Assimilating Dual Panel Surveys to Generate Population Estimates

Marcin Hitczenko

Consumer Payment Research Center
Federal Reserve Bank of Boston

August 9, 2015

- Goal: estimate a population proportion, p.
  Example: Proportion of U.S. adults who own a credit card.
- Data come from two separate methodologies of collecting data.
- Samples *do not* represent dual frames $\Rightarrow$ dual frame methods do not apply.
- No a priori knowledge of differences in sampling distributions.

Example 1: 2012 Survey of Consumer Payment Choice (SCPC)

- $\sim$ 2,000 American Life Panel (ALP) panelists who participated in SCPC of previous years
- $\sim$ 1,000 ALP panelists who were newly recruited in 2012

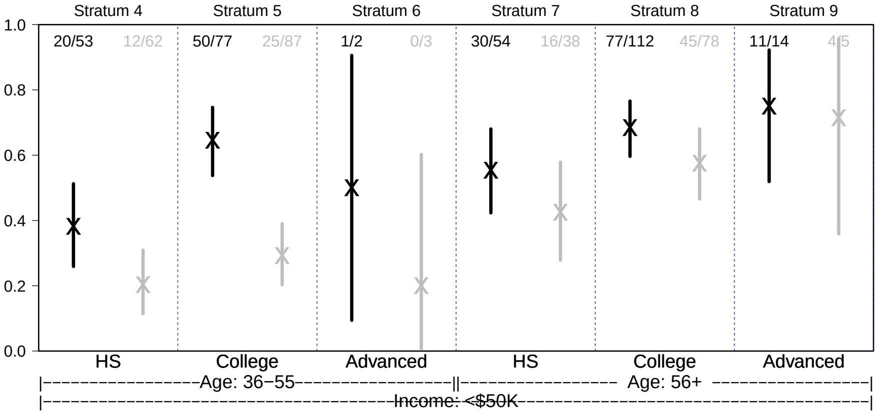
Example 2: 2014 Survey of Consumer Payment Choice (SCPC)

- $\sim$1,800 ALP panelists
- $\sim$1,300 Understanding America Study (UAS) panelists
- Surveys in each were coded and administered using different software.

- A common approach is to post-stratify by demographics:

$$\hat{p} = \sum_s f_s \hat{p}_s,$$

- f_s = proportion of population in stratum s.
- $\hat{p}_s$ = sample-based estimate of the proportion in stratum s.



We focus on stratum $s = 5$. Different results could be due to unaccounted-for demographic variables:

	Sample 1	Sample 2
% Male (M=1)	35.1	36.8
% White (W=1)	48.1	54.5
% Employed (J=1)	88.3	65.9

- We fit a logistic-regression model to each sample:

$$P(\text{credit card adopter}) \quad = \quad logit^{-1}(M \times \ W \times \ J)$$

- We use results from sample 1 to predict sample 2, and results of sample 2 to predict sample 1:

	Predicted	Observed
Sample 1	0.68	0.65
Sample 2	0.28	0.29

Consider stratum s, with data from two samples:

Sample 1 (n=77)	X_1 1	X_2 0	... ...	... ...	X_n 1	$X = \sum_{i=1}^{n} X_i$ 50
Sample 2 (m=87)	Y_1 1	Y_2 1	... ...	Y_m 0		$Y = \sum_{i=1}^{n} Y_i$ 25

- Estimates based on each sample alone are

$$\hat{p}_s(x) = \frac{X}{n} = 0.65 \quad \text{and} \quad \hat{p}_s(y) = \frac{Y}{m} = 0.29.$$

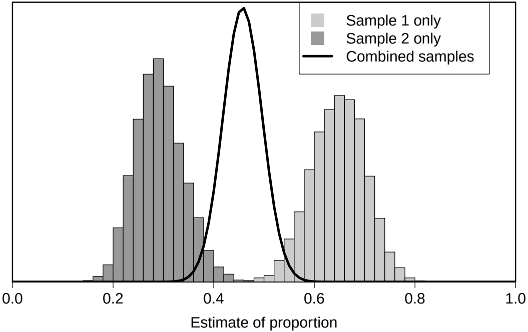
- Simply combining the data yields the estimate:

$$\hat{p}_s = \frac{X+Y}{n+m} = 0.46$$

- Uncertainty can be defined by distributions:

$$\mathrm{Beta}(X+Y, n+m\text{-}X\text{-}Y) \quad \text{similar to} \quad \mathrm{Normal}\left(\hat{p}_s, \frac{\hat{p}_s(1-\hat{p}_s)}{n+m}\right).$$

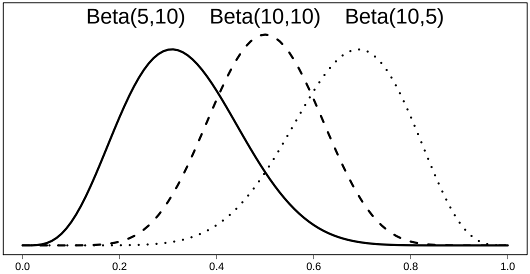
Looking at the uncertainty intervals:



- Not an intuitive result.
- The error corresponds to frequentist assessment of distributions for sample means for samples collected using same methodology.
- We really want to assess uncertainty of true stratum proportion, p_s, given our data.

We consider the following multi-level model:

- True stratum proportion: p_s.
- Every unique data collection methodology, c, produces sample with expected proportion $p_s(c) \sim \mathrm{Beta}(\alpha, \beta)$.
- Then, $\mathrm{E}\left[p_s(c)\right] = \frac{\alpha}{\alpha+\beta}$.

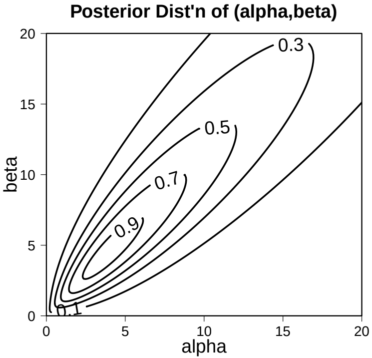


- Response in each sample are Bernoulli with probability $p_s(c)$:

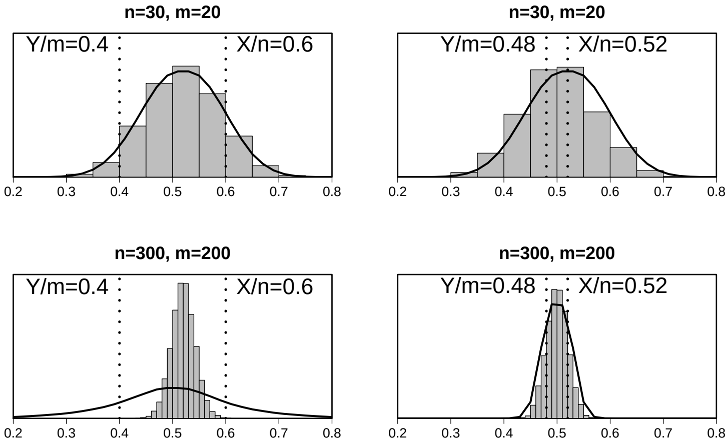
$$X \sim \mathrm{Binomial}(n, p_s(x)) \quad \text{and} \quad Y \sim \mathrm{Binomial}(m, p_s(y)).$$

We want to estimate $\mathrm{P}(p_s \mid X, Y, n, m)$:

- For given α, β, $\hat{p}_s = \frac{\alpha}{\alpha+\beta}$
- Uncertainty about α, β corresponds to uncertainty about p_s.
- Flat priors on α, β (slight shrinkage of p_s toward 0.5).
- Posterior of (α, β) in our stratum example:

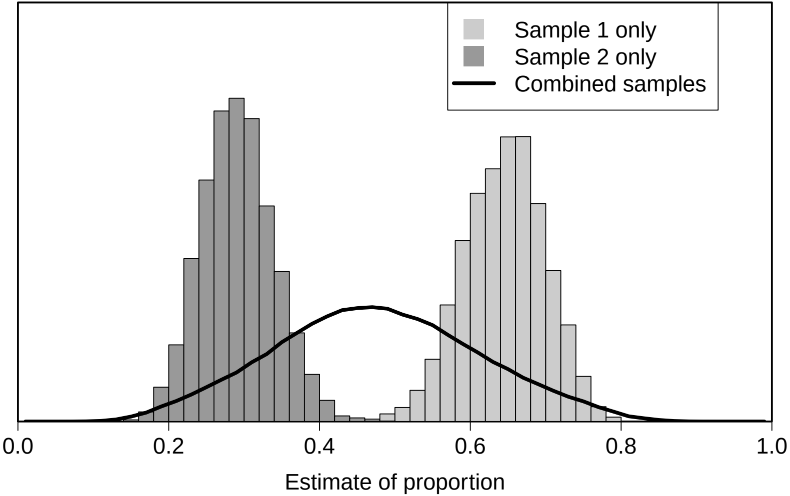


- Can be tricky to sample when posterior of (α, β) is diffuse.



- If two samples are consistent with one another, posterior distribution of p_s resembles frequentist combining.
- As samples get less consistent wit one another, posterior distribution of p_s diffuses.

Using this approach for our example stratum:

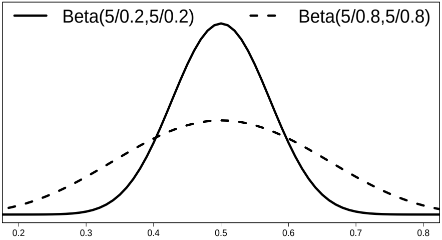


Seems a more reasonable assessment of uncertainty.

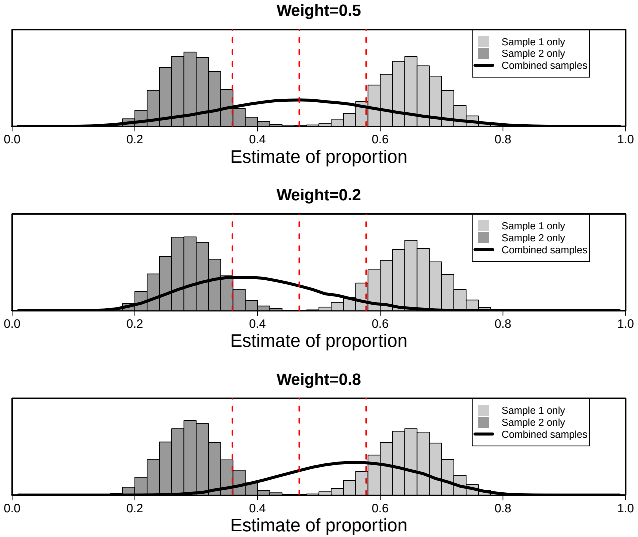
- $p_s(x), p_s(y) \sim \mathrm{Beta}(\alpha, \beta)$ assumes exchangeability of our samples; either is equally likely to have sample proportion closer to true stratum mean.
- What if we have additional information that tells us that one sample is more likely to be a better representation of the stratum?
- Consider $w \in [0, 1]$, and

$$p_s(x) \sim \mathrm{Beta}\left(\frac{\alpha}{w}, \frac{\beta}{w}\right) \quad \text{and} \quad p_s(y) \sim \mathrm{Beta}\left(\frac{\alpha}{1-w}, \frac{\beta}{1-w}\right).$$

- Weight w keeps the same mean, but changes the variance around p_s for the two samples.



For given w, we run our algorithm to to estimate p_s. A few examples; vertical lines at $w\hat{p}_s(x) + (1-w)\hat{p}_s(y)$.



- Information about relative quality of two samples can be incorporated into model to improve inference: smaller mean-squared errors, shorter uncertainty intervals.
- w represents how much more likely we believe the methodology in sample 1 to generate estimates closer to the true mean than the methodology in sample 2.
- $w \neq 0.5$ pushes posterior estimates of p_s closer to observed proportions in favored sample.

Future Work:

- How well do we need to accurately choose w to make sizeable gains in inference?
- How do we determine w, or distribution of w? Ask questions with known distributions under desirable sampling scheme?
  Example: Distribution of whites in sample 1(68/77) is different than in sample 2(58/88). Which is closer to the truth? Perhaps:

$$\frac{w}{1-w} = \frac{\mathrm{P}(\text{observing } 68/77 \text{ whites in stratum under SRS})}{\mathrm{P}(\text{observing } 58/88 \text{ whites in stratum under SRS})}?$$

- Ideas? Suggestions?