

different sampling designs produces similar results.

References

Carcagno, G.J., Carlson, B.L., Cohen, R., and Eisenhower, D. (1997) *Community Health Center User and Visit Study: Survey Methodology Report*. Princeton, NJ: Mathematica Policy Research, Inc.

Korn, E.L., and Graubard, B.I. (1999) *Analysis of Health Surveys*. New York: John Wiley and Sons, Inc.

Research Triangle Institute (2001). SUDAAN User's Manual, Release 8.0. Research Triangle Park, NC: Research Triangle Institute.

Table 1. Degrees of Freedom Comparison

Domain by Health Indicators	Variable PSU Method	Fixed PSU Method
Poverty Level (< 100%)		
Hypertension	16.9	17.2
Diabetes	16.9	17.2
Coronary Heart Disease	16.9	17.2
Asian Pacific Islanders		
Hypertension	2.2	17.2
Diabetes	2.2	17.2
Coronary Heart Disease	2.2	17.2
Hispanics of Age > 65		
Hypertension	1.0	17.2
Diabetes	1.0	17.2
Coronary Heart Disease	1.0	17.2

Table 2. Null Hypothesis Rejection Rates

Domain by Health Indicators	Variable PSU Method	Fixed PSU Method
Poverty Level < 100%		
Hypertension	6.4%	6.5%
Diabetes	7.6%	7.7%
Coronary Heart Disease	10.9%	11.0%
Asian Pacific Islanders		
Hypertension	6.7%	23.3%
Diabetes	6.2%	19.9%
Coronary Heart Disease	4.5%	14.4%
Hispanics of Age > 65		
Hypertension	6.9%	40.9%
Diabetes	11.0%	53.0%
Coronary Heart Disease	3.3%	18.1%

`