Sampling in developing countries: Five challenges from the field

Macartan Humphreys, Associate Professor, Department of Political Science, Columbia University
Outline

1. Sampling for balance: Approach to sub-district selection used in Aceh
2. Sampling without a frame: Approach to XC sampling used in Aceh
3. Sampling and Control: Ups and downs of tying enumerators hands on subject characteristics
4. The problem with clusters
5. Other things:
   1. Sampling and Replacement: To stay local or start again
   2. Sampling and Manipulation: Who should make the village lists --- us or them?
Key challenges

• Poor frames
• Difficulty of relying on enumerators to implement complex designs
• Difficulty of doing power analysis when distribution of outcomes not well known
• Various logistic concerns
• Spillover concerns of various forms
1 Indonesia: Sampling Controls

The lack of randomization makes identification hard.

We are interested in sampling to estimate causal effects

1. **Identifying quasi-controls**

   Used the known assignment rule and a simulation to generate a control group most similar to the treated group.
The Assignment Rule

**ASSIGNMENT VARIABLES:** Conflict intensity and Spending capacity

**RULE**
- Set a target number of subdistricts to be treated in each district equal the number of high conflict-affected subdistricts in the district. If there are no high conflict affected, set the target equal the number of medium conflict affected districts; if there are no high or medium, let the target equal one.
- Select the most conflict affected subdistricts in each district up to the target in each district, conditional on a subdistrict meeting the 60 percent spending criterion.
- If insufficient subdistricts meet the spending criterion, select subdistricts with the best spending performance.

This rule
- (a) is complex
- (b) is not perfectly replicable, and
- (c) doesn’t tell us which untreated subdistricts where ‘closest’ to being treated for a control group
- (d) assignment of one unit depends on characteristics of others
1 Identifying quasi-controls

THE ASSIGNMENT RULE

(a) is complex

(b) is not perfectly replicable, and

(c) doesn’t tell us which untreated subdistricts were ‘closest’ to being treated for a control group.

(d) assignment of one unit depends on characteristics of others
1 Identifying quasi-controls
1 Identifying quasi-controls
1 Identifying quasi-controls

This unit is actually “close” to being treated

This unit is actually “close” to not being treated
1 Identifying quasi-controls

OUR APPROACH

1) Use simulation to administer small shock to all units on both assignment variables
2) Apply the rule to “perturbed” data to select treatment subdistricts
3) Repeat (10,000 times)
4) Calculate a “propensity” to assignment when uncertainty derives from (simulated) noise in data
5) Take as the control group the 67 subdistricts that had the highest propensity. (also possible: match on propensities)
6) This gives a guide to sampling
7) It also gives a guide to estimation, insofar as we believe propensities we can do IPW, we can also see how robust results are to variation in propensities
1 Matched Controls
1 Final Locations
2 Sampling without a frame

• We wanted to sample ex-combatants in Aceh.
  – But we did not have a list of ex-combatants
  – And we did not know where they were concentrated
  – They are fairly “rare” so we cannot afford to do triage
2 Sampling without a frame

- **Strategy:**
  - Randomly sample share $q_1$ of villages.
  - Generate a complete list of population of interest
  - Sample share $q_2$ of subjects from this list.
  - So probability that an individual is sampled is $q_1 q_2$ which is independent of village size and of number of combatants in village; so self weighting
2 Sampling without a frame

• Worries:
  – Ex post worry: Lists may not be complete. We have no good solution to this.
  – Ex ante worries:
    • There may be too much clustering and not enough power.
    • There is uncertainty about how many people we will actually survey!
2 Sampling without a frame

• We considered two possibilities:
• 1. Fixed share of population, $pn$
• 2. Some concave function of population e.g. $pn^{0.5}$

• The advantage of 1 is that the survey would be self-weighting; the advantage of 2 is that it reduces variance (we won’t have small numbers of places with huge samples)
2 Sampling without a frame

- Empirical distribution of #GAM at the subdistrict level
2 Sampling without a frame

Negative Binomial Distributions for different size parameters
2 Sampling without a frame

Distribution of Sample Size for Each Population Distribution Type
(Linear Approach: Taking .5n in each village)
2 Sampling without a frame

Distribution of Sample Size for Each Population Distribution Type
(Concave Approach; Taking $n^{0.5}$ in each village)
3 The problem with clustering

• It too us a long time to realise that cluster sampling is not good if you want to construct measure like ethnic fragmentation.

• Individual level measures may be unbiased; but social measures are not!!

• Conversely, systematic random sampling may not be good if you are interest in networks
4 Within household sampling

• Statistics versus management?
• Costs of stratification?

• We use dictionaries to remove discretion from sampling and to maximize balance
• Dictionaries specify: ID, date, enumerator, household, gender of respondent, any variations in treatment. These are hardwired for balance.
4 Within household sampling

Dictionaries

<table>
<thead>
<tr>
<th>ID</th>
<th>V_ID</th>
<th>LLU_NAME</th>
<th>D</th>
<th>CHEF</th>
<th>Equipe</th>
<th>Etape A</th>
<th>GENDER</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
<th>S10</th>
<th>S11</th>
<th>S12</th>
</tr>
</thead>
<tbody>
<tr>
<td>130610</td>
<td>13061</td>
<td>KAPULULA</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>11-Oct-10</td>
<td>H</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>130611</td>
<td>13061</td>
<td>KAPULULA</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>11-Oct-10</td>
<td>F</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>130612</td>
<td>13061</td>
<td>KAPULULA</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>11-Oct-10</td>
<td>H</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>130613</td>
<td>13061</td>
<td>KAPULULA</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>11-Oct-10</td>
<td>F</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>130614</td>
<td>13061</td>
<td>KAPULULA</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>11-Oct-10</td>
<td>H</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>130510</td>
<td>13051</td>
<td>KIMONO</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>14-Oct-10</td>
<td>F</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>130511</td>
<td>13051</td>
<td>KIMONO</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>14-Oct-10</td>
<td>H</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>130512</td>
<td>13051</td>
<td>KIMONO</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>14-Oct-10</td>
<td>F</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>130513</td>
<td>13051</td>
<td>KIMONO</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>14-Oct-10</td>
<td>H</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>10</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>130514</td>
<td>13051</td>
<td>KIMONO</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>14-Oct-10</td>
<td>F</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>11</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>130110</td>
<td>13011</td>
<td>KAFUBU</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>18-Oct-10</td>
<td>F</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>7</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>130111</td>
<td>13011</td>
<td>KAFUBU</td>
<td>HK</td>
<td>KAPONDA</td>
<td>1</td>
<td>18-Oct-10</td>
<td>H</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>V_ID</th>
<th>Equipe</th>
<th>LLU_NAME</th>
<th>D</th>
<th>CHEF</th>
<th>EtapeA</th>
<th>EtapeB</th>
<th>EtapeC</th>
<th>EtapeD</th>
<th>EXTERNE</th>
<th>INTERNE</th>
<th>MALPUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>13061</td>
<td>1</td>
<td>KAPULULA</td>
<td>HK</td>
<td>KAPONDA</td>
<td>11-Oct-10</td>
<td>18-Oct-10</td>
<td>21-Oct-10</td>
<td>6-Dec-10</td>
<td>EXTERNE</td>
<td>INTERNE</td>
<td>DIA PUB</td>
</tr>
<tr>
<td>13051</td>
<td>1</td>
<td>KIMONO</td>
<td>HK</td>
<td>KAPONDA</td>
<td>14-Oct-10</td>
<td>21-Oct-10</td>
<td>25-Oct-10</td>
<td>9-Dec-10</td>
<td>PAS EXTERNE</td>
<td>INTERNE</td>
<td>DIA PUB</td>
</tr>
<tr>
<td>13011</td>
<td>1</td>
<td>KAFUBU</td>
<td>HK</td>
<td>KAPONDA</td>
<td>18-Oct-10</td>
<td>25-Oct-10</td>
<td>28-Oct-10</td>
<td>13-Dec-10</td>
<td>EXTERNE</td>
<td>INTERNE</td>
<td>DIA PRIV</td>
</tr>
<tr>
<td>13041</td>
<td>1</td>
<td>KASANGULA</td>
<td>HK</td>
<td>KAPONDA</td>
<td>21-Oct-10</td>
<td>28-Oct-10</td>
<td>1-Nov-10</td>
<td>16-Dec-10</td>
<td>EXTERNE</td>
<td>PAS INTERNE</td>
<td>DIA PUB</td>
</tr>
<tr>
<td>13071</td>
<td>1</td>
<td>KINANDU</td>
<td>HK</td>
<td>KAPONDA</td>
<td>25-Oct-10</td>
<td>1-Nov-10</td>
<td>4-Nov-10</td>
<td>20-Dec-10</td>
<td>EXTERNE</td>
<td>INTERNE</td>
<td>DIA PUB</td>
</tr>
<tr>
<td>11371</td>
<td>1</td>
<td>LUFIRA 1</td>
<td>HK</td>
<td>LUFIRA</td>
<td>28-Oct-10</td>
<td>4-Nov-10</td>
<td>8-Nov-10</td>
<td>3-Jan-11</td>
<td>EXTERNE</td>
<td>PAS INTERNE</td>
<td>DIA PUB</td>
</tr>
<tr>
<td>11161</td>
<td>1</td>
<td>DITENGA Q1</td>
<td>HK</td>
<td>BASANGA</td>
<td>1-Nov-10</td>
<td>8-Nov-10</td>
<td>11-Nov-10</td>
<td>6-Jan-11</td>
<td>EXTERNE</td>
<td>PAS INTERNE</td>
<td>DIA PUB</td>
</tr>
<tr>
<td>11221</td>
<td>1</td>
<td>ILUNGA ET KYANGWE</td>
<td>HK</td>
<td>LUFIRA</td>
<td>4-Nov-10</td>
<td>11-Nov-10</td>
<td>15-Nov-10</td>
<td>10-Jan-11</td>
<td>PAS EXTERNE</td>
<td>PAS INTERNE</td>
<td>DIA PRIV</td>
</tr>
<tr>
<td>11391</td>
<td>1</td>
<td>MULUBWA</td>
<td>HK</td>
<td>LUFIRA</td>
<td>8-Nov-10</td>
<td>15-Nov-10</td>
<td>18-Nov-10</td>
<td>13-Jan-11</td>
<td>EXTERNE</td>
<td>INTERNE</td>
<td>DIA PRIV</td>
</tr>
<tr>
<td>11301</td>
<td>1</td>
<td>QUARTIER 8</td>
<td>HK</td>
<td>LUFIRA</td>
<td>11-Nov-10</td>
<td>18-Nov-10</td>
<td>22-Nov-10</td>
<td>17-Jan-11</td>
<td>EXTERNE</td>
<td>INTERNE</td>
<td>DIA PRIV</td>
</tr>
<tr>
<td>11461</td>
<td>1</td>
<td>KABUNGWE1</td>
<td>HK</td>
<td>LUFIRA</td>
<td>15-Nov-10</td>
<td>22-Nov-10</td>
<td>25-Nov-10</td>
<td>20-Jan-11</td>
<td>PAS EXTERNE</td>
<td>PAS INTERNE</td>
<td>DIA PRIV</td>
</tr>
<tr>
<td>11271</td>
<td>1</td>
<td>QUARTIER 1</td>
<td>HK</td>
<td>LUFIRA</td>
<td>18-Nov-10</td>
<td>25-Nov-10</td>
<td>29-Nov-10</td>
<td>24-Jan-11</td>
<td>PAS EXTERNE</td>
<td>PAS INTERNE</td>
<td>DIA PRIV</td>
</tr>
<tr>
<td>11331</td>
<td>1</td>
<td>QUARTIER 1ER</td>
<td>HK</td>
<td>LUFIRA</td>
<td>22-Nov-10</td>
<td>29-Nov-10</td>
<td>2-Dec-10</td>
<td>27-Jan-11</td>
<td>EXTERNE</td>
<td>PAS INTERNE</td>
<td>DIA PUB</td>
</tr>
</tbody>
</table>
4 Within household sampling

• But at some **cost** which we are not yet too clear about:
• Eg for gender: we hardwire whether in household X we will survey a man or a woman.
• Without this there are real risks
• What oddities does this introduce?
  – For a given household size, men are less likely to be sampled if the live in households with more men than women – **account for this with sampling weights**.
  – Whole households are less likely to be surveyed (and drop out) if they are mixed gender – **how to account for this**?
5 Questions we are wondering about:

• Local replacement or back to lists
  – We rely on local between households but household replacement in case of individual refusals

• Sampling and Manipulation: Who should make the village lists --- us or them?
  – Approach: relying on them but neighbor checking (imperfect, but...)
Wish list

• Would like some general code for making dictionaries
• Would like more general sampling code
• Would like more power analysis suites