Standard Error of Estimates in Complex Surveys:
Estimating Village Household Milk Production at a sub-district level

Using Stata

Subir Mitra, B.E.(Electrical), PGDRM(IRMA)

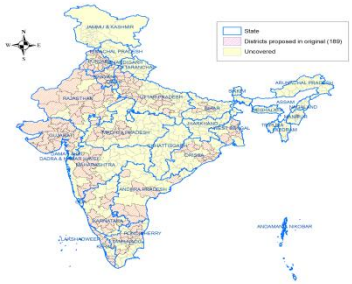
Senior Manager (Sectoral Analysis & Studies),
National Dairy Development Board, Anand

Aug 1, 2013

Outline

- One Stage Stratified Cluster Sampling Design, utilising GIS
- Introduction to the dataset
- Getting results using EXCEL & STATA
- Why Stata is better & quicker
- Acknowledging Statalist Forum
- Looking ahead!

Using GIS tool for areal stratification

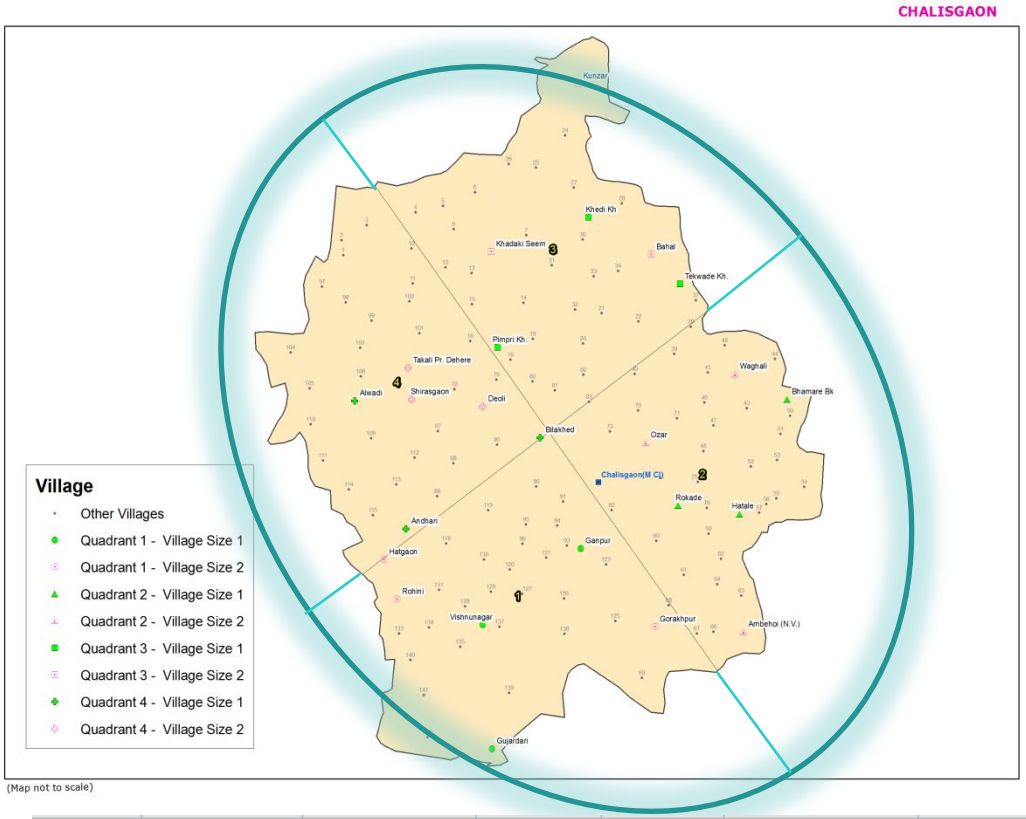


First, we use the **Directional Ellipse tool** of GIS to give geographic or areal representation, through GIS (ESRI ArcGIS) using digital maps of the sub-district plotted with village centroids.

Next, intersection of minor & major axis of the ellipse, gives us **4 geographical quadrants** within the ellipse, which we use as the first level of stratification.

Lastly, we stratify each of these 4 geographical strata into 2 further strata by grouping those villages which are higher & lower than the average number of households among all the villages in that quadrant.

Thus, now, we have **8 strata** - From each strata , we choose **2 sample villages** (clusters) randomly AND interview **all households** in those villages! In statistical jargon , this is called **Stratified One Stage Cluster Sampling**!



GEOQUAD	Average Households per Village	STRATA	Number of Villages	Number of Households	SAMPLE VILLAGE	Number of Households
		CHALISGAON-1-1			Ganpur	170
		CHALISGAON-1-1	18	2762	Gujardari	70
		CHALISGAON-1-2			Hatgaon	369
1	288	CHALISGAON-1-2	17	7333	Rohini	393
		CHALISGAON-2-1			Bhamare Bk	288
		CHALISGAON-2-1	29	5160	Rokade	103
		CHALISGAON-2-2			Ozar	630
2	356	CHALISGAON-2-2	11	9062	Waghali	1415
		CHALISGAON-3-1			Khedi Kh	330
		CHALISGAON-3-1	20	4334	Tekwade Kh.	154
		CHALISGAON-3-2			Bahal	1450
3	441	CHALISGAON-3-2	14	10671	Kunzar	651
		CHALISGAON-4-1			Andhari	263
		CHALISGAON-4-1	21	4274	Bilakhed	351
		CHALISGAON-4-2			Deoli	812
4	376	CHALISGAON-4-2	13	8508	Shirasgaon	548
		TOTAL	143	52104		

Data Definitions & Some Village Household data samples…..

SN	NAME OF VARIABLE	DEFINITION	REMARKS
1	TAHSIL	Name of tahsil	
2	VILLAGE	Name of Village	
3	SDTQVS	State District Tahsil Quadrant Village Sample Category	Quadrant Identification Number (each taluka cut into 4 quadrants)*
4	VILL CD	Village Code (16 digit Census 2001)	
5	HHNO	Household Number	
6	FMLYMEM	Number of family members in the household	
7	OPERLAND	Operation Land (acres)	
8	CASTE	Caste	General=1, SC=2, ST=3, OBC=4
9	OCCU	Occupation	Agri.=1, Agri. Labourer=2, Dairying=3, Service=4, Business=5, Others=6
10	ECOCATG	Economic Category	Above Poverty Line (APL)=1, Below Poverty Line (BPL)=2
11	LCIM	Number of Local Cow In Milk	
12	LCDRY	Number of Local Cow Dry	
13	LCPROD	Milk Production (In Ltrs) of Local Cows	
14	CBIM	Number of Crossbred Cow In Milk	
15	CBDRY	Number of Crossbred Cow Dry	
16	CBPROD	Milk Production (In Ltrs) of Crossbred Cows	
17	BFIM	Number of Buffalo In Milk	
18	BFDRY	Number of Buffalo Dry	
19	BFPROD	Milk Production (In Ltrs) of Buffaloes	
20	TOTPROD	Total Milk Production (In Ltrs)	
21	PUR	Purchase of Milk (In Ltrs)	
22	CONS	Consumption by the Household (Ltrs)	
23	SLDCS	Sale to DCS	
24	SLLOCAL	Sale to Local consumers within village	
25	SLDUDHIA	Sale to Dudhia	
26	SLOUTVILL	Sale outside village	
27	SLPVTDAIRY	Sale to Private Dairy	

TAHSIL	VILLAGE	VILLCD	STRQUAD	HHNO	FMLYMEM	OPERLAND	CASTE	OCCU	ECOCATG	LCIM	LCDRY	LCPROD	CBIM	CBDRY	CBPROD	BFIM	BFDRY	BFPROD	TOTPROD	PUR	CONS	SLDCS	SLLOCAL	SLDUDHIA	SLOUTVILL	SLPVTDAIRY
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	1	10	17	4	1	1	0	0	0	7	0	40	0	0	0	40	0	5	35	0	0	0	0
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	2	7	6	4	1	1	2	1	4.5	0	0	0	0	0	0	4.5	0	4.5	0	0	0	0	0
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	3	4	5.5	4	1	1	0	0	0	0	0	0	1	0	4	4	0	2	0	0	2	0	0
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	4	3	6.25	4	1	1	0	0	0	0	0	0	3	0	20	20	0	2	0	0	18	0	0
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	5	10	14	4	1	1	1	0	3	0	0	0	1	0	4	7	0	1	0	0	6	0	0
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	6	4	3	4	1	1	0	0	0	0	0	0	0	0	0	0	0.5	0.5	0	0	0	0	0
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	7	5	5	4	1	1	0	0	0	1	0	8	0	0	0	8	0	1	0	0	7	0	0
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	8	5	0	4	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	9	4	13	4	1	1	0	0	0	0	0	0	1	0	7	7	0	1	0	1	5	0	0
CHALISGAON	Bahal	27030013002751	CHALISGAON-3-2	10	4	3	4	1	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0

Getting Results Using Excel

$$\mathrm{var}(r_C) = \sum_{h=1}^{L} \left(\frac{M_h}{M}\right)^2 \frac{(1-f_h)}{m_h \mu_N^2} (s_{yh}^2 + r^2 s_{Nh}^2 - 2rs_{yNh})$$

Using Taylor's Method for finding standard error of estimates for milk production per household

Name of Tahsil Quadrant	Number of actual households in the village	Number of Villages in the Quadrant	Weight of the sample village	Number of census households in the sample village	Number of Milch Animal Owning Households in the sample village	Volume of milk production in the sample village	Number of sample villages in the quadrant	$\sum N_i$	$\sum N_i^2$	$\sum y_i$	$\sum y_i^2$	$\sum N_i y_i$	$\hat{\mu}_N = \frac{\sum N_i y_i}{\sum N_i / m_h}$	$\hat{\mu}_h = \frac{\sum y_i / m_h}{\sum N_i / m_h}$	$\hat{\tau}_y = \sum M_h \hat{\mu}_h$	$\hat{\tau}_N = \sum M_h \hat{\mu}_N$	Combined stratified ratio estimator (milk production per census households)	Sampling Fraction, which is the inverse of Weight	Variance of milk production in sample villages	Variance of census households in sample villages	Co-variance of milk production and census households in the sample villages	Variance of the estimate : Note carefully the expression on the extreme right bracket, which can be explained as the joint variation of 2 random variables viz. milk production and households. If these are independent of each other then the covariance term will vanish!
		M_h		N_i		y_i	m_h											$\frac{f_h}{m_h/M_h}$ =	s_{yh}^2	s_{Nh}^2	s_{yNh}	
STRQUAD	ACTH	QUADVILL	VGT	CENHH	MAH	TOTPROD																
CHALISGAON-1-1	270	18	9	170	52	231	2	240	57600	371	137270	12410	120	185	3335	2160		0.11	4186	5000	4575	0.0004
CHALISGAON-1-1	96		9	70	51	140																
CHALISGAON-1-2	450	17	9	393	105	275	2	762	580644	539	290521	91818	381	270	4582	6477		0.12	50	288	120	0.0000
CHALISGAON-1-2	528		9	369	137	265																
CHALISGAON-2-1	150	29	15	103	61	1542	2	391	152881	3728	13897984	62443	196	1864	54056	5670		0.07	208013	17113	59663	0.0024
CHALISGAON-2-1	413		15	288	195	2187																
CHALISGAON-2-2	1389	11	6	1415	232	2825	2	2045	4182025	3272	10702712	384350	1023	1636	17993	11248		0.18	2828631	308113	933561	0.0020
CHALISGAON-2-2	690		6	630	89	447																
CHALISGAON-3-1	411	20	10	330	69	318	2	484	234256	409	167281	27082	242	205	4090	4840		0.10	25765	15488	19976	0.0008
CHALISGAON-3-1	190		10	154	28	91																
CHALISGAON-3-2	787	14	7	651	295	1703	2	2101	4414201	4720	22273680	647345	1051	2360	33037	14707		0.14	862641	319201	524743	0.0036
CHALISGAON-3-2	1717		7	1450	314	3017																
CHALISGAON-4-1	443	21	11	351	75	448	2	614	376996	777	602952	48943	307	388	8153	6447		0.10	7021	3872	5214	0.0002
CHALISGAON-4-1	371		11	263	86	329																
CHALISGAON-4-2	642	13	7	548	143	1143	2	1360	1849600	2496	6230016	243200	680	1248	16224	8840		0.15	22261	34848	27852	0.0016
CHALISGAON-4-2	887		7	812	203	1354																
		M											$\hat{\mu}_N = \frac{\hat{\tau}_y}{M}$			$\hat{\tau}_N$	$r_C = \frac{\hat{\tau}_y}{\hat{\tau}_N}$					$\mathrm{var}(r_C)$
		143											422.29			60388	2.343					0.0111
																						Standard Error of the Estimate
																						0.1053

1. Summarise household data at village level, using Pivot Table
2. Feed the summarised data in the worksheet as above, carefully!
3. Repeat the process for all sub-districts, within one district, for one parameter of interest , at one time!
4. What if you have many such districts and couple of parameters to be estimated for each?

Getting Quicker & Better Results Using Stata!

```
use hh.dta
gen LC = LCIM+LCDRY
gen CB=CBIM+CBDRY
gen BF=BFIM+BFDRY
gen MILCH=LC+CB+BF
gen INMILK=LCIM+CBIM+BFIM
gen SALE= SLDCS+SLLOCAL+SLDUDHIA+SLOUTVILL+SLPVTDAIRY
gen MAH=1 if MILCH>0
replace MAH=0 if MAH==.
gen NMAH=1 if MILCH==0
replace NMAH=0 if NMAH==.
gen MPH=1 if TOTPROD>0
gen MSH=1 if SALE>0
gen CONSMAH=CONS if MAH==1
gen CONSNMAH=CONS if MAH==0
gen SURPLUS = TOTPROD - CONS
replace CASTE=5 if CASTE <1
replace OCCU=7 if OCCU<1
replace ECOCATG=4 if ECOCATG<1
tab  CASTE, gen( CAST)
tab  OCCU, gen( OCC)
tab  ECOCATG, gen( ECO)
rename CAST1 GEN
....
rename OCC1  AGRI
....
rename ECO1 APL
.....
collapse (first) STRQUAD (first) VILLCD (count)  ACTHH=HHNO  (sum) GEN  (sum),....., by (VILLAGE)
save villraw.dta,replace
merge 1:1  VILLCD using base.dta
drop _merge
sort STRQUAD
save vill.dta, replace
```

```
use vill.dta
gen pstrata=1
svyset VILLAGE [pw=WGT] , strata(STRQUAD)  poststrata(pstrata)  postweight(THSLHH) fpc( QUADVILL)
svy:ratio MILCH CENHH /* estimated milch animal holding in the tahsil per hh*/
svy:ratio MAH  CENHH /* estimated milk producing HHs as a ratio of total HHs  in the tahsil */
svy:ratio TOTPROD CENHH /* estimated milk production in the tahsil per hh*/
svy:ratio SURMAH CENHH /* estimated producer's  milk surplus in the tahsil per hh*/
svy:ratio SURPLUS CENHH /* estimated net milk surplus in the tahsil per hh*/
svy:ratio CONS TOTPROD /* estimated milk consumption all HHs to production ratio in tahsil*/
svy:ratio CONSMAH  TOTPROD /* estimated milk consumption of producing HHs to production ratio in tahsil*/
```

Estimating population parameters for the sub-district from the village level data

Summarising Household level data to the Village Level

Comparing results between Excel & Stata

EXCEL

We can observe that the milk production per census household->	2.3427 Ltr/day
Standard Error of estimate->	0.1053 Ltr/day
Therefore, assuming normal distribution, for a 95% confidence interval	
the upper limit is	2.5491 Ltr/day
and the lower limit is	2.1362 Ltr/day

STATA

```
. svy:ratio TOTPROD CENHH
(running ratio on estimation sample)

Survey: Ratio estimation

Number of strata =         8          Number of obs     =        16
Number of PSUs   =        16          Population size   =     52104
N. of poststrata =         1          Design df         =         8

      _ratio_1: TOTPROD/CENHH

------------------------------------------------------------------------------
             |                Linearized
             |       Ratio   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _ratio_1 |  2.342667   .1053275       2.099782    2.585553
------------------------------------------------------------------------------
```

The estimate and the standard error is the same, only the width of the limits are higher in Stata since it assumes a t-distribution, which is more accurate statistically speaking, as we have a very small sample size!!

Discovering & Using Stata : Acknowledging excellent support from Statalist Forum

From: owner-statalist@hsphsun2.harvard.edu on behalf of Stas Kolenikov <skolenik@gmail.com>
Sent: Wednesday, January 30, 2013 8:00 PM
To: statalist@hsphsun2.harvard.edu
Subject: Re: st: one stage cluster with preliminary stratification

```
*** assuming n1, n2, m1, m2, N are contained in the identically named scalars

gen wgt = scalar(n1)/ 10 if stratum== 1
replace wgt = scalar(n2)/ 8 if stratum==2 assert !missing(wgt)
* cluster size does not matter

*** option 1: poststrata
gen pstrata = 1
gen popsize = scalar(N)
svyset cluster [pw=wgt], strata( stratum ) poststrata( pstrata ) postweight( popsize )

*** option 2: rescale weights
sum wgt
generate wgt2 = wgt*scalar(N)/r(sum)
svyset cluster [pw=wgt2], strata( stratum )

—
— Stas Kolenikov, PhD, PStat (SSC) :: http://stas.kolenikov.name
— Senior Survey Statistician, Abt SRBI :: work email kolenikovs at srbi dot com
— Opinions stated in this email are mine only, and do not reflect the position of my employer
```

On Wed, Jan 30, 2013 at 6:32 AM, Subir Mitra <subir@nddb.coop> wrote:
> ONE STAGE CLUSTER WITH PRELIMINARY STRATIFICATION -I stratify population N (members living in clusters , which is known) into 2 strata and randomly pick up 10 clusters from 1st stratum and 8 clusters from 2nd stratum (stratum population n1 & n2 and total clusters m1 & m2 in both stratum also known) and all members of the clusters are sampled.
>
> Any guidance to me to find the svyset command in this case, assuming N, m1,m2,n1 and n2 known and I want to make use of it? (The problem is from Schaeffer et al 1996-328 problem 8.19)

Summary

- Stata is very user friendly and much easy to learn. *Inspiration from World Bank household survey analyses!*
- Excellent support exists from practitioners.
- We look foreword to work together with Stata Users and exchange ideas!
- We are exploring using Stata in Monitoring & Evaluation- Baseline Survey & Follow-ups

  (National Dairy Plan 2012-17)

  Propensity Score Matching or Regression?