splitsample — Split data into random samples

Description

splitsample splits data into random samples based on a specified number of samples and specified proportions for each sample. Splitting can also be done based on clusters. Sample splitting can also be balanced across specified variables. Balanced splitting can be used for matched treatment assignment.

Quick start

Split data into two random samples of equal sizes and generate sample ID variable svar with values 1 and 2

splitsample, generate(svar)

As above, but with sample ID variable svar having values 0 and 1

splitsample, generate(svar) values(0 1)

Split data into three random samples of equal sizes and generate sample ID variable svar with values 1, 2, and 3

splitsample, generate(svar) nsplit(3)

As above, but with sample ID variable svar equal to missing (.) whenever any of y or x1-x100 have missing values

splitsample y x1-x100, generate(svar) nsplit(3)

Split data into three random samples with the first sample having 25% of the observations, the second having 25%, and the third having 50%

splitsample, generate(svar) split(0.25 0.25 0.5)

Same sample split as above, but specify the split using ratios rather than proportions

splitsample, generate(svar) split(1 1 2)

As above, but maintain the specified sample-size ratios in each group defined by the variables agegrp and gender

splitsample, generate(svar) split(1 1 2) balance(agegrp gender)

As above, but randomly round sample sizes when samples within an agegrp by gender group cannot be chosen to satisfy the specified sample-size ratios exactly

splitsample, generate(svar) split(1 1 2) balance(agegrp gender) rround

Split data into three samples based on clusters defined by clustvar

splitsample, generate(svar) nsplit(3) cluster(clustvar)
As above, but maintain the specified sample proportions based on clusters in each group defined by the variables \texttt{agegrp} and \texttt{gender}, randomly round cluster sample sizes, and display a table showing the cluster sample sizes

\begin{verbatim}
  splitsample, generate(svar) nsplit(3) cluster(clustvar) //
  balance(agegrp gender) rround show
\end{verbatim}

**Menu**

Data $>$ Create or change data $>$ Other variable-creation commands $>$ Split data into random samples

**Syntax**

\begin{verbatim}
splitsample [ \texttt{varlist} ] [ \texttt{if} ] [ \texttt{in} ], \texttt{generate(newvar[, replace])} [ \texttt{options} ]
\end{verbatim}

\texttt{varlist} is checked for missing values, and the sample ID variable \texttt{newvar} is set to missing for observations where any variable in \texttt{varlist} is missing. \_all or \* may be specified for \texttt{varlist}.

**options**

<table>
<thead>
<tr>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{generate(newvar[, replace])} creates a new variable containing ID values for the random samples. The variable \texttt{newvar} is valued 1, 2, \ldots by default. The option \texttt{values(numlist)} can be used to specify different ID values. \texttt{generate()} is required.</td>
</tr>
<tr>
<td>\texttt{values(numlist)} specifies \texttt{numlist} of values for sample ID variable</td>
</tr>
<tr>
<td>\texttt{nsplit(#)} splits into # random samples of equal size</td>
</tr>
<tr>
<td>\texttt{split(numlist)} specifies \texttt{numlist} of proportions or ratios for the split</td>
</tr>
<tr>
<td>\texttt{rround} randomly rounds sample sizes when an exact split cannot be made</td>
</tr>
<tr>
<td>\texttt{cluster(clustvar)} splits by clusters defined by \texttt{clustvar}, not observations</td>
</tr>
<tr>
<td>\texttt{balance(balvars)} splits each group defined by the distinct values of \texttt{balvars} independently based on the specified sample proportions</td>
</tr>
<tr>
<td>\texttt{strok} evaluates string variables in \texttt{varlist} for missing values; by default, string variables are ignored</td>
</tr>
<tr>
<td>\texttt{rseed(#)} specifies random-number seed</td>
</tr>
<tr>
<td>\texttt{show} displays a table showing the sample sizes of the split</td>
</tr>
<tr>
<td>\texttt{percent} displays percentages in the table showing the split</td>
</tr>
</tbody>
</table>

*\texttt{generate()} is required.

 collect is allowed; see [U] 11.1.10 Prefix commands.
nsplit(#) splits the data into # random samples of equal size, or as close to equal as possible. If
neither nsplit() nor split() is specified, the data are split into two samples.

split(numlist) is an alternative to nsplit() for specifying the split. This option splits the data into
samples whose sizes are proportional to the values of numlist. The values of numlist can be any
positive number. You can specify proportions that sum to 1, or you can specify integers that define
ratios for the sample sizes. Regardless of whether you specify decimals less than 1 or integers,
the proportions of the split are given by the values in numlist divided by their sum.

rround specifies that sample sizes be randomly rounded when an exact split cannot be made. When
an exact split can be made, this option does nothing. When split(numlist) is specified with
rround, numlist must consist of integers, and the integers should contain no common factors. For
instance, use split(1 1 2), not split(25 25 50). See Methods and formulas for an explanation.

By default, the sample sizes of the splits are calculated using a deterministic rounding formula.
That is, if you repeat the splitting with a different random-number seed, you will get exactly
the same sample sizes. Specifying rround creates randomly rounded sample sizes such that the
expected values of the sample sizes match the specified split proportions exactly.

The option rround is designed for use with the balance() option when the number of observations
in each of the balance groups is small. When group sizes are small (especially when smaller than
the number of splits), rround ensures that the overall actual sample split proportions closely match
the specified split proportions.

values(numlist) specifies that numlist be used for the values of the sample ID variable rather than the
default of 1, 2, .... The number of values in numlist must correspond to the number of samples
into which the data are split and must be ascending nonnegative integers.

cluster(clustvar) specifies that the data be split by the clusters defined by clustvar. That is, all
observations in a cluster are kept together in the same split sample. The proportions of the split
are based on numbers of clusters, not numbers of observations. clustvar can be a numeric or string
variable.

balance(balvars) specifies that each group defined by the distinct values of balvars be split indepen-
dently based on the specified sample proportions. This ensures a balanced, or roughly balanced,
distribution of the balvars values across the split samples. When the number of observations (or
clusters) in each group is about the same as (or smaller than) the number of split samples, the
option rround is recommended. balvars can be numeric or string variables.

strokr (applies only when a varlist is specified) specifies to check any string variables in varlist for
missing values. For observations with missing values, the generated sample ID variable is set to
missing. By default, string variables in varlist are ignored.
rseed(#) sets the random-number seed. This option can be used to reproduce results. rseed(#) is
equivalent to typing set seed # prior to running splitsample. See [R] set seed.

show displays a table showing the sample sizes of the split. When cluster() is specified, it shows
the numbers of clusters in the samples. When balance(balvars) is specified, it displays a table in
which each row corresponds to a distinct set of values of balvars and shown across the columns
are the numbers of observations (or clusters) belonging to each split sample for that balance group.

percent specifies to display percentages rather than the number of observations (or clusters) in the
table. percent can only be specified with the option show.
 Remarks and examples

splitsample is useful for dividing data into training, validation, and testing samples for machine learning and automated model-building procedures such as those performed by the lasso, stepwise, and nestreg commands.

splitsample with the options balance() and rround can also be used to do random treatment assignment with matching. See example 3.

Example 1: Splitting by observations

Let’s create a dataset with 101 observations and run splitsample without any options except the required option giving the name of the sample ID variable to generate. Then we tabulate the newly created variable.

```
. set obs 101
   Number of observations (N) was 0, now 101.
. splitsample, generate(svar)
. tabulate svar
```

<table>
<thead>
<tr>
<th>svar</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>51</td>
<td>50.50</td>
<td>50.50</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>49.50</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 101 | 100.00 |

By default, splitsample splits the data into two samples, with the samples as equal in size as possible.

The option nsplit(#) can be used to split the data into as many samples as you want—in this case, three samples.

```
. splitsample, generate(svar, replace) nsplit(3)
. tabulate svar
```

<table>
<thead>
<tr>
<th>svar</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34</td>
<td>33.66</td>
<td>33.66</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>32.67</td>
<td>66.34</td>
</tr>
<tr>
<td>3</td>
<td>34</td>
<td>33.66</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 101 | 100.00 |

The option split(numlist) can be specified in place of nsplit() to split the data into any proportions you want. Here we specify that we want 25% of the observations in sample 1, 25% in sample 2, and 50% in sample 3.

```
. splitsample, generate(svar, replace) split(0.25 0.25 0.50) show
```

<table>
<thead>
<tr>
<th>svar</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>24.75</td>
<td>24.75</td>
</tr>
<tr>
<td>2</td>
<td>26</td>
<td>25.74</td>
<td>50.50</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>49.50</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 101 | 100.00 |

It split the data as close as it could to 25% : 25% : 50%. The option show displayed the tabulation for us.
Example 2: Splitting by clusters

`splitsample` can also split the data by clusters. Let’s create a cluster variable `clustvar` and split the data into three samples with proportions 25% : 25% : 50% for the numbers of clusters. We also specify the option `show`, which gives a convenient tabulation by numbers of clusters rather than numbers of observations.

```
. set seed 12345
. generate clustvar = runiformint(1, 20)
. splitsample, generate(svar, replace) split(0.25 0.25 0.50) cluster(clustvar)
> show
```

<table>
<thead>
<tr>
<th>svar</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>25.00</td>
<td>25.00</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>25.00</td>
<td>50.00</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>50.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total is number of clusters.

Because we had 20 clusters, the split into 25% : 25% : 50% yielded cluster sample sizes that met the specified proportions exactly.

The resulting split by number of observations is, of course, different.

```
. tabulate svar
```

<table>
<thead>
<tr>
<th>svar</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>34</td>
<td>33.66</td>
<td>33.66</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>20.79</td>
<td>54.46</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>45.54</td>
<td>100.00</td>
</tr>
</tbody>
</table>

When splitting by clusters, the size of each cluster is ignored.

Example 3: Balanced splitting and treatment assignment

`splitsample` can split the data independently within groups using the option `balance()`. Let’s create two fake categorical variables, one `agegrp` representing eight age–group categories, and a 0/1 variable `gender`.

```
. set seed 12345
. generate agegrp = runiformint(1, 8)
. generate gender = runiformint(0, 1)
```

We want to split the data into four samples, where the first three samples are the same size, and the fourth sample is twice the size of each of the others. We specify `split(1 1 1 2)` using integer ratios. We specify the option `balance(agegrp gender)` to ensure that the distribution of `agegrp × gender` is roughly balanced across the four samples. The option `show` is useful for seeing the actual splits of the numbers of observations within each `agegrp × gender` group.
. splitsample, generate(svar, replace) split(1 1 1 2)
> balance(agegrp gender) show
note: some groups defined by balance() do not contain every sample value.

<table>
<thead>
<tr>
<th>agegrp</th>
<th>gender</th>
<th>svar 1</th>
<th>svar 2</th>
<th>svar 3</th>
<th>svar 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>11</td>
</tr>
</tbody>
</table>

We get a message “some groups defined by balance() do not contain every sample value”. Indeed, all the groups of size three have no observations in sample 2. Because we are splitting the data into four samples, obviously we need at least four observations in a group for every sample to contain at least one observation.

Second, we notice that all groups of the same size are split into the four samples with exactly the same number of observations in each sample. For example, the two groups of size eight (agegrp = 1, gender = 0 and agegrp = 8, gender = 0) both have two observations in each of samples 1 and 3, one observation in sample 2, and three observations in sample 4.

Groups of the same size have exactly the same sample-size splits because, by default, the sample sizes for the splits are calculated using a deterministic formula. If the sizes of the groups vary, this typically would not be an issue. Overall, one would expect the actual split proportions to be close to the specified split proportions. But imagine if all, or almost all, the group sizes were the same. What if the size of each group were eight observations in this example? Every group would be split $2 : 1 : 2 : 3$ by observations, yielding actual split proportions of $25\% : 12.5\% : 25\% : 37.5\%$, which are rather different from the specified split proportions of $20\% : 20\% : 20\% : 40\%$. 
The option `rround` provides a solution for this problem. It randomly rounds the split sample sizes when the split cannot be made exactly.

```
. splitsample, generate(svar, replace) split(1 1 1 2) 
> balance(agegrp gender) rround rseed(54321) show 
```

Note: some groups defined by `balance()` do not contain every sample value.

<table>
<thead>
<tr>
<th>agegrp</th>
<th>gender</th>
<th>svar 1</th>
<th>svar 2</th>
<th>svar 3</th>
<th>svar 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>11</td>
</tr>
</tbody>
</table>

We see that the groups of sizes three, eight, and nine now have different splits by numbers of observations. The groups of size five have exactly the same splits by size because they could be divided exactly based on the specified split ratios of 1:1:1:2.

The option `rround` with `balance()` thus does a “more random” assignment of observations (or clusters), which is important when the sizes of the balance groups are small. When the sizes of the balance groups are large, and the sizes of the groups vary, splits made with or without `rround` will be similar.

Note that `rround` with `balance()` is suitable for random treatment assignment with matching defined by values of the balance variables.

The computational procedure for option `rround` first randomly assigns as many observations to the split samples as it can to match the specified split proportions exactly. Leftover observations are assigned to samples by dividing them randomly based on the specified split ratios. Splitting ratios must be specified as integers to facilitate this method of splitting the leftovers. See Methods and formulas.

### Example 4: Missing values

`varlist` can be specified with `splitsample` to handle missing values. Let’s say we want to divide our data into training and validation samples for a lasso or other procedure. Imagine that the variables in the lasso have more than a few missing values. Specifying these variables as `varlist` for
splitsample means that the sample ID variable created will have missing values whenever any of the variables in `varlist` are missing.

Here’s an illustration. We create a couple of variables with missing values.

```
  . set seed 1234
  . generate y = runiform()
  . replace y = . if runiform() < 0.1
    (11 real changes made, 11 to missing)
  . generate x = runiform()
  . replace x = . if runiform() < 0.1
    (15 real changes made, 15 to missing)
```

Then split the data specifying these variables to be checked for missing:

```
  . splitsample y x, generate(svar, replace)
  . tabulate svar, miss
```

```
  svar | Freq. | Percent | Cum.  |
  -----|-------|---------|-------|
  1    | 38    | 37.62   | 37.62 |
  2    | 38    | 37.62   | 75.25 |
  .     | 25    | 24.75   | 100.00|
  Total | 101   | 100.00  |
```

The split was done exactly for the observations without missing values.

### Stored results

`splitsample` stores the following in `r()`:

Scalars

- `r(N)` total number of observations
- `r(N_clust)` total number of clusters
- `r(n_samples)` number of split samples

Macros

- `r(clustvar)` name of cluster variable
- `r(balancevars)` names of balance variables
- `r(rngstate)` random-number state used

### Methods and formulas

Let $r_1, r_2, \ldots, r_K$ be the arguments to `split(numlist)`. If the split is specified using `nssplit(#), then we set each $r_k = 1$, and the number of split samples is $K = #$. The split sample proportions are

$$ p_k = \frac{r_k}{R} \quad \text{where} \quad R = \sum_{i=1}^{K} r_i $$

The cumulative proportions are

$$ s_k = \sum_{i=1}^{k} p_i $$
For the default deterministic rounding, we calculate cumulative sample sizes:

\[ M_k = \text{round}(N s_k) \]

where \( N \) is the total number of observations or the number of clusters, and \text{round}() is Stata's \text{round()} function. When the option \text{balance()} is specified, \( N \) is the number of observations or clusters in a single balance group. The sample sizes \( N_1, N_2, \ldots, N_K \) are given by

\[
N_1 = M_1 \\
N_k = M_k - M_{k-1} \quad \text{for} \ k = 2, \ldots, K
\]

When the option \text{rround} is specified for random rounding, we first divide \( N \), the number of observations or clusters, as follows:

\[ N = cR + d \]

where \( R \) is the sum of \( r_1, r_2, \ldots, r_K \); \( c \) is a nonnegative integer; and \( 0 \leq d < R \). In other words, \( cR \) observations can be split into \( K \) samples matching the specified split proportions exactly. We randomly pick \( cR \) observations and assign them to the samples. The leftover \( d \) observations are randomly placed in \( R \) bins without replacement, where the first \( r_1 \) bins represent sample 1, the next \( r_2 \) bins represent sample 2, and so on.

The computational procedure for random rounding thus requires \( r_1, r_2, \ldots, r_K \) to be integers and also requires \( R \leq N \). To reduce the variance of the random rounding, the integers \( r_1, r_2, \ldots, r_K \) should have no common factors.

\textbf{Also see}

[D] \text{sample} — Draw random sample