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:

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

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

```
svyset VILLAGE [pweight=invweight], strata(QUADSTRATA)
fpc(FPC) poststrata(PSTRATA) postweight(THLACTHH)
vce(linearized)
```

- Multi-Stage Design:

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

- Stages are delimited by "**||**"
- SSU – secondary / subsequent sampling units

Estimation of Important Variables

- 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

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

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

Through Pivot table in excel

Table 3.2: Percentage distribution of households by social category

Taluka	General	SC	ST	OBC	All
Amalner	9%	11%	12%	67%	100%
Bhadgaon	21%	20%	15%	44%	100%
Bhusawal	12%	21%	4%	63%	100%
Bodvad	16%	28%	10%	46%	100%
Chalisgaon	26%	14%	14%	45%	100%
Chopda	2%	9%	35%	53%	100%
Dharangaon	19%	13%	14%	52%	100%
Erandol	26%	13%	14%	47%	100%
Jalgaon	21%	10%	11%	58%	100%
Jamner	12%	24%	17%	47%	100%
Muktainagar	8%	19%	14%	58%	100%
Pachora	9%	13%	20%	57%	100%
Parola	30%	12%	13%	44%	100%
Raver	7%	24%	23%	46%	100%
Yawal	3%	25%	22%	49%	100%
Jalgaon district	**15%**	**17%**	**16%**	**52%**	**100%**

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

Tahsil	Milch animals (%)			
	Local cow	CB cow	Buffalo	All
Amalner	35%	7%	59%	100%
Bhadgaon	27%	17%	55%	100%
Bhusawal	24%	9%	67%	100%
Bodvad	46%	5%	48%	100%
Chalisgaon	30%	32%	38%	100%
Chopda	43%	7%	49%	100%
Dharangaon	33%	10%	57%	100%
Erandol	32%	11%	57%	100%
Jalgaon	41%	7%	52%	100%
Jamner	50%	6%	44%	100%
Muktainagar	35%	11%	54%	100%
Pachora	15%	30%	55%	100%
Parola	31%	15%	54%	100%
Raver	22%	15%	63%	100%
Yawal	19%	21%	59%	100%
Jalgaon district	**31%**	**15%**	**54%**	**100%**

Problem: Estimated Milk Production Species-wise

Through Stata command

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

```
svy linearized : total
LCPROD CBPROD BFPROD
TOTPROD, over(PSTRATA)
cformat(%9.0f)
```

Through Pivot table in excel

Table 3.7: Specie-wise milk production

Taluka	*Local cow*	*CB cow*	*Buffalo*	*All*
Amalner	20%	10%	71%	100%
Bhadgaon	14%	24%	62%	100%
Bhusawal	10%	12%	77%	100%
Bodvad	8%	8%	84%	100%
Chalisgaon	14%	42%	44%	100%
Chopda	25%	10%	65%	100%
Dharangaon	16%	14%	70%	100%
Erandol	13%	15%	72%	100%
Jalgaon	24%	9%	67%	100%
Jamner	16%	7%	77%	100%
Muktainagar	5%	19%	76%	100%
Pachora	5%	41%	54%	100%
Parola	13%	21%	66%	100%
Raver	6%	21%	73%	100%
Yawal	7%	26%	68%	100%
Jalgaon district	**13%**	**22%**	**65%**	**100%**

Problem: Milk Procured by different Agencies

Through Stata command

```
tabstat sldcsper  sllocalper
sldudhiaper sloutvillper
slpvtdairyper totsaleper,
by(Tehsil) columns(variables)
format(%9.2f)
```

```
svy linearized : total
SLDCS SLLOCAL SLDUDHIA
SLOUTVILL SLPVTDAIRY
totalsale, over(PSTRATA)
cformat(%9.2f)
```

Through Pivot table in excel

Table 3.9: Percentage volume of milk procured by different agencies

Taluka	DCS	Local	Dudhia	Outside village	Private dairy	All
Amalner	23%	16%	51%	6%	4%	100%
Bhadgaon	28%	12%	11%	2%	46%	100%
Bhusawal	22%	13%	34%	12%	19%	100%
Bodvad	50%	18%	7%	0%	25%	100%
Chalisgaon	43%	10%	12%	7%	27%	100%
Chopda	42%	19%	3%	17%	19%	100%
Dharangaon	10%	9%	36%	26%	19%	100%
Erandol	11%	11%	36%	3%	38%	100%
Jalgaon	14%	14%	50%	6%	16%	100%
Jamner	50%	43%	6%	0%	0%	100%
Muktainagar	47%	19%	0%	3%	31%	100%
Pachora	40%	8%	1%	1%	50%	100%
Parola	24%	7%	32%	3%	33%	100%
Raver	37%	26%	1%	3%	34%	100%
Yawal	44%	25%	20%	3%	8%	100%
Jalgaon district	**30%**	**13%**	**21%**	**6%**	**29%**	**100%**

Automation of Survey data analysis

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

Do-file Editor - vinay

File Edit Tools View

vinay | Untitled.do

```
*       FOR HH LEVEL
clear
cd "C:\Users\vinayap\Desktop\Stata Do files\Final Do files\Milk Production and Surplus Study"
use jalgaonHH.dta
set more off
label define caste_labels 1 "GEN" 2 "SC" 3 "ST" 4 "OBC" 0 "NR", replace
label define occu_label 1 "AGRI." 2 "AGRI. LABOUR" 3 "DARYING" 4 "JOB" 5 "BUSINESS" 6 "OTHERS" 0 "NR"
label define ecocatg_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-max)"
label define animals_label 0 "No animals" 1111 "Marginal(0-2)" 2222 "Small(2-5)" 3333 "Medium(5-10)" 4444 "Large(10-Max)"
label values CASTE caste_labels
label values OCCU occu_label
label values ECOCATG ecocatg_label
label values MAH mah_label
cd "C:\Users\vinayap\Desktop\stata_output"
*************** Table 3.2: Percentage distribution of households by social category
tabulate Tehsil CASTE if CASTE > 0, nofreq row
tabout 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, nofreq row
tabout Tehsil OCCU if OCCU > 0 using table2_occu.xls, c(row)
*************** Table 3.4: Percentage distribution of households by economic category
tabulate Tehsil ECOCATG, nofreq row
tabout Tehsil ECOCATG using table3_ecocatg.xls, c(row)
*************** Table 3.8: Incidence of non-milking
gen LC_NON = LCIM if LCIM > 0 & LCPROD ==0
gen CB_NON = CBIM if CBIM > 0 & CBPROD ==0
gen BF_NON = BFIM if BFIM > 0 & BFPROD ==0
tabstat LC_NON LCIM CB_NON CBIM BF_NON BFIM, statistics( sum ) by(Tehsil) columns(variables) save
tabstatmat A
matrix table4=A
xml_tab 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 ltblue ) 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