AGRODEP

Household survey data course

Dakar, 8-10 October 2012

Design effects

IFPRI®

HarvestChoice

BETTER CHOICES, BETTER LIVES

THE WORLD BANK

LSMS

Living Standards Measurement Study

Design Effects

- Several (probability) sampling designs
- Any departure from SRS may affect precision
- Design effect (DEFF) is a measure of relative efficiency of sample design in relation to SRS

Design effects (cont'd)

- DEFF defined as the ratio between variance from a particular complex sample design and variance from SRS of same size
- DEFT = square root of DEFF, expressed in terms of the standard error of an estimate
- Takes into account the effects of stratification and clustering

Importance of Design Effects

- Necessary to calculate sampling errors and design effects based on actual sample design
  - If *deff* is ignored, the sampling errors will be underestimated, and the conclusions from any test of hypothesis or analysis will be biased
- Statistical software will always assume simple random sampling unless told otherwise
- In Stata standard errors and deff for complex designs can be calculated using the `svy` commands.

Suggestion…

- In publication of survey results, important to include an annex describing the accuracy of survey results
  - Sources of nonsampling error
  - Tables of standard errors and other measures of precision for most important survey estimates, at different levels of disaggregation
- Tables of standard errors should include CVs, confidence intervals and design effects

Variable	Value (R)	Standard error (SE)	Design effect (DEFT)	Relative error (SE/R)	Confidence limits R-2SE	R+2SE
Urban	0.284	0.011	2.518	0.039	0.262	0.306
Literate	0.673	0.012	2.682	0.018	0.648	0.697
No education	0.242	0.012	2.794	0.049	0.219	0.266
Secondary education	0.086	0.006	2.096	0.067	0.074	0.097
Net attendance ratio	0.731	0.011	2.050	0.015	0.709	0.753
Never married	0.230	0.007	1.607	0.029	0.216	0.243
Currently married	0.673	0.007	1.491	0.010	0.659	0.687
Married before age 20	0.645	0.008	1.566	0.013	0.628	0.661
Currently pregnant	0.105	0.004	1.249	0.036	0.098	0.113
Children ever born	2.912	0.037	1.365	0.013	2.837	2.987
Total Fertility Rate (3 years)	5.659	0.137	1.827	0.024	5.385	5.933
Children surviving	2.473	0.031	1.312	0.012	2.412	2.535
Children ever born to women 40-49	6.367	0.095	1.337	0.015	6.177	6.557
Knows any contraceptive method	0.979	0.003	1.818	0.003	0.972	0.985

Design Effects

- The cluster effect measures the inefficiency of two-stage sampling relative to SRS

$$ceff = \frac{e_{TSS}^2}{e_{SRS}^2}$$

- In complex sample design (with many stages, stratification, etc.)

$$deff = \frac{e_{complex}^2}{e_{SRS}^2}$$

- The design effect can also be interpreted as

$$deff = \frac{n}{n_{SRS}}$$

where n_{SRS} is the size of an SRS sample with the sample level of error

- Some researchers use :

$$deft = \sqrt{deff} = \frac{e_{complex}}{e_{SRS}}$$

Example # 1: Malawi Third Integrated Household Survey (IHS3) 2010/11

- 31 strata (districts)
  - Representative at the national-, urban/rural-, regional, & district-levels
- 768 primary sampling units (PSU)
- 16 HHs per PSU; Total sample:12,288 HHs

Main Messages

- Necessary information to construct correct *point estimates* from complex surveys is contained in sampling weights
- Knowing sampling weights alone not enough to construct correct *standard errors*
- Stratification generally yields more precision per observation unit than SRS, while clustering usually does the opposite
- Design effect: Measure of precision gained or lost by use of the more complex design instead of SRS
- In surveys with both stratification & clustering, the overall design effect depends on whether more precision is lost by clustering than gained by stratification

Example # 2: Tanzania National Panel Survey (TZNPS) 2008/09

- Given a fixed budget, what are the implications of cluster size (# HHs/cluster) on total sample & standard error of household per capita consumption expenditures estimates?
- Using the Tanzania Household Budget Survey 2000/01 for estimates of between & within-group SD of consumption

TZNPS 2008/09 (Cont'd)

- Alternative sample designs with different implications for survey costs & precision

# of Households/ Cluster	*# of Clusters*	*Total # of Households*	*SE of Consumption (Tshillings)*
24	193	4,625	725
20	218	4,368	688
16	252	4,032	650
12	298	3,574	612
8	328	3,276	594
6	364	2,912	578

Design Effects

- In order to calculate design effects for a particular dataset, you first need to define the complex design for Stata.
- This example uses the common two-stage cluster sample, but other more complicated designs are also supported.

```
svyset clusterid [w= hh_weight_trimmed], strata(strataid)
```

To simply calculate the design effects for the overall sample, use the following commands:

```
svy: mean hhsize
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      16          Number of obs    =     3265
Number of PSUs   =     409          Population size  =  7245851
                                    Design df        =      393

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      hhsize |   5.166811   .0668891      5.035306    5.298316
--------------------------------------------------------------

estat effects

-----------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.       DEFF      DEFT
-------------+---------------------------------------
      hhsize |   5.166811   .0668891     1.76149   1.32721
-----------------------------------------------------
```

Over subpopulations:

```
svy: mean hhsize, over (rural)
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      16          Number of obs    =     3265
Number of PSUs   =     409          Population size  =  7245851
                                    Design df        =      393

            1: rural = 1
            2: rural = 2

--------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
hhsize       |
           1 |   5.436019   .0809182      5.276932    5.595105
           2 |    4.41353     .10837      4.200473    4.626587
--------------------------------------------------------------
```

```
estat effects, srssubpop

            1: rural = 1
            2: rural = 2

-------------------------------------------------------------
             |             Linearized
        Over |       Mean   Std. Err.       DEFF      DEFT
-------------+-----------------------------------------------
hhsize       |
           1 |   5.436019   .0809182     1.58701   1.25977
           2 |    4.41353     .10837     2.04029   1.42839
-------------------------------------------------------------
```

Design Effects

Then to calculate ϱ, use the following formula:

$$deff = \left[1 + \rho(m-1)\right]$$

where *m* is the cluster size. You will know the cluster size either from the survey documentation or it can be calculated from the data:

```
gen n=1

collapse (sum) n, by (clusterid)

sum n
```

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           n |       409    7.982885    .1298585          7          8
```