QUICK CLUSTER

QUICK CLUSTER is available in the Statistics Base option.

```
QUICK CLUSTER {varlist}
              {ALL    }

 [/MISSING=[{LISTWISE**}] [INCLUDE]]
            {PAIRWISE  }
            {DEFAULT   }

 [/FILE='savfile'|'dataset']

 [/INITIAL=(value list)]

 [/CRITERIA=[CLUSTER({2**})][NOINITIAL][MXITER({10**})] [CONVERGE({0**})]]
                     {n  }                     {n   }             {n  }

 [/METHOD=[{KMEANS[(NOUPDATE)]**}]
           {KMEANS(UPDATE)}     }
           {CLASSIFY            }

 [/PRINT=[INITIAL**] [CLUSTER] [ID(varname)] [DISTANCE] [ANOVA] [NONE]]

 [/OUTFILE='savfile'|'dataset']

 [/SAVE=[CLUSTER[(varname)]] [DISTANCE[(varname)]]]
```

**Default if subcommand or keyword is omitted.

This command reads the active dataset and causes execution of any pending commands. For more information, see the topic Command Order on p. 41.

Example

```
QUICK CLUSTER V1 TO V4
  /CRITERIA=CLUSTER(4)
  /SAVE=CLUSTER(GROUP).
```

Overview

When the desired number of clusters is known, QUICK CLUSTER groups cases efficiently into clusters. It is not as flexible as CLUSTER, but it uses considerably less processing time and memory, especially when the number of cases is large.

Options

Algorithm Specifications. You can specify the number of clusters to form with the CRITERIA subcommand. You can also use CRITERIA to control initial cluster selection and the criteria for iterating the clustering algorithm. With the METHOD subcommand, you can specify how to update cluster centers, and you can request classification only when working with very large data files.

Initial Cluster Centers. By default, QUICK CLUSTER chooses the initial cluster centers. Alternatively, you can provide initial centers on the INITIAL subcommand. You can also read initial cluster centers from IBM® SPSS® Statistics data files using the FILE subcommand.

Optional Output. With the PRINT subcommand, you can display the cluster membership of each case and the distance of each case from its cluster center. You can also display the distances between the final cluster centers and a univariate analysis of variance between clusters for each clustering variable.

Saving Results. You can write the final cluster centers to a data file using the OUTFILE subcommand. In addition, you can save the cluster membership of each case and the distance from each case to its classification cluster center as new variables in the active dataset using the SAVE subcommand.

Basic Specification

The basic specification is a list of variables. By default, QUICK CLUSTER produces two clusters. The two cases that are farthest apart based on the values of the clustering variables are selected as initial cluster centers and the rest of the cases are assigned to the nearer center. The new cluster centers are calculated as the means of all cases in each cluster, and if neither the minimum change nor the maximum iteration criterion is met, all cases are assigned to the new cluster centers again. When one of the criteria is met, iteration stops, the final cluster centers are updated, and the distance of each case is computed.

Subcommand Order

- The variable list must be specified first.
- Subcommands can be named in any order.

Operations

The procedure generally involves four steps:

- First, initial cluster centers are selected, either by choosing one case for each cluster requested or by using the specified values.
- Second, each case is assigned to the nearest cluster center, and the mean of each cluster is calculated to obtain the new cluster centers.
- Third, the maximum change between the new cluster centers and the initial cluster centers is computed. If the maximum change is not less than the minimum change value and the maximum iteration number is not reached, the second step is repeated and the cluster centers are updated. The process stops when either the minimum change or maximum iteration criterion is met. The resulting clustering centers are used as classification centers in the last step.
- In the last step, all cases are assigned to the nearest classification center. The final cluster centers are updated and the distance for each case is computed.

When the number of cases is large, directly clustering all cases may be impractical. As an alternative, you can cluster a sample of cases and then use the cluster solution for the sample to classify the entire group. This can be done in two phases:

- The first phase obtains a cluster solution for the sample. This involves all four steps of the QUICK CLUSTER algorithm. OUTFILE then saves the final cluster centers to a data file.

- The second phase requires only one pass through the data. First, the FILE subcommand specifies the file containing the final cluster centers from the first analysis. These final cluster centers are used as the initial cluster centers for the second analysis. CLASSIFY is specified on the METHOD subcommand to skip the second and third steps of the clustering algorithm, and cases are classified using the initial cluster centers. When all cases are assigned, the cluster centers are updated and the distance of each case is computed. This phase can be repeated until final cluster centers are stable.

Example

```
QUICK CLUSTER
  zlnlong zlntoll zlnequi zlncard zlnwire zmultlin zvoice zpager zinterne
  zcallid zcallwai zforward zconfer zebill
  /MISSING=PAIRWISE
  /CRITERIA= CLUSTER(3) MXITER(20) CONVERGE(0)
  /METHOD=KMEANS(NOUPDATE)
  /PRINT INITIAL ANOVA DISTAN.
```

- The procedure clusters cases based upon their values for variables *zlnlong* through *zebill*.
- The MISSING subcommand specifies that a case is assigned to clusters based upon all clustering variables for which the case has nonmissing values.
- The CRITERIA subcommand specifies that three clusters will be formed, and that 20 iterations are allowed for updating cluster centers.
- The PRINT subcommand specifies that the initial cluster centers, distances between final cluster centers, and an ANOVA table of descriptive *F* tests for the clustering variables should be displayed.
- All other options are set to their default values.

Variable List

The variable list identifies the clustering variables.

- The variable list is required and must be the first specification on QUICK CLUSTER.
- You can use keyword ALL to refer to all user-defined variables in the active dataset.
- QUICK CLUSTER uses squared Euclidean distances, which equally weight all clustering variables. If the variables are measured in units that are not comparable, the procedure will give more weight to variables with large variances. Therefore, you should standardize variables measured on different scales using procedure DESCRIPTIVES before performing QUICK CLUSTER.

CRITERIA Subcommand

CRITERIA specifies the number of clusters to form and controls options for the clustering algorithm. You can use any or all of the keywords below.

- The NOINITIAL option followed by the remaining steps of the default QUICK CLUSTER algorithm makes QUICK CLUSTER equivalent to MacQueen's *n*-means clustering method.

CLUSTER(n) *Number of clusters.* QUICK CLUSTER assigns cases to *n* clusters. The default is 2.

NOINITIAL *No initial cluster center selection.* By default, initial cluster centers are formed by choosing one case (with valid data for the clustering variables) for each cluster requested. The initial selection requires a pass through the data to ensure that the centers are well separated from one another. If NOINITIAL is specified, QUICK CLUSTER selects the first *n* cases without missing values as initial cluster centers.

MXITER(n) *Maximum number of iterations for updating cluster centers.* The default is 10. Iteration stops when the maximum number of iterations has been reached. MXITER is ignored when METHOD=CLASSIFY.

CONVERGE(n) *Convergence criterion controlling minimum change in cluster centers.* The default value for *n* is 0. The minimum change value equals the convergence value (*n*) times the minimum distance between initial centers. Iteration stops when the largest change of any cluster center is less than or equal to the minimum change value. CONVERGE is ignored when METHOD=CLASSIFY.

METHOD Subcommand

By default, QUICK CLUSTER recalculates cluster centers after assigning all the cases and repeats the process until one of the criteria is met. You can use the METHOD subcommand to recalculate cluster centers after each case is assigned or to suppress recalculation until after classification is complete. When METHOD=KMEANS is specified, QUICK CLUSTER displays the iteration history table.

KMEANS(NOUPDATE) *Recalculate cluster centers after all cases are assigned for each iteration.* This is the default.

KMEANS(UPDATE) *Recalculate a cluster center each time a case is assigned.* QUICK CLUSTER calculates the mean of cases currently in the cluster and uses this new cluster center in subsequent case assignment.

CLASSIFY *Do not recalculate cluster centers.* QUICK CLUSTER uses the initial cluster centers for classification and computes the final cluster centers as the means of all the cases assigned to the same cluster. When CLASSIFY is specified, the CONVERGE or MXITER specifications on CRITERIA are ignored.

INITIAL Subcommand

INITIAL specifies the initial cluster centers. Initial cluster centers can also be read from a data file (see FILE Subcommand on p. 1674).

- One value for each clustering variable must be included for each cluster requested. Values are specified in parentheses cluster by cluster.

Example

```
QUICK CLUSTER  A B C D
  /CRITERIA = CLUSTER(3)
  /INITIAL = (13 24  1  8
               7 12  5  9
              10 18 17 16).
```

- This example specifies four clustering variables and requests three clusters. Thus, twelve values are supplied on `INITIAL`.
- The initial center of the first cluster has a value of 13 for variable *A*, 24 for variable *B*, 1 for *C*, and 8 for *D*.

FILE Subcommand

Use `FILE` to obtain initial cluster centers from an external IBM® SPSS® Statistics data file or currently open dataset. (`DATASET DECLARE` command).

- The only specification is the quoted file specification or dataset name.

Example

```
QUICK CLUSTER  A B C D
  /FILE='/data/init.sav'
  /CRITERIA = CLUSTER(3).
```

- In this example, the initial cluster centers are read from file *init.sav*. The file must contain cluster centers for the same four clustering variables specified (*A*, *B*, *C*, and *D*).

PRINT Subcommand

`QUICK CLUSTER` always displays in a Final Cluster Centers table listing the centers used to classify cases and the mean values of the cases in each cluster and a Number of Cases in Each Cluster table listing the number of weighted (if weighting is on) and unweighted cases in each cluster. Use `PRINT` to request other types of output.

- If `PRINT` is not specified or is specified without keywords, the default is `INITIAL`.

INITIAL *Initial cluster centers.* When `SPLIT FILES` is in effect, the initial cluster center for each split file is displayed. This is the default.

CLUSTER *Cluster membership.* Each case displays an identifying number or value, the number of the cluster to which it was assigned, and its distance from the center of that cluster. This output is extensive when the number of cases is large.

ID(varname) *Case identification.* The value of the specified variable is used in addition to the case numbers to identify cases in output. Case numbers may not be sequential if cases have been selected.

DISTANCE *Pairwise distances between all final cluster centers.* This output can consume a great deal of processing time when the number of clusters requested is large.

ANOVA *Descriptive univariate* F *tests for the clustering variables.* Since cases are systematically assigned to clusters to maximize differences on the clustering variables, these tests are descriptive only and should not be used to test the null hypothesis that there are no differences between clusters. Statistics after clustering are also available through procedure DISCRIMINANT or GLM (GLM is available in the Advanced Statistics option).

NONE *No additional output.* Only the default output is displayed. NONE overrides any other specifications on PRINT.

Example

```
QUICK CLUSTER A B C D E
  /CRITERIA=CLUSTERS(6)
  /PRINT=CLUSTER ID(CASEID) DISTANCE.
```

- Six clusters are formed on the basis of the five variables *A*, *B*, *C*, *D*, and *E*.
- For each case in the file, cluster membership and distance from cluster center are displayed. Cases are identified by the values of the variable *CASEID*.
- Distances between all cluster centers are printed.

OUTFILE Subcommand

OUTFILE saves the final cluster centers in an external IBM® SPSS® Statistics data file or a previously declared dataset in the current session. You can later use these final cluster centers as initial cluster centers for a different sample of cases that use the same variables. You can also cluster the final cluster centers themselves to obtain clusters of clusters.

- The only specification is a filename or previously declared dataset name for the file. Filenames should be enclosed in quotes and are stored in the working directory unless a path is included as part of the file specification. Datasets are available during the current session but are not available in subsequent sessions unless you explicitly save them as data files.
- The program displays the name of the saved file in the procedure information notes.

Example

```
QUICK CLUSTER A B C D
  /CRITERIA = CLUSTER(3)
  /OUTFILE = '/data/QC1.sav'.
```

- QUICK CLUSTER writes the final cluster centers to the file *QC1.sav*.

SAVE Subcommand

Use SAVE to save results of cluster analysis as new variables in the active dataset.

- You can specify a variable name in parentheses following either keyword. If no variable name is specified, QUICK CLUSTER forms unique variable names by appending an underscore and a sequential number to the rootname *QCL*. The number increments with each new variable saved.

- The program displays the new variables and a short description of each in the procedure information notes.

CLUSTER[(varname)] *The cluster number of each case.* The value of the new variable is set to an integer from 1 to the number of clusters.

DISTANCE[(varname)] *The distance of each case from its classification cluster center.*

Example

```
QUICK CLUSTER A B C D
  /CRITERIA=CLUSTERS(6)
  /SAVE=CLUSTER DISTANCE.
```

- Six clusters of cases are formed on the basis of the variables *A*, *B*, *C*, and *D*.
- A new variable *QCL_1* is created and set to an integer between 1 and 6 to indicate cluster membership for each case.
- Another new variable *QCL_2* is created and set to the Euclidean distance between a case and the center of the cluster to which it is assigned.

MISSING Subcommand

`MISSING` controls the treatment of cases with missing values.

- `LISTWISE`, `PAIRWISE`, and `DEFAULT` are alternatives. However, each can be used with `INCLUDE`.

LISTWISE *Delete cases with missing values listwise.* A case with a missing value for any of the clustering variables is deleted from the analysis and will not be assigned to a cluster. This is the default.

PAIRWISE *Assign each case to the nearest cluster on the basis of the clustering variables for which the case has nonmissing values.* Only cases with missing values for *all* clustering variables are deleted.

INCLUDE *Treat user-missing values as valid.*

DEFAULT *Same as LISTWISE.*