Subpopulation estimation focuses on part of the population. This entry discusses subpopulation estimation and explains why you should use the `subpop()` option instead of `if` and `in` for your survey data analysis.

Subpopulation estimation involves computing point and variance estimates for part of the population. This is not the same as restricting the estimation sample to the collection of observations within the subpopulation because variance estimation for survey data measures sample-to-sample variability, assuming that the same survey design is used to collect the data; see Methods and formulas for a detailed explanation. West, Berglund, and Heeringa (2008) provides further information on subpopulation analysis.

The `svy` prefix command’s `subpop()` option performs subpopulation estimation. The `svy: mean`, `svy: proportion`, `svy: ratio`, and `svy: total` commands also have the `over()` option to perform estimation for multiple subpopulations.

The following examples illustrate how to use the `subpop()` and `over()` options.

### Example 1

Suppose that we are interested in estimating the proportion of women in our population who have had a heart attack. In our NHANES II dataset (McDowell et al. 1981), the female participants can be identified using the `female` variable, and the `heartatk` variable indicates whether an individual has ever had a heart attack. Below we use `svy: mean` with the `heartatk` variable to estimate the proportion of individuals who have had a heart attack, and we use `subpop(female)` to identify our subpopulation of interest.

```stata
use https://www.stata-press.com/data/r16/nhanes2d
.svy, subpop(female): mean heartatk
(running mean on estimation sample)
Survey: Mean estimation
Number of strata = 31 Number of obs = 10,349
Number of PSUs = 62 Population size = 117,131,111
Subpop. no. obs = 5,434 Subpop. size = 60,971,631
Design df = 31

Linearized
        Mean Std. Err.     [95% Conf. Interval]
heartatk  .0193276  .0017021  .0158562   .0227991
```

1
The `subpop(varname)` option takes a 0/1 variable, and the subpopulation of interest is defined by `varname = 1`. All other members of the sample not in the subpopulation are indicated by `varname = 0`.

If a person’s subpopulation status is unknown, `varname` should be set to missing (.), so those observations will be omitted from the analysis. For instance, in the preceding analysis, if a person’s sex was not recorded, `female` should be coded as missing rather than as male (`female = 0`).

### Technical note

Actually, the `subpop(varname)` option takes a zero/nonzero variable, and the subpopulation is defined by `varname ≠ 0` and not missing. All other members of the sample not in the subpopulation are indicated by `varname = 0`, but 0, 1, and missing are typically the only values used for the `subpop()` variable.

Furthermore, you can specify an `if` qualifier within `subpop()` to identify a subpopulation. The result is the same as generating a variable equal to the conditional expression and supplying it as the `subpop()` variable. If a `varname` and an `if` qualifier are specified within the `subpop()` option, the subpopulation is identified by their logical conjunction (logical `and`), and observations with missing values in either are dropped from the estimation sample.

### Example 2: Multiple subpopulation estimation

Means, proportions, ratios, and totals for multiple subpopulations can be estimated using the `over()` option with `svy: mean`, `svy: proportion`, `svy: ratio`, and `svy: total`, respectively. Here is an example using the NMIHS data (Gonzalez, Krauss, and Scott 1992), estimating mean birthweight over the categories of the `race` variable.

```
. use https://www.stata-press.com/data/r16/nmihs
. svy: mean birthwgt, over(race)
(running mean on estimation sample)
Survey: Mean estimation
Number of strata =  6 Number of obs =  9,946
Number of PSUs   =  9,946 Population size =  3,895,562
Design df        =  9,940

Linearized

  Mean   Std. Err.     [95% Conf. Interval]
---------  -----------  ----------------------
c.birthwgt@race
  nonblack  3402.32  7.609532   3387.404   3417.236
  black    3127.834  6.529814   3115.035   3140.634
```


More than one variable can be used in the `over()` option.

```
. svy: mean birthwgt, over(race marital)
(running mean on estimation sample)
```

Survey: Mean estimation

<table>
<thead>
<tr>
<th></th>
<th>Linearized Mean</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>c.birthwgt@race#marital</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>nonblack#single</td>
<td>3291.045</td>
<td>20.18795</td>
<td>3251.472 3330.617</td>
</tr>
<tr>
<td>nonblack#married</td>
<td>3426.407</td>
<td>8.379497</td>
<td>3409.982 3442.833</td>
</tr>
<tr>
<td>black#single</td>
<td>3073.122</td>
<td>8.752553</td>
<td>3055.965 3090.279</td>
</tr>
<tr>
<td>black#married</td>
<td>3221.616</td>
<td>12.42687</td>
<td>3197.257 3245.975</td>
</tr>
</tbody>
</table>

Here the `race` and `marital` variables have value labels. `race` has the value 0 labeled “nonblack” (that is, white and other) and 1 labeled “black”; `marital` has the value 0 labeled “single” and 1 labeled “married”. Value labels on the `over()` variables make for a more informative legend above the table of point estimates. See [U] 12.6.3 Value labels for information on creating value labels.

We can also combine the `subpop()` option with the `over()` option.

```
. generate nonblack = (race == 0) if !missing(race)
. svy, subpop(nonblack): mean birthwgt, over(marital age20)
(running mean on estimation sample)
```

Survey: Mean estimation

<table>
<thead>
<tr>
<th></th>
<th>Linearized Mean</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>c.birthwgt@marital#age20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>single#age20+</td>
<td>3312.012</td>
<td>24.2869</td>
<td>3264.398 3359.625</td>
</tr>
<tr>
<td>single#age&lt;20</td>
<td>3244.709</td>
<td>36.85934</td>
<td>3172.448 3316.971</td>
</tr>
<tr>
<td>married#age20+</td>
<td>3434.923</td>
<td>8.674633</td>
<td>3417.916 3451.929</td>
</tr>
<tr>
<td>married#age&lt;20</td>
<td>3287.301</td>
<td>34.15988</td>
<td>3220.332 3354.271</td>
</tr>
</tbody>
</table>

Note: 3 strata omitted because they contain no subpopulation members.

This time, we estimated means for the marital status and age (<20 or ≥20) subpopulations for `race == 0` (nonblack) only. We carefully define `nonblack` so that it is missing when `race` is missing. If we omitted the `if !missing(race)` in our `generate` statement, then `nonblack` would be 0 when `race` was missing. This would improperly assume that all individuals with a missing value for `race` were black and could cause our results to have incorrect standard errors. The standard errors could be incorrect because those observations for which `race` is missing would be counted as part of the estimation sample, potentially inflating the number of PSUs used in the formula for the variance estimator. For this reason, observations with missing values for any of the `over()` variables are omitted from the analysis.
Methods and formulas

The following discussion assumes that you are already familiar with the topics discussed in [SVY] Variance estimation.

Cochran (1977, sec. 2.13) discusses a method by which you can derive estimates for subpopulation totals. This section uses this method to derive the formulas for a subpopulation total from a simple random sample (without replacement) to explain how the subpop() option works, shows why this method will often produce different results from those produced using an equivalent if (or in) qualifier (outside the subpop() option), and discusses how this method applies to subpopulation means, proportions, ratios, and regression models.

Methods and formulas are presented under the following headings:

- Subpopulation totals
- Subpopulation estimates other than the total
- Subpopulation with replication methods

Subpopulation totals

Let $Y_j$ be a survey item for individual $j$ in the population, where $j = 1, \ldots, N$ and $N$ is the population size. Let $S$ be a subset of individuals in the population and $I_S(j)$ indicate if the $j$th individual is in $S$, where

$$I_S(j) = \begin{cases} 
1, & \text{if } j \in S \\
0, & \text{otherwise}
\end{cases}$$

The subpopulation total is

$$Y_S = \sum_{j=1}^{N} I_S(j) Y_j$$

and the subpopulation size is

$$N_S = \sum_{j=1}^{N} I_S(j)$$

Let $y_j$ be the items for those individuals selected in the sample, where $j = 1, \ldots, n$ and $n$ is the sample size. The number of individuals sampled from the subpopulation is

$$n_S = \sum_{j=1}^{n} I_S(j)$$

The estimator for the subpopulation total is

$$\hat{Y}_S = \sum_{j=1}^{n} I_S(j) w_j y_j \quad (1)$$
where \( w_j = N/n \) is the unadjusted sampling weight for this design. The estimator for \( N_S \) is

\[
\hat{N}_S = \sum_{j=1}^{n} I_S(j)w_j
\]

The replicate values for the BRR and jackknife variance estimators are computed using the same method.

The linearized variance estimator for \( \hat{Y}_S \) is

\[
\hat{V}(\hat{Y}_S) = \left(1 - \frac{n}{N}\right) \frac{n}{n-1} \sum_{j=1}^{n} \left\{ I_S(j)w_jy_j - \frac{1}{n} \hat{Y}_S \right\}^2
\]

The covariance estimator for the subpopulation totals \( \hat{Y}_S \) and \( \hat{X}_S \) (notation for \( X_S \) is defined similarly to that of \( Y_S \)) is

\[
\hat{Cov}(\hat{Y}_S, \hat{X}_S) = \left(1 - \frac{n}{N}\right) \frac{n}{n-1} \sum_{j=1}^{n} \left\{ I_S(j)w_jy_j - \frac{1}{n} \hat{Y}_S \right\} \left\{ I_S(j)w_jx_j - \frac{1}{n} \hat{X}_S \right\}
\]

Equation (2) is not the same formula that results from restricting the estimation sample to the observations within \( S \). The formula using this restricted sample (assuming a `svyset` with the corresponding FPC) is

\[
\tilde{V}(\hat{Y}_S) = \left(1 - \frac{n_S}{N_S}\right) \frac{n_S}{n_S-1} \sum_{j=1}^{n} I_S(j) \left\{ w_jy_j - \frac{1}{n_S} \hat{Y}_S \right\}^2
\]

These variance estimators, (2) and (4), assume two different survey designs. In (2), \( n \) individuals are sampled without replacement from the population comprising the \( N_S \) values from the subpopulation with \( N - N_S \) additional zeros. In (4), \( n_S \) individuals are sampled without replacement from the subpopulation of \( N_S \) values. We discourage using (4) by warning against using the `if` and `in` qualifiers for subpopulation estimation because this variance estimator does not accurately measure the sample-to-sample variability of the subpopulation estimates for the survey design that was used to collect the data.

For survey data, there are only a few circumstances that require using the `if` qualifier. For example, if you suspected laboratory error for a certain set of measurements, then using the `if` qualifier to omit these observations from the analysis might be proper.

### Subpopulation estimates other than the total

To generalize the above results, note that the other point estimators—such as means, proportions, ratios, and regression coefficients—yield a linearized variance estimator based on one or more (equation level) score variables. For example, the weighted sample estimation equations of a regression model for a given subpopulation (see (3) from [SVY] Variance estimation) is

\[
\hat{G}(\beta_S) = \sum_{j=1}^{n} I_S(j)w_jS(\beta_S; y_j, x_j) = 0
\]

You can write \( \hat{G}(\beta_S) \) as

\[
\hat{G}(\beta_S) = \sum_{j=1}^{n} I_S(j)w_jd_j
\]

which is an estimator for the subpopulation total \( G(\beta_S) \), so its variance estimator can be computed using the design-based variance estimator for a subpopulation total.
Subpopulation with replication methods

The above comparison between the variance estimator from the `subpop()` option and the variance estimator from the `if` and `in` qualifiers is also true for the replication methods.

For the BRR method, the same number of replicates is produced with or without the `subpop()` option. The difference is how the replicate values are computed. Using the `if` and `in` qualifiers may cause an error because `svy brr` checks that there are two PSUs in every stratum within the restricted sample.

For the jackknife method, every PSU produces a replicate, even if it does not contain an observation within the subpopulation specified using the `subpop()` option. When the `if` and `in` qualifiers are used, only the PSUs that have at least 1 observation within the restricted sample will produce a replicate.

For methods using replicate weight variables, every weight variable produces a replicate, even if it does not contain an observation within the subpopulation specified using the `subpop()` option. When the `if` and `in` qualifiers are used, only the PSUs that have at least 1 observation within the restricted sample will produce a replicate.

References


Also see

[SVY] Survey — Introduction to survey commands
[SVY] svy — The survey prefix command
[SVY] svy postestimation — Postestimation tools for svy
[SVY] svyset — Declare survey design for dataset