Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 21

11-1-2003

On Treating A Survey Of Convenience Sample As A Simple Random Sample

W. Gregory Thatcher
University of West Florida, wthatcher@uwf.edu

J. Wanzer Drane
University of South Carolina

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 478-480

1538 – 9472/03/$30.00

On Treating A Survey Of Convenience Sample As A Simple Random Sample

W. Gregory Thatcher
Department of Health
University of West Florida

J. Wanzer Drane
Department of Epidemiology & Biostatistics
University of South Carolina

Threat of bias has kept many from using data gathered in less than optimal conditions. We maintain that when convenience sampling represents race and gender at nearly correct proportions and can be beneficial, as these two variables are quite often used as stratification variables. We compared a convenience sample with a proven sample. Race and Sex were nearly proportional as was found in the proven sample. We conclude that the convenience sample can be used as though it is simple random.

Key words: Simple random sampling, convenience sampling

Introduction

From the first semester of Introduction to Statistics through our career as scientists by whatever names, we are warned of the sampling and non-sampling errors and how to overcome them. Recently a question was asked: “May I treat my convenience sample as a simple random sample?” To answer the question we employed a sample of known qualities, SCYRBS99, the South Carolina Youth Risk Behavior Survey of 1999.

Representative coverage of Gender and Race is paramount, if the sample is to be instructive when formulating health policy, and we know that SCYRBS99 and earlier YRBS samples are constructed so that the estimates of prevalence among these two variables, as well as others, are nearly unbiased (CDC, 1999).

If we can show that the estimates of the percentages of gender and race are nearly the same in the convenience sample as are in the weighted estimates of the YRBS sample of the same year, then we can at least increase our confidence in the treatment of our sample as simple random. Such a comparison does not, nor will it ever, PROVE the convenience sample to be totally unbiased and simple random, but it will go a long way toward our believing the prevalence calculated are nearly unbiased.

Results

The estimates of gender and race prevalence will be compared to those obtained from the SCYRBS99 sample, which are treated as population constants. Tables 1 and 2 display those values together with the estimates from the convenience sample.

Remembering that X^2 is directly proportional to the sample size, which is 4421 in this case, then a Chi-square of 9.43 is not large at all. In order to reach a significance of only 0.05, N had to be at least (4421/9.43)*3.84) = 1800. This is a case in which we have too much power. From an administrative point of view we would require alpha to be equivalent to about four standard errors or 0.0001. Therefore, we are able to accept a difference of 46.66-44.36=2.30% as non-significant and administratively not important. Further, we can treat this sample as a simple random sample.

W. Gregory Thatcher, Department of Health, Leisure and Exercise Science. University of West Florida, 11000 University Parkway Pensacola, FL 32514. Phone: 850-474-2598, Fax: 850-474-2106. Email: wthatcher@uwf.edu

Table 1: SCMS (Convenience Sample) with expected percentages and numbers obtained from SCYRBS99. Variable = GENDER. Expected F = P (F|SCYRBS99)*4733. $X^2 = (2409-2376.91)^2/2376.91 + (2324-2356.09)^2/2356.09 = 0.87$, df = 1, p-value = 0.35.

GENDER	SCMS Percent Count	SCYRBS99 Percent Expected number
F	50.90	50.22
	2409	2376.91
M	49.10	49.78
	2324	2356.09
Total	4733	4733

Table 2: SCMS (Convenience Sample) with expected percentages and numbers obtained from SCYRBS99. Variable = RACE. Expected B = P (B|SCYRBS99)*4733. $X^2 = (1961-2022.41)^2/2022.41 + (2460-2310.65)^2/2310.65 + (312-399.94)^2/399.94 = 30.85$ df = 2, p-value = 0.0000002.

RACE	SCMS Percent Count	SCYRBS99 Percent Expected
B	41.43	42.73
	1961	2022.41
W	51.98	48.82
	2460	2310.65
O	6.59	8.45
	312	399.94
Total	4733	4733

Table 3: A repeat of Table 2 with the O category excluded. Expected B = P(B|SCYRBS99)*4421. $X^2 = (1961-2062.84)^2/2062.84 + (2460-2358.16)^2/2358.16 = 9.43$, df=1, p-value = 0.0021.

RACE	SCMS Percent Count	SCYRBS99 Percent Expected
B	44.36	46.66
	1961	2062.84
W	55.64	53.34
	2460	2358.16
Total	4421	4421

Between female and male distribution the convenience sample is right on target, but the p-value of the chi-square among the three racial groups indicates a noticeable difference. An examination of actual count versus the expectations show there is an excess of white students at the expense of those captured as ‘O’ or other than Black or White. If those are omitted, as usually is the case because of small numbers in more complex analyses, we have the results in Table 3.

Conclusion

The convenience sample has nearly the same gender and racial compositions as is estimated from the SCYRBS99 data. It can then be treated as a simple random sample. For the skeptic or purist, caution should be used when generalizing across racial lines when using the SCMS data.

If stratification is made along the four categories (B,F), (B,M), (W,F) and (W,M), estimates within category should be nearly unbiased. From those four strata, comparisons could still be made without hesitation. If you insist on a larger alpha, then the RACE variable should not appear in a regression, linear or logistic, in conjunction with a set of risk and confounder variables.

References

Centers for Disease Control and Prevention. (1999). Division of Adolescent and School Health Youth Risk Behavior Survey, 1999. http://www.cdc.gov/nccdphp/dash/yrbs/