This report is part of the RAND Corporation research report series. RAND reports present research findings and objective analysis that address the challenges facing the public and private sectors. All RAND reports undergo rigorous peer review to ensure high standards for research quality and objectivity.
Capacity Building at the Kurdistan Region Statistics Office Through Data Collection

Appendixes

Sponsored by the Kurdistan Regional Government

Shmuel Abramzon
Nicholas Burger
Bonnie Ghosh-Dastidar
Peter Glick

Krishna B. Kumar
Francisco Perez-Arce
Alexandria C. Smith
This research was undertaken within RAND Labor and Population. RAND Labor and Population has built an international reputation for conducting objective, high-quality, empirical research to support and improve policies and organizations around the world. Its work focuses on labor markets, social welfare policy, demographic behavior, immigration, international development, and issues related to aging and retirement with a common aim of understanding how policy and social and economic forces affect individual decision-making and the well-being of children, adults, and families.

*The authors of this report are listed in alphabetical order.*
## Contents

### APPENDIXES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>C.</td>
<td>Kurdistan Region Labor Force Survey Questionnaire</td>
<td>1</td>
</tr>
<tr>
<td>D.</td>
<td>Public Reporting Template</td>
<td>25</td>
</tr>
<tr>
<td>E.</td>
<td>Sampling Workshop</td>
<td>28</td>
</tr>
<tr>
<td>F.</td>
<td>Questionnaire Design Workshop</td>
<td>96</td>
</tr>
<tr>
<td>G.</td>
<td>Data Collection and Entry Workshop</td>
<td>148</td>
</tr>
<tr>
<td>H.</td>
<td>Data Cleaning Workshop</td>
<td>197</td>
</tr>
<tr>
<td>I.</td>
<td>Labor Market Concepts Workshop</td>
<td>239</td>
</tr>
<tr>
<td>J.</td>
<td>Statistical Analysis Workshop</td>
<td>260</td>
</tr>
</tbody>
</table>
This appendix reproduces the questionnaire created for the survey described in the companion report.
A. Introductory Data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Governorate:</td>
</tr>
<tr>
<td>A2</td>
<td>District:</td>
</tr>
<tr>
<td>A3</td>
<td>Sub district:</td>
</tr>
<tr>
<td>A4</td>
<td>Environment: 1. Urban 2. Rural</td>
</tr>
<tr>
<td>A5</td>
<td>Quarter’s name and number:</td>
</tr>
<tr>
<td>A6</td>
<td>Section’s name and number:</td>
</tr>
<tr>
<td>A7</td>
<td>Block number:</td>
</tr>
<tr>
<td>A8</td>
<td>Street number:</td>
</tr>
<tr>
<td>A9</td>
<td>Building number in census:</td>
</tr>
<tr>
<td>A10</td>
<td>GPS coordinates:</td>
</tr>
<tr>
<td></td>
<td>Longitude: : ___ ___ ° ___ ___ . ___ ___ ___’</td>
</tr>
<tr>
<td></td>
<td>Latitude: : ___ ___ ° ___ ___ . ___ ___ ___’</td>
</tr>
<tr>
<td>A11</td>
<td>Stage number:</td>
</tr>
<tr>
<td>A12</td>
<td>Cluster number:</td>
</tr>
<tr>
<td>A13</td>
<td>Family number in the cluster:</td>
</tr>
<tr>
<td>A14</td>
<td>Name and Phone number of Head of the family:</td>
</tr>
<tr>
<td>A15</td>
<td>Number of the family members:</td>
</tr>
<tr>
<td>A16</td>
<td>Number of family members aged 5-17 years:</td>
</tr>
<tr>
<td>A17</td>
<td>Name of responder: _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _</td>
</tr>
<tr>
<td>A18</td>
<td>Responder relation with the HHD:</td>
</tr>
<tr>
<td>A19</td>
<td>Date of interview: / / 2012</td>
</tr>
<tr>
<td>20A</td>
<td>Interview result:</td>
</tr>
<tr>
<td></td>
<td>1. Meeting conducted</td>
</tr>
<tr>
<td></td>
<td>2. Household not present</td>
</tr>
<tr>
<td></td>
<td>3. House doesn’t exist anymore</td>
</tr>
<tr>
<td></td>
<td>4. Meeting refused</td>
</tr>
<tr>
<td></td>
<td>5. Others ___________________</td>
</tr>
</tbody>
</table>
### B. Household Roster

- Please tell me the names of all individuals living in this household.

**LIST NAMES OF ALL HH MEMBERS (INDIVIDUALS LIVING HERE AND SHARING MEALS) BEFORE ANSWERING B2–B9**

<table>
<thead>
<tr>
<th>ID</th>
<th>Is NAME</th>
<th>What is the relation of NAME to head of the family?</th>
<th>How old is NAME and what is NAME’S birthdate?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>[1. Male 2. Female]</td>
<td>[ENTER AGE IN FULL YEARS AS OF TODAY’S DATE]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>[NOTE: If it is more than 99 years, just write 99]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B1-A</th>
<th>B1-B</th>
<th>B1-C</th>
<th>B2</th>
<th>B3</th>
<th>B4-A</th>
<th>B4-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>What is the marital status of NAME? [ONLY ASK IF 12 OR OLDER]</td>
<td>Where was NAME born?</td>
<td>How long NAME has lived here?</td>
<td>Where did NAME live before 9/4/2003? (Ask only the Head of Household)</td>
<td>What is the main reason NAME was displaced?</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>---------------------------------------------------------------</td>
<td>-----------------------</td>
<td>--------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>-----------------------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>If the answer is not (1) please write the name of place below.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B5</th>
<th>B6</th>
<th>B7-A</th>
<th>B7-B</th>
<th>B8-A</th>
<th>B8-B</th>
<th>B9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### C. Educational Characteristics of Members of the Family – AGE 5 AND OVER ONLY

<table>
<thead>
<tr>
<th>ID #</th>
<th>Can NAME read and write?</th>
<th>Has NAME ever attended school?</th>
<th>Is NAME currently going to school, college or university, or other educational foundation?</th>
<th>What is NAME’s last completed year of schooling?</th>
<th>What is NAME’s highest degree attained?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Yes - full time 2. Yes - part time 3. No</td>
<td>[SEE CODE LIST #1]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### D. Labor Force Participation – AGE 12 AND OVER ONLY

<table>
<thead>
<tr>
<th>ID</th>
<th>Has NAME done any paid work (for cash or in kind) within the last 7 days for at least one hour, whether in a wage job or his/her own enterprise or farm?</th>
<th>Has NAME done any unpaid work in the last 7 days for at least one hour in a family business or farm?</th>
<th>Has NAME done any unpaid work in the last 7 days on another person’s farm or business (including a relative’s)?</th>
<th>Although NAME didn’t work last week, did NAME have any job or business from which he/she was temporarily absent or on a break?</th>
</tr>
</thead>
</table>
|    | 1. Yes, ➔ skip to Section F  
2. No                                                                                                                                 | 1. Yes, ➔ Skip to Section F  
2. No                                                                                      | 1. Yes, ➔ Skip to Section F  
2. No                                                                                      | 1. Yes  
2. No ➔ Skip to Section E                                                                 |

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>D4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID #</td>
<td>What is the main reason NAME didn’t work last week?</td>
<td>During this absence period, has NAME received any wages (in cash or in kind) or enterprise profits?</td>
<td>How long will it take NAME to get back to this work?</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Strike or labor stoppage</td>
<td>1. Yes → skip to Section F 2. No</td>
<td>1. Will be back this same week → SKIP TO SECTION F</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Technical breaks, machinery break-downs</td>
<td></td>
<td>2. In four weeks or less → SKIP TO SECTION F</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Temporary lay-off</td>
<td></td>
<td>3. More than 4 weeks with certain return date→ SKIP TO SECTION F</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Have a job but have not started yet</td>
<td></td>
<td>4. Uncertain</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>The work is seasonal</td>
<td></td>
<td>5. Will not return</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Bad weather or natural phenomena</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Illness or injury</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Leave or vacation/non work related travel</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Lack of raw materials, financial shortage or lack of customers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Education or training leave</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Maternity or parental leave</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>A teacher</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Other______</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Don’t know</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>D5</th>
<th>D6</th>
<th>D7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>2</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>3</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>4</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>5</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>6</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>7</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>8</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>9</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>10</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>11</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>12</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>13</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>14</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>15</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>16</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>17</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>18</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>ID</td>
<td>Was NAME willing and able to work if an opportunity would have been available in the last 7 days?</td>
<td>Has NAME actively sought work (or tried opening a business) in the last four weeks?</td>
<td>How many months has NAME been actively seeking work?</td>
</tr>
<tr>
<td>----</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>1. Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. No ➔ SKIP TO E5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1. Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2. No ➔ SKIP TO E5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1. Yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2. No ➔ SKIP TO E5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**E. Non-Working Cases**
What is the main reason that NAME did not look for work in the last four weeks?

1.  Waiting to go back to the previous employment
2.  Found job and waiting to start
3.  NAME is a full time student
4.  NAME is retired
5.  NAME is a housewife committed to housework
6.  NAME is disabled by physical or mental disease that prevents him/her from working
7.  NAME has an income without needing to work
8.  Believing that there is no work
9.  Tired of looking
10. Not knowing how to look for job
11. Can’t find appropriate work
12. Not qualified for work
13. Temporary family reasons or illness prevented search in last 4 weeks, otherwise would look
14. Other (specify)
99. Doesn’t know

GO TO NEXT HHD MEMBER

<table>
<thead>
<tr>
<th>ID #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1. Waiting to go back to the previous employment</td>
</tr>
<tr>
<td>2</td>
<td>2. Found job and waiting to start</td>
</tr>
<tr>
<td>3</td>
<td>3. NAME is a full time student</td>
</tr>
<tr>
<td>4</td>
<td>4. NAME is retired</td>
</tr>
<tr>
<td>5</td>
<td>5. NAME is a housewife committed to housework</td>
</tr>
<tr>
<td>6</td>
<td>6. NAME is disabled by physical or mental disease that prevents him/her from working</td>
</tr>
<tr>
<td>7</td>
<td>7. NAME has an income without needing to work</td>
</tr>
<tr>
<td>8</td>
<td>8. Believing that there is no work</td>
</tr>
<tr>
<td>9</td>
<td>9. Tired of looking</td>
</tr>
<tr>
<td>10</td>
<td>10. Not knowing how to look for job</td>
</tr>
<tr>
<td>11</td>
<td>11. Can’t find appropriate work</td>
</tr>
<tr>
<td>12</td>
<td>12. Not qualified for work</td>
</tr>
<tr>
<td>13</td>
<td>13. Temporary family reasons or illness prevented search in last 4 weeks, otherwise would look</td>
</tr>
<tr>
<td>14</td>
<td>14. Other (specify)</td>
</tr>
<tr>
<td>99</td>
<td>99. Doesn’t know</td>
</tr>
</tbody>
</table>
Now I'd like to ask you some questions about your work. Many people engage in more than one activity. If NAME does more than one activity, let's start with the one where NAME typically works the most hours in a week.

**NOTE:** REFERS TO MAIN ACTIVITY, EVEN IF THE INDIVIDUAL IS TEMPORARILY AWAY FROM THAT ACTIVITY IN THE LAST WEEK.

Where is the main location where this work takes place?
1. Offices
2. Family home
3. Home other than family's
4. Factory or plant
5. Farm
6. Construction and building
7. Shop, market or kiosk
8. Mobile in various places
9. Others

What is the main activity of this workplace (firm/organization)?
USE CODELIST #2

- 1. 10 or less
- 2. 11-24
- 3. 25-49
- 4. 50-249
- 5. 250-499
- 6. 500 or more
- 99. Doesn't know

[IF MORE THAN 10 OR DOES NOT KNOW SKIP TO FA4]

How many persons are employed in this workplace (firm/organization)?

- 1. Government
- 2. State owned enterprise
- 3. Mixed private-public (e.g., public agency working for a private company; or publicly owned but privately managed)
- 4. Private
- 5. Kurdish or Iraqi NGO
- 6. Foreign private
- 7. Foreign government/foreign NGO
- 99. Doesn't know

What is the status of this workplace?

<table>
<thead>
<tr>
<th>ID #</th>
<th>FA1</th>
<th>FA2</th>
<th>FA3.1</th>
<th>FA3.2</th>
<th>FA4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>2</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>3</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>4</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>5</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>6</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>7</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>8</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>9</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>10</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>11</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>12</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>13</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>14</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>15</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>16</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>17</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>18</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
</tbody>
</table>
## Working Cases: FA – cont.

**What is the main occupation of NAME in this job? Please describe:**

1. Managers
2. Professionals
3. Technicians and associate professionals
4. Clerical support workers
5. Service and sales workers
6. Skilled agricultural, forestry and fishery workers
7. Craft and related trades workers
8. Plant and machine operators, and assemblers
9. Elementary occupations
10. Armed forces occupations

**Is this job:**

1. Permanent
2. Temporary (fixed term; for example, a two-month job NAME might not return to)
3. Seasonal (expect to return to it)
4. Casual (For the day or a few days)
5. Task-based (depends on finishing a particular task, no matter how long)

**In this job/work activity, is NAME a...?**

1. Paid worker
2. Business owner
3. Self-employed
4. Unpaid worker in a different business or farm.
5. Unpaid worker in a family business or farm. voluntary work (other than apprentice)
6. Apprentice (unpaid)

<table>
<thead>
<tr>
<th>ID</th>
<th>FA5</th>
<th>FA6</th>
<th>FA7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID #</td>
<td>How many hours does NAME usually work in this activity per week?</td>
<td>Does NAME receive any of the following benefits? Employer contributes to Social Guarantee Fund for [NAME]?</td>
<td>Does NAME receive any of the following benefits? Health Insurance or access to free or lower cost care at clinics?</td>
</tr>
<tr>
<td>------</td>
<td>---------------------------------------------------------------</td>
<td>---------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>1. Less than 10</td>
<td>1. Yes</td>
<td>1. Yes</td>
</tr>
<tr>
<td>2</td>
<td>2. 11-20</td>
<td>2. No</td>
<td>2. No</td>
</tr>
<tr>
<td>4</td>
<td>4. 31-40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>5. 41-50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>6. 51 or more</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID #</td>
<td>Did NAME sign a contract for this job?</td>
<td>Taking into account wages, profits, tips, fees, and other payments, what is the typical income that NAME earns at this work activity?</td>
<td>Does NAME have another work activity?</td>
</tr>
<tr>
<td>------</td>
<td>--------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>------------------------------------------</td>
</tr>
</tbody>
</table>

**FA13**

<table>
<thead>
<tr>
<th></th>
<th>FA14.1: AMOUNT</th>
<th>FA14.2: TIME UNIT</th>
<th>FA15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>2</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>3</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>4</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>5</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>6</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>7</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>8</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>9</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>10</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>11</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>12</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>13</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>14</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>15</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>16</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>17</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>18</td>
<td>___</td>
<td>___</td>
<td>___</td>
</tr>
</tbody>
</table>
## Working Cases: FB  SECOND ACTIVITY

<table>
<thead>
<tr>
<th>ID</th>
<th>Where is the main location where this work takes place?</th>
<th>What is the main activity of this workplace (firm/organization)?</th>
<th>How many persons are employed in this workplace (firm/organization)?</th>
<th>What is the status of this workplace?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Offices</td>
<td>USE CODELIST #2</td>
<td>1. 10 or less</td>
<td>1. Government</td>
</tr>
<tr>
<td>2</td>
<td>Family home</td>
<td></td>
<td>2. 11-24</td>
<td>2. State owned enterprise</td>
</tr>
<tr>
<td>3</td>
<td>Home other than family’s</td>
<td></td>
<td>3. 25-49</td>
<td>3. Mixed private-public (e.g., public agency working for a private company; or publicly owned but privately managed)</td>
</tr>
<tr>
<td>4</td>
<td>Factory or plant</td>
<td></td>
<td>4. 50-24</td>
<td>4. Private</td>
</tr>
<tr>
<td>5</td>
<td>Farm</td>
<td></td>
<td>5. 250-499</td>
<td>5. Kurdish or Iraqi NGO</td>
</tr>
<tr>
<td>6</td>
<td>Construction and building</td>
<td></td>
<td>6. 500 or more</td>
<td>6. Foreign private</td>
</tr>
<tr>
<td>7</td>
<td>Shop, market or kiosk</td>
<td></td>
<td>9. Doesn’t know</td>
<td>7. Foreign government/foreign NGO</td>
</tr>
<tr>
<td>8</td>
<td>Mobile in various places</td>
<td></td>
<td>IF 10 OR LESS, DOES NOT KNOW</td>
<td>9. Doesn’t know</td>
</tr>
<tr>
<td>9</td>
<td>Others</td>
<td></td>
<td>SKIP TO F4A</td>
<td></td>
</tr>
</tbody>
</table>

*IF MORE THAN 10 OR DOES NOT KNOW, ASK RESPONDENT TO SPECIFY HOW MANY*

### Table

<table>
<thead>
<tr>
<th>ID</th>
<th>FB1</th>
<th>FB2</th>
<th>FB3.1</th>
<th>FB3.2</th>
<th>FB4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Working Cases: FB – cont.

<table>
<thead>
<tr>
<th>ID</th>
<th>FB5</th>
<th>FB6</th>
<th>FB7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**What is the main occupation of NAME in this job?**

1. Managers
2. Professionals
3. Technicians and associate professionals
4. Clerical support workers
5. Service and sales workers
6. Skilled agricultural, forestry and fishery workers
7. Craft and related trades workers
8. Plant and machine operators, and assemblers
9. Elementary occupations
10. Armed forces occupations

**Is this job:**

1. Permanent
2. Temporary (fixed term; for example, a two-month job NAME might not return to)
3. Seasonal (expect to return to it)
4. Casual (for the day or a few days)
5. Task-based (depends on finishing a particular task, no matter how long)

**In this job/work activity, is NAME a...?**

1. Paid worker
2. Business owner
3. Self-employed
4. Unpaid worker in a family business or farm.
5. Unpaid worker in a different business or voluntary work (other than apprentice)
6. Apprentice (unpaid)
<table>
<thead>
<tr>
<th>ID #</th>
<th>How many hours does NAME usually work in this activity per week?</th>
<th>Does NAME receive any of the following benefits?</th>
<th>Does NAME receive any of the following benefits?</th>
<th>Does NAME receive any of the following benefits?</th>
<th>Does NAME receive any of the following benefits?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>How many hours does NAME usually work in this activity per week?</td>
<td>Does NAME receive any of the following benefits?</td>
<td>Does NAME receive any of the following benefits?</td>
<td>Does NAME receive any of the following benefits?</td>
<td>Does NAME receive any of the following benefits?</td>
</tr>
<tr>
<td></td>
<td>Employer contributes to Social Guarantee Fund for [NAME]?</td>
<td>Health Insurance or access to free or lower cost care at clinics?</td>
<td>Paid Vacation?</td>
<td>Paid sick leave?</td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>Less than 10</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>2.</td>
<td>11-20</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>3.</td>
<td>21-30</td>
<td>Doesn’t know</td>
<td>Doesn’t know</td>
<td>Doesn’t know</td>
<td>Doesn’t know</td>
</tr>
<tr>
<td>4.</td>
<td>31-40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>41-50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>51 or more</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.</td>
<td>Doesn’t know</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ASK FOR NUMBERS OF DAYS USUALLY WORKED PER WEEK AND USUAL HOURS PER DAY AND MULTIPLY

IF RESPONDENT IS A PAID WORKER (FB7=1), ASK FB9-FB13; OTHERWISE, SKIP TO FB14
### Working Cases: FB – cont.

<table>
<thead>
<tr>
<th>ID #</th>
<th>Did NAME sign a contract for this job?</th>
<th>Taking into account wages, profits, tips, fees, and other payments, what is the typical income that NAME earns at this work activity?</th>
<th>Does NAME have another work activity?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1. Yes</td>
<td>USE ONE OF THE FOLLOWING TIME PERIODS BASED ON WHAT IS EASIEST FOR RESPONDENT</td>
<td>1. Yes</td>
</tr>
<tr>
<td></td>
<td>2. No</td>
<td>1. Monthly (in thousand IDs)</td>
<td>2. No, → Section G</td>
</tr>
<tr>
<td></td>
<td>9. Does not know</td>
<td>2. Bi-Weekly (in thousand IDs)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Weekly (in thousand IDs)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. Daily (in thousand IDs)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[SPECIFY AMOUNT and TIME UNIT CODE (1, 2, 3 or 4)]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>2</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>3</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>4</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>5</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>6</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>7</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>8</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>9</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>10</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>11</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>12</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>13</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>14</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>15</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>16</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>17</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
<tr>
<td>18</td>
<td>___</td>
<td>___ ___ ___ ___</td>
<td>___</td>
<td>___</td>
</tr>
</tbody>
</table>
### Working Cases: FC (Third Activity).

<table>
<thead>
<tr>
<th>ID</th>
<th>Where is the main location where this work takes place?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Offices</td>
</tr>
<tr>
<td>2.</td>
<td>Family home</td>
</tr>
<tr>
<td>3.</td>
<td>Home other than family's</td>
</tr>
<tr>
<td>4.</td>
<td>Factory or plant</td>
</tr>
<tr>
<td>5.</td>
<td>Farm</td>
</tr>
<tr>
<td>6.</td>
<td>Construction and building</td>
</tr>
<tr>
<td>7.</td>
<td>Shop, market or kiosk</td>
</tr>
<tr>
<td>8.</td>
<td>Mobile in various places</td>
</tr>
<tr>
<td>9.</td>
<td>Others</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>What is the main activity of this workplace (firm/organization)?</th>
</tr>
</thead>
</table>
| USE CODELIST #2 | 1. 10 or less  
2. 11-24  
3. 25-49  
4. 50-249  
5. 250-499  
6. 500 or more  
9. Doesn’t know |

<table>
<thead>
<tr>
<th>ID</th>
<th>How many persons are employed in this workplace (firm/organization)?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>10 or less, ASK RESPONDENT TO SPECIFY HOW MANY</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>What is the status of this workplace?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Government</td>
</tr>
<tr>
<td>2.</td>
<td>State owned enterprise</td>
</tr>
<tr>
<td>3.</td>
<td>Mixed private-public (e.g., public agency working for a private company; or publicly owned but privately managed)</td>
</tr>
<tr>
<td>4.</td>
<td>Private</td>
</tr>
<tr>
<td>5.</td>
<td>Kurdish or Iraqi NGO</td>
</tr>
<tr>
<td>6.</td>
<td>Foreign private</td>
</tr>
<tr>
<td>7.</td>
<td>Foreign government/foreign NGO</td>
</tr>
<tr>
<td>9.</td>
<td>Doesn’t know</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FC1</th>
<th>FC2</th>
<th>FC3.1</th>
<th>FC3.2</th>
<th>FC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td>What is the main occupation of NAME in this job?</td>
<td>Is this job:</td>
<td>In this job/work activity is NAME a...?</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-----------------------------------------------</td>
<td>--------------</td>
<td>----------------------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1. Managers</td>
<td>1. Permanent</td>
<td>1. Paid worker</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2. Professionals</td>
<td>2. Temporary (fixed term; for example, a two-month job NAME might not return to)</td>
<td>2. Business owner</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Technicians and associate professionals</td>
<td>3. Seasonal (expect to return to it)</td>
<td>3. Self-employed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Clerical support workers</td>
<td>4. Casual (For the day or a few days)</td>
<td>4. Unpaid worker in a family business or farm.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. Service and sales workers</td>
<td>5. Task-based (depends on finishing a particular task, no matter how long)</td>
<td>5. Unpaid worker in a different business or voluntary work</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6. Skilled agricultural, forestry and fishery workers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7. Craft and related trades workers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8. Plant and machine operators, and assemblers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>9. Elementary occupations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10. Armed forces occupations</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|    | 1
|    | 2
|    | 3
|    | 4
|    | 5
|    | 6
|    | 7
|    | 8
|    | 9
|    | 10
|    | 11
|    | 12
|    | 13
|    | 14
|    | 15
|    | 16
|    | 17
|    | 18

<table>
<thead>
<tr>
<th></th>
<th>FC5</th>
<th>FC6</th>
<th>FC7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID #</td>
<td>How many hours does NAME usually work in this activity per week?</td>
<td>Does NAME receive any of the following benefits? Employer contributes to Social Guarantee Fund for [NAME]?</td>
<td>Does NAME receive any of the following benefits? Health insurance or access to free or lower cost care at clinics?</td>
</tr>
<tr>
<td>------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------</td>
</tr>
<tr>
<td>1</td>
<td>Less than 10</td>
<td>1. Yes</td>
<td>1. Yes</td>
</tr>
<tr>
<td>2</td>
<td>11-20</td>
<td>2. No</td>
<td>2. No</td>
</tr>
<tr>
<td>4</td>
<td>31-40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>41-50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>51 or more</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Doesn’t know</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Working Cases: FC – cont.

<table>
<thead>
<tr>
<th>ID #</th>
<th>Did NAME sign a contract for this job?</th>
<th>Taking into account wages, profits, tips, fees, and other payments, what is the typical income that NAME earns at this work activity?</th>
<th>Does NAME have more than three jobs or work-related activities?</th>
<th>How many hours does NAME typically spend in all these other work activities per week? (NOT COUNTING THE FIRST 3 ACTIVITIES)</th>
</tr>
</thead>
</table>
|      | 1. Yes  
2. No  
9. Does not know | USE ONE OF THE FOLLOWING TIME PERIODS BASED ON WHAT IS EASIEST FOR RESPONDENT  
1. Monthly (in thousand IDs)  
2. Bi-Weekly (in thousand IDs)  
3. Weekly (in thousand IDs)  
4. Daily (in thousand IDs)  
[SPECIFY AMOUNT and TIME UNIT CODE (1, 2, 3 or 4)] | 1. Yes,  
2. No -> SKIP TO SECTION G, |  |

<table>
<thead>
<tr>
<th></th>
<th>FC13</th>
<th>FC14.1: AMOUNT</th>
<th>FC14.2: TIME UNIT</th>
<th>FC15</th>
<th>FC16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**G. Part-Time and Under Employment: IF NAME HAD MORE THAN ONE ACTIVITY, ASK G1; IF ONLY ONE ACTIVITY, START AT G2**

<table>
<thead>
<tr>
<th></th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
<th>G4</th>
<th>G5</th>
<th>G6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ADD THE HOURS IN ALL JOBS IN A TYPICAL WEEK, FROM F8A, F8B, F8C AND F16C. INDICATE THOSE HOURS:

THEN ASK: Taking together the average hours per week you told me you worked in each activity, it seems that your total normal hours per week is:

**Is this correct?**
1. Correct
2. Incorrect

IF INCORRECT, ASK RESPONDENT TO CONFIRM THE HOURS IN EACH ACTIVITY AND CHANGE F8A, F8B, F8C AND F16C AS NECESSARY, THEN RE-ENTER TOTAL BELOW

**Would NAME want to work more hours per week than his/her current situation?**
1. Yes
2. No (GO TO NEXT HHD MEMBER)

**Since NAME would like to work more, what is the MAIN reason working less than what he/she would like to?**

IF MULTIPLE, CHOOSE MOST IMPORTANT
1. Slack work/business conditions
2. Could only find part-time work
3. Seasonal work
4. Child care problems
5. Other family/personal obligations
6. Health/medical limitations
7. School/training
8. Pay is not enough
9. Other

Was NAME willing and able to work an additional job/work activity or more hours at the current one if an opportunity to do so would have been available in the last 7 days?

1. Yes
2. No
Has NAME sought additional work, a job with more hours—or sought to open a business in the last four weeks?

<table>
<thead>
<tr>
<th>ID</th>
<th>Has NAME sought additional work, a job with more hours—or sought to open a business in the last four weeks?</th>
<th>How many months has NAME been seeking additional work or a job with more hours?</th>
<th>Which of the following has NAME done to find additional work or a job with more hours?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td></td>
<td>1. Registration with job-seeking agencies (such as a public employment service)</td>
</tr>
<tr>
<td>2</td>
<td>No (GO TO NEXT HHD MEMBER)</td>
<td></td>
<td>2. Visiting governmental in institutions</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. Directly attend a workplace (factory, shop, facility)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. Sent application to employers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5. Look at advertisements in newspapers</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6. Look at job advertisements in the internet</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>7. Respond to newspaper advertisements in the internet</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8. Ask relatives and friends to inform you about jobs</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>9. Requested a loan in order to start a business</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>10. Sought land, buildings, equipment or machinery to start a business</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11. Applied for a permit or license to start a business</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>12. Others (specify)</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>G7</th>
<th>G8</th>
<th>G9a</th>
<th>G9b</th>
<th>G9c</th>
<th>G9d</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Researcher's Name and Signature:</td>
<td>Researcher's Code:</td>
<td>Date:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
<td>-------------------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>_________________</td>
<td>2012 //</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Supervisor's Name and Signature:</th>
<th>Supervisor's Code:</th>
<th>Date:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>_________________</td>
<td>2012 //</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Auditor's Name and Signature:</th>
<th>Auditor's Code:</th>
<th>Date:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>_________________</td>
<td>2012 //</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Entry's Name and Signature:</th>
<th>Data Entry's Code:</th>
<th>Date:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>_________________</td>
<td>2012 //</td>
</tr>
</tbody>
</table>
This appendix reproduces the sample report described in the companion report.
Unemployment Rate in the Kurdistan Region of Iraq at 7.4%, Labor Force Survey Finds

The Kurdistan Regional Statistics Office (KRSO) announced today that the overall unemployment rate across the Kurdistan Region of Iraq (KRI) was 7.4%. The unemployment rate is the percentage of individuals in the labor force who have not been working but are looking for work. These findings are based on a new periodic survey of Kurdistan’s labor force conducted by KRSO in July 2012. The survey is in line with international best practices and will be conducted on a quarterly basis, providing up-to-date data for policymakers and the public on the KRI Labor Market. The other main findings from the survey are given below.

Men’s Unemployment Rate is 5.0%, Compared to 19.9% for Women

The unemployment rate differs significantly between men (5.0%) and women (19.9%). The male unemployment rate is at least 10 percentage points lower for age groups between 15 and 44 years old.

Unemployment Rate is Lowest in 55-64 Age Group

The unemployment rate is lowest for those in the 55-64 age group (1.5%). The group between 15 and 24 has the highest rate (17.6%), approximately 10 percentage points higher than the average unemployment rate.
Labor Force Participation Rate is 38.3%, Varies by Age Group and Gender

The labor force participation rate is the percentage of employed and unemployed workers (those that are either working or looking for work) as a fraction of the population 15 and older. As of July 2012, the rate for the Kurdistan Region of Iraq stood at 38.4.

The participation rate of those aged 25-34, 35-44, and 45-54 is above average, at 51.4%, 55.4%, and 45.4%, respectively. The participation rate of persons aged 65 and older is 9% and of aged 15-24 equaled 22%. Total labor force participation is significantly higher for men (65.8%) when compared to women (12.2%).

Both Labor Force Participation and Unemployment Rates are Lower in Rural Areas and Among the Less Educated

<table>
<thead>
<tr>
<th></th>
<th>Labor Force Participation (%)</th>
<th>Unemployment (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rural areas</td>
<td>35.3%</td>
<td>6.0%</td>
</tr>
<tr>
<td>Urban areas</td>
<td>39.0%</td>
<td>7.7%</td>
</tr>
<tr>
<td>Primary education</td>
<td>53.6%</td>
<td>6.9%</td>
</tr>
<tr>
<td>Secondary education</td>
<td>41.3%</td>
<td>8.5%</td>
</tr>
<tr>
<td>College degree</td>
<td>80.3%</td>
<td>10.9%</td>
</tr>
</tbody>
</table>

All correspondence should be directed to:
Kurdistan Regional Statistics Office, Erbil
Phone: 00964(0)66 222 01 70 • Email: info@mop-krg.org
This appendix reproduces the briefing materials created to introduce Kurdistan Regional Statistics Office Personnel to sampling theory, procedures, and related issues, as described in the companion report.
An Introduction to Sampling

Bonnie Ghosh-Dastidar

Kurdistan Data Project

KRSO Workshop

Erbil

February 2012

Do not distribute without permission of authors
Outline

• Sampling
  – Probability sampling
  – Stratified and Clustered Designs

• Construction and use of sampling weights

• Sample size considerations
  – Precision (Margin of error), Power
Why Sample?

• Census is expensive!
  – Cost of implementation
  – Respondent burden
  – Large sample sizes often means shorter questionnaire

• Sampling can produce reliable estimates of the underlying population using smaller sample sizes
Key Sampling Concepts

• Who do you want to generalize to?

• Who can you access?

• How can you get access to them?

• Who is in the study?

• Theoretical, underlying population

• The study population

• The sampling frame

• The study sample
Considerations for Sampling

• **Coverage**: Quality of sampling frame

• **Generalizability**: Probability sample with known prob. for each element in the study

• **Minimum error**: Sampling and non-sampling errors

• **Sample Size/Precision**: Adequate sample sizes to ensure reliable estimates
Sampling Process

Units of Analysis (people)

- Target Population
  Population of Interest

Actual Population to Which Generalizations Are Made
Defined/Listed by Sampling Frame

- Sample
  The people actually studied

List or Procedure

- Sampling Frame
  List or Rule
  Defining the Population

- List of Target Sample

Generalization

Response Rate
Two Main Types of Samples

- Probability sample
  - You *know* the probability of inclusion for all members of the population (don’t have to be equal)
  - No one in the population has zero prob. of inclusion
  - Statistical inference valid
    - Household survey from sampling frame
    - Mail survey from list
    - Random digit dial (RDD) of phone banks
Two Main Types of Samples

- Judgment/convenience sample
  - Volunteerism or unsystematic approach makes probabilities of inclusion *unknown*
  - Statistical inference not valid
  - Mall intercept, inbound calls, etc.
    - Network sample
    - Sample of immigrants through locations
Some Types of Probability Samples

- Simple random sample (SRS)
- Systematic sample
- Stratified sample
- Cluster sample
- Combinations
Estimation Error

· *Bias*: Systematic (*non-sampling*) error
  – Expected difference between sample and population parameter

· *Variance*: Unsystematic (*sampling*) error

· *Mean squared error* (*MSE*): Total estimation error
  = Variance + squared bias
**Bias and Sample Size**

• Bias is typically “invisible” to software packages.

• Larger sample sizes reduce variance component, but not bias.

• As sample sizes get bigger, bias dominates variance and becomes most of the MSE.
Types of Non-sampling Error

- Whom you select: Selection bias
- Who responds: Non-response bias
- How they respond: Response bias
Whom you select: Selection bias

- Are some people systematically omitted, over-represented, or under-represented?
- Example: Rich vs poor respondents
- Solutions
  - Alter sampling approach
  - Design weights
  - Re-define population
Who responds: Nonresponse bias

– Are the people who respond like those who do not in important ways?

– Solutions
  · Drive response rate up
  · Alter survey approach
  · Non-response weights
How they respond: Response bias

– When people answer, do the answers themselves systematically differ from the truth?

Solutions

• Redesign questions to deal with social desirability, etc.
• Regression adjustment/calibration
A little bias goes a long way

• If we use $\bar{x}$ to represent standardized bias ($\bar{x} = b / \sigma$)

  $b = \text{bias}$; $\sigma = \text{standard deviation}$ (for a proportion variable = $\sqrt{p(1-p)}$)

• It can be shown that the effective sample size (ESS) for an estimate is

  $$\text{ESS} = n / (1 + n \sigma^2)$$

  $e = 0.02$ (1% bias on 50% outcome, 0.6% bias on 10% outcome)

  -kills 4% of the “efficiency” (MSE-based ESS) at $n=100$

  -kills 67% of the efficiency at $n=5000$

  $e = 0.10$ (5% bias on 50% outcome, 3% bias on 10% outcome)

  -kills 67% of the efficiency at $n=100$

  -kills 98% of the efficiency at $n=5000$
Stratified Sampling
Simple random sampling

• *Epsem sampling:* *(Equal Probability Sampling)* Every element in the study population is equally likely to be selected

• Once you’ve identified a listing, draw a sample of size $N$ without replacement

• Often not ideal
  – No guarantee that every sub-population of interest is in the sample
  – May be too much geographic dispersion
Systematic sampling

• Involves the selection of elements from an ordered list of elements (frame).

• The most common is an equal-probability method, in which every $k^{th}$ element in the frame is selected

• $K=$ sampling interval is set to be $N/n$
  – *Where* $n=$ sample size, $N=$ population size

• More efficient than SRS
**Stratification**

- Stratification is the process of dividing the sampling frame into sub-groups (strata) of elements or homogeneous (similar) PSUs.

- Advantages: better precision, flexible design, sub-national estimates for smaller domains (differential sampling rates)
  - Reduced variance within stratum given similarity of units.

- Example of stratification: Region, Urban/rural.
**Stratification and Sample Allocation**

- Proportional allocation
  - Effective for precision of estimates at the national level

- Equal allocation to each domain (form of disproportionate sampling)
  - Used when each domain requires same level of precision

- Optimum allocation – takes into account differential variance and costs by stratum
  - For example, variability may be higher in urban areas and enumeration costs may be higher in rural areas – use higher sampling rate for urban areas
Stratified Sampling Using Proportionate Allocation

- Simple approach

- Sample sizes selected per stratum are proportionate to the percentage of the population they represent

- Efficient: produces a self-weighting sample

- Usually improves precision as compared to an SRS (can’t worsen)
  - Depends on relationship of stratification variable to what you are estimating; i.e., is there variation in the outcome of interest by Strata.
Example: Unemployment rate

• 15.6% is national unemployment rate (UR)
  - Under SRS (using binomial dist.)
    • Variance(Unemp. Rate) = (.156)(1-.156) = .1316
• Say,
  • 30% of a 10% sub-population are unemployed,
  • Remaining 90% of population have 14% unemp. rate
• To reduce the overall variance, we can allocate sample proportionately so that our sample has exactly 10% from the high unemployment area
• Note: we need to identify/know the sub-populations
Relationship between SRS and PSS Variances

• For proportionate stratification sampling (PSS), the variance is \( \text{Var}_{PSS} \), the weighted average of the within strata variances.

• The relationship between the SRS variance and the PSS variance for strata \( i=1,\ldots,k \) is

\[
\text{Var}_{SRS} = \text{Var}_{PSS} + \sum_{i=1}^{k} w_i (UR_i - UR)^2
\]
Proportionate Stratification Results in Smaller Variance

If UR = Unemployment Rate, and w=Percentage of Population, then if

\[ w_1 = .1 \quad w_2 = .9 \]
\[ UR = .1(30) + .9(13) = 15.6 \]
\[ UR_1 = .30 \quad UR_2 = .14 \]

Recall SRS \( \sigma^2 = .1316 \)

For PSS \( \sigma^2_w = (.1)(.30)(1-.30) + (.9)(.13)(1-.14) = .1294 \)

\[ \sum_{i=1}^{k} w_i (UR_i - UR)^2 = .1(.30 - .156)^2 + .9(.14 - .156)^2 = .0023 \]

and \( .1316 = .1294 + .0023 \) (relationship seen on previous slide)

\[ Var_{PSS}^2 = .983(Var_{SRS}^2) \]
Pros/Cons of Proportionate Sampling

• Pros:
  – Proportionate stratification produces *self weighted* estimates; no need for weights
  – Efficient with smaller variances than Disproportionate

• Cons:
  – Disproportionate allocation may be needed to ensure min. sample sizes per stratum
  – You cannot (generally) show reliable strata-level estimates for the smaller strata
**Stratified Sample Using Disproportionate Allocation**

- Fully divide the population of interest into mutually exclusive strata

- Sample is not allocated proportionately to stratum size

- Sample weight for an observation = inverse of the sampling probability of the observation
Iraq Household Socio-economic Survey 2007

- Objectives: Provide national estimates of poverty; establish new CPI; produce estimates of consumption expenditure
  - Comparisons across governorates
  - Comparisons across rural and urban
  - Minimum sample size per governorate
**IHSES (2007)**

- Produce nationally representative sample, also regional coverage

- Large sample necessary to provide precision at smaller levels

- Sensitive data > face-to-face survey

- Updating of frame > 1997 population census frame (10 years old) and didn’t include 3 governorates in Kurdistan
IHSES (2007)

• Launched November 1, 2006

• Sample size = 18,144 households

• Strata
  - Governorate (18)
  - Strata: Urban, rural, metropolitan (3)
  - $18 \times 3 = 54 + 2$ (Baghdad had 5 strata)
**IHSES (2007)**

- Disproportionate sampling
  - Sample not allocated proportional to stratum size
  - Instead, equal sample sizes per stratum guarantee equal precision
    - 324 hh per stratum (Gov. -- Rural/Urban/Met)
    - Or, 972 hh per Governorate across 3 strata
    - 1,620 hh in Baghdad with 5 strata
Cluster Sampling
Pros & Cons of Cluster Sampling

· Advantages
  – Feasibility (two-stage access to units)
  – Cost-effective (travel)
  – Want clusters for analysis (e.g. market, community)

· Disadvantage:
  – Loss of information because of homogeneity within groups (variance inflation)
Design Effects of Complex Surveys

• Design effect (DEFF) of a complex design =

  \( \frac{\text{variance of the estimate obtained via the complex design}}{\text{variance of the estimate obtained via a SRS with the same sample size}} \)

• May come from stratification/weighting (from disproportionate sampling)
  – Applies to all outcomes equally – same weight for all units in the same stratum across all outcomes (no clustered sampling involved)

• May come from clustering
  – Affects all outcomes, but not equally (depends on ICC for the outcome, defined below).
Meaning of Design Effects

• DEFF>1 ---> loss of precision relative to an SRS
  – Most common for complex survey design

• DEFF=1 ---> precision is equivalent to that of an SRS

• DEFF<1 ---> gain in precision relative to an SRS
  – may happen with a proportionately stratified sample
Effective Sample Size

• Effective Sample Size (ESS) is the sample size of simple random sample to which the current sample size is equivalent given the DEFF from the complex survey design.

• ESS = N / DEFF

• Effective Sample Size Translates Design effects (DEFF) into Sample Size terms.
  – The ESS of a SRS is the nominal Sample Size
  – Allows comparison alternative design in common terms
    • Minimize cost per ESS
    • Maximize ESS under cost constraints
Design Effects in Cluster Sampling

• Size of the clusters (B) and the degree of similarity of items within a cluster, as measured by the *intra-class correlation coefficient* =\(ICC=r\), increase the loss of precision in a cluster sample

• ICC is the proportion of variance of individual outcomes attributable to clusters

  \[ r = 0 \quad \rightarrow \quad \text{outcomes are as heterogeneous within clusters as between (random assignment to clusters)} \]

  \[ r > 0 \quad \rightarrow \quad \text{outcomes are more homogeneous within clusters than between (most common)} \]

  Typical values of \( r \) are between 0.01 and 0.15, but can be higher
Formula for DEFF for Cluster Samples

\[ r \approx \frac{B(\text{Var(cluster means)}) - \text{Var}(X)}{(B - 1)\text{Var}(X)} \]

\[ \text{DEFF} \approx 1 + (B - 1)r \]

- If \( \text{Var(cluster means)} = \text{Var}(X) \), then \( r = 0 \)
- Related to F-Stat in 1-way ANOVA (with clusters as groups)
- Can derive from PROC VARCOMP (SAS) or simple hierarchical models
Multi-stage sampling

• Two-stage sampling is a type of Cluster Sampling
  – **First stage**: select a sample of clusters (with or without stratification); these are the Primary Sampling Unit (PSU).

  – **Second stage**: Instead of selecting every element in each cluster, you may choose to sample within cluster.
Iraq Household Socio-economic Survey (2007)

• Primary Sampling Unit
  – Each PSU is defined to have 70-100 Households
  – Combine villages in Rural areas; combine majals in Urban areas
  – PSU selected with probability proportionate to size

• Within PSU, households were sampled using updated 2006 frame.
  – The 1997 Population Census frame was updated in 2006 for all sampled PSUs
IHSES (2007)

- **First stage:** 54 sample points (PSUs) were selected per strata (56 strata)
  - 3,024 sampled points
  - Sampled with *Probability proportionate to size*
  - Systematic sampling used to draw sample

- **Second stage:** 6 households per PSU

- Total sample of 18,144 households
  - 6 households per 3,024 primary sampling unit
**IHSES (2007) -- Implementation**

- Listing re-done in 3,024 sampled points in 18 governorates
  - 270 in Baghdad
  - 162 in each of the other 17 Governorates

- Organized neighbor sampled points into groups of 3 or “trios” (*fieldwork efficiency*)

- One year period was broken in 18 waves w/ 3 weeks per wave (*accountability*)

- If each team interviews 1 trio per wave, you need 56 teams working the 18 waves.
Clustering in IHSES (2007)

- Suppose \( r = 0.05 \) within each PSU; a sample of 6 households is drawn from each of the 3,024 clusters.

- Total sample size = 18,144

- \( \text{DEFF} = 1 + (6-1)(0.05) = 1.25 \)

- The variance of the sample mean estimated from the cluster design is 1.25 times as large as it would be based on an SRS of 18,144 students.

- Note this is not a big difference, because \( r \) is low.

- Effective sample = \( 18,144 / 1.25 = 14,515 \) households.

- If 24 households were chosen from each of 756 clusters, for a total sample size of 18,144, then \( \text{DEFF} = 2.2 \) and effective sample = 8,247.
ICC limits usefulness of more observations within a cluster

• Let \( r \) be the ICC, \( B \) be the cluster size

• \( 1/r \) is the maximum ESS per cluster

• \( Br \) is a measure of how “saturated” the clusters are relative to the ICC

• As \( Br \) increases, the marginal value of observations added to a cluster drops rapidly
  - >50% of maximum ESS is achieved at \( Br=1 \); >75% at \( Br=3 \); >90% at \( Br=9 \)

• For example, at \( r=0.05 \), 20 is the maximum ESS per cluster
  - \( B=19 \rightarrow \) ESS of 10 /cluster; \( B=57 \rightarrow \) ESS of 15/cluster; \( B=171 \rightarrow \) ESS of 18/cluster
  - If extra observations within a cluster have marginal cost, they are wasteful beyond a certain point
Weighting
Weighting

• Calculating and using sample weights

• Design weights and non-response weights

• “Design effects” of weighting (variance inflation)

• Trade-offs in weighting
  • Fixing problems
  • When not to use weights
Purpose of Weights-1

• Imagine that we want to know the proportion of recent inpatients who would recommend their hospital to friends and family.

• From a list, we send out a survey to a subset, which is completed by a subset of that group

• How do we estimate the above parameter from our survey responses?
Purpose of Weights-2

• If all on the list were equally likely to be surveyed and equally likely to respond, we could simply average the outcome among respondents.

• If, however, the above conditions are not met, and there is some association between these probabilities and the characteristics we are measuring, simple averages will be biased.

• Weights can reduce or eliminate these biases.
Weights Have Limitations

• Weights cannot turn a convenience sample into a probability sample

• In a probability sample, weights are not always needed when sampling probabilities are unequal (sometimes you do not need to generalize to the population).

• Poorly designed weights can make inference less accurate
Types of Weights-1

• Design weights
  – Correct for *known* differing probabilities of selection of population members into the sample we attempt to contact
  – Used with disproportionate stratified random sampling (e.g., attempt to contact 10% of one subgroup but 20% of another subgroup)

• Non-response weights
  – Correct for *estimated* differing probabilities of participation among those we attempt to contact, using information available for both non-respondents and respondents
  – Example: perhaps 74% of females and 52% of males respond
Types of Weights-2

• Post-stratification weights

• Correct for *estimated* differing probabilities of population members into the sample of respondents using characteristics that are known for the general population but which are not known about the individual members of the population until they respond
  – Example: suppose we did not know the race/ethnicity of non-respondents, but we did know the true distribution of race/ethnicity for the population from a separate data source, and we want our sample to be representative of race/ethnicity
Creating Weights

– In a typical survey with non-response, both design and non-response weights are involved.
  • Design weights reflect the probability of selection from the population into the subset we attempt to contact \(p_{sel}\)
    – \(DW=1/ p_{sel}\)
  • Non-response weights reflect the estimated probability of people like the respondent responding, given that we attempted to contact them \(p_{res}\)
    – \(NRW=1/ p_{res}\)
Creating Weights

- Overall weights reflect the probability of selection from the population into sample of respondents \((p_{sel})*(p_{res})\)
- \(OW=1/((p_{sel})*(p_{res}) )= DW*NRW\)
- Overall weights can be estimated directly as the inverse of the fraction of population members who respond within each stratum
- They can also be created in stages: design weights, then non-response weights, then multiply to create overall weight; This approach often has advantages
Design Weights IHSES (2007)

• The selection probability of a Household in PSU $i$ within Stratum $h$ ($H_{ij}$), is given by
  
  $P(h_{ij}) = k_h \times (S_{hi} / N_h) \times (n_{hi} / S'_{hi})$ .... (1)

  $k_h =$ number of PSUs in stratum $h$

  $S_{hi} / N_h =$ Size of PSU $i$ in 1997 frame to Size of stratum $h$ in 1997 frame

  $n_{hi} / S'_{hi} =$ Number of sampled households in PSU $i$

  relative to Size of PSU $i$ in 2006 Listing

• The design weight is the inverse of the selection probability (1)
Creation of Non-response Weights

• Design weights are based on known probabilities (you chose the number per strata), so these are just ratios in a simple stratified sample. Often simple.

• Non-response weights involve estimating the probability that people like the respondent respond when contacted.
  – These estimates are based on an explicit model of non-response. (e.g. Logistic Regression).
  – Incorporating additional variables that are predictive of non-response can further decrease bias. For example, a standard logistic or probit regression model to predict Response status controlling for observed covariates.
  – Incorporating un-predictive variables (including strata that are too fine for non-response) can add noise to the weights and the estimates.
Response Rates and Non-response Bias

• The main reason there is a push for high response rates (beyond the desire for larger sample sizes) is that higher response rates leave less room for differential non-response by personal characteristics.

• While it is possible that non-response is entirely random with a 10% response rate and highly selective with a 70% response rate, the reverse is more often true.

• But the relationship is weak (Groves and Petcheya 2008); good probability sampling process is most important.
**Statistical Significance of Non-response Models**

- A statistically significant non-response model can correct for differential non-response on the basis of the predictor variables in the model
  - For example, if rich, white males have a different prob. of response than poor, black males, you can include these covariates in an explicit model of non-response (Logistic regression)

- A statistically non-significant non-response model can provide some evidence that non-response does not differ on the basis of the predictor variables in the model if the sample size is large enough for adequate power
Sample Sizes
Sample Size, Power, Precision, and Confidence Intervals

• What sample size do I need (for a certain amount of power/precision)?

• What can I say with a given sample size?

• What matters and how much?
  – Proportions vs. Means
  – Effects of N, confidence level, significance level, power level, allocation/balance, pairing
Power Versus Precision

· **Power** has to do with the ability to detect differences of a given magnitude in hypothesis testing

· **Precision** refers to the amount of variability present in point estimates in estimation

· Precision is a more basic concept, if we understand it, we understand power
  · The power you have is largely determined by the precision you have
Sample Size and Precision

Width of a confidence interval is inversely related to the square root of the sample size.

For a Confidence Interval that is 1/3 as wide, multiply n by 9

1/2  4
2/3  2.25
3/4  1.78

Cutting sample size by 10% multiplies Conf. Interval width by 1.05

20%  1.12
30%  1.20
40%  1.29

• 50%  1.41
Precision/Power “Worse” for Proportions

• Standard deviations are “large” for proportions

• 50% at 50% Prevalence rate

• 40% at 20% or 80% Prevalence rates

• 30% at 10% or 90% Prevalence rates

• “Small” / “Medium” / “Large” effect sizes are 0.2 / 0.5 / 0.8 standard deviations (Cohen)

• These are 6-10% / 15-25% / 24-40% for proportions in the 10-90% range (ie when prevalence rate lies between 10-90%, you call a 6-10% change as Small; 15-25% Medium, 24-40% Large)
Precision and $p$

- Precision is greatest when the proportion $p$ is near 0 or 1 and least when $p$ is near 0.5 in terms of percentages points, but NOT relative to $p$.

- Let $n$ be the sample size required for a CI of width $E$ when $p=0.25$ and of width $ap$ when $p=0.25$. $Ap$ captures relative margin of error. Defined as percentage $a$ of the Prevalence rate $p$.

<table>
<thead>
<tr>
<th>$p$</th>
<th>Obs for CI width $E$</th>
<th>Obs for CI Width $ap$</th>
</tr>
</thead>
<tbody>
<tr>
<td>.05</td>
<td>0.25n</td>
<td>6.33n</td>
</tr>
<tr>
<td>.1</td>
<td>0.48n</td>
<td>3.00n</td>
</tr>
<tr>
<td>.2</td>
<td>0.85n</td>
<td>1.33n</td>
</tr>
<tr>
<td>.25</td>
<td>1.00n</td>
<td>1.00n</td>
</tr>
<tr>
<td>.3</td>
<td>1.12n</td>
<td>0.78n</td>
</tr>
<tr>
<td>.4</td>
<td>1.28n</td>
<td>0.50n</td>
</tr>
<tr>
<td>.5</td>
<td>1.33n</td>
<td>0.33n</td>
</tr>
</tbody>
</table>
**Greater Confidence Levels are Costly**

- A 99% CI is 32% wider than a 95% CI
  
  A 90% CI is 84% as wide as a 95% CI
  
  An 80% CI is 65% as wide as a 95% CI

- To have the same width as a given 95% CI with sample size $n$ you need
  
  1.73$n$ for a 99% CI
  
  0.71$n$ for a 90% CI
  
  0.43$n$ for an 80% CI
Sample Size Calculation IHSES (2007)

- \( Z_{1-\alpha/2} = 1.96 \) (95% confidence level)
- \( P = 0.5 \) (most conservative)
- \( \text{DEFF} = 2 \)
- \( E = \text{Margin of error} \ 7.7\% \)
- Estimate that \( n = 323.96 \)

\[
n = \frac{Z_{1-\alpha/2}^2 \times P(1-P) \times \text{DEFF}}{E^2}
\]
Define ME from previous slide:

ME = Margin of error of a (1-Alpha)% Conf. Interval

\[ ME = E = Z_{1-\alpha/2} \times \sqrt{\frac{P(1-P)}{n/DEFF}} \]
This appendix reproduces the briefing materials created to introduce Kurdistan Regional Statistics Office Personnel to the principles of questionnaire design, as described in the companion report.
Designing Survey Questionnaires

Peter Glick
Kurdistan Data Project
KRSO Workshop
Erbil
February 2012
Outline

I. Survey Conceptualization – Key Issues
II. A Typology of Surveys
III. Questionnaire Design: Maximizing Usefulness and Data Quality
IV. Other Topics
I. Survey Conceptualization – Key Issues
(1) What Are The Objectives Of The Survey: What Questions Will It Try To Answer?

1. To understand the characteristics of a population
   
   *For example:*
   - Poverty rate by region
   - Unemployment by age group
   - Efficiency by firm size

2. To understand who is benefiting from policies or programs

   - Are anti-poverty programs well targeted to the poor?
   - Who uses public health services?
   - What kinds of firms get formal credit?
(1) Objectives (cont.)

3. To monitor changes over time
   - Is poverty increasing?
   - Is the unemployment rate changing and for what groups?
   - Is the vaccination rate improving?

4. To understand the *causal determinants* of outcomes
   - What factors cause malnutrition or obesity?
   - What is the impact of microcredit on enterprise growth?
   - How does schooling affect occupation and earnings?

The survey objectives will determine the type of survey, how it is designed, and its implementation.

Feasibility and costs determine this as well.
(2) What kind of survey is needed to meet the objectives?

Broad categories:
- Household surveys
- Labor force surveys
- Firm or establishment surveys
- Facilities surveys (health care, schools)
- User/client surveys

Also:
- One year or multi-year (repeated) surveys
- Panel (longitudinal) surveys
(3) What kind of survey is feasible (including, what sample size)?

Have to consider:

- Financial cost
- Human resource capacity (available expertise in sampling, survey design, implementation)
- Households or other respondents’ willingness to participate
  - In some contexts, households may be too suspicious
  - Or survey may be too long/complex
(4) What institutions will be involved—which agencies will be responsible for which aspects?

(5) Who will ultimately be the users of the data?

(5) How will the survey be organized (design, training, field-work)?

(6) How will data quality be ensured?
II. Typology Of Surveys
Key dimensions of surveys

- Purpose (poverty measurement, monitoring trends, policy analysis, focus on specific groups)
- Sampled population: households, individuals, firms, farms, schools, health facilities,
- Complexity and duration (respondent burden)
- Sample Size
- Single year or Repeated? Panel?
- Sampling design (representative or convenience sample; multistage design and weighting, etc.)

We focus here on all but sampling issues, covered separately. Discuss household surveys now, other surveys later.
Household-based surveys

- Censuses
- Multi-topic household surveys (Integrated Surveys)
- Welfare Monitoring Surveys
- Monitoring and Measurement surveys: MICS, DHS
- Household budget surveys, Income/Expenditure surveys
- Labor Force Surveys/Unemployment surveys

How do these vary along the dimensions listed earlier?

(Skip over census)
### Multi-topic household surveys (Integrated Surveys)

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Measure multiple aspects of well being; create monetary welfare measures and definition of poverty line; estimate determinants of welfare, education, health, type of work, etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examples</td>
<td>LSMS, IHSES</td>
</tr>
<tr>
<td>Complexity</td>
<td>High—measures many factors (health, employment/income, housing). Includes detailed consumption and agricultural production modules.</td>
</tr>
<tr>
<td>Survey Length/ burden</td>
<td>High</td>
</tr>
<tr>
<td>Sample Size</td>
<td>Moderate (usually 2000-5000)</td>
</tr>
<tr>
<td>Repeated</td>
<td>Not necessarily</td>
</tr>
<tr>
<td>Advantages</td>
<td>Many indicators, &amp; permits analysis of relationships—gets at the determinants of welfare (the “why”, not just the “what”).</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Smaller sample, long time to generate outputs for policymakers, complex training and design process.</td>
</tr>
</tbody>
</table>
## **Priority surveys/Welfare monitoring surveys**

<table>
<thead>
<tr>
<th><strong>Purpose</strong></th>
<th>Monitor key indicators of the well-being of different groups over time, to help target policies.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Examples</strong></td>
<td>World Bank’s Core Welfare Indicator Questionnaire (CWIQ) Survey.</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td>Low—usually focusses on several key measures such as food security, health, use of services. No consumption measure.</td>
</tr>
<tr>
<td><strong>Survey Length/burden</strong></td>
<td>Low</td>
</tr>
<tr>
<td><strong>Sample Size</strong></td>
<td>Large (e.g., 8000)</td>
</tr>
<tr>
<td><strong>Advantages</strong></td>
<td>Good for monitoring; short so allows for large sample (&amp; subgroup precision); quick to implement and to prepare data, so provides results quickly for policy makers.</td>
</tr>
<tr>
<td><strong>Disadvantages</strong></td>
<td>Limited in scope. Cannot use to analyze determinants of outcomes (answers “who” but not “why”).</td>
</tr>
</tbody>
</table>
## Monitoring & Measurement Surveys: Multiple Indicator Cluster Surveys, Demographic and Health surveys

| Purpose | Measures well being of women and children especially: MICS tracks progress to Millennium Development Goals; DHS collects detailed information on fertility and family planning, HIV/AIDS. |
| Complexity | Medium; not ‘light’ surveys. No consumption measure. Special modules for women, children. |
| Survey Length/burden | Medium-high |
| Sample Size | Large (e.g., 5000-8000) |
| Repeated | Yes |
| Advantages | Good for monitoring; large samples allow subgroup comparisons; standardized so allows international comparisons. |
| Disadvantages | More limited in scope than standard LSMS integrated survey, while more complex than lighter monitoring survey. More “what” than “why”. |
# Household Budget Surveys

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Collect information on household expenditures (and income) to produce or update the weights for consumer price indices as well as to provide inputs for national accounts.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>Low - focus on hh expenditures and sometimes income (for an Income Expenditure survey). Basic info on other things like household structure, education, employment.</td>
</tr>
<tr>
<td>Survey Length/burden</td>
<td>Low</td>
</tr>
<tr>
<td>Sample Size</td>
<td>Large</td>
</tr>
<tr>
<td>Repeated</td>
<td>Yes</td>
</tr>
<tr>
<td>Advantages</td>
<td>Good for the specific uses noted. Not difficult to administer.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Limited scope, little opportunity for analysis of determinants of consumption or income. More “what” than “why”.</td>
</tr>
</tbody>
</table>
# Labor Force/Employment Surveys

<table>
<thead>
<tr>
<th>Purpose</th>
<th>Measure and monitor employment and unemployment patterns and trends.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>low/varying - focus on labor topics such as employment, unemployment, earnings, hours of work, occupation, industry. Additional questions sometimes added on previous work experience, job search health, etc.</td>
</tr>
<tr>
<td>Survey Length/ burden</td>
<td>Low</td>
</tr>
<tr>
<td>Sample Size</td>
<td>Large</td>
</tr>
<tr>
<td>Repeated</td>
<td>Yes, and frequent (e.g., monthly or quarterly)</td>
</tr>
<tr>
<td>Advantages</td>
<td>Good for the specific uses noted. Not difficult to administer.</td>
</tr>
<tr>
<td>Disadvantages</td>
<td>Typically has limited scope, little opportunity for analysis of determinants of earnings or employment or sector. More “what” than “why”.</td>
</tr>
</tbody>
</table>
Should we collect panel data?

Panel surveys collect data from same units (households, firms, students) at different points in time. i.e., they are longitudinal data.

Advantages:
- For measuring changes over time, less sampling error (more precise estimates) because we are using the same units, which reduces variance.

- Best for understanding individual dynamics & response to changes
  - Effect of household economic shocks on labor supply
  - Effect of change in regulations on business behavior & profits
  - Effects of health and other inputs on child growth and development

- Some very important econometric benefits: can control for unobservable confounding factors (a form of omitted variables), because these are constant for the same person or household over time.
Panel data (cont.)

Disadvantages:

- Attrition: non-response gets large through cumulatively. If each year the same % drop out, eventually you can have substantial sample reduction.
- This process likely is non-random, so sample becomes non-representative.
  - Finding replacement households can restore sample size but does not solve the selectivity problem
- Costs and complexity of tracking (re-locating) respondents.
- For understanding changes in population or subgroup values for key variables, repeated cross section surveys are usually better:
  - Child nutrition
  - Use of health services
  - Household consumption and poverty

- Alternative: rolling panels combine cross-section and panel approaches. Typically half of the households from the previous survey are retained for the new survey while the other half are replaced with a new random sample.
Including a community survey with a household survey

- Often implemented in conjunction with multi-purpose household surveys.
- Gathers information on market wages and prices, local facilities (schools and health clinics), and infrastructure, employment.

Benefits:

- Price data can be used to construct spatial price deflators to adjust for regional differences in the cost of living.
- Can provide direct information on the local availability and quality of local services, which are themselves important correlates of well-being.
- Community information expands possibilities for policy-relevant analysis. For ex., permits analysis of the factors affecting the demand for health services, enrollment and attainment decisions and school choice.
- Usually not expensive to collect (survey teams are in the communities anyway).
III. Questionnaire Design: Maximizing Usefulness And Data Quality
Developing the Survey Questionnaire

We want to maximize:

1. Usefulness to policymakers, researchers, and other users.
2. Data quality

Steps in the process:

1. Initial Steps
   (form team, data analysis plan)
2. Formative Research
3. Questionnaire Design
4. Pre-testing, Pilot
5. Questionnaire revisions

Note: activities related to sampling design and sampling proceed at the same time.
1. Initial Steps

1. Assemble a survey design team including:
   o Technical experts (subject area experts, statistician, questionnaire design experts, data entry expert)
   o Policy-makers and data users
   o Fieldwork experts

2. Write a data analysis plan
   o Outlines how the data collected in a survey will be used to answer the key questions (attain the objectives); or, how will these questions be operationalized.

3. Create an advisory board
2. Don’t overlook the Formative Research

- Expert interviews
- Stakeholder interviews
- Focus groups/in depth interviews with individuals in the target population for the survey
  
  Ex. Palestine survey of youth and health uses FGs to learn:
  
  o What are the issues that are important to them (aspirations, jobs, religion, marriage, respect)
  o What is the appropriate terminology to ask about these subjects
  o What mode of interview would be most appropriate, especially on sensitive topics

- Literature review
- Review of other surveys in the region or on the same topic
3. Getting down to questionnaire design

Organization of the questionnaire:

HH surveys are built up from modules (sections) for different topics, given to different person. Ideally:

- Roster, housing → head of HH
- Education, migration, health, employment/earnings, fertility, anthropometrics → individual(s)
- Agricultural production, family enterprises, HH expenditures, credit, land/property, other income: → Best informed member

Modules on more sensitive subjects come last
- assets, loans, transfers, fertility in a HH survey
- use of illegal substances, sexual activity in a health or youth

RAND survey
3. Design (cont.)
Overall length of the questionnaire

It’s very easy to create a questionnaire that is too long but must consider the tradeoffs:

1. **Respondent burden**
   Longer questionnaire means a higher refusal rate (non-response bias) or lower quality responses due to fatigue (response bias) → **trade-off between quantity and quality**
   Use field test to try to assess this

2. **Costs**
   For a given sample size it is less expensive to do a survey with a 1 hour questionnaire than a comprehensive 3 hour questionnaire.
   → **cost/comprehensiveness trade-off**

   Or, for a given survey budget, can have a larger sample (smaller sampling errors, more precision) with a shorter questionnaire than with a more comprehensive questionnaire.
   → **sample size /comprehensiveness trade-off**
3. Design (cont.)
Formulating the questions on a topic

Is the full scope of information desired included?
Examples:

- Education: do we include not just enrollment but school type (public, private)? Subject of study? Not just highest grade attained but also reason for dropout?

- Employment: just hours worked and pay, or also ask about job benefits? Current work only, or job history? How much detail on sector of work and occupation?
3. Design (cont.)

Formulating the questions on a topic

Are the questions as consistent with standard approaches used in other surveys?

We don’t have to be slaves to previous examples but:

- Following standard approaches to modules and individual questions greatly helps place national findings in international perspective.
- Standard modules e.g., on employment reflect the long experience of experts at World Bank, ILO, Unisex, etc.
- Of course, must be sure the questions are adapted to the local context.
- At one extreme, the DHS and MICS are very highly standardized across countries.
3. Design (cont.)

Formulating the questions on a topic

Is the right format used for the question?

A. Explicit Questions
B. Pre-coded responses
C. Likert Scales
D. Open ended Questions
3. Design (cont.)

Formulating the questions on a topic

A. Explicit questions

- How old is [NAME]?
- What was the highest grade completed?
- What is the amount of your usual monthly earnings at your main job?

Each asks for specific amounts (years, grades, dollars/dinars)
Not always appropriate for the respondent—ranges sometimes better:

7. What is the amount of your monthly earnings at your main job?

CODES FOR Q. 7:

LESS THAN 40,000 CFA...............1
BETWEEN 40000 CFA AND 59000 CFA.....2
BETWEEN 60000 CFA AND 79000 CFA.....3
BETWEEN 80000 CFA AND 99000 CFA.....4
BETWEEN 100000 CFA AND 119000 CFA..5
BETWEEN 120000 CFA AND 149000 CFA..6
150000 CFA OR MORE..................7
3. Design (cont.)
Formulating the questions on a topic

B. Pre-coded responses
(Like the last example.) Sometimes the need for them is more obvious:

2. Did [NAME] ever attend school?
   Yes...1
   No...2

13. What kind of school did [NAME] last attend?
   PUBLIC.................      1
   PRIVATE NON-RELIGIOUS....2
   PRIVATE RELIGIOUS........3
   COMMUNITY SCHOOL........4
   OTHER (SPECIFY)_________5

Be careful of using the ‘Other’ response—better to include all relevant choices explicitly.
3. Design (cont.)

Formulating the questions on a topic

Pre-coded (Cont.)

Try to minimize use of the “Don’t know” option

13. What was the highest grade your father completed?

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>NONE</td>
</tr>
<tr>
<td>1</td>
<td>PRIMARY</td>
</tr>
<tr>
<td>2</td>
<td>CI</td>
</tr>
<tr>
<td>3</td>
<td>CP</td>
</tr>
<tr>
<td>4</td>
<td>CE1</td>
</tr>
<tr>
<td>5</td>
<td>CE2</td>
</tr>
<tr>
<td>6</td>
<td>CM1</td>
</tr>
<tr>
<td>7</td>
<td>CM2</td>
</tr>
<tr>
<td>8</td>
<td>LOWER SECONDARY</td>
</tr>
<tr>
<td>9</td>
<td>UPPER SECONDARY</td>
</tr>
<tr>
<td>10</td>
<td>6EME</td>
</tr>
<tr>
<td>11</td>
<td>5EME</td>
</tr>
<tr>
<td>12</td>
<td>4EME</td>
</tr>
<tr>
<td>13</td>
<td>3EME</td>
</tr>
<tr>
<td>14</td>
<td>UNIVERSITY</td>
</tr>
<tr>
<td>15</td>
<td>1ERE</td>
</tr>
<tr>
<td>16</td>
<td>2NDE</td>
</tr>
<tr>
<td>17</td>
<td>TERMINALE</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>22</td>
<td>5</td>
</tr>
<tr>
<td>23</td>
<td>6+</td>
</tr>
<tr>
<td>99</td>
<td>DON’T KNOW</td>
</tr>
</tbody>
</table>

- Many people may not know the exact years of their parents’ schooling, so you will get a lot of “Don’t knows”. We can’t use these in analysis.
- Also, the “Don’t know” option allows the interviewer to be lazy and not press the respondent.
3. Design (cont.)  
Formulating the questions on a topic

Better to formulate questions so you have the best chance of getting good responses (and train interviewers to get the respondents to think carefully)

13. What was the highest school level your father attended?

<table>
<thead>
<tr>
<th>Level</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>NONE</td>
<td>0</td>
</tr>
<tr>
<td>PRIMARY</td>
<td>1</td>
</tr>
<tr>
<td>LOWER SECONDARY</td>
<td>2</td>
</tr>
<tr>
<td>UPPER SECONDARY</td>
<td>3</td>
</tr>
<tr>
<td>UNIVERSITY</td>
<td>4</td>
</tr>
</tbody>
</table>

Or, ask this question, then ask specific years of schooling as before, allowing Don’t Know. This way you get at least some information even if respondent does not know the exact years.
3. Design (cont.)

Formulating the questions on a topic

C. Likert Scales

These are coded rating scales. Examples:

<table>
<thead>
<tr>
<th>Question</th>
<th>Likert Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. How satisfied were you with the service you received at this facility?</td>
<td></td>
</tr>
<tr>
<td>VERY SATISFIED .... .....1</td>
<td></td>
</tr>
<tr>
<td>SOMEWHAT SATISFIED .....2</td>
<td></td>
</tr>
<tr>
<td>NEITHER SATISFIED</td>
<td></td>
</tr>
<tr>
<td>OR DISSATISFIED ... ....3</td>
<td></td>
</tr>
<tr>
<td>SOMEWHAT DISSATISFIED...4</td>
<td></td>
</tr>
<tr>
<td>VERY DISSATISFIED... ...5</td>
<td></td>
</tr>
</tbody>
</table>

6. Please state whether you strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, strongly disagree with the following statements:

1. I enjoy working with others
2. I am well organized
3. I think about the future a lot

While quantitative responses are hard to evaluate because they are subjective; being “satisfied” can mean different things to different people.

Anchoring vignettes: technique to make these interpersonal comparison possible (still fairly new).
3. Design (cont.)

Formulating the questions on a topic

D. Open ended questions

All of the preceding are “closed-ended” questions. Open-ended questions are of the form:

- Please describe what you liked about this health facility
- What didn’t you like about it?
- Tell me about how you would look for a job.
- What do you think are the main barriers to getting a good job for people like you?

These are qualitative questions. They are rarely asked in large, quantitative household or other surveys of the type we are concerned with.

However, they are important for the formative research (focus groups or interviews) for surveys. The first two can inform design of a health services survey; the last two, an employment or youth survey.
### 3. Design (cont.)

Use of *skips and filtering* permit complex conditional branching

| 1. Did [NAME] have an illness or injury during the past *four* weeks? |
| 2. What kind of illness or injury did [NAME] have? |
| 3. How many days of work/school/other regular activities did [NAME] miss in the past 4 weeks due to this illness/injury? |
| 4. Does [NAME] have a chronic (long term) or permanent health condition or disability? |
| 5. What is this condition? |
| 6. How many years has [NAME] had this condition? |
| 7. Does this condition prevent [NAME] from working, being active, going to school, etc? |
| 8. CHECK AGE. IF [NAME] IS LESS THAN 15, >> NEXT PERSON |
| 9. If you had to walk 5 kilometers, could you do it easily, with difficulty, or not at all? |

In paper and pencil surveys, complicated skips may lead to interviewer errors. In computer based interviewing this problem is eliminated.
3. Design (cont.)

Many questions rely on recall over a reference period

Examples

How many hours did you work in this activity in the last seven days?
Did [NAME] have an illness or injury in the last four weeks?
In the last week, did your household consume any of the following…

Choosing the right reference period is hard:

- If too long, you risk recall errors
- If too short, you miss events that are less frequent but could be important (illnesses, purchases of consumer durables, etc.)
3. Design (cont.)

Recall is particularly an issue for collecting expenditure information.
Researchers use aggregate expenditures as a measure of welfare rather than HH income because it is usually less subject to variation (especially for agricultural HHs). But still have problems of:

- Infrequently purchased items
- Seasonality—some food items not consumed year round
- Food consumed outside the home less regular, harder to remember

Standard approach:
- Short (e.g., 1 week) ref period for frequently purchased items (bread, rice)
  - Or make two visits 2 weeks apart and ask “Since my last visit…”
- Longer (1 year) recall for infrequent purchases (including clothing, services)
- Seasonality: one year ref period is impractical. Instead, visit same HH multiple times over the year, or different HHs randomly throughout the year
## 3. Design (cont.)

### Reference period in employment surveys

- Usually ask about hours in last week (or sometimes, in a typical week for the job you worked in last week).
- But given seasonality of employment or temporary employment, want to know about past year too:

<table>
<thead>
<tr>
<th>SR. N0.</th>
<th>During the last 7 days did you work at least one hour for pay in cash or in kind, or in your own business activity or your own agricultural or livestock activity?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. YES...1</td>
<td>During the last 7 days did you do any of the following activities for at least one hour: unpaid help in a family business; unpaid work on family farm?</td>
</tr>
<tr>
<td>NO...2</td>
<td>YES...1 (&gt;&gt;8) NO...2</td>
</tr>
</tbody>
</table>

### 2. During the last 7 days did you do any of the following activities for at least one hour: unpaid help in a family business; unpaid work on family farm? |

<table>
<thead>
<tr>
<th>SR. N0.</th>
<th>During the last 12 months did you work at all for pay in cash or in kind, or in your own business activity or your own agricultural or livestock activity?</th>
</tr>
</thead>
<tbody>
<tr>
<td>26. YES...1</td>
<td>Now I'd like to ask about any work you did in the last 12 months. During the last 12 months did you work at all for pay in cash or in kind, or in your own business activity or your own agricultural or livestock activity?</td>
</tr>
<tr>
<td>NO...2</td>
<td>YES...1 (&gt;&gt;29) NO...2</td>
</tr>
</tbody>
</table>

### 27. Why did you not work? |

- ATTENDING SCHOOL...1
- HOUSEHOLD DUTIES/ CARING FOR SICK FAMILY MEMBER...2
- MATERNITY...3
- ILL/INJURED...4
- STRIKE...5
- LAID OFF/FIRED...6
- PERMANENT CLOSURE...7
- OTHER...8

The above asks 1\textsuperscript{st} about last week, then about last year for those not working in last week.

**RAND**
4. Pre-testing and Pilot Testing

These 2 steps together are referred to as Field Testing.

Pre-testing

- Involves trying out selected sections or modules on a small number of households (for example, 10-15), to see how the questions are working:
  - Do respondents understand the questions (wording correct?)
  - Are all relevant responses coded?
  - Is the sequence and flow of the questions appropriate?
  - Are the skip patterns appropriate?

- Can be done more than once during the design process, allowing for refinements to the questionnaire.
4. Pre-testing and Pilot Testing

Pilot testing

- This is a formal “dress rehearsal” for the survey.
- In addition to the issues investigated in pre-test, in the pilot we evaluate:
  - Overall duration of questionnaire/respondent burden
  - Ordering of the modules
  - Procedures for interacting with households (or firms, etc.) and local community leaders if relevant
  - All field logistics
  - Procedures for data entry if done in the field, and the data entry program

- For a household survey, ideally choose at least 100 households, in several locations (e.g., rural and urban).
- Make sure all modules (e.g., family enterprises) are tested on at least 50 HHs.
4. Pre-testing and Pilot Testing

Who does the pilot test and where does training fit in?

Two possibilities:
1. Have team supervisors and senior staff conduct the pilot.
   The steps are:
   1) Train supervisors
   2) Conduct the pilot
   3) Revise questionnaires (allow time for this!)
   4) Prepare interviewer (and supervisor) manuals
   5) Train interviewers
   6) Carry out survey

This is a good approach if the survey is complex or innovative and is likely to require significant revisions after the pilot.
(if we trained all the interviewers before the pilot and had to do major changes, significant retraining might be necessary).
4. Pre-testing and Pilot Testing

2. Or, have all survey staff, including interviewers, participate in the pilot. The steps:

1) Prepare interviewer manuals
2) Train supervisors and interviewers together
3) Conduct the pilot
4) Revise questionnaires (allow time for this!) and manuals
5) Carry out survey

Advantages:
- Interviewers are involved in pilot and in questionnaire revisions so have greater stake in the project.
- Interviewers may have insights that improve the questionnaire.
- Can select the best trainees to be the supervisors.

Disadvantages
- If survey is complex or innovative, may have to make significant revisions to questionnaire and manuals, potentially have 2nd training.
IV. Other Topics

1. Interview and Data Entry Methods
2. Other Surveys (Facility and enterprise surveys)
Interview and Data Entry Methods

They are not very related to questionnaire design but they ARE very important for data quality (minimizing non-sampling errors)

<table>
<thead>
<tr>
<th>Interview Mode</th>
<th>Data Entry Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper and Pencil (PAPI)</td>
<td>1. Data Entry in office</td>
</tr>
<tr>
<td></td>
<td>or</td>
</tr>
<tr>
<td></td>
<td>2. Data Entry in the field on a laptop</td>
</tr>
<tr>
<td>Computer-Assisted Personal Interview (CAPI)</td>
<td>3. Data Entry simultaneous with interview</td>
</tr>
</tbody>
</table>
Interview and Data Entry Methods (cont.)

Above all, data entry should be integrated with survey implementation (=linked in real time). This permits staff to return to households to correct errors while survey is still in field.

If doing PAPI and integrating:
- Option 2 means including data entry person on mobile teams.
- Option 1 means rapid transport of filled questionnaires to office for entry and problem detection—logistically complex.

If doing CAPI:
- Data entry occurs as part of computer based interview—automatically integrated.
Advantages of CAPI

- Eliminates interviewer errors related to skips, filters, sequencing—software program routes the interviewer.
- Data can be checked immediately (and corrected) via internal consistency, validity, and range checks in the software.
- Interviewer controls can be built in—interview duration, GPS location.

Potential disadvantages:
- Interviewers (and respondents) must be comfortable with computers.
- Significant up front costs for software programming.
Other Surveys: Facility or Provider Surveys (Schools, Health Care Providers)

- Like community surveys can be done with HH surveys but are of significant interest on their own
- If not linked to HH survey, usually a representative sample with the facility as the unit of observation, often stratified by sector (private, public) and size or level (e.g., clinic and hospital)

- Uses:
  1. Measure quality of services and characteristics of services:
     - Health services: staffing, absenteeism, utilization, equipment, drug shortages, financial management and oversight.
     - Schools: teacher qualifications, absenteeism, facilities, classroom supplies, teacher practices, class size etc.
  2. Compare quality and practices in public and private facilities, including pricing and incentives for employees
  3. Identify regions or types of services where greater investment is needed

RAND
Expand the scope of information on quality in health facility surveys

- They collect the standard ‘structural’ measures of quality—equipment, facilities, medicines and other supplies, staff number and qualifications.
- Also collect “process” measures of quality – that is, the care practices of practitioners in the facility, through direct observation or ‘vignettes’ approach.
- Also, information on incentives and accountability, supervision mechanisms which vary by public, private, non-governmental.

Used both to measure quality and its variation and to understand the impacts of reforms in health sector.

Funded by World Bank, standardized across countries.
Other Surveys: Enterprise/Establishment surveys

Traditional Establishment Surveys:

- Designed to be an input into compilation of national accounts and provide information on performance by sector.
- Stratified by geographical areas, industrial sector, size
  - Representation of small Informal firms often inadequate
- Collects info on expenditure, revenue, assets and liabilities, and employment, hours of work, and wages & salaries of each establishment.
- Used to calculate national output (Gross Domestic Product); also provide employment data, both by sector.
- Not designed for more in depth research on enterprise activity or problems.
Enterprise surveys

World Bank’s Enterprise Surveys (standardized, 125 countries) are more complex than traditional establishment surveys.

- Also uses stratified representative sample of private sector firms
- Covers broader range of topics:
  - Access to finance
  - Corruption
  - Competition
  - Performance
  - Business environment, regulation
- Specialized country specific modules can be added to standard ones
- Standardization permits cross country comparisons (e.g., of productivity)
- Detail permits more analytical work on determinants of firms performance

Doing Business surveys: more limited, focus on regulation & business environment.
This appendix reproduces the briefing materials created to introduce Kurdistan Regional Statistics Office Personnel to steps involved in preparing and conducting a survey, then collating the results in a form appropriate for assessment, as described in the companion report.
Data Collection and Entry

Nicholas Burger

Kurdistan Data Project
KRSO Workshop
Erbil
May 2012
Survey Process

Sampling → Questionnaire Design

Data Collection & Entry
Dr. Nicholas Burger

Data Processing
- Data cleaning & integrity
- Protection of Data
- Documentation

Data Analysis
Goals of This Session

• To review and discuss all the steps in survey data collection, from start to finish
• To provide insights from a “researcher perspective,” which isn’t always the same at the survey implementer’s perspective
• To better understand KRSO’s current practices and work to collaboratively identify areas for improvement
• To develop training tools for use with future KRSO employees
Outline

I. Survey preparation

II. Preparing to Collect Data

III. Survey Implementation

IV. Data Entry and Verification
Steps Leading Up to Starting A Survey

• Need to answer some guiding questions
  – What is the objective of the survey?
  – What type of data (collection) is needed?
  – Who will implement the survey? Are there any partners?
  – Who will use the data, when do they need it, and how to get it to them?
  – *What other issues do you consider?*
The Type of Survey One Needs Will Affect Survey Planning

• What type of data will be collected?
  – Administrative
  – Census Data
  – Surveys
    • Household
    • Establishment

• How often will data be collected?
  – Once
  – Many times
    • If more than once, should the same people be surveyed each time (panel data)?
Important Differences Between Household and Enterprise Surveys

• Households:
  – Numerous
  – Relatively easy to access
  – Often can handle longer surveys
  – Usually do not keep ‘records’, so recall is a problem

• Enterprises (firms, providers, organizations):
  – Are busy and may require multiple visits
  – Wary of releasing sensitive information
  – Keep records, aiding data accuracy

• Need to tailor the survey design and implementation accordingly
Questionnaire Design

• Observed vs. self-reported data
• Open- vs. closed-ended questions
• Uniformity and standards
• Minimizing biases
  – Respondent fatigue (within/across surveys)
  – Response bias
  – Recall bias
  – Respondents within households
• Keep in mind data entry error and costs
• Thinking ahead to the data entry and analysis stages can be very useful
Activities to Support Questionnaire Design

- Literature review
- Expert interviews
- Stakeholder interviews
- Focus groups
- Look at other surveys for insight
  - Benefits: harmonized questions can enable analysis across data sets
  - Drawbacks: someone else’s version of a question may not meet your needs
## Paper-and-Pencil (PAPI) vs. Computer-Aided (CAPI) Surveys

<table>
<thead>
<tr>
<th></th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PAPI</strong></td>
<td>• Low-cost</td>
<td>• Higher risk of data loss</td>
</tr>
<tr>
<td></td>
<td>• Easy to implement</td>
<td>• Data archiving is hard</td>
</tr>
<tr>
<td><strong>CAPI</strong></td>
<td>• Improves data quality by reducing errors</td>
<td>• High-cost</td>
</tr>
<tr>
<td></td>
<td>• Speeds data delivery</td>
<td>• May require additional training</td>
</tr>
<tr>
<td></td>
<td>• Can include automatic checks (e.g. location)</td>
<td>• May increase risk to interviewers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Hard to use in some environments</td>
</tr>
</tbody>
</table>
Look for Ways to Be Efficient with Data Collection

• May be possible to combined multiple surveys into one effort to reduce costs
  – Example: adding a community survey to collect information on community resources or characteristics
  – Example: Two ministries want to collect household survey data at about the same time

• There are limitations
  – Survey length, respondent fatigue
  – Combining types of data collection that aren’t compatible (e.g., short large-sample surveys with long small-sample surveys)
Sampling

- At an early stage determine sampling approach
- Will have implications for survey timing and affects background research needs
  - If you have access to a sample frame, need to determine sampling approach and draw samples
  - If no access to a sample frame, need to budget more time to find households (especially if sample is specialized)
Sampling for Enterprise Surveys Has Similar and Different Elements

• Sample frames may be easier to get
  – The relevant ministry may have firm registration records (e.g., Ministry of Trade and Industry, Health Ministry)
  – Lists may be out of date and inaccurate: more likely to include firms that no longer exist

• If no sample frame, then need to construct one—multiple ways to identify firms
  – “windshield” sample
  – Full enterprise/facility census

• Can be difficult to locate informal sector firms
Used sampling techniques we discussed to design the KRSO Labor Force Survey 2012

- Explicitly stratify by each of the 33 districts in the Kurdistan region
- But we will also calculate statistics for the following subgroups:
  - Governorate
  - Age groups (15-30, 31-49, 50-64)
  - Gender
  - Urban/Rural
  - Urban/Rural by governorate, age group, gender
- Two-stage design
  - Blocks as primary sampling units
  - Households within blocks as secondary sampling units
Given this sampling design how to calculate the required sample size and allocate sample across districts?

• First calculate the sample size if we had done a simple random sampling
• Then adjust for the design effect arising from clustering
• Next explore considerations involved in calculating estimates at the national or subgroup levels
  • Leads to a discussion of whether the total sample should be allocated equally across the districts or proportionate to population
• These sample size calculations are to provide guidance
  • Art as well as a science (use assumptions and previous estimates, and an iterative process)
  • Actual data collected from the survey will ultimately determine estimates and their precision
First calculate Effective Sample Size (ESS)

• ESS is the sample size needed if we did Simple Random Sampling (SRS):

\[
\frac{z_{.95}^2 \times p(1 - p)}{e^2}
\]

• We use:
  • \(p\): 42.4% for labor force participation rate and 11.9% for unemployment rate, from IHSES 2007
  • \(e\): desire 5% margin of error
  • \(z\)-statistic: 1.96 for a 95% confidence interval

• Works out to about 375 individuals (for labor force participation rate, for which larger sample needed)
  • 3.77 individuals per household who are 15 and older in IHSES 2007
  • Therefore, need 375 / 3.77, or about 100 households per district (our stratum)
Next adjust for design effect of clustering (DEFF$_c$)

- DEFF$_c$ is the ratio of variance of estimate with clustering to ratio of variance with SRS:
  \[
  \text{DEFF}_c = 1 + (m-1)*r
  \]
- We use:
  - $m$ = individuals per cluster; assuming 10 households per cluster and 3.77 individuals per household 15 and older $\rightarrow$ (about 38 individuals)
  - $r$ = ICC of 0.0268 for labor force participation rate and 0.0335 for unemployment rate
- DEFF$_c$ works out to 1.99 (for labor force participation rate)
- District level nominal or total sample size ($n$) = ESS x DEFF$_c$ = 100 x 1.99, or about 200 households
- **Exercise:** What would this sample size be if we cared only about the unemployment rate? Use average number of individuals per household in labor force of 1.60.
If we cared only about district-level outcomes...

- ...we are done
  - We will sample 200 x 33 = 6,600 households in the Kurdistan region
  - Sample size very similar to previous surveys done in the region

- But what if we want governorate or national estimates (total or by subgroups such as gender)?

- With the above equal allocation scheme we will need to weight estimates to get aggregate (say national) outcomes
  - Smaller districts (for example, Mawat) will get lower weights and larger districts (for example, Hawler center) will get higher weights
  - The differences in weights across districts will introduce another design effect that will decrease efficiency
Design effect due to weighting affects our sample allocation strategy

• This design effect is given by:

\[ DEFF_w = 1 + (CV_w)^2 \]

• Here, \( CV_w \) is the coefficient of variation (standard deviation divided by the mean) of the district weights

Final design effect is given by:

\[ DEFF = DEFF_c \times DEFF_w \]

• With equal allocation across districts, the design effect including weighting is a high value of 2.5
  • Would need 250 households per district and a total of 33 x 250, or 8,250 households to get acceptable precision at the national level
Compromise between equal allocation and proportionate allocation

• Equal allocation ensures equal precision across districts, is simple, and has been used in all previous surveys in the Kurdistan region; but decreases efficiency for national and subgroup-level estimates

• Proportionate allocation ensures better efficiency since weighting is not needed; but allocates too few households to smallest district and too many to largest

• We round up the 6,600 households from equal allocation to 7,000 households

• We adopt a compromise (a “modified Kish approach”) that allocates more households to high-population districts
  • But ensures a minimum of 160 households for the smallest districts and caps the maximum at 400 households for the largest
  • Get excellent precision at national level and acceptable precision for the smallest districts; design effect 1.58 instead of the 2.5 with equal allocation
Proposed sample allocation across districts

<table>
<thead>
<tr>
<th></th>
<th>Governorate</th>
<th>Population</th>
<th>District</th>
<th>Number of households</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Erbil</td>
<td>852329</td>
<td>Hawler center</td>
<td>400</td>
</tr>
<tr>
<td>2</td>
<td>Sulaymaniya</td>
<td>761557</td>
<td>Sulaymaniya center</td>
<td>400</td>
</tr>
<tr>
<td>3</td>
<td>DuhoK</td>
<td>323400</td>
<td>DuhoK center</td>
<td>350</td>
</tr>
<tr>
<td>4</td>
<td>DuhoK</td>
<td>237236</td>
<td>Zakho</td>
<td>280</td>
</tr>
<tr>
<td>5</td>
<td>Erbil</td>
<td>203072</td>
<td>Deshti Hawler</td>
<td>250</td>
</tr>
<tr>
<td>6</td>
<td>Sulaymaniya</td>
<td>198518</td>
<td>Raniya</td>
<td>250</td>
</tr>
<tr>
<td>7</td>
<td>Erbil</td>
<td>178319</td>
<td>Makhmur</td>
<td>240</td>
</tr>
<tr>
<td>8</td>
<td>Sulaymaniya</td>
<td>170624</td>
<td>Kalar</td>
<td>230</td>
</tr>
<tr>
<td>9</td>
<td>DuhoK</td>
<td>162058</td>
<td>Semel</td>
<td>230</td>
</tr>
<tr>
<td>10</td>
<td>Erbil</td>
<td>159969</td>
<td>Soran</td>
<td>220</td>
</tr>
<tr>
<td>11</td>
<td>DuhoK</td>
<td>152124</td>
<td>Akre</td>
<td>220</td>
</tr>
<tr>
<td>12</td>
<td>Sulaymaniya</td>
<td>146358</td>
<td>Chamchamal</td>
<td>210</td>
</tr>
<tr>
<td>13</td>
<td>DuhoK</td>
<td>145043</td>
<td>Shekhan</td>
<td>210</td>
</tr>
<tr>
<td>14</td>
<td>Erbil</td>
<td>131660</td>
<td>Shaqlawa</td>
<td>210</td>
</tr>
<tr>
<td>15</td>
<td>DuhoK</td>
<td>118841</td>
<td>Bardarash</td>
<td>200</td>
</tr>
<tr>
<td>16</td>
<td>Sulaymaniya</td>
<td>114731</td>
<td>Pishdar</td>
<td>200</td>
</tr>
<tr>
<td>17</td>
<td>DuhoK</td>
<td>95797</td>
<td>Amedi</td>
<td>190</td>
</tr>
<tr>
<td>18</td>
<td>Erbil</td>
<td>95746</td>
<td>Koye</td>
<td>190</td>
</tr>
<tr>
<td>19</td>
<td>Erbil</td>
<td>95148</td>
<td>Khabat</td>
<td>190</td>
</tr>
<tr>
<td>20</td>
<td>Sulaymaniya</td>
<td>91611</td>
<td>Halabja</td>
<td>180</td>
</tr>
<tr>
<td>21</td>
<td>Sulaymaniya</td>
<td>73010</td>
<td>Sayid Sadiq</td>
<td>180</td>
</tr>
<tr>
<td>22</td>
<td>Sulaymaniya</td>
<td>62881</td>
<td>Dukan</td>
<td>170</td>
</tr>
<tr>
<td>23</td>
<td>Sulaymaniya</td>
<td>58536</td>
<td>Sharezur</td>
<td>170</td>
</tr>
<tr>
<td>24</td>
<td>Erbil</td>
<td>52865</td>
<td>Mergasur</td>
<td>170</td>
</tr>
<tr>
<td>25</td>
<td>Sulaymaniya</td>
<td>47250</td>
<td>Kifri</td>
<td>170</td>
</tr>
<tr>
<td>26</td>
<td>Sulaymaniya</td>
<td>43297</td>
<td>Derbendikhan</td>
<td>170</td>
</tr>
<tr>
<td>27</td>
<td>Sulaymaniya</td>
<td>40475</td>
<td>Penjwin</td>
<td>160</td>
</tr>
<tr>
<td>28</td>
<td>Erbil</td>
<td>28404</td>
<td>Choman</td>
<td>160</td>
</tr>
<tr>
<td>29</td>
<td>Erbil</td>
<td>22608</td>
<td>Rawenduz</td>
<td>160</td>
</tr>
<tr>
<td>30</td>
<td>Sulaymaniya</td>
<td>18628</td>
<td>Sharbajar</td>
<td>160</td>
</tr>
<tr>
<td>31</td>
<td>Sulaymaniya</td>
<td>11967</td>
<td>Khanaqin</td>
<td>160</td>
</tr>
<tr>
<td>32</td>
<td>Sulaymaniya</td>
<td>7983</td>
<td>Qaradakh</td>
<td>160</td>
</tr>
<tr>
<td>33</td>
<td>Sulaymaniya</td>
<td>7839</td>
<td>Mawat</td>
<td>160</td>
</tr>
</tbody>
</table>

Total households: 4909884
Total population: 7000

Exercise: Compare sample allocation to a fully proportionate allocation
Ethics and Human Subjects Protection

• Survey respondents have rights or provided protections by law—also best practice

• Key issues include
  – Protecting privacy and sensitive information
  – Ensuring sensitive data are handled carefully
  – Reporting problems when they occur

• Some organizations have human subjects review committees to manage this process
  – Even if no committee exists, organizations can implement standards
How To Assess Risks and Plan Protections

• Voluntary vs non-voluntary participation
  – Ensuring valid informed consent
  – Safeguarding confidentiality

• Ethical best practices
  – Records of ethical conduct
  – Data safeguarding plan
  – Deidentifying data
  – Risk transparency

• Minimizing risks in the field
  – Only collect the information you need
  – Keep data secure at all times
Think About Documentation Early in the Survey process

• The goal should be to inform others who are using the data in the future or who are trying to do further data collection
• Planning for documentation early can avoid headaches later
  – Example: Forgetting why you chose to include or exclude a particular survey question
• Should record detailed descriptions of:
  – Questionnaire design
  – Sampling
  – Variables and their definitions
  – Other information necessary to interpret data or reproduce data collection process
Other issues

• Geographic location data
  – Will this be included?
  – Who will be responsible for collecting the data?
    • If a different/separate entity, how to coordinate on data collection and sharing?
  – What type of data will be required and how will they be used?
  – Is the necessary equipment available?
    • GPS devices are relatively expensive, but they can be shared
Outline

I. Survey preparation

II. Preparing to Collect Data

III. Survey Implementation

IV. Data Entry and Verification
Selecting the Field Work Team

• In-house data collection or a third party?
• How many enumerators do you need?
  – Number of households/firms/etc.
  – Time constraints
  – Geographic constraints
• What type of enumerators do you need?
  – Specialized skill sets or languages?
  – Local knowledge?
• What are KRSO’s previous experiences? Challenges?
Determining the Survey Timeline

• An well-thought out and realistic timeline will facilitate effective survey implementation
• The length of field work will be determined by
  – Sample size
  – Survey length
  – Field team size
  – Unexpected events
• Need to balance ambition with the tendency for surveys to take longer than expected
• Timeline should be published and revisited periodically
• What are the biggest issues for KRSO timelines?
Questionnaire Formatting and Preparation (especially paper surveys)

- A well-formatted questionnaire will dramatically improve data collection quality
- Templates can ease formatting and ensure clarity
  - Review templates periodically to improve them
- Questionnaires should be closely reviewed by two separate people before final printing or dissemination
- Format questionnaires to reduce confusion, data entry errors, and comprehension errors
Example: How to Indicate Multiple Responses

• Q3.1 Which of the following challenges does your firm face?

• (A) Enter all responses: ________________ Q 3.1
  1. Difficulty with tax registration
  2. Difficulty accessing finance or a loan
  3. Demands for informal payments from officials

• (B)
  1. Difficulty with tax registration          Y[ ]  N[ ] Q 3.1A
  2. Difficulty accessing finance or a loan   Y[ ]  N[ ] Q 3.1B
  3. Demands for informal payments            Y[ ]  N[ ] Q 3.1C
Pre- and Pilot-testing (field testing) To Improve Survey Quality at Low-cost

• Pre-testing involves trying out key parts of a survey on a small sample
  – *KRSO examples? Benefits?*

• Pilot-testing is a more formal test of the full survey to judge length, module order, field logistics, etc.
  – *KRSO examples? Benefits?*
Outline

I. Survey preparation

II. Preparing to Collect Data

III. Survey Implementation

IV. Data Entry and Verification
Field Work Management

• A critical step to effective survey implementation

• Team typically consists of primary data collection manager, one or more field managers, interviewers, and support staff

• *What is KRSO’s approach to field work management? What works well and what could be improved?*
Finding or Identifying Respondents

• Locating respondents in a pre-selected sample
  – Address/maps/GIS

• For field-selected samples, important to have a clear rule that all interviewers understand
  – E.g., start in town/village center, interview every third household
  – Determine skip patterns in advance
Multi-round survey issues

• Tracking respondents across rounds
  – Cell phone numbers
  – Neighbors

• Dealing with attrition
  – Decide whether respondents need to be replaced
  – Replace respondents with similar individuals
Human Subjects Protection in the Field

• Field work is where human subjects protection can be most challenging

• Protections be included in field work plan—all team members should know protocols and contingency plans and should be included in training
  – Helps to ensure consistency

• Basic privacy mechanisms include:
  – Not using names on field work documents except as needed (e.g., ID numbers in panel surveys)
  – Record data as soon as possible or transfer documents to KRSO offices (through secure means)
  – If electronic data collection is used, ensure data are stored on secure devices
When Human Subjects Problems Arise

• Documentation and reporting
  – All problems/violations should be documented and reported
  – Severe issues may warrant halting data collection
  – Field staff should be made to feel comfortable about reporting problems

• Respondents should be informed if their information is released
Where Human Subjects Protection Meets Reality

• There may be times when it is hard to implement strict protocols
• Not always clear where line is between acceptable deviation and violation
• For example: group interviews or interviews in public settings (e.g., China survey)
• For example: subjects may volunteer more information than asked (e.g., Mexico survey)
Steps to Ensure High-Quality Data

• Effective training to ensure all interviewers are on the same page
  – What questions mean & what the survey goals are
  – How data are to be recorded (units, standards)
  – What to do when things go wrong

• Help interviewers to learn to spot errors

• Document everything (within reason)
  – If something goes wrong, always better to know when, how, and why

• Do a 10% field sample, enter data, and quickly assess
Documenting Field Work Activities

• Have interviewers keep a log
• Where possible, develop standard protocols for reporting common issues
  – Some of these will be incorporated into the survey (e.g., desired respondent not available, interview interrupted or halted and why)
  – Others should be reported to supervisor or central system (e.g., village head denies permission to survey, an unusual event that might affect data quality)
Outline

I. Survey preparation

II. Preparing to Collect Data

III. Survey Implementation

IV. Data Entry and Verification
Data Entry Procedures for Paper-Based Surveys

• Decide on data entry process early and stick to the selected approach

• Entry can be done
  – In the field
  – In field offices
  – In central offices

• Data entry in the field reduces risk of data loss but can be more expensive less standardized
Options for Entering Data into the Computer

• Simplest data entry option is Excel, but this approach can lead to errors
• Programs like CSPro can speed data entry while reduce errors
• Key tool is automated error checking (i.e., flagging or disallowing entries that do not meet the constraints for that data point
• *KRSO’s experience with CSPro?*
Double Data Entry

• Having more than one person enter data independently helps to reduce errors
  – But it is also time-consuming/expensive

• Typical standards for double data entry include 10% or 20% of all surveys
  – Two people enter entire survey, each on their own
  – Compare results and identify discrepancies
  – Look for patterns that suggest widespread issues or errors
  – Follow up on one-off discrepancies to determine where data entry error was made
Random Household Verification

• Another approach to ensure data quality is to randomly verify data for a subset of households

• One approach:
  – Randomly select a small (N = 10) subset of households for each interviewer
  – For each household, randomly selected one or more survey modules
  – Contact the household (phone or in person) and verify data for the selected modules

• Where discrepancies are found, look for patterns across interviewers or modules

• If systematic problems, consider re-collecting problematic data for all households

• *Does KRSO do something like this currently?*
Other Data Quality Assessment Approaches

• Statistical analysis (to be covered by Alex)

• Cross-reference with other data sources
  – An option when more than one data source for the same information (e.g., survey and administrative data)

• Automated error checking (ideally built into data entry software)
Other issues and questions?
APPENDIX H

Data Cleaning Workshop

This appendix reproduces the briefing materials created to introduce Kurdistan Regional Statistics Office Personnel to steps involved in processing and organizing the data collected and providing appropriate data collection, as described in the companion report.
Overview of Data Processing

Workshop 3

Alexandria Smith
Survey Process

Sampling

Questionnaire Design

Data Collection & Entry
Nicholas Burger

Data Processing
- Data cleaning & integrity
- Protection of Data
- Documentation

Data Analysis
Introduction to Data Processing

- Proper data collection and entry is a key step in ensuring data quality
  - Topics discussed by Dr. Nicholas Burger

- Also important is correct processing of the raw data
  - Protection of our raw data sources and personal information
  - Effective cleaning practices
  - Proper organizational structure
  - Documentation
Outline of Today’s Workshop

Data Protection
Effective Cleaning Practices
Proper Directory Structure
Program Documentation
**Protection of Participants Should Occur Throughout the Research**

- US government has recognized how important it is to protect participants
  - All organizations that receive research funds review any research that includes human subjects

- Organizations can set up their own independent review: An Institutional Review Board (IRB)

- The IRB is responsible for
  - Reviewing and approving studies that involve people
  - Ensuring that people are protected during the research study.

- RAND has its own IRB.
  - All studies at RAND must be approved before research can begin

- Studies that do not follow the rules lose their funding
What is Considered Human Subjects Research

- **Interactions:** data are collected from people in person, through the phone, focus groups, even online surveys
  - Labor Force Survey
  - IHSES
  - Educational studies that observe teachers or students

- **Interventions:** data collected from people who have been assigned to different experimental groups
  - Educational studies that assign classrooms to different teaching styles
  - Health interventions that assign individuals to different treatment groups

- **Getting identifiable data any other way:** existing records
  - Medical records
  - Insurance data
  - Educational records
Institutional Review Board Process

Research project is submitted to IRB

• Determine if the research project involves
  • Interactions with human subjects
  • Interventions
  • Getting identifiable data

IRB Review

• Review and approval of
  • Survey/Research design and sampling
  • Consent Form
  • Data Protection Plan

Approval of Project

• Project can begin
The Consent Form

- Prior to interview or written survey, participants should give oral or written approval to take part in the research

- Consent form should contain the following information
  - Purpose of the study
  - What is involved
  - Risks and benefits of participating
  - Contact information for researchers
  - Reassurance that participation is completely voluntary
Data Safeguarding Plan
Methods to protect Participants

- Explains how the study will keep the confidentiality of all participants
- Level of protection depends on what personal information is collected
- Paper documentation should be locked in cabinets when not being used
- Electronic data
  - Creation of a delinked file (crosswalk) so you can remove personal information from the analytic file but you can still link it back to the raw data
  - Encrypt **ALL** data that contains identifiable information
    - Raw data files
    - Delinked file
**Data Protection Process**

1. **Paper Survey**
2. **Data Entry**
3. **Original Raw Data**
   - Name: Alex
   - Address: Main St.
   - Responses: Age 30, Employer RAND
4. **Add Random Identifier**
5. **Delinked File (cross-walk)**
6. **De-Identified Raw Data**
7. **Stored datasets - Encrypted**
8. **Files Used in Cleaning and Building Analytic File**

Random ID: 58TZ8
Responses: Age 30, Employer RAND
Introduction to Data Processing

Data Protection

Effective Cleaning Practices

Proper Directory Structure

Program Documentation
Data Cleaning and Manipulation

- Raw data, those entered directly from the survey, need to be cleaned prior to analysis.
- There are many ways to approach data cleaning and processing.
- Missing responses
- Skip patterns
- Outliers
- Inconsistent response

Everyone working with the data needs to make the same decisions.
Missing Data

- Before beginning analysis, you need to understand how much data is missing and whether it is systematic.

- Data can be missing for many reasons:
  - Respondent should answer a question but does not
  - Skip patterns
  - Error in data entry

- It is important to identify the cause of the missing data since corrections require different responses:
  - Individuals not responding to a question
    - Question or skip patterns are problematic
    - Specific groups not responding to a question
Uniformity in Cleaning Missing Data

- Decide on a uniform way to handle missing values
- Create a standardized program which you can use for every wave of data
- Standardize missing values
  - Select a code for missing [ 999 ] [.M ] [. ]
  - Never use zero
- Distinguish types of missing data. This is typically decided on prior to fielding survey.
  - Nonresponse [.M]
  - Illegible response [.I]
  - Respondent doesn’t know [.D]
Skip Patterns

- In the pilot phase, skip patterns should be carefully examined
  - Problems should be fixed prior to wider distribution
  - Problems can be caused by confusing language or interviewer training

- Understanding skip patterns key for dealing with missing data

- Decide whether to set questions out of order to missing
  - This is typically done on a case by case basis and depends on the nature of the question
**Example: Educational Attainment**

Will use the following three questions from the IHSES 2007 data to discuss missing data and skip patterns

1. **Question 402**
   - Have you ever attended School
     - (1) Yes, Attended in Past [Skip to 406]
     - (2) Yes, I am currently Attending [Skip to 407]
     - (3) No [403]

2. **Question 403**
   - Reason for not Attending
     - (1-16) reasons

3. **Question 406**
   - Highest Certificate attained
     - (1-13) Certificate Level

4. **Questions 407**
   - Number of years in school
     - (1-24) Years in school
Important to understand initial distribution

- Only individuals age 6+ are supposed to respond to the question

<table>
<thead>
<tr>
<th>q0103(age)</th>
<th>q0402(ever attended school)</th>
<th>Missing</th>
<th>0</th>
<th>1-Yes, Attended in past</th>
<th>2-Yes, Currently Attending</th>
<th>3-No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>under 6</td>
<td></td>
<td>2900</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2904</td>
</tr>
<tr>
<td></td>
<td></td>
<td>99.86</td>
<td>0</td>
<td>0.03</td>
<td>0.07</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>6-50</td>
<td></td>
<td>23</td>
<td>1</td>
<td>5143</td>
<td>5574</td>
<td>2378</td>
<td>13119</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.18</td>
<td>0.01</td>
<td>39.2</td>
<td>42.49</td>
<td>18.13</td>
<td></td>
</tr>
<tr>
<td>50+</td>
<td></td>
<td>4</td>
<td>0</td>
<td>478</td>
<td>5</td>
<td>1445</td>
<td>1932</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.21</td>
<td>0</td>
<td>24.74</td>
<td>0.26</td>
<td>74.79</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>2927</td>
<td>1</td>
<td>5622</td>
<td>5581</td>
<td>3824</td>
<td>17955</td>
</tr>
</tbody>
</table>

- Use age as the base: individuals under 6 are not included in the educational analysis

- Decide to keep or remove invalid responses

- Missing responses are <1% for individuals who should have responded to q0402 so this is not a serious problem

RAND
**Reason for Not Attending School**

<table>
<thead>
<tr>
<th>q0403(reason to never attend)</th>
<th>q0402(ever attended school)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.Missing</td>
</tr>
<tr>
<td>.Missing</td>
<td>27</td>
</tr>
<tr>
<td>1 There is no school</td>
<td>0</td>
</tr>
<tr>
<td>2 There are no seats at school</td>
<td>0</td>
</tr>
<tr>
<td>3 There are no sanitary utilities at school</td>
<td>0</td>
</tr>
<tr>
<td>4 Transport is very difficult and unsafe</td>
<td>0</td>
</tr>
<tr>
<td>5 There is no female teacher</td>
<td>0</td>
</tr>
<tr>
<td>6 There is absolutely no teacher at all</td>
<td>0</td>
</tr>
<tr>
<td>7 Household cannot afford school expenses</td>
<td>0</td>
</tr>
<tr>
<td>8 Work for the family</td>
<td>0</td>
</tr>
<tr>
<td>9 Work for someone else</td>
<td>0</td>
</tr>
<tr>
<td>10 Disability or disease</td>
<td>0</td>
</tr>
<tr>
<td>11 Marriage</td>
<td>0</td>
</tr>
<tr>
<td>12 Family not interested</td>
<td>0</td>
</tr>
<tr>
<td>13 [NAME] not interested</td>
<td>0</td>
</tr>
<tr>
<td>14 There are no documents</td>
<td>0</td>
</tr>
<tr>
<td>15 Social reasons</td>
<td>0</td>
</tr>
<tr>
<td>16 Other</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
</tr>
</tbody>
</table>

RAND
# Highest Diploma Attained

<table>
<thead>
<tr>
<th>Highest Diploma Attained</th>
<th>q0402 (ever attended school)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.Missing</td>
</tr>
<tr>
<td>.Missing</td>
<td>26</td>
</tr>
<tr>
<td>No Certificate</td>
<td>0</td>
</tr>
<tr>
<td>Elementary</td>
<td>1</td>
</tr>
<tr>
<td>Intermediate</td>
<td>0</td>
</tr>
<tr>
<td>Preparatory</td>
<td>0</td>
</tr>
<tr>
<td>Vocational</td>
<td>0</td>
</tr>
<tr>
<td>Diploma from an institution</td>
<td>0</td>
</tr>
<tr>
<td>Bachelor Degree or Higher</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
</tr>
</tbody>
</table>

- Missing responses <1% for q0406
# Years of Schooling

All individuals who attended school respond to 407

<table>
<thead>
<tr>
<th>q0407_a(years of schooling)</th>
<th>q0402(ever attended school)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>.Missing</td>
</tr>
<tr>
<td>.Missing</td>
<td>24</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1-24yrs of schooling</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>27</td>
</tr>
</tbody>
</table>

- Missing responses <1% for q0407
- Zero years of schooling could indicate an incorrect value or less than 1 year of school
Error could be due to data collection

28% of the data is missing for values on literacy

<table>
<thead>
<tr>
<th>Scores</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing</td>
<td>305</td>
<td>28%</td>
</tr>
<tr>
<td>Low</td>
<td>208</td>
<td>19%</td>
</tr>
<tr>
<td>Average</td>
<td>160</td>
<td>15%</td>
</tr>
<tr>
<td>High</td>
<td>400</td>
<td>37%</td>
</tr>
<tr>
<td>Total</td>
<td>1073</td>
<td>100%</td>
</tr>
</tbody>
</table>

By location, you can see that site 4 has a higher level of missing.

<table>
<thead>
<tr>
<th>Location</th>
<th>. Missing</th>
<th>Low</th>
<th>Average</th>
<th>High</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>85</td>
<td>97</td>
<td>56</td>
<td>77</td>
<td>315</td>
</tr>
<tr>
<td>Site 2</td>
<td>93</td>
<td>56</td>
<td>51</td>
<td>172</td>
<td>372</td>
</tr>
<tr>
<td>Site 3</td>
<td>61</td>
<td>39</td>
<td>35</td>
<td>63</td>
<td>198</td>
</tr>
<tr>
<td>Site 4</td>
<td>81</td>
<td>16</td>
<td>18</td>
<td>73</td>
<td>188</td>
</tr>
<tr>
<td>Total</td>
<td>305</td>
<td>208</td>
<td>160</td>
<td>400</td>
<td>1073</td>
</tr>
</tbody>
</table>

If data collection is still in process, you can investigate data collection in field
Outliers

- Two basic types of outliers
  - Invalid Responses
    - Age > 200
  - Possible, yet extreme values
    - Earnings

- Impossible outliers can be set to missing

- Extreme outliers can be dealt with in the analysis
  - These are possible true values and provide a data point for an accurate snapshot of your population
Examples of possible, yet extreme values

The oldest individual in the IHSES survey is 103 years of age

<table>
<thead>
<tr>
<th>Age at interview</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>17955</td>
</tr>
</tbody>
</table>

For wage jobs in the IHSES, extremely high salary above the 99th percentile

<table>
<thead>
<tr>
<th>Biweekly Salary (1000 Dinar)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
</tr>
<tr>
<td>----</td>
</tr>
<tr>
<td>3275</td>
</tr>
</tbody>
</table>
Inconsistent Response

- Inconsistency in data include
  - Basic discrepancies
  - Differences in participant response across waves
  - Errors in questions due to incorrect skip logic

- Specific checks will need to be conducted on waves of data.

- In the Labor Force Survey the following questions should stay the same
  - Gender
  - Relation to head of household

- Check for appropriate changes
  - Individuals can gain education but not lose education
  - Age should increment by no more than the elapsed time between surveys
Outline of Today’s Workshop

Data Protection

Effective Cleaning Practices

Proper Directory Structure

Program Documentation
Data warehouse design

- File structure and layout key for organization
- There are a number of ways to create a directory structure

<table>
<thead>
<tr>
<th>Folder Name</th>
<th>Type of Data Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw</td>
<td>Raw data sets</td>
</tr>
<tr>
<td>Analytic Files</td>
<td>Produced analytic data files</td>
</tr>
<tr>
<td>Programs</td>
<td>Analysis Programs</td>
</tr>
<tr>
<td>Catalogs</td>
<td>Format and Macro Libraries</td>
</tr>
<tr>
<td>Documents</td>
<td>Program and Project Documentation</td>
</tr>
<tr>
<td>Output</td>
<td>Tables, Frequencies, Lists</td>
</tr>
</tbody>
</table>
Data warehouse design

Inside the primary folder LaborForceSurvey

Analytic Files
Catalogs
Documents
Output
Programs
Raw
Code detailing data warehouse design

- Code is included in one program
  - Inc.sas
  - Inc.do

- Reference the file at the top of each program

```sas
%let basedir=C:\DATA\LaborForceSurvey;
libname raw "&basedir\rawdata";
libname sasdata "&basedir\analyticfiles";
libname output "&basedir\output";
```
Outline of Today’s Workshop

Data Protection

Effective Cleaning Practices

Proper Directory Structure

Program Documentation
The Truth about Documentation

- No one likes to document and it is tedious
  - Documentation tends to get put off to the end
  - This is particularly problematic for multi-year projects

- It is invaluable in the long run and necessary for reproducibility

- Documentation should include
  - External documentation
    - Documentation of the overall process
    - Should have some standardization across the group
  - Internal program documentation
**External Program Documentation**

- External documentation consists of several documents
  - Data Dictionary with variable description
  - Readme file – description of programs used to build the analytic file
  - Description of Sampling
  - Description of field procedures
  - Documentation of changes made to the data

Store documentation in a central location

Documentation should allow any individual to recreate the analysis file
Internal Program Documentation

- There are three main ways to internally document a program
  - Imbedded Comments
  - Self Documenting Code (Logical Code)
  - Layout

- Use code to self document a program
  - If code is “tricky or complicated” then its not easy to follow
  - The purpose should be clear even before adding comments

- Layout of a program is important
  - Use of formatting (indentation or spacing) to make readable

- Warning: Commenting can be overdone; balance is needed
Good internal documentation starts with a description of the program

Example of a program header

/***********************************************************/

Program name: 
Author:
Date created : 
Project : 
Purpose: 
Inputs: 
Outputs : 
Updated by: 
Notes: 
/***********************************************************/
Imbedded Comments

- Should be strategically placed inside a program
  - They explain the intent
  - Should be readable, short, and to the point

```c
if xgov in (11,13,15);
```

Too little text

```c
/*
Only keep observations who are over the age of six.
When looking at the literacy levels, there are a large number of observations
who are under the age of 6 but have a high level of literacy. These observations are
skewing the overall rates
*/
if Q0103 < 6 then delete;
```

Way too much text
Imbedded Comments

- Straight forward comments

```c
/*
  keep gov Sulaimaniya/Erbil/Dohuk
*/
if xgov in (11,13,15);

/*
  drop under age 6 skews results for language
*/
if Q0103 <6 then delete;
```

- Some extra detail helps when working with different individuals

- This is readable to users of SAS, STATA, SPSS
Example of Poorly Written Code

%let tablname = Adherencia| AutoEvalucion | ComposicionFamiliar | ConocimientosNutricionales |
CONOCIMIENTO |Depresion | Dieta | EstigmaDiscrimi |EvaluacionNutrici | FichaResumen | Maestro |
Percepcion | EficaciPercepciones | Seguridad |SituSocioEcono | UsoAlcohol |ViviendaServicio ;
%macro macroname;;
%LET ntab = 1;
%DO %WHILE(%scan(%STR(&tablname), &ntab, |) ^= );
%LET ntab = %EVAL(&ntab + 1);
%END;
%LET ntables = %EVAL(&ntab - 1);
%PUT There are &ntables Access databases.;
%Do i=1 %To &ntables ;
%let dt=%upcase(%scan(&tablname,&i,'|'));
PROC IMPORT OUT= WORK.&dt
DATATABLE= "&dt"
DBMS=ACCESS REPLACE;
DATABASE=&dblocation;
SCANMEMO=YES;
USEDATE=YES;
SCANTIME=YES;
RUN;
.....
%mend ImportAccess;
Code Written Clearer

%let tablename = Adherencia | AutoEvalucion | ComposicionFamiliar | ConocimientosNutricionales | CONOCIMIENTO | Depresion | Dieta | EstigmaDiscrimi | EvaluacionNutrici | FichaResumen | Maestro | Percepcion | EficaciPercepciones | Seguridad | SituSocioEcono | UsoAlcohol | ViviendaServicio ;

- Comment explains the purpose of the list
- Use alignment to clearly see each of the dataset names

```plaintext
/*
 name of tables to import
*/
%let tablename = Adherencia
AutoEvalucion
ComposicionFamiliar
ConocimientosNutricionales
Depresion
Dieta
EficaciPercepciones
EstigmaDiscrimi
EvaluacionNutrici
FichaResumen
Maestro
Percepcion
Seguridad
SituSocioEcono
UsoAlcohol
VIHCONOCIMIENTO
ViviendaServicio ;
```
Example of the code written more clearly

```sas
/* creates an index counting number of tables */
%LET ntab = 1;

%DC %WHILE(%scan(%STR(&tablename), &ntab, |) ^= );
   %LET ntab = %EVAL(&ntab + 1);
%END;

/* set variable equals total number of tables */
%LET ntables = %EVAL(&ntab - 1);

/* transform each access table to dataset */
%Do i=1 %To &ntables ;
   %let dt=%upcase(%scan(&tablename,&i,'|'));
   PROC IMPORT OUT= WORK.&dt
   DATATABLE= "&dt"
   DBMS=ACCESS REPLACE;
   DATABASE=&dblocation;
   SCANMEMO=YES;
   USEDATE=YES;
   SCANTIME=YES;
RUN;

.......
%mend ImportAccess
```
Data protection depends on the level of sensitivity

<table>
<thead>
<tr>
<th>Sensitivity Level</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reportable Private</td>
<td>Information that if lost or compromised would require reporting the loss to the data subjects</td>
<td>☐ Government number</td>
</tr>
<tr>
<td></td>
<td></td>
<td>☐ Driver's license number</td>
</tr>
<tr>
<td></td>
<td></td>
<td>☐ Protected health information</td>
</tr>
<tr>
<td>Sensitive</td>
<td>Information if lost does not need to be reported but could cause personal damage</td>
<td>☐ Survey respondents' names and minor medical conditions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>☐ Roster information (such as student names and grade level)</td>
</tr>
<tr>
<td>Public</td>
<td>Non-individually identified data that has been publicly released</td>
<td>☐ Publicly released datasets</td>
</tr>
</tbody>
</table>
Example: The Labor Force Survey

- There is no reportable private information, but there is sensitive information

<table>
<thead>
<tr>
<th>Sensitivity Level</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitive</td>
<td>NAME</td>
</tr>
<tr>
<td></td>
<td>ADDRESS</td>
</tr>
<tr>
<td></td>
<td>EDUCATION</td>
</tr>
<tr>
<td></td>
<td>EMPLOYMENT</td>
</tr>
</tbody>
</table>

Good Rule: It never hurts to have more protection than you need
This appendix reproduces the briefing materials created to introduce Kurdistan Regional Statistics Office Personnel to the principles of the analysis of labor market indicators.
Analysis of Labor Market Indicators

Workshop

July 30, 2012
Overview

- **OBJECTIVE:** How to calculate indicators from the survey data using examples from the Labor Force Survey

1. How to choose indicators?

2. Steps to calculate indicators: from concept to software commands

3. Examples:
   A. Total unemployment
   B. Total employment
   C. Labor force participation
   D. Informal employment
   E. And others
Choosing Indicators

• 18 key labor market indicators are described by the International Labour Organization (ILO):

• Key indicators of the labour market (KILM)

• Similarly, there are other international guides for such topics as:
  – Education: unicef, OECD
  – Other: United Nations
Choosing Indicators (continued)

<table>
<thead>
<tr>
<th>18 KILM INDICATORS</th>
</tr>
</thead>
<tbody>
<tr>
<td>KILM 1: Labor force participation rate</td>
</tr>
<tr>
<td>KILM 2: Employment-to-population ratio</td>
</tr>
<tr>
<td>KILM 3: Status in employment</td>
</tr>
<tr>
<td>KILM 4: Employment by sector</td>
</tr>
<tr>
<td>KILM 5: Employment by occupation</td>
</tr>
<tr>
<td>KILM 6: Part-time workers</td>
</tr>
<tr>
<td>KILM 7: Hours of work</td>
</tr>
<tr>
<td>KILM 8: Employment in the informal economy</td>
</tr>
<tr>
<td>KILM 9: Unemployment</td>
</tr>
<tr>
<td>KILM 10: Youth unemployment</td>
</tr>
<tr>
<td>KILM 11: Long-term unemployment</td>
</tr>
<tr>
<td>KILM 12: Time-related underemployment</td>
</tr>
<tr>
<td>KILM 13: Inactivity</td>
</tr>
<tr>
<td>KILM 14: Educational attainment and illiteracy</td>
</tr>
<tr>
<td>KILM 15: Average monthly wages</td>
</tr>
<tr>
<td>KILM 16: Hourly compensation costs</td>
</tr>
<tr>
<td>KILM 17: Labour productivity</td>
</tr>
<tr>
<td>KILM 18: Poverty, income distribution and the working poor</td>
</tr>
</tbody>
</table>

Choose indicators based on these guidelines but:
- Works only for indicators for which the survey (questionnaire and sample) was already designed for
Choosing Indicators (continued)

• Focus of this presentation is on the most important concepts and indicators:
  – EMPLOYMENT
    • KILM employment to population ratio
  – UNEMPLOYMENT
    • KILM youth unemployment
  – LABOR FORCE PARTICIPATION
    • Labor force participation rate
Overview: from Concept to Number

Choose
- **Choose**: concept and indicator

Define
- **Define**: what is and what is not to be counted in that definition

Implement
- **Implement**: definition according to questionnaire items
  - Which set of responses determines an outcome

Code
- **Code**: program that translates variables in dataset (responses from questionnaire) into indicators

Present
- **Present** results: create tables and describe main findings for report
  - Keep in mind why an indicator was chosen
Example 1: Employment

1. Choose:

Why calculate employment?
- Important for indicators such as employment to population ratio (KILM 2)
- Other reasons?
Example 1: Employment (continued)

2. Define:

Use “standard” definitions from international guidelines as much as possible:

(1) The "employed" comprise all persons above a specific age who during a specified brief period, either one week or one day, were in the following categories:
   (a) "paid employment":
      (a1) "at work": persons who during the reference period performed some work for wage or salary, in cash or in kind;
      (a2) "with a job but not at work": persons who, having already worked in their present job, were temporarily not at work during the reference period and had a formal attachment to their job. This formal job attachment should be determined in the light of national circumstances, according to one or more of the following criteria:
         (i) the continued receipt of wage or salary;
         (ii) an assurance of return to work following the end of the contingency, or an agreement as to the date of return;
         (iii) the elapsed duration of absence from the job which, wherever relevant, may be that duration for which workers can receive compensation benefits without obligations to accept other jobs.
   (b) "self-employment":
      (b1) "at work": persons who during the reference period performed some work for profit or family gain, in cash or in kind;
      (b2) "with an enterprise but not at work": persons with an enterprise, which may be a business enterprise, a farm or a service undertaking, who were temporarily not at work during the reference period for any specific reason.
Example 1: Employment (continued)

2. Define (continued):

It is necessary to adapt and make decisions, for example:

(1) The "employed" comprise all persons above a specific age who during a specified brief period, either one week or one day, were in the following categories:
   (b) "self-employment":
      (b1) "at work": persons who during the reference period performed some work for profit or family gain, in cash or in kind;

(2) For operational purposes, the notion "some work" may be interpreted as work for at least one hour

In some cases, these decisions will have been implicitly taken since questionnaire was drafted
Example 1: Employment (continued)

3. Implementing the definition:

paid-employed: use D1 or a combination of D4, D6 and D7 to determine if individual is paid-employed

self-employed: use D2 and D3 to determine if individual is self-employed

above a certain age: use Q. B4-A (or age variable in cleaned dataset) and discard observations where age is less than 15
Example 1: Employment (continued)

4. Coding the definition

**paid-employed:** use D1 or a combination of D4, D6 and D7 to determine if individual is paid-employed

if D1=1 employed=1
if D4=1 and D6=1 employed=1.
if D4=1 and (D7=1 or D7=2).

**self-employed:** use D2 and D3 to determine if individual is self-employed

if (D2=1 or D3=1) employed=1

**above a certain age:** use Q. B4-A and discard observations where age is less than 15

if age<15 employed=$SYSMIS.
Example 2: Unemployment

1. Why is unemployment important?
Example 2: Unemployment (continued)

1. Choose indicator:

E.g., Total number of unemployed individuals

2. State definition:

“An unemployed person is someone who is of working age, is not employed but was willing and looking for work”
Example 2: Unemployment (continued)

3. Implement the definition:

**working age**: use Q. B4-A and discard any observations where age is:
- less than 15;
- missing

**is not employed**: use questions in section (declare *not unemployed* if:
- either responded Yes to questions D1-D3
- said Yes to D4 and D6
- said Yes to D4 and responded code 1 or 2 to D7
Example 2: Unemployment (continued)

3. Implement the definition (continued):

- willing and able to work use Q. E1 (was ... willing and able to work?)

- looking for work use Q.E2 (.....looked for work?)
**Example 2: Unemployment (continued)**

4. **Code:**

   **working age:** use Q. B4-A (or age variable) and discard observations where age is:
   a) less than 15; or c) missing
   
   ```
   compute unemployed=1.
   if missing(age) unemployed=$SYSMIS
   if age<15 unemployed=$SYSMIS
   ```

   **is not employed:** use questions in section D. (declare not unemployed if:
   a) either responded yes to either of D1-D3
   b) said yes to D4 and D6
   c) said yes to D4 and responded code 1 or 2 to D7
   
   ```
   if (D1=1 or D2=1 or D3=1) unemployed=0.
   if (D4=1 and D6=1) unemployed=0.
   if D4=1 and (D7=1 or D7=2) unemployed=0
   ```
Example 2: Unemployment (continued)

4. Implement the definition (continued):

willing and able to work use Q. E1 (was … willing and able to work?)
if E1=2 unemployed=0.

looking for work use Q.E2 (…..looked for work?
if E2=2 unemployed=0.
Example 3: Youth Unemployment

Choose
- KILM 10, *Youth unemployment*
- Discuss: why is youth unemployment particularly important?

Define
- A respondent falls in this category if he/she is *aged 15 to 24* and is *unemployed* (see definition above)

Implement
- Use the “unemployment” category created above and “age”

Code
- Compute youth_unemployed=$SYSMIS
- if (age>=15 and age<=24 and unemployed=1) youth_unemployed=1.
Example 4: Labor Force Participation

Choose
• Whether an individual works or looks for work
• Discuss: why is labor force participation important?

Define
• An *individual* participates in the labor force if he/she is either *employed* or *unemployed*

Implement
• Use previously created definitions of *unemployed* and *employed*

Code
• compute laborforcepart=0.
• if (unemployed=1 or employed=1) laborforcepart=1.

Present

---

Code:
```python
laborforcepart = 0.
if unemployed == 1 or employed == 1:
    laborforcepart = 1.
```
This appendix reproduces the briefing materials created to introduce Kurdistan Regional Statistics Office Personnel to methods of working with survey data, including statistical analysis methods.
Working with Survey Data

Peter Glick
Francisco Perez Arce

Kurdistan Data Project
KRSO Workshop
Erbil
July 2012
Outline

I. Implications of Complex Surveys: Weighting and variances

II. Statistical Analysis Methods
II. Implications of Complex Surveys:

Weighting and variances
Complex Survey Design in the KRI Employment Survey

1. Stratification

Rather than a simple random sample, we group population into *Districts* (*strata*) to ensure adequate representation of each, then sample within each district.

Further, we use disproportionate sampling across the Districts:

- *We oversample* (select more clusters from) smaller districts.
- This allocation is based on Kish rule (1988).
Complex Survey Design (continued)

2. Two-stage (cluster) design

Within each district we:

• Randomly select clusters with probability proportional to size (PPS) (1st stage)

• Randomly select 10 households per cluster for interview (2nd stage)
More on stratification

• In this survey, our stratum is District \((n=33)\)

• If the percentage of the population selected in each District is the same in all districts this is a proportionate stratification
  – Produces a self-weighting sample
  – Design weights are not needed

• Usually improves precision as compared to an SRS
Instead we use
Disproportionate Stratification

• Total sample (7,000 households) is not allocated to Districts proportionate to pop.

• Use modified Kish rule: small districts oversampled to ensure precision (and large ones undersampled)

• Means the sample does not reflect the true distribution of the population across districts—must use weights to achieve this
More on clustering

As noted, within each district we randomly select clusters with probability proportional to size (PPS) and choose the same number of hhs from each.

Means each hh in the district has the same probability of selection, as in a simple random sample.

This produces a self-weighting sample within the district. Hence clustering (with PPS) does not create the need for sampling weights.
Implications of this Complex Survey Design for the Analysis

1. Weighting
   – We need to create the correct sample weights so that the sample is representative of the population (Design weights)

2. Variances of estimates:
   – Are affected by the survey design (both from disproportionate sampling from strata and clustering)
   – Need to be corrected to understand the true precision of our estimates, and to do statistical tests correctly.
Weighting-1

Design Weights (DW)

Complex survey design leads to use of weights. (there are other reasons for weighting, discussed next)

- Design weights correct for differing probabilities of selection of population members into the sample as result of the survey design

- Used to adjust for disproportionate stratified random sampling (e.g., we sample 1% of pop. of one district but 2% of another district)

General concept: weight for each observation = inverse of its sampling probability (P). So weight = 1/P
Weighting-2

Non-response weights (NRW)

After the fieldwork, we need to correct for differing probabilities of participation among those we attempt to contact. Example:

• In district 1, 94% of households responded (i.e., 6% were not found or refused to be interviewed)
• In district 2, only 90% responded
• If we do nothing, then district 1 is over represented, 2 is underrepresented relative to their true population shares
• So adjust design weights for district 1 by factor of 1/.94 and adjust design weights of district 2 by factor of 1/.90 (i.e., 1/P again)
Creating Weights-1

In a typical survey with non-response, both design and non-response weights are involved.

- **Design weights** reflect the probability of selection from the population into the subset we attempt to contact ($p_{sel}$):
  \[ DW = 1 / p_{sel} \]

- **Non-response weights** reflect the estimated probability of people like the respondent responding, given that we attempted to contact them ($p_{res}$):
  \[ NRW = 1 / p_{res} \]
Creating weights -2

Step 1: \( P_{sel} \)

\[ P_1 = \frac{\text{Total \# HH in sample}}{\text{Total \# HH in KRI Population}} \] (overall sampling rate for survey)

\[ P_2 = \frac{\text{# HH sampled in District } i}{\text{Total \# HH in District } i} \] (sampling rate for district \( i \))

Therefore the probability of selection \( P_{sel} = P_2/P_1 \)

It is \( >1 \) if District \( i \) is oversampled

\( <1 \) if District \( i \) is undersampled

Design weight (DW) = \( 1/P_{sel} \)
Creating weights-3

Step 2: $P_{\text{res}}$

Probability of responding in cluster j

$= \frac{\# \text{ HH interviewed in Cluster } j}{\# \text{ HH in sample in Cluster } j}$

For example, if 9 of the 10 selected households in the cluster agree to the interview, the probability of response $P_{\text{res}} = .9$

Non-response weight (NRW) = $1/P_{\text{res}}$
Step 3: Combine to create overall weights

- Overall weights (OW) reflect the probability of selection from the population into sample of respondents \((p_{\text{sel}})(p_{\text{res}})\)

- \(OW = 1/((p_{\text{sel}})(p_{\text{res}})) = DW*NRW\)

- Overall weights can be estimated directly as the inverse of the fraction of population members who respond within each stratum
Next: Implications of complex survey design for variances

- Design effect (DEFF) of a complex design is the ratio:

\[
\frac{\text{Variance of the estimate obtained via the complex design}}{\text{Variance of the estimate obtained via a Simple Random Sample with the same sample size}}
\]

- DEFF may come from both aspects of complex design:
  1. Stratification (via the weighting)
     - Effect applies to all outcomes equally
  2. Clustering
     - Effects all outcomes, but unequally
Design Effect of Stratification/Weighting

• Weighting may correct bias, but generally at a price—increased variance of the estimates

• For disproportionately stratified designs the design effect is

\[
DEFF = 1 + \text{Var}(\text{weights}).
\]

(If the weights are standardized to have mean 1)

• Therefore the more variable are the weights (the more disproportionate the sampling across strata), the higher the variance relative to no weighting

• For ex., if you want to sample the same # households in very small and large districts, you are sampling very disproportionately —result is a large DEFF
Design Effect of Clustering

For given overall sample size, the design effect from clustering is larger:

• The larger the cluster (#hhs), or equivalently given sample size, the smaller the number of clusters

• The greater the degree of similarity of households or individuals within a cluster, as measured by the *intra-class correlation coefficient* = ICC
  - ICC is the proportion of variance of individual outcomes attributable to clusters
How do we calculate DEFF from these two sources?

• Possible to do ‘by hand’. Calculate DEFF from stratification/weighting and from clustering

• Fortunately, this complex adjustment is made automatically by SPSS or other software programs using the appropriate commands
  – SPSS ‘complex samples’ module
  – STATA ‘Survey’ commands

• We still need to calculate the weights ourselves
II. Statistical Analysis Methods

1. Confidence intervals

2. Significance tests

3. Correlation
Types of Analysis

“Descriptive” statistics: presentation of means, totals, proportions of individual variables:
   - Ex: unemployment rate, share of informal employment
   - The main focus of reports

“Analytic” statistics: examines relationships among variables
   - 2-way tables: unemployment by governorate, or by rural/urban (also typically in reports, and usually also called ‘descriptive’)
   - Correlation analysis – how one variable changes with another
   - Regression analysis – how one variable changes with another, holding other variables constant.

Since we are dealing with samples, all of the above involves variances, and also sampling weights and design effects.
1. **Confidence Intervals (CIs)**

For descriptive statistics

- CIs indicate the **precision** of an estimate
  - For example, if the calculated unemployment rate is 12%, how reliable is this estimate?

- Since this is an estimate from a sample, not a population, there is always uncertainty about it—and the smaller the sample, the more uncertainty

- This is very important, BUT most survey reports do not report CIs (or margin of error). They should report them.
**Confidence Intervals**

Cls have 3 elements:

1. Confidence level
2. Sample statistic or estimate
3. Margin of error (ME)

Say the statistic is the unemployment rate estimated from the survey (10%), the confidence level is 95%, and the margin of error is 2%

The CI is the interval defined by:

\[
\text{Estimate } \pm \text{ margin of error}
\]

Or 10% plus or minus 2%, that is, 8% to 12%

This means, we can say with 95% probability that the TRUE unemployment rate is between 8% and 12%.
The mean is our estimate (10%). Given the sampling distribution we can say there is a 95% probability the true value falls within the interval 8.0%, 12.0%. (and 2.5% probability it is above 12%, 2.5% probability it is below 8%)

For a normal distribution, 2.5% of the sample lies above 1.96 *standard deviations* above the mean, and 2.5% below. So the limits are +/- 2 s.d. from the mean →

RAND
Confidence Intervals

Therefore the 95% confidence interval is:

\[ \text{Estimate} \pm 1.96 \text{s.d.} \]

Where s.d. = standard error of the estimate

Recall s.d. = square root of variance; in practice we replace it with the estimate of the s.d., the standard error.

So margin of error (ME) = 1.96 s.d. (or about 2 s.d.)

The bigger the s.d. or variance, the bigger the ME/the wider the confidence interval. And the less precise the estimate.

And the smaller the sample, the larger the variance (or s.d.):

• Hence small sample means larger ME

• Which is why we are so concerned with sample size when designing a survey.
Confidence Intervals

So to get the CI, we just need the estimate (mean unemployment) and the standard error of the estimate.

SPSS and other software automatically compute these (and the CI too).

BUT it’s not so simple with complex survey designs

Recall from the discussion of sampling design, several aspects of the sampling affect the variance of the estimates:
Confidence Intervals

• The *two-stage design* involves clustering. This increases the variance and standard error relative to a simple random sample.

• *Stratification* of the sample (by district) is efficient: it reduces the variance

• But *disproportionate* stratification (we over sample smaller districts, undersample large ones) requires the use of differential weights. This increases the variance
Confidence Intervals

So if \( s.d. \) represents the standard error from a simple random sample, the CI we need is actually:

\[
\text{Estimate } + / - \ 1.96 \times \text{deft} \times s.d.
\]

The multiplier \( \text{deft} \) is the ‘design factor’. It adjusts the standard deviation for the effects of all the design features above.

Fortunately, this complex adjustment is made automatically by SPSS or other software programs using the appropriate commands (e.g., ‘complex samples’ modules in SPSS).

However, it is important to be aware of this effect.
Confidence Intervals

Presenting confidence intervals in tables and reports

### Male Urban Labor Force Participation Rate by Age, 2010

<table>
<thead>
<tr>
<th>Age</th>
<th>Number</th>
<th>Mean (%)</th>
<th>95% Confidence Interval for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-14</td>
<td>425</td>
<td>16.2</td>
<td>12.7 - 19.7</td>
</tr>
<tr>
<td>15-49</td>
<td>1025</td>
<td>71.6</td>
<td>68.8 - 74.4</td>
</tr>
<tr>
<td>50-65</td>
<td>500</td>
<td>65.3</td>
<td>61.1 - 69.5</td>
</tr>
<tr>
<td>65+</td>
<td>125</td>
<td>41.7</td>
<td>33.0 - 50.4</td>
</tr>
</tbody>
</table>

Or more compactly:

### Male Urban Labor Force Participation Rate by Age, 2010 (% and 95% confidence interval)

<table>
<thead>
<tr>
<th>Age</th>
<th>Number</th>
<th>Mean</th>
<th>Mean Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>7-14</td>
<td>425</td>
<td>16.2 (+/-3.5)</td>
<td></td>
</tr>
<tr>
<td>15-49</td>
<td>1025</td>
<td>71.6 (+/-2.8)</td>
<td></td>
</tr>
<tr>
<td>50-65</td>
<td>500</td>
<td>65.3 (+/-4.2)</td>
<td></td>
</tr>
<tr>
<td>65+</td>
<td>125</td>
<td>41.7 (+/-8.7)</td>
<td></td>
</tr>
</tbody>
</table>
Confidence Intervals

For larger tables with many estimates, it can be awkward to present CIs all the time.

But it is essential for readers to understand the reliability of the estimates. For example, for disaggregated presentations (e.g., type of employment by rural women), numbers will get small and estimates correspondingly imprecise.

Options:

• Note at bottom of table something like ‘Margin of errors are between +/- 2% and +/- 5%’ (since there is no single ME for all the estimates shown).

• Within a table, when specific estimates are not reliable, use ‘—’ and note below table: ‘(-) indicates sample too small for reliable estimate.’

• Indicate in text that certain sample divisions are not shown, for same reason
2. Significance Tests

Beyond simply knowing the precision (CIs) of individual estimates:

“What is the labor force participation (LFP) rate for female secondary graduates?”

…we want to know how this statistic compares between groups:

“Is LFP rate for university graduates the same, higher, or lower than for secondary graduates?”

We are always making these comparisons, for example, on the table seen earlier. This is “Analytic” statistics -- examining relationships among variables, in this case, age and LFP.

• However, because we are dealing with estimates from samples, not the population, we can only make statements about probabilities

• So we need to test statistically for differences, not just compare the point estimates.
Significance Tests

The test we use depends on the nature of the variables we are comparing

**Continuous Variables**: *(interval or ratio variables).* Values are ranked low to high and we have a scale to interpret the differences. Examples:

- Age (years)
- Monthly income (Dinars)
- Value of assets (Dinars)

We know that someone who is 40 is twice as old as someone who is 20

**Ordered Variables**: Values are ranked in order of low to high, but we have no scale to measure the differences. Examples:

- School level: primary, secondary, university
- Satisfaction: very dissatisfied, somewhat dissatisfied, neither satisfied or dissatisfied, somewhat satisfied, very satisfied
Significance Tests

Nominal Variables: Values are categories that are not ordered.

Examples:
- Labor force participation status: 0=No, 1 =Yes
- Sector of Work: public, private wage, private self-employed
- Occupation: Manager/Professional, Technical, Clerical, Manual etc.

Also known as proportions variables (or frequency or percentage variables)—e.g., proportion of adult women in the labor force

For variables with multiple categories such as sector, we may assign numbers to responses in the survey (1=public, 2=pvt wage, 3=pvt self-employed) but these numbers do not imply any ordering or ranking.
Significance Tests

To compare two means of a continuous variable, we use the \textit{t-test}: 

- Requires that the variables be normally distributed or approximately so

- Usually we use the independent samples t-test, since the two groups are usually independent (non-overlapping) e.g., males vs females
  - The independent samples t-test is probably the most widely used test in statistics.

Example: We want to test whether monthly earnings in public sector are the same as in private wage sector.
Significance Tests

What we need:

Mean and variance of each variable:

Define:

Mean earnings in public sector = \( X_1 \)

Mean earnings in private sector = \( X_2 \)

Sample variance of \( X_1 = S^2_1 \)

Sample variance of \( X_2 = S^2_2 \)

(Note: the variances incorporate all the sample design factors that affect variance)

Also need:

\( n_1 = \) sample size for public sector earners

\( n_2 = \) sample size for private wage earners
Significance Tests

The t-statistic is then:

\[ t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S^2_{x_1}}{n_1 - 1} + \frac{S^2_{x_2}}{n_2 - 1}}} \]

The numerator, \( \bar{X}_1 - \bar{X}_2 \), is the difference in mean earnings in the two sectors. We want to know if this is statistically different from zero.

- Essentially, if the difference \( \bar{X}_1 - \bar{X}_2 \) is ‘large enough’ it is likely to indicate a true difference, rather than occurring just by chance.

- How large is ‘large enough’ depends on the variance of \( \bar{X}_1 - \bar{X}_2 \). This variance (or more precisely the estimated standard deviation) is given by the term in the denominator.
Significance Tests

How large is large enough for the difference to be considered significant? We define a *critical value* for the t-statistic. Assume we want 95% confidence, as before.

- We find the value such that if the t-stat exceeds this value, we can say with 95% probability the difference in the two means is not due to chance.

- The critical value depends on samples size, or relatedly, *degrees of freedom*
Significance Tests

Critical values and P-values

Think of the distribution—the further away \((X_1 - X_2)\) is from zero, the more likely there is indeed a difference in the means, and we can reject the null hypothesis that they are the same.

The probability the true difference is zero when the t-stat exceeds the critical value in either direction (for 2-tailed test) is just 5%.

The \(p\)-value is the probability the true value is zero. At the critical value it is 5%. If the t-stat exceeds this, p-value is less than 5%.
Significance Tests

In practice, software such as SPSS makes conducting t-tests very simple as everything is calculated automatically. For example, let ‘public’ indicate wage sector (0=private, 1=public), and wage = monthly wage.

To test if the average wage is the same in the two sectors:

```
T-TEST
/GROUPS=public(0 1)
/VARIABLES=wage.
```

SPSS will report p-values (also called ‘significance level’)

- If p < .05 it is significantly different from zero at 5% level
- If p < .01 it is significantly different from zero at 1% level
- If p < .10 it is significantly different from zero at 10% level
**Significance Tests**

Second most common test: To compare *proportions* for two groups, we use the chi-square test:

Example: is the unemployment rate higher for male or female youth?

Unemployment is a *proportion* or *percent* - the proportion of the group that is unemployed.

Note we are testing a relationship of two categorical variables—gender and unemployment. Hence we test the null hypothesis that unemployment rate is *independent* of gender.

(Why can’t we just use the t-test? Because that test requires that the variable has a normal or nearly normal distribution. This does not hold for nominal variables).
**Significance Tests**

Chi-square test commands in SPSS:

```
ANALYZE -> DESCRIPTIVE STATISTICS -> CROSSTABS
/TABLES=Unemployed BY sex
/STATISTICS=CHISQ.
```

As with the t-statistic, the Chi-Square test statistic is a measure of the difference of the two values (the proportions for males and females), and we are testing if this difference is greater than zero.

SPSS output will report the value of the Chi-Square test statistic and the \( p \)-value, e.g, **Chi-Square (1) = 6.1, \( p=0.02 \).**

If \( p < 0.05 \), there is less than a 5% chance this difference would occur by chance, so we can say the proportions are significantly different (with 95% probability)
Back to the t-test for continuous variables. It depends on key assumptions, one of which is that the variables are distributed normally:

Note the mean, shown here, and the median (middle value) are the same or close because the normal distribution is symmetric.
Significance Tests

But what if the distribution was not normal? For example, income and earnings data often have a distribution that is skewed to the right:

![Graph showing a skewed distribution with mode, median, and mean marked.]

Most people earn relatively low pay, but a small number have very high pay (in the right tail).

Now the median is below the mean—and here is a better measure of ‘central location’ or ‘where are most of the people found?’
**Implications of non-normal distribution:**

First, always look at the distribution of your variables:

- Is the distribution skewed? (is the median very different from the mean?)

- This may be due to outliers or bad values as well, not just the shape of the distribution, so is important to check

If this is the case, it may be more appropriate to report the median than the mean, or report both. The mean may be misleading.

- And for testing if the means of two variables are the same, the t-test is invalid, because it assumes normal distributions →
The Wilcoxon-Mann-Whitney test is an alternative to the independent samples t-test for cases where the variable is not normally distributed.

It makes no assumption at all about the distribution. It tests whether the medians for the two groups are the same.

Example: are hourly earnings in the public sector the same as in private sector?

SPSS code:
```
npar test
/m-w = wage by public(0 1).
```
3. Correlation Analysis

The 3rd major statistical test of the relation of two variables

Here we are looking to see how one variable changes with another.

For example, how are earnings (Y) related to years experience (or age) (X)?

For continuous (normally distributed variables) we use ‘Pearsons r’:

\[
 r(X,Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}
\]
Correlation Analysis

- If these variables vary perfectly with each other positively, \( r=1 \) (the two variables are essentially the same)
- If these variables vary perfectly with each other negatively \( r=-1 \)
- If these variables are completely unrelated, they do not vary with each other at all so \( r=0 \)

We normally expect values between 0,1 or -1,0. for example a likely correlation of earnings and experience might be .2

**Statistical test:** we test whether there is any relationship, so we test whether \( r \) is different from zero.

**CORRELATIONS**
/VARIABLES=earnings experience

Reports the correlation and p-values for test that correlation \( =0 \)