Title

stata.com

Direct standardization — Direct standardization of means, proportions, and ratios

Description

Direct standardization is an estimation method that allows comparing rates that come from different frequency distributions. The `mean`, `proportion`, and `ratio` commands can estimate means, proportions, and ratios by using direct standardization.

See [SVY] **Calibration** and [SVY] **Poststratification** for similar estimation methods when some population totals are known.

Remarks and examples

stata.com

In direct standardization, estimated rates (means, proportions, and ratios) are adjusted according to the frequency distribution of a standard population. The standard population is partitioned into categories, called standard strata. The stratum frequencies for the standard population are called standard weights. In the standardizing frequency distribution, the standard strata are most commonly identified by demographic information such as age, sex, and ethnicity.

Stata's `mean`, `proportion`, and `ratio` estimation commands have options for estimating means, proportions, and ratios by using direct standardization. The `stdize()` option takes a variable that identifies the standard strata, and the `stdweight()` option takes a variable that contains the standard weights.

The standard strata (specified using `stdize()`) from the standardizing population are not the same as the strata (specified using `svyset`'s `strata()` option) from the sampling design. In the output header, "Number of strata" is the number of strata in the first stage of the sampling design, and "N. of std strata" is the number of standard strata.

In the following example, we use direct standardization to compare the death rates between two districts of London in 1840.

▷ Example 1: Standardized rates

Table 3.12-6 of Korn and Graubard (1999, 156) contains enumerated data for two districts of London for the years 1840–1841. The `age` variable identifies the age groups in 5-year increments, `bgliving` contains the number of people living in the Bethnal Green district at the beginning of 1840, `bgdeaths` contains the number of people who died in Bethnal Green that year, `hsliving` contains the number of people living in St. George's Hanover Square at the beginning of 1840, and `hsdeaths` contains the number of people who died in Hanover Square that year.

```
. use https://www.stata-press.com/data/r17/stdize
. list, noobs sep(0) sum
```

	age	bgliving	bgdeaths	hsliving	hsdeaths
	0-5	10739	850	5738	463
	5-10	9180	76	4591	55
	10-15	8006	38	4148	28
	15-20	7096	37	6168	36
	20-25	6579	38	9440	68
	25-30	5829	51	8675	78
	30-35	5749	51	7513	64
	35-40	4490	56	5091	78
	40-45	4385	47	4930	85
	45-50	2955	66	2883	66
	50-55	2995	74	2711	77
	55-60	1644	67	1275	55
	60-65	1835	64	1469	61
	65-70	1042	64	649	55
	70-75	879	68	619	58
	75-80	366	47	233	51
	80-85	173	39	136	20
	85-90	71	22	48	15
	90-95	21	6	10	4
	95-100	4	2	2	1
	unknown	50	1	124	0
Sum		74088	1764	66453	1418

We can use `svy: ratio` to compute the deathrates for each district in 1840. Because this dataset is identified as census data, we will create an FPC variable that will contain a sampling rate of 100%. This method will result in zero standard errors, which are interpreted to mean no variability—appropriate because our point estimates came from the entire population.

```
. generate fpc = 1
. svyset, fpc(fpc)
Sampling weights: <none>
             VCE: linearized
     Single unit: missing
        Strata 1: <one>
 Sampling unit 1: <observations>
           FPC 1: fpc
. svy: ratio (Bethnal: bgdeaths/bgliving) (Hanover: hsdeaths/hsliving)
(running ratio on estimation sample)
Survey: Ratio estimation
Number of strata =  1                     Number of obs   = 21
Number of PSUs   = 21                     Population size = 21
                                          Design df       = 20
      Bethnal: bgdeaths/bgliving
      Hanover: hsdeaths/hsliving
```

	Ratio	Linearized std. err.	[95% conf.	interval]
Bethnal	.0238095	0	.	.
Hanover	.0213384	0	.	.

```
Note: Zero standard errors because of 100% sampling rate
      detected for FPC in the first stage.
```

The deathrates are 2.38% for Bethnal Green and 2.13% for St. George's Hanover Square. These observed deathrates are not really comparable because they come from two different age distributions. We can standardize based on the age distribution from Bethnal Green. Here `age` identifies our standard strata and `bgliving` contains the associated population sizes.

```
. svy: ratio (Bethnal: bgdeaths/bgliving) (Hanover: hsdeaths/hsliving),
> stdize(age) stdweight(bgliving)
(running ratio on estimation sample)

Survey: Ratio estimation

Number of strata =  1                    Number of obs    = 21
Number of PSUs   = 21                    Population size = 21
N. of std strata = 21                    Design df       = 20

      Bethnal: bgdeaths/bgliving
      Hanover: hsdeaths/hsliving
```

	Ratio	Linearized std. err.	[95% conf.	interval]
Bethnal	.0238095	0	.	.
Hanover	.0266409	0	.	.

```
Note: Zero standard errors because of 100% sampling rate
      detected for FPC in the first stage.
```

The standardized deathrate for St. George's Hanover Square, 2.66%, is larger than the deathrate for Bethnal Green.

For this example, we could have used `dstdize` to compute the deathrates; however, `dstdize` will not compute the correct standard errors for survey data. Furthermore, `dstdize` is not an estimation command, so `test` and the other postestimation commands are not available.

◁

❑ Technical note

The values in the variable supplied to the `stdweight()` option are normalized so that (1) is true; see *Methods and formulas*. Thus the `stdweight()` variable can contain either population sizes or population proportions for the associated standard strata.

❑

Methods and formulas

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

In direct standardization, a weighted sum of the point estimates from the standard strata is used to produce an overall point estimate for the population. This section will show how direct standardization affects the ratio estimator. The mean and proportion estimators are special cases of the ratio estimator.

Suppose that you used a complex survey design to sample m individuals from a population of size M. Let D_g be the set of individuals in the sample that belong to the gth standard stratum, and let $I_{D_g}(j)$ indicate if the jth individual is in standard stratum g, where

$$I_{D_g}(j) = \begin{cases} 1, \text{ if } j \in D_g \\ 0, \text{ otherwise} \end{cases}$$

Also let L_D be the number of standard strata, and let π_g be the proportion of the population that belongs to standard stratum g.

$$\sum_{g=1}^{L_D} \pi_g = 1 \tag{1}$$

In subpopulation estimation, π_g is set to zero if none of the individuals in standard stratum g are in the subpopulation. Then the standard stratum proportions are renormalized.

Let y_j and x_j be the items of interest and w_j be the sampling weight for the jth sampled individual. The estimator for the standardized ratio of $R = Y/X$ is

$$\widehat{R}^D = \sum_{g=1}^{L_D} \pi_g \frac{\widehat{Y}_g}{\widehat{X}_g}$$

where

$$\widehat{Y}_g = \sum_{j=1}^{m} I_{D_g}(j)\, w_j y_j$$

with $\widehat{X}_g$ similarly defined.

For replication-based variance estimation, replicates of the standardized values are used in the variance formulas.

The score variable for the linearized variance estimator of the standardized ratio is

$$z_j(\widehat{R}^D) = \sum_{g=1}^{L_D} \pi_g I_{D_g}(j) \frac{\widehat{X}_g y_j - \widehat{Y}_g x_j}{\widehat{X}_g^2}$$

This score variable was derived using the method described in [SVY] **Variance estimation** and is a direct result of the methods described in Deville (1999), Demnati and Rao (2004), and Shah (2004).

For the `mean` and `proportion` commands, the mean estimator is a ratio estimator with the denominator variable equal to one ($x_j = 1$) and the proportion estimator is the mean estimator with an indicator variable in the numerator ($y_j \in \{0, 1\}$).

References

Demnati, A., and J. N. K. Rao. 2004. Linearization variance estimators for survey data. *Survey Methodology* 30: 17–26.

Deville, J.-C. 1999. Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology* 25: 193–203.

Korn, E. L., and B. I. Graubard. 1999. *Analysis of Health Surveys*. New York: Wiley.

Shah, B. V. 2004. Comment [on Demnati and Rao (2004)]. *Survey Methodology* 30: 29.

Also see

[SVY] **Survey** — Introduction to survey commands

[SVY] **svy** — The survey prefix command

[SVY] **svyset** — Declare survey design for dataset

[SVY] **Calibration** — Calibration for survey data

[SVY] **Poststratification** — Poststratification for survey data