How can I draw a random sample of my data in Stata?

Authored by

stats writer

June 30, 2024

To draw a random sample of data in Stata, you can use the "sample" command. This command allows you to specify the size of the sample, as well as any specific variables or conditions you want to include or exclude in the sampling process. By default, the "sample" command will draw a random sample without replacement, meaning that each observation can only appear once in the sample. You can also use the "sample" command with the "by" option to draw random samples from different groups within your data. Overall, the "sample" command provides a straightforward and flexible way to obtain a random sample of data in Stata.

How can I draw a random sample of my data? | Stata FAQ

There are many instances when you may want to take a random sample of your dataset. For example, you may have a very large data set, and you are trying to work out the code to clean your data or to analyze it. Running the code on many observations can take a while, so testing the code on a subset of the data is a good way to save some time. However, you may not want to take just the first 100 or so cases, as they may be different in some important way than cases that occur later in the data set. Hence, taking a random sample of the data would be desirable.

There are two commands in Stata that can be used to take a random sample of your data set. Use the sample command to draw a sample without replacement, meaning that once an observation (i.e., case, element) has been selected into the sample, it is not available to be selected into the sample again. Use the bsample command if you want to draw a sample with replacement, meaning that once the observation has been selected into the sample, it is replaced into the pool of observations from which the sample is being drawn. Theoretically, it can be selected a second, third, fourth, etc. time. If your data set is very large, the results from the two commands probably will not differ (assuming that you used the same seed for both, see below). This is because the probability of any given observation being selected into the sample is low in a large data set, so the odds of being

selected a second time is also low. (Please note that the probability of being selected into the sample will not change whether or not an observation has previously been selected into the data set. In other words, observations that have already been selected into the sample have the same probability of selection has observations that have not yet been selected into the sample.)

Sampling without replacement

Let's suppose that we want to create a sample of 10% of our current data set. After opening our data set, hsb2, we will use the count command to see how many observations are in the data set. Next, we will issue the sample command and then use the count command again to see how many observations are in the data set.

```
use https://stats.idre.ucla.edu/stat/stata/notes/hsb2,
```

clear

count
200

sample 10
(180 observations deleted)

count
20

As you can see, only 20 of the original 200 observations remain in the data
set (20 is 10% of 200). You may want to save this smaller data set with a
new name, so that you do not overwrite your original data set.

Now let's specify the number of observations, say 50, that we want in our
sample, instead of the percentage of the data set. To do this, we will use
the count option for the sample command.

use https://stats.idre.ucla.edu/stat/stata/notes/hsb2,

```
clear

sample 50, count
(150 observations deleted)

count
50
```

What will happen if we specify a number that is larger than the number of observations in the data set?

```
use https://stats.idre.ucla.edu/stat/stata/notes/hsb2, clear

sample 250, count
count
200
```

As you can see, all of the observations from the data set were kept, but none were sampled a second time to increase the sample size the desired number. Notice also that Stata did not issue an error message when the sample size

exceeded the number of observations in the data set.

You can also select a sample with a given percentage or number from each of
level of a grouping variable. (This might be a strata variable.) In the
hsb2 data set, the variable prog
is a three-level categorical (grouping) variable that indicates the type of
school program each student is in (1= general, 2=academic, 3=vocational). We
can select a sample such that, say 15%, of each of those categories are selected
into the sample. Note that you need to sort the data on the grouping
variable before using the by: prefix. We will start with a count of all of the cases in each level of prog so that we can compare
these numbers to those we have after issuing the sample command. Also note that you can use
the count option with the by: prefix if you want to specify the
number of observations to be included in the sample.

```
use https://stats.idre.ucla.edu/stat/stata/notes/hsb2, clear

sort prog
by prog: count

-------------------------------------------------------------------------------
-> prog = general
45
-------------------------------------------------------------------------------
-> prog = academic
105
-------------------------------------------------------------------------------
-> prog = vocation
50

by prog: sample 15
(169 observations deleted)

count
31

by prog: count
```

-> prog = general

7

-> prog = academic

16

-> prog = vocation

8

You can also also specify conditions by which the sample should be selected.
For example, consider the code below.

```
use https://stats.idre.ucla.edu/stat/stata/notes/hsb2, clear

sort prog
by prog: count
```

-> prog = general

45

-> prog = academic

105

-> prog = vocation

50

sample 12 if prog == 3

(44 observations deleted)

count

156

sort prog

by prog: count

-> prog = general

45

-> prog = academic
105

-> prog = vocation
6

As you can see, all of the observations from the non-vocation
(general and academic) categories were included in the sample, as well as approximately 12% of the
cases from the vocation category were included (.12*50 = 6). Now let's
consider writing the code as shown below.

use https://stats.idre.ucla.edu/stat/stata/notes/hsb2, clear

sample 12 if prog != 3
(132 observations deleted)

count
68

sort prog

```
by prog: count
```

```
-> prog = general
7
```

```
-> prog = academic
11
```

```
-> prog = vocation
50
```

We can see that all 50 cases from the vocation category were included, as well as approximately 12% from each of the other categories.

Sampling with replacement

To illustrate how to sample with replacement, we will create a little data set, as shown below.

```
clear
input id wt strata1 cluster1 x
1 4 1 1 15
2 4 1 1 29
3 4 2 2 14
4 4 2 2 25
5 4 3 2 17
6 5 3 3 19
7 5 4 3 20
8 5 4 3 27
9 5 5 4 26
10 5 5 4 28
end

save "d:wrsample.dta", replace
```

The basic command is bsample followed by the number of observations
that you want in the sample. Note that sample size cannot exceed the
number of observations in the data set.

```
bsample 5
list
```

```
+---------------------------------+
| id wt strata1 cluster1 x |
|---------------------------------|
1. | 10 5 5 4 28 |
2. | 3 4 2 2 14 |
3. | 5 4 3 2 17 |
4. | 10 5 5 4 28 |
5. | 9 5 5 4 26 |
+---------------------------------+
```

You can use the weight option to see the frequency weights. Note
that you need to have a "weight" variable in the data set.

```
use "d:wrsample.dta", clear

bsample 4, weight(wt)
list
```

```
+---------------------------------+
| id wt strata1 cluster1 x |
|---------------------------------|
1. | 5 0 3 2 17 |
2. | 2 2 1 1 29 |
```

```
3. | 8 1 4 3 27 |
4. | 3 1 2 2 14 |
5. | 9 0 5 4 26 |
|---------------------------------|
6. | 10 0 5 4 28 |
7. | 1 0 1 1 15 |
8. | 4 0 2 2 25 |
9. | 7 0 4 3 20 |
10. | 6 0 3 3 19 |
+---------------------------------+
```

In this example, observation number 2 was selected twice, and observations 8 and 3 were each selected once.

You still have all 10 observations, but the weights have been changed to reflect which observations should be included in the sample. Try running the code multiple times and you will see that you get different results each time that you run it.

If your data are stratified, you can sample from each of

the strata.
You need to provide Stata with the number of observations that you want
from each strata, not the total number of observations that you want in the
sample. In the following example, we will ask for one observation from
each strata, giving us a total sample size of 5.

```
use "d:wrsample.dta", clear

bsample 1, strata(strata1)
list

+---------------------------------+
| id wt strata1 cluster1 x |
|---------------------------------|
1. | 2 4 1 1 29 |
2. | 3 4 2 2 14 |
3. | 5 4 3 2 17 |
4. | 8 5 4 3 27 |
5. | 9 5 5 4 26 |
+---------------------------------+
```

You can also sample clusters of your data using the

cluster option.

Note that Stata will select as many clusters as you request, not that many observations.

```
use "d:wrsample.dta", clear

bsample 3, cluster(cluster1)
list

+---------------------------------+
| id wt strata1 cluster1 x |
|---------------------------------|
1. | 6 5 3 3 19 |
2. | 7 5 4 3 20 |
3. | 8 5 4 3 27 |
4. | 6 5 3 3 19 |
5. | 7 5 4 3 20 |
|---------------------------------|
6. | 8 5 4 3 27 |
7. | 1 4 1 1 15 |
8. | 2 4 1 1 29 |
+---------------------------------+
```

In this example, Stata chose cluster 3 twice and cluster

1 once for a total
of three clusters.

Setting the seed

When taking a random sample of your data, you may want to do so in a way that
is reproducible. In other words, you can generate the same sample if you
need to. To do this, you will need to set the seed. The seed is the
number with which Stata (or any other program) starts its algorithm to generate
the pseudo-random numbers. If you do not set the seed, Stata will start
its algorithm with the seed 123456789. To set the seed, use the set seed command followed by a number. The number can be very
large, including 30 or more digits. Remember to do this in a .do file or to
write the seed number down somewhere. Please see the Stata Data Management Manual for more information.

set seed 2038947