How can I sample from a dataset with frequency weights?

Authored by

stats writer

July 1, 2024

Sampling from a dataset with frequency weights refers to the process of selecting a subset of data points from a larger dataset, where the probability of selecting a particular data point is proportional to its frequency in the original dataset. This technique is commonly used in statistics and data analysis to ensure that the selected subset is representative of the entire dataset. It involves assigning weights to each data point based on its frequency and then using these weights to guide the selection process. This approach allows for a more accurate and unbiased representation of the original dataset, making it a valuable tool in data sampling and analysis.

How can I sample from a dataset with frequency weights? | Stata FAQ

If you are working with a dataset that contains frequency weights, you may wish to sample from your dataset according to these weights.

Example 1: Using expand and sample

In Stata, you can easily sample from your dataset using these weights by
using expand to create a dataset with an observation for each unit and
then sampling from your expanded dataset.

We will be looking at a dataset with 200 frequency-weighted observations. The
frequency weights (fw) range from 1 to 20.

use https://stats.idre.ucla.edu/stat/data/hsb2_fw, clear

tabstat fw, stat(mean sum n)

variable | mean sum N
----------+----------------------------
fw | 10.325 2065 200
--

In this small example, we can see that by expanding our dataset, we would go
from 200 to 2065 observations. We can do so with the code below.

expand fw
(1865 observations created)

We can see that adding the number of observations created to our original
count of 200 observations arrives at 2065 observations. Now each item in
our population is represented by an observation. At this point, we can use
the sample command to draw a simple random sample with the size set to 20% of

our population.

sample 20
(1652 observations deleted)

The new dataset in memory now contains (2065 - 1652) = 413 observations-20%
of 2065. This process has been quite straightforward. However, if you start with a dataset that is already very large, you may wish to
avoid generating a much larger dataset with this step.

Example 2: Using gsample to sample with equal or unequal probabilities

In the above
example, we increased our observation count by a factor of 10. This was
very manageable with our small dataset, but it is not always a reasonable
option. The user-written
gsample command allows you to sample from your dataset with using frequency
weights or other unequal probability schemes.

To download
gsample, enter
search gsample in your Stata command
window and install the needed files. You might also need to enter
ssc install
moremata if you never have before in order for
gsample commands to
run.

Let's load our original 200 observations again. We can use
gsample to generate a simple
random sample of 20 observations from our set of 200 observations in the same
way the sample command would.

```
use https://stats.idre.ucla.edu/stat/data/hsb2_fw, clear

preserve

gsample 20

tabstat fw, stat(mean sum n)
```

variable | mean sum N

----------+----------------------------

fw | 9.95 199 20

--

restore

In the above sample of 20, all of the 200 observations in our data were
sampled with equal probability, regardless of their frequency weights. If we
wish to sample with greater probabilities the observations with higher frequency
weights and with lower probabilities the observations with lower frequency
weights, that is easily done with
gsample. When we specify aw = fw, the sampling probability of
an observation is proportional to its fw value.

preserve

gsample 20

tabstat fw, stat(mean sum n)

```
variable | mean sum N
-------------+-----------------------------
fw | 12.45 249 20
-----------------------------------------
```

```
restore
```

We can see that while the number of sampled units is the same in this sample
as in the previous sample, the mean
fw value is noticeably higher.
If we repeated this experiment over and over, we would expect this to be the
case since we are sampling observations with higher fw values with
greater probability.

Example 3: Drawing a simple random sample from population using gsample

If we wish to generate a simple random sample from our population that is 40%
of its size, we can do this using gsample as well. When we sampled
20 observations with the sampling probabilities

proportional to their frequency
weights, that was equivalent to drawing a 20 observation simple random sample
from our population. We can first calculate how many observations we would need to
sample in this way first and then draw a sample of the calculated size.

dis 2065 * .4
826

gsample 826

Other gsample features

Gsample is also capable of stratified and cluster sampling and these
can be combined with the weightsoption. While the default is to
replace the existing dataset with the sampled dataset, you can opt instead to
generate a variable in the existing dataset with the sampled frequencies.

gsample 20 , gen(sfreq)

tab(sfreq)

sfreq | Freq. Percent Cum.

----------+----------------------------------

0 | 181 90.50 90.50

1 | 18 9.00 99.50

2 | 1 0.50 100.00

----------+----------------------------------

Total | 200 100.00

By default,
gsample
samples with replacement. We can see that our sample of 20 includes 18 observations sampled once and 1 observation sampled twice. Using the
wor option, you can indicate that you want to sample without replacement.

gsample 20 , wor gen(sfwor)

tab(sfwor)

sfwor | Freq. Percent Cum.

----------+----------------------------------

0 | 180 90.00 90.00

1 | 20 10.00 100.00

----------+----------

Total | 200 100.00