mean — Estimate means

Description | Quick start | Menu | Syntax
Options | Remarks and examples | Stored results | Methods and formulas
References | Also see

Description

`mean` produces estimates of means, along with standard errors.

Quick start

Mean, standard error, and 95% confidence interval for `v1`

```
mean v1
```

Also compute statistics for `v2`

```
mean v1 v2
```

Same as above, but for each level of categorical variable `catvar1`

```
mean v1 v2, over(catvar1)
```

Weighting by probability weight `wvar`

```
mean v1 v2 [pweight=wvar]
```

Population mean using `svyset` data

```
svy: mean v3
```

Subpopulation means for each level of categorical variable `catvar2` using `svyset` data

```
svy: mean v3, over(catvar2)
```

Test equality of two subpopulation means

```
svy: mean v3, over(catvar2)
test v3@1.catvar2 = v3@2.catvar2
```

Menu

Statistics > Summaries, tables, and tests > Summary and descriptive statistics > Means

Syntax

`mean` *varlist* [*if*] [*in*] [*weight*] [, *options*]

options	Description
Model	
`stdize(`*varname*`)`	variable identifying strata for standardization
`stdweight(`*varname*`)`	weight variable for standardization
`nostdrescale`	do not rescale the standard weight variable
if/in/over	
`over(`*varlist*$_o$`)`	group over subpopulations defined by *varlist*$_o$
SE/Cluster	
`vce(`*vcetype*`)`	*vcetype* may be `analytic`, `cluster` *clustvar*, `bootstrap`, or `jackknife`
Reporting	
`level(#)`	set confidence level; default is `level(95)`
`noheader`	suppress table header
display_options	control column formats, line width, display of omitted variables and base and empty cells, and factor-variable labeling
`coeflegend`	display legend instead of statistics

varlist may contain factor variables; see [U] **11.4.3 Factor variables**.

`bayesboot`, `bootstrap`, `collect`, `jackknife`, `mi estimate`, `rolling`, `statsby`, and `svy` are allowed; see [U] **11.1.10 Prefix commands**.

`vce(bootstrap)` and `vce(jackknife)` are not allowed with the `mi estimate` prefix; see [MI] **mi estimate**.

Weights are not allowed with the `bootstrap` prefix; see [R] **bootstrap**.

`aweight`s are not allowed with the `jackknife` prefix; see [R] **jackknife**.

`vce()` and weights are not allowed with the `svy` prefix; see [SVY] **svy**.

`fweight`s, `aweight`s, `iweight`s, and `pweight`s are allowed; see [U] **11.1.6 weight**.

`coeflegend` does not appear in the dialog box.

See [U] **20 Estimation and postestimation commands** for more capabilities of estimation commands.

Options

Model

`stdize(`*varname*`)` specifies that the point estimates be adjusted by direct standardization across the strata identified by *varname*. This option requires the `stdweight()` option.

`stdweight(`*varname*`)` specifies the weight variable associated with the standard strata identified in the `stdize()` option. The standardization weights must be constant within the standard strata.

`nostdrescale` prevents the standardization weights from being rescaled within the `over()` groups. This option requires `stdize()` but is ignored if the `over()` option is not specified.

if/in/over

`over(`*varlist*$_o$`)` specifies that estimates be computed for multiple subpopulations, which are identified by the different values of the variables in *varlist*$_o$. Only numeric, nonnegative, integer-valued variables are allowed in `over(`*varlist*$_o$`)`.

SE/Cluster

vce(*vcetype*) specifies the type of standard error reported, which includes types that are derived from asymptotic theory (analytic), that allow for intragroup correlation (cluster *clustvar*), and that use bootstrap or jackknife methods (bootstrap, jackknife); see [R] ***vce_option***.

vce(analytic), the default, uses the analytically derived variance estimator associated with the sample mean.

Reporting

level(#); see [R] **Estimation options**.

noheader prevents the table header from being displayed.

display_options: noomitted, vsquish, noemptycells, baselevels, allbaselevels, nofvlabel, fvwrap(#), fvwrapon(*style*), cformat(%*fmt*), and nolstretch; see [R] **Estimation options**.

The following option is available with mean but is not shown in the dialog box:

coeflegend; see [R] **Estimation options**.

Remarks and examples

▷ Example 1

Using the fuel data from example 3 of [R] **ttest**, we estimate the average mileage of the cars without the fuel treatment (mpg1) and those with the fuel treatment (mpg2).

```
. use https://www.stata-press.com/data/r19/fuel
. mean mpg1 mpg2
Mean estimation                          Number of obs = 12
```

	Mean	Std. err.	[95% conf.	interval]
mpg1	21	.7881701	19.26525	22.73475
mpg2	22.75	.9384465	20.68449	24.81551

Using these results, we can test the equality of the mileage between the two groups of cars.

```
. test mpg1 = mpg2
 ( 1)  mpg1 - mpg2 = 0
       F(  1,    11) =    5.04
            Prob > F =    0.0463
```

◁

▷ Example 2

In example 1, the joint observations of mpg1 and mpg2 were used to estimate a covariance between their means.

```
. matrix list e(V)

symmetric e(V)[2,2]
             mpg1        mpg2
mpg1   .62121212
mpg2    .4469697   .88068182
```

If the data were organized this way out of convenience but the two variables represent independent samples of cars (coincidentally of the same sample size), we should reshape the data and use the over() option to ensure that the covariance between the means is zero.

```
. use https://www.stata-press.com/data/r19/fuel

. stack mpg1 mpg2, into(mpg) clear

. rename _stack trt

. label define trt_lab 1 "without" 2 "with"

. label values trt trt_lab

. label var trt "Fuel treatment"

. mean mpg, over(trt)

Mean estimation                            Number of obs = 24

─────────────────────────────────────────────────────────────
             │       Mean   Std. err.     [95% conf. interval]
─────────────┼───────────────────────────────────────────────
   c.mpg@trt │
     without │         21    .7881701      19.36955    22.63045
        with │      22.75    .9384465      20.80868    24.69132
─────────────────────────────────────────────────────────────

. matrix list e(V)

symmetric e(V)[2,2]
                 c.mpg@      c.mpg@
                 1.trt       2.trt
c.mpg@1.trt   .62121212
c.mpg@2.trt           0   .88068182
```

Now, we can test the equality of the mileage between the two independent groups of cars.

```
. test mpg@1.trt = mpg@2.trt

 ( 1)  c.mpg@1bn.trt - c.mpg@2.trt = 0

       F(  1,    23) =    2.04
            Prob > F =    0.1667
```

◁

▷ Example 3: standardized means

Suppose that we collected the blood pressure data from example 2 of [R] **dstdize**, and we wish to obtain standardized high blood pressure rates for each city in 1990 and 1992, using, as the standard, the age, sex, and race distribution of the four cities and two years combined. Our rate is really the mean of a variable that indicates whether a sampled individual has high blood pressure. First, we generate the strata and weight variables from our standard distribution, and then use `mean` to compute the rates.

```
. use https://www.stata-press.com/data/r19/hbp, clear
. egen strata = group(age race sex) if inlist(year, 1990, 1992)
(675 missing values generated)
. by strata, sort: gen stdw = _N
. mean hbp, over(city year) stdize(strata) stdweight(stdw)
Mean estimation
N. of std strata = 24                    Number of obs = 455
```

	Mean	Std. err.	[95% conf.	interval]
c.hbp@city#year				
1 1990	.058642	.0296273	.0004182	.1168657
1 1992	.0117647	.0113187	-.0104789	.0340083
2 1990	.0488722	.0238958	.0019121	.0958322
2 1992	.014574	.007342	.0001455	.0290025
3 1990	.1011211	.0268566	.0483425	.1538998
3 1992	.0810577	.0227021	.0364435	.1256719
5 1990	.0277778	.0155121	-.0027066	.0582622
5 1992	.0548926	0	.	.

The standard error of the high blood pressure rate estimate is missing for city 5 in 1992 because there was only one individual with high blood pressure; that individual was the only person observed in the stratum of white males 30–35 years old.

By default, `mean` rescales the standard weights within the `over()` groups. In the following, we use the `nostdrescale` option to prevent this, thus reproducing the results in [R] **dstdize**.

```
. mean hbp, over(city year) stdize(strata) stdweight(stdw) nostdrescale
Mean estimation
N. of std strata = 24                    Number of obs = 455
```

	Mean	Std. err.	[95% conf.	interval]
c.hbp@city#year				
1 1990	.0073302	.0037034	.0000523	.0146082
1 1992	.0015432	.0014847	-.0013745	.004461
2 1990	.0078814	.0038536	.0003084	.0154544
2 1992	.0025077	.0012633	.000025	.0049904
3 1990	.0155271	.0041238	.007423	.0236312
3 1992	.0081308	.0022772	.0036556	.012606
5 1990	.0039223	.0021904	-.0003822	.0082268
5 1992	.0088735	0	.	.

◁

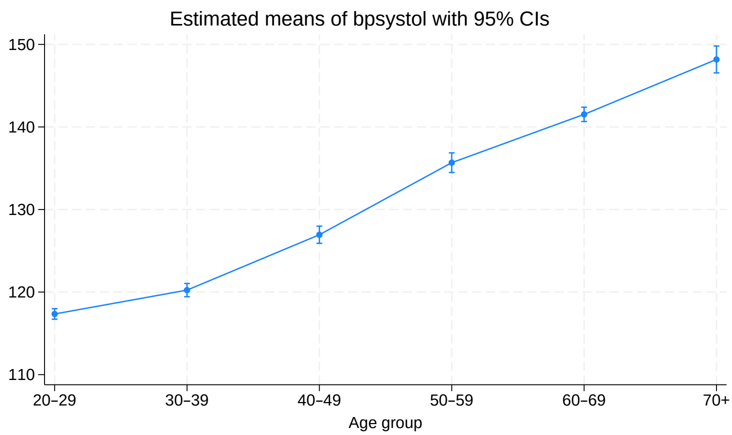
▷ Example 4: profile plots and contrasts

The first example in [R] **marginsplot** shows how to use `margins` and `marginsplot` to get profile plots from a linear regression. We can similarly explore the data using `marginsplot` after `mean` with the `over()` option. Here we use `marginsplot` to plot the means of systolic blood pressure for each age group.

```
. use https://www.stata-press.com/data/r19/nhanes2, clear
. mean bpsystol, over(agegrp)
Mean estimation                          Number of obs = 10,351
```

	Mean	Std. err.	[95% conf.	interval]
c.bpsystol@agegrp				
20–29	117.3466	.3247329	116.71	117.9831
30–39	120.2374	.4095845	119.4345	121.0402
40–49	126.9442	.532033	125.9013	127.9871
50–59	135.6754	.6061842	134.4872	136.8637
60–69	141.5227	.4433527	140.6537	142.3918
70+	148.1765	.8321116	146.5454	149.8076

```
. marginsplot
Variables that uniquely identify means: agegrp
```



We see that the mean systolic blood pressure increases with age. We can use `contrast` to formally test whether each mean is different from the mean in the previous age group using the `ar.` contrast operator; see [R] **contrast** for more information on this command.

```
. contrast ar.agegrp#c.bpsystol, effects nowald
Contrasts of means
```

	Contrast	Std. err.	t	P>|t|	[95% conf.	interval]
agegrp#c.bpsystol						
(30-39 vs 20-29)	2.89081	.5226958	5.53	0.000	1.866225	3.915394
(40-49 vs 30-39)	6.706821	.6714302	9.99	0.000	5.390688	8.022954
(50-59 vs 40-49)	8.731263	.8065472	10.83	0.000	7.150275	10.31225
(60-69 vs 50-59)	5.847282	.7510133	7.79	0.000	4.375151	7.319413
(70+ vs 60-69)	6.653743	.9428528	7.06	0.000	4.80557	8.501917

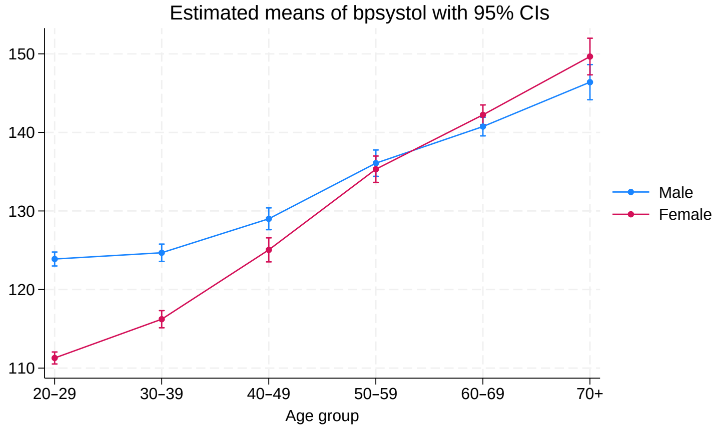
The first row of the output reports that the mean systolic blood pressure for the 30–39 age group is 2.89 higher than the mean for the 20–29 age group. The mean for the 40–49 age group is 6.71 higher than the mean for the 30–39 age group, and so on. Each of these differences is significantly different from zero.

We can include both `agegrp` and `sex` in the `over()` option to estimate means separately for men and women in each age group.

```
. mean bpsystol, over(agegrp sex)
Mean estimation                                  Number of obs = 10,351
```

	Mean	Std. err.	[95% conf.	interval]
c.bpsystol@agegrp#sex				
20-29#Male	123.8862	.4528516	122.9985	124.7739
20-29#Female	111.2849	.3898972	110.5206	112.0492
30-39#Male	124.6818	.5619855	123.5802	125.7834
30-39#Female	116.2207	.5572103	115.1284	117.3129
40-49#Male	129.0033	.7080788	127.6153	130.3912
40-49#Female	125.0468	.7802558	123.5174	126.5763
50-59#Male	136.0864	.855435	134.4096	137.7632
50-59#Female	135.3164	.8556015	133.6393	136.9935
60-69#Male	140.7451	.6059786	139.5572	141.9329
60-69#Female	142.2368	.6427981	140.9767	143.4968
70+#Male	146.3951	1.141126	144.1583	148.6319
70+#Female	149.6599	1.189975	147.3273	151.9924

```
. marginsplot
Variables that uniquely identify means: agegrp sex
```



Are the means different for men and women within each age group? We can again perform the tests using `contrast`. This time, we will use `r.sex` to obtain contrasts comparing men and women and use `@agegrp` to request that the tests are performed for each age group.

```
. contrast r.sex#c.bpsystol@agegrp, effects nowald
Contrasts of means
```

	Contrast	Std. err.	t	P>|t|	[95% conf.	interval]
sex@agegrp#c.bpsystol						
(Female vs Male) 20–29	-12.60132	.5975738	-21.09	0.000	-13.77268	-11.42996
(Female vs Male) 30–39	-8.461161	.7913981	-10.69	0.000	-10.01245	-6.909868
(Female vs Male) 40–49	-3.956451	1.053648	-3.76	0.000	-6.021805	-1.891097
(Female vs Male) 50–59	-.7699782	1.209886	-0.64	0.525	-3.141588	1.601631
(Female vs Male) 60–69	1.491684	.8834022	1.69	0.091	-.2399545	3.223323
(Female vs Male) 70+	3.264762	1.648699	1.98	0.048	.0329927	6.496531

Using a 0.05 significance level, we find that the mean systolic blood pressure is different for men and women in all age groups except the fifties and sixties.

◁

Video example

Descriptive statistics in Stata

Stored results

`mean` stores the following in `e()`:

Scalars

`e(N)`	number of observations
`e(N_over)`	number of subpopulations
`e(N_stdize)`	number of standard strata
`e(N_clust)`	number of clusters
`e(k_eq)`	number of equations in `e(b)`
`e(df_r)`	sample degrees of freedom
`e(rank)`	rank of `e(V)`

Macros

`e(cmd)`	`mean`
`e(cmdline)`	command as typed
`e(varlist)`	*varlist*
`e(stdize)`	*varname* from `stdize()`
`e(stdweight)`	*varname* from `stdweight()`
`e(wtype)`	weight type
`e(wexp)`	weight expression
`e(title)`	title in estimation output
`e(clustvar)`	name of cluster variable
`e(over)`	*varlist* from `over()`
`e(vce)`	*vcetype* specified in `vce()`
`e(vcetype)`	title used to label Std. err.
`e(properties)`	`b V`
`e(estat_cmd)`	program used to implement `estat`
`e(marginsnotok)`	predictions disallowed by `margins`

Matrices

`e(b)`	vector of mean estimates
`e(V)`	(co)variance estimates
`e(sd)`	vector of standard deviation estimates
`e(_N)`	vector of numbers of nonmissing observations
`e(_N_stdsum)`	number of nonmissing observations within the standard strata
`e(_p_stdize)`	standardizing proportions
`e(error)`	error code corresponding to `e(b)`

Functions

`e(sample)`	marks estimation sample

In addition to the above, the following is stored in `r()`:

Matrices

`r(table)`	matrix containing the coefficients with their standard errors, test statistics, p-values, and confidence intervals

Note that results stored in `r()` are updated when the command is replayed and will be replaced when any r-class command is run after the estimation command.

Methods and formulas

Methods and formulas are presented under the following headings:

The mean estimator

Let y be the variable on which we want to calculate the mean and y_j an individual observation on y, where $j = 1, \dots, n$ and n is the sample size. Let w_j be the weight, and if no weight is specified, define $w_j = 1$ for all j. For `aweights`, the w_j are normalized to sum to n. See *The survey mean estimator* for `pweighted` data.

Let W be the sum of the weights

$$W = \sum_{j=1}^{n} w_j$$

The mean is defined as

$$\overline{y} = \frac{1}{W} \sum_{j=1}^{n} w_j y_j$$

The default variance estimator for the mean is

$$\widehat{V}(\overline{y}) = \frac{1}{W(W-1)} \sum_{j=1}^{n} w_j (y_j - \overline{y})^2$$

The standard error of the mean is the square root of the variance.

If x, x_j, and $\overline{x}$ are similarly defined for another variable (observed jointly with y), the covariance estimator between $\overline{x}$ and $\overline{y}$ is

$$\widehat{\mathrm{Cov}}(\overline{x}, \overline{y}) = \frac{1}{W(W-1)} \sum_{j=1}^{n} w_j (x_j - \overline{x})(y_j - \overline{y})$$

Survey data

See [SVY] **Variance estimation**, [SVY] **Direct standardization**, and [SVY] **Poststratification** for discussions that provide background information for the following formulas. The following formulas are derived from the fact that the mean is a special case of the ratio estimator where the denominator variable is one, $x_j = 1$; see [R] **ratio**.

The survey mean estimator

Let Y_j be a survey item for the jth individual in the population, where $j = 1, \ldots, M$ and M is the size of the population. The associated population mean for the item of interest is $\overline{Y} = Y/M$ where

$$Y = \sum_{j=1}^{M} Y_j$$

Let y_j be the survey item for the jth sampled individual from the population, where $j = 1, \ldots, m$ and m is the number of observations in the sample.

The estimator for the mean is $\overline{y} = \widehat{Y}/\widehat{M}$, where

$$\widehat{Y} = \sum_{j=1}^{m} w_j y_j \qquad \text{and} \qquad \widehat{M} = \sum_{j=1}^{m} w_j$$

and w_j is a sampling weight. The score variable for the mean estimator is

$$z_j(\overline{y}) = \frac{y_j - \overline{y}}{\widehat{M}} = \frac{\widehat{M} y_j - \widehat{Y}}{\widehat{M}^2}$$

The standardized mean estimator

Let D_g denote the set of sampled observations that belong to the gth standard stratum and define $I_{D_g}(j)$ to indicate if the jth observation is a member of the gth standard stratum; where $g = 1, \ldots, L_D$ and L_D is the number of standard strata. Also, let π_g denote the fraction of the population that belongs to the gth standard stratum, thus $\pi_1 + \cdots + \pi_{L_D} = 1$. π_g is derived from the `stdweight()` option.

The estimator for the standardized mean is

$$\overline{y}^D = \sum_{g=1}^{L_D} \pi_g \frac{\widehat{Y}_g}{\widehat{M}_g}$$

where

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

The score variable for the standardized mean is

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

The poststratified mean estimator

Let P_k denote the set of sampled observations that belong to poststratum k and define $I_{P_k}(j)$ to indicate if the jth observation is a member of poststratum k; where $k = 1, \dots, L_P$ and L_P is the number of poststrata. Also let M_k denote the population size for poststratum k. P_k and M_k are identified by specifying the `poststrata()` and `postweight()` options on `svyset`; see [SVY] **svyset**.

The estimator for the poststratified mean is

$$\overline{y}^P = \frac{\widehat{Y}^P}{\widehat{M}^P} = \frac{\widehat{Y}^P}{M}$$

where

$$\widehat{Y}^P = \sum_{k=1}^{L_P} \frac{M_k}{\widehat{M}_k}\widehat{Y}_k = \sum_{k=1}^{L_P} \frac{M_k}{\widehat{M}_k} \sum_{j=1}^{m} I_{P_k}(j)\, w_j y_j$$

and

$$\widehat{M}^P = \sum_{k=1}^{L_P} \frac{M_k}{\widehat{M}_k}\widehat{M}_k = \sum_{k=1}^{L_P} M_k = M$$

The score variable for the poststratified mean is

$$z_j(\overline{y}^P) = \frac{z_j(\widehat{Y}^P)}{M} = \frac{1}{M}\sum_{k=1}^{L_P} I_{P_k}(j)\frac{M_k}{\widehat{M}_k}\left(y_j - \frac{\widehat{Y}_k}{\widehat{M}_k}\right)$$

The standardized poststratified mean estimator

The estimator for the standardized poststratified mean is

$$\overline{y}^{DP} = \sum_{g=1}^{L_D} \pi_g \frac{\widehat{Y}_g^P}{\widehat{M}_g^P}$$

where

$$\widehat{Y}_g^P = \sum_{k=1}^{L_p} \frac{M_k}{\widehat{M}_k}\widehat{Y}_{g,k} = \sum_{k=1}^{L_p} \frac{M_k}{\widehat{M}_k} \sum_{j=1}^{m} I_{D_g}(j) I_{P_k}(j)\, w_j y_j$$

and

$$\widehat{M}_g^P = \sum_{k=1}^{L_p} \frac{M_k}{\widehat{M}_k}\widehat{M}_{g,k} = \sum_{k=1}^{L_p} \frac{M_k}{\widehat{M}_k} \sum_{j=1}^{m} I_{D_g}(j) I_{P_k}(j)\, w_j$$

The score variable for the standardized poststratified mean is

$$z_j(\overline{y}^{DP}) = \sum_{g=1}^{L_D} \pi_g \frac{\widehat{M}_g^P z_j(\widehat{Y}_g^P) - \widehat{Y}_g^P z_j(\widehat{M}_g^P)}{(\widehat{M}_g^P)^2}$$

where

$$z_j(\widehat{Y}_g^P) = \sum_{k=1}^{L_P} I_{P_k}(j) \frac{M_k}{\widehat{M}_k} \left\{ I_{D_g}(j) y_j - \frac{\widehat{Y}_{g,k}}{\widehat{M}_k} \right\}$$

and

$$z_j(\widehat{M}_g^P) = \sum_{k=1}^{L_P} I_{P_k}(j) \frac{M_k}{\widehat{M}_k} \left\{ I_{D_g}(j) - \frac{\widehat{M}_{g,k}}{\widehat{M}_k} \right\}$$

Subpopulation estimation

Let S denote the set of sampled observations that belong to the subpopulation of interest, and define $I_S(j)$ to indicate if the jth observation falls within the subpopulation.

The estimator for the subpopulation mean is $\overline{y}^S = \widehat{Y}^S / \widehat{M}^S$, where

$$\widehat{Y}^S = \sum_{j=1}^{m} I_S(j)\, w_j y_j \qquad \text{and} \qquad \widehat{M}^S = \sum_{j=1}^{m} I_S(j)\, w_j$$

Its score variable is

$$z_j(\overline{y}^S) = I_S(j) \frac{y_j - \overline{y}^S}{\widehat{M}^S} = I_S(j) \frac{\widehat{M}^S y_j - \widehat{Y}^S}{(\widehat{M}^S)^2}$$

The estimator for the standardized subpopulation mean is

$$\overline{y}^{DS} = \sum_{g=1}^{L_D} \pi_g \frac{\widehat{Y}_g^S}{\widehat{M}_g^S}$$

where

$$\widehat{Y}_g^S = \sum_{j=1}^{m} I_{D_g}(j) I_S(j)\, w_j y_j \qquad \text{and} \qquad \widehat{M}_g^S = \sum_{j=1}^{m} I_{D_g}(j) I_S(j)\, w_j$$

Its score variable is

$$z_j(\overline{y}^{DS}) = \sum_{g=1}^{L_D} \pi_g I_{D_g}(j) I_S(j) \frac{\widehat{M}_g^S y_j - \widehat{Y}_g^S}{(\widehat{M}_g^S)^2}$$

The estimator for the poststratified subpopulation mean is

$$\overline{y}^{PS} = \frac{\widehat{Y}^{PS}}{\widehat{M}^{PS}}$$

where

$$\widehat{Y}^{PS} = \sum_{k=1}^{L_P} \frac{M_k}{\widehat{M}_k} \widehat{Y}_k^S = \sum_{k=1}^{L_P} \frac{M_k}{\widehat{M}_k} \sum_{j=1}^{m} I_{P_k}(j) I_S(j)\, w_j y_j$$

and

$$\widehat{M}^{PS} = \sum_{k=1}^{L_P} \frac{M_k}{\widehat{M}_k} \widehat{M}_k^S = \sum_{k=1}^{L_P} \frac{M_k}{\widehat{M}_k} \sum_{j=1}^{m} I_{P_k}(j) I_S(j)\, w_j$$

Its score variable is

$$z_j(\overline{y}^{PS}) = \frac{\widehat{M}^{PS} z_j(\widehat{Y}^{PS}) - \widehat{Y}^{PS} z_j(\widehat{M}^{PS})}{(\widehat{M}^{PS})^2}$$

where

$$z_j(\widehat{Y}^{PS}) = \sum_{k=1}^{L_P} I_{P_k}(j) \frac{M_k}{\widehat{M}_k} \left\{ I_S(j)\, y_j - \frac{\widehat{Y}_k^S}{\widehat{M}_k} \right\}$$

and

$$z_j(\widehat{M}^{PS}) = \sum_{k=1}^{L_P} I_{P_k}(j) \frac{M_k}{\widehat{M}_k} \left\{ I_S(j) - \frac{\widehat{M}_k^S}{\widehat{M}_k} \right\}$$

The estimator for the standardized poststratified subpopulation mean is

$$\overline{y}^{DPS} = \sum_{g=1}^{L_D} \pi_g \frac{\widehat{Y}_g^{PS}}{\widehat{M}_g^{PS}}$$

where

$$\widehat{Y}_g^{PS} = \sum_{k=1}^{L_p} \frac{M_k}{\widehat{M}_k} \widehat{Y}_{g,k}^S = \sum_{k=1}^{L_p} \frac{M_k}{\widehat{M}_k} \sum_{j=1}^{m} I_{D_g}(j) I_{P_k}(j) I_S(j)\, w_j y_j$$

and

$$\widehat{M}_g^{PS} = \sum_{k=1}^{L_p} \frac{M_k}{\widehat{M}_k} \widehat{M}_{g,k}^S = \sum_{k=1}^{L_p} \frac{M_k}{\widehat{M}_k} \sum_{j=1}^{m} I_{D_g}(j) I_{P_k}(j) I_S(j)\, w_j$$

Its score variable is

$$z_j(\overline{y}^{DPS}) = \sum_{g=1}^{L_D} \pi_g \frac{\widehat{M}_g^{PS} z_j(\widehat{Y}_g^{PS}) - \widehat{Y}_g^{PS} z_j(\widehat{M}_g^{PS})}{(\widehat{M}_g^{PS})^2}$$

where

$$z_j(\widehat{Y}_g^{PS}) = \sum_{k=1}^{L_P} I_{P_k}(j) \frac{M_k}{\widehat{M}_k} \left\{ I_{D_g}(j) I_S(j)\, y_j - \frac{\widehat{Y}_{g,k}^S}{\widehat{M}_k} \right\}$$

and

$$z_j(\widehat{M}_g^{PS}) = \sum_{k=1}^{L_P} I_{P_k}(j) \frac{M_k}{\widehat{M}_k} \left\{ I_{D_g}(j) I_S(j) - \frac{\widehat{M}_{g,k}^S}{\widehat{M}_k} \right\}$$

References

Bakker, A. 2003. The early history of average values and implications for education. *Journal of Statistics Education* 11: art. 3. https://doi.org/10.1080/10691898.2003.11910694.

Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. New York: Wiley.

Manski, C. F., and M. Tabord-Meehan. 2017. Evaluating the maximum MSE of mean estimators with missing data. *Stata Journal* 17: 723–735.

Stuart, A., and J. K. Ord. 1994. *Distribution Theory*. Vol. 1 of *Kendall's Advanced Theory of Statistics*, 6th ed. London: Arnold.

Also see

[R] **mean postestimation** — Postestimation tools for mean

[R] **ameans** — Arithmetic, geometric, and harmonic means

[R] **proportion** — Estimate proportions

[R] **ratio** — Estimate ratios

[R] **summarize** — Summary statistics

[R] **total** — Estimate totals

[MI] **Estimation** — Estimation commands for use with mi estimate

[SVY] **Direct standardization** — Direct standardization of means, proportions, and ratios

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

[SVY] **Subpopulation estimation** — Subpopulation estimation for survey data

[SVY] **svy estimation** — Estimation commands for survey data

[SVY] **Variance estimation** — Variance estimation for survey data

[U] 20 Estimation and postestimation commands