Sampling in the CLSA

Lauren Griffith
Harry Shannon
McMaster University

CLSA Methods Seminar
29 November 2011
Outline of presentation

• Background on sampling
• Participants in the CLSA
• Sampling approaches in the CLSA
• CCHS participants
• Sampling from provincial health registries
• Principles of Random Digit Dialing
• Issues with RDD
• Conclusion
Principles of sampling

• Population vs Sample
• Want representative sample of some target population
• Need every member of the population to have non-zero probability of being sampled
• Must be able to estimate the probability of sampling any individual chosen
Simple random sampling

- All units in target population are known
- Sample is chosen randomly
- Each unit has an equal probability of being chosen
Simple random sampling

- Unit sampled
- Unit not sampled
Stratified random sampling

- Population of interest is divided into strata (e.g., male and female; young, middle-aged, old)
- Simple random sample is chosen from each stratum
- Probabilities of selection between the strata can vary
Stratified random sampling

Stratum 1

Stratum 2

Unit sampled

Unit not sampled
More complex designs

- Stratification
- Clustering
- Multi-stage
- Combinations
Sampling in difficult situations

• E.g., disaster areas, war zones, Low Income Countries
• Various alternative methods
• E.g., EPI, Extended Program on Immunization
• Methods typically have some limitations
• May have to balance bias, precision, speed, cost

Back to CLSA ...
Aims of sampling in CLSA

• Choose representative sample of eligible Canadians
  – 20K Tracking cohort; 30K Comprehensive cohort
  – Specified numbers in age-sex groups by province
Potential Sampling Frames

- Canadian Community Health Survey Participants
- Provincial Health Registration Databases
- Random Digit Dialling

ALL OF THE ABOVE
Canadian Longitudinal Study on Aging

Sampling Frame:
- CCHS, provincial health registration databases, and RDD
- CLSA Tracking (n=20,000)
  - 45-54: 6,000
  - 55-64: 6,000
  - 65-74: 4,000
  - 75-85: 4,000

Sampling Frame:
- provincial health registration databases, and RDD
- CLSA Comprehensive (n=30,000)
  - 45-54: 9,000
  - 55-64: 9,000
  - 65-74: 6,000
  - 75-85: 6,000
• CCHS provided first part of sample
• Options for methods of selection of remaining participants:
  – Using provincial health registries - *preferred*
  – Random digit dialing
• In several provinces, we cannot use registries, so need to do RDD
Recruitment from the CCHS

• CLSA collaborated with Statistics Canada to develop the CCHS Healthy Aging Questionnaire
• **Target population:** People aged 45 and over living in private occupied dwellings in the ten provinces
• **Excluded:**
  • Residents of the three territories
  • Persons living on Indian reserves or Crown lands
  • Persons living in institutions
  • Full-time members of the Canadian Forces
  • Residents of some remote regions
Recruitment from the CCHS, *ctd.*

Multi-stage sampling

- Sampling frame 2006 Census
- Selection
  - Clusters based on Census dissemination area blocks
  - Dwellings within cluster
  - Person within dwelling
- Response Rate
  - Household-level 80.8%
  - Person-level 92.1%
  - Overall 74.4%
Recruitment from the CCHS, *ctd.*

Participants were asked to share:

- Their contact information with the CLSA (for recruitment)
- Their survey responses with the CLSA (for analysis)

12,269 (47%) of Eligible Participants provided Contact Information to CLSA

- Contact + Survey
- Contact Only
- Survey Only
- Neither

N=30,865

45-85

N=26,248

N=11,742

Contact + Survey

N=8,345

Survey Only

N=5,634

Neither

N=527

Contact Only

N=4,617

>85
Recruitment from the CCHS, *ctd.*

**Canadian Longitudinal Study on Aging**

**Sampling Frame:**
- CCHS, provincial health registration databases, and RDD
- CLSA Tracking
  - (n=20,000)

**Sampling Frame:**
- provincial health registration databases, and RDD
- CLSA Comprehensive
  - (n=30,000)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>CCHS</th>
<th>Remainder</th>
</tr>
</thead>
<tbody>
<tr>
<td>45-54</td>
<td>617</td>
<td>5,383</td>
</tr>
<tr>
<td>55-64</td>
<td>1,704</td>
<td>4,296</td>
</tr>
<tr>
<td>65-74</td>
<td>1,350</td>
<td>2,650</td>
</tr>
<tr>
<td>75-85</td>
<td>791</td>
<td>3,209</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age Group</th>
<th>CLSA Tracking</th>
<th>CLSA Comprehensive</th>
</tr>
</thead>
<tbody>
<tr>
<td>45-54</td>
<td>9,000</td>
<td>9,000</td>
</tr>
<tr>
<td>55-64</td>
<td>9,000</td>
<td>6,000</td>
</tr>
<tr>
<td>65-74</td>
<td>6,000</td>
<td>6,000</td>
</tr>
<tr>
<td>75-85</td>
<td>791</td>
<td>3,209</td>
</tr>
</tbody>
</table>

Canadian Longitudinal Study on Aging Sampling Frame:
- CCHS, provincial health registration databases, and RDD
- CLSA Tracking
  - (n=20,000)

CLSA Comprehensive
- (n=30,000)
Recruitment from Provincial health registration databases

• 2005
  – Feasibility study to explore practical, methodological and ethical aspects of accessing Health Care Utilization data from Provincial databases (published 2009)

• 2009-2011
  – Several meetings with Provincial Data Stewards and Privacy Commissioners to negotiate access to health registration databases for sampling
MOH Mailout
- Letter(s) from MOH and/or CLSA
- Brochure, Information Package
- Consent to contact form
  + Postcard reminder in 20 days

Participant returns consent to contact form

NCC
- Assigns unique ID
- Sends:
  - Participant consent form
  - Additional study materials

Participant contacted

CATI
- Assess eligibility
- Answers participant questions

Participant interested and ready
- Participant consent
- Conducts telephone interview
- Collects Provincial Health Number (if participant provides consent)

Participant interested but not ready
- Schedule Interview
MOH Mailout
• Letter(s) from MOH and/or CLSA
• Brochure, IP
• Consent to contact form
  + Postcard reminder in 20 days

Participant returns consent to contact form

NCC
• Assigns unique ID
• Sends additional study materials if requested

Participant contacted

DCS - CATI
• Verifies eligibility
• Answers participant questions
• Schedules In-home interview

Participant eligible and interested

In-Home
• Recruits participant
• Participant consent obtained
• Conducts in-home interview
• Collects Provincial Health Number (if participant provides consent)
• Schedules DCS visit

DCS
• Participant receives printed copy of consent
• Physical assessment
• Neuropsychological battery
• Blood and urine (if participant provides consent)
RDD – Tracking + Telephone Administered Questionnaires Pilot

42,860 calls → 300 pilot participants

<table>
<thead>
<tr>
<th></th>
<th>Mean Age (SD)</th>
<th>Language</th>
<th>Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Injury Module (n=200)</td>
<td>70.5 y (11.2 y)</td>
<td>Fr=100</td>
<td>F=92</td>
</tr>
<tr>
<td></td>
<td></td>
<td>En=100</td>
<td>M=108</td>
</tr>
<tr>
<td>Tracking Baseline (n=50)</td>
<td>64.3 y (10.6 y)</td>
<td>Fr=23</td>
<td>F=33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>En=27</td>
<td>M=17</td>
</tr>
<tr>
<td>Maintaining Contact - Comp (n=25)</td>
<td>61.3 y (9.0 y)</td>
<td>Fr=12</td>
<td>F=12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>En=13</td>
<td>M=13</td>
</tr>
<tr>
<td>Maintaining Contact - Tracking (n=25)</td>
<td>63.1 y (10.0 y)</td>
<td>Fr=15</td>
<td>F=13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>En=10</td>
<td>M=12</td>
</tr>
<tr>
<td>TOTAL (n=300)</td>
<td>62.7 y (10.8 y)</td>
<td>Fr=150</td>
<td>F=150</td>
</tr>
<tr>
<td></td>
<td></td>
<td>En=150</td>
<td>M=150</td>
</tr>
</tbody>
</table>
RDD – Comprehensive Pilot

Total number of calls made (RDD) n=12,267

Recruitment of participants who agreed to be contacted by CLSA n=131

Hamilton: n=66
Mean Age = 65.0 y (11.0 y)
Language = En
Female = 33; male = 32
45 Completed Interviews

Montreal: n=65
Mean Age = 65.0 y (12.0 y)
Language = Fr
Female = 32; male = 33
41 Completed Interviews
Plan for Additional Recruitment

P=Provincial Health Registration Databases
R=RDD Only
H=Hybrid: RDD then Provincial Health Registration Databases
# Example of requirement by province

## Tracking cohort

<table>
<thead>
<tr>
<th></th>
<th>45-54</th>
<th>55-64</th>
<th>65-74</th>
<th>75-85</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td># Required</td>
<td>306</td>
<td>306</td>
<td>306</td>
<td>306</td>
<td>2,040</td>
</tr>
<tr>
<td># Providing Contact Info</td>
<td>121</td>
<td>128</td>
<td>153</td>
<td>193</td>
<td>1,022</td>
</tr>
<tr>
<td># Anticipated through CCHS</td>
<td>28</td>
<td>35</td>
<td>56</td>
<td>82</td>
<td>376</td>
</tr>
<tr>
<td># Additional Participants</td>
<td>278</td>
<td>271</td>
<td>250</td>
<td>224</td>
<td>1,664</td>
</tr>
<tr>
<td># Phone numbers to Sample*</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

* This will depend on the recruitment rate per number sampled
RDD approach

• In principle, idea is simple
• Randomly sample numbers as far as possible in specified area codes and with next 3 digits in relevant area
• Identify eligible people at each number
• Randomly choose one person
• Recruit willing participants until ‘quota’ filled
Issues in using RDD

• Identifying numbers in specified area
• Having up-to-date list of numbers for target population
• Ability to compute sample weights
• Presence of landlines and/or cellphones
• Eligibility within household – changes over time
• Method of initial contact
• Households without phones
• Numbers may be businesses, out of order, etc.
• People away from home (snowbirds, etc.)
Cell phones and landlines

• Statistics Canada survey December 2010
• Supplement to Labour Force Survey
• Households using cell phones exclusively:
  – Overall: 13%
  – Age 18-34 50%
  – Over 35 8%
  – Over 55 4%
• Increasing over time
• Landlines reach nearly all our eligibles
Combining samples from cell phones and landlines

• Methods have been described
• Need to determine all phones in each household
• Keep logs of unfilled quotas (age-sex numbers)
• Interviewers construct rosters of eligibles within households and randomly choose one
Some issues with cell phones

• Ethical: incoming calls may cost user; privacy; activity when answering (driving, etc); children
• Cost: AAPOR states at least 2x, maybe 3-4x cost of landline survey
• Getting addresses
• Quality of data (may be similar to landlines)

Source: AAPOR
‘Cold calling’ vs prior contact/letters

• Time and expense of mailing letters (only possible when we have name and address)
• May increase willingness to talk to interviewers (call display)
• However, many households will not include any eligible people
Contacting subjects

• On average, anticipate making many calls to recruit a single person
  – Up to 7-10 calls to obtain response
  – Leave message?
  – Willingness to participate

• Working on assumption of 20% ‘recruitment rate’ for health registry data (15% in 75-85 age group)

• Exclude households without a phone
Estimation of sampling weights

- Calculate probability of selecting sampling unit (in CLSA, unit = person)
- Account for different sampling bases
- Allow for non-response
- Weight = $1 / P(\text{selection})$
- Use to ‘weight up’ the sample to get estimates of parameters (means, proportions, etc) for the target population
- Various assumptions required
Sources of the CLSA sample

• Tracking cohort:
  – CCHS
  – Health registries
  – RDD

• Comprehensive cohort
  – Health registries
  – RDD
Probabilities for the CCHS

- Provided by StatsCan
- Must allow for non-response in the CLSA
- Some issues on confidentiality – information sharing
Probabilities for the health registries

- Health registries (HR) have list of (virtually) all target population
- HR can provide numbers of people in each age-sex group for the province (denominators)
- Mail-outs from HRs will lead to estimation of proportion of ineligibles and adjustment of denominators
- Estimate probability of participation
Probabilities for RDD

- Phone numbers in range (population) identified
- For tracking, all numbers in province
- For comprehensive, some eligibility established during contact call
- Eligibility: private residence, geography, age, competent to interview, quota not filled, other
- Probability of selection is product of various probabilities
N: Number of TNs in sampling frame

N1: # of TNs called to achieve quota

N2: # of TNs not out of order

N3: # of TNs answered

N4: # of those TNs we find out if eligible as residence

N5: # of those TNs that are residences

N6: # of those we find out if geographically eligible

N7: # of those that are geographically eligible

N8: # of those for which we identify the eligible people in household

N9: # of those with an eligible person in household (from stratum)
N10: # of those with person from stratum chosen for the sample

N11a: # of those where chosen person is already on phone

N11b: # of those where chosen person is not already on phone

N12: # of those where person on phone asks chosen person to come to phone

N13: # of those where chosen person comes to phone

N14: # where chosen person agrees to participate
Some probabilities estimated

\[ P_{noo} = \frac{TNs \text{ not out of order}}{TNs \text{ called to achieve quota}} \]

\[ P_{res} = \frac{TNs \text{ that are residences}}{TNs \text{ we find out if eligible as residence}} \]

\[ P_{part} = \frac{\text{number agreeing to participate}}{\text{number selected to participate}} \]
Combining samples from different sources

- Want overall $P(\text{Participation})$
- Use addition rule of probability
- E.g., for someone chosen via RDD, need $P(\text{Selected by RDD}) \text{ AND } P(\text{Selected in CCHS})$
- Latter is an average probability, not an individual one
- Similarly for selection through HR
Additional issues

• When $P(\text{Participation})$ is based on the product of probabilities, have to assume independence of probabilities

• Confidentiality conditions may mean, e.g., we call people in RDD who were in the CCHS and did not want to participate in the CLSA

• In RDD, have to allow for multiple phones in the household

• At some point, likely to fill some age/sex quotas; then only recruit unfilled quotas
Summary

• Various sources of participants for CLSA
• Each has its own strengths and limitations
• Need to estimate sampling probabilities for each source
• Aiming for representativeness – but ...
• Various assumptions must be made