Survey Data Analysis in Stata

Example

Real-world, publicly available survey data is often very complex (see the DHS example). Consequently, we will contrive an example for this tutorial, estimating p, the prevalence of a disease, say malaria, in a hypothetical country, called "Inventia".

Country profile:

Province	Population size	Number of districts
1	225,000	50
2	150,000	42
3	100,000	32
4	25,000	23
Total	500,000	146

In Inventia, the climate differs between provinces; for instance, province 4 is more arid and at a higher altitudes than the rest of the country. Consequently, the prevalence of malaria p differs between different provinces. Also, access to malaria prevention is not consistent across the country, and subsequently p may also vary somewhat between districts. (For instance, urban populations may have more resources to prevent malaria, and thus a lower prevalence.) The true prevalence of malaria in Inventia is 13.1%.

Today, we review how to analyze data from several different survey designs:

- **Simple Random Sampling** - We randomly sample 1,000 people from Inventia.
- **Stratified Sampling** - We randomly sample 250 people from each of the 4 provinces of Inventia.
- **Cluster Sampling** - We randomly sample 25 districts from Inventia and randomly sample 40 people within each district.
- **Stratified Cluster sampling** - For each of the 4 provinces, we randomly sample 5 districts. Within these 20 districts, we randomly sample 50 people.

Analyzing Survey Data in Stata

In order to analyze survey data in Stata, you must first `svyset` your data. This command tells Stata what survey design was used to obtain the data. This includes specification of survey weights, the finite population correction(s), and levels of clustering and stratification.

Once Stata has this information, it incorporates the specified design elements into its calculations. You can then use the survey estimation procedures in Stata. For example, `svy: mean var_name, svy: proportion var_name, svy: regress ...`.

Before analyzing your survey data, you need to be able to answer the following questions:

1. What is the design of my survey?
2. Am I using a finite population correction? At which stage of the design?
3. What are the survey weights used in the design?

Once you know these things, you can start analyzing your data in Stata.

1 Simple Random Sampling

Design: We randomly sample 1,000 people from the entire country of Inventia.

Notation:

- N is the total population size
- n is the number of individuals sampled from the population without replacement

In our case, $n = 1,000, N = 500,000$.

Finite Population Correction: $1 - f = \left(1 - \frac{n}{N}\right)$

Survey Weights $w_i = P($ individual i is included in the survey$)^{-1} = \frac{N}{n}$

Exercise: Estimate the prevalence of malaria in Inventia.

```
use "srs.dta", clear

generate weight_srs = pop_size/1000
generate fpc = 1000/pop_size * note that this does not match the definition above
svyset id [pweight=weight_srs], fpc(fpc)
svy: proportion malaria

svyset id [pweight=weight_srs]
svy: proportion malaria
estat effects, deff

proportion malaria
```

Under simple random sampling (SRS), when will `proportion malaria` and `svy: proportion malaria` give you the same results? Why?

Why does it not matter much if you use the finite population correction in this example?

Exercise: Estimate the prevalence of malaria in each of the four provinces.

```
svy, sub(if province==1): proportion malaria
svy, sub(if province==2): proportion malaria
svy, sub(if province==3): proportion malaria
svy, sub(if province==4): proportion malaria
```

Is there evidence of province-level variation in malaria prevalence?

2 Stratified Sampling

Design: We randomly sample 250 people from each of the 4 provinces of Inventia.

Notation:

- N is the total population size
- N_j is the population in province j, $j = \{1, 2, 3, 4\}$
- n_j individuals are sampled from province j

The important design question in stratified sampling is how to choose the sample size within each stratum. In our case, $N_1 = 225,000, N_2 = 150,000, N_3 = 100,000$ and $N_4 = 25,000$. $n_j = 250$ for each j.

Finite Population Correction: $1 - f_j = \left(1 - \frac{n_j}{N_j}\right)$

Survey Weights: $w_{ij} = P($ individual i in strata j is in the survey$)^{-1} = \frac{N_j}{n_j}$

Exercise: Estimate the prevalence of malaria in Inventia.

```
use "stratified.dta", clear

proportion malaria
proportion malaria, over(province)
generate weight_stratified = prov_size/250
generate fpc_stratified = 1/weight_stratified
svyset id [pweight=weight_stratified], strata(province) fpc(fpc_stratified)
svydescribe weight
svy: proportion malaria
estat effects, deff
```

Exercise: Why is our estimate of p too low when we do not specify the survey design?

3 Cluster Sampling

Design: We randomly sample 25 districts (clusters) from Inventia; within each district, we randomly sample 40 people.

Notation:

- N is the total population size
- N_k is the population size in district k, $k = \{1, ..., 146\}$
- n_I out of N_I total districts are sampled for inclusion in the survey (primary sampling unit)
- We sample n_k individuals in district k are selected for inclusion in the survey (secondary sampling unit)

In our survey, $n_I = 25$, $N_I = 146$, $n_k = 40$, and N_k is the population size in district k.

Finite Population Correction:

Stage I: $1 - f_I = \left(1 - \frac{n_I}{N_I}\right)$

Stage II: $1 - f_k = \left(1 - \frac{n_k}{N_k}\right)$

Survey Weights':

$$\begin{aligned} w_{ik} &= P(\text{individual } i \text{ in cluster } k \text{ is in the survey})^{-1} \\ &= [P(\text{cluster } k \text{ selected}) * P(\text{ individual } i \text{ in cluster } k \text{ selected} \mid \text{cluster} k \text{ selected})]^{-1} \\ &= \frac{N_I}{n_I} * \frac{N_k}{n_k} \end{aligned}$$

Exercise: Estimate the prevalence of malaria in Inventia, using only the first stage finite population correction.

```
use "cluster.dta", clear

generate fpc1 = 25/146
generate fpc2 = 40/districtsize
generate weight_cluster = (fpc1*fpc2)^-1
svyset district [pweight=weight_cluster], fpc(fpc1) || id, fpc(fpc2)
svy: proportion malaria
estat effects, deff
```

4 Stratified Cluster Sampling

We could combine stratified, cluster and simple random sampling all into one design!

Design: For each of the 4 provinces, we randomly sample 5 districts. Within each of the 20 districts, we randomly sample 50 people.

Survey weights: As an example, for province 2:

$$\begin{aligned} P(\text{person } i \text{ in district } & j \text{ in province } 2 \text{ in survey }) \\ &= P(\text{district } j \text{ in survey} \mid \text{province } 2) P(\text{person } i \text{ in survey} \mid \text{district } j) \\ &= \frac{5}{42} * \frac{50}{districtsize_j} \end{aligned}$$

Finite population correction:

Stage I: $\frac{\#\text{sampled districts}}{\text{total}\#\text{districts in the province}}$

Stage II: $\frac{\#\text{sampled per district}}{\text{district population}} = \frac{50}{\text{districtsize}_j}$ for district j.

Exercise: Estimate the prevalence of malaria in Inventia.

```
use "stratifiedcluster.dta", clear

generate fpc1 = 5/ndistrict
generate fpc2 = 50/districtsize
generate weight_stratcluster = (fpc1*fpc2)^-1
svyset district [pweight=weight_stratcluster], fpc(fpc1) strata(province) || id, fpc(fpc2)
svy: proportion malaria
estat effects, deff
```