Title

> **bsample —** Sampling with replacement

Syntax	Menu	Description	Options
Remarks and examples	References	Also see	

Syntax

> bsample $\big[$ *exp* $\big]$ $\big[$ *if* $\big]$ $\big[$ *in* $\big]$ $\big[$, *options* $\big]$

where *exp* is a standard Stata expression; see [U] **13 Functions and expressions**.

options	Description
<u>str</u>ata(*varlist*)	variables identifying strata
cluster(*varlist*)	variables identifying resampling clusters
idcluster(*newvar*)	create new cluster ID variable
<u>weight</u>(*varname*)	replace *varname* with frequency weights

Menu

Statistics > Resampling > Draw bootstrap sample

Description

bsample draws bootstrap samples (random samples with replacement) from the data in memory.

exp specifies the size of the sample, which must be less than or equal to the number of sampling units in the data. The observed number of units is the default when *exp* is not specified.

For bootstrap sampling of the observations, *exp* must be less than or equal to _N (the number of observations in the data; see [U] **13.4 System variables (_variables)**).

For stratified bootstrap sampling, *exp* must be less than or equal to _N within the strata identified by the strata() option.

For clustered bootstrap sampling, *exp* must be less than or equal to N_c (the number of clusters identified by the cluster() option).

For stratified bootstrap sampling of clusters, *exp* must be less than or equal to N_c within the strata identified by the strata() option.

Observations that do not meet the optional if and in criteria are dropped (not sampled).

Options

strata(*varlist*) specifies the variables identifying strata. If strata() is specified, bootstrap samples are selected within each stratum.

cluster(*varlist*) specifies the variables identifying resampling clusters. If cluster() is specified, the sample drawn during each replication is a bootstrap sample of clusters.

idcluster(*newvar*) creates a new variable containing a unique identifier for each resampled cluster.

weight(*varname*) specifies a variable in which the sampling frequencies will be placed. *varname* must be an existing variable, which will be replaced. After bsample, *varname* can be used as an fweight in any Stata command that accepts fweights, which can speed up resampling for commands like regress and summarize. This option cannot be combined with idcluster().

By default, bsample replaces the data in memory with the sampled observations; however, specifying the weight() option causes only the specified *varname* to be changed.

Remarks and examples

Below is a series of examples illustrating how bsample is used with various sampling schemes.

▷ Example 1: Bootstrap sampling

We have data on the characteristics of hospital patients and wish to draw a bootstrap sample of 200 patients. We type

```
. use http://www.stata-press.com/data/r13/bsample1
. bsample 200
. count
  200
```

◁

▷ Example 2: Stratified samples with equal sizes

Among the variables in our dataset is female, an indicator for the female patients. To get a bootstrap sample of 200 female patients and 200 male patients, we type

```
. use http://www.stata-press.com/data/r13/bsample1, clear
. bsample 200, strata(female)
. tabulate female
```

female	Freq.	Percent	Cum.
male	200	50.00	50.00
female	200	50.00	100.00
Total	400	100.00	

◁

▷ Example 3: Stratified samples with unequal sizes

To sample 300 females and 200 males, we must generate a variable that is 300 for females and 200 for males and then use this variable in *exp* when we call bsample.

```
. use http://www.stata-press.com/data/r13/bsample1, clear
. generate nsamp = cond(female,300,200)
. bsample nsamp, strata(female)
. tabulate female
```

female	Freq.	Percent	Cum.
male	200	40.00	40.00
female	300	60.00	100.00
Total	500	100.00	

◁

▷ Example 4: Stratified samples with proportional sizes

Our original dataset has 2,392 males and 3,418 females.

```
. use http://www.stata-press.com/data/r13/bsample1, clear
. tabulate female
```

female	Freq.	Percent	Cum.
male	2,392	41.17	41.17
female	3,418	58.83	100.00
Total	5,810	100.00	

To sample 10% from females and males, we type

```
. bsample round(0.1*_N), strata(female)
```

bsample requires that the specified size of the sample be an integer, so we use the round() function to obtain the nearest integer to 0.1×2392 and 0.1×3418. Our sample now has 239 males and 342 females:

```
. tabulate female
```

female	Freq.	Percent	Cum.
male	239	41.14	41.14
female	342	58.86	100.00
Total	581	100.00	

◁

▷ Example 5: Samples satisfying a condition

For a bootstrap sample of 200 female patients, we type

```
. use http://www.stata-press.com/data/r13/bsample1, clear
. bsample 200 if female
. tabulate female
```

female	Freq.	Percent	Cum.
female	200	100.00	100.00
Total	200	100.00	

◁

▷ Example 6: Generating frequency weights

To identify the sampled observations using frequency weights instead of dropping unsampled observations, we use the `weight()` option (we will need to supply it an existing variable name) and type

```
. use http://www.stata-press.com/data/r13/bsample1, clear
. set seed 1234
. generate fw = .
(5810 missing values generated)
. bsample 200 if female, weight(fw)
. tabulate fw female
```

| | | female | |
fw	male	female	Total
0	2,392	3,221	5,613
1	0	194	194
2	0	3	3
Total	2,392	3,418	5,810

Note that $(194 \times 1) + (3 \times 2) = 200$.

◁

▷ Example 7: Oversampling observations

`bsample` requires the expression in *exp* to evaluate to a number that is less than or equal to the number of observations. To sample twice as many male and female patients as there are already in memory, we must expand the data before using `bsample`. For example,

```
. use http://www.stata-press.com/data/r13/bsample1, clear
. set seed 1234
. expand 2
(5810 observations created)
. bsample, strata(female)
. tabulate female
```

female	Freq.	Percent	Cum.
male	4,784	41.17	41.17
female	6,836	58.83	100.00
Total	11,620	100.00	

◁

▷ Example 8: Stratified oversampling with unequal sizes

To sample twice as many female patients as male patients, we must expand the records for the female patients because there are less than twice as many of them as there are male patients, but first put the number of observed male patients in a local macro. After expanding the female records, we generate a variable that contains the number of observations to sample within the two groups.

```
. use http://www.stata-press.com/data/r13/bsample1, clear
. set seed 1234
. count if !female
 2392
. local nmale = r(N)
. expand 2 if female
(3418 observations created)
. generate nsamp = cond(female,2*`nmale',`nmale')
. bsample nsamp, strata(female)
. tabulate female
```

female	Freq.	Percent	Cum.
male	2,392	33.33	33.33
female	4,784	66.67	100.00
Total	7,176	100.00	

◁

▷ Example 9: Oversampling of clusters

For clustered data, sampling more clusters than are present in the original dataset requires more than just expanding the data. To illustrate, suppose we wanted a bootstrap sample of eight clusters from a dataset consisting of five clusters of observations.

```
. use http://www.stata-press.com/data/r13/bsample2, clear
. tabstat x, stat(n mean) by(group)
Summary for variables: x
     by categories of: group
```

group	N	mean
A	15	-.3073028
B	10	-.00984
C	11	.0810985
D	11	-.1989179
E	29	-.095203
Total	76	-.1153269

bsample will complain if we simply expand the dataset.

```
. use http://www.stata-press.com/data/r13/bsample2
. expand 3
(152 observations created)
. bsample 8, cluster(group)
resample size must not be greater than number of clusters
r(498);
```

Expanding the data will only partly solve the problem. We also need a new variable that uniquely identifies the copied clusters. We use the `expandcl` command to accomplish both these tasks; see [D] **expandcl**.

```
. use http://www.stata-press.com/data/r13/bsample2, clear
. set seed 1234
. expandcl 2, generate(expgroup) cluster(group)
(76 observations created)
. tabstat x, stat(n mean) by(expgroup)
Summary for variables: x
    by categories of: expgroup
```

expgroup	N	mean
1	15	-.3073028
2	15	-.3073028
3	10	-.00984
4	10	-.00984
5	11	.0810985
6	11	.0810985
7	11	-.1989179
8	11	-.1989179
9	29	-.095203
10	29	-.095203
Total	152	-.1153269

```
. generate fw = .
(152 missing values generated)
. bsample 8, cluster(expgroup) weight(fw)
. tabulate fw group
```

	group					
fw	A	B	C	D	E	Total
0	15	10	0	0	29	54
1	15	10	22	22	0	69
2	0	0	0	0	29	29
Total	30	20	22	22	58	152

The results from `tabulate` on the generated frequency weight variable versus the original cluster ID (`group`) show us that the bootstrap sample contains one copy of cluster A, one copy of cluster B, two copies of cluster C, two copies of cluster D, and two copies of cluster E ($1 + 1 + 2 + 2 + 2 = 8$).

◁

▷ Example 10: Stratified oversampling of clusters

Suppose that we have a dataset containing two strata with five clusters in each stratum, but the cluster identifiers are not unique between the strata. To get a stratified bootstrap sample with eight clusters in each stratum, we first use `expandcl` to expand the data and get a new cluster ID variable. We use `cluster(strid group)` in the call to `expandcl`; this action will uniquely identify the $2 * 5 = 10$ clusters across the strata.

```
. use http://www.stata-press.com/data/r13/bsample2, clear
. set seed 1234
. tabulate group strid
```

group	strid 1	2	Total
A	7	8	15
B	5	5	10
C	5	6	11
D	5	6	11
E	14	15	29
Total	36	40	76

```
. expandcl 2, generate(expgroup) cluster(strid group)
(76 observations created)
```

Now we can use `bsample` with the expanded data, stratum ID variable, and new cluster ID variable.

```
. generate fw = .
(152 missing values generated)
. bsample 8, cluster(expgroup) str(strid) weight(fw)
. by strid, sort: tabulate fw group
```

-> strid = 1

fw	A	B	group C	D	E	Total
0	0	5	0	5	14	24
1	14	5	10	5	0	34
2	0	0	0	0	14	14
Total	14	10	10	10	28	72

-> strid = 2

fw	A	B	group C	D	E	Total
0	8	10	0	6	0	24
1	8	0	6	6	15	35
2	0	0	6	0	15	21
Total	16	10	12	12	30	80

The results from `by strid: tabulate` on the generated frequency weight variable versus the original cluster ID (`group`) show us how many times each cluster was sampled for each stratum. For stratum 1, the bootstrap sample contains two copies of cluster A, one copy of cluster B, two copies of cluster C, one copy of cluster D, and two copies of cluster E $(2 + 1 + 2 + 1 + 2 = 8)$. For stratum 2, the bootstrap sample contains one copy of cluster A, zero copies of cluster B, three copies of cluster C, one copy of cluster D, and three copies of cluster E $(1 + 0 + 3 + 1 + 3 = 8)$.

◁

References

Gould, W. W. 2012a. Using Stata's random-number generators, part 2: Drawing without replacement. The Stata Blog: Not Elsewhere Classified.
http://blog.stata.com/2012/08/03/using-statas-random-number-generators-part-2-drawing-without-replacement/.

———. 2012b. Using Stata's random-number generators, part 3: Drawing with replacement. The Stata Blog: Not Elsewhere Classified. http://blog.stata.com/2012/08/29/using-statas-random-number-generators-part-3-drawing-with-replacement/.

Also see

[R] **bootstrap** — Bootstrap sampling and estimation

[R] **bstat** — Report bootstrap results

[R] **simulate** — Monte Carlo simulations

[D] **sample** — Draw random sample