Analyses Based on Combining Similar Information from Multiple Surveys

Georgia Roberts[1], David Binder[2]
[1]Statistics Canada, Ottawa Ontario Canada K1A 0T6
[2] Statistics Canada, Ottawa Ontario Canada K1A 0T6

Abstract

Many researchers have access to different survey sources, each with similar variables. These researchers are often interested in the appropriateness of bringing together the data from the different sources for the purpose of data analysis, particularly when each source has a small sample size for the question being studied. We address a variety of topics that the researchers should be aware of such as the comparability of the variables across surveys, and the suitability of positing a model for the variables in the different surveys. We discuss possible approaches for combining the information.

Key Words: Pooling, target population, design-based analysis, model-design-based framework

1. Introduction

With the increasing availability of more than one survey containing the same or similar variables, more attention is being paid to whether and how to combine the data from the different surveys to improve estimates. It seems reasonable to think that one should usually be able to improve the estimate of a quantity of interest (with respect to either accuracy or precision) by combining the samples, *provided that an appropriate approach is used to form the new estimate*. However, which method is appropriate is not always clear.

There are several reasons why analysts would want to combine the data from two or more surveys. A major reason is that the sample sizes for the phenomenon under study are small in each of the data sources either due to each survey having a small sample size or due to the domain of interest being rare in the population(s) targeted by each of the surveys. The combining of samples for increasing number of observations is used not only in the case of separate surveys, but also for combining rolling samples of the same survey and for combining data from overlapping panels in a repeated panel survey. In all cases, it is expected that increasing the overall sample size should lead to reduced sampling errors.

Having small sample sizes in each of the data sources is not the only reason for wanting to combine the data from two or more surveys. Instead, an analyst may wish to bring together the data from periodic surveys on the same topic in order to estimate change. Or, in cases where there may be frame deficiencies, combining surveys with similar variables using multiple frame methods may be used to improve the coverage. As well as the coverage problem, Schenker and Raghunathan (2007) discuss other types of non-sampling errors for which combining information from multiple surveys could be beneficial.

Underlying all of the reasons given above for wanting to combine is a common problem - the data from any single survey are limited in some sense for addressing the analytic problem at hand. However, combining the data from more than one source raises a number of issues that need to be addressed before reasonable decisions may be made on whether and how estimation can be carried out using the different sources. The first of these is the comparability of the information obtained from the different surveys. Schenker et al. (2002) and Schenker and Raghunathan (2007) discuss a number of potential sources of incomparability that could affect whether variables recorded in different surveys are actually measuring the same quantities: differences in the types of respondents and/or the sources of the respondents' information, differences in the modes of interviewing, differences in the survey contexts, differences in the sample designs and differences in survey questions. However, an additional important comparability question relates to how the target populations of the data sources compare: whether they are similar for both target group[1] and time, whether the target groups are similar but times differ (which is the most common case), or whether they differ substantively with respect to both target group and time.

In this paper we address a number of other topics that are important to making decisions on how estimation can be carried out using the different survey sources. In Section 2 there is a description of the three main types of quantities that analysts are interested in estimating from combined data sources of similar variables. Section 3 begins with definitions of the two usual approaches to estimation when combining data from multiple surveys and then provides descriptions of randomization frameworks that could assist an analyst in deciding which approach might be most suitable for which type of quantity of interest. An illustration of combining the data from two Canadian health surveys is described in Section 4, with the paper finishing with a number of points of discussion in Section 5.

2. Three Types of Quantities of Interest

Before introducing the general randomization frameworks within which the properties of various estimators can be discussed, we present three general categories of what is frequently estimated from data arising from multiple surveys. In each case, we consider the unknown quantities that are being estimated and to which target population these quantities refer. It should be noted that the analyst's target population has two components – the target group (i.e., the attributes of the units being targeted) and the reference time(s) (for example, a single time point such as December 31, 2008 or a number of time periods such as both 2004 and 2005).

1) *Simple Descriptive*: We say that the quantities of interest are simple descriptive when they are characteristics of a single finite target population or when they are a fixed function of the characteristics of more than one finite population. Finite population characteristics are quantities such as means, proportions and totals.

There would be a single finite target population, for example, if all surveys being combined covered the same target group at the same point in time. A single finite

[1] The target group is defined as the set of units having the targeted attributes – say, females aged 25 to 34 living in California.

population would also be the case when the surveys being combined each target a different piece of the full target population and the quantities of interest.

If each survey refers to a different finite population, we may be interested in a simple or weighted average of the characteristics of the different populations. For example, if the prevalence of a disease is P_1 and P_2 in populations 1 and 2 respectively, our quantity of interest may be $(P_1 + P_2)/2$. In some cases, we may prefer a weighted average, such as weighting by population size, so that our quantity of interest is $(N_1P_1 + N_2P_2)/(N_1 + N_2)$, where N_1 and N_2 are the respective population sizes. Other weighted averages can also be considered. Another form of a descriptive characteristic is the difference between two population means - $(\bar{Y}_2 - \bar{Y}_1)$. Note that for any of these examples, the two populations could involve entirely different population groups (say, different age groups) or they could be the same population group at different points in time. It should also be noted that, in the case of simple descriptive quantities, the characteristics of interest are defined without a model justification.

2) *Descriptive under an assumed relationship*: Rather than being a simple descriptive quantity, it is not uncommon that the parameter of interest is based on an assumed relationship among the characteristics of the finite target populations of the different surveys. For example, we could suppose that the prevalence rate of a particular disease is the same for each population and it is this common rate that we wish to estimate. Another example would be the case where we want to estimate a quantity for a time point that is midway between two survey periods; assuming a linear trend over time, the quantity of interest would be a simple average of the individual population quantities.

3) *Analytic quantities*: When the quantities of interest are characteristics or relationships that hold beyond the specific finite populations surveyed (such as the parameters of a superpopulation), we say that these quantities are analytic. Often parameters of a model are used to summarize such characteristics or relationships; for example, a logistic model might be used to describe a prevalence rate that is measured in each survey.

3. Approaches to Estimation

As we have described in Section 2, when combining similar information from multiple surveys, we need to first consider which population quantity is being estimated. Once this is established, the properties of estimators should be assessed in the context of the randomization framework for selecting the sample. We first describe the two usual approaches to estimation when combining data from multiple surveys.

The separate approach: In the separate approach to estimation, an estimate is obtained from each survey separately, and then the overall estimator is a function of the separate estimates. The most common method here is to take some linear combination of the separate estimates to form the overall estimator. The particular linear combination chosen can depend on whether the quantity of interest is descriptive or analytic. The linear combination can also depend on whether the separate survey estimates are independent, and whether one can achieve an adequate reduction in the variances of the overall estimate for the most important quantities of interest. (Note that in a multipurpose survey there are usually several quantities that the researcher wishes to estimate.)

As an example of the separate approach, suppose that $\hat{\theta}_1$ and $\hat{\theta}_2$ are unbiased estimates of the same unknown descriptive parameter θ from each of two surveys, and that $\hat{\theta}_1$ and $\hat{\theta}_2$ are independently distributed with known variances σ_1^2 and σ_2^2, respectively. The separate approach estimator $\hat{\theta}_c = \alpha\hat{\theta}_1 + (1-\alpha)\hat{\theta}_2$ will be unbiased for θ regardless of the value of fixed composite weight α and will have minimum variance when $\alpha = \sigma_2^2 / (\sigma_1^2 + \sigma_2^2)$. If $\sigma_i^2 = \sigma^2 / n_i$, where n_1 and n_2 are the respective survey sample sizes, the minimum variance estimator is $\hat{\theta}_c = (n_1\hat{\theta}_1 + n_2\hat{\theta}_2)/(n_1 + n_2)$.

The pooled approach: On the other hand, in the pooled approach, the individual records from all the surveys are combined, the original weights may be modified, and estimation is based on the pooled sample using the new weights and using techniques appropriate to a single sample. Typically, for the observations in each individual survey, the modified weights are proportional to the original weights. The choice of rescaling factors can depend on criteria similar to those used for choosing a linear combination in the separate approach.

Now, to study the properties of these approaches, we need to establish which randomization process led to the observed data. This is not necessarily straightforward, especially when each survey is taken at a different time point.

3.1 Design-based Randomization

For estimating a descriptive quantity, it is common to assume that the underlying randomization framework is design-based. This means that statistical inferences (such as construction of confidence intervals and performing tests of hypotheses) are based only on the probabilities used to select the samples from the finite populations. For the separate approach to estimation, in Figure 1, we illustrate the design-based framework when there are two finite populations (where these populations could overlap). We see that the samples are taken from each of the two populations, separate estimates are formed from each, and then an overall estimate, based on these separate estimates, is derived.

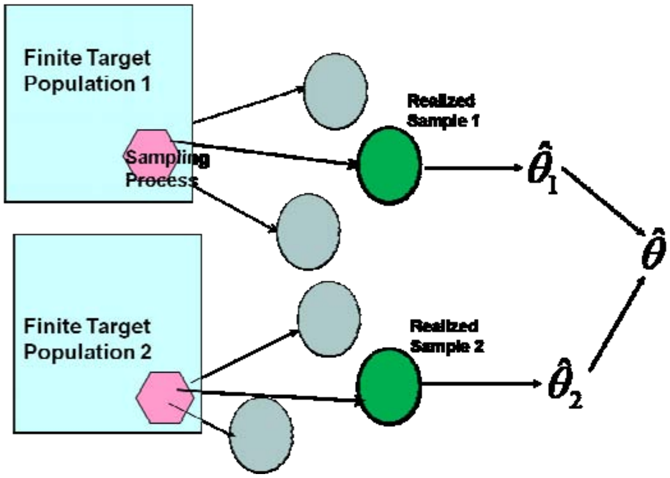

Figure 1: Descriptive estimation – separate approach

The pooled approach in the design-based framework is illustrated in Figure 2. In this approach, the samples from each of the surveys are combined into one large sample, possibly with some weight adjustments, and an overall estimate is obtained from the pooled data. Again, in a design-based framework, the only randomness is the sample selection process for each of the finite populations.

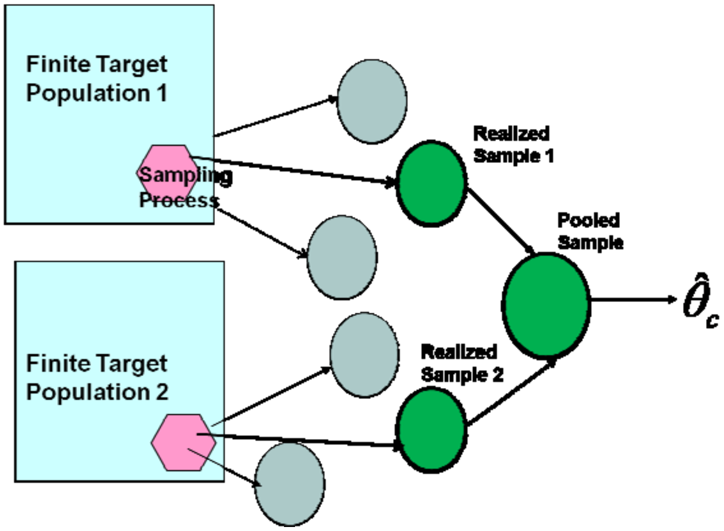

Figure 2 Descriptive estimation – pooled approach

In general, the pooled approach and the separate approach lead to different estimates. These estimates may not even have the same expected values. For example, if the prevalence of a disease in each of two populations is estimated by $\hat{P}_1$ and $\hat{P}_2$, using the samples from each, a separate approach might be to take the simple average of these estimates, which has an expected value of $(P_1 + P_2)/2$. On the other hand, if we take an analogous pooled approach, and rescale the weights of the observations from each sample by $1/2$, the quantity being estimated would be $(N_1P_1 + N_2P_2)/(N_1 + N_2)$. Unless $N_1 = N_2$, the separate and pooled approaches are estimating different quantities. On the other hand, if, under a model, both P_1 and P_2 are measuring a common overall prevalence rate, then the separate and pooled approaches are estimating the same prevalence rate.[2]

A variety of methods has been proposed for rescaling weights for use with combined surveys (see, for example, Korn and Graubard, 1999). One approach adopted by some (see, for example, Thomas (2007)) is to rescale the weights by the factor n_i / D_i for the *i*th survey, where n_i is the sample size and D_i is some "average design effect" for the *i*th survey. This rescaling is motivated by the fact that for the separate approach this can yield minimum variance estimates when the individual survey estimates are unbiased.

[2] In more complex cases, such as the fitting of regression models, there are analogous differences between the separate and pooled approaches.

However, for the pooled approach, if the population sizes are very different, this may not be best, even when the design effects for all the quantities are equal within each survey. As an example, suppose that $\hat{P}_i$ is an unbiased estimate from the *i*th survey of a common overall prevalence rate P, with variance $D_i P(1-P)/n_i$. A pooled approach estimate using two samples and rescaling factor n_i / D_i is

$$\hat{P}_c = \frac{n_1 N_1 \hat{P}_1 / D_1 + n_2 N_2 \hat{P}_2 / D_2}{n_1 N_1 / D_1 + n_2 N_2 / D_2}\,. \tag{3.1}$$

On the other hand, if the separate approach estimator is defined as $\hat{P}_c^{(\alpha)} = \alpha \hat{P}_1 + (1-\alpha)\hat{P}_2$, the optimal value for the composite weight α would be

$$\alpha_{opt} = \frac{n_1 / D_1}{n_1 / D_1 + n_2 / D_2}\,.$$

yielding the minimum-variance separate approach estimator

$$\hat{P}_c^{(\alpha)} = \frac{n_1 \hat{P}_1 / D_1 + n_2 \hat{P}_2 / D_2}{n_1 / D_1 + n_2 / D_2}\,. \tag{3.2}$$

Therefore, if the population sizes are very different, the minimum-variance separate approach estimate given by (3.2) will not be close in value to the pooled estimate in (3.1). Also, as is typical when combining surveys, if the sample size for each survey is small, the estimates of the design effects may not be very accurate for either approach.

3.2 Model-design-based Randomization

Often the quantity of interest to a researcher can be formulated in terms of parameters of a model. For example, the probability of being diagnosed with a particular disease may be thought of as an outcome from a logistic regression model. In this case, a suitable randomization framework for statistical inference may be given by assuming that (i) the study variables in each finite population are realizations of random variables of a model, and (ii) a probability-based sample is selected from each resulting finite population. This is illustrated in Figure 3.

Whereas our previous description of separate and pooled estimates was given in the context of estimating finite population quantities, these estimates (and others that are model-motivated) can be assessed under a model-design-based framework. As pointed out by Binder and Roberts (2003), when the sampling fractions are small, weighted estimates can be used to obtain model-design-based (approximately) unbiased estimates for the model parameters of interest. However, when the sample size (or the number of psu’s in the case of a multi-stage survey) is not large, care may be required in making appropriate inferences.

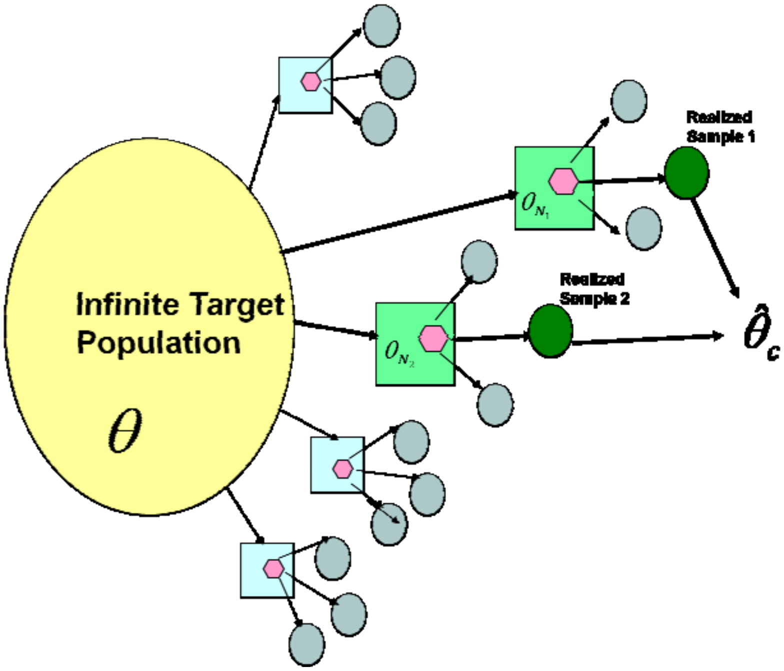

Figure 3: Analytic study – a model-design-based view

4. Use of Health Care for Non-heterosexual Males – An Example from the Canadian Community Health Survey

Suppose that an analyst is interested in studying whether gay and bisexual men differ in their use of health care. Two surveys are proposed as data sources for the analysis – the Canadian Community Health Surveys of 2003 and 2005[3] - since the sample sizes of gay and bisexual men are relatively small in each survey. These are independent cross-sectional surveys of the non-institutional Canadian population aged 12 and over. Both surveys contain the following same question that would identify people aged 18 to 59 who self-report as being homosexual or bisexual: “Do you consider yourself to be heterosexual (sexual relations with people of the opposite sex), homosexual, that is lesbian or gay (sexual relations with people of your own sex), or bisexual (sexual relations with people of both sexes)?” As well, both surveys contain the same set of socio-demographic and health-related variables that the analyst would like to use in his study. The two surveys also seem comparable with respect to other aspects that could influence results, such as sample designs, survey questions and modes of interviewing.

Since the surveys occur just two years apart, the analyst initially expects that the characteristics of his target group should be very similar at the two time points, and that he will be able to make assumptions of equality of characteristics when estimating descriptive quantities. However, when he does some initial investigation of his two data sources, he finds that, while responding sample sizes overall and of males 18-59 are quite similar at the two time points, sample sizes of gay and bisexual men are up 25% and 12%

[3] See Béland (2002) and the Statistics Canada website (www.statcan.gc.ca) for more information about these surveys and also Tjepkema (2008), for a motivating study for this example.

in 2005, as compared to 2003. As well, while estimated population size of males 18-59 is fairly steady, estimates for gay and bisexual men are up approximately 20%. (See Table 1) Furthermore, the estimated distributions of some demographic characteristics (see Table 2) appear to differ more than what might be expected if the target groups actually are the same. In particular, there are higher estimated percentages in the older age groups and in the married/common-law category for both gay and bisexual men in 2005 and the estimated regional concentrations differ between years. Because of these observations, the analyst should suspect differences in his target groups at the two times[4], and thus be wary of making assumptions of equality over time periods when doing his estimations.

Since the objective of the analyst is to study the use of health care in his target group, consider, now, an investigation of whether gay men differ from bisexual men in their probability of not having a regular doctor. The estimated percentages without a regular doctor in 2003 and 2005 respectively were 24 and 20 for gay men and 33 and 21 for bisexual men. The decision is made to take a pooling approach for model estimation. The analyst prepares a data file that includes the observations from both time points and a weight variable that consists of the unmodified weights of the original surveys. Also included on the file are the additional variables required for variance estimation, which will be straightforward since the two surveys are independent. If the analyst should then fit a logistic model to his data, including just a 0/1 time indicator and a 0/1 gay/bisexual indicator he would obtain the results illustrated in Table 3. It appears as if the probability of not having a regular doctor does differ between the two time periods but no significant difference is found between gay and bisexual men. This significant time difference would have been missed if the analyst had pooled the data and ignored the 2 sources in his analysis.

Table 1: Sample sizes and population estimates from the two surveys

	2003		*2005*	
	Sample Size	*Population Estimate*	*Sample Size*	*Population Estimate*
Both sexes	134,072		132,947	
Males 18-59	39,299	9,412,400	38,936	9,507,300
Gay	490	118,400	613	141,600
Bisexual	235	54,200	263	64,500

[4] In fact, there was a series of changes in provincial and federal legislation over the 2003 to 2005 time period that gave same-sex unions legal recognition and a number of other rights. These events might have had an impact on who was willing to self-identify as gay or bisexual.

Table 2: Age, region and marital status breakdowns of target groups

	Gay men		Bisexual men	
	2003	***2005***	***2003***	***2005***
Age 18-24	11	9	28	20
25-34	26	20	24	13
35-44	38	35	20	24
45-59	26	36	28	42
	100	100	100	100
Toronto/Montreal/Vancouver	49	62	47	47
Other CMA	34	24	32	21
Not CMA	17	15	21	34
	100	100	100	100
Married/Common law	26	37	29	49
Previously married	3	5	6	8
Single	71	59	64	43
	100	100	100	100

Table 3: Estimated logistic model

Variable	**Coefficient**	**p-value of t-test**
Intercept	-0.84	
Gay/Bisexual	-0.22	0.24
2003/2005	-0.36	0.04

5. Discussion

Combining of similar data from more than one survey does seem to be a possibility in a number of different situations, but often it is not appropriate. However, when appropriate, care is required in determining an approach that is suitable and has the intended properties. A number of points to keep in mind are outlined briefly below.

As noted in Section 3.1, it may be possible to estimate the same quantity by separate and pooled approaches, but the estimates themselves may not be the same. Furthermore, the "best" composite weighting for the separate approach estimate is different from the "best" survey-weight adjustment for the pooled estimate (where "best" means minimum variance). However, you cannot actually calculate the minimum variance separate estimator since you do not know the variances required for that estimator and frequently you cannot estimate them well if you have only small sample sizes from each survey source. Regardless of the sample sizes, however, using estimated variances in the estimates affects their mean values and variances

A suitable way to apply the separate approach for a vector of quantities of interest (such as a vector of model coefficients) does raise questions. What composite weighting makes sense? Should each component of the vector be weighted equally? More study is required here.

Using a pooled approach when fitting models can be justified under a model-design-based view. Possible differences between the finite target populations generated by the model can be included as variables in the model and tested for significance. However, there may be an issue as to which models are suitable to be used, particularly if dealing with small sample sizes.

Pooled samples do not necessarily need to have weights adjusted. That depends on the target population to which the combined estimate refers. It is possible, for example, that the target group for the analysis actually includes units from the two or more different time points from which the samples were taken.

Frequently, an analyst wishes to include a number of different analyses in his study. He thus needs to consider whether a "multipurpose" pooled file can actually estimate all quantities of interest by the desired approach(es). What is optimal for the estimator of one quantity may not be optimal for others.

Statistical tests about assumptions (such as equality of a characteristic in the finite populations from which the different samples are drawn) may have little power if sample sizes are small. Other sources of information could be more valuable in deciding whether assumptions seem reasonable.

It may not be straightforward to use software tools designed for a pooled approach to produce estimates for the separate approach.

Suitable variance estimation may be difficult for either the separate or the pooled approach, especially if samples are not independently selected.

References

Béland Y. (2002), "Canadian Community Health Survey - Methodological Overview," *Health Reports* , Statistics Canada, Catalogue 82-003-X, Ottawa, 9-14.

Binder, D. A. and G. R. Roberts (2003), "Design-Based and Model-Based Methods for Estimating Model Parameters," in: R.L. Chambers and C.J Skinner eds., *Analysis of Survey Data*, Wiley, Chichester, 29-48.

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

Schenker, N., Gentleman, J.F., Rose, D., Hing, E., and I.M. Shimizu (2002). "Combining estimates from complementary surveys: a case study using prevalence estimates from national health surveys of households and nursing homes," *Public Health Reports 2002*, 117, 393-407.

Schenker, N. and T.E. Raghunathan, (2007), "Combining Information from Multiple Surveys to Enhance Estimation of Measures of Health," *Statistics in Medicine, 26*, 1802-1811.

Thomas, S. (2007), "Combining Cycles of the Canadian Community Health Survey," *Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health*, Statistics Canada, Catalogue 11-522-XIE, Ottawa.

Tjepkema, M. (2008), "Health Care Use Among Gay, Lesbian and Bisexual Canadians," *Health Reports, 19(1),*Statistics Canada, Catalogue 82-003, Ottawa, 53-64.