Automated Analysis of Survey Data using STATA

Vinay A. Patel
Deputy Manager
National Dairy Development Board
1. Background
2. Prior to using STATA
3. Survey Data Characteristics
4. Estimation of Important Variables
5. Automation of Survey data analysis
6. Important Commands-need to know
7. Interface between GIS and STATA
Background

• Types of Activities
• Various Surveys
• Repetitive Surveys
• Survey procedures
Prior to using STATA

- Time required @ every stage
- Data Collection
- Data Analysis in EXCEL
- Hypothesis Testing
- Report Preparation
Survey Data Characteristics

• Single-Stage Design:

```bash
svyset [ psu ] [ weight ] [ , strata (varname) fpc (varname) ]
```

- Primary Sampling Unit (psu)
- Sampling weights-pweight
- Strata
- Finite Population Correction (fpc)

```bash
svyset VILLAGE [pweight=invweight], strata(QUADSTRATA) fpc(FPC) poststrata(PSTRATA) postweight(THLACTHH) vce(linearized)
```
• Multi-Stage Design:

```plaintext
svyset psu [weight] [ , strata (varname) fpc (varname) ]
|| ssu [ , strata (varname) fpc (varname) ]
```

- Stages are delimited by "||"
- SSU – secondary / subsequent sampling units
• Important Variables

- Social Category
- Economic Category
- Village Surplus
- Milch Animals
- Milch Animals’ Yield
- Producer Surplus
- Milk Production
- Size of Family
- Milch Animal Households
**Problem:** % Distribution of Households by Social Category

**Through Stata command**

```stata
tabulate Tehsil CASTE if CASTE > 0, nofreq row

svy linearized: proportion CASTE,over(PSTRATA) cformat(%9.2f)
```

**Through Pivot table in excel**

<table>
<thead>
<tr>
<th>Tehsil</th>
<th>General</th>
<th>SC</th>
<th>ST</th>
<th>OBC</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amalner</td>
<td>9%</td>
<td>11%</td>
<td>12%</td>
<td>67%</td>
<td>100%</td>
</tr>
<tr>
<td>Bhadgaon</td>
<td>21%</td>
<td>20%</td>
<td>15%</td>
<td>44%</td>
<td>100%</td>
</tr>
<tr>
<td>Bhujawal</td>
<td>12%</td>
<td>21%</td>
<td>4%</td>
<td>63%</td>
<td>100%</td>
</tr>
<tr>
<td>Bodvad</td>
<td>16%</td>
<td>28%</td>
<td>10%</td>
<td>46%</td>
<td>100%</td>
</tr>
<tr>
<td>Chalingaon</td>
<td>26%</td>
<td>14%</td>
<td>14%</td>
<td>45%</td>
<td>100%</td>
</tr>
<tr>
<td>Chopda</td>
<td>2%</td>
<td>9%</td>
<td>35%</td>
<td>53%</td>
<td>100%</td>
</tr>
<tr>
<td>Dharangao</td>
<td>19%</td>
<td>13%</td>
<td>14%</td>
<td>52%</td>
<td>100%</td>
</tr>
<tr>
<td>Eradol</td>
<td>26%</td>
<td>13%</td>
<td>14%</td>
<td>47%</td>
<td>100%</td>
</tr>
<tr>
<td>Jalgaon</td>
<td>21%</td>
<td>10%</td>
<td>11%</td>
<td>58%</td>
<td>100%</td>
</tr>
<tr>
<td>Jamner</td>
<td>12%</td>
<td>24%</td>
<td>17%</td>
<td>47%</td>
<td>100%</td>
</tr>
<tr>
<td>Muktainagar</td>
<td>8%</td>
<td>19%</td>
<td>14%</td>
<td>58%</td>
<td>100%</td>
</tr>
<tr>
<td>Pachora</td>
<td>9%</td>
<td>13%</td>
<td>20%</td>
<td>57%</td>
<td>100%</td>
</tr>
<tr>
<td>Parola</td>
<td>30%</td>
<td>12%</td>
<td>13%</td>
<td>44%</td>
<td>100%</td>
</tr>
<tr>
<td>Ravar</td>
<td>7%</td>
<td>24%</td>
<td>23%</td>
<td>46%</td>
<td>100%</td>
</tr>
<tr>
<td>Yawal</td>
<td>8%</td>
<td>25%</td>
<td>22%</td>
<td>49%</td>
<td>100%</td>
</tr>
<tr>
<td>Jalgaon district</td>
<td>15%</td>
<td>17%</td>
<td>16%</td>
<td>52%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Problem: Composition of Milch Animals (%)

Through Stata command

```
tabstat lcper cbper bfper totaniper, by(Tehsil) columns(variables) format(%9.0f)
```

svy linearized : total LCIM CBIM BFIM inmilk lc cb bf MilchAnimal, over (PSTRATA) cformat(%9.2f)

Through Pivot table in excel

![Table 3.5: Composition of milch animals](image)
**Problem:** Estimated Milk Production Species-wise

Through Stata command

```stata
tabstat lcshare cbshare bfshare allshare, by(Tehsil) columns(variables) format(%9.0f)
```

Through Pivot table in excel

```stata
svy linearized : total LCPROD CBPROD BFPROD TOTPROD, over(PSTRATA) cformat(%9.0f)
```
Problem: Milk Procured by different Agencies

Through Stata command

```stata
tabstat sldcisper sllocalper sludhiaper sloutvillper
  slpvtdairyper totsaleper, by(Tehsil) columns(variables)
  format(%9.2f)
```

```stata
svy linearized: total
  SLDCS SLLOCAL SLDUDHIA
  SLOUTVILL SLPVTDairy
  totsale, over(PSTRATA)
  cformat(%9.2f)
```

Through Pivot table in excel

Table 3.9: Percentage volume of milk procured by different agencies

<table>
<thead>
<tr>
<th>Tahuka</th>
<th>DCS</th>
<th>Local</th>
<th>Dudhin</th>
<th>Outside village</th>
<th>Private dairy</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amlaner</td>
<td>23%</td>
<td>16%</td>
<td>51%</td>
<td>6%</td>
<td>4%</td>
<td>100%</td>
</tr>
<tr>
<td>Bhadgaon</td>
<td>28%</td>
<td>12%</td>
<td>11%</td>
<td>2%</td>
<td>36%</td>
<td>100%</td>
</tr>
<tr>
<td>Hushawal</td>
<td>22%</td>
<td>13%</td>
<td>34%</td>
<td>12%</td>
<td>19%</td>
<td>100%</td>
</tr>
<tr>
<td>Bodva</td>
<td>50%</td>
<td>18%</td>
<td>7%</td>
<td>0%</td>
<td>25%</td>
<td>100%</td>
</tr>
<tr>
<td>Chalisaon</td>
<td>43%</td>
<td>10%</td>
<td>12%</td>
<td>7%</td>
<td>27%</td>
<td>100%</td>
</tr>
<tr>
<td>Chopda</td>
<td>42%</td>
<td>19%</td>
<td>8%</td>
<td>17%</td>
<td>19%</td>
<td>100%</td>
</tr>
<tr>
<td>Dharangon</td>
<td>10%</td>
<td>9%</td>
<td>36%</td>
<td>26%</td>
<td>19%</td>
<td>100%</td>
</tr>
<tr>
<td>Erandol</td>
<td>11%</td>
<td>11%</td>
<td>36%</td>
<td>3%</td>
<td>38%</td>
<td>100%</td>
</tr>
<tr>
<td>Jalgaon</td>
<td>14%</td>
<td>14%</td>
<td>50%</td>
<td>6%</td>
<td>16%</td>
<td>100%</td>
</tr>
<tr>
<td>Jamnan</td>
<td>50%</td>
<td>43%</td>
<td>6%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>Multainagar</td>
<td>47%</td>
<td>19%</td>
<td>0%</td>
<td>3%</td>
<td>31%</td>
<td>100%</td>
</tr>
<tr>
<td>Fachoara</td>
<td>40%</td>
<td>8%</td>
<td>1%</td>
<td>1%</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>Farola</td>
<td>24%</td>
<td>7%</td>
<td>32%</td>
<td>3%</td>
<td>33%</td>
<td>100%</td>
</tr>
<tr>
<td>Raver</td>
<td>37%</td>
<td>26%</td>
<td>1%</td>
<td>3%</td>
<td>34%</td>
<td>100%</td>
</tr>
<tr>
<td>Yawal</td>
<td>44%</td>
<td>25%</td>
<td>20%</td>
<td>3%</td>
<td>8%</td>
<td>100%</td>
</tr>
<tr>
<td>Jalgoan district</td>
<td>30%</td>
<td>13%</td>
<td>21%</td>
<td>6%</td>
<td>29%</td>
<td>100%</td>
</tr>
</tbody>
</table>
Automation of Survey data analysis

- Create a **Do-file**
- Convert into ado-file
- Use same do or ado-file for repetitive use

```stata
* FOR NH LEVEL
clear
cd "C:/Users/vinayap/Desktop/Stata Do files/Unified Do Files/Milk Production and Surplus Study"
use jsjgsonNH.dta
set more off
label define caste_labels 1 "GEN" 2 "SC" 3 "ST" 4 "OBC" 5 "NR", replace
label define occu_label 1 "AGRI." 2 "AGRI. LABOUR" 3 "DARYING" 4 "JOB" 5 "BUSINESS" 6 "OTHERS" 7 "NR"
label define eoccatg_label 0 "NR" 1 "APL" 2 "BPL" 3 "ANT"
label define mah_label 0 "NON-MAH" 1 "MAH"
label define land_label 0 "Land less" 1111 "Marginal(0-2)" 2222 "Small(2-4)" 3333 "Medium(4-10)" 4444 "Large(10-15)"
label define animals_label 0 "No animals" 1111 "Marginal(0-2)" 2222 "Small(2-4)" 3333 "Medium(5-10)" 4444 "Large(10-15)"
lvole values CASTE Caste_labels
lvole values OCCU occu_label
lvole values EOCATG eoccatg_label
lvole values MAH mah_label
od "C:/Users/vinayap/Desktop/Stata/Unified stata_file stata_output"
*************** Table 3.2: Percentage distribution of households by social category
tabulate Tehsil CASTE if CASTE > 0, nolabel row
tabouta Tehsil CASTE if CASTE > 0 using table1_caste.xls, c(row)
*************** Table 3.3: Distribution of households by primary occupation
tabulate Tehsil OCCU if OCCU > 0, nolabel row
tabouta Tehsil OCCU if OCCU > 0 using table2_occu.xls, c(row)
*************** Table 3.4: Percentage distribution of households by economic category
tabulate Tehsil EOCATG, nolabel row
tabouta Tehsil EOCATG using table3_eocatg.xls, c(row)
*************** Table 3.8: Incidence of non-milking
gen LC NON = LCM if LCM > 0 & LCEROD ==0
gen CB NON = CBIM if CBIM > 0 & CBEROD ==0
gen BF NON = BFIM if BFIM > 0 & BFEROD ==0
teststat LCM CB NON CBIM BFIM BF NON BFIM, statistics(sum) by(Tehsil) columns(Variables) save
tabstatmat A
matrix table=A
xml file table4 using table4_nonmilking.xls
```
Important Commands-need to know

- For graphs
  - tableplot
  - vioplot
  - tabplot
  - catplot

- For Export of tables
  - tabout
  - xml_tab
  - outreg2
  - collapse
Interface between GIS and STATA

- **shp2dta** - Converts shape boundary files to Stata datasets

  ```
  shp2dta using shpfilename, database(filename) coordinates(filename) [options]
  ```

- **spmap** - Visualization of spatial data

  ```
  spmap DCSSALE using thlscoor.dta, id(_ID) title("Sold to DCS as a % of Total Sale in Tehsil") fcolor(white gs15 ltkhaki ltbblue) point( xcoord( x_c) ycoord( y_c) proportional( DCSSALE) size(*2) fcolor(eltgreen) ocolor(white)) label( xcoord( x_c) ycoord( y_c) label( teh_dcs) size(*.5))
  ```
THANK YOU