# Title

stata.com

**cluster stop —** Cluster-analysis stopping rules



## Description

`cluster stop` and `clustermat stop` compute the stopping-rule value for each cluster solution. The commands currently provide two stopping rules, the Caliński and Harabasz pseudo-$F$ index and the Duda–Hart Je(2)/Je(1) index. For both rules, larger values indicate more distinct clustering. Presented with the Duda–Hart Je(2)/Je(1) values are pseudo-$T$-squared values. Smaller pseudo-$T$-squared values indicate more distinct clustering.

Users can add more `stop` rules; see [MV] **cluster programming subroutines**.

## Quick start

*Cluster analysis of data*

Caliński–Harabasz pseudo-$F$ index stopping rule for the most recent `cluster` analysis

```
cluster stop
```

Duda–Hart Je(2)/Je(1) index stopping rule

```
cluster stop, rule(duda)
```

As above, but use results for the 5–20-group solutions instead of the default 1–15-group solutions

```
cluster stop, rule(duda) groups(5/20)
```

As above, but for cluster analysis results named `myclus`

```
cluster stop myclus, rule(duda) groups(5/20)
```

As above, but use variables `v1` and `v2` to compute the stopping rule instead of the variables used in `myclus`

```
cluster stop myclus, rule(duda) groups(5/20) variables(v1 v2)
```

*Cluster analysis of dissimilarity matrix*

Caliński–Harabasz pseudo-$F$ index stopping rule computed using `v1`, `v2`, and `v3` from results named `mymatclus`

```
clustermat stop mymatclus, variables(v1 v2 v3)
```

## Menu

Statistics > Multivariate analysis > Cluster analysis > Postclustering > Cluster analysis stopping rules

# Syntax

*Cluster analysis of data*

```
cluster stop [clname] [, options]
```

*Cluster analysis of a dissimilarity matrix*

```
clustermat stop [clname] , variables(varlist) [ options ]
```

where *clname* is the name of the cluster analysis. The default is the most recently performed cluster analysis, which can be reset using the `cluster use` command; see [MV] **cluster utility**.

| *options* | Description |
|---|---|
| `rule(calinski)` | use Caliński–Harabasz pseudo-$F$ index stopping rule; the default |
| `rule(duda)` | use Duda–Hart Je(2)/Je(1) index stopping rule |
| `rule(`*rule_name*`)` | use *rule_name* stopping rule; see *Options* for details |
| `groups(`*numlist*`)` | compute stopping rule for specified groups |
| `matrix(`*matname*`)` | save results in matrix *matname* |
| * `variables(`*varlist*`)` | compute the stopping rule using *varlist* |

* `variables(`*varlist*`)` is required with a `clustermat` solution and optional with a `cluster` solution.
`rule(`*rule_name*`)` is not shown in the dialog box. See [MV] **cluster programming subroutines** for information on how to add stopping rules to the `cluster stop` command.

# Options

`rule(calinski | duda | `*rule_name*`)` indicates the stopping rule. `rule(calinski)`, the default, specifies the Caliński–Harabasz pseudo-$F$ index. `rule(duda)` specifies the Duda–Hart Je(2)/Je(1) index.

`rule(calinski)` is allowed for both hierarchical and nonhierarchical cluster analyses. `rule(duda)` is allowed only for hierarchical cluster analyses.

You can add stopping rules to the `cluster stop` command (see [MV] **cluster programming subroutines**) by using the `rule(`*rule_name*`)` option. [MV] **cluster programming subroutines** illustrates how to add stopping rules by showing a program that adds a `rule(stepsize)` option, which implements the simple step-size stopping rule mentioned in Milligan and Cooper (1985).

`groups(`*numlist*`)` specifies the cluster groupings for which the stopping rule is to be computed. `groups(3/20)` specifies that the measure be computed for the three-group solution, the four-group solution, ..., and the 20-group solution.

With `rule(duda)`, the default is `groups(1/15)`. With `rule(calinski)` for a hierarchical cluster analysis, the default is `groups(2/15)`. `groups(1)` is not allowed with `rule(calinski)` because the measure is not defined for the degenerate one-group cluster solution. The `groups()` option is unnecessary (and not allowed) for a nonhierarchical cluster analysis.

If there are ties in the hierarchical cluster-analysis structure, some (or possibly all) of the requested stopping-rule solutions may not be computable. `cluster stop` passes over, without comment, the `groups()` for which ties in the hierarchy cause the stopping rule to be undefined.

`matrix(`*matname*`)` saves the results in a matrix named *matname*.

With `rule(calinski)`, the matrix has two columns, the first giving the number of clusters and the second giving the corresponding Caliński–Harabasz pseudo-$F$ stopping-rule index.

With `rule(duda)`, the matrix has three columns: the first column gives the number of clusters, the second column gives the corresponding Duda–Hart Je(2)/Je(1) stopping-rule index, and the third column provides the corresponding pseudo-$T$-squared values.

`variables(`*varlist*`)` specifies the variables to be used in the computation of the stopping rule. By default, the variables used for the cluster analysis are used. `variables()` is required for cluster solutions produced by `clustermat`.

## Remarks and examples

stata.com

Cluster-analysis stopping rules are used to determine the number of clusters. A stopping-rule value (also called an index) is computed for each cluster solution (for example, at each level of the hierarchy in a hierarchical cluster analysis). Larger values (or smaller, depending on the particular stopping rule) indicate more distinct clustering. See [MV] **cluster** for background information on cluster analysis and on the `cluster` and `clustermat` commands.

Everitt et al. (2011) and Gordon (1999) discuss the problem of determining the number of clusters and describe several stopping rules, including the Caliński–Harabasz (1974) pseudo-$F$ index and the Duda–Hart (2001, sec. 10.10) Je(2)/Je(1) index. There are many cluster stopping rules. Milligan and Cooper (1985) evaluate 30 stopping rules, singling out the Caliński–Harabasz index and the Duda–Hart index as two of the best rules.

Large values of the Caliński–Harabasz pseudo-$F$ index indicate distinct clustering. The Duda–Hart Je(2)/Je(1) index has an associated pseudo-$T$-squared value. A large Je(2)/Je(1) index value and a small pseudo-$T$-squared value indicate distinct clustering. See *Methods and formulas* at the end of this entry for details.

Example 2 of [MV] **clustermat** shows the use of the `clustermat stop` command.

Some stopping rules such as the Duda–Hart index work only with a hierarchical cluster analysis. The Caliński–Harabasz index, however, may be applied to both nonhierarchical and hierarchical cluster analyses.

### ▷ Example 1

Previously, you ran kmeans cluster analyses on data where you measured the flexibility, speed, and strength of the 80 students in your physical education class; see example 1 of [MV] **cluster kmeans and kmedians**. Your original goal was to split the class into four groups, though you also examined the three- and five-group kmeans cluster solutions as possible alternatives.

Now out of curiosity, you wonder what the Caliński–Harabasz stopping rule shows for the three-, four-, and five-group solutions from a kmedian clustering of this dataset.

```
. use http://www.stata-press.com/data/r15/physed
. cluster kmed flex speed strength, k(3) name(kmed3) measure(abs) start(lastk)
. cluster kmed flex speed strength, k(4) name(kmed4) measure(abs)
> start(kr(93947))
. cluster kmed flex speed strength, k(5) name(kmed5) measure(abs)
> start(prand(16872))
```

```
. cluster stop kmed3
```

| Number of clusters | Calinski/ Harabasz pseudo-F |
|---|---|
| 3 | 132.75 |

```
. cluster stop kmed4
```

| Number of clusters | Calinski/ Harabasz pseudo-F |
|---|---|
| 4 | 337.10 |

```
. cluster stop kmed5
```

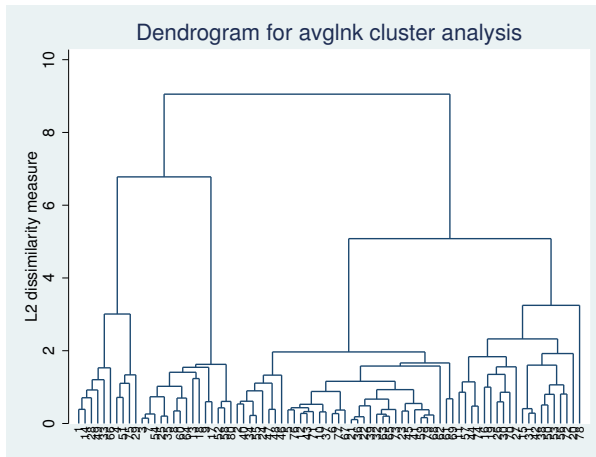
| Number of clusters | Calinski/ Harabasz pseudo-F |
|---|---|
| 5 | 300.45 |

The four-group solution with a Caliński–Harabasz pseudo-$F$ value of 337.10 is largest, indicating that the four-group solution is the most distinct compared with the three-group and five-group solutions.

The three-group solution has a much lower stopping-rule value of 132.75. The five-group solution, with a value of 300.45, is reasonably close to the four-group solution.

Though you do not think it will change your decision on how to split your class into groups, you are curious to see what a hierarchical cluster analysis might produce. You decide to try an average-linkage cluster analysis using the default Euclidean distance; see [MV] **cluster linkage**. You examine the resulting cluster analysis with the `cluster tree` command, which is an easier-to-type alias for the `cluster dendrogram` command; see [MV] **cluster dendrogram**.

```
. cluster averagelink flex speed strength, name(avglnk)
. cluster tree avglnk, xlabel(, angle(90) labsize(*.75))
```



You are curious to see how the four- and five-group solutions from this hierarchical cluster analysis compare with the four- and five-group solutions from the kmedian clustering.

```
. cluster gen avgg = groups(4/5), name(avglnk)
. table kmed4 avgg4
```

| kmed4 | avgg4 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 10 | | | |
| 2 | | | 35 | |
| 3 | | 15 | | |
| 4 | | | | 20 |

```
. table kmed5 avgg5
```

| kmed5 | avgg5 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | | 15 | | |
| 2 | | 15 | | | |
| 3 | | | | 19 | 1 |
| 4 | | | 20 | | |
| 5 | 10 | | | | |

The four-group solutions are identical, except for the numbers used to label the groups. The five-group solutions are different. The kmedian clustering split the 35-member group into subgroups having 20 and 15 members. The average-linkage clustering instead split one member off from the 20-member group.

Now you examine the Caliński–Harabasz pseudo-$F$ stopping-rule values associated with the kmedian hierarchical cluster analysis.

```
. cluster stop avglnk, rule(calinski)
```

| Number of clusters | Calinski/ Harabasz pseudo-F |
|---|---|
| 2 | 131.86 |
| 3 | 126.62 |
| 4 | 337.10 |
| 5 | 269.07 |
| 6 | 258.40 |
| 7 | 259.37 |
| 8 | 290.78 |
| 9 | 262.86 |
| 10 | 258.53 |
| 11 | 249.93 |
| 12 | 247.85 |
| 13 | 247.53 |
| 14 | 236.98 |
| 15 | 226.51 |

Because `rule(calinski)` is the default, you could have obtained this same table by typing

```
. cluster stop avglnk
```

or, because `avglnk` was the most recent cluster analysis performed, by typing

```
. cluster stop
```

You did not specify the number of groups to examine from the hierarchical cluster analysis, so it defaulted to examining up to 15 groups. The highest Caliński–Harabasz pseudo-$F$ value is 337.10 for the four-group solution.

What does the Duda–Hart stopping rule produce for this hierarchical cluster analysis?

```
. cluster stop avglnk, rule(duda) groups(1/10)
```

| Number of clusters | Duda/Hart Je(2)/Je(1) | Duda/Hart pseudo T-squared |
|---|---|---|
| 1 | 0.3717 | 131.86 |
| 2 | 0.1349 | 147.44 |
| 3 | 0.2283 | 179.19 |
| 4 | 0.8152 | 4.08 |
| 5 | 0.2232 | 27.85 |
| 6 | 0.5530 | 13.74 |
| 7 | 0.5287 | 29.42 |
| 8 | 0.6887 | 3.16 |
| 9 | 0.4888 | 8.37 |
| 10 | 0.7621 | 7.80 |

This time, we asked to see the results for one to 10 groups. The largest Duda–Hart Je(2)/Je(1) stopping-rule value is 0.8152, corresponding to four groups. The smallest pseudo-$T$-squared value is 3.16 for the eight-group solution, but the pseudo-$T$-squared value for the four-group solution is also low, with a value of 4.08.

Distinct clustering is characterized by large Caliński–Harabasz pseudo-$F$ values, large Duda–Hart Je(2)/Je(1) values, and small Duda–Hart pseudo-$T$-squared values.

The conventional wisdom for deciding the number of groups based on the Duda–Hart stopping-rule table is to find one of the largest Je(2)/Je(1) values that corresponds to a low pseudo-$T$-squared value that has much larger $T$-squared values next to it. This strategy, combined with the results from the Caliński–Harabasz results, indicates that the four-group solution is the most distinct from this hierarchical cluster analysis.

◁

## ❑ Technical note

There is a good reason that the word "pseudo" appears in "pseudo-$F$" and "pseudo-$T$-squared". Although these index values are based on well-known statistics, any $p$-values computed from these statistics would not be valid. Remember that cluster analysis searches for structure.

If you were to generate random observations, perform a cluster analysis, compute these stopping-rule statistics, and then follow that by computing what would normally be the $p$-values associated with the statistics, you would almost always end up with significant $p$-values.

Remember that you would expect, on average, five of every 100 groupings of your random data to show up as significant when you use .05 as your threshold for declaring significance. Cluster-analysis methods search for the best groupings, so there is no surprise that $p$-values show high significance, even when none exists.

Examining the stopping-rule index values relative to one another is useful, however, in finding relatively reasonable groupings that may exist in the data.

❑

❑ Technical note

As mentioned in *Methods and formulas*, ties in the hierarchical cluster structure cause some of the stopping-rule index values to be undefined. Discrete (as opposed to continuous) data tend to cause ties in a hierarchical clustering. The more discrete the data, the more likely it is that ties will occur (and the more of them you will encounter) within a hierarchy.

Even with so-called continuous data, ties in the hierarchical clustering can occur. We say "so-called" because most continuous data are truncated or rounded. For instance, miles per gallon, length, weight, etc., which may really be continuous, may be observed and recorded only to the tens, ones, tenths, or hundredths of a unit.

You can have data with no ties in the observations and still have many ties in the hierarchy. Ties in distances (or similarities) between observations and groups of observations cause the ties in the hierarchy.

Thus, do not be surprised when some (many) of the stopping-rule values that you request are not presented. Stata has decided not to break the ties arbitrarily, because the stopping-rule values may differ widely, depending on which split is made.

❑

❑ Technical note

The stopping rules also become less informative as the number of elements in the groups becomes small, that is, having many groups, each with few observations. We recommend that if you need to examine the stopping-rule values deep within your hierarchical cluster analysis, you do so skeptically.

❑

## Stored results

`cluster stop` and `clustermat stop` with `rule(calinski)` stores the following in `r()`:

Scalars

| | |
|---|---|
| `r(calinski_#)` | Caliński–Harabasz pseudo-$F$ for # groups |

Macros

| | |
|---|---|
| `r(rule)` | `calinski` |
| `r(label)` | `C-H pseudo-F` |
| `r(longlabel)` | `Calinski & Harabasz pseudo-F` |

`cluster stop` and `clustermat stop` with `rule(duda)` stores the following in `r()`:

Scalars

| | |
|---|---|
| `r(duda_#)` | Duda–Hart Je(2)/Je(1) value for # groups |
| `r(dudat2_#)` | Duda–Hart pseudo-$T$-squared value for # groups |

Macros

| | |
|---|---|
| `r(rule)` | `duda` |
| `r(label)` | `D-H Je(2)/Je(1)` |
| `r(longlabel)` | `Duda & Hart Je(2)/Je(1)` |
| `r(label2)` | `D-H pseudo-T-squared` |
| `r(longlabel2)` | `Duda & Hart pseudo-T-squared` |

## Methods and formulas

The Caliński–Harabasz pseudo-$F$ stopping-rule index for $g$ groups and $N$ observations is

$$\frac{\text{trace}(\mathbf{B})/(g-1)}{\text{trace}(\mathbf{W})/(N-g)}$$

where $\mathbf{B}$ is the between-cluster sum of squares and cross-products matrix, and $\mathbf{W}$ is the within-cluster sum of squares and cross-products matrix.

Large values of the Caliński–Harabasz pseudo-$F$ stopping-rule index indicate distinct cluster structure. Small values indicate less clearly defined cluster structure.

The Duda–Hart Je(2)/Je(1) stopping-rule index value is literally Je(2) divided by Je(1). Je(1) is the sum of squared errors within the group that is to be divided. Je(2) is the sum of squared errors in the two resulting subgroups.

Large values of the Duda–Hart pseudo-$T$-squared stopping-rule index indicate distinct cluster structure. Small values indicate less clearly defined cluster structure.

The Duda–Hart Je(2)/Je(1) index requires hierarchical clustering information. It needs to know at each level of the hierarchy which group is to be split and how. The Duda–Hart index is also local because the only information used comes from the group's being split. The information in the rest of the groups does not enter the computation.

In comparison, the Caliński–Harabasz rule does not require hierarchical information and is global because the information from each group is used in the computation.

A pseudo-$T$-squared value is also presented with the Duda and Hart Je(2)/Je(1) index. The relationship is

$$\frac{1}{\text{Je}(2)/\text{Je}(1)} = 1 + \frac{T^2}{N_1 + N_2 - 2}$$

where $N_1$ and $N_2$ are the numbers of observations in the two subgroups.

Je(2)/Je(1) will be zero when Je(2) is zero, that is, when the two subgroups each have no variability. An example of this is when the cluster being split has two distinct values that are being split into singleton subgroups. Je(1) will never be zero because we do not split groups that have no variability. When Je(2)/Je(1) is zero, the pseudo-$T$-squared value is undefined.

Ties in splitting a hierarchical cluster analysis create an ambiguity for the Je(2)/Je(1) measure. For example, to compute the measure for the case of going from five clusters to six, you need to identify the one cluster that will be split. With a tie in the hierarchy, you would instead go from five clusters directly to seven (just as an example). Stata refuses to produce an answer in this situation.

## References

Caliński, T., and J. Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics—Theory and Methods* 3: 1–27.

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. *Pattern Classification*. 2nd ed. New York: Wiley.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl. 2011. *Cluster Analysis*. 5th ed. Chichester, UK: Wiley.

Gordon, A. D. 1999. *Classification*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Milligan, G. W., and M. C. Cooper. 1985. An examination of procedures for determining the number of clusters in a dataset. *Psychometrika* 50: 159–179.

## Also see

[MV] **cluster** — Introduction to cluster-analysis commands

[MV] **clustermat** — Introduction to clustermat commands