Appendix C

HOW EFFECTIVE IS USING A CONVENIENCE SAMPLE TO SUPPLEMENT A PROBABILITY SAMPLE?

The appeal of Web-based convenience samples lies in the potentially very low marginal cost per respondent. Attracting respondents to a Web site does not require expensive labor (as phone calling does) or expensive materials (as mailings do). Furthermore, marginal processing costs per respondent are also reduced because the data are already recorded electronically.

But the disadvantage of convenience samples is obvious—potentially large and unmeasured bias. One solution to this problem may be to use a combined probability/convenience sample.

The idea behind this combined-sample concept is that the same survey would be administered to both a traditional probability sample (with or without a Web-based response mode) and a Web-based convenience sample. For example, obtaining a probability sample with 4,000 individuals and a convenience sample with 10,000 individuals might be no more expensive than obtaining a probability sample with 5,000 individuals (assuming that convenience observations are one-tenth the cost of probability observations).

The probability sample will provide a means of measuring the bias present in the convenience sample, parameter by parameter. With an estimate of the amount of bias, one could then combine information from the convenience and probability samples to yield more-precise estimates than would be possible from the probability sample alone. If the convenience sample is very biased, then it will be nearly useless. This implies that the probability portion of the sample

would have to be large enough to stand on its own in a worst-case scenario.

If the bias is so large that it renders the convenience sample useless, then there is a moderate loss in precision. (In the example just given, the standard errors would be increased by 10 percent, hypothetically, because only 4,000 observations were available instead of 5,000.) However, if the bias is small, then there is a "precision windfall," allowing subgroup analyses that otherwise would not have been affordable.

USING THE PROBABILITY SAMPLE TO ADJUST THE CONVENIENCE SAMPLE

Assume a probability sample with X_{1i} that are independently and identically distributed (iid) with mean μ, variance σ_1^2, $i = 1,\ldots,n_1$. Also assume a convenience sample of X_{2j} that are iid with mean $\mu + \varepsilon$, variance σ_2^2, $j = 1,\ldots,n_2$; the X_{1i} and X_{2j} are independent; ε, σ_1^2, and σ_2^2 are known; and μ is the unknown parameter of interest.

One would naturally consider using information in the probability sample to attempt to remove the bias from the convenience sample prior to combining the data from the two samples to estimate μ. That is, one can estimate the bias as $\hat{\varepsilon} = \bar{X}_1 - \bar{X}_2$, where

$$\bar{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} \text{ and } \bar{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j},$$

and then use the estimate to adjust each of the convenience sample observations: $X_{2j}^* = X_{2j} - \hat{\varepsilon}$. Having adjusted each of the convenience sample observations, the mean can be estimated as

$$\hat{\mu} = \frac{1}{n_1 + n_2}\left[\sum_{i=1}^{n_1} X_{1i} + \sum_{j=1}^{n_2} X_{2j}^*\right].$$

For this estimator, one could then ask, what is the optimal allocation of the sample between n_1 and n_2 that would minimize the variance of $\hat{\mu}$? The unfortunate reality is that $\text{Var}(\hat{\mu}) = \sigma_1^2 / n_1$. Hence, the variance of the estimator depends *only* on the sample size of the probability sample, which means that the variance is minimized *a priori* by allocating everything to the probability sample. That is, after adjustment, the convenience sample contains no information to contribute to the estimation of the sample mean, so there is no point in allocating resources to collecting the convenience sample, no matter how inexpensive the convenience sample observations are to obtain.

INITIAL BIAS REDUCTION

If attempting to remove the bias from the convenience sample will prove ineffective, then the only alternative is to use the (potentially) biased data in the estimation. However, as we show later in this appendix, and as one might expect, the bias of the convenience sample must be small. One way to respond to this limitation may be to focus on estimating parameters that are less subject to bias, such as within-subject differences or regression coefficients, rather than population estimates of proportions or means. One can also use post-stratification to reduce bias as much as possible. For example, a small set of items can be included in both the convenience and probability samples that are (1) associated with likelihood of participation in the Web-based convenience sample (for example, age, education, computer use, and other such factors) and (2) likely to be associated with the parameters being measured.

To use the post-stratification variables, one should treat the characteristics of the probability sample as the target and model the relative response probabilities of members of the "convenience sample pool" with given values of post-stratification variables. Weights inversely proportional to these estimated relative probabilities are then applied to the convenience sample only. The design effect from this process will reduce the effective sample size (ESS) of the convenience sample, but the low cost of these observations makes compensating for moderate design effects on the convenience sample affordable.

LINEAR COMBINATIONS OF BIASED AND UNBIASED ESTIMATORS OF A POPULATION MEAN

The previous discussion prompts a specific estimation problem: What is the most efficient estimator that is a linear combination of an *unbiased estimator* (the sample mean of the population of interest) and a *biased estimator* (the sample mean of a population that is biased with respect to the population of interest)?

The notation and initial assumptions are as follows: Let n_1 be the number of observations in the unbiased (probability) sample. Let n_2^* be the number of observations in the biased (convenience) sample. Let DEFF be the design effect of post-stratification weights on the convenience sample. Let $n_2 = n_2^*/\text{DEFF}$ be the ESS of the convenience sample. As earlier, assume that X_{1i} are iid with mean μ, variance $\sigma_1^2, i = 1,\ldots,n_1$ and assume that X_{2j} are iid with mean $\mu + \varepsilon$, variance $\sigma_2^2, j = 1,\ldots,n_2$. Also, as earlier, assume that X_{1i} and X_{2j} are independent; ε, σ_1^2,and σ_2^2 are known; and μ is the unknown parameter of interest. Thus, ε is the residual bias after post-stratification.

We are interested in the estimator $\hat{\mu} = \lambda\overline{X}_2 + (1-\lambda)\overline{X}_1$ where

$$\overline{X}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} \text{ and } \overline{X}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}.$$

Therefore, the bias and variance of this estimator are:

$$\text{bias}(\hat{\mu}) = \lambda\varepsilon; \text{var}(\hat{\mu}) = \lambda^2\sigma_2^2/n_2 + (1-\lambda)^2\sigma_1^2/n_1.$$

As shorthand notation, let $\Sigma_1^2 = \sigma_1^2/n_1$ and $\Sigma_2^2 = \sigma_2^2/n_2$. Note that Σ_1^2 is the mean squared error (MSE) of the probability sample and Σ_2^2 is what the MSE of the convenience sample would be if post-stratification had removed all bias. In this notation, $\text{MSE}(\hat{\mu}) = (\Sigma_1^2 + \Sigma_2^2 + \varepsilon^2)\lambda^2 - 2\Sigma_1^2\lambda + \Sigma_1^2$. The value of λ that minimizes $\text{MSE}(\hat{\mu})$ is $\lambda = \Sigma_1^2/(\Sigma_1^2 + \Sigma_2^2 + \varepsilon^2)$, which means that the preferred estimator is of the following form:

$$\hat{\mu} = \frac{\Sigma_1^2 \bar{X}_2 + \left(\Sigma_2^2 + \varepsilon^2\right)\bar{X}_1}{\Sigma_1^2 + \varepsilon_2^2 + \varepsilon^2}.$$

The intuition for the form just shown is that observations are weighted inversely to the MSE per observation from each sample. Again, as shorthand, let $\Omega = \Sigma_1^2 + \Sigma_2^2 + \varepsilon^2$ so that $\lambda = \Sigma_1^2 / \Omega$ and $1 - \lambda = (\Sigma_2^2 + \varepsilon^2)/\Omega$. Then, one can write $\text{MSE}(\hat{\mu}) = \Sigma_1^2 (\Sigma_2^2 + \varepsilon^2)/(\Sigma_1^2 + \Sigma_2^2 + \varepsilon^2)$.

Note that as $\varepsilon \rightarrow 0$,

$$\text{MSE} \rightarrow \frac{1}{\left(1/\Sigma_1^2 + 1/\Sigma_2^2\right)},$$

and as $\varepsilon \rightarrow \infty$, $\text{MSE} \rightarrow \Sigma_1^2$, the MSE of the probability sample.

Also, as $n_2 \rightarrow \infty$,

$$\text{MSE} \rightarrow \frac{1}{\left(1/\varepsilon^2 + 1/\Sigma_1^2\right)},$$

which is the minimum MSE possible for a given bias.

QUANTIFYING THE CONTRIBUTIONS OF THE CONVENIENCE SAMPLE

Let UESS be the sample size of an unbiased sample mean with the same MSE as the pooled estimator. Then, one can express it as the following:

$$\text{UESS} = \left(\frac{\Omega}{\Omega - \Sigma_1^2}\right) n_1.$$

Let IUESS be the increment to UESS added by the convenience sample. Then, the equivalent probability sample size increment can be expressed as

$$\text{IUESS} = \left(\frac{\Sigma_1^2}{\Omega - \Sigma_1^2} \right) n_1 = \frac{\sigma_1^2}{\Sigma_2^2 + \varepsilon^2}.$$

Next, define the bias in terms of standard deviations of the probability sample $E = \varepsilon / \sigma_1$. Now consider the simplified respondent where $\sigma_1^2 = \sigma_2^2$ so that

$$\text{IUESS} = \frac{1}{1/n_2 + E^2}.$$

As $E \to \infty$, IUESS $\to 0$, and as $n_2 \to \infty$, IUESS $\to 1/E^2 = \text{MIUESS}$, the maximum possible increment to effective sample size. Note the striking ceiling on the IUESS. It means that an uncorrected bias of 1/100 of a standard error limits the IUESS to 10,000. This is a pretty sobering result—an unbiased sample of 150 is preferable to a sample with 10,000 observations and a standard deviation bias of 0.1.

CONCLUSIONS

We have shown that there is no point in using a probability sample to remove the bias from a convenience sample. Furthermore, the use of an unadjusted convenience sample to supplement a probability sample may be practical only under limited circumstances:

- The probability sample is large (at least 2,000).
- The convenience sample is inexpensive (no more than 20 percent of the cost per observation).
- The convenience sample is large (at least as large as the probability sample).
- The bias after post-stratification is very low (no more than three percentage points).

From a practical point of view, it is also not clear what the source would be for an estimate of the bias parameter.