Analysing Complex Social Surveys

Scottish Social Survey Network, Master Class

Stirling, 25 March 2010

Peter Lynn

University of Essex

INSTITUTE FOR SOCIAL & ECONOMIC RESEARCH

University of Essex

What is a Complex Survey?

Features of importance to analysts:

Sample Design: Clustered sample

Stratified sample

Variable selection probabilities

Unit non-response

Item non-response

How Does Complex Design Affect Estimates?

Effects on both bias and variance of sample statistics

$$MSE(y) = E(y-Y)^2$$

$$= E[y-E(y)]^2 + (E(y)-Y)^2$$

$$= Var(y) + Bias^2;$$

On bias: variable selection probabilities (disproportionate stratified sampling)

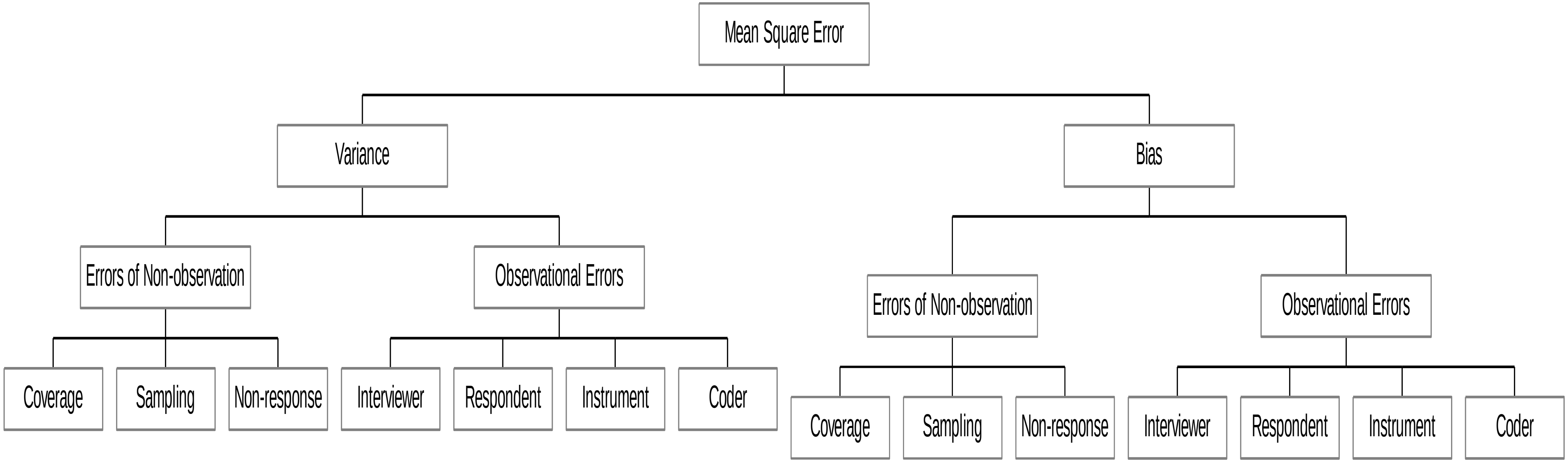
On variance: stratification (proportionate stratified sampling)

variable selection probabilities

clustering (multi-stage sampling)

Also: non-response

Complex Design in the Context of Survey Error



Disproportionate Stratified Sampling

Probability sampling does not require all units to have an equal probability of selection

Disproportionate sampling involves:

-Selecting a sample independently from each stratum;

-Allowing the sampling fraction, $f_i = \frac{n_i}{N_i}$, to vary between strata

Motivation 1: to increase sample size from particular strata of interest without unduly increasing overall sample size.

Effect 1: to increase precision for estimation within that stratum but, usually, to reduce precision for total sample estimation.

Motivation 2: to over-sample strata with particularly high variance.

Effect 2: to increase precision for total sample estimation.

Effect of VSPs on Estimates: Example

Population of 6 individuals with associated measures:

Men			Women		
A	B	D	C	E	F
2	6	10	8	10	12

Consider a sample design: to sample two men and one woman (disproportionate stratified SRS).

There are 9 possible samples of *n*=3 from *N*=6

Sample	1	2	3	4	5	6	7	8	9
Members of sample	A B C	A D C	B D C	A B E	A D E	B D E	A B F	A D F	B D F
Values in sample	2 6 8	2 10 8	6 10 8	2 6 10	2 10 10	6 10 10	2 6 12	2 10 12	6 10 12

Example Continued

Design weights are $w_1 = \frac{3}{2}$; $w_2 = \frac{3}{1}$.

For each sample, we can calculate both an unweighted and weighted sample mean.

e.g. for sample #1, we have these data:

Sample Member	X	Stratum (l)	Design weight (w_l)
A	2	1	1.50
B	6	1	1.50
C	8	2	3.00

So, $\hat{\bar{X}} = \frac{(1.5\times2)+(1.5\times6)+(3.0\times8)}{(1.5+1.5+3.0)} = \frac{36}{6} = 6.0$

Example Continued

For all samples, we obtain:

Sample	1	2	3	4	5	6	7	8	9	
Members of sample	A B C	A D C	B D C	A B E	A D E	B D E	A B F	A D F	B D F	
Values in sample	2 6 8	2 10 8	6 10 8	2 6 10	2 10 10	6 10 10	2 6 12	2 10 12	6 10 12	
Unwtd Sample mean	5.33	6.67	8.00	6.00	7.33	8.67	6.67	8.00	9.33	*7.33*
Wtd sample mean	6.0	7.0	8.0	7.0	8.0	9.0	8.0	9.0	10.0	*8.0*

Note $\bar{X} = 8.0.$ Unweighted mean is biased. Effect of design weighting is, for all possible samples except one, to increase the estimate of the mean.

Disproportionate Stratified Sampling: Estimation

For unbiased estimation, we can no longer use the direct sample analogue of the population parameter. Instead, we should use the Horvitz-Thompson estimator, which in the case of a mean is:

$$\hat{\hat{X}} = \frac{\Sigma_{l=1}^{n} w_l x_l}{\Sigma_{l=1}^{n} w_l}$$

Where w_l is the *design weight* (or *sampling weight*) assigned to sample unit l .

Design weights proportional to the inverse of the selection probability: $w_{il} = \frac{N_i}{n_i}$.

Special case of unweighted (epsem) data (wl=1): $\hat{\hat{X}} = \frac{\Sigma_{l=1}^{n} x_l}{n}$

So, weights are likely to affect the estimate of a mean. They will do so if there is an association between *w* and *x* and the effect will be greater the stronger the association.

Effects of Sample Design on Variance of Estimates

Variance of estimates under *Simple Random Sampling (SRS)*:

$Var(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n},$ where $S^2 = Var(y)$

However, typical sample designs are not SRS, but instead involve stratification, clustering (multi-stage designs), and variable selection probabilities.

These features also all affect variance of estimates.

Design Effects

The effect of sample design on sampling variance.

The effect can (and will) be different for different estimates from the same survey.

We should not refer to *the* design effect for a particular sample design.

DEFF is the ratio of the actual sampling variance to SRS sampling variance for sample of same size:

$$Deff = \frac{Var_c(\bar{y})}{Var_{SRS}(\bar{y})}$$

where

$Var_c(\bar{y})$ is the sampling variance of the complex design under consideration, and

$Var_{SRS}(\bar{y})$ is the sampling variance of a simple random sample of the same size.

Design Factor; Effective Sample Size

The equivalent ratio of standard errors is known as the *design factor*, *DEFT*:

$$DEFT = \frac{S.E._C}{S.E._{SRS}} = \sqrt{DEFF}$$

The *effective sample size*, *neff*, is the size of a simple random sample that would have produced the same precision as the actual (complex) sample design under consideration:

$$neff = \frac{n}{DEFF}; \qquad DEFF = \frac{n}{neff}.$$

e.g. if *DEFF* = 2 and n = 1000, then *neff* = 500

Standard Errors for Stratified Sampling

In general: $Var(\bar{x}) = \sum_{i=1}^{I} \frac{N_i^2 S_i^2}{N^2 n_i}(1 - \frac{n_i}{N_i})$,

where $S_i^2 = Var_i(x)$ is the variance of x within stratum *i*

Note:

1. Differences between strata do not contribute to $Var_i(x)$.
2. Sampling variance will be reduced if strata are homogeneous (small S_i^2).
3. In the case of proportionate sampling, $\frac{n_i}{N_i} = \frac{n}{N}$, so

$$Var(\bar{x}) = \frac{1}{n}(1 - \frac{n}{N}) \sum_{i=1}^{I} N_i S_i^2$$

Deff due to Proportionate Stratified Sampling: Example

GHS 1996

Insalaco (2000) compared effect of alternative stratification on variance of estimates

Used regression models at postcode sector level

‘Best’ stratification (applied from 2000 GHS onwards) consisted of:

- Government office regions, with subdivisions for Scotland, Wales, met areas of England, London (8 quadrants)
- % households with no car (3 categories)
- % households with head prof, employer, mgr (3 categories)

This stratification resulted in the following estimated *deffs*:

Estimate	*Deff*
% adults heavy drinkers	0.88
% households with elderly person	0.77
% respondents seen GP in last 2 weeks	0.91
% divorced HoH	0.89
% ethnic minority HoH	0.52
% inpatient in last year	0.89
% lone parent families	0.83
% hhds below bedroom standard	0.84
% employed with a pension scheme	0.87
% stepchildren (of children)	0.95

(Insalaco, 2000)

Deff due to Variable Selection Probabilities

Recall (ignoring f.p.c.): $Var(\bar{x}) = \sum_{i=1}^{I} \frac{N_i^2 S_i^2}{N^2 n_i}$

In many situations, it will be the case that S_i^2 varies little, so it is instructive to consider $Var(\bar{x})$ when $S_i^2 = S^2$. Then, $Var(\bar{x}) = \frac{S^2}{N^2} \sum_{i=1}^{I} \frac{N_i^2}{n_i}$

So, $Deff(\bar{x}) = \frac{n}{N^2} \sum_{i=1}^{I} \frac{N_i^2}{n_i}$

$$= \frac{n}{N^2} \sum n_i (w_i^2)$$

Note: the effect on the variance is not dependent on applying design weights in analysis. It is dependent on the sample design.

Deff due to VSPs: Example

Austria	AT	1.25
Switzerland	CH	1.21
Czech Republic	CZ	1.25
Spain	ES	1.22
France	FR	1.23
Greece	GR	1.22
Ireland	IE	1.04
Israel	IL	1.56
Italy	IT	1.16
Luxembourg	LU	1.26
The Netherlands	NL	1.19
Portugal	PT	1.83
United Kingdom	UK	1.22

Lynn et al, 2007, JOS 23: 107-124

Design Effect due to Clustering

Clustering tends to increase sampling variance. This is because elements within a cluster tend to be more homogeneous than elements as a whole.

Clustering therefore tends to have the opposite effect to stratification.

The design effect due to clustering takes the form:

$$Deff_{cl} = 1 + (b - 1)\rho$$

where b is sample size per cluster (in practice b may vary – see next page), and ρ (roh) is the intra-cluster correlation.

$\rho = 0$: randomly sorted clusters

$\rho = 1$: perfectly homogeneous clusters

Note: ρ is a population characteristic relating to the chosen definition of PSU; b is chosen by the researcher as part of the sample design

e.g. b = 10: if $\rho = 0$ then $Deff_{cl} = 1$; if $\rho = 1$ then $Deff_{cl} = 10$; more realistically, if $\rho = 0.05$ then $Deff_{cl} = 1.45$

A Note about Cluster Sample Size

$Deff_{cl} = 1 + (b - 1)\rho$ strictly holds only with no variation in cluster sample size, i.e. $n_j = b \; \forall \, j$. For complex surveys, where n_j may vary and, additionally, unequal selection probabilities may be used, the design effect due to clustering is:

$$Deff_{cl} = 1 + (b^* - 1)\rho$$

where $b^* = \frac{\sum_{j=1}^{J}\left(\sum_{l=1}^{n_j} w_{jl}\right)^2}{\sum_{j=1}^{J}\sum_{l=1}^{n_j} w_{jl}^2}$.

Note that for an epsem design, this gives $b^* = \frac{\sum_{j=1}^{J}(n_j)^2}{\sum_{j=1}^{J} n_j}$

In some situations, notably when variation in n_j is small, mean cluster size, $\bar{b} = \bar{n}_j$, may provide an adequate approximation. But often it is a poor approximation: see Lynn & Gabler (2005)

Lynn P & Gabler S (2005) Approximations to b* in the prediction of design effects due to clustering, *Survey Methodology*, **31**, 101-104

Example of Intra-Cluster Correlations

From the British Social Attitudes Survey:

Variable	$\hat{\rho}$	b	$\widehat{Deft}$	$\widehat{Deft}$ if $b = 10$
Household size	0.070	16.6	1.45	1.28
Owner-occupier	0.231	16.5	2.14	1.75
Has telephone	0.102	16.5	1.61	1.38
Asian	0.334	8.3	1.86	1.53
Roman Catholic	0.037	16.4	1.25	1.15
Not racially prejudiced	0.021	8.4	1.08	1.03
Extra-marital sex wrong	0.044	8.3	1.15	1.08
Dodging VAT is OK	0.021	8.2	1.07	1.04

Note: small $\hat{\rho}$ for attitudinal variables, so design effects small. But large $\hat{\rho}$ for variables related to ethnicity and housing type. The most effective degree of clustering might be greater for an attitude survey (fewer clusters, larger n_j) than for a housing survey.

Effect of Non-Response on Estimates

Population → Selected Sample → Responding Sample

Non-response as an extension of sampling

$$y_r = y_n + \left(\frac{y_{nr}}{y_n}\right)(y_r - y_{nr})$$

Non-response error (bias) is product of two components:

- Non-response rate
- Difference between respondents and non-respondents

Note: deterministic and stochastic models of non-response lead to same expression for realised error

Weighting for Non-response

Analogous to design weighting, but:

- For well-specified sample designs, selection probabilities are *known*;
- Response probabilities must be *estimated*

Inclusion probability is product of (known) selection probability and (unknown) response probability: $\pi_i = \pi_i^s \times \pi_i^{r|s}$.

Usual approach is to estimate response probability and then adjust design weight:

$$w_i = \frac{1}{\pi_i} \times \frac{1}{\hat{\pi}_i^{r|s}}$$

Estimation of Response Probability

Two broad approaches:

- Model predictions
- Observed response rates for subgroups

Key issues:

- How to develop the model / define the subgroups
- Auxiliary data / covariates

Sources of Auxiliary Data

Sample Frame

Linked geographical data

Other linked data

Interviewer observations

Survey process data

Defining Weighting Classes

Desirable criteria:

- Response rates vary between classes (hence, “response homogeneity groups”)
- Sample statistics vary between classes
- Sample statistics similar for respondents and non-respondents within classes
- Class sample sizes not too small

Effect of Non-Response on Estimates, ctd.

Adjusted weight (“combined weight”) can be used in standard way

Bias will be reduced but not removed

Proportion of bias removed depends on extent to which criteria for weighting classes are met

This cannot be known and will vary between estimates

Accounting for Design in Analysis

Re. bias: use weights

Recall earlier example of effect of VSPs on estimates

Effects applies to any statistic (estimate) that is associated with inclusion propensity

Re. variance: use appropriate estimation method that reflects the relevant sources of variance (VSPs, clusters, strata)

Mis-Specification Effects

The mis-specification effect summarises the impact (on estimated variances of sample estimates) of failing to take account of sample design in estimation.

$$MEFF = \frac{Var(\bar{x})_{C(SRS)}}{Var(\bar{x})_{C(C)}}$$

where $Var(\bar{x})_{A(B)}$ is the variance of $\bar{x}$ using estimator B with data collected under design A.

Note: this is closely related to (the reciprocal of) *DEFF*, but it is not the same. The denominator of *MEFF* is the same as the numerator of *DEFF*. But the numerator of *MEFF* is not the same as the denominator of *DEFF* (denominator of *DEFF* is $Var(\bar{x})_{SRS(SRS)}$).

Meff due to Proportionate Stratified Sampling

Usually > 1.0, i.e. standard errors over-estimated

Reason is that proportionate stratification tends to reduce standard errors (design effect < 1.0)

But effects are typically modest and s.e. estimates are at least conservative if mis-specified

Meff: Example

Returning to earlier example of $N = 6$, $n = 3$:

If we estimated as if the sample was SRS, we would get biased estimates of $\bar{x}$ (already shown), but also biased estimates of $Var(\bar{x})$.

We would estimate $Var(\bar{x}) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$

The mean of the 9 possible sample values of s^2 is 10.07, so on average we would obtain:

$$Var(\bar{x})_{C(SRS)} = \frac{10.07}{3}\left(1 - \frac{3}{6}\right) = 1.68.$$

Whereas,

$$Var(\bar{x})_{C(C)} = \sum_{i=1}^{I} \frac{N_i^2 S_i^2}{N^2 n_i}\left(1 - \frac{n_i}{N_i}\right)$$

$$= \frac{3 \times 32}{6^2 \times 2}\left(1 - \frac{2}{3}\right) + \frac{3 \times 8}{6^2 \times 1}\left(1 - \frac{1}{3}\right) = 0.89$$

$$So, MEFF = \frac{1.68}{0.89} = 1.89.$$

We would have over-estimated the variance by 89%. This is an unusual case in which the stratification effect is greater than the effect of variable sampling fractions (and there is no clustering).

Meff due to Clustering

$1+\left(b^{*}-1\right)\rho$ can be large, so failure to specify clustering can result in a serious *meff*.

Estimated design effects from World Fertility Survey

Deff	Thailand	Columbia	Nepal
% currently married	1.02	1.38	1.14
Number of marriages	1.32	1.84	-
Children ever born	1.47	1.28	2.08
Months breast fed	2.04	1.74	2.08
% wanting no more children	1.18	1.16	2.13
% expressing boy-preference	1.15	1.11	2.37
% knowing condoms	1.96	3.22	2.44
% using modern contraceptive	2.53	2.90	2.10

(Verma, Scott & O’Muircheartaigh, 1980, JRSSA 143: 431-473)

Software and Methods

Complex standard errors can be estimated “correctly” in many general software packages, including:

Stata

SPSS

SAS

The basic necessity is for the survey data file to include variables indicating:

Design weight

Stratum

Cluster

Implementation in Stata

Command to indicate the design:

```
svyset psuid [pweight=weight], strata(stratumid)
```

Then, all commands pre-fixed by “svy:”, e.g.:

```
svy: mean income

svy: logistic y x1 x1
```

Special Topic 1: Multi-Domain Designs

Differences in numbers or definitions of strata or PSUs are dealt with standardly

Differences in numbers of stages of selection are generally dealt with standardly as only first stage (PSUs) needs to be specified

Multi-stage design in one domain and single-stage in another is not dealt with standardly: User should derive a NEWPSU=PSU for multi-stage domain and NEWPSU=PID for single-stage domain, then specify NEWPSU as PSU variable

(e.g. for combining GB and NI components of BHPS)

(Gabler, Häder & Lynn, 2006, Design effects for multiple design samples, *Survey Methodology* 32: 115-120)

Special Topic 2: Public-Use Data Files

Beware PSU variable for multi-domain design.

E.g. EU-SILC for Norway, “PSU” is in fact a municipality indicator for all cases, despite municipalities being used as PSUs only in one domain; other domain is unclustered

Strata variable not always supplied, but sometimes REGION can be used if it formed part of the stratification

PSU variable sometimes not supplied: big problem!

Special Topic 3: Longitudinal Designs

Relatively little is known about design effects for longitudinal estimates

Some suggestion that they may be:

Smaller for $\overline{\Delta}y$ than for $\bar{y}$ (Vieira & Skinner, 2005)

Smaller over waves for some $\bar{y}$ - declustering effect (Lynn & Fumagalli, 2008)

Increasing over waves for regression coefficients (Vieira & Skinner, 2005)

Special Topic 4: Domain Comparison

For independent domains, can manually add variance estimates for each domain

Or, can use appropriate estimation methods in standard packages, e.g. lincom in Stata

But, lincom assumes domains independent of design, i.e. domain n is random

User should specify domains as strata before using lincom