Survey Research Methods (2019)
Vol. 13, No. 1, pp. 95-101
doi:10.18148/srm/2019.v13i1.7222

© *European Survey Research Association*
ISSN 1864-3361
http://www.surveymethods.org

A Partially Successful Attempt to Integrate a Web-Recruited Cohort into an Address-Based Sample

Phillip S. Kott
RTI International
Rockville, USA

We use a web-and-mail survey on attitudes towards and use of marijuana to demonstrate how a web-recruited cohort could be integrated into an address-based sample using a calibration-weighting procedure in the software language SUDAAN 11$^{\text{TM}}$. A Holm-Bonferroni procedure is employed to test whether a pivotal assumption underlying the integration is supported by the data for individual survey items as well as for the survey as a whole. Delete-a-group jackknife weights for the integrated sample are then developed.

Keywords: Selection model; Calibration weighting; WTADJX; Holm-Bonferroni procedure; Composite; Delete-a-group jackknife

1 Introduction

A web-and-mail survey was conducted in the US state of Oregon in 2015 on attitudes towards and use of marijuana. The drug had recently been legalized in that state for both medicinal and non-public recreational use by adults 21 or over. Roughly two-thirds of the respondent sample was selected via a simple random sample of Oregon addresses. Randomly sampled individuals, one per address, were encouraged to respond by web, but about half of those respondents returned a mail questionnaire instead. Another third of the respondent sample was purely nonprobability, recruited via Facebook to increase the sample size and (it was hoped) the precision of the estimates. Facebook recruits responded by the internet.

Thus, there were three respondent cohorts: a mail cohort, a mail-to-web cohort, and a recruit cohort. Preliminary investigations revealed that the recruit cohort did not look like the mail cohort, but that the recruit cohort might be comparable to the mail-to-web cohort. The paper shows how the SUDAAN 11 (RTI International, 2012) procedure WTADJX was used to calibrate the randomly-selected respondents to variable totals from the 2014 American Community Survey (ACS) while the recruit cohort was calibrated to the mail-to-web cohort using the same ACS variables and adding the survey item political affiliation (procedures analogous to WTADJX are available in R). WTADJX was also used to assess whether differences between calibrated estimates from the mail-to-web and recruit cohorts were statistically significant. The calibrated weights for these cohorts were then scaled so that the population they represented was single-counted. Finally, delete-a-group jackknife weights (Kott, 2001) were developed for estimates computed from the entire respondent sample so that future inferences could be drawn from the combined sample.

Unfortunately, the survey data were collected before serious thought was given to the estimation to be derived from them. Because of this, the methods described in this paper may be more useful than the results on marijuana attitudes and use in Oregon.

2 Some Survey Details

Respondents to an address-based sample (ABS) of one adult (18 years or older) per Oregon household were each given a 20-minute questionnaire on marijuana use and attitudes. Roughly half responded via mail, half via Internet. The expectation of a poor response rate lead to the recruitment of additional Oregon adults via Facebook.

In all, there was a respondent sample size of 1,989 adults: 722 mail responses, 640 mail-to-web responses, and 627 Facebook recruits. We could not create standard (quasi-)probability weights for the ABS respondents because 745 responding addresses did not provide the number of adults in the household (a survey item). In addition, over 1,300 of the respondents across all the cohorts did not provide their race.

Because of the large fraction missing values for many survey items, only sex, age group (six levels: 18-24, 25-34, 35-44, 45-54, 55-64, 65+), and education (three levels: not a college graduate, a college graduate, other) were used to calibrate the ABS sample to the Oregon adult population totals in the 2014 American Community Survey. Missing values for these survey items were imputed via a hot deck proce-

Phillip S. Kott, RTI International, 6010 Executive Blvd., Rockville, MD USA 20852 (email:pkott@rti.org)

dure, first for age (3 missing), then for sex (76 missing) using imputed sex as a classifying variable, and finally for education (173 missing) using imputed sex and imputed age as classifying variables.

3 The Selection Models

We did not know nor could we estimate the conditional probability that a particular adult in a sampled ABS household was selected to respond to the marijuana questionnaire. Consequently, even our ABS sample was not strictly speaking a probability sample with a conventional adjustment for nonresponse. Instead we assume a selection model where the probability that an Oregon adult was selected for—that is, was sampled and then responded to—the ABS survey is a logistic function of three categorical variables: age group, sex, and education level. Moreover, the selections of individual respondents are assumed to be independent.

A selection model is fit like a response (propensity) model. It is used to estimate the probabilities of selection (which is our case includes response). The parameters of our selection model will be estimated using a calibration equation as we shall see in Section 4 rather than a maximum-likelihood method because Kim and Riddles (2012) showed that the former tends to be more statistically efficient in this context.

The probability that an Oregon adult was recruited into the sample via Facebook is similarly assumed to be an independent logistic function of the above three categorical variables above and political affiliation, a survey-collected categorical variable with five levels: Republican, Democrat, Independent, No preference, and No or an invalid answer. This means that the population that would respond by Internet when given the chance (represented by the mail-to-web cohort) is assumed to be the same as the population that could be recruited via Facebook. This consistency of this assumption with the data will be tested in Section 6.

4 WTADJX

The WTADJX procedure in SUDAAN 11 and analogous procedures in R, such as 'Sampling' (Tille & Matei, 2013) create a set of calibration weights w_k from pre-calibration weights d_k by using Newton's method of repeated linearizations (see, for example, Kott, 2006) to find a vector $\boldsymbol{g}$, if one exists, that satisfies that calibration equation:

$$\sum_{S} d_k \left(1 + \exp(\boldsymbol{x}_k^T \boldsymbol{g})\right) z_k = \boldsymbol{T}_z \quad , \tag{1}$$

where S is the set of respondents, $w_k = d_k \left(1 + \exp(\boldsymbol{x}_k^T \boldsymbol{g})\right)$ is the calibrated weight for respondent k, $\boldsymbol{x}_k$ is a vector of model variables, that is to say, selection is a logistic function of these variables, $\boldsymbol{z}_k$ is a vector of calibration variables (ideally with as many components as $\boldsymbol{x}_k$, and $\boldsymbol{T}_z$ is the known or estimated population total for the z_k (i.e., $\sum_U z_k$, where U is the population or a consistent estimate for this total based on a sample larger than S).

For the Oregon marijuana sample, 0/1 indicator variables were constructed for the respondent's sex (male), age group (age1,... age6), education level (edu1, edu2, edu3), political affiliation (party1, ..., party5), whether the respondent was selected for the ABS sample (abs), and whether the respondent was a Facebook recruit (rec). In addition, a variable (zz) was created. It was set equal to 1 for Facebook recruits, to -1 for respondents in the mail-to-web cohort, and to 0 otherwise.

Letting a×b denote a variable with a value equal to the product of the values of variables a and b, the model vector ($\boldsymbol{x}_k$) consisted of these components: male×abs, age1×abs, ..., age6×abs, edu1×abs, edu2×abs, edu3×abs, male×rec, age1×rec, ..., age6×rec, edu1×rec, edu2×rec, edu3×rec, and party1×rec, ..., *party*5×*rec*. The calibration-variable vector (z_k) consisted of these components: male×abs, age1×abs, ..., age6×abs, edu1×abs, edu2×abs, edu3×abs, male×zz, age1×zz, ..., age6×zz, edu1×zz, edu2×zz, edu3×zz, and party1×zz, ..., party5×zz, Some of these components of each vector were linear combinations of other components. Fortunately, SUDAAN will remove unnecessary components from the calibration vector.

The calibration variable totals ($\boldsymbol{T}_z$) are the ACS totals for males, the six age groups, and three education levels, followed by 17 zeroes. This is because the ABS respondent sample is calibrated to the ACS totals, while the recruit cohort is calibrated to the mail-to-web cohort of the ABS respondent sample. The use of 0 totals in calibration weighting and 1's and -1's in associated calibration variables was introduced by Singh, Dever, and Iannacchione (2004).

Finally, we set all d_k to 1. WTADJX was developed to calibrate the weights in probability samples with design weights. Here we set the "design weights"—the d_k in equation (1)—to 1 in each cohort. The probability of selection (including response) for k is assumed to be

$$p_k = \frac{1}{\left(1 + \exp(\boldsymbol{x}_k^T \boldsymbol{\gamma})\right)} \quad . \tag{2}$$

An estimate for the inverse of the right-hand side of equation (2) is the weight adjustment function

$$\alpha\left(\boldsymbol{x}_k^T \boldsymbol{g}\right) = 1 + \exp(\boldsymbol{x}_k^T \boldsymbol{g}) \quad ,$$

where $\boldsymbol{g}$ is a consistent estimator for $\boldsymbol{\gamma}$.

A weight-adjustment function in WTADJX can be expressed as

$$\alpha\left(\boldsymbol{x}_k^T \boldsymbol{g}\right) = \frac{L + \exp(\boldsymbol{x}_k^T \boldsymbol{g})}{1 + U^{-1} \exp(\boldsymbol{x}_k^T \boldsymbol{g})} \quad . \tag{3}$$

In our selection model, the upper bound U is infinite (the default), while the lower bound L is 1.

The key to establishing the consistency of $\boldsymbol{g}$ under an assumed selection model (shown in Kott, 2006 and elsewhere for a response model) is the mean value theorem:

$$\alpha\left(\boldsymbol{x}_k^T\boldsymbol{g}\right) = \alpha\left(\boldsymbol{x}_k^T\boldsymbol{\gamma}\right) + \alpha'(\theta)\boldsymbol{x}_k^T(\boldsymbol{g}-\boldsymbol{\gamma})$$

for some θ between $\boldsymbol{x}_k^T\boldsymbol{\gamma}$ and $\boldsymbol{x}_k^T\boldsymbol{g}$ when $\alpha(\theta)$ is twice differentiable everywhere. After inserting the right-hand side of the last equation into the calibration equation $\sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{g})\boldsymbol{z}_k = \boldsymbol{T}_z$, we see that

$$\boldsymbol{g}-\boldsymbol{\gamma} = -\frac{\sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{\gamma})\boldsymbol{z}_k - \boldsymbol{T}_z}{\sum_S d_k\alpha'(\theta)\boldsymbol{z}_k\boldsymbol{x}_k^T} \quad .$$

Assuming $\frac{\sum_S d_k\alpha'(\theta)\boldsymbol{z}_k\boldsymbol{x}_k^T}{N}$ and its limit as the population size N grows arbitrarily large is finite and invertible, $\boldsymbol{g}$ converges to $\boldsymbol{\gamma}$ (while $\boldsymbol{x}_k^T\boldsymbol{g}$ and θ converge to $\boldsymbol{x}_k^T\boldsymbol{\gamma}$) in probability as $\frac{\sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{\gamma})\boldsymbol{z}_k - \boldsymbol{T}_z}{N}$ converges in probability to 0.

Observe that the error in the calibrated estimator $t = \sum_S w_k y_k = \sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{g})y_k$ can be expressed as

$$\begin{aligned}\sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{g})y_k - T_y &= \sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{g})\boldsymbol{z}_k^T\,\mathrm{plim}(\boldsymbol{b})\\ &\quad + \sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{g})e_k^*\\ &\quad - (\boldsymbol{T}_z^T\,\mathrm{plim}(\boldsymbol{b}) + T_{e^*})\\ &= \sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{g})e_k^* - T_{e^*}, \qquad (4)\end{aligned}$$

where

$$e_k^* = y_k - \boldsymbol{z}_k^T\,\mathrm{plim}(\boldsymbol{b}) \quad ,$$

$$\boldsymbol{b} = \frac{\sum_S d_k\alpha'(\boldsymbol{x}_k^T\boldsymbol{g})\boldsymbol{x}_k y_k}{\sum_S d_k\alpha'(\boldsymbol{x}_k^T\boldsymbol{g})\boldsymbol{x}_k\boldsymbol{z}_k^T} \quad ,$$

and the probability limit (plim) is taken over the expected respondent sample size n. From equation (4), the mean squared error of t is nearly equal to the mean squared error of

$$\begin{aligned}\sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{g})e_k^* &= \sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{\gamma})e_k^* + \sum_S d_k\alpha'(\theta)e_k^*\boldsymbol{x}_k^T(\boldsymbol{g}-\boldsymbol{\gamma})\\ &\approx \sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{\gamma})e_k^*\end{aligned}$$

because both $(\boldsymbol{g}-\boldsymbol{\gamma})$ and the components of $\frac{\sum_S d_k\alpha'(\theta)e_k^*\boldsymbol{x}_k^T}{N}$ are small (technically, $O_P(1/\sqrt{n})$ under mild conditions). To estimate the variance of t, in a nearly (asymptotically) unbiased fashion WTADJX replaces e_k^* with $e_k = y_k - \boldsymbol{z}_k^T b$ and $\boldsymbol{\gamma}$ with $\boldsymbol{g}$ in $\mathrm{Var}\left(\sum_S d_k\alpha(\boldsymbol{x}_k^T\boldsymbol{\gamma})e_k^*\right)$, treating the respondent selections modeled by equation (2) as independent across respondents.

5 Calibrating on Political Affiliation

Tables 1 and 2 show the impact of calibration on political affiliation. One striking observation is the difference in the pre-calibration fractions of each cohort that provided no (or an invalid) answer to the political affiliation question. This reflects the general tendency of Facebook recruits to not answer items, which is the reason we treated "no answer" as its own level. After calibration, the columns for Facebooks recruits and mail-to-web respondents match exactly as they should. They calibrate to the same number of Oregon adults, which are double counted in the overall total (4,642,817). We address how we corrected for the double counting in Section 7.

6 The Holm-Bonferroni Procedure

The DIFFVAR statement in WTADJX was used to analyze differences in calibrated estimates for the recruit cohort and mail-to-web cohort (the theory supporting the use of this linearization-based technique can be found in Kott, 2006). 40 variables were investigated based on the 20 survey items listed in the appendix. Half concerned whether (or not) there was a valid response to each of the 20 items. The other half were the responses to the 20 items when valid.

A Holm-Bonferroni procedure (Holm, 1979) was used to assess the significance of differences between the cohort estimates. The procedure not only test the assumption that the two cohorts represent the same population for every variable of interest, but also tests the assumption for individual variables.

To perform this procedure, we sorted the 40 differences by their p-values, smallest to largest. The smallest difference is deemed significant at the 10% level when its p-value is less than $\frac{0.1}{40}$. If it is, the second smallest difference is deemed significant at the 10% level if its p-value is less than $\frac{0.1}{39}$, and so forth until a difference is not deemed significant. We assess differences primarily at the 10% level rather than the conventional 5% because a Bonferroni procedure is conservative.

When 40 differences were investigated simultaneously, only the difference in the item "In your opinion, does the legalization of recreational marijuana lead to more people driving under the influence of marijuana?" was deemed significant at the 10% level (barely, and not at the 5% level). This is because 0.00247 in the p-value column is less than 0.00250 in the HB40_.1 column but not 0.0010 in the HB40_.05 column.

Only one response/no-response variable is included in Table 3 containing the nine lowest values, and it features the highest p-value among the nine. If we only investigate the 20 differences for items with valid responses, then the difference in the answer to the question "In your opinion should people be allowed to use edible marijuana in places they are not allowed to smoke it?" would also be significant at the 10%

Table 1
Political Affiliation Before Calibration Weighting

	Cohort			
Political Affiliation	Facebook %	Mail-to-Web %	Mail %	Total
No answer	15	4	6	-
Republican	14	18	22	-
Democrat	26	34	30	-
Independent	19	23	21	-
No preference	27	21	21	-
Total	627	640	722	1,989

Table 2
Political Affiliation After Calibration Weighting

	Cohort			
Political Affiliation	Facebook %	Mail-to-Web %	Mail %	Total
No answer	3	3	5	-
Republican	18	18	22	-
Democrat	29	29	26	-
Independent	24	24	21	-
No preference	27	27	26	-
Total	1,531,798	1,531,798	1,579,221	4,642,817

level (reading from the HB20_.1 column). That only one or two (out of 40 or 20) variables had significant differences at the 10% level is the reason for the "partially successful" in the title. Nevertheless, that a single difference was significant means that the overall Bonferroni test for the equivalence of the two cohorts failed.

7 Jackknife Weights

For future analysis needing standard errors, we created replicate weights using a delete-a-group (dag) jackknife methodology (Kott, 2001). This jackknife requires N to be much larger than n, which is the case here, so that T_{e^*} in equation (4) can be ignored in variance estimation.

To compute dag jackknife weights for those survey items for which the assumptions in Section 3 appear to hold, we randomly sorted the ABS and recruit respondent samples, and then systematically assigned respondents to one of $n_G = 30$ jackknife groups. We created the r$^{\text{th}}$ set of jackknife replicate weights by setting the replicate weights of a respondent k in the r$^{\text{th}}$ group to $w_{k(r)} = 0$ zero. For a respondent outside the r$^{\text{th}}$ group, the $w_{k(r)} = 0$ were determined by finding the $g_{(r)}$ that solved the calibration equation:

$$\left(\frac{n_G}{n_G-1}\right)\sum_{k\in S_{(r)}}\left[w_k\left(\exp\left(\frac{w_k-1}{w_k}x_k^T g(r)\right)\right)\right]z_k = \boldsymbol{T}_z$$

$$\left(\frac{n_G}{n_G-1}\right)\sum_{S_{(r)}} w_{k(r)}z_k = \boldsymbol{T}_z \quad , \qquad (5)$$

where $S_{(r)}$ is the summation is over the respondents not in the r$^{\text{th}}$ group. Dag jackknife variance estimates have the form:

$$\mathrm{Var}_{JK}(t) = \frac{n_G-1}{n_G}\sum_{r=1}^{n_G}(\tau-\tau_{(r)})^2 \quad ,$$

where τ is a smooth function of estimators like $t = \sum_S w_k y_k$ each computed with the calibrated weights, and $\tau_{(r)}$ is the analogous function of estimators computed with the rth set of replicate weights.

The first line of equation (4) is asymptotically equivalent (as n_G grows arbitrarily large) to a linearized version of the dag jackknife, where the rth set of dag jackknife weights are determined by finding a $g_{(r)}$ that solves the linear calibration equation:

$$\frac{n_G}{n_G-1}\sum_{S_{(r)}}\left[w_k\left(1+\frac{w_k-1}{w_k}\boldsymbol{x}_k^T\boldsymbol{g}_{(r)}\right)\right]z_k = T_z \quad . \qquad (6)$$

Table 3
Smallest 9 p-values of the Holm-Bonferroni Results for the Difference Between the Mail-to-Web and Facebook Cohorts

	Estimated difference	p-value	HB40_.1 HB20_.05	HB20_.1	HB40_.05
More DUI?	0.11	0.00247	0.00250	0.00500	0.001000
Edible MJ in public?	–0.23	0.00371	0.00256	0.00526	0.001026
How legal?	0.11	0.00658	0.00263	0.00556	0.001053
Adult frequency?	–0.13	0.01619	0.00270	0.00588	0.001081
Is edible MJ safer?	–0.17	0.02260	0.00278	0.00625	0.001111
Guest use in home?	–0.18	0.04079	0.00286	0.00667	0.001143
Is vaping safer?	0.10	0.05260	0.00294	0.00714	0.001176
More teenage use?	0.12	0.08722	0.00303	0.00769	0.001212
Response to vaping Q	0.05	0.09704	0.00313	0.00833	0.001250

(Kott, 2006, under a design with one stratum and 30 primary sampling units). This is because when $n_G < n$ is large and the calibration equation in (1) holds, each $g_{(r)}$ must be close to zero (technically $O_P(1/\sqrt{n_G})$ under mild conditions), so

$$\exp\left(\frac{w_k - 1}{\boldsymbol{w}_k}\boldsymbol{x}_k^T\boldsymbol{g}_{r)}\right) \approx 1 + \frac{w_k - 1}{w_k}\boldsymbol{x}_k^T\boldsymbol{g}_{(r)} \quad .$$

When $\alpha(\boldsymbol{x}_k^T\boldsymbol{g}) = 1 + \exp(\boldsymbol{x}_k^T\boldsymbol{g}), \alpha'(\boldsymbol{x}_k^T\boldsymbol{g}) = \exp(\boldsymbol{x}_k^T\boldsymbol{g})$, which is the $w_k - 1$ in equation (5). With some work, we can see that the linearized dag jackknife in equation (6) works for an estimator like $t = \sum_S w_k y_k$. Observe

$$\begin{aligned} t - t_{(r)} &= \sum_S w_k e_k^* - \sum_S w_{k(r)} e_k^* \\ &= \sum_{S_r} w_k e_k^* - \frac{1}{30}\sum_{S_{(r)}} w_k e_k^* - \frac{29}{30}\sum_{S(r)} (w_k - 1)\boldsymbol{x}_k^T\boldsymbol{g}_{(r)} e_k^* \\ &= \sum_{S_r} w_k e_k^* - \frac{1}{30}\sum_{S_{(r)}} w_k e_k^* \\ &\quad - \left(\frac{29}{30}\boldsymbol{g}_{(r)}^T \sum_S \alpha'(\boldsymbol{x}_k^T\boldsymbol{g}) x_k e_k^* - \frac{29}{30}\sum_{S_r}(w_k - 1)\boldsymbol{x}_k^T\boldsymbol{g}_{(r)} e_k^*\right) \\ &\approx \sum_{S_r} w_k e_k^* \quad , \end{aligned}$$

where S_r is the r^th^ jackknife group (because the respondent selections are independent, $\frac{1}{30}\sum_{S_{(r)}} w_k e_k^*$ is $O_P(1/\sqrt{n_G})$ under mild conditions; in addition the components of $\sum_S \alpha'(\boldsymbol{x}_k^T\boldsymbol{g})\boldsymbol{x}_k e_k^*$ and $\boldsymbol{g}_{(r)}$ are $O_P(1/\sqrt{n_G})$ under mild conditions. The independence of the selections implies $E\left((t - t_{(r)})^2\right) \approx E\left(\left(\sum_{S_r} w_k e_k^*\right)^2\right) = \sum_{S_r} E\left(\left(w_k e_k^*\right)^2\right)$, so $E\left(\mathrm{Var}_{JK}(t)\right) \approx \sum_S E\left(\left(w_k e_k^*\right)^2\right)$.

A solution to the linear calibration equation in (6) almost always exists. When the standard approach for creating jackknife weights is used (i.e., employing equation (1), replacing S and $\boldsymbol{g}$ with $S_{(r)}$ and $\boldsymbol{g}_{(r)}$ by contrast, a solution might not exist for some r. The advantage of using equation (5) to create dag jackknife weights over its nearly-identical linear alternative in equation (6) is that jackknife weights will always be positive, which is not assured using equation (5). With the Oregon marijuana data, all 30 jackknife calibration equations in (5) had a solution.

To use WTADJX to compute dag jackknife weights with equation (5), we set L in equation (3) to 0 and the model variables to $\frac{w_k - 1}{w_k}\boldsymbol{x}_k$, where the $\boldsymbol{x}_k$ were as defined for equation (1).

We needed to scale the calibrated and dag jackknife weights assigned to the mail-to-web (by 0.65) and recruit (by 0.35) cohorts to eliminate the double counting noted in Section 5. Scaling factors were chosen that roughly minimized the estimated standard error of the "What is your opinion about legalizing the use of marijuana by adults?" item.

Computing the estimated standard errors of the 40 differences with jackknife weights (and DIFFVAR) rather than through WTADJX (which uses linearization; see Kott, 2006) increased SE measures by 4.8% on average ($\log\left(\frac{\mathrm{SE}_{\mathrm{JK}}}{\mathrm{SE}_{\mathrm{WTADJX}}}\right)$); 6.0% was the median, while interquartile range extended from 1.0% to 12.1%. This is consistent with theory (linearization tends to underestimate calibrated estimates' SEs, replication to overestimate; see Kott, 2006.

Incorporating the recruit cohort into the ABS sample decreased estimated SEs by 8.6% on average (comparing jackknife estimated SE to jackknife estimated SE); 7.5% was the median, interquartile range extended from 4.2% to 12.1%. These despite the nearly 46% increase in respondent sample size realized by the nonprobability recruiting.

8 Concluding Remarks

We offer the following concluding remarks. One needs to think about the analysis to be done before data are collected. In the Oregon Marijuana sample, for example, certain items values were frequently missing that would have been useful in weighting the sample. More insistence should have been placed on collecting them.

Using nonprobability samples relies on assumptions. These assumptions need to be stated clearly to users and tested when possible as they were in Sections 3 and Section 6 respectively.

Statistical testing, however, has its limits. An estimated difference not being statistically significant at the .1 level does not assure one that the actual difference, in this case between the estimand for the Facebook and mail-to-web cohorts, is 0. The recent literature has been bombarded with diatribes against treating statistical tests based on p-values as ultimate authorities (see, for example, Wasserstein & Lazar, 2016). Here Bonferroni-adjusted p-values sensitive to the multitude of comparison being made were not used to analyze the data per se but to assess the reasonableness of an assumption that samples from the two cohorts were estimating the same things. Moreover, the reader should remember that the overall Bonferroni test failed: the assumption that the two cohorts always estimated the same things failed for (at least) one variable at the 10% level, which is all that is necessary for an overall Bonferroni test to fail. The Holm variation allowed us to assess whether the difference estimates from the two population were significant for individual variables.

When appropriately calibrated (using WTADJX or an equivalent program in R) the decrease in estimated SE under the assumed selection model from adding nonprobability samples is often considerably less than the sample-size increase implies as we saw here.

We considered other items for calibrating the recruit cohort to the mail-to-web cohort in our Oregon marijuana sample. One such was whether the respondent ever tried marijuana. The results in terms number of failures of the Holm-Bonferroni were near identical. Recall that the survey data were collected before serious thought was given to the estimation to be derived from them. Because of this, the methods described in this paper are more relevant than the results on marijuana attitudes and use in Oregon.

Acknowledgements

The author wishes to thank Matthew Farrelly and Kian Kamyab for their work developing and implementing the Oregon Marijuana Survey and Joseph McMichael for his help in developing the analysis weights for the ABS sample. Jane Allen helped me understand the survey a little better. All are with RTI International. Unfortunately, the data used in this analysis are preliminary and not available to the public.

References

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Kim, J. & Riddles, M. (2012). Some theory for propensity scoring adjustment estimator. *Survey Methodology*, *38*(2), 57–165.

Kott, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, *17*(4), 521–526.

Kott, P. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, *32*(6), 133–142.

RTI International. (2012). *SUDAAN language manual. Release 11.0*. Research Triangle Park, NC: RTI International.

Singh, A., Dever, J., & Iannacchione, V. (2004). Efficient estimation for surveys with nonresponse follow-up using dual-frame calibration. In *Proceedings of the American Statistical Association Survey Research Methods section* (pp. 3919–3930).

Tille, Y. & Matei, A. (2013). Package ‘Sampling’ (procedure: gencalib). Software. Retrieved from http://cran.r-project.org/web/packages/sampling/sampling.pdf

Wasserstein, R. & Lazar, N. (2016). The ASA’s statement on p-value: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133.

Appendix
Items Used in Holm-Bonferroni Analysis

The following 20 items were used in the Holm-Bonferroni analysis in Section 6. Whether there was a valid response to the item was treated as a variable. When there was a response, the numeric value of the response given below was treated as a continuous variable (the valuess of the responses to I20 were reordered to make them ordinal).

I1 Do you now smoke cigarettes

I2 Do you now smoke electronic cigarettes

I3 Do you now drink alcohol

$\square_1$ Every day
$\square_2$ Some days
$\square_3$ Rarely
$\square_4$ Not at all
$\square_.$ Don't know/ prefer not to answer

I4 When you drink, how many drinks do you usually have?

$\square_1$ One
$\square_2$ Two
$\square_3$ Three
$\square_4$ Four or more
$\square_.$ Don't know/ prefer not to answer

I5 What is your opinion about legalizing the use of marijuana by adults?

I6 What do most people in your state think about legalizing the use of marijuana use by adults?

$\square_1$ It should not be legal for any purpose
$\square_2$ It should be legal only for medical use
$\square_3$ It should also be legal recreational use
$\square_.$ Don't know/ prefer not to answer

I7 What is your opinion about the use of marijuana by adults

I8 What is your opinion about the use of marijuana by teenagers?

$\square_1$ It is okay to use every day
$\square_2$ It is okay to use some days
$\square_3$ It is not okay to use at all
$\square_.$ Don't know/ prefer not to answer

I9 Would you allow guests to use marijuana in your home?

I10 Would it bother you if people were smoking marijuana in public?

I11 In your opinion should people be allowed to use edible marijuana in places they are not allowed to smoke it?

I12 In your opinion is edible marijuana, such as food or candy, safer to use than marijuana that is smoked?

I13 In your opinion is vaping marijuana, such as through an e-cig or e-vaporizer device, safer than smoking marijuana in a joint or pipe?

I14 In your opinion, does legalization of medical marijuana lead to more teenagers trying marijuana?

I15 In your opinion, does the legalization of recreational marijuana lead to more teenagers trying marijuana?

$\square_1$ Definitely yes
$\square_2$ Probably yes
$\square_3$ Probably not
$\square_4$ Definitely not
$\square_.$ Don't know/ prefer not to answer

I16 Have you ever tried marijuana, even one time?

I17 In your opinion, does the legalization of recreational marijuana lead to more people driving under the influence of marijuana?

I18 Do you think people convicted of possessing more than an allowable amount of marijuana should serve time in jail?

I19 Are you aware of any stores or shops in or near your community that sell marijuana?

$\square_1$ Yes
$\square_2$ No
$\square_.$ Don't know/ prefer not to answer

I20 Now that recreational marijuana is legal in Oregon, will your usage...

$\square_1$ Increase
$\square_2$ Stay the same
$\square_3$ Decrease
$\square_.$ Don't know/prefer not to respond