COMBINING RANDOM AND NON-RANDOM SAMPLES

DOUGLAS RIVERS, VICKI HUGGINS, AND DANIEL SLOTWINER

ABSTRACT. Methods for combining data from random and non-random (convenience) samples are investigated for estimating characteristics of low incidence populations. In each type of sample, logistic regressions are used to summarize the conditional distribution of responses given some covariates whose marginal distribution is known. The population proportion of responses is obtained by weighting the predicted probabilities from logistic regression by the known population marginals. The estimates based on the random and non-random samples are then weighted to minimize the mean square error of the combinated estimate. Two applications illustrate the method.

1. Introduction

It can be said, without too much exaggeration, that the world of survey research is divided into two camps. On one side, primarily in the government and academic world, are those who insist upon proper probability sampling designs, who calculate response rates, and worry about design effects. On the other side, mostly in the world of commercial market research, are those willing to use convenience samples, who fill quotas, and care little about sampling theory. One tends to encounter the former group at JSM and the latter at CASRO. These two groups rarely talk to one another and sometimes even appear to be unaware of the existence of the other. To simplify the discussion (and to avoid perjorative labels), we shall refer to the approaches as *random* and *non-random* sampling.

Nonrandom samples refer to samples where the sampling frame is not well-defined and there is no known probability of selection. In other words, there is not a full accounting of the population of interest such that a representative sample can be drawn. Mall intercept samples, most e-mail lists, commercial mail panels, and home scanner panels are non-probability samples and for which conventional sampling theory is inapplicable. Because

---

Date: August 2, 2003.
Prepared for the 2003 Joint Statistical Meetings of the American Statistical Association.

little is known about the sample selection mechanism, the sampling distribution of sample statistics is unknown. Nonetheless, it is common practice to compute standard errors under the (incorrect) assumption of random sampling.

A few years ago, in the early days of Internet surveys, the two camps debated what was the right approach. Then and now, most Internet surveys were based non-randomly recruited panels. Insofar as those from the random sampling camp contemplated using the Internet, it was for applications where a complete listing (including e-mail addresses) of the population existed or for Internet measurement (using RDD to recruit Internet users, such as the NetRatings panel).

The Knowledge Networks (KN) panel was unusual in that it attempted to use probability sampling and Web-based interviewing for general population studies. The KN sample is a representative sample of the U.S. population; households without Internet access were provided with an inexpensive Web access device, solving the coverage problem. A probability sample of phone numbers (random within pre-identified strata) is selected out of all possible phone numbers in the United States. Sample weights are calculated to reflect the probability of selecting a household or person into the sample. By using the sample weights, the estimates calculated from the panel are representative of the U.S. population and can project for any subpopulation within such as Blacks, Hispanics, households with young children, people that consume coffee, etc.

The main drawback of the KN approach is its expense. The cost of providing hardware and Internet access, in addition to the cost of RDD recruitment, makes the cost of maintaining the KN panel substantially higher than non-random Internet panels (though somewhat less than the cost of maintaining an equivalent quality RDD phone panel). As a consequence, the KN panel is much smaller than competing nonrandom Internet panels, which often claim to have a million or more active members.[1]

While we locate ourselves in the random sampling camp, we have some appreciation for why many commercial market research surveys are conducted using non-random methods. When time is short, budgets are limited, and incidence rates are low, the choice is, as a practical matter, not between a random and a non-random sample, but between a non-random sample and not doing the study. Commercial market research is often targeted at small, hard to reach populations (*e.g.*, premium Tequilla drinkers). The

[1]There is some question about the effective size of the nonrandom Internet panels, as the vendors tend to be secretive about the size, composition, and response rates of their panels.

client would like to track the target group "continuously" (not by reinterviewing the same respondents, but by drawing a fresh sample every month or quarter). And they would like to do all of this for a fraction of what it would cost using random sampling.

Faced with these tradeoffs, commercial clients often choose to dispense with random sampling, especially since non-random sampling vendors claim that their sampling methods have been "validated" and work perfectly well. When one scratches the surface of these claims, however, it often turns out that the validation consists in showing that the same method yields the same results over time or gave similar results as some other method (also with unknown properties). Frequently the "validation" turns out to be that the estimates from one method are "strongly correlated" with the estimates produced by another method.[2]

What we propose in this paper is an alternative approach that attempts to combine random and non-random samples. Quota sampling, post-stratification, and propensity score weighting all depend on an implicit assumption of ignorability of the unknown sampling mechanism (discussed in detail below). A random sample—even a very small random sample—is very useful because it allows us to test the assumption of ignorability. If such as assumption is warranted, then there is useful information in the non-random sample that might be used to improve estimation.

The procedure involves three steps. First, the conditional distribution of response variables in the random and non-random samples is estimated using logistic regression. This reduces a potentially high-dimensional estimation problem into something more manageable and provides a simple test of ignorability. (This test can be done non-parametricly, but we follow the approach conventionally used in propensity score estimation.) Next, we use the *estimated* conditional distribution of responses weighted by known population marginals (taken from the Current Population Survey) to obtain a post-stratified estimate from each subsample. (This idea is similar to that used in Gelman *et al.*, 2002.) Finally, the two post-stratified estimates are weighted to minimize the mean square error of the combined estimate.

The approach does rely upon some modelling assumptions, but we have attempted to make these assumptions explicit and, as much as possible, to minimize their impact. We do not as yet have much experience in using this methodology. Two sample applications are provided to illustrate the technique. It appears that the assumption of ignorability is critical and may

[2]We have also encountered clients who prefer to obtain consistent results which they know to be inaccurate, to switching to a new method that is likely to give less biased estimates. Presumably, they are employing some informal bias correction adjustment and are unsure of how to make such adjustments with the new method, even though the required adjustments would be smaller.

not be satisfied, even as a rough approximation, for the non-random samples available to us.

The plan of the paper is as follows. Section 2 describes a common notation and framework for understanding the problem. The critical assumption of ignorability is defined. Section 3 reviews the traditional approach of quota sampling as well as the more recent technique of propensity score weighting. Both depend upon an implicit assumption of ignorability. Section 4 describes our proposed approach to combining random and non-random samples, which recognizes that the ignorability assumption needs to be tested and, in most cases, will not be exactly satisfied. Finally, Section 5 contains two illustrative applications of the proposed technique.

2. Notation, Definitions and Assumptions

In this section, we explain the notation and terminology that will be utilized in the remainder of the paper. We will assume that all variables are discrete. Continuous variables can be accommodated using density estimators, but the discussion is simpler in the discrete case which applies to the preponderance of variables encountered in survey work. In most cases, the response variable of interest will be assumed to be dichotomous, so the relevant parameter is the population fraction or success probability. Again, generalizations are feasible and fairly obvious, so these assumptions are not restrictive.

Let Y denote some variable of interest, X a set of covariates whose marginal distribution is available from a census, and S a sample indicator ($S = r$ for a random sample, $S = n$ for a sample with unknown selection probabilities—the "non-random" sample). In most cases Y can be taken to be dichotomous, so $Y \in \mathcal{Y} = \{0, 1\}$. We are able to make repeated independent draws according to each sampling process. For the random sample, this will usually be a result of the design (*e.g.*, if we have random sampling either with or without replacement, equal probability of selection, and the underlying population is large). For the non-random sample, the assumption can usually be justified by an exchangeability argument.

Let $\pi(x)$ denote the marginal distribution of the covariates in the population. Let $f_s(x)$ denote the (marginal) frequency function for X in sample $s \in \{r, n\}$. We shall assume that $\pi(x) > 0$ for $x \in \mathcal{X}$, the support of X. As discussed above, $f_r(x) = \pi(x)$ for $x \in \mathcal{X}$. Little will be assumed about $f_n(x)$ (the distribution of the covariates in the non-random sample) other than regularity ($f_n(x) > 0$ for all $x \in \mathcal{X}$). The latter assumption is not altogether trivial. For example, if X includes a measure of Internet access, Internet convenience samples will not satisfy this assumption (since they fail to represent those without Internet access).

Both the joint and conditional distributions $f_s(x,y)$ and $f_s(y|x)$ can be estimated from respective samples (because of the independence assumption). By design, the joint distribution of X and Y in the random sample, $f_r(x,y)$, matches the (unknown) joint distribution in the population $\pi(x,y)$. Conventional sampling theory concerns estimation and inference about $\pi(x,y)$ when the sample design involves stratification, clustering, multi-stage selection, and so forth. Again, to simplify the discussion, we shall assume simple random sampling with replacement. It is, of course, knowledge of the selection mechanism (equal probability of selection and independence of draws with SRS with replacement) that guarantees that the estimates will be unbiased and the inferences correct.

In the non-random sample, nothing is really known about the "sampling mechanism" $f_n(x,y)$. It is, nonetheless, a well-defined probability distribution and the more relevant question is what is its relationship to the population distribution of interest $\pi(x,y)$. Furthermore, since the marginal distribution of X, denoted $\pi(x)$, is known (from a census or other source), we only need to estimate $\pi(y|x) = \pi(x,y)/\pi(x)$. A sampling mechanism is said to be *ignorable* conditional upon covariates X if

$$f_s(y|x) = \pi(y|x) \quad \text{for all } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}.$$

For the random sample, $f_r(y|x) = \pi(y|x)$ by design. For the non-random sample, the assumption of ignorability is clearly much more a leap of faith.

Because of cost considerations, the random sample is usually small, so estimates based on it, although unbiased, are likely to have high variance. The non-random sample is much cheaper and therefore usually larger. Estimates based on it will be biased but have smaller variance. In combining the samples, we confront a bias-variance tradeoff. This is the basis for our method of combining samples in Section 4, but first we consider how alternative methods can be understood within this framework.

3. Quota Sampling and Propensity Score Weighting

Among non-random Internet surveys, it appears that the most popular methods for dealing with non-random selection are quota sampling and propensity score weighting. There is little technical discussion of these methods available, but both can be understood as methods for weighting for non-random selection.

In quota sampling, one establishes "quotas" for population groups $[X = x]$ proportional to their size in the population $\pi(x)$ and samples until the quota is filled. This is equivalent to (though much cheaper than) running the sampling mechanism to completion and weighting observations with $X = x$ equally until the quota is filled and zero thereafter. This involves two key assumptions. First is an assumption of exchangeability (the sequence in which sample draws occurs is uninformative). This would be violated if the first respondents with $X = x$ differ in some way from later respondents with the same characteristic. (For example, in a telephone survey, quotas for women are likely to be filled quickly with non-working women, leading to an underrepresentation of working women.) Second, is an assumption of ignorability. The whole point of quota sampling is to obtain a sample that matches the known population marginals $\pi(x)$ (though the quotas are usually imposed only on the marginal distribution of each component of X rather than on the joint distribution of X). If this is done, then $f_n(y) = \pi(x)$ and

$$f_n(x, y) = f_n(y|x)f_n(x) = \pi(y|x)\pi(x) = \pi(x, y),$$

provided ignorability holds ($f_n(y|x) = \pi(y|x)$). These are, of course, very strong assumptions and, at least in our experience, rarely is much effort made to verify or justify such assumptions.

One practical problem of quota sampling is the difficulty of dealing with more than a small number of covariates. The plausibility of the ignorability assumption depends rather criticstyally upon controlling for the myriad of factors which affect sample selection. It becomes time-consuming and expensive to fill a large number of small quotas. It is also wasteful to reject respondents who belong to groups whose quotas have already been filled. An obvious alternative is post-stratification weighting.[3] However, if cell sizes are small, post-stratification weights can become unstable.

Recently, some Internet surveys have been weighted by *propensity scores*. Propensity scores were introduced by Rosenbaum and Rubin (1984) for matching treatment and control groups in observational studies. As applied to surveys, the sample from a random sample (usually a telephone survey) is treated as the control group and the sample from a non-random Internet panel is the treatment group. The treatment group (*i.e.*, the non-random Internet panel) is then matched to the control group (*i.e.*, the random sample) by values of the propensity score.

The propensity score is usually estimated by a logistic regression of sample source (random vs. non-random) on some covariates[4] The random sample is then divided into quintiles (or sometimes deciles) and the non-random sample is divided using the same cutpoints. The non-random sample is then

[3]We have learned that many commercial clients are so accustomed to quota samples that they do not expect samples to come with weights. Any departures from population marginals, however small, is viewed as a sign of poor quality, rather than normal sampling fluctuations. At Knowledge Networks, we have often used stratification for within-panel sample selection to obtain random samples that look like quota samples. This greatly complicates panel management.

[4]The covariates need not be restricted to census demographics.

weighted inversely proportional to the number of observations falling into each of the propensity score quintiles. Thus, for example, if 10% of the non-random sample has a propensity score value that would fall into the top quintile of the control group propensity scores, these observations would be weighted by a factor of two.

Note that the logistic regression does *not* correspond to the probability of sample selection. If the random sample uses equal probability of selection, selection into the random sample does not depend upon the covariates X. Further, the size of each subsample is determined by the design and is not random. Nonetheless, we proceed formally as if

$$\text{Prob}\{S = s\} = \frac{n_s}{n_r + n_n}$$

where n_r and n_n are the sizes of the random and non-random samples, respectively. By Bayes' Theorem (taking $f_s(x)$ as the conditional distribution of X given $S = s$),

$$\begin{aligned} p(x) &= \text{Prob}\{S = r | X = x\} \\ &= \frac{\frac{n_r}{n_r + n_n} f_r(x)}{\frac{n_r}{n_r + n_n} f_r(x) + \frac{n_n}{n_r + n_n} f_n(x)} \\ &= \frac{n_r}{n_r + n_n \frac{f_n(x)}{f_r(x)}} \end{aligned}$$

where $p(x)$ is the *propensity score*. This shows that the propensity score is a monotone decreasing function of the ratio of selection probabilities $f_n(x)/f_r(x)$. This means that matching on propensity scores is equivalent to poststratification weighting. Matching on propensity score quintiles is approximately equivalent to post-stratification weighting. The approximation arises because all observations within the same quintile of propensity scores are weighted equally.

To summarize, conventional methods for adjusting non-random samples, such as quota sampling and propensity score weighting all involve an implicit (and usually untested) assumption of ignorability. The adjustment amounts to either exact or approximate poststratification upon the population marginals or an estimate of the population marginals. The primary drawback of such methods in practice is that the ignorability assumption is untested and assumed to hold exactly. Such procedures inevitably involve some degree of bias. When a random sample is available, it is possible to estimate the amount of bias and to improve estimation by seeking an optimal bias-variance tradeoff.

4. Combining Samples to Minimize Mean Square Error

The primary advantage of non-random methods is that they offer larger samples at lower cost. The primary disadvantage is that they involve unknown biases. We propose a method for combining random and non-random samples that we think is a reasonable compromise between the purist position (only samples with known selection probabilities determined by the design) and pure expediency (ignore how the data were collected and hope for the best).

Our approach relies upon a parametric model for the conditional frequency functions $f_r(y|x)$ and $f_s(y|x)$. We use logit models

$$\text{Prob}\{Y = 1|X = x, S = s\} = f_s(1|x) = \Lambda(x^T \beta_s) \qquad s \in \{r, n\}$$

where $\Lambda(u) = e^u/(1 + e^u)$ to model these conditional distributions. The advantage of this approach is that it reduces estimation in each subsample to an easily-implemented and tractable parametric problem. A fully non-parametric approach would require estimation probabilities for each element of $\mathcal{X} \times \{0, 1\}$. In most applications, $\mathcal{X}$ will be high-dimensional. (Of course, the logistic regression can include functions of x and interactions, but the specification is admittedly *ad hoc*.) Let $\hat{\beta}_r$ and $\hat{\beta}_n$ denote the MLEs of β_r and β_n. The asymptotic covariance matrix of $\hat{\beta}_s$ is

$$\widehat{\boldsymbol{V}}_s = \text{var}\{\hat{\beta}_s\} = \left(\boldsymbol{X}_s^T \boldsymbol{D}_s \boldsymbol{X}_s\right)^{-1}$$

where $\boldsymbol{X}_s$ is the matrix of covariates in sample s (with n_s rows) and $\boldsymbol{D}_s$ is a diagonal matrix with typical element

$$d_{ii}^{(s)} = \Lambda(x_{is}^T \hat{\beta}_s) \left(1 - \Lambda(x_{is}^T \hat{\beta}_s)\right).$$

With this simplification, the frequency function for Y in the random sample is

$$f_r(y) = \sum_{x \in \mathcal{X}} f_r(y|x) f_r(x) = \sum_{x \in \mathcal{X}} \pi(y|x)\pi(x).$$

In the non-random sample, the marginal distribution of the covariates $f_n(x)$ will usually differ from that of the population $\pi(x)$. However, if sample selection is ignorable conditional upon X, then we can still obtain an estimate of the marginal distribution of Y given X using

$$\begin{aligned}\sum_{x \in \mathcal{X}} \Lambda(x^T \beta_n)\pi(x) &= \sum_{x \in \mathcal{X}} f_n(y|x)\pi(x) \\ &= \sum_{x \in \mathcal{X}} \pi(y|x)\pi(x) = f_r(y).\end{aligned}$$

Let $\theta = \text{Prob}\{Y = 1\}$ denote the parameter of interest. We have two estimates of θ:

(1) From the random sample, we estimate the conditional frequency function $f_r(1|x) = \Lambda(x^T\beta_r)$ using logistic regression and apply the poststratification weights,

$$\hat{\theta}_r = \sum_{x \in \mathcal{X}} \Lambda(x^T\hat{\beta}_r)\pi(x).$$

(2) From the non-random sample, we estimate the conditional frequency function $f_n(1|x) = \Lambda(x^T\beta_n)$ using logistic regression and apply the poststratification weights,

$$\hat{\theta}_n = \sum_{x \in \mathcal{X}} \Lambda(x^T\hat{\beta}_n)\pi(x).$$

Within this setup, it is easy to test for ignorability of sample selection in the non-random sample. The hypothesis of ignorability holds if and only if $\beta_r = \beta_s$, which can be tested using the likelihood ratio statistic (or equivalent test statistic). These tests usually have good power, even when the random sample is quite small. In most examples that we have seen, ignorability is rejected (and usually rejected overwhelmingly).

If ignorability holds, then certainly we can get a better estimate of θ by taking averaging the two estimates $\hat{\theta}_r$ and $\hat{\theta}_n$. However, in general, selection will not be ignorable and the pooling will cause bias. But even when selection is nonignorable, an appropriate average will still be a better estimate in the sense of reducing the mean square error of estimation. Denote the weighted estimator by

$$\hat{\theta}(w) = w\hat{\theta}_r + (1-w)\hat{\theta}_n.$$

The mean square error of the weighted estimate is

$$\begin{aligned}\text{MSE}\{\hat{\theta}(w)\} &= w^2\text{MSE}\{\hat{\theta}_r\} + (1-w)^2\text{MSE}\{\hat{\theta}_n\} \\ &\quad + 2w(1-w)\text{Bias}\{\hat{\theta}_r\}\text{Bias}\{\hat{\theta}_n\} \\ &= w^2\text{MSE}\{\hat{\theta}_r\} + (1-w)^2\text{MSE}\{\hat{\theta}_n\},\end{aligned}$$

since $\hat{\theta}_r$ is unbiased. Since $\text{MSE}\{\theta(w)\}$ is a convex function of w, the minimum mean square error is attained for w equal to

$$w^* = \frac{\text{MSE}\{\hat{\theta}_0\}}{\text{MSE}\{\hat{\theta}_0\} + \text{MSE}\{\hat{\theta}_n\}}.$$

Thus, an alternative and possibly superior estimator would be of the form

$$\hat{\theta} = \hat{w}\hat{\theta}_r + (1-\hat{w})\hat{\theta}_n.$$

where $\hat{w}$ is an estimate of w^*. The remainder of this section is devoted to how w^* should be estimated.

Before proceeding, it is helpful to introduce some additional notation. Let $\tilde{\boldsymbol{X}}$ be a matrix whose rows consist of the elements of $\mathcal{X}$. Let $\boldsymbol{\pi}$ be a vector with typical element $\pi(x)$ (so $\boldsymbol{\pi}$ has the same number of rows as $\boldsymbol{X}$). The number of rows of $\boldsymbol{X}$ and $\boldsymbol{\pi}$ will depend upon the number of variables included in X. For example, if X includes indicators for race (three categories), education (four categories), and gender (two categories), then $\mathcal{X}$ has $3 \times 4 \times 2 = 24$ elements, and both $\tilde{\boldsymbol{X}}$ and $\boldsymbol{\pi}$ have 24 rows. Let $\hat{\boldsymbol{\lambda}}_s$ be a column vector with typical element $\Lambda(x^T\hat{\beta}_s)$ arranged in the same order as $\boldsymbol{\pi}$. Thus, $\hat{\boldsymbol{\lambda}}_r$ contains the predicted probabilities for each possible $x \in \mathcal{X}$ based on the random sample. Similarly, $\hat{\boldsymbol{\lambda}}_n$ contains the predicted probabilities based on the logistic regression from the non-random sample.

In this notation, the two estimates of θ are

$$\hat{\theta}_s = \boldsymbol{\pi}^T \boldsymbol{\lambda}_s \qquad s \in \{r, n\}$$

and, using the delta method,

$$\begin{aligned}
\hat{V}_s &= \text{var}\{\hat{\theta}_s)\} \\
&\approx \boldsymbol{\pi}^T \frac{\partial \boldsymbol{\lambda}_s}{\partial \hat{\beta}_s^T} \text{var}\{\hat{\beta}_s\} \frac{\partial \boldsymbol{\lambda}_s^T}{\partial \hat{\boldsymbol{\beta}}_s} \boldsymbol{\pi} \\
&\approx \boldsymbol{\pi}^T \tilde{\boldsymbol{D}}_s \tilde{\boldsymbol{X}}^T (\boldsymbol{X}_s^T \boldsymbol{D}_s \boldsymbol{X}_s)^{-1} \tilde{\boldsymbol{X}} \tilde{\boldsymbol{D}}_s \boldsymbol{\pi}
\end{aligned}$$

where $\tilde{\boldsymbol{D}}_s$ is a diagonal matrix with the same number of rows as $\tilde{\boldsymbol{X}}$ and typical element $\Lambda(x^T\hat{\beta}_s)(1 - \Lambda(x^T\hat{\beta}_s))$.

Let us return to the primary case of interest, where Y is a dichotomous indicator of some attribute and we wish to estimate the proportion of respondents θ) possessing the attribute. The random-sample estimator is approximately unbiased, so its mean square error is (aside from a term of order smaller than $1/\sqrt{n}$)

$$\hat{V}_r = \text{var}\{\hat{\theta}_r\} = \boldsymbol{\pi}^T \tilde{\boldsymbol{D}}_r \tilde{\boldsymbol{X}}^T (\boldsymbol{X}_r^T \boldsymbol{D}_r \boldsymbol{X}_r)^{-1} \tilde{\boldsymbol{X}} \tilde{\boldsymbol{D}}_r \boldsymbol{\pi}$$

Since the random-sample based estimate is approximately unbiased, the variance and mean square error coincide.

The estimate $\hat{\theta}_n$ from the non-random sample may be biased, so its approximate mean square error is given by

$$\begin{aligned}
\text{MSE}\{\hat{\theta}_n\} &\approx \left(E\hat{\theta}_r - E\hat{\theta}_n\right)^2 + \text{var}\{\hat{\theta}_n\} \\
&\approx \left(E\hat{\theta}_r - E\hat{\theta}_n\right)^2 + \hat{\boldsymbol{V}}_n
\end{aligned}$$

where

$$\hat{V}_n = \boldsymbol{\pi}^T \tilde{\boldsymbol{D}}_n \tilde{\boldsymbol{X}}^T (\boldsymbol{X}_n^T \boldsymbol{D}_n \boldsymbol{X}_n)^{-1} \tilde{\boldsymbol{X}} \tilde{\boldsymbol{D}}_n \boldsymbol{\pi}.$$

Estimation of the squared bias raises some complications. $(\hat{\theta}_r - \hat{\theta}_n)^2$ is consistent, but in small samples will usually be too large. Instead, we estimate the squared bias using

$$\hat{\delta}^2 = (\hat{\theta}_r - \hat{\theta}_n)^2 - \hat{V}_r - \hat{V}_n$$

when this quantity is positive and by zero otherwise.

Finally, the optimal weight w^* is estimated by replacing the unknown quantities by the estimates given above, suggesting the estimator

$$\hat{\theta} = \hat{w}\hat{\theta}_r + (1 - \hat{w})\hat{\theta}_n$$

where

$$\hat{w} = \frac{\hat{\delta}^2 + \hat{V}_n}{\hat{\delta}^2 + \hat{V}_n + \hat{V}_r}.$$

The weight is between zero and one, with higher values indicating how much weight is given to the random estimate. As sample size increases, the two variance terms tend to zero, so $\hat{w}$ is determined primarily by the degree of bias. However, in small samples, the formula may give considerable weight to the non-random sample despite its bias.

5. Example Applications

We have conducted two cases studies where we utilized e-mail convenience samples to increase the sample size for certain subpopulations or incidence groups. In each case, the random sample was provided by the Knowledge Networks panel, while the nonrandom sample came from two independent vendors using convenience samples. The descriptions below illustrate the process, methodology employed, analyses conducted, and results relevant to each study. The first case study required a significant oversample of a subpopulation and the target population for the second study was estimated to comprise only 2% of the total population.

5.1. *Alcohol Incidence and Consumption Study (ACS)*

An alcohol awareness and usage study, required a substantial oversample of males aged 21–27 who consumed certain alcoholic beverages. We were particularly interested in boosting the sample size for Black and Hispanic males aged 21–27. The national incidence rate for consumption of spirits was estimated to be about 36%. We requested potential vendors to bid on data collection for this study under two strategies: one with a quota sampling strategy (which we will refer to as ACS-S1) and one without quotas (called ACS-S2). For the quota sampling, we supplied estimated incidence rates based upon an initial survey of KN panelists. (The methodology described in section 4 is applicable if the quotas do not match the population marginals, since all estimates are post-stratified using CPS marginals.)

Table 1
ACS Logistic Regressions

Variable	Random Sample	Nonrandom Sample	Pooled
Intercept	0.01	1.57	-0.93
	(0.27)	(0.43)	(0.30)
Education	-0.22	-0.03	-0.23
	(0.18)	(0.39)	(0.16)
Age	-0.22	-0.04	0.12
	(0.17)	(0.25)	0.14)
White	0.04	-0.23	-0.08
	(0.25)	(0.32)	(0.20)
Black	-0.67	-1.01	-0.85
	(0.37)	(0.37)	(0.26)
$-2\log L$	8.5	10.7	59.9
n	541	318	859

The response of interest in the ACS was consumption of specific types of alcoholic beverages. Logistic regressions for one type of beverage are shown in Table 1. The covariates are indicators for education (college graduate), age (over 24), and race (white, black, and other). The test for ignorability yields a likelihood ratio statistic of 40.7 with 5 d.f. ($p < .001$). The estimates from the non-random sample, even when corrected give much higher estimates of consumption of this type of beverage (74% vs. 48%).

In this example, the random sample was somewhat larger than the non-random sample ($n_r = 541$ vs. $n_n = 318$) and, given the evident bias in the non-random sample, it is not surprising that the estimated optimal weight was $\hat{w} = 0.826$. (See Table 2.) In this case, little was gained by supplementing the KN panel with the nonrandom sample.

Table 2
Combined Estimates

	ACS	FPS
$\hat{\theta}_r$	0.483	0.016
$\hat{\theta}_n$	0.741	0.040
$\hat{\theta}$	0.502	0.030
$\hat{w}$	0.826	0.432

5.2. *Facial Products Usage Study (FPS)*

The second study, a facial products awareness and usage study, required females aged 13–69 who used a certain types of facial products. The incidence was estimated to be about 2% in the population. Again, we requested potential vendors to bid on data collection for this study under a quota sampling strategy with incidence rates based upon an initial survey of KN panelists.

The logistic regression estimates for the FPS are shown in Table 3 below. The logistic regressions are fairly similar except for an intercept difference. (This suggests that it might be useful to adjust the nonrandom sample for a level effect, but we have not pursued this.) Nonetheless, the likelihood ratio test easily rejects the hypothesis of ignorability (a test statistic of 51.3 with 5 d.f. ($p < .001$).

The estimates when poststratified on the covariates differ substantially between the two samples. The random sample gives an estimated usage rate of 1.6% versus the nonrandom sample estimate of 4.0%. The estimated weight was 0.432, yielding a combined estimate of 3.0%. (See Table 2 above.)

Table 3
FPS Logistic Regressions

Variable	Random Sample	Nonrandom Sample	Pooled
Intercept	-2.87	-1.19	-3.15
	(0.46)	(0.33)	(0.36)
Education	-0.17	-0.07	-0.13
	(0.22)	(0.19)	(0.14)
Age	-0.24	-0.53	-0.41
	(0.09)	(0.07)	0.06)
White	-0.44	-0.36	-0.39
	(0.38)	(0.31)	(0.36)
Black	-0.67	-0.65	-0.66
	(0.52)	(0.51)	(0.36)
$-2\log L$	10.2	65.1	136.7
n	7,426	4,069	11,495

DEPARTMENT OF POLITICAL SCIENCE, STANFORD UNIVERSITY, STANFORD, CA 94305

KNOWLEDGE NETWORKS, INC., 1350 WILLOW ROAD, MENLO PARK, CA 94025

KNOWLEDGE NETWORKS, INC., 212 FIFTH AVENUE, NEW YORK, NY 10010