Title

stata.com

sample — Draw random sample

Syntax | Menu | Description | Options
Remarks and examples | References | Also see

Syntax

```
sample # [ if ] [ in ] [ , count by(groupvars) ]
```

by is allowed; see [D] **by**.

Menu

Statistics > Resampling > Draw random sample

Description

sample draws random samples of the data in memory. "Sampling" here is defined as drawing observations without replacement; see [R] **bsample** for sampling with replacement.

The size of the sample to be drawn can be specified as a percentage or as a count:

- sample without the count option draws a #% pseudorandom sample of the data in memory, thus discarding (100 − #)% of the observations.
- sample with the count option draws a #-observation pseudorandom sample of the data in memory, thus discarding _N − # observations. # can be larger than _N, in which case all observations are kept.

In either case, observations not meeting the optional if and in criteria are kept (sampled at 100%).

If you are interested in reproducing results, you must first set the random-number seed; see [R] **set seed**.

Options

count specifies that # in sample # be interpreted as an observation count rather than as a percentage. Typing sample 5 without the count option means that a 5% sample be drawn; typing sample 5, count, however, would draw a sample of 5 observations.

Specifying # as greater than the number of observations in the dataset is not considered an error.

by(*groupvars*) specifies that a #% sample be drawn within each set of values of *groupvars*, thus maintaining the proportion of each group.

count may be combined with by(). For example, typing sample 50, count by(sex) would draw a sample of size 50 for men and 50 for women.

Specifying by *varlist*: sample # is equivalent to specifying sample #, by(*varlist*); use whichever syntax you prefer.

Remarks and examples

stata.com

Example 1

We have NLSY data on young women aged 14–26 years in 1968 and wish to draw a 10% sample of the data in memory.

```
. use http://www.stata-press.com/data/r13/nlswork
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. describe, short
Contains data from http://www.stata-press.com/data/r13/nlswork.dta
  obs:        28,534                          National Longitudinal Survey.
                                                Young Women 14-26 years of age
                                                in 1968
 vars:            21                          27 Nov 2012 08:14
 size:       941,622
Sorted by:  idcode  year
. sample 10
(25681 observations deleted)
. describe, short
Contains data from http://www.stata-press.com/data/r13/nlswork.dta
  obs:         2,853                          National Longitudinal Survey.
                                                Young Women 14-26 years of age
                                                in 1968
 vars:            21                          27 Nov 2012 08:14
 size:        94,149
Sorted by:
     Note:  dataset has changed since last saved
```

Our original dataset had 28,534 observations. The sample-10 dataset has 2,853 observations, which is the nearest number to 0.10×28534.

◁

Example 2

Among the variables in our data is `race`. By typing `label list`, we see that `race` = 1 denotes whites, `race` = 2 denotes blacks, and `race` = 3 denotes other races. We want to keep 100% of the nonwhite women but only 10% of the white women.

```
. use http://www.stata-press.com/data/r13/nlswork, clear
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. tab race
```

race	Freq.	Percent	Cum.
white	20,180	70.72	70.72
black	8,051	28.22	98.94
other	303	1.06	100.00
Total	28,534	100.00	

```
. sample 10 if race == 1
(18162 observations deleted)
```

```
. describe, short
Contains data from http://www.stata-press.com/data/r13/nlswork.dta
  obs:        10,372                          National Longitudinal Survey.
                                                Young Women 14-26 years of age
                                                in 1968
 vars:            21                          27 Nov 2012 08:14
 size:       342,276
Sorted by:
     Note:  dataset has changed since last saved
. display .10*20180 + 8051 + 303
10372
```

◁

▷ Example 3

Now let's suppose that we want to keep 10% of each of the three categories of `race`.

```
. use http://www.stata-press.com/data/r13/nlswork, clear
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
. sample 10, by(race)
(25681 observations deleted)
. tab race
```

race	Freq.	Percent	Cum.
white	2,018	70.73	70.73
black	805	28.22	98.95
other	30	1.05	100.00
Total	2,853	100.00	

This differs from simply typing `sample 10` in that with `by()`, `sample` holds constant the percentages of white, black, and other women.

◁

❑ Technical note

We have a large dataset on disk containing 125,235 observations. We wish to draw a 10% sample of this dataset without loading the entire dataset (perhaps because the dataset will not fit in memory). `sample` will not solve this problem—the dataset must be loaded first—but it is rather easy to solve it ourselves. Say that `bigdata.dct` contains the dictionary for this dataset; see [D] **import**. One solution is to type

```
. infile using bigdata if runiform()<=.1
dictionary {
     etc.
}
(12,580 observations read)
```

The `if` modifier on the end of `infile` drew uniformly distributed random numbers over the interval 0 and 1 and kept each observation if the random number was less than or equal to 0.1. This, however, did not draw an exact 10% sample—the sample was expected to contain only 10% of the observations, and here we obtained just more than 10%. This is probably a reasonable solution.

If the sample must contain precisely 12,524 observations, however, after getting too many observations, we could type

```
. generate u=runiform()
. sort u
. keep in 1/12524
(56 observations deleted)
```

That is, we put the resulting sample in random order and keep the first 12,524 observations. Now our only problem is making sure that, at the first step, we have more than 12,524 observations. Here we were lucky, but half the time we will not be so lucky—after typing `infile ... if runiform()<=.1`, we will have less than a 10% sample. The solution, of course, is to draw more than a 10% sample initially and then cut it back to 10%.

How much more than 10% do we need? That depends on the number of records in the original dataset, which in our example is 125,235.

A little experimentation with `bitesti` (see [R] **bitest**) provides the answer:

```
. bitesti 125235 12524 .102
        N   Observed k    Expected k   Assumed p   Observed p
------------------------------------------------------------
   125235       12524      12773.97     0.10200      0.10000

Pr(k >= 12524)                 = 0.990466  (one-sided test)
Pr(k <= 12524)                 = 0.009777  (one-sided test)
Pr(k <= 12524 or k >= 13025) = 0.019584  (two-sided test)
```

Initially drawing a 10.2% sample will yield a sample larger than 10% 99 times of 100. If we draw a 10.4% sample, we are virtually assured of having enough observations (type `bitesti 125235 12524 .104` for yourself).

❑

References

Cox, N. J. 2001. dm86: Sampling without replacement: Absolute sample sizes and keeping all observations. *Stata Technical Bulletin* 59: 8–9. Reprinted in *Stata Technical Bulletin Reprints*, vol. 10, pp. 38–39. College Station, TX: Stata Press.

——. 2005. Software Updates: Sampling without replacement: Absolute sample sizes and keeping all observations. *Stata Journal* 5: 139.

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/.

Weesie, J. 1997. dm46: Enhancement to the sample command. *Stata Technical Bulletin* 37: 6–7. Reprinted in *Stata Technical Bulletin Reprints*, vol. 7, pp. 37–38. College Station, TX: Stata Press.

Also see

[R] **bsample** — Sampling with replacement