Statistical Analysis Using Combined Data Sources: Discussion
2011 JPSM Distinguished Lecture
University of Maryland

Michael Elliott[1]

[1]University of Michigan School of Public Health

April 2011

"Complete" (Ideal) vs. Observed (Messy) Data

- "Complete" data $\mathbf{d}_U$ vs. observed data $\mathbf{d}_s$:
  - Data with measurement error (X vs. X^*): last month vs. this month employee count.
  - Partial Z: "large" area IDs vs. small area IDs.
  - Non-overlapping covariates: health outcomes X vs. behavioral risk factors W.
  - Multiple surveys.

"Complete Data"

	Y	X	W	Z
1	■	■	■	■
2	■	■	■	■
3	■	■	■	■
4	■	■	■	■
5	■	■	■	■
6	■	■	■	■
⋮	⋮	⋮	⋮	⋮
N	■	■	■	■

"Observed Data"

Survey A

	Y	X	X*	W	Z
1	■	■			■
2	■				
3		■			■
4					
5					■
6					■
⋮	⋮	⋮	⋮	⋮	⋮
N					■

Survey B

	Y	X	X*	W	Z
1	■		■	■	■
2					■
3					■
4					■
5	■			■	■
6	■		■	■	■
⋮	⋮	⋮	⋮	⋮	⋮
N					■

...

Missing Information Principle

- Would like to make inference about $Q(Y_1, ..., Y_N, X_1, ..., X_N) = Q(\mathbf{X}, \mathbf{Y})$ using $\mathbf{d}_U$
- Stuck with making inference about $Q(\mathbf{X}, \mathbf{Y})$ using $\mathbf{d}_s$
- Chambers: Use "missing information principle" to obtain likelihood inference about $\theta \equiv \theta_N \equiv Q(\mathbf{X}, \mathbf{Y})$ using $\mathbf{d}_s$ by replacing sufficient statistics $S(\mathbf{d}_U)$ for θ_N by $E(S(\mathbf{d}_U) \mid \mathbf{d}_s)$.
    - EM algorithm
    - Replace likelihood inference with pseudo-likelihood inferences to accommodate unequal probability sample designs.

Missing Information Principle

- Alternative to calibration: calibration data becomes part of $\mathbf{d}_s$, and replace data in MLE/PMLE score equations with expected value conditional on calibration totals.
  - Chambers shows that this can improve efficiency over uncalibrated estimators, and avoid bias due to failure of the implied calibration model.
- Probabilistic linkage: allows for bias correction in estimating equations using probabilistic linked datasets.
- Combining data from multiple surveys using MIP extension of missing data algorithms.

Bayesian Survey Inference

Focus on inference about $Q(\mathbf{X}, \mathbf{Y})$ using posterior predictive distribution based on $p(\mathbf{Y}_{nobs}, \mathbf{X}_{nobs} \mid \mathbf{y}, \mathbf{x})$:

$$p(\mathbf{Y}_{nobs}, \mathbf{X}_{nobs} \mid \mathbf{y}, \mathbf{x}) = \frac{p(\mathbf{Y}, \mathbf{X})}{p(\mathbf{y}, \mathbf{x})} =$$

$$\frac{\int p(\mathbf{Y}, \mathbf{X} \mid \theta) p(\theta) d\theta}{p(\mathbf{y}, \mathbf{x})} =$$

$$\frac{\int p(\mathbf{Y}_{nobs}, \mathbf{X}_{nobs} \mid \mathbf{y}, \mathbf{x}, \theta) p(\mathbf{y}, \mathbf{x} \mid \theta) p(\theta) d\theta}{p(\mathbf{y}, \mathbf{x})} =$$

$$\int p(\mathbf{Y}_{nobs}, \mathbf{X}_{nobs} \mid \mathbf{y}, \mathbf{x}, \theta) p(\theta \mid \mathbf{y}, \mathbf{x}) d\theta$$

(Ericson 1969; Scott 1977; Rubin 1987).

Bayesian Survey Inference vs. MIP

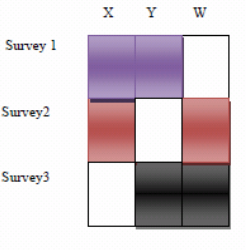
- Both Bayesian survey inference and the missing information principle focus on prediction of missing elements in the population conditional on observed data.
- Bayesian approach obtains full posterior predictive distribution of $Q(\mathbf{X}, \mathbf{Y})$, rather than point estimate and asymptotic normality assumptions
  - EM vs. MCMC ("stochastic EM")
- Bayesian survey inference should yield similar inference to MIP, at least in large samples.

Applications of Bayesian Survey Inference

- (Item level) missing data (Rubin 1987, Little and Rubin 2002).
- Weight trimming (Elliott and Little 2000, Elliott 2007, 2008, 2009).
- Disclosure risk (Raghunathan et al. 2003, Reither 2005).
- Combining data from multiple surveys (Raghunathan et al. 2007, Davis et al. 2010).

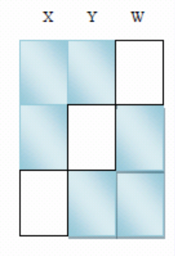
Combining Data from Multiple Complex Sample Design Surveys (Dong 2011)

- Motivating example: obtain inference about $Q(Y, X, W)$ when only two of the three variables are contained in any one survey.
- Surveys use different designs and data collection methods $\longrightarrow$ different sampling and nonsampling error properties
    - Cannot simply pool data for analysis.



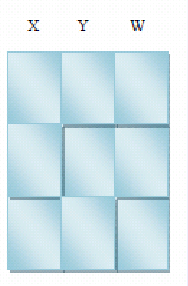
Combining Data from Multiple Complex Sample Design Surveys (Dong 2011)

- Borrow from disclosure risk literature to generate synthetic populations using data from each survey.
- Each generated population "uncomplexes" sample design to create what is effectively a simple random sample.



Combining Data from Multiple Complex Sample Design Surveys (Dong 2011)

- Pool data and use standard imputation approaches to fill in missing variables for the data from each survey.



Combining Estimates from Synthetic Populations Generated by A Single Survey (Raghunathan et al. 2003)

- Suppose we have synthetic populations $\mathcal{P}_l$, $l = 1, ..., L$ generated from $P(\mathbf{Y}_{nobs}, \mathbf{X}_{nobs} \mid \mathbf{y}, \mathbf{x})$.
- Estimate $Q = Q(X, Y)$ with $\overline{Q}_L = L^{-1} \sum_l Q_l$, where Q_l is an (asymptotically) unbiased estimator of Q generated from $\mathcal{P}_l$.
- $\mathrm{var}(\overline{Q}_L) = T_L = (1 + L^{-1})B_L - \overline{U}_L$, where $B_L = (L-1)^{-1} \sum_l (Q_l - \overline{Q}_L)(Q_l - \overline{Q}_L)^T$ is between-imputation variance and $\overline{U}_L = L^{-1} \sum_l U_l$ is the average of the within-imputation variances U_l of Q_l.
  - $T_L \approx B_L$ if generated sample large enough that U_L can be ignored and number of synthetics L large enough that $1/L$ can be ignored
- Need a bit more care if synthetic populations are generated from different surveys, since posterior predictive distribution models from different survey may not be the same.

Combining Estimates from Synthetic Populations Generated by Multiple Surveys (Dong et al. 2011)

- Generate synthetic populations $\mathcal{P}_l^s$, $l = 1, ..., L$ for surveys $s = 1, ..., S$ from $P(\mathbf{Y}_{nobs}^s, \mathbf{X}_{nobs}^s \mid \mathbf{y}^s, \mathbf{x}^s)$.
  - Account for complex sample design features; regard synthetic data as SRS from population.
  - No need to actually generate entire population, just sample large enough that between-imputation variance swamps within-imputation variance.

Combining Estimates from Synthetic Populations Generated by Multiple Surveys (Dong et al. 2011)

- For each survey, obtain $\overline{Q}_L^s$ as the synthetic population point estimator of Q for survey s and B_L^s as its variance. If
- $\hat{Q}^s \sim N(Q, U^s)$
- $Q_l^s \mid \mathbf{y}^s, \mathbf{x}^s \sim N(\hat{Q}^s, B_s)$, then
  - As $L \rightarrow \infty$, posterior predictive distribution of Q approximately $N(\overline{Q}_\infty, B_\infty)$, $\overline{Q}_\infty = \frac{\sum_s \overline{Q}_\infty^s / B_\infty^s}{\sum_s 1/B_\infty^s}$, $B_\infty = \frac{1}{\sum_s 1/B_\infty^s}$.
  - t approximation available for small L.

Combining Estimates from Synthetic Populations Generated by Multiple Surveys: Extension to Missing Data (Dong et al. 2011)

- Synthesize, then impute.
- Imputation for missing components in each survey obtained by stacking $\mathcal{P}_l^s$, $s = 1, ..., S$ and treating as SRS from population.
- Multiply impute $m = 1, ..., M$ complete datasets for each of the L synthetic populations.
- Reseparate into the surveys and obtain Q_{ml}^s for each of the multiply imputed synthetic datasets.
  - $\overline{Q}_L^s = L^{-1} \sum_l \overline{Q}_{M,l}^s$ for $\overline{Q}_{M,l}^s = M^{-1} \sum_m Q_{ml}^s$.
- $B_L^s = \frac{1+L^{-1}}{L-1} \sum_l (\overline{Q}_{M,l}^s - \overline{Q}_L^s)^2 + \frac{1+M^{-1}}{L} \sum_l (M-1)^{-1} \sum_m (Q_{ml}^s - \overline{Q}_{M,l}^s)^2$
- Combine survey-level predictive distributions as in the complete data case on previous slide.

Generating Synthetic Populations from Posterior Predictive Distribution

Derivation of predictive distribution ignores sampling indicator I; this requires (Rubin 1987)

- Unconfounded sampling
  - $P(\mathbf{I} \mid \mathbf{Y}, \mathbf{X}) = P(\mathbf{I} \mid \mathbf{y}, \mathbf{x})$
- Independence of $\mathbf{I}$ and $(\mathbf{Y}_{nobs}, \mathbf{X}_{nobs})$ given $\mathbf{y}$,$\mathbf{x}$, and $\boldsymbol{\theta}$
  - $P(\mathbf{Y}_{nobs}, \mathbf{X}_{nobs} \mid \mathbf{y}, \mathbf{x}, \mathbf{I}, \boldsymbol{\theta}) = P(\mathbf{Y}_{nobs}, \mathbf{X}_{nobs} \mid \mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$

Maintaining these assumption requires:

- Probability sample
- Model $p(\mathbf{Y}, \mathbf{X})$ attentive to design features and robust enough to sufficiently capture all aspects of the distribution of $\mathbf{Y}, \mathbf{X}$ relevant to $Q(\mathbf{Y}, \mathbf{X})$.

Non-parametric Posterior Predictive Distribution

Generate the lth synthetic population as follows:

- Account for stratification and clustering by drawing a Bayesian bootstrap sample of the clusters within each stratum.
  - For stratum h with C_h clusters, draw $C_h - 1$ random variables from $U(0,1)$ and order $a_1, .., a_{C_h-1}$; sample C_h clusters with replacement with probability $a_c - a_{c-1}$, where $a_0 = 0$ and $a_{C_h} = 1$.
- Use finite population Bayesian bootstrap Polya urn scheme (Lo 1988) extended to account for selection weights (Cohen 1997) to generate unobserved elements of the population within each cluster c in stratum h:
  - Draw a sample of size $N_{ch} - n_{ch}$ by drawing (y_k, x_k) from the ith unit among the n_{ch} sampled elements with probability $\frac{w_i - 1 + l_{i,k-1} * (N_{ch} - n_{ch})}{N_{ch} - n_{ch} + (k-1) * (N_{ch} - n_{ch})}$ where $l_{i,k-1}$ is the number of bootstrap draws of the ith unit among the previous $k-1$ bootstrap selections.
  - Repeat F times for each boostrapped cluster.

Application: Distribution of Insurance Coverage

- Estimating 2006 insurance coverage using Behavior Risk Factor Surveillance Survey (BRFSS), National Health Interview Survey (NHIS), Medical Expenditure Panel Survey (MEPS).
- NHIS and MEPS split insured into private and public: impute BRFSS using gender, race, region, education, age, and income.
- Generate synthetic populations using non-parametric Bayesian bootstrap method.

Application: Distribution of Insurance Coverage

	Type of coverage	NHIS (n=76K)	BRFSS (n=356K)	MEPS (n=34K)	Combined
Point Est. (%)	Private	74.6		73.4	75.9
	Public	7.5		13.3	8.6
	None	17.8	15.4	13.2	15.3
SE	Private	.50		.53	.23
	Public	.25		.38	.16
	None	.43	.18	.38	.16

The Role of the Model

- Bayesian bootstrap is very robust: in most settings it doesn't provide efficiency gains over design-based methods, but "combining data" situations are likely exceptions.
- In application, some issues arise:
  - BRFSS estimator may be biased (low response rate, telephone-only sampling frame).
  - MEPS is a subsample of previous year's NHIS.
  - More traditional modeling approaches needed?
- MIP uses $E\left\{\partial_\theta \log f(\mathbf{d}_u) \mid \mathbf{d}_s\right\}$
  - Target quantity of interest if model is misspecified? Solution to population score equation? Can we think of $\theta \equiv \theta_N \equiv Q(\mathbf{X}, \mathbf{Y})$?
  - Use estimating equation methodology to obtain consistent estimators of θ.

Thanks

- Qi Dong
- Trivellore Raghunathan
- NCI grant R01-CA129101

References

Cohen, M.P. (1997). The Bayesian bootstrap and multiple imputation for unequal probability sample designs, Proceedings of the Survey Research Methods Section, American Statistical Association, p. 635-638.

Davis, W.W., Parsons, V.L., Xie, D., Schenker, N., Town, M., Raghunathan, T.E., and Feuer, E.J. (2010). State-based estimates of mammography screening rates based on information from two health surveys. *Public Health Reports*, **125**, 567-578.

Dong, Q., (2011). A principled method to combine information from multiple complex surveys. Unpublished Ph.D. thesis, University of Michigan.

Elliott, M.R., and Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, **16**, 191-209.

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, **33**, 23-34.

Elliott, M.R. (2008). Model averaging methods for weight trimming, *Journal of Official Statistics*, **24**, 517-540.

Elliott, M.R. (2009). Model averaging methods for weight trimming in generalized linear regression models, J*ournal of Official Statistics*, **25**, 1-20.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society*, **B31**, 195-234.

Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2^nd Ed., New York: Wiley.

Lo, A.Y. (1987). A large sample study of the Bayesian bootstrap, *Annals of Statistics*, **15**, 360-375.

Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. (2003) . Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics*, **19**, 1-16.

Raghunathan, T.E., Xie, D., Schenker, N., Parsons, V.L., Davis, W.W., Dodd, K.W., and Feuer, E.J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk Factors and Screening, *Journal of the American Statistical Association*, **102**, 474-486.

Reiter, J.P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, **168**, 185-205.

Rubin, D.B. (1987). *Multiple Imputation for Non-response in Surveys*, New York: Wiley.

Scott, A.J. (1977). Large sample posterior distributions in finite populations. *The Annals of Mathematical Statistics*, **42**, 1113-1117.