Improving Robustness Of Estimates From Non-Probability Online Samples

Dina Neiger, Andrew C. Ward and Darren W. Pennay
Workshop on Robust Inference for Sample Surveys,
University of Wollongong, 7 July 2017
Motivation

- **Weighting adjustments for non-probability samples:**
  - Usual approach: standard demographic post-stratification (with or without RIM)
    - Relatively little benefit for samples with demographic quotas
    - Can increase bias
  - Blending
    - Very effective but expensive and not always practical
  - **Model-based design weights**
    - Worth the effort?

- **Calibration**
  - What variables work best?
  - Independent benchmarks not always available
  - Difference between panels
  - Value proposition?
The Online Panels Benchmarking Study

3 x surveys
➢ probability samples
➢ Australian pop aged 18+

5 x surveys
➢ Members non-probability online panels
➢ Aged 18+
➢ Oct-Dec 15

Questionnaire included range of demographic questions and questions about health, wellbeing and use of technology

Items were chosen because there were high quality (e.g. ABS) population benchmarks available for these measures

Standardised questions were used to mitigate mode effects

Data and documentation available from the Australian Data Archive
www.ada.edu.au/ada/01329
Life in Australia, Design Features

Recruitment Oct-Nov 2016
- Dual-frame RDD (60% mobile frame/40% landline frame)
- National probability proportional to size sample design
- Trialling different recruitment methods (combinations of incentives, materials, one and two-stage recruitment)

Response maximisation
- Non-contingent incentive
- Extended routine / Reminders

Panel member
Advance notification postcard
3-week enumeration/multiple reminders (email and text)

Panel dimensions
- N=3,000 approx.
- Aged 18 years and over
- Online and offline population
- English-speaking

Panel maintenance
- Sample top up via a single – frame mobile phone survey
Weighting (standard approach)

Design weight (Probability surveys only):
- Adjusting for overlapping chance of selection (single-frame approach), number of landlines, number of in-scope persons in household
- LinA includes propensity weight based on response probabilities
- Design weight of ‘1’ for the non-probability panels

Post stratification (Raking/RIM):
- Gender
- Education by age (18-24 years, 25-34 years with/out university degree, 35-44 years with/out university degree, 45-54 years with/out university degree, 55-64 years with/out university degree, 65-74 years with/out university degree, 75+ years with/out university degree)
- Telephone status (Landline-only, Dual-users, Mobile-only)
- Volunteer (Yes/No)
- Future work to improve LinA estimates: include additional educational levels, State, investigate optimal post-stratification variables post Census 2016 data release
Comparison

Method


➢ Average absolute error to compare to benchmarks
  o Average of absolute differences (percentage points) across all measures between the benchmark and the survey estimate

Results

➢ Probability and non-probability surveys perform similarly well with respect to demographics

➢ for substantive characteristics the probability surveys are more accurate
## Results – substantive characteristics

<table>
<thead>
<tr>
<th>Substantive variables</th>
<th>Benchmark value (%)</th>
<th>Distance from benchmarks (percentage point difference from benchmark)</th>
<th>LinA</th>
<th>Dual Frame Prob.</th>
<th>Non-probability Panels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Life satisfaction (8 out of 10)</td>
<td>32.6</td>
<td></td>
<td>-1.4</td>
<td>2.2</td>
<td>-11.1</td>
</tr>
<tr>
<td>Psychological distress - Kessler 6 (Low)</td>
<td>82.2</td>
<td></td>
<td>-21.3</td>
<td>-9.8</td>
<td>-26.5</td>
</tr>
<tr>
<td>General Health Status - SF1 (Very good)</td>
<td>36.2</td>
<td></td>
<td>-3.8</td>
<td>-5.0</td>
<td>-4.2</td>
</tr>
<tr>
<td>Private Health Insurance</td>
<td>57.1</td>
<td></td>
<td>2.6</td>
<td>3.9</td>
<td>-8.0</td>
</tr>
<tr>
<td>Daily smoker</td>
<td>13.5</td>
<td></td>
<td>-1.0</td>
<td>2.1</td>
<td>9.0</td>
</tr>
<tr>
<td>Consumed alcohol in the last 12 months</td>
<td>81.9</td>
<td></td>
<td>2.7</td>
<td>3.0</td>
<td>-1.1</td>
</tr>
<tr>
<td>Enrolled to vote</td>
<td>78.5</td>
<td></td>
<td>9.0</td>
<td>8.3</td>
<td>8.4</td>
</tr>
</tbody>
</table>
Initial calibration

Method
- Blending probability and non-probability surveys
- Use measures of early adopter behaviours in calibration
- Experiment with different weighting variables

Results
- Reduction in bias is mostly due to blending but not always practical
- Use the best probability sample to combine with non probability regardless of mode and response rate
Extend calibration methodology to

**Model probability of appearing in non-probability sample**
- Valliant, R., Dever, J. A. (2011)
- Use propensity-response model to calculate probability-based design weight for non-probability sample

**Include other significant differentiators in the calibration**
  Scientific Surveys Based on Incomplete Sampling Frames and High Rates of Nonresponse. *Survey Practice*. 8 (5)
- Find demographic, behavioural and attitudinal measures that differentiate between probability and non-probability panels

**Compare effectiveness of methods using multiple samples**
  A model based approach for achieving a representative sample
- Average of the “average absolute errors” and “mean square errors” across multiple panels when comparing weighting methods
Implementation steps

1 Post-stratify reference sample

Adjust for coverage and non-response bias
- Raking ("rim weighting")
- Include Volunteerism
- Keep number of variables to a minimum
Implementation steps

1. Propensity score model

2. Post-stratify reference sample

Calculate probability based design weights for non-probability samples:
- Non-probability cases=1, reference cases=0
- Design weight - inverse of estimated probability of inclusion in non-probability sample given weighted reference sample
- Use as many covariates as available (exclude measures being evaluated)

Issues encountered:
- Extremely small probability for some cases => to extremely large weights (even when using propensity class mean)
- Solution: trim design weights
Implementation steps

<table>
<thead>
<tr>
<th>Adjustment Type</th>
<th>Average absolute error</th>
<th>% Change in bias compared with Design weights=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted</td>
<td>8.34</td>
<td>-</td>
</tr>
<tr>
<td>Standard post-stratification adjustments</td>
<td>8.68</td>
<td>4.1</td>
</tr>
<tr>
<td>Design weights=1</td>
<td>8.68</td>
<td>4.1</td>
</tr>
<tr>
<td>Design weights from propensity model – all covariates</td>
<td>8.46</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Small improvement but benefit of estimating probability design weights to enable inference using traditional probability based theory.
Implementation steps

3 Post-stratify with mode-based design weights

4 Find significant differentiators

Compare estimates post-stratified reference sample with the post-stratified non-probability sample and test for significant differences.

Significantly different for all 5 panels (compared to the reference sample)

✓ Early adopter variables
✓ Internet usage variables
✓ Income
✓ Employment
✓ Remoteness
  (was not a factor in RDD, LinA – larger sample)
✓ Home ownership
Implementation steps

Evaluate

Add key differentiators

Compare average absolute error and root mean square error across five non-probability panels for 7 outcome measures:

- General health status (ABS National Health Survey)
- Psychological distress (ABS National Health Survey)
- Life satisfaction (ABS General Social Survey)
- Private health insurance coverage (ABS National Health Survey)
- Daily smoking status (AIHW National Drug Strategy Household Survey)
- Alcohol consumption in the last 12 months (AIHW National Drug Strategy Household Survey)
- Enrolled to vote (Australian Electoral Commission)
Early Adopter Characteristics

DiSogra et al. (2011) “Early adopters (EA) – consumers who embrace new technology and products sooner than most others”

EA1: I usually try new products before other people do
EA2: I often try new brands because I like variety and get bored with the same old thing
EA3: When I shop I look for what is new
EA4: I like to be the first among my friends and family to try something new
EA5: I like to tell others about new brands or technology
Early adopters in calibration

Composite is better than individual variables

Impact of EA variables on bias (% change)
Internet usage measures

Look for information:
How often look for information over the internet

Access a home:
Type of internet connection

Post to blogs etc:
How often - Post to blog/forums/interest groups

Financial transactions:
How often - Conduct financial transactions such as banking over the Internet

Social media:
How often - Comment or post images to social media sites (Facebook, Twitter, etc.)

Frequency of use:
How often - Use the Internet at work, home or elsewhere
Internet usage measures

Not useful when calibrating online to online

Impact of Internet Usage measures on bias (% change)

- Look for information: 2.6
- Access at home, Look for information: 5, 0.9
- Look for information, blogs etc, Financial transactions, Social Media: 4.5, 0.3
- Post to blogs etc, Financial transactions, Social Media: 6.4
- Access at home, Frequency of use: 5.5, 1.3, 1.9

- Ave abs error compared to unweighted
- Ave RMSE compared to std adjustment
#### Other differentiators

**Income:** Annual income pre-tax:

<table>
<thead>
<tr>
<th>Income Range</th>
<th>Annual Income Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2,000 or more per week</td>
<td>($104,000 or more per year)</td>
</tr>
<tr>
<td>$1,500 - $1,999 per week</td>
<td>($78,000 - $103,999 per year)</td>
</tr>
<tr>
<td>$1,250 - $1,499 per week</td>
<td>($65,000 - $77,999 per year)</td>
</tr>
<tr>
<td>$1,000 - $1,249 per week</td>
<td>($52,000 - $64,999 per year)</td>
</tr>
<tr>
<td>$800 - $999 per week</td>
<td>($41,600 - $51,999 per year)</td>
</tr>
<tr>
<td>$600 - $799 per week</td>
<td>($31,200 - $41,599 per year)</td>
</tr>
<tr>
<td>$400 - $599 per week</td>
<td>($20,800 - $31,199 per year)</td>
</tr>
<tr>
<td>$300 - $399 per week</td>
<td>($15,600 - $20,799 per year)</td>
</tr>
<tr>
<td>$200 - $299 per week</td>
<td>($10,400 - $15,599 per year)</td>
</tr>
<tr>
<td>$1 - $199 per week</td>
<td>($1 - $10,399 per year)</td>
</tr>
<tr>
<td>Nil income/Negative income</td>
<td></td>
</tr>
</tbody>
</table>

**Employment:**

In the last week, any work done in a job, business or farm, without pay in a family business, including away from because of holidays, sickness or any other reason.

**Home ownership:**

- Owned outright
- Owned with a mortgage
- Being purchased under a rent/buy scheme
- Being rented
- Being occupied rent free
- Being occupied under a life tenure scheme
- Other

**Remoteness:**

- Major Cities
- Inner Regional
- Other (Outer Regional, Remote, Very Remote, Not answered)
Other Differentiators

Income – single most influential variable to reduce bias and RMSE
Benefit from including both: income and employment

Impact of Income, Employment and Home Ownership variables on bias (% change)

-7.3 -11.0
-8.5 -8.5
-10.3 -13.9
-10.6 -10.3
-7.3 -7.3
-5.8 -5.8
-5.9 -5.9
-9.6 -9.6

-16.0
-14.0
-12.0
-10.0
-8.0
-6.0
-4.0
-2.0
0.0

Ave abs error compared to unweighted
Ave abs error compared to std adjustment
Ave RMSE compared to std adjustment
Combine EA, Income and Employment

Std adjustments: Telephone status, Age by Education, Gender, Volunteer

Design weights model: Primary demographics, Secondary demographics, Telephone status, Volunteer, EA (individual), Internet use (all)

Key differentiations: Income, Employment, EA total agreement score

<table>
<thead>
<tr>
<th>Adjustment Type</th>
<th>Average absolute error</th>
<th>% Change in bias cf unweighted</th>
<th>% Change in bias cf weighted with std adjustments</th>
<th>Average RMSE</th>
<th>% Change in RMSE cf weighted with std adjustments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted</td>
<td>8.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted with std adjustments</td>
<td>8.68</td>
<td>4.10</td>
<td></td>
<td>9.10</td>
<td></td>
</tr>
<tr>
<td>Weighted with design weights and std adjustments</td>
<td>8.46</td>
<td>1.40</td>
<td>-2.60</td>
<td>8.96</td>
<td>-1.50</td>
</tr>
<tr>
<td>Weighted with design weights and key differentiators</td>
<td>7.47</td>
<td>-10.40</td>
<td>-13.90</td>
<td>8.31</td>
<td>-8.60</td>
</tr>
</tbody>
</table>
Variability among 5 non probability panels (LinA calibration)

<table>
<thead>
<tr>
<th></th>
<th>Average Absolute Error</th>
<th>% Change in bias</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unweighted</td>
<td>Std adjustments</td>
</tr>
<tr>
<td>Panel 1</td>
<td>9.2</td>
<td>9.8</td>
</tr>
<tr>
<td>Panel 2</td>
<td>10.0</td>
<td>10.4</td>
</tr>
<tr>
<td>Panel 3</td>
<td>7.7</td>
<td>7.7</td>
</tr>
<tr>
<td>Panel 4</td>
<td>7.4</td>
<td>8.1</td>
</tr>
<tr>
<td>Panel 5</td>
<td>7.4</td>
<td>7.4</td>
</tr>
<tr>
<td>All Panels</td>
<td>8.3</td>
<td>8.7</td>
</tr>
</tbody>
</table>

Benefit of adjustment is sample dependent!
Variability among 5 non probability panels (RDD calibration)

<table>
<thead>
<tr>
<th></th>
<th>Unweighted</th>
<th>Std adjustments</th>
<th>Design weights and key differentiators</th>
<th>From unweighted</th>
<th>From std adjustment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel 1</td>
<td>9.2</td>
<td>9.8</td>
<td>6.7</td>
<td>-27.3</td>
<td>-31.4</td>
</tr>
<tr>
<td>Panel 2</td>
<td>10.0</td>
<td>10.4</td>
<td>8.3</td>
<td>-17.2</td>
<td>-20.5</td>
</tr>
<tr>
<td>Panel 3</td>
<td>7.7</td>
<td>7.7</td>
<td>5.5</td>
<td>-29.0</td>
<td>-29.6</td>
</tr>
<tr>
<td>Panel 4</td>
<td>7.4</td>
<td>8.1</td>
<td>7.5</td>
<td>2.4</td>
<td>-6.7</td>
</tr>
<tr>
<td>Panel 5</td>
<td>7.4</td>
<td>7.4</td>
<td>7.1</td>
<td>-4.0</td>
<td>-3.8</td>
</tr>
<tr>
<td>All Panels</td>
<td>8.3</td>
<td>8.7</td>
<td>7.0</td>
<td>-15.8</td>
<td>-19.1</td>
</tr>
</tbody>
</table>

Benefit of adjustment is sample dependent!
Comparison with independent benchmarks vs reference sample benchmarks

Mostly change in bias is consistent but there can be large differences for some samples.
In conclusion…

Combination of

- A robust reference sample with enough variables in common
- Application of propensity scores to calculate design based probabilities
- Extending calibration variables to include key differentiators (beyond EA and internet use)
- Good population benchmarks
- A bit of time and statistical “know how”

Will improve quality of inference from a non-probability sample but not a magic bullet

- Adjustments need to be developed for each sample
- Plan ahead to include sufficient calibration variables in reference and non-probability sample
- Aim to include at least some independent benchmarks
Where to next...
Now that probability online panel is available

Repeat analysis with Life in Australia as the reference sample
- ~2,500 respondents in the first two waves
- No mode issues

Scope to
- Systematise and standardise approach
- Trial other key differentiators identified in the literature (e.g. media exposure, alternative income measures)
- Optimisation (Terhanian et al) for selection of best combination of calibration variables

Ideas and suggestions from the workshop
Thank you

Contact: Andrew Ward (andrew.ward@srcentre.com.au)