Inferences from Nonrandom Samples with Model-Implied Randomization [1]

Vladislav Beresovsky (hvy4@cdc.gov)

National Center for Health Statistics, CDC

[1] Disclaimer: The findings and conclusions in this presentation are those of the author(s) and do not necessarily represent the views of the Centers for Disease Control and Prevention.

Main question and available data

Is it possible to make reliable inferences of NHIS variables of interest from a sample collected by a web survey?

In 2015 NCHS commissioned a commercial vendor to collect responses on a subset of NHIS questions using web survey distributed to the members of its online panel.

For the IV Quarter were collected:

- *Regular random* sample $S^{\text{rand}}\left(\mathbf{Y}, \mathbf{X}^d, \mathbf{X}^c, W\right)$ of size $n^{\text{rand}} = 7,723$;
- *Nonrandom* sample from web survey $S^{\text{non}}\left(\mathbf{Y}, ..., \mathbf{X}^c, ...\right)$ of size $n^{\text{non}} = 2,304$;

- $\mathbf{Y}$ - variables of interest: insurance, chronic conditions, alcohol/tobacco/physical exercise, access to health care, food security, ...;
- $\mathbf{X}^c$ - demographic variables + BMI, self assessed health status;
- $\mathbf{X}^d$ - sample design, such as strata and PSU identifiers;
- W - NHIS sampling weights.

What we tried to accomplish?

Propose general inferential methodology ...

- for robust point estimation of general population means;
- to estimate variances;
- to validate usability of data collected from a web survey by comparing estimates from S^{rand} and S^{non} samples, which means testing some hypothesis.

What will be presented?

- Model-implied conditional randomization as the foundation for inferences from nonrandom samples (an attempt of justification);
- Point estimators, variance estimators and hypothesis testing;
- Simulation study to show how it all works out;
- Application to NHIS chronic conditions variables;
- Conclusions and directions for further work.

CDC
Centers for Disease Control and Prevention

Randomization as the foundation for inferences

- *Unconditional* randomization is usually assumed for testing hypothesis of no treatment effect in Clinical Trials and estimating population characteristics from SRS data;
- *Conditional* randomization underlines testing hypothesis in Observational Studies and estimating population characteristics from Stratified SRS and PPS sampled data;
- Following analogy with biostatistics, we propose using response propensity (RP) and outcome response (OR) models to impose conditional randomization between S^{rand} and S^{non}
  - Calculate IPW from the scores of weighted RP model $e_W(\mathbf{x}^c)$ and use them as PPS weights;
  - Impose strata s_q using the scores of RP model and treat S^{rand} as StrSRS;
  - Impose strata s_{qy} with the scores of OR model $e_y(\mathbf{x}^c)$;
  - Impose more sophisticated stratification using both RP and OR models;

From observational studies to nonrandom samples

How to extend conditional randomization to allow population-level inferences from a nonrandom sample?

From observational studies to nonrandom samples

How to extend conditional randomization to allow population-level inferences from a nonrandom sample?

$$\left.\begin{array}{r} e\left(\mathbf{x}^c\right), e_y\left(\mathbf{x}^c\right) \quad \text{model-implied} \\ s_q, w = e^{-1}\left(\mathbf{x}^c\right), s_{yq} \text{ randomization} \end{array}\right\} \Rightarrow \begin{array}{c} \text{hypothesis testing} \\ \text{in observational studies} \end{array}$$

From observational studies to nonrandom samples

How to extend conditional randomization to allow population-level inferences from a nonrandom sample?

$$\left.\begin{array}{r} e\left(\mathbf{x}^c\right), e_y\left(\mathbf{x}^c\right) \quad \text{model-implied} \\ s_q, w = e^{-1}\left(\mathbf{x}^c\right), s_{yq} \text{ randomization} \end{array}\right\} \Rightarrow \begin{array}{c}\text{hypothesis testing} \\ \text{in observational studies}\end{array}$$

$$\left.\begin{array}{c} e\left(\mathbf{x}^c\right), e_y\left(\mathbf{x}^c\right) \\ s_q, w = e^{-1}\left(\mathbf{x}^c\right), s_{yq} \\ \mathbf{X}^d, W \end{array} \begin{array}{c}\text{model-implied} \\ \text{randomization} \\ \text{survey} \\ \text{design}\end{array}\right\} \not\Rightarrow \begin{array}{c}\text{population-level} \\ \text{inferences}\end{array}$$

Population inferences from nonrandom samples (IPW)

Randomization with IPW.
Folowing Beresovsky(2016), IPW for S^{non} may be defined by scores of *weighted* RP model, using S^{rand} as a reference

$$w_i^{\text{pop,IPW}} = [1 - e_W(\mathbf{x}_i^c)]/e_W(\mathbf{x}_i^c)$$

S^{non} is treated as PPS sample with no strata.

$$\hat{\bar{Y}}^{\text{IPW}} = \frac{1}{N} \sum_{i \in S^{\text{non}}} w_i^{\text{pop,IPW}} y_i$$

$$\hat{V}\left(\hat{\bar{Y}}^{\text{IPW}}\right) = \frac{n^{\text{non}}}{N^2 (1 - n^{\text{non}})} \sum_{i \in S^{\text{non}}} \left(w_i^{\text{pop,IPW}} y_i - \frac{\sum_{i \in S^{\text{non}}} w_i^{\text{pop,IPW}} y_i}{n^{\text{non}}} \right)^2 \times$$

Population inferences from nonrandom samples (StrSRS)

Randomization within quantiles s_q of weighted RP or unweighted OR scores $e_W(\mathbf{x}^c), e_y(\mathbf{x}^c)$.
Quantile weights

$$w_q^{\text{pop,sub}} = N_q / n_q^{\text{non}} ,$$

where n_q^{non} is the number of S^{non} units and $N_q = \sum_{i \in s_q} W_i$ is estimated population count in a quantile q.
Point and variance estimators

$$\hat{\bar{Y}}_{\text{RP}}^{\text{StSRS}} = \frac{1}{N} \sum_{q=1}^{Q} N_q \bar{y}_{s_q}, \quad \bar{y}_{s_q} = \sum_{i \in s_q} y_i / n_q^{\text{non}}$$

$$\hat{V}\left(\hat{\bar{Y}}_{\text{RP}}^{\text{StSRS}}\right) = \frac{1}{N^2} \sum_{q=1}^{Q} \frac{N_q^2}{n_q^{\text{non}}} S_{y,s_q}^2, \quad S_{y,s_q}^2 = \frac{\sum_{i \in s_q} \left(y_i - \bar{y}_{s_q}\right)^2}{n_q^{\text{non}} - 1}$$

Comparing estimates from S^{non} and S^{rand} samples

Cochran (1954) and Mantel-Haenszel (1959) large-sample test of stratum-adjusted independence in 2x2xQ contingency tables for binomial outcome variable Y within Q strata

$$\text{CMH}=\left[\sum_{q=1}^{Q}\left(\hat{Y}_q^{\text{non}}-E\left(\hat{Y}_q^{\text{non}}\right)\right)\right]^2\Big/\hat{V}\left[\sum_{q=1}^{Q}\left(\hat{Y}_q^{\text{non}}-E\left(\hat{Y}_q^{\text{non}}\right)\right)\right]$$

where $\hat{Y}_q^{\text{non}}=\sum_{i\in S_q^{\text{non}}} w_i^{\text{pop}} y_i$ and $E\left(\hat{Y}_q^{\text{non}}\right)=\hat{Y}_q^{\text{non+rand}}\frac{\hat{N}_q^{\text{non}}}{\hat{N}_q^{\text{non+rand}}}$.

Test statistic is χ_1^2, see Agresti (2002) Expr 6.6 and SUDAAN documentation Chap. 14.9.3.3.

It is implemented in SUDAAN, so it can fully account for survey design of S^{rand} and be applied for IPW conditional randomization of S^{non}. Lumley and Scott (2013) propose Wilcoxon rank sum test for continuous Y accounting for complex sample design.

Pairing point estimators with conditional randomization

Each considered point estimator corresponds to a certain conditional randomization depending on covariates $\mathbf{x}_i^c$.

Estimator	Conditional randomization
$\hat{\bar{Y}}^{\text{IPW}}$	with IPW, $w_i^{\text{pop,IPW}}$
$\hat{\bar{Y}}_{\text{RP}}^{\text{StSRS}}$	quantiles of weighted RP scores $e_W\left(\mathbf{x}_i^c\right)$
$\hat{\bar{Y}}_{\text{OR}}^{\text{StSRS}}$	quantiles of unweighted OR scores $e_y\left(\mathbf{x}_i^c\right)$
$\hat{\bar{Y}}_{\text{RP}\times\text{OR}}^{\text{StSRS}}$	*intersecting* quantiles of RP and OR scores;
$\hat{\bar{Y}}_{\text{OR,RP-align}}^{\text{StSRS}}$	quantiles of scores of the OR model $e_{\text{yp}}\left(\mathbf{X}^{c,al}\right) = E\left(Y \mid \bar{Y}_{\text{p}}, \mathbf{X}^{\text{c,al}}\right)$ with intersect $\bar{Y}_{\text{p}} = \sum_{i \in s_q} Y_i / n_q^{\text{non}}$ defined for quantiles of RP model Testing with aligned blocks by subtracting their means was proposed by Hodges and Lehmann (1962). Rosenbaum (2002) proposed in addition to regress on aligned covariates. Using random intercept model may work even better.

Simulated population and sampling of S^{non} and S^{rand}

Simulated population of size $N = 50,000$ with identically distributed normal covariates $U_{01-04} \sim N(0,1)$.
Added some noise $U_{1-4} = U_{01-04} + N\left(0, \sigma_n^2\right)$, $\sigma_n^2 = 1.44$.
Covariates X_{1-4} are obtained by nonlinear transformation:
$X_1 = \exp(U_1/2)$, $X_2 = U_2/(1 + \exp(U_1)) + 10$,
$X_3 = (U_1 + U_3/25 + 0.6)^3$, $X_4 = (U_2 + U_4 + 20)^2$.
X_{1-4} are considered "observed", but population and sampling characteristics depend on "true" covariates U_{01-04}, see Schafer and Kang (2007).
Bernoulli outcome variable Y

$$\text{logit}(p_y) = -2 + 2U_{01} + U_{02} + U_{03} + U_{04}$$

S^{non} of size $n^{\text{non}} = 300$ is sampled with PPS

$$\text{logit}(p^{\text{non}}) = -1 + U_{01} - 0.5U_{02} + 0.25U_{03} + 0.1U_{04}$$

S^{rand} of size $n^{\text{rand}} = 900$ is also PPS with known measure of size.

Measured model misspecification

$I_i^{yc} \sim$ Bernoulli (P^{yc}) identifier specifies which model is correctly specified for each sampled unit.
If $I_i^{yc} = 1$ OR model uses "true" covariates U_{01-04} and RP uses "observed" covariates X_{01-04}. If $I_i^{yc} = 0$, it's *vice versa*.
Simulations were conducted for $P^{yc} = 0, 0.5, 1$.
$P^{yc} = 0$ - RP model is correct, OR is misspecified for all units;
$P^{yc} = 1$ - OR model is correct, RP is misspecified for all units;
$P^{yc} = 0.5$ - both models are misspecified for $\sim 50\%$ of the sampled units. But either one of the models is correct for every sampled unit.

Bias reduction (BR) of point estimates

BR of an estimator $\hat{\bar{Y}}^{\text{est}}$ shows percent of bias of a direct estimator $\hat{\bar{Y}}^{\text{web}}$ removed by this estimator

$$\text{BR} = \left(\text{1-abs}\left(\frac{\hat{\bar{Y}}^{\text{est}} - \bar{Y}^{\text{pop}}}{\hat{\bar{Y}}^{\text{web}} - \bar{Y}^{\text{pop}}}\right)\right) * 100\%$$

P^{yc}	$\hat{\bar{Y}}^{\text{IPW}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{RP}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{OR}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{RP}\times\text{OR}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{OR,RP-align}}$
0	87%	98%	24%	93%	96%
0.5	58%	73%	78%	96%	96%
1	18%	20%	98%	96%	96%

Estimates of standard error (SE)

SE^{MC} -SD of MC variability of estimates.
$\widehat{SE}^{q}$ - estimated SE using quantiles of conditional randomization (when applicable). Underestimates SE for quantiles of OR score because model variability is unaccounted.
$\widehat{SE}^{P}$ - estimated SE using quantiles of RP score. Less underestimated, but a bit *ad hoc*. Rigorous method for variance estimation is still in order.

	P^{yc}	$\hat{\hat{Y}}^{\text{IPW}}$	$\hat{\hat{Y}}^{\text{StSRS}}_{\text{RP}}$	$\hat{\hat{Y}}^{\text{StSRS}}_{\text{OR}}$	$\hat{\hat{Y}}^{\text{StSRS}}_{\text{RP}\times\text{OR}}$	$\hat{\hat{Y}}^{\text{StSRS}}_{\text{OR,RP-align}}$
SE^{MC}	0.0	0.026	0.028	0.028	0.027	0.030
$\widehat{SE}^{q}/SE^{MC}$		1.09	0.99	0.95	0.97	0.69
$\widehat{SE}^{P}/SE^{MC}$				0.97	1.02	0.79
SE^{MC}	0.5	0.028	0.030	0.028	0.027	0.029
$\widehat{SE}^{q}/SE^{MC}$		0.98	0.91	0.90	0.93	0.76
$\widehat{SE}^{P}/SE^{MC}$				0.95	0.99	0.84
SE^{MC}	1.0	0.027	0.031	0.023	0.024	0.025
$\widehat{SE}^{q}/SE^{MC}$		1.05	1.01	0.78	0.85	0.72
$\widehat{SE}^{P}/SE^{MC}$				1.08	1.12	1.00

CDC
Centers for Disease Control and Prevention

Validating estimates from S^{non} by hypotheses testing

t-test for an estimate of the mean from S^{non} – $H_0^t : \hat{\bar{Y}}^{\text{non}} = \bar{Y}^{\text{pop}}$

CMH test of independence between S^{non} and S^{rand} – $H_0^{CMH} : \hat{\bar{Y}}^{\text{non}} = \hat{\bar{Y}}^{\text{rand}}$

t-test tests exactly what we like to know, but is impossible with real data.

CMH test is practically possible, but indirect. Simulations help to justify its use to validate estimates from S^{non}.

	P^{yc}	$\hat{\bar{Y}}^{\text{IPW}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{RP}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{OR}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{RPxOR}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{OR,RP-align}}$
t-test, $\widehat{SE}^q$	0.0	0.95*	0.94	0.10	0.95	0.83
t-test, $\widehat{SE}^p$				0.11	0.96	0.88
CMH		0.97	0.97	0.17	0.97	0.90
t-test, $\widehat{SE}^q$	0.5	0.59	0.81	0.82	0.93	0.86
t-test, $\widehat{SE}^p$				0.85	0.95	0.91
CMH		0.68	0.83	0.86	0.94	0.90
t-test, $\widehat{SE}^q$	1.0	0.08	0.14	0.88	0.91	0.85
t-test, $\widehat{SE}^p$				0.97	0.97	0.95
CMH		0.16	0.23	0.92	0.93	0.91

* Test acceptance rate over the simulations. 95% is nominal value.

Odds ratio (OR) between estimates from S^{non} and S^{rand}

Graubard, Fears and Gail (1989) adapted Mantel-Haenszel common odds ratio estimator to population-based studies

$$\widehat{\text{MHOR}} = \frac{\sum_{q=1}^{Q} \left(\hat{N}_q^{Y=1,\text{non}} \hat{N}_q^{Y=0,\text{rand}} / \hat{N}_q^{\text{rand+non}} \right)}{\sum_{q=1}^{Q} \left(\hat{N}_q^{Y=0,\text{non}} \hat{N}_q^{Y=1,\text{rand}} / \hat{N}_q^{\text{rand+non}} \right)}$$

SUDAAN also calculates its lower and upper confidence bounds, see Chapt. 14.9.5.

	P^{yc}	$\hat{\bar{Y}}^{\text{IPW}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{RP}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{OR}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{RP}\times\text{OR}}$	$\hat{\bar{Y}}^{\text{StSRS}}_{\text{OR,RP-align}}$
OR	0.0	1.09	1.03	1.63	1.07	1.05
CI Lower		0.79	0.72	1.19	0.73	0.75
CI Upper		1.49	1.46	2.24	1.55	1.47
OR	0.5	1.27	1.20	1.18	1.04	1.04
CI Lower		0.95	0.86	0.85	0.74	0.75
CI Upper		1.71	1.68	1.64	1.48	1.45
OR	1.0	1.55	1.57	1.05	1.06	1.07
CI Lower		1.16	1.15	0.70	0.71	0.73
CI Upper		2.09	2.14	1.56	1.60	1.58

Chronic Conditions from NHIS Web Survey Sample

Variable	Description
Diabetes (R)	Have you been told by a doctor that you have diabetes or sugar diabetes?
Taking Pill	Are you NOW taking diabetic pills to lower your blood sugar?
Taking Insulin	Are you NOW taking insulin?
Hypertension	Have you been told by a doctor that you have hypertension ?
Meds Ever	Has a doctor ever prescribed any medicine for your high blood pressure?
Meds Now	Are you NOW taking any medicine for your high blood pressure?
Asthma Ever	Have you been told by a doctor that you have asthma?
Asthma Still	Do you still have asthma?
Asthma Attack	During the past 12 months have you had an episode of asthma, or an asthma attack?
Asthma ER	During the past 12 months have you had to visit an ER because of asthma?
Lungs Problem (R)	Do you have any one of the chronic lung problems?
Emphysema	Have you ever been told by a doctor that you had emphysema?
COPD	Have you ever been told by a doctor that you had chronic obstructive pulmonary disease?
Emphysema/COPD (R)	Have you ever been told by a doctor that you had either Emphysema or COPD?
Chronic Bronchitis	Have you ever been told by a doctor that you had chronic bronchitis?

Covariates: Age, Gender, Region, Race/Ethnicity, Education, Income Group, Marital Status, General Health, BMI, (Gateway variables for Follow ups)

CDC
Centers for Disease Control and Prevention

PRB $\left(\hat{Y}^{est}\right)$ $\left(\left(\hat{Y}^{est} - \hat{Y}^{NHIS}\right)/\hat{Y}^{NHIS}\right)$ and CMH test for chronic conditions

	$\hat{\hat{Y}}^{WEB}$	$\hat{\hat{Y}}^{IPW}$	$\hat{\hat{Y}}^{StSRS}_{RP}$	$\hat{\hat{Y}}^{StSRS}_{OR}$	$\hat{\hat{Y}}^{StSRS}_{RPxOR}$	$\hat{\hat{Y}}^{StSRS}_{OR,RP\text{-}align}$
Diabetes	0.13	-0.09	-0.07	0.06	-0.12	0.03
Taking Pill	0.22			0.13	0.09	0.14
Taking Insulin	0.03		-0.19			-0.20
Hypertension	0.16	0.03	0.03	0.05	0.02	0.04
Meds Ever	0.17	0.01	0.03	0.01	0.00	0.01
Meds Now	0.17	-0.01	0.0	-0.02	-0.02	-0.01
Asthma Ever	0.32	0.23	0.22	0.23	0.19	0.20
Asthma Still	0.34	-0.02	0.02	-0.01	0.01	0.01
Asthma Attack	0.84	0.36	0.45	0.40	0.34	0.35
Asthma ER	0.52	0.26	0.41	0.23	0.37	0.33
Lungs Problem (R)	0.44	0.38	0.36	0.51	0.31	0.56
Emphysema	-0.12	-0.36	-0.31	-0.15	-0.33	-0.35
COPD	0.15	-0.20	-0.24	-0.04	-0.13	-0.02
Emphesyma/ COPD (R)	0.16	-0.20	-0.22	-0.05	-0.15	-0.14
Chronic Bronchitis	0.68	0.24	0.18	0.09	0.08	0.06

CMH $p > 0.2$, green: Estimates are reliable;
$0.05 < p \leq 0.2$, yellow: Estimates are partially reliable;

$p \leq 0.05$, red: Estimates are unreliable.

CDC
Centers for Disease Control and Prevention

Reasons for biased estimates from nonrandom samples

Rosenbaum and Rubins (1983, 1984) defined "conditionally strongly ignorable treatment assignment" as

$$\Pr\left(r_1, r_0, z|\mathbf{X}^c\right) = \Pr\left(r_1, r_0|\mathbf{X}^c\right)\Pr\left(z|\mathbf{X}^c\right)$$

If RP, OR models are misspecified on population level because observed covariates are incorrect, then it becomes

$$\Pr\left(Y^{\text{WEB}}, Y^{\text{NHIS}}, z|\mathbf{X}^{\text{obs}}\right) = \Pr\left(Y^{\text{WEB}}, Y^{\text{NHIS}}|\mathbf{U}^{\text{corr}}\right)\Pr\left(z|\mathbf{U}^{\text{corr}}\right) + \Pr\left(Y^{\text{WEB}}, Y^{\text{NHIS}}, z|\left(\mathbf{X}^{\text{obs}} - \mathbf{U}^{\text{corr}}\right)\right)$$

The first term is bias due to the difference in survey responses depending on "survey mode", like "treatment effect" in biostatistics.

The second term comes from unresolved correlations between outcome variable and survey mode of data collection, because of insufficient information in the observed covariates to impose conditional randomization.

Conclusions and further work

- A principal contribution to producing reliable estimates from nonrandom samples will come from **survey methodologists and sociologists**. Their fine work must ensure that "treatment effect" is minimized and covariates necessary to impose conditional randomization are collected;
- **Math stats** must provide rigorous statistical justification for using model-implied conditional randomization for population-level inferences from nonrandom samples. Particularly, variance estimation must be better worked out;
- Employ techniques for robust estimation for both RP and OR models, as was suggested by Rosenbaum (2002).

CDC
Centers for Disease Control and Prevention

The Quiz

One word was never mentioned in my slides. What is it?

The Quiz

One word was never mentioned in my slides. What is it?

IMPUTATION

The Quiz

One word was never mentioned in my slides. What is it?

IMPUTATION

If there are any questions, please contact:

`Vberesovsky@cdc.gov`