RESUME: Discussion of inference from Population Samples

NHIS Sample Design

Cross-sectional household interview survey

- Sampling and interviewing continuous thru year.
- Multistage area probability design
- Representative sample of households & group qtrs
- Sampled clusters of addresses in primary sampling units (PSU's). PSU consists of — county, a small group of contiguous counties, or a metropolitan statistical area.
- NO oversample race/ethnicity groups at household level.
- BUT persons aged 65 or older, blacks, Hispanics, Asians have higher chance to be selected in household/group quarters
- ALSO representative sample of children

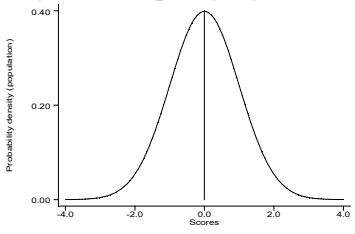
Basic Sampling Theory for Simple and Cluster Samples

Malcolm Rosier
Survey Design and Analysis Services Pty Ltd
http://survey-design.com.au

Sample design

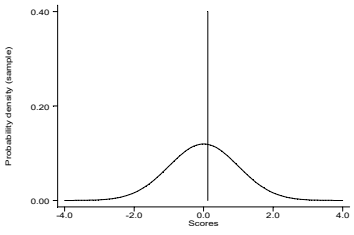
- The focus of the design for a sample must be on the magnitude of the standard errors of sampling not than on an arbitrary percentage of the target population.
- The standard errors are used to calculate confidence intervals around the sample data.

Graph: Target population



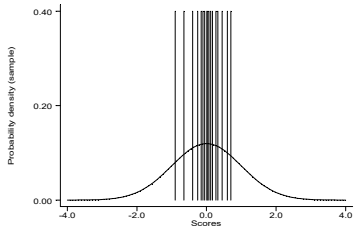
- Population: mean = μ, standard deviation = σ

Graph: Sample



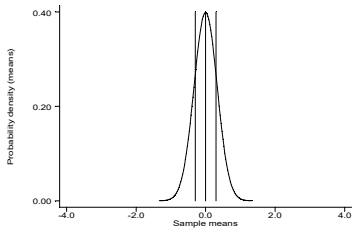
- Sample: mean = x, standard deviation = s

Graph: Means from many samples



- However we could get many different samples with different sample means from the population.

Graph: Distribution of sample means



- This gives us a sampling distribution of sample means:

Sampling distribution of sample means

- normal distribution
- mean = μ = mean of underlying population distribution
- standard deviation = $\sigma / \sqrt{n}$

Standard error of a population mean

The standard deviation of the sampling distribution of sample means is termed the standard error of a mean.

standard error of population mean = $\sigma / \sqrt{n}$

Standard error of a proportion (srs)

The standard error of a percentage (proportion) is:

$se(prop) = \sqrt{[p(1-p)/n]}$

Confidence intervals

- Confidence intervals are usually expressed at the 95 per cent level (1.96 standard errors of sampling for a proportion)

Table: Effect of sample size on standard error

Size	se(p50)	lower 95 ci	upper 95 ci
100	0.050	0.402	0.598
200	0.035	0.431	0.569
500	0.022	0.456	0.544
1000	0.016	0.469	0.531
2000	0.011	0.478	0.522

Two stage samples

- The most efficient method is usually sampling at the first stage with probability proportional to size (pps).
- This produces a self-weighting sample.
- Easier logistics for administration.

Two stage samples

Stage 1

Primary sampling units (psu) are selected with a probability proportional to the size of the target population in the psu.

Example of psu: schools

Two stage samples

Stage 2

A random cluster of secondary sampling units (ssu) is selected at random from each of the psu.

Example of ssu: students in schools

Deff

- Two-stage sampling is less efficient than a simple random sample (srs) of the same size.

deff = (standard error of sampling for complex sample)^2 / (standard error of sampling for srs)^2

Deft

- The square root of deff is deft, which gives the ratio of the standard errors of sampling.

deft = (standard error of sampling for complex sample) / (standard error of sampling for srs)

Simple equivalent sample

- The simple equivalent sample (ses) is the size of a simple random sample which has the same standard error as the complex sample.
- We sometime use the term effective sample (n_{eff})

Table: Values for deff and simple equivalent sample

	psu	ssu	total	rho	deff	ses
1	50	20	1000	0.05	1.95	513
2	50	20	1000	0.10	2.90	345
3	50	20	1000	0.20	4.80	208
4	20	50	1000	0.05	3.45	290
5	20	50	1000	0.10	5.90	169
6	20	50	1000	0.20	10.80	93

ESTIMATE	DEFT	Clustered Sample Size Needed to equal SRS Sample of 1,000
% Strong Republicans	1.24	1,538
% Libras	1.03	1,061
Mean Years of Education	1.92	3,686
% Black	2.57	6,605

WEIGHTING

First-stage weights are typically inverse of selection probability

Second-stage weights provide additional precision, e.g., ratio-adjustment to known Census marginals

Weights may apply at several levels, e.g., household weights, person weights, etc.

SVYSET command for NHIS

- *SETUP for NHIS weighted analysis taking account of Complex Sample Design
- gen year_strata = (10000*year) + strata
- format year_strata %12.1g
- summarize year_strata
- svyset psu,[pweight=sampweight] strata(year_strata)

- **PLACE SVYSET COMMAND BEFORE THE FIRST svy: ANALYSIS**

Homework Assignment #3
Due by 11-59pm on Sunday. 9/20

- Tabulate BMI using CDC categories of BMI by AgeC
- FIRST, do analysis first ignoring the NHIS complex sample design
- SECOND, re-do the analysis using the **svyset** command and **svy:** prefix to take account of the NHIS complex sample design