Multivariate Analysis

Table Of Contents

© 2003 Minitab Inc.

Multivariate Analysis

Overview

Multivariate Analysis Overview

Use Minitab's multivariate analysis procedures to analyze your data when you have made multiple measurements on items or subjects. You can choose to:

- Analyze the data covariance structure to understand it or to reduce the data dimension
- Assign observations to groups
- Explore relationships among categorical variables

Because Minitab does not compare tests of significance for multivariate procedures, interpreting the results is somewhat subjective. However, you can make informed conclusions if you are familiar with your data.

Analysis of the data structure

Minitab offers two procedures for analyzing the data covariance structure:

- Principal Components helps you to understand the covariance structure in the original variables and/or to create a smaller number of variables using this structure.
- Factor Analysis, like principal components, summarizes the data covariance structure in a smaller number of dimensions. The emphasis in factor analysis is the identification of underlying "factors" that might explain the dimensions associated with large data variability.

Grouping observations

Minitab offers three cluster analysis methods and discriminant analysis for grouping observations:

- Cluster Observations groups or clusters observations that are "close" to each other when the groups are initially unknown. This method is a good choice when no outside information about grouping exists. The choice of final grouping is usually made according to what makes sense for your data after viewing clustering statistics.
- Cluster Variables groups or clusters variables that are "close" to each other when the groups are initially unknown. The procedure is similar to clustering of observations. You may want to cluster variables to reduce their number.
- Cluster K-Means, like clustering of observations, groups observations that are "close" to each other. K-means clustering works best when sufficient information is available to make good starting cluster designations.
- Discriminant Analysis classifies observations into two or more groups if you have a sample with known groups. You can use discriminant analysis to investigate how the predictors contribute to the groupings.

Correspondence Analysis

Minitab offers two methods of correspondence analysis to explore the relationships among categorical variables:

- Simple Correspondence Analysis explores relationships in a 2-way classification. You can use this procedure with 3-way and 4-way tables because Minitab can collapse them into 2-way tables. Simple correspondence analysis decomposes a contingency table similar to how principal components analysis decomposes multivariate continuous data. Simple correspondence analysis performs an eigen analysis of data, breaks down variability into underlying dimensions, and associates variability with rows and/or columns.
- Multiple Correspondence Analysis extends simple correspondence analysis to the case of 3 or more categorical variables. Multiple correspondence analysis performs a simple correspondence analysis on an indicator variables matrix in which each column corresponds to a level of a categorical variable. Rather than a 2-way table, the multi-way table is collapsed into 1 dimension.

Multivariate

Stat > Multivariate

Allows you to perform a principal components analysis, factor analysis, cluster analysis, discriminant analysis, and correspondence analysis.

Select one of the following options:

Principal Components – performs principal components analysis

Factor Analysis – performs factor analysis

Cluster Observations – performs agglomerative hierarchical clustering of observations

Cluster Variables – performs agglomerative hierarchical clustering of variables

© 2003 Minitab Inc.

Cluster K-Means – performs K-means non-hierarchical clustering of observations

Discriminant Analysis – performs linear and quadratic discriminant analysis

Simple Correspondence Analysis – performs simple correspondence analysis on a two-way contingency table

Multiple Correspondence Analysis – performs multiple correspondence analysis on three or more categorical variables

Minitab offers the following additional multivariate analysis options:

Balanced MANOVA

General MANOVA

Multivariate control charts

Examples of Multivariate Analysis

The following examples illustrate how to use the various multivariate analysis techniques available. Choose an example below:

Principal Components Analysis

Factor Analysis

Cluster Observations

Cluster Variables

Cluster K-Means

Discriminant Analysis

Simple Correspondence Analysis

Multiple Correspondence Analysis

References – Multivariate Analysis

[1] T.W. Anderson (1984). *An Introduction to Multivariate Statistical Analysis*, Second Edition. John Wiley & Sons.

[2] W. Dillon and M. Goldstein (1984). *Multivariate Analysis: Methods and Applications*. John Wiley & Sons.

[3] S.E. Fienberg (1987). *The Analysis of Cross-Classified Categorical Data*. The MIT Press.

[4] M. J. Greenacre (1993). *Correspondence Analysis in Practice.* Academic Press, Harcourt, Brace & Company.

[5] H. Harmon (1976). *Modern Factor Analysis*, Third Edition. University of Chicago Press.

[6] R. Johnson and D. Wichern (1992). *Applied Multivariate Statistical Methods*, Third Edition. Prentice Hall.

[7] K. Joreskog (1977). "Factor Analysis by Least Squares and Maximum Likelihood Methods," *Statistical Methods for Digital Computers,* ed. K. Enslein, A. Ralston and H. Wilf, John Wiley & Sons.

[8] J. K. Kihlberg, E. A. Narragon, and B. J. Campbell. (1964). *Automobile crash injury in relation to car size*. Cornell Aero. Lab. Report No. VJ-1823-R11.

[9] G.N. Lance and W.T. Williams (1967). "A General Theory of Classificatory Sorting Strategies, I. Hierarchical systems," *Computer Journal*, 9, 373–380

[10] G. W. Milligan (1980). "An Examination of the Effect of Six Types of Error Pertubation on Fifteen Clustering Algorithms," *Psychometrika*, 45, 325-342.

[11] S.J. Press and S. Wilson (1978). "Choosing Between Logistic Regression and Discriminant Analysis," *Journal of the American Statistical Association*, 73, 699-705.

[12] A. C. Rencher (1995). *Methods of Multivariate Analysis*, John Wiley & Sons.

Principal Components

Principal Components Analysis

Stat > Multivariate > Principal Components

Use principal component analysis to help you to understand the underlying data structure and/or form a smaller number of uncorrelated variables (for example, to avoid multicollinearity in regression).

An overview of principal component analysis can be found in most books on multivariate analysis, such as [5].

 © 2003 Minitab Inc.

Dialog box items

Variables: Choose the columns containing the variables to be included in the analysis.

Number of components to compute: Enter the number of principal components to be extracted. If you do not specify the number of components and there are p variables selected, then p principal components will be extracted. If p is large, you may want just the first few.

Type of Matrix

Correlation: Choose to calculate the principal components using the correlation matrix. Use the correlation matrix if it makes sense to standardize variables (the usual choice when variables are measured by different scales).

Covariance: Choose to calculate the principal components using the covariance matrix. Use the covariance matrix if you do not wish to standardize variables.

<Graphs>

<Storage>

Data – principal components analysis

Set up your worksheet so that each row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement (response). If a missing value exists in any column, Minitab ignores the whole row. Missing values are excluded from the calculation of the correlation or covariance matrix.

To perform Principal Component Analysis

1 Choose **Stat > Multivariate > Principal Components**.

2 In **Variables**, enter the columns containing the measurement data.

3 If you like, use any dialog box options, then click **OK**.

Nonuniqueness of Coefficients

The coefficients are unique (except for a change in sign) if the eigenvalues are distinct and not zero. If an eigenvalue is repeated, then the "space spanned" by all the principal component vectors corresponding to the same eigenvalue is unique, but the individual vectors are not. Therefore, the coefficients that Minitab prints and those in a book or another program may not agree, though the eigenvalues (variances) will always be the same.

If the covariance matrix has rank r < p, where p is the number of variables, then there will be p - r eigenvalues equal to zero. Eigenvectors corresponding to these eigenvalues may not be unique. This can happen if the number of observations is less than p or if there is multicollinearity.

Principal Components Analysis – Graphs

Stat > Multivariate > Principal Components> Graphs

Displays plots for judging the importance of the different principal components and for examining the scores of the first two principal components.

Dialog box items

Scree plot: Check to display a Scree plot (eigenvalue profile plot). Minitab plots the eigenvalue associated with a principal component versus the number of the component. Use this plot to judge the relative magnitude of eigenvalues.

Score plot for first 2 components: Check to plot the scores for the second principal component (y-axis) versus the scores for the first principal component (x-axis). To create plots for other components, store the scores and use Graph > Scatterplot.

Loading plot for first 2 components: Check to plot the loadings for the second component (y-axis) versus the loadings for the first component (x-axis). A line is drawn from each loading to the (0, 0) point.

Principal Components Analysis – Storage

Stat > Multivariate > Principal Components > Storage

Stores the coefficients and scores.

Dialog box items

Coefficients: Enter the storage columns for the coefficients of the principal components. The number of columns specified must be less than or equal to the number of principal components calculated.

© 2003 Minitab Inc.

Scores: Enter the storage columns for the principal components scores. Scores are linear combinations of your data using the coefficients. The number of columns specified must be less than or equal to the number of principal components calculated.

Example of Principal Components Analysis

You record the following characteristics for 14 census tracts: total population (Pop), median years of schooling (School), total employment (Employ), employment in health services (Health), and median home value (Home). The data were obtained from [6], Table 8.2.

You perform principal components analysis to understand the underlying data structure. You use the correlation matrix to standardize the measurements because they are not measured with the same scale.

1 Open the worksheet EXH_MVAR.MTW.

2 Choose **Stat > Multivariate > Principal Components**.

3 In **Variables**, enter *Pop-Home*.

4 Under **Type of Matrix**, choose **Correlation**.

5 Click **Graphs** and check **Scree plot**.

6 Click **OK** in each dialog box.

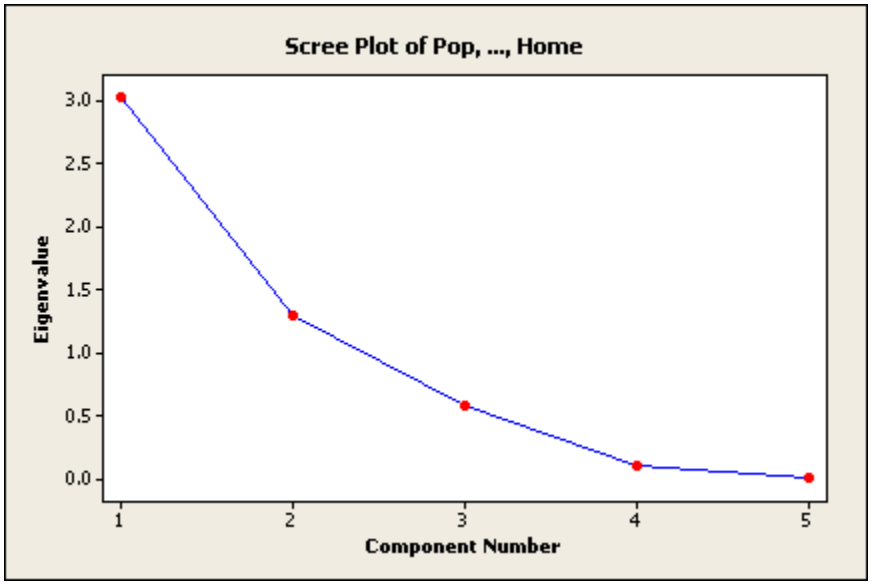
Session window output

Principal Component Analysis: Pop, School, Employ, Health, Home

```
Eigenanalysis of the Correlation Matrix

Eigenvalue  3.0289  1.2911  0.5725  0.0954  0.0121
Proportion   0.606   0.258   0.114   0.019   0.002
Cumulative   0.606   0.864   0.978   0.998   1.000


Variable      PC1     PC2     PC3     PC4     PC5
Pop        -0.558  -0.131   0.008   0.551  -0.606
School     -0.313  -0.629  -0.549  -0.453   0.007
Employ     -0.568  -0.004   0.117   0.268   0.769
Health     -0.487   0.310   0.455  -0.648  -0.201
Home        0.174  -0.701   0.691   0.015   0.014
```


© 2003 Minitab Inc.

Graph window output



Interpreting the results

The first principal component has variance (eigenvalue) 3.0289 and accounts for 60.6% of the total variance. The coefficients listed under PC1 show how to calculate the principal component scores:

PC1 = –.558 Pop – .313 School – .568 Employ – .487 Health + .174 Home

It should be noted that the interpretation of the principal components is subjective, however, obvious patterns emerge quite often. For instance, one could think of the first principal component as representing an overall population size, level of schooling, employment level, and employment in health services effect, because the coefficients of these terms have the same sign and are not close to zero.

The second principal component has variance 1.2911 and accounts for 25.8% of the data variability. It is calculated from the original data using the coefficients listed under PC2. This component could be thought of as contrasting level of schooling and home value with health employment to some extent.

Together, the first two and the first three principal components represent 86.4% and 97.8%, respectfully, of the total variability. Thus, most of the data structure can be captured in two or three underlying dimensions. The remaining principal components account for a very small proportion of the variability and are probably unimportant. The Scree plot provides this information visually.

Factor Analysis

Factor Analysis

Stat > Multivariate > Factor Analysis

Use factor analysis, like principal components analysis, to summarize the data covariance structure in a few dimensions of the data. However, the emphasis in factor analysis is the identification of underlying "factors" that might explain the dimensions associated with large data variability.

Dialog box items

Variables: Choose the columns containing the variables you want to use in the analysis. If you want to use a stored correlation or covariance matrix, or the loadings from a previous analysis instead of the raw data, click <Options>.

Number of factors to extract: Enter number of factors to extract (required if you use maximum likelihood as your method of extraction). If you don't specify a number with a principal components extraction, Minitab sets it equal to the number of variables in the data set. If you choose too many factors, Minitab will issue a warning in the Session window.

Method of Extraction:

Principal components: Choose to use the principal components method of factor extraction.

Maximum likelihood: Choose to use maximum likelihood for the initial solution.

Type of Rotation: Controls orthogonal rotations.

© 2003 Minitab Inc.

None: Choose not to rotate the initial solution.

Equimax: Choose to perform an equimax rotation of the initial solution (gamma = number of factors / 2).

Varimax: Choose to perform a varimax rotation of the initial solution (gamma = 1).

Quartimax: Choose to perform a quartimax rotation of the initial solution (gamma = 0).

Orthomax with gamma: Choose to perform an orthomax rotation of the initial solution, then enter value for gamma between 0 and 1.

<Options>

<Graphs>

<Storage>

<Results>

Data – Factor Analysis

You can have three types of input data:

- Columns of raw data
- A matrix of correlations or covariances
- Columns containing factor loadings

The typical case is to use raw data. Set up your worksheet so that a row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement (response). Minitab automatically omits rows with missing data from the analysis.

Usually the factor analysis procedure calculates the correlation or covariance matrix from which the loadings are calculated. However, you can enter a matrix as input data. You can also enter both raw data and a matrix of correlations or covariances. If you do, Minitab uses the matrix to calculate the loadings. Minitab then uses these loadings and the raw data to calculate storage values and generate graphs. See To perform factor analysis with a correlation or covariance matrix.

If you store initial factor loadings, you can later input these initial loadings to examine the effect of different rotations. You can also use stored loadings to predict factor scores of new data. See To perform factor analysis with stored loadings.

To perform factor analysis with a correlation or covariance matrix

You can choose to calculate the factor loadings and coefficients from a stored correlation or covariance matrix rather than the raw data. In this case, the raw data will be ignored. (Please note that this means scores can not be calculated.)

If it makes sense to standardize variables (usual choice when variables are measured by different scales), enter a correlation matrix; if you do not wish to standardize, enter a covariance matrix.

1 Choose **Stat > Multivariate > Factor Analysis**.

2 Click **Options**.

3 Under **Matrix to Factor**, choose **Correlation** or **Covariance**.

4 Under **Source of Matrix**, choose **Use matrix** and enter the matrix. Click **OK**.

To perform factor analysis with raw data

There are three ways that you might carry out a factor analysis in Minitab. The usual way, described below, is to enter columns containing your measurement variables, but you can also use a matrix as input (See To perform factor analysis with a correlation or covariance matrix) or use stored loadings as input (See To perform factor analysis with stored loadings).

1 Choose **Stat > Multivariate > Factor Analysis**.

2 In **Variables**, enter the columns containing the measurement data.

3 If you like, use any dialog box options, then click **OK**.

To perform factor analysis with stored loadings

If you store initial factor loadings from an earlier analysis, you can input these initial loadings to examine the effect of different rotations. You can also use stored loadings to predict factor scores of new data.

1 Cick **Options** in the Factor Analysis dialog box.

2 Under **Loadings for Initial Solution**, choose **Use loadings**. Enter the columns containing the loadings. Click **OK**.

© 2003 Minitab Inc.

3 Do one of the following, and then click **OK**:

- To examine the effect of a different rotation method, choose an option under **Type of Rotation**. See Rotating the factor loadings for a discussion of the various rotations>Main.
- To predict factor scores with new data, in **Variables**, enter the columns containing the new data.

Factor analysis in practice

The goal of factor analysis is to find a small number of factors, or unobservable variables, that explains most of the data variability and yet makes contextual sense. You need to decide how many factors to use, and find loadings that make the most sense for your data.

Number of factors

The choice of the number of factors is often based upon the proportion of variance explained by the factors, subject matter knowledge, and reasonableness of the solution [6]. Initially, try using the principal components extraction method without specifying the number of components. Examine the proportion of variability explained by different factors and narrow down your choice of how many factors to use. A Scree plot may be useful here in visually assessing the importance of factors. Once you have narrowed this choice, examine the fits of the different factor analyses. Communality values, the proportion of variability of each variable explained by the factors, may be especially useful in comparing fits. You may decide to add a factor if it contributes to the fit of certain variables. Try the maximum likelihood method of extraction as well.

Rotation

Once you have selected the number of factors, you will probably want to try different rotations. Johnson and Wichern [6] suggest the varimax rotation. A similar result from different methods can lend credence to the solution you have selected. At this point you may wish to interpret the factors using your knowledge of the data. For more information see Rotating the factor loadings.

Rotating the factor loadings

There are four methods to orthogonally rotate the initial factor loadings found by either principal components or maximum likelihood extraction. An orthogonal rotation simply rotates the axes to give you a different perspective. The methods are equimax, varimax, quartimax, and orthomax. Minitab rotates the loadings in order to minimize a simplicity criterion [5]. A parameter, gamma, within this criterion is determined by the rotation method. If you use a method with a low value of gamma, the rotation will tend to simplify the rows of the loadings; if you use a method with a high value of gamma, the rotation will tend to simplify the columns of the loadings. The table below summarizes the rotation methods.

Rotation method	Goal is ...	Gamma
equimax	to rotate the loadings so that a variable loads high on one factor but low on others	number of factors / 2
varimax	to maximize the variance of the squared loadings	1
quartimax	simple loadings	0
orthomax	user determined, based on the given value of gamma	0-1

Factor Analysis – Options

Stat > Multivariate > Factor Analysis > Options

Allows you to specify the matrix type and source, and the loadings to use for the initial extraction.

Dialog box items

Matrix to Factor

Correlation: Choose to calculate the factors using the correlation matrix. Use the correlation matrix if it makes sense to standardize variables (the usual choice when variables are measured by different scales).

Covariance: Choose to calculate the factors using the covariance matrix. Use the covariance matrix if you do not wish to standardize variables. The covariance matrix cannot be used with a maximum likelihood estimation.

Source of Matrix:

Compute from variables: Choose to use the correlation or covariance matrix of the measurement data.

Use matrix: Choose to use a stored matrix for calculating the loadings and coefficients. (Note: Scores can not be calculated if this option is chosen.) See To perform factor analysis with a correlation or covariance matrix.

© 2003 Minitab Inc.

Loadings for Initial Solution

Compute from variables: Choose to compute loadings from the raw data.

Use loadings: Choose to use loadings which were previously calculated, then specify the columns containing the loadings. You must specify one column for each factor calculated. See To perform factor analysis with stored loadings.

Maximum Likelihood Extraction

Use initial communality estimates in: Choose the column containing data to be used as the initial values for the communalities. The column should contain one value for each variable.

Max iterations: Enter the maximum number of iterations allowed for a solution (default is 25).

Convergence: Enter the criterion for convergence (occurs when the uniqueness values do not change very much). This number is the size of the smallest change (default is 0.005).

Factor Analysis – Graphs

Stat > Multivariate > Factor Analysis > Graphs

Displays a Scree plot, and score and loading plots for the first two factors.

To create simple loading plots for other factors, store the loadings and use Graph > Scatterplot. If you want to connect the loading point to the zero point, add a zero to the bottom of each column of loadings in the Data window, then add lines connecting the loading points to the zero point with the graph editor. See graph editing overview.

Dialog box items

Scree plot: Check to display a Scree plot (eigenvalue profile plot). Minitab plots the eigenvalue associated with a factor versus the number of the factor.

Score plot for first 2 factors: Check to plot the scores for the second factor (y-axis) versus the scores for the first factor (x-axis). Scores are linear combinations of your data using the coefficients. To create plots for other factors, store the scores and use Graph > Scatterplot. (Note: Scores must be calculated from raw data, therefore this graph can not be generated if the **Use matrix** option is selected. See <Options>.)

Loading plot for first 2 factors: Check to plot the loadings for the second factor (y-axis) versus the loadings for the first factor (x-axis). A line is drawn from each loading to the (0, 0) point.

Factor Analysis – Storage

Stat > Multivariate > Factor Analysis > Storage

Allows you to store factor loadings, factor score coefficients, factor or standard scores, rotation matrix, residual matrix, eigenvalues, and eigenvectors. You can then use this information for further analysis.

Dialog box items

Storage

Loadings: Enter storage columns for the factor loadings. You must enter one column for each factor. If a rotation was specified, Minitab stores the values for the rotated factor loadings These can be input using <Options> and specifying the columns under Loadings for initial solutions.

Coefficients: Enter storage columns for the factor score coefficients. You must enter one column for each factor.

Scores: Enter storage columns for the scores. You must enter one column for each factor. Minitab calculates factor scores by multiplying factor score coefficients and your data after they have been centered by subtracting means. (Note: Scores must be calculated from raw data, therefore the **Use matrix** option must not be selected. See <Options>.)

Rotation matrix: Enter a location to store the matrix used to rotate the initial loadings. You may enter a matrix name or number (for example, M3). The rotation matrix is the matrix used to rotate the initial loadings. If L is the matrix of initial loadings and M is the rotation matrix, LM is the matrix of rotated loadings.

Residual matrix: Enter a location to store the residual matrix. The residual matrix for the initial and rotated solutions are the same. You may enter a matrix name or number (for example, M3). The residual matrix is (A-LL'), where A is the correlation or covariance matrix and L is a matrix of loadings. The residual matrix is the same for initial or rotated solutions.

Eigenvalues: Enter a column to store the eigenvalues of the matrix that was factored. The eigenvalues are stored in numerical order from largest to smallest. To store eigenvalues, you must do the initial extraction using principal components. You can plot the eigenvalues to obtain a Scree plot.

Eigenvector matrix: Enter a matrix to store the eigenvectors of the matrix that was factored. Each vector is stored as a column of the matrix, in the same order as the eigenvalues.

© 2003 Minitab Inc.

Factor analysis storage

To store loadings, factor score coefficients, or factor scores, enter a column name or column number for each factor that has been extracted. The number of storage columns specified must be equal in number to the number of factors calculated. If a rotation was specified, Minitab stores the values for the rotated solution. Minitab calculates factor scores by multiplying factor score coefficients and your data after they have been centered by subtracting means.

You can also store the rotation matrix and residual matrix. Enter a matrix name or matrix number. The rotation matrix is the matrix used to rotate the initial loadings. If L is the matrix of initial loadings and M is the rotation matrix that you store, LM is the matrix of rotated loadings. The residual matrix is (A-LL), where A is the correlation or covariance matrix and L is a matrix of loadings. The residual matrix is the same for initial and rotated solutions.

You can also store the eigenvalues and eigenvectors of the correlation or covariance matrix (depending on which is factored) if you chose the initial factor extraction via principal components. Enter a single column name or number for storing eigenvalues, which are stored from largest to smallest. Enter a matrix name or number to store the eigenvectors in an order corresponding to the sorted eigenvalues.

Factor Analysis – Results

Stat > Multivariate > Factor Analysis > Results

Controls the display of Session window results.

Dialog box items

Display of Results:

Do not display: Choose to suppress the display of results. All requested storage is done.

Loadings only: Choose to display loadings (and sorted loadings if requested) for the final solution.

Loadings and factor score coefficients: Choose to display factor loadings and scores.

All and MLE iterations: Choose to display the factor loadings, factor scores and information on the iterations if a maximum likelihood estimation was used.

Sort loading: Check to sort the loadings in the Session window (within a factor if the maximum absolute loading occurs there).

Zero loading less than: Check to enter a value. Loadings less than this value will be displayed as zero.

Example of Factor Analysis, Using Maximum Likelihood and a Rotation

Two factors were chosen as the number to represent the census tract data of the Example of Factor Analysis Using Principal Components. You perform a maximum likelihood extraction and varimax rotation to interpret the factors.

1 Open the worksheet EXH_MVAR.MTW.

2 Choose **Stat > Multivariate > Factor Analysis**.

3 In **Variables**, enter *Pop-Home*.

4 In **Number of factors to extract**, enter *2*.

5 Under **Method of Extraction**, choose **Maximum likelihood**.

6 Under **Type of Rotation**, choose **Varimax**.

7 Click **Graphs** and check **Loading plot for first 2 factors**.

8 Click **Results** and check **Sort loadings**. Click **OK** in each dialog box.

Session window output

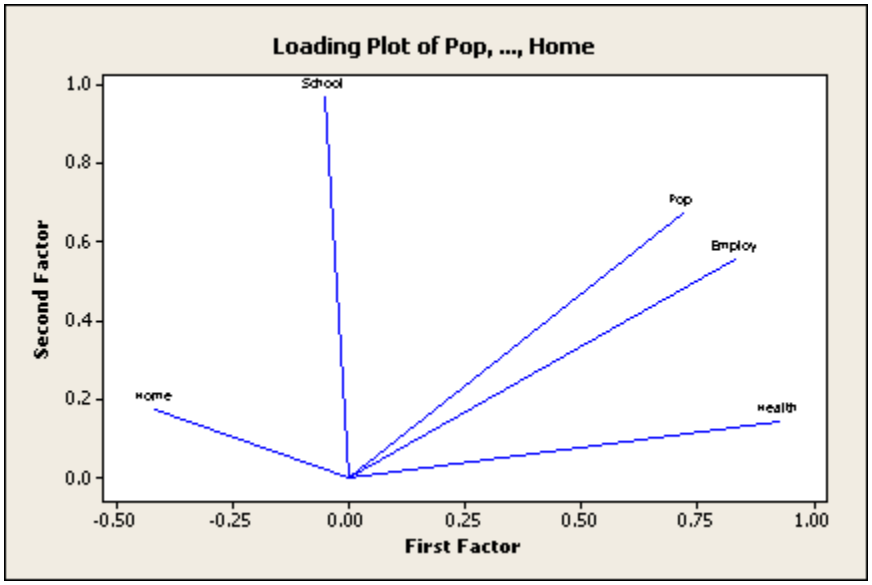
Factor Analysis: Pop, School, Employ, Health, Home

```
Maximum Likelihood Factor Analysis of the Correlation Matrix

* NOTE * Heywood case

Unrotated Factor Loadings and Communalities

Variable  Factor1  Factor2  Communality
Pop         0.971    0.160        0.968
School      0.494    0.833        0.938
Employ      1.000    0.000        1.000
Health      0.848   -0.395        0.875
Home       -0.249    0.375        0.202
```

© 2003 Minitab Inc.

```
Variance   2.9678   1.0159       3.9837
% Var       0.594    0.203        0.797


Rotated Factor Loadings and Communalities
Varimax Rotation

Variable  Factor1  Factor2  Communality
Pop         0.718    0.673        0.968
School     -0.052    0.967        0.938
Employ      0.831    0.556        1.000
Health      0.924    0.143        0.875
Home       -0.415    0.173        0.202

Variance   2.2354   1.7483       3.9837
% Var       0.447    0.350        0.797


Sorted Rotated Factor Loadings and Communalities

Variable  Factor1  Factor2  Communality
Health      0.924    0.143        0.875
Employ      0.831    0.556        1.000
Pop         0.718    0.673        0.968
Home       -0.415    0.173        0.202
School     -0.052    0.967        0.938

Variance   2.2354   1.7483       3.9837
% Var       0.447    0.350        0.797


Factor Score Coefficients


Variable  Factor1  Factor2
Pop        -0.165    0.246
School     -0.528    0.789
Employ      1.150    0.080
Health      0.116   -0.173
Home       -0.018    0.027
```

Graph window output



© 2003 Minitab Inc.

Interpreting the results

The results indicates that this is a Heywood case. There are three tables of loadings and communalities: unrotated, rotated, and sorted and rotated. The unrotated factors explain 79.7% of the data variability (see last line under Communality) and the communality values indicate that all variables but Home are well represented by these two factors (communalities are 0.202 for Home, 0.875-1.0 for other variables). The percent of total variability represented by the factors does not change with rotation, but after rotating, these factors are more evenly balanced in the percent of variability that they represent, being 44.7% and 35.0%, respectfully.

Sorting is done by the maximum absolute loading for any factor. Variables that have their highest absolute loading on factor 1 are printed first, in sorted order. Variables with their highest absolute loadings on factor 2 are printed next, in sorted order, and so on. Factor 1 has large positive loadings on Health (0.924), Employ (0.831), and Pop (0.718), and a -0.415 loading on Home while the loading on School is small. Factor 2 has a large positive loading on School of 0.967 and loadings of 0.556 and 0.673, respectively, on Employ and Pop, and small loadings on Health and Home.

You can view the rotated loadings graphically in the loadings plot. What stands out for factor 1 are the high loadings on the variables Pop, Employ, and Health and the negative loading on Home. School has a high positive loading for factor 2 and somewhat lower values for Pop and Employ.

Let's give a possible interpretation to the factors. The first factor positively loads on population size and on two variables, Employ and Health, that generally increase with population size. It negatively loads on home value, but this may be largely influenced by one point. We might consider factor 1 to be a "health care - population size" factor. The second factor might be considered to be a "education - population size" factor. Both Health and School are correlated with Pop and Employ, but not much with each other.

In addition, Minitab displays a table of factor score coefficients. These show you how the factors are calculated. Minitab calculates factor scores by multiplying factor score coefficients and your data after they have been centered by subtracting means.

You might repeat this factor analysis with three factors to see if it makes more sense for your data.

Example of Factor Analysis, Using Principal Components

You record the following characteristics of 14 census tracts: total population (Pop), median years of schooling (School), total employment (Employ), employment in health services (Health), and median home value (Home) (data from [6], Table 8.2). You would like to investigate what "factors" might explain most of the variability. As the first step in your factor analysis, you use the principal components extraction method and examine an eigenvalues (scree) plot in order to help you to decide upon the number of factors.

1 Open the worksheet EXH_MVAR.MTW.

2 Choose **Stat > Multivariate > Factor Analysis**.

3 In **Variables**, enter *Pop-Home*.

4 Click **Graphs** and check **Scree plot**. Click **OK** in each dialog box.

Session window output

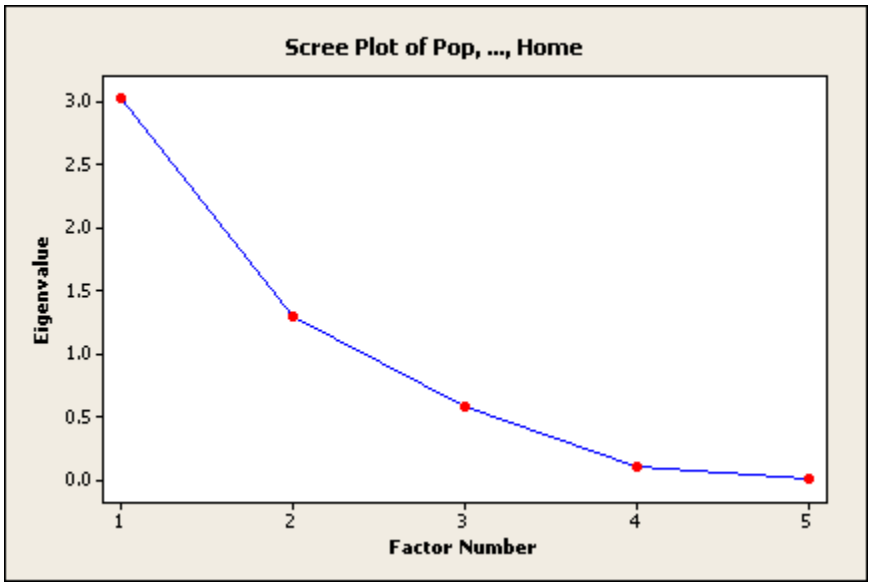
Factor Analysis: Pop, School, Employ, Health, Home

```
Principal Component Factor Analysis of the Correlation Matrix


Unrotated Factor Loadings and Communalities

Variable  Factor1  Factor2  Factor3  Factor4  Factor5  Communality
Pop        -0.972   -0.149    0.006    0.170   -0.067        1.000
School     -0.545   -0.715   -0.415   -0.140    0.001        1.000
Employ     -0.989   -0.005    0.089    0.083    0.085        1.000
Health     -0.847    0.352    0.344   -0.200   -0.022        1.000
Home        0.303   -0.797    0.523    0.005    0.002        1.000

Variance   3.0289   1.2911   0.5725   0.0954   0.0121       5.0000
% Var       0.606    0.258    0.114    0.019    0.002        1.000


Sorted Unrotated Factor Loadings and Communalities

Variable  Factor1  Factor2  Factor3  Factor4  Factor5  Communality
Employ     -0.989   -0.005    0.089    0.083    0.085        1.000
Pop        -0.972   -0.149    0.006    0.170   -0.067        1.000
Health     -0.847    0.352    0.344   -0.200   -0.022        1.000
Home        0.303   -0.797    0.523    0.005    0.002        1.000
```

© 2003 Minitab Inc.

```
School      -0.545   -0.715   -0.415   -0.140    0.001        1.000

Variance    3.0289   1.2911   0.5725   0.0954   0.0121       5.0000
% Var        0.606    0.258    0.114    0.019    0.002        1.000


Factor Score Coefficients

Variable  Factor1  Factor2  Factor3  Factor4  Factor5
Pop        -0.321   -0.116    0.011    1.782   -5.511
School     -0.180   -0.553   -0.726   -1.466    0.060
Employ     -0.327   -0.004    0.155    0.868    6.988
Health     -0.280    0.272    0.601   -2.098   -1.829
Home        0.100   -0.617    0.914    0.049    0.129
```

Graph window output



Interpreting the results

Five factors describe these data perfectly, but the goal is to reduce the number of factors needed to explain the variability in the data. Examine the Session window results line of % Var or the eigenvalues plot. The proportion of variability explained by the last two factors is minimal (0.019 and 0.002, respectively) and they can be eliminated as being important. The first two factors together represent 86% of the variability while three factors explain 98% of the variability. The question is whether to use two or three factors. The next step might be to perform separate factor analyses with two and three factors and examine the communalities to see how individual variables are represented. If there were one or more variables not well represented by the more parsimonious two factor model, you might select a model with three or more factors.

See the example below for a rotation of loadings extracted by the maximum likelihood method with a selection of two factors.

Cluster Observations

Cluster Observations

Stat > Multivariate > Cluster Observations

Use clustering of observations to classify observations into groups when the groups are initially not known.

This procedure uses an agglomerative hierarchical method that begins with all observations being separate, each forming its own cluster. In the first step, the two observations closest together are joined. In the next step, either a third observation joins the first two, or two other observations join together into a different cluster. This process will continue until all clusters are joined into one, however this single cluster is not useful for classification purposes. Therefore you must decide how many groups are logical for your data and classify accordingly. See Determining the final cluster grouping for more information.

© 2003 Minitab Inc.

Dialog box items

Variables or distance matrix: Enter either the columns containing measurement data or a stored distance matrix on which to perform the hierarchical clustering of observations.

Linkage Method: Choose the linkage method that will determine how the distance between two clusters is defined.

Distance Measure: Choose the distance measure to use if you selected columns as input variables.

Standardize variables: Check to standardize all variables by subtracting the means and dividing by the standard deviation before the distance matrix is calculated–a good idea if variables are in different units and you wish to minimize the effect of scale differences. If you standardize, cluster centroids and distance measures are in standardized variable space.

Specify Final Partition by

Number of Clusters: Choose to determine the final partition by a specified number of clusters. Enter this number in the box. See Determining the final cluster grouping.

Similarity Level: Choose to determine the final partition by the specified level of similarity. Enter this value in the box. See Determining the final cluster grouping.

Show Dendrogram: Check to display the dendrogram or tree diagram, showing the amalgamation steps. Use <Customize> to change the default display of the dendrogram.

<Customize>

<Storage>

Data – Cluster Observations

You can have two types of input data: columns of raw data or a matrix of distances.

Typically, you would use raw data. Each row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement. You must delete rows with missing data from the worksheet before using this procedure.

If you store an n x n distance matrix, where n is the number of observations, you can use this matrix as input data. The (i, j) entry in this matrix is the distance between observations i and j. If you use the distance matrix as input, statistics on the final partition are not available.

To perform clustering of observations

1 Choose **Stat > Multivariate > Cluster Observations**.

2 In **Variables or distance matrix**, enter either columns containing the raw (measurement) data or a matrix of distances.

3 If you like, use any dialog box options, then click **OK**.

Cluster Observations – Dendrogram – Customize

Stat > Multivariate > Cluster Observations > Show Dendrogram > Customize

Allows you to add a title and control y-axis labeling and displaying for the dendrogram.

Double-click the dendrogram after you create it to specify the line type, color, and size for the cluster groups. See Graph Editing Overview.

Dialog box items

Title: To display a title above the dendrogram, type the desired text in this box.

Case labels: Enter a column of case labels. This column must be the same length as the data column.

Label Y Axis with

Similarity: Choose to display similarities on the y-axis.

Distance: Choose to display distances on the y-axis.

Show Dendrogram in

One graph: Choose to display the dendrogram in a single graph window.

Maximum number of observations per graph (without splitting a group): Choose to display a specified number of observation per graph and enter an integer greater than or equal to 1.

© 2003 Minitab Inc.

Deciding Which Distance Measures and Linkage Methods to Use – Cluster Observations

Distance Measures

If you do not supply a distance matrix, Minitab's first step is to calculate an n x n distance matrix, D, where n is the number of observations. The matrix entries, d(i, j), in row i and column j, is the distance between observations i and j.

Minitab provides five different methods to measure distance. You might choose the distance measure according to properties of your data.

- The Euclidean method is a standard mathematical measure of distance (square root of the sum of squared differences).
- The Pearson method is a square root of the sum of square distances divided by variances. This method is for standardizing.
- Manhattan distance is the sum of absolute distances, so that outliers receive less weight than they would if the Euclidean method were used.
- The squared Euclidean and squared Pearson methods use the square of the Euclidean and Pearson methods, respectfully. Therefore, the distances that are large under the Euclidean and Pearson methods will be even larger under the squared Euclidean and squared Pearson methods.

Tip If you choose Average, Centroid, Median, or Ward as the linkage method, it is generally recommended [9] that you use one of the squared distance measures.

Linkage methods

The linkage method that you choose determines how the distance between two clusters is defined. At each amalgamation stage, the two closest clusters are joined. At the beginning, when each observation constitutes a cluster, the distance between clusters is simply the inter-observation distance. Subsequently, after observations are joined together, a linkage rule is necessary for calculating inter-cluster distances when there are multiple observations in a cluster.

You may wish to try several linkage methods and compare results. Depending on the characteristics of your data, some methods may provide "better" results than others.

- With single linkage, or "nearest neighbor," the distance between two clusters is the minimum distance between an observation in one cluster and an observation in the other cluster. Single linkage is a good choice when clusters are clearly separated. When observations lie close together, single linkage tends to identify long chain-like clusters that can have a relatively large distance separating observations at either end of the chain [6].
- With average linkage, the distance between two clusters is the mean distance between an observation in one cluster and an observation in the other cluster. Whereas the single or complete linkage methods group clusters based upon single pair distances, average linkage uses a more central measure of location.
- With centroid linkage, the distance between two clusters is the distance between the cluster centroids or means. Like average linkage, this method is another averaging technique.
- With complete linkage, or "furthest neighbor," the distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster. This method ensures that all observations in a cluster are within a maximum distance and tends to produce clusters with similar diameters. The results can be sensitive to outliers [10].
- With median linkage, the distance between two clusters is the median distance between an observation in one cluster and an observation in the other cluster. This is another averaging technique, but uses the median rather than the mean, thus downweighting the influence of outliers.
- With McQuitty's linkage, when two clusters are be joined, the distance of the new cluster to any other cluster is calculated as the average of the distances of the soon to be joined clusters to that other cluster. For example, if clusters 1 and 3 are to be joined into a new cluster, say 1*, then the distance from 1* to cluster 4 is the average of the distances from 1 to 4 and 3 to 4. Here, distance depends on a combination of clusters rather than individual observations in the clusters.
- With Ward's linkage, the distance between two clusters is the sum of squared deviations from points to centroids. The objective of Ward's linkage is to minimize the within-cluster sum of squares. It tends to produce clusters with similar numbers of observations, but it is sensitive to outliers [10]. In Ward's linkage, it is possible for the distance between two clusters to be larger than dmax, the maximum value in the original distance matrix. If this happens, the similarity will be negative.

Determining the Final Grouping of Clusters

The final grouping of clusters (also called the final partition) is the grouping of clusters which will, hopefully, identify groups whose observations or variables share common characteristics. The decision about final grouping is also called cutting the dendrogram. The complete dendrogram (tree diagram) is a graphical depiction of the amalgamation of observations or variables into one cluster. Cutting the dendrogram is akin to drawing a line across the dendrogram to specify the final grouping.

 © 2003 Minitab Inc.

How do you know where to cut the dendrogram? You might first execute cluster analysis without specifying a final partition. Examine the similarity and distance levels in the Session window results and in the dendrogram. You can view the similarity levels by placing your mouse pointer over a horizontal line in the dendrogram. The similarity level at any step is the percent of the minimum distance at that step relative to the maximum inter-observation distance in the data. The pattern of how similarity or distance values change from step to step can help you to choose the final grouping. The step where the values change abruptly may identify a good point for cutting the dendrogram, if this makes sense for your data.

After choosing where you wish to make your partition, rerun the clustering procedure, using either **Number of clusters** or **Similarity level** to give you either a set number of groups or a similarity level for cutting the dendrogram. Examine the resulting clusters in the final partition to see if the grouping seems logical. Looking at dendrograms for different final groupings can also help you to decide which one makes the most sense for your data.

Note For some data sets, average, centroid, median and Ward's methods may not produce a hierarchical dendrogram. That is, the amalgamation distances do not always increase with each step. In the dendrogram, such a step will produce a join that goes downward rather than upward.

Cluster Observations – Storage

Stat > Multivariate > Cluster Observations > Storage

Allows you to store cluster membership for each observation, the distance between each observation and each cluster centroid, and the distance matrix.

Dialog box items

Cluster membership column: Enter a single column to store for cluster membership for each observation. This column can then be used as a categorical variable in other Minitab commands.

Distance between observations and cluster centroids (Give a column for each cluster group): Enter storage column(s) for the distance between each observation and each cluster centroid. The number of columns specified must equal the number of cluster in the final partition. The distances stored are always Euclidean distances.

Distance matrix: Enter a storage matrix (M) for the N x N distance matrix, where N is the number of observations. The stored distance matrix can then be used in subsequent commands.

Example of Cluster Observations

You make measurements on five nutritional characteristics (protein, carbohydrate, and fat content, calories, and percent of the daily allowance of Vitamin A) of 12 breakfast cereal brands. The example and data are from p. 623 of [6]. The goal is to group cereal brands with similar characteristics. You use clustering of observations with the complete linkage method, squared Euclidean distance, and you choose standardization because the variables have different units. You also request a dendrogram and assign different line types and colors to each cluster.

1 Open the worksheet CEREAL.MTW.

2 Choose **Stat > Multivariate > Cluster Observations**.

3 In **Variables or distance matrix**, enter *Protein-VitaminA*.

4 From **Linkage Method**, choose **Complete** and from **Distance Measure** choose **Squared Euclidean**.

5 Check **Standardize variables**.

6 Under **Specify Final Partition by**, choose **Number of clusters** and enter *4*.

7 Check **Show dendrogram**.

8 Click **Customize**. In **Title**, enter *Dendrogram for Cereal Data*.

9 Click **OK** in each dialog box.

Session window output

Cluster Analysis of Observations: Protein, Carbo, Fat, Calories, VitaminA

```
Standardized Variables, Squared Euclidean Distance, Complete Linkage
Amalgamation Steps


                                                                Number
        Number                                                of obs.
            of  Similarity  Distance  Clusters      New   in new
Step  clusters       level     level   joined   cluster  cluster
   1        11     100.000    0.0000  5     12        5        2
   2        10      99.822    0.0640  3      5        3        3
```

© 2003 Minitab Inc.

3	9	98.792	0.4347	3	11	3	4
4	8	94.684	1.9131	6	8	6	2
5	7	93.406	2.3730	2	3	2	5
6	6	87.329	4.5597	7	9	7	2
7	5	86.189	4.9701	1	4	1	2
8	4	80.601	6.9810	2	6	2	7
9	3	68.079	11.4873	2	7	2	9
10	2	41.409	21.0850	1	2	1	11
11	1	0.000	35.9870	1	10	1	12

Final Partition
Number of clusters: 4

	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	2	2.48505	1.11469	1.11469
Cluster2	7	8.99868	1.04259	1.76922
Cluster3	2	2.27987	1.06768	1.06768
Cluster4	1	0.00000	0.00000	0.00000

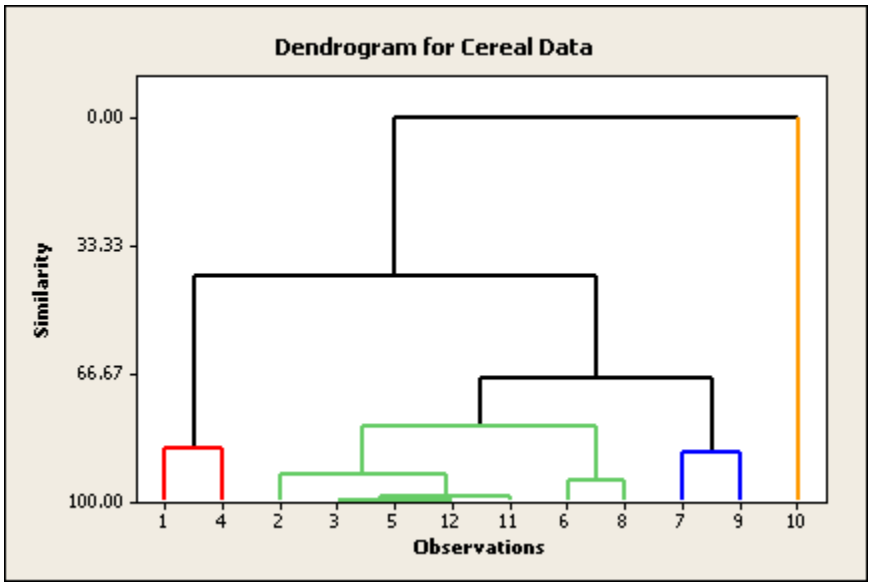
Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Cluster4	Grand centroid
Protein	1.92825	-0.333458	-0.20297	-1.11636	0.0000000
Carbo	-0.75867	0.541908	0.12645	-2.52890	0.0000000
Fat	0.33850	-0.096715	0.33850	-0.67700	0.0000000
Calories	0.28031	0.280306	0.28031	-3.08337	-0.0000000
VitaminA	-0.63971	-0.255883	2.04707	-1.02353	-0.0000000

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3	Cluster4
Cluster1	0.00000	2.67275	3.54180	4.98961
Cluster2	2.67275	0.00000	2.38382	4.72050
Cluster3	3.54180	2.38382	0.00000	5.44603
Cluster4	4.98961	4.72050	5.44603	0.00000

Graph window output



© 2003 Minitab Inc.

Interpreting the results

Minitab displays the amalgamation steps in the Session window. At each step, two clusters are joined. The table shows which clusters were joined, the distance between them, the corresponding similarity level, the identification number of the new cluster (this number is always the smaller of the two numbers of the clusters joined), the number of observations in the new cluster, and the number of clusters. Amalgamation continues until there is just one cluster.

The amalgamation steps show that the similarity level decreases by increments of about 6 or less until it decreases by about 13 at the step from four clusters to three. This indicates that four clusters are reasonably sufficient for the final partition. If this grouping makes intuitive sense for the data, then it is probably a good choice.

When you specify the final partition, Minitab displays three additional tables. The first table summarizes each cluster by the number of observations, the within cluster sum of squares, the average distance from observation to the cluster centroid, and the maximum distance of observation to the cluster centroid. In general, a cluster with a small sum of squares is more compact than one with a large sum of squares. The centroid is the vector of variable means for the observations in that cluster and is used as a cluster midpoint. The second table displays the centroids for the individual clusters while the third table gives distances between cluster centroids.

The dendrogram displays the information in the amalgamation table in the form of a tree diagram. In our example, cereals 1 and 4 make up the first cluster; cereals 2, 3, 5, 12, 11, 6, and 8 make up the second; cereals 7 and 9 make up the third; cereal 10 makes up the fourth.

Cluster Variables

Cluster Variables

Stat > Multivariate > Cluster Variables

Use Clustering of Variables to classify variables into groups when the groups are initially not known. One reason to cluster variables may be to reduce their number. This technique may give new variables that are more intuitively understood than those found using principal components.

This procedure is an agglomerative hierarchical method that begins with all variables separate, each forming its own cluster. In the first step, the two variables closest together are joined. In the next step, either a third variable joins the first two, or two other variables join together into a different cluster. This process will continue until all clusters are joined into one, but you must decide how many groups are logical for your data. See Determining the final grouping.

Dialog box items

Variables or distance matrix: Enter either the columns containing measurement data or a distance matrix on which to perform the hierarchical clustering of variables.

Linkage Method: Choose the linkage method that will determine how the distance between two clusters is defined.

Distance Measure: If you selected columns as input variables, choose the desired distance measure.

- **Correlation:** Choose to use the correlation distance measure.
- **Absolute correlation:** Choose to use the absolute correlation distance measure.

Specify Final Partition by

- **Number of clusters:** Choose to determine the final partition by a specified number of clusters. Enter this number in the box.
- **Similarity level:** Choose to determine the final partition by the specified level of similarity. Enter a value between 0 and 100 in the box.

Show dendrogram: Check to display the dendrogram (tree diagram), showing the amalgamation steps. Use <Customize> to change the default display of the dendrogram.

<Customize>

<Storage>

Data – Cluster Variables

You can have two types of input data to cluster variables: columns of raw data or a matrix of distances.

Typically, you use raw data. Each row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement. Delete rows with missing data from the worksheet before using this procedure.

If you store a p x p distance matrix, where p is the number of variables, you can use the matrix as input data. The (i, j) entry in the matrix is the distance between observations i and j. If you use the distance matrix as input, final partition statistics are not available.

© 2003 Minitab Inc.

To perform clustering of variables

1 Choose **Stat > Multivariate > Cluster Variables**.

2 In **Variables or distance matrix**, enter either columns containing the raw (measurement) data or a matrix of distances.

3 If you like, use any dialog box options, then click **OK**.

Clustering variables in practice

You must make similar decisions to cluster variables as you would to cluster observations. Follow the guidelines in Determining the final grouping to help you determine groupings. However, if the purpose behind clustering of variables is data reduction, you may decide to use your knowledge of the data to a greater degree in determining the final clusters of variables.

Deciding Which Distance Measures and Linkage Methods to Use – Cluster Variables

Distance Measures

You can use correlations or absolute correlations for distance measures. With the correlation method, the (i,j) entry of the distance matrix is $d_{ij} = 1 - \rho_{ij}$ and for the absolute correlation method, $d_{ij} = 1 - |\rho_{ij}|$, where ρ_{ij} is the (Pearson product moment) correlation between variables i and j. Thus, the correlation method will give distances between 0 and 1 for positive correlations, and between 1 and 2 for negative correlations. The absolute correlation method will always give distances between 0 and 1.

- If it makes sense to consider negatively correlated data to be farther apart than positively correlated data, then use the correlation method.
- If you think that the strength of the relationship is important in considering distance and not the sign, then use the absolute correlation method.

Linkage methods

The linkage method that you choose determines how the distance between two clusters is defined. At each amalgamation stage, the two closest clusters are joined. At the beginning, when each variables constitutes a cluster, the distance between clusters is simply the inter-variables distance. Subsequently, after observations are joined together, a linkage rule is necessary for calculating inter-cluster distances when there are multiple variables in a cluster.

You may wish to try several linkage methods and compare results. Depending on the characteristics of your data, some methods may provide "better" results than others.

- With single linkage, or "nearest neighbor," the distance between two clusters is the minimum distance between a variable in one cluster and a variable in the other cluster. Single linkage is a good choice when clusters are clearly separated. When variables lie close together, single linkage tends to identify long chain-like clusters that can have a relatively large distance separating variables at either end of the chain [6].
- With average linkage, the distance between two clusters is the mean distance between a variable in one cluster and a variable in the other cluster. Whereas the single or complete linkage methods group clusters based upon single pair distances, average linkage uses a more central measure of location.
- With centroid linkage, the distance between two clusters is the distance between the cluster centroids or means. Like average linkage, this method is another averaging technique.
- With complete linkage, or "furthest neighbor," the distance between two clusters is the maximum distance between a variable in one cluster and a variable in the other cluster. This method ensures that all variables in a cluster are within a maximum distance and tends to produce clusters with similar diameters. The results can be sensitive to outliers [10].
- With median linkage, the distance between two clusters is the median distance between a variable in one cluster and a variable in the other cluster. This is another averaging technique, but uses the median rather than the mean, thus downweighting the influence of outliers.
- With McQuitty's linkage, when two clusters are be joined, the distance of the new cluster to any other cluster is calculated as the average of the distances of the soon to be joined clusters to that other cluster. For example, if clusters 1 and 3 are to be joined into a new cluster, say 1*, then the distance from 1* to cluster 4 is the average of the distances from 1 to 4 and 3 to 4. Here, distance depends on a combination of clusters rather than individual variables in the clusters.
- With Ward's linkage, the distance between two clusters is the sum of squared deviations from points to centroids. The objective of Ward's linkage is to minimize the within-cluster sum of squares. It tends to produce clusters with similar numbers of variables, but it is sensitive to outliers [10]. In Ward's linkage, it is possible for the distance between two clusters to be larger than dmax, the maximum value in the original distance matrix. If this happens, the similarity will be negative.

 © 2003 Minitab Inc.

Cluster Variables – Dendrogram – Customize

Stat > Multivariate > Cluster Variables > Show Dendrogram > Customize

Allows you to add a title and control y-axis labeling and displaying for the dendrogram.

Double-click the dendrogram after you create it to specify the line type, color, and size for the cluster groups. See Graph Editing Overview.

Dialog box items

Title: To display a title above the dendrogram, type the desired text in this box.

Label Y Axis with

Similarity: Choose to display similarities on the y-axis.

Distance: Choose to display distances on the y-axis.

Show Dendrogram in

One graph: Choose to display the dendrogram in a single graph window.

Maximum number of variables per graph (without splitting a group): Choose to display a specified number of variables per graph and enter an integer greater than or equal to 1.

Cluster Variables – Storage

Stat > Multivariate > Cluster Variables > Storage

Allows you to store the distance matrix.

Dialog box items

Distance matrix: Specify a storage matrix (M) for the P x P distance matrix (D), where P is the number of variables. The stored distance matrix can then be used in subsequent commands.

Example of Cluster Variables

You conduct a study to determine the long-term effect of a change in environment on blood pressure. The subjects are 39 Peruvian males over 21 years of age who had migrated from the Andes mountains to larger towns at lower elevations. You recorded their age (Age), years since migration (Years), weight in kg (Weight), height in mm (Height), skin fold of the chin, forearm, and calf in mm (Chin, Forearm, Calf), pulse rate in beats per minute (Pulse), and systolic and diastolic blood pressure (Systol, Diastol).

Your goal is to reduce the number of variables by combining variables with similar characteristics. You use clustering of variables with the default correlation distance measure, average linkage and a dendrogram.

1 Open the worksheet PERU.MTW.

2 Choose **Stat > Multivariate > Cluster Variables**.

3 In **Variables or distance matrix**, enter *Age-Diastol*.

4 For **Linkage Method**, choose **Average**.

5 Check **Show dendrogram**. Click **OK**.

Session window output

Cluster Analysis of Variables: Age, Years, Weight, Height, Chin, Forearm, ...

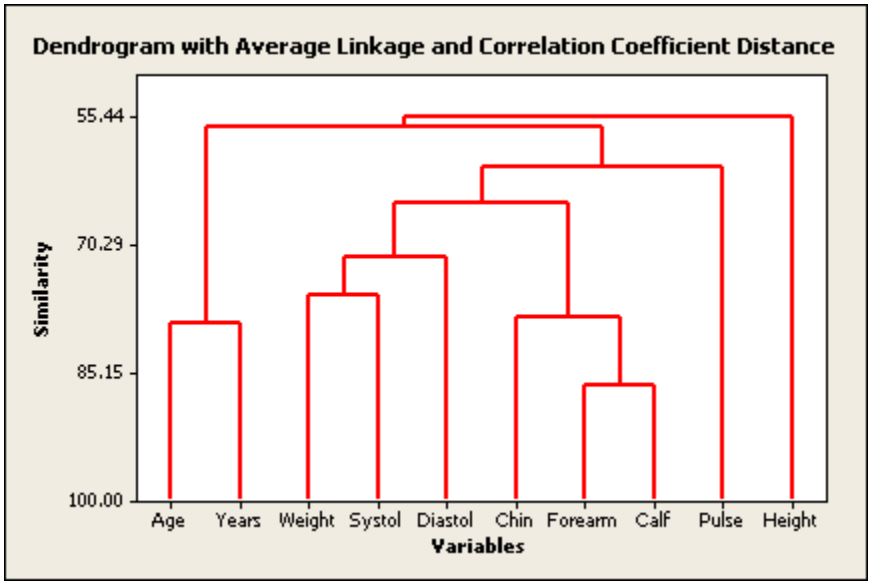
```
Correlation Coefficient Distance, Average Linkage
Amalgamation Steps
```

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	9	86.7763	0.264474	6	7	6	2
2	8	79.4106	0.411787	1	2	1	2
3	7	78.8470	0.423059	5	6	5	3
4	6	76.0682	0.478636	3	9	3	2
5	5	71.7422	0.565156	3	10	3	3
6	4	65.5459	0.689082	3	5	3	6

© 2003 Minitab Inc.

```
7          3      61.3391   0.773218   3       8          3          7
8          2      56.5958   0.868085   1       3          1          9
9          1      55.4390   0.891221   1       4          1         10
```

Graph window output



Interpreting the results

Minitab displays the amalgamation steps in the Session window. At each step, two clusters are joined. The table shows which clusters were joined, the distance between them, the corresponding similarity level, the identification number of the new cluster (this is always the smaller of the two numbers of the clusters joined), the number of variables in the new cluster and the number of clusters. Amalgamation continues until there is just one cluster.

If you had requested a final partition you would also receive a list of which variables are in each cluster.

The dendrogram displays the information printed in the amalgamation table in the form of a tree diagram. Dendrogram suggest variables which might be combined, perhaps by averaging or totaling. In this example, the chin, forearm, and calf skin fold measurements are similar and you decide to combine those. The age and year since migration variables are similar, but you will investigate this relationship. If subjects tend to migrate at a certain age, then these variables could contain similar information and be combined. Weight and the two blood pressure measurements are similar. You decide to keep weight as a separate variable but you will combine the blood pressure measurements into one.

Cluster K-Means

Cluster K-Means

Stat > Multivariate > Cluster K-Means

Use K-means clustering of observations, like clustering of observations, to classify observations into groups when the groups are initially unknown. This procedure uses non-hierarchical clustering of observations according to MacQueen's algorithm [6]. K-means clustering works best when sufficient information is available to make good starting cluster designations.

Dialog box items

Variables: Enter the columns containing measurement data on which to perform the K-means non-hierarchical clustering of observations.

Specify Partition by: Allows you to specify the initial partition for the K-means algorithm.

Number of clusters: Choose to specify the number of clusters to form. If you enter the number 5, for example, Minitab uses the first 5 observations as initial cluster centroids. Each observation is assigned to the cluster whose centroid it is closest to. Minitab recalculates the cluster centroids each time a cluster gains or loses an observation.

Initial partition column: Choose to specify a column containing cluster membership to begin the partition process.

© 2003 Minitab Inc.

Standardize variables: Check to standardize all variables by subtracting the means and dividing by the standard deviation before the distance matrix is calculated. This is a good idea if the variables are in different units and you wish to minimize the effect of scale differences. If you standardize, cluster centroids and distance measures are in standardized variable space before the distance matrix is calculated.

<Storage>

Data – Cluster K Means

You must use raw data as input to K-means clustering of observations. Each row contains measurements on a single item or subject. You must have two or more numeric columns, with each column representing a different measurement. You must delete rows with missing data from the worksheet before using this procedure.

To initialize the clustering process using a data column, you must have a column that contains a cluster membership value for each observation. The initialization column must contain positive, consecutive integers or zeros (it should not contain all zeros). Initially, each observation is assigned to the cluster identified by the corresponding value in this column. An initialization of zero means that an observation is initially unassigned to a group. The number of distinct positive integers in the initial partition column equals the number of clusters in the final partition.

To perform K-means clustering of observations

1 Choose **Stat > Multivariate > Cluster K-Means**.

2 In **Variables**, enter the columns containing the measurement data.

3 If you like, use any dialog box options, then click **OK**.

Initializing Cluster K-Means Process

K-means clustering begins with a grouping of observations into a predefined number of clusters.

1 Minitab evaluates each observation, moving it into the nearest cluster. The nearest cluster is the one which has the smallest Euclidean distance between the observation and the centroid of the cluster.

2 When a cluster changes, by losing or gaining an observation, Minitab recalculates the cluster centroid.

3 This process is repeated until no more observations can be moved into a different cluster. At this point, all observations are in their nearest cluster according to the criterion listed above.

Unlike hierarchical clustering of observations, it is possible for two observations to be split into separate clusters after they are joined together.

K-means procedures work best when you provide good starting points for clusters [10]. There are two ways to initialize the clustering process: specifying a number of clusters or supplying an initial partition column that contains group codes.

You may be able to initialize the process when you do not have complete information to initially partition the data. Suppose you know that the final partition should consist of three groups, and that observations 2, 5, and 9 belong in each of those groups, respectively. Proceeding from here depends upon whether you specify the number of clusters or supply an initial partition column.

- If you specify the number of clusters, you must rearrange your data in the Data window to move observations 2, 5 and 9 to the top of the worksheet, and then specify 3 for Number of clusters.
- If you enter an initial partition column, you do not need to rearrange your data in the Data window. In the initial partition worksheet column, enter group numbers 1, 2, and 3, for observations 2, 5, and 9, respectively, and enter 0 for the other observations.

The final partition will depend to some extent on the initial partition that Minitab uses. You might try different initial partitions. According to Milligan [10], K-means procedures may not perform as well when the initializations are done arbitrarily. However, if you provide good starting points, K-means clustering may be quite robust.

To initialize the process by specifying the number of clusters

1 Choose **Stat > Multivariate > Cluster K-Means**.

2 In **Variables**, enter the columns containing the measurement data.

3 Under **Specify Partition by**, choose **Number of clusters** and enter a number, k, in the box. Minitab will use the first k observations as initial cluster seeds, or starting locations.

4 Click **OK**.

© 2003 Minitab Inc.

To initialize the process using a data column

1 Choose **Stat > Multivariate > Cluster K-Means**.

2 In **Variables**, enter the columns containing the measurement data.

3 Under **Specify Partition by**, choose **Initial partition column**. Enter the column containing the initial cluster membership for each observation.

4 Click **OK**.

Cluster K-Means – Storage

Stat > Multivariate > Cluster K-Means > Storage

Allows cluster membership for each observation and the distance between each observation and each cluster centroid.

Dialog box items

Cluster membership column: Enter a single storage column for final cluster membership for each observation. This column can then be used as a categorical variable in other Minitab commands, such as Discriminant Analysis or Plot.

Distance between observations and cluster centroids (Give a column for each cluster group): Enter storage columns for the distance between each observation and each cluster centroid. The number of columns specified must equal the number of clusters specified for the initial partition. The distances stored are Euclidean distances.

Example of Cluster K-Means

You live-trap, anesthetize, and measure one hundred forty-three black bears. The measurements are total length and head length (Length, Head.L), total weight and head weight (Weight, Weight.H), and neck girth and chest girth (Neck.G, Chest.G). You wish to classify these 143 bears as small, medium-sized, or large bears. You know that the second, seventy-eighth, and fifteenth bears in the sample are typical of the three respective categories. First, you create an initial partition column with the three seed bears designated as 1 = small, 2 = medium-sized, 3 = large, and with the remaining bears as 0 (unknown) to indicate initial cluster membership. Then you perform K-means clustering and store the cluster membership in a column named BearSize.

1 Open the worksheet BEARS.MTW.

2 To create the initial partition column, choose **Calc > Make Patterned Data > Simple Set of Numbers**.

3 In **Store patterned data in**, enter *Initial* for the storage column name.

4 In both **From first value** and **From last value**, enter *0*.

5 In **List each value**, type *143*. Click **OK**.

6 Go to the Data window and type *1*, *2*, and *3* in the second, seventy-eighth, and fifteenth rows, respectively, of the column named *Initial*.

7 Choose **Stat > Multivariate > Cluster K-Means**.

8 In **Variables**, enter **'Head.L'**–**Weight**.

9 Under **Specify Partition by**, choose **Initial partition column** and enter *Initial*.

10 Check **Standardize variables**.

11 Click **Storage**. In **Cluster membership column**, enter *BearSize*.

12 Click **OK** in each dialog box.

Session window output

K-means Cluster Analysis: Head.L, Head.W, Neck.G, Length, Chest.G, Weight

```
Standardized Variables


Final Partition


Number of clusters: 3

                     Within   Average   Maximum
                    cluster  distance  distance
          Number of   sum of      from      from
       observations  squares  centroid  centroid
```


© 2003 Minitab Inc.

```
Cluster1            41   63.075     1.125     2.488
Cluster2            67   78.947     0.997     2.048
Cluster3            35   65.149     1.311     2.449


Cluster Centroids

                                          Grand
Variable  Cluster1  Cluster2  Cluster3  centroid
Head.L     -1.0673    0.0126    1.2261   -0.0000
Head.W     -0.9943   -0.0155    1.1943    0.0000
Neck.G     -1.0244   -0.1293    1.4476   -0.0000
Length     -1.1399    0.0614    1.2177    0.0000
Chest.G    -1.0570   -0.0810    1.3932   -0.0000
Weight     -0.9460   -0.2033    1.4974   -0.0000


Distances Between Cluster Centroids

          Cluster1  Cluster2  Cluster3
Cluster1    0.0000    2.4233    5.8045
Cluster2    2.4233    0.0000    3.4388
Cluster3    5.8045    3.4388    0.0000
```

Interpreting the results

K-means clustering classified the 143 bears as 41 small bears, 67 medium-size bears, and 35 large bears. Minitab displays, in the first table, the number of observations in each cluster, the within cluster sum of squares, the average distance from observation to the cluster centroid, and the maximum distance of observation to the cluster centroid. In general, a cluster with a small sum of squares is more compact than one with a large sum of squares. The centroid is the vector of variable means for the observations in that cluster and is used as a cluster midpoint.

The centroids for the individual clusters are displayed in the second table while the third table gives distances between cluster centroids.

The column BearSize contains the cluster designations.

Discriminant Analysis

Discriminant Analysis

Stat > Multivariate > Discriminant Analysis

Use discriminant analysis to classify observations into two or more groups if you have a sample with known groups. Discriminant analysis can also used to investigate how variables contribute to group separation.

Minitab offers both linear and quadratic discriminant analysis. With linear discriminant analysis, all groups are assumed to have the same covariance matrix. Quadratic discrimination does not make this assumption but its properties are not as well understood.

In the case of classifying new observations into one of two categories, logistic regression may be superior to discriminant analysis [3], [11].

Dialog box items

Groups: Choose the column containing the group codes. There may be up to 20 groups.

Predictors: Choose the column(s) containing the measurement variables or predictors.

Discriminant Function

Linear: Choose to perform linear discriminant analysis. All groups are assumed to have the same covariance matrix.

Quadratic: Choose to perform quadratic discriminant analysis. No assumption is made about the covariance matrix; its properties are not as well understood.

Use cross validation: Check to perform the discrimination using cross-validation. This technique is used to compensate for an optimistic apparent error rate.

Storage

Linear discriminant function: Enter storage columns for the coefficients from the linear discriminant function, using one column for each group. The constant is stored at the top of each column.

Fits: Check to store the fitted values. The fitted value for an observation is the group into which it is classified.

Fits from cross validation: Check to store the fitted values if discrimination was done using cross-validation.

<Options>

© 2003 Minitab Inc.

Data – Discriminant Analysis

Set up your worksheet so that a row of data contains information about a single item or subject. You must have one or more numeric columns containing measurement data, or predictors, and a single grouping column containing up to 20 groups. The column of group codes may be numeric, text, or date/time. If you wish to change the order in which text groups are processed from their default alphabetized order, you can define your own order. (See Ordering Text Categories). Minitab automatically omits observations with missing measurements or group codes from the calculations.

If a high degree of multicollinearity exists (i.e., if one or more predictors is highly correlated with another) or one or more of the predictors is essential constant, discriminant analysis calculations cannot be done and Minitab displays a message to that effect.

To perform linear discriminant analysis

1 Choose **Stat > Multivariate > Discriminant Analysis**.

2 In **Groups**, enter the column containing the group codes.

3 In **Predictors**, enter the column or columns containing the measurement data.

4 If you like, use any dialog box options, then click **OK**.

Linear discriminant analysis

An observation is classified into a group if the squared distance (also called the Mahalanobis distance) of observation to the group center (mean) is the minimum. An assumption is made that covariance matrices are equal for all groups. There is a unique part of the squared distance formula for each group and that is called the linear discriminant function for that group. For any observation, the group with the smallest squared distance has the largest linear discriminant function and the observation is then classified into this group.

Linear discriminant analysis has the property of symmetric squared distance: the linear discriminant function of group i evaluated with the mean of group j is equal to the linear discriminant function of group j evaluated with the mean of group i.

We have described the simplest case, no priors and equal covariance matrices. If you consider Mahalanobis distance a reasonable way to measure the distance of an observation to a group, then you do not need to make any assumptions about the underlying distribution of your data.

See Prior Probabilities for more information.

Quadratic discriminant analysis

There is no assumption with quadratic discriminant analysis that the groups have equal covariance matrices. As with linear discriminant analysis, an observation is classified into the group that has the smallest squared distance. However, the squared distance does not simplify into a linear function, hence the name quadratic discriminant analysis.

Unlike linear distance, quadratic distance is not symmetric. In other words, the quadratic discriminant function of group i evaluated with the mean of group j is not equal to the quadratic discriminant function of group j evaluated with the mean of group i. On the results, quadratic distance is called the generalized squared distance. If the determinant of the sample group covariance matrix is less than one, the generalized squared distance can be negative.

Cross-Validation

Cross-validation is one technique that is used to compensate for an optimistic apparent error rate. The apparent error rate is the percent of misclassified observations. This number tends to be optimistic because the data being classified are the same data used to build the classification function.

The cross-validation routine works by omitting each observation one at a time, recalculating the classification function using the remaining data, and then classifying the omitted observation. The computation time takes approximately four times longer with this procedure. When cross-validation is performed, Minitab displays an additional summary table.

Another technique that you can use to calculate a more realistic error rate is to split your data into two parts. Use one part to create the discriminant function, and the other part as a validation set. Predict group membership for the validation set and calculate the error rate as the percent of these data that are misclassified.

Prior Probabilities

Sometimes items or subjects from different groups are encountered according to different probabilities. If you know or can estimate these probabilities a priori, discriminant analysis can use these so-called prior probabilities in calculating the posterior probabilities, or probabilities of assigning observations to groups given the data. With the assumption that the

 © 2003 Minitab Inc.

data have a normal distribution, the linear discriminant function is increased by ln(p_i), where p_i is the prior probability of group i. Because observations are assigned to groups according to the smallest generalized distance, or equivalently the largest linear discriminant function, the effect is to increase the posterior probabilities for a group with a high prior probability.

Now suppose we have priors and suppose $f_i(x)$ is the joint density for the data in group i (with the population parameters replaced by the sample estimates).

The posterior probability is the probability of group i given the data and is calculated by

$$\frac{p_i f_i(x)}{\sum_i p_i f_i(x)}$$

The largest posterior probability is equivalent to the largest value of $\ln [p_i f_i (x)]$.

If $f_i(x)$ is the normal distribution, then

$$\ln [p_i f_i (x)] = -0.5 [d_i^2 (x) - 2 \ln p_i] - \text{(a constant)}$$

The term in square brackets is called the generalized squared distance of x to group i and is denoted by $d_i^2 (x)$. Notice,

$$d_i^2 (x) = -2 [\mathbf{m}_i' \mathbf{S}_p^{-1} x - 0.5 \mathbf{m}_i' \mathbf{S}_p^{-1} \mathbf{m}_i + \ln p_i] + x' \mathbf{S}_p^{-1} x$$

The term in square brackets is the linear discriminant function. The only difference from the non-prior case is a change in the constant term. Notice, the largest posterior is equivalent to the smallest generalized distance, which is equivalent to the largest linear discriminant function.

Predicting group membership for new observations

Generally, discriminant analysis is used to calculate the discriminant functions from observations with known groups. When new observations are made, you can use the discriminant function to predict which group that they belong to. You can do this by either calculating (using **Calc > Calculator**) the values of the discriminant function for the observation(s) and then assigning it to the group with the highest function value or by using Minitab's discriminant procedure. See To predict group membership for new observations.

To predict group membership for new observations

1 Choose **Stat > Multivariate > Discriminant Analysis**.

2 In **Groups**, enter the column containing the group codes from the original sample.

3 In **Predictors**, enter the column(s) containing the measurement data of the original sample.

4 Click **Options**. In **Predict group membership for**, enter constants or columns representing one or more observations. The number of constants or columns must be equivalent to the number of predictors.

5 If you like, use any dialog box options, and click **OK**.

Discriminant Analysis – Options

Stat > Multivariate > Discriminant Analysis > Options

Allows you to specify prior probabilities, predict group membership for new observations, and control the display of Session window output.

Dialog box items

Prior probabilities: Enter prior probabilities. You may type in the probabilities or specify constants (K) that contain stored values. There should be one value for each group. The first value will be assigned to the group with the smallest code, the second to the group with the second smallest code, etc. If the probabilities do not sum to one, Minitab normalizes them. See Prior Probabilities.

Predict group membership for: Enter values for predicting group membership for new observations.

Display of Results:

Do not display: Choose to suppress all results. Requested storage is done.

Classification matrix: Choose to display only the classification matrix.

Above plus ldf, distances, and misclassification summary: Choose to display the classification matrix, the squared distance between group centers, linear discriminant function, and a summary of misclassified observations.

© 2003 Minitab Inc.

Above plus mean, std. dev., and covariance summary: Choose to display the classification matrix, the squared distance between group centers, linear discriminant function, a summary of misclassified observations, means, standard deviations, and covariance matrices, for each group and pooled.

Above plus complete classification summary: Choose to display the classification matrix, the squared distance between group centers, linear discriminant function, a summary of misclassified observations, means, standard deviations, covariance matrices, for each group and pooled, and a summary of how all observations were classified. Minitab notes misclassified observations with two asterisks beside the observation number.

Example of Discriminant Analysis

In order to regulate catches of salmon stocks, it is desirable to identify fish as being of Alaskan or Canadian origin. Fifty fish from each place of origin were caught and growth ring diameters of scales were measured for the time when they lived in freshwater and for the subsequent time when they lived in saltwater. The goal is to be able to identify newly-caught fish as being from Alaskan or Canadian stocks. The example and data are from [6], page 519-520.

1 Open the worksheet EXH_MVAR.MTW.

2 Choose **Stat > Multivariate > Discriminant Analysis**.

3 In **Groups**, enter *SalmonOrigin*.

4 In **Predictors**, enter *Freshwater Marine*. Click **OK**.

Session window output

Discriminant Analysis: SalmonOrigin versus Freshwater, Marine

```
Linear Method for Response: SalmonOrigin


Predictors: Freshwater, Marine


Group     Alaska    Canada
Count         50        50


Summary of classification

                   True Group
Put into Group  Alaska  Canada
Alaska              44       1
Canada               6      49
Total N             50      50
N correct           44      49
Proportion       0.880   0.980

N = 100           N Correct = 93           Proportion Correct = 0.930


Squared Distance Between Groups

          Alaska   Canada
Alaska  0.00000  8.29187
Canada  8.29187  0.00000


Linear Discriminant Function for Groups

             Alaska  Canada
Constant    -100.68  -95.14
Freshwater     0.37    0.50
Marine         0.38    0.33


Summary of Misclassified Observations

                                                 Squared
Observation    True Group  Pred Group   Group  Distance  Probability
          1**      Alaska      Canada  Alaska     3.544        0.428
```


© 2003 Minitab Inc.

			Canada	2.960	0.572
2**	Alaska	Canada	Alaska	8.1131	0.019
			Canada	0.2729	0.981
12**	Alaska	Canada	Alaska	4.7470	0.118
			Canada	0.7270	0.882
13**	Alaska	Canada	Alaska	4.7470	0.118
			Canada	0.7270	0.882
30**	Alaska	Canada	Alaska	3.230	0.289
			Canada	1.429	0.711
32**	Alaska	Canada	Alaska	2.271	0.464
			Canada	1.985	0.536
71**	Canada	Alaska	Alaska	2.045	0.948
			Canada	7.849	0.052

Interpreting the results

As shown in the Summary of Classification table, the discriminant analysis correctly identified 93 of 100 fish, though the probability of correctly classifying an Alaskan fish was lower (44/50 or 88%) than was the probability of correctly classifying a Canadian fish (49/50 or 98%). To identify newly-caught fish, you could compute the linear discriminant functions associated with Alaskan and Canadian fish and identify the new fish as being of a particular origin depending upon which discriminant function value is higher. You can either do this by using **Calc > Calculator** using stored or output values, or performing discriminant analysis again and predicting group membership for new observations.

The Summary of Misclassified Observations table shows the squared distances from each misclassified point to group centroids and the posterior probabilities. The squared distance value is that value from observation to the group centroid, or mean vector. The probability value is the posterior probability. Observations are assigned to the group with the highest posterior probability.

Simple Correspondence Analysis

Simple Correspondence Analysis

Stat > Multivariate > Simple Correspondence Analysis

Simple correspondence analysis helps you to explore relationships in a two-way classification. Simple correspondence analysis can also operate on three-way and four-way tables because they can be collapsed into two-way tables. This procedure decomposes a contingency table in a manner similar to how principal components analysis decomposes multivariate continuous data. An eigen analysis of the data is performed, and the variability is broken down into underlying dimensions and associated with rows and/or columns.

Dialog box items

Input Data

Categorical variables: Choose to enter the data as categorical variables. If you do not use the Combine subdialog box, enter two worksheet columns. The first is for the row categories; the second is for the column categories. Minitab then forms a contingency table from the input data.

Columns of a contingency table: Choose to enter the data as columns of a contingency table. Each worksheet column you enter will be used as one column of the contingency table. All values in the contingency columns must be positive integers or zero.

Row names: Enter a column that contains names for the rows of the contingency table. The name column must be a text column whose length matches the number of rows in the contingency table. Minitab prints the first 8 characters of the names in tables, but prints the full name on graphs. If you do not enter names here, the rows will be named Row1, Row2, etc.

Column names: Enter a column that contains names for the columns of the contingency table. The name column must be a text column whose length matches the number of columns in the contingency table. Minitab prints the first 8 characters of the names in tables, but prints the full name on graphs. If you do not enter names here, the columns will be named Column1, Column2, etc.

Number of components: Enter the number of components to calculate. The minimum number of components is one. The maximum number of components for a contingency table with r rows and c columns is the smaller of (r-1) or (c-1), which is equivalent to the dimension of the subspace onto which you project the profiles. The default number of components is 2.

<Combine>

<Supp Data>

<Results>

<Graphs>

© 2003 Minitab Inc.

<Storage>

Data – Simple Correspondence Analysis

Worksheet data may be arranged in two ways: raw or contingency table form. See Arrangement of Input Data. Worksheet data arrangement determines acceptable data values.

- If your data are in raw form, you can have two, three, or four classification columns with each row representing one observation. All columns must be the same length. The data represent categories and may be numeric, text, or date/time. If the categories in a column are text data, the levels are used in the order of first occurrence, i.e., the first level becomes the first row (column) of the table, the next distinct level becomes the second row (column) of the table, and so on. If you wish to change the order in which text categories are processed from their default alphabetized order, you can define your own order. See Ordering Text Categories. You must delete missing data before using this procedure. Because simple correspondence analysis works with a two-way classification, the standard approach is to use two worksheet columns. However, you can obtain a two-way classification with three or four variables by crossing variables within the simple correspondence analysis procedure. See Crossing variables to create a two-way table.
- If your data are in contingency table form, worksheet columns must contain integer frequencies of your category combinations. You must delete any rows or columns with missing data or combine them with other rows or columns. Unlike the χ^2 test for association procedure, there is no set limit on the number of contingency table columns. You could use simple correspondence analysis to obtain χ^2 statistics for large tables.

Supplementary data

When performing a simple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or supplementary data in the same form as the main set, because you can see how these supplementary data are "scored" using the results from the main set. These supplementary data may be further information from the same study, information from other studies, or target profiles [4]. Minitab does not include these data when calculating the components, but you can obtain a profile and display supplementary data in graphs.

You can have row supplementary data or column supplementary data. Row supplementary data constitutes an additional row(s) of the contingency table, while column supplementary data constitutes an additional column(s) of the contingency table. Supplementary data must be entered in contingency table form. Therefore, each worksheet column of these data must contain c entries (where c is the number of contingency table columns) or r entries (where r is the number of contingency table rows).

To perform a simple correspondence analysis

1 Choose **Stat > Multivariate > Simple Correspondence Analysis**.

2 How you enter your data depends on the form of the data and the number of categorical variables.
   - If you have two categorical variables, do one of the following:
     - For raw data, enter the columns containing the raw data in **Categorical variables**.
     - For contingency table data, enter the columns containing the data in **Columns of a contingency table**.
   - If you have three or four categorical variables, you must cross some variables before entering data as shown above. See Crossing variables to create a two-way table.

3 If you like, use any dialog box options, then click **OK**.

Crossing variables to create a two-way table

Crossing variables allows you to use simple correspondence analysis to analyze three-way and four-way contingency tables. You can cross the first two variables to form rows and/or the last two variables to form columns. You must enter three categorical variables to perform one cross, and four categorical variables to perform two crosses.

The following example illustrates row crossing. Column crossing is similar. Suppose you have two variables. The row variable, Sex, has two levels: male and female; the column variable, Age, has three levels; young, middle aged, old.

Crossing Sex with Age will create 2 x 3 = 6 rows, ordered as follows:

male / young
male / middle aged
male / old
female / young
female / middle aged
female / old

Simple Correspondence Analysis – Combine

Stat > Multivariate > Simple Correspondence Analysis > Combine

© 2003 Minitab Inc.

Crossing variables allows you to use simple correspondence analysis to analyze three-way and four-way contingency tables. You can cross the first two variables to form rows and/or the last two variables to form columns. You must enter three categorical variables to perform one cross, and four categorical variables to perform two crosses.

In order to cross columns, you must choose **Categorical variables** for **Input Data** rather than **Columns of a contingency table** in the main dialog box. If you want to cross for either just the rows or for just the columns of the contingency table, you must enter three worksheet columns in the **Categorical variables** text box. If you want to cross both the rows and the columns of the table, you must specify four worksheet columns in this text box.

Dialog box items

Define Rows of the Contingency Table Using:

First variable: Choose to use the first input column to form the rows of the contingency table. Thus, the rows of the contingency table are not formed by crossing variables.

First 2 variables crossed: Choose to cross the categories in the first two input columns to form the rows of the contingency table. For example, if the first variable is Sex (with 2 levels, male and female) and the second variable is Age (with 3 levels, young, middle aged, old), then there will be 2 x 3 = 6 rows, ordered as follows:

males / young
males / middle aged
males / old
females / young
females / middle aged
females / old

Define Columns of Contingency Table Using:

Last variable: Choose to use the last input column to form the columns of the contingency table.

Last 2 variables crossed: Choose to cross the categories in the last two input columns to form the columns of the contingency table.

Simple Correspondence Analysis – Supplementary Data

Stat > Multivariate > Simple Correspondence Analysis > Supp Data

When performing a simple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or supplementary data in the same form as the main set, because you can see how these supplementary data are "scored" using the results from the main set. See What are Supplementary Data?

Dialog box items

Supplementary Rows: Enter one or more columns containing additional rows of the contingency table.

Supplementary Columns: Enter one or more columns containing additional columns of the contingency table.

Row names: Enter a column containing text names for the supplementary rows.

Column names: Enter a column containing text names for the supplementary columns.

What are Supplementary Data?

When performing a simple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or supplementary data in the same form as the main set, because you can see how these supplementary data are "scored" using the results from the main set. These supplementary data may be further information from the same study, information from other studies, or target profiles [4]. Minitab does not include these data when calculating the components, but you can obtain a profile and display supplementary data in graphs.

You can have row supplementary data or column supplementary data. Row supplementary data constitutes an additional row(s) of the contingency table, while column supplementary data constitutes an additional column(s) of the contingency table. Supplementary data must be entered in contingency table form. Therefore, each worksheet column of these data must contain c entries (where c is the number of contingency table columns) or r entries (where r is the number of contingency table rows).

Simple Correspondence Analysis – Results

Stat > Multivariate > Simple Correspondence Analysis > Results

Allows you to control displayed output.

Dialog box items

Contingency table: Check to display the contingency table.

Row profiles: Check to display a table of row profiles and row masses.

© 2003 Minitab Inc.

Columns profiles: Check to display a table of column profiles and column masses.

Expected frequencies: Check to display a table of the expected frequency in each cell of the contingency table.

Observed - expected frequencies: Check to display a table of the observed minus the expected frequency in each cell of the contingency table.

Chi-square values: Check to display a table of the χ^2 value in each cell of the contingency table.

Inertias: Check to display the table of the relative inertia in each cell of the contingency table.

Simple Correspondence Analysis – Graphs

Stat > Multivariate > Simple Correspondence Analysis > Graphs

Allows you display various plots to complement your analysis. See Simple correspondence analysis graphs.

In all plots, row points are plotted with red circles--solid circles for regular points, and open circles for supplementary points. Column points are plotted with blue squares--solid squares for regular points, and open squares for supplementary points.

The aspect ratio of the plots is one-to-one so that a unit on the x-axis is equal to a unit on the y-axis.

Dialog box items

Axis pairs for all plots (Y then X): Enter between 1 and 15 axis pairs for each requested plot. The axes you list must be axes in the subspace you defined in the main dialog box. For example, if you entered 4 in number of components, you can only list axes 1, 2, 3, and 4.

The first axis in a pair will be the Y or vertical axis of the plot; the second axis will be the X or horizontal axis of the plot. For example, if you enter 2 1 3 1 plots component 2 versus component 1, and component 3 versus component 1.

Show supplementary points in all plots: Check to display supplementary points on all plots.

Plots:

Symmetric plot showing rows only: Check to display a plot that shows the row principal coordinates.

Symmetric plot showing columns only: Check to display a plot that shows the column principal coordinates.

Symmetric plot showing rows and columns: Check to display a symmetric plot that shows both row principal coordinates and column principal coordinates overlaid in a joint display.

Asymmetric row plot showing rows and columns: Check to display an asymmetric row plot.

Asymmetric column plot showing rows and columns: Check to display an asymmetric column plot.

To display simple correspondence analysis plots

1 Perform steps 1–2 of To perform a simple correspondence analysis.

2 Click **Graphs** and check all of the plots that you would like to display.

3 If you like, you can specify the component pairs and their axes for plotting. Enter between 1 and 15 component pairs in **Axis pairs for all plots (Y then X)**. Minitab plots the first component in each pair on the vertical or y-axis of the plot; the second component in the pair on the horizontal or x-axis of the plot.

4 If you have supplementary data and would like to include this data in the plot(s), check **Show supplementary points in all plots**. Click **OK** in each dialog box.

In all plots, row points are plotted with red circles–solid circles for regular points, and open circles for supplementary points. Column points are plotted with blue squares–blue squares for regular points, and open squares for supplementary points.

Choosing a simple correspondence analysis graph

You can display the following simple correspondence-analysis plots:

- A row plot or a column plot
- A symmetric plot
- An asymmetric row plot or an asymmetric column plot

A row plot is a plot of row principal coordinates. A column plot is a plot of column principal coordinates.

A symmetric plot is a plot of row and column principal coordinates in a joint display. An advantage of this plot is that the profiles are spread out for better viewing of distances between them. The row-to-row and column-to-column distances are approximate χ^2 distances between the respective profiles. However, this same interpretation cannot be made for row-to-column distances. Because these distances are two different mappings, you must interpret these plots carefully [4].

© 2003 Minitab Inc.

An asymmetric row plot is a plot of row principal coordinates and of column standardized coordinates in the same plot. Distances between row points are approximate χ^2 distances between the row profiles. Choose the asymmetric row plot over the asymmetric column plot if rows are of primary interest.

An asymmetric column plot is a plot of column principal coordinates and row standardized coordinates. Distances between column points are approximate χ^2 distances between the column profiles. Choose an asymmetric column plot over an asymmetric row plot if columns are of primary interest.

An advantage of asymmetric plots is that there can be an intuitive interpretation of the distances between row points and column points, especially if the two displayed components represent a large proportion of the total inertia [4]. Suppose you have an asymmetric row plot, as shown in Example of simple correspondence analysis. This graph plots both the row profiles and the column vertices for components 1 and 2. The closer a row profile is to a column vertex, the higher the row profile is with respect to the column category. In this example, of the row points, Biochemistry is closest to column category E, implying that biochemistry as a discipline has the highest percentage of unfunded researchers in this study. A disadvantage of asymmetric plots is that the profiles of interest are often bunched in the middle of the graph [4], as happens with the asymmetric plot of this example.

Simple Correspondence Analysis – Storage

Stat > Multivariate > Simple Correspondence Analysis > Storage

Allows you to store results. In the four cases that store coordinates, the coordinate for the first component is stored in the first column, the coordinate for the second component in the second column, and so on. If there are supplementary points, their coordinates are stored at the ends of the columns.

Dialog box items

Columns of the contingency table: Enter one worksheet column for each column of the contingency table. Minitab does not store supplementary rows and columns.

Row principal coordinates: Check to store the row principal coordinates. Minitab stores the coordinate for the first component in a column named RPC1, the coordinate for the second component in a column that named RPC2, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Row standardized coordinates: Check to store the row standardized coordinates. Minitab stores the coordinate for the first component in a column named RSC1, the coordinate for the second component in a column that named RSC2, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Column principal coordinates: Check to store the column principal coordinates. Minitab stores the coordinate for the first component in a column named CPC1, the coordinate for the second component in a column that named CPC2, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Column standardized coordinates: Check to store the column standardized coordinates. Minitab stores the coordinate for the first component in a column named CSC1, the coordinate for the second component in a column that named CSC2, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Example of Simple Correspondence Analysis

The following example is from Correspondence Analysis in Practice, by M. J. Greenacre, p.75. Seven hundred ninety-six researchers were cross-classified into ten academic disciplines and five funding categories, where A is the highest funding category, D is the lowest, and category E is unfunded. Here, disciplines are rows and funding categories are columns. You wish to see how the disciplines compare to each other relative to the funding categories so you perform correspondence analysis from a row orientation. Supplementary data include: a row for museum researchers not included in the study and a row for mathematical sciences, which is the sum of Mathematics and Statistics.

1 Open the worksheet EXH_TABL.MTW.

2 Choose **Stat > Multivariate > Simple Correspondence Analysis**.

3 Choose **Columns of a contingency table,** and enter *CT1-CT5*. In **Row names**, enter *RowNames*. In **Column names**, enter *ColNames*.

4 Click **Results** and check **Row profiles**. Click **OK**.

5 Click **Supp Data**. In **Supplementary Rows**, enter *RowSupp1 RowSupp2*. In **Row names**, enter *RSNames*. Click **OK**.

6 Click **Graphs**. Check **Show supplementary points in all plots**. Check **Symmetric plot showing rows only** and **Asymmetric row plot showing rows and columns**.

7 Click **OK** in each dialog box.

© 2003 Minitab Inc.

Session window output

Simple Correspondence Analysis: CT1, CT2, CT3, CT4, CT5

Row Profiles

	A	B	C	D	E	Mass
Geology	0.035	0.224	0.459	0.165	0.118	0.107
Biochemistry	0.034	0.069	0.448	0.034	0.414	0.036
Chemistry	0.046	0.192	0.377	0.162	0.223	0.163
Zoology	0.025	0.125	0.342	0.292	0.217	0.151
Physics	0.088	0.193	0.412	0.079	0.228	0.143
Engineering	0.034	0.125	0.284	0.170	0.386	0.111
Microbiology	0.027	0.162	0.378	0.135	0.297	0.046
Botany	0.000	0.140	0.395	0.198	0.267	0.108
Statistics	0.069	0.172	0.379	0.138	0.241	0.036
Mathematics	0.026	0.141	0.474	0.103	0.256	0.098
Mass	0.039	0.161	0.389	0.162	0.249	

Analysis of Contingency Table

Axis	Inertia	Proportion	Cumulative	Histogram
1	0.0391	0.4720	0.4720	******************************
2	0.0304	0.3666	0.8385	**********************
3	0.0109	0.1311	0.9697	********
4	0.0025	0.0303	1.0000	*
Total	0.0829			

Row Contributions

					Component 1			
ID	Name	Qual	Mass	Inert	Coord	Corr	Contr	
1	Geology	0.916	0.107	0.137	-0.076	0.055	0.016	
2	Biochemistry	0.881	0.036	0.119	-0.180	0.119	0.030	
3	Chemistry	0.644	0.163	0.021	-0.038	0.134	0.006	
4	Zoology	0.929	0.151	0.230	0.327	0.846	0.413	
5	Physics	0.886	0.143	0.196	-0.316	0.880	0.365	
6	Engineering	0.870	0.111	0.152	0.117	0.121	0.039	
7	Microbiology	0.680	0.046	0.010	-0.013	0.009	0.000	
8	Botany	0.654	0.108	0.067	0.179	0.625	0.088	
9	Statistics	0.561	0.036	0.012	-0.125	0.554	0.014	
10	Mathematics	0.319	0.098	0.056	-0.107	0.240	0.029	

		Component 2		
ID	Name	Coord	Corr	Contr
1	Geology	-0.303	0.861	0.322
2	Biochemistry	0.455	0.762	0.248
3	Chemistry	-0.073	0.510	0.029
4	Zoology	-0.102	0.083	0.052
5	Physics	-0.027	0.006	0.003
6	Engineering	0.292	0.749	0.310
7	Microbiology	0.110	0.671	0.018
8	Botany	0.039	0.029	0.005
9	Statistics	-0.014	0.007	0.000
10	Mathematics	0.061	0.079	0.012

Supplementary Rows

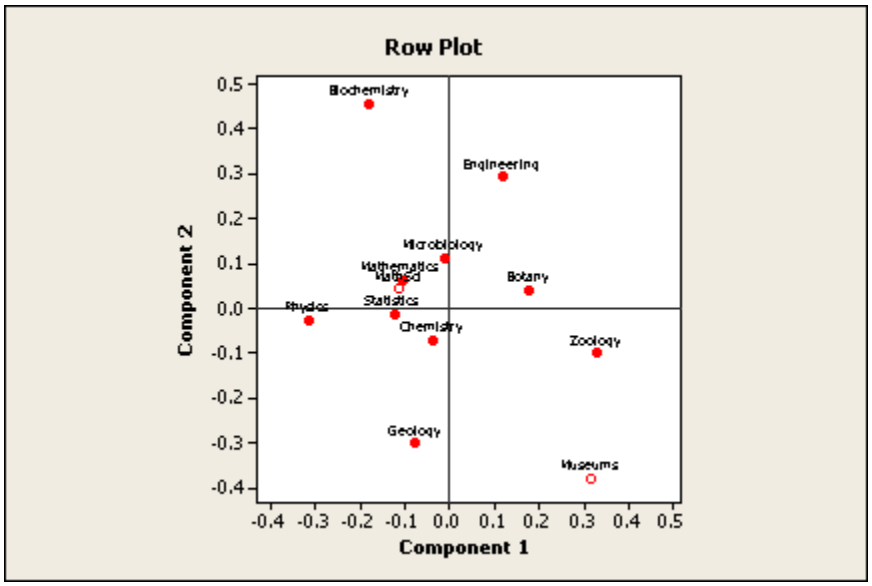
					Component 1			Component 2		
ID	Name	Qual	Mass	Inert	Coord	Corr	Contr	Coord	Corr	Contr
1	Museums	0.556	0.067	0.353	0.314	0.225	0.168	-0.381	0.331	0.318
2	MathSci	0.559	0.134	0.041	-0.112	0.493	0.043	0.041	0.066	0.007

Column Contributions

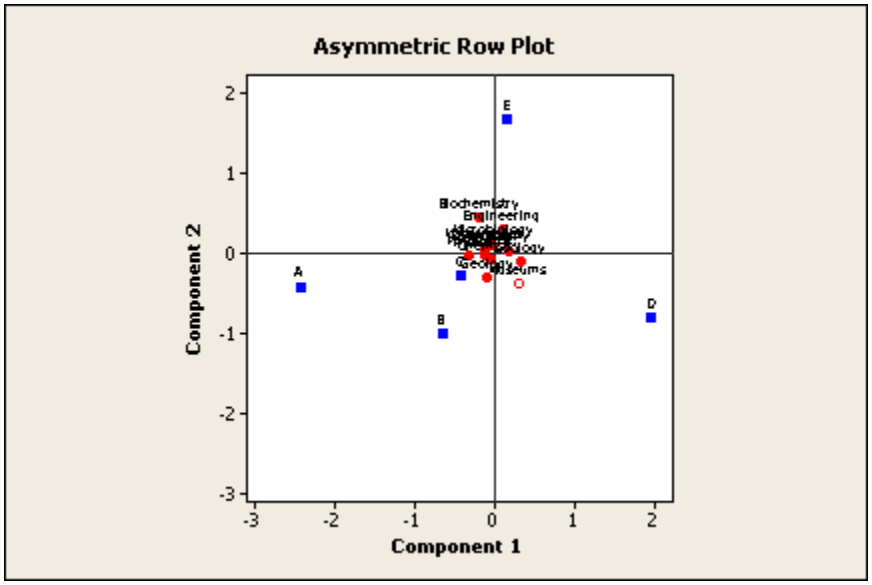
© 2003 Minitab Inc.

					Component 1			Component 2		
ID	Name	Qual	Mass	Inert	Coord	Corr	Contr	Coord	Corr	Contr
1	A	0.587	0.039	0.187	-0.478	0.574	0.228	-0.072	0.013	0.007
2	B	0.816	0.161	0.110	-0.127	0.286	0.067	-0.173	0.531	0.159
3	C	0.465	0.389	0.094	-0.083	0.341	0.068	-0.050	0.124	0.032
4	D	0.968	0.162	0.347	0.390	0.859	0.632	-0.139	0.109	0.103
5	E	0.990	0.249	0.262	0.032	0.012	0.006	0.292	0.978	0.699

Graph window output



Graph window output



Interpreting the results

Row Profiles. The first table gives the proportions of each row category by column. Thus, of the class Geology, 3.5% are in column A, 22.4% are column B, etc. The mass of the Geology row, 0.107, is the proportion of all Geology subjects in the data set.

Analysis of Contingency Table. The second table shows the decomposition of the total inertia. For this example, the table gives a summary of the decomposition of the 10 x 5 contingency table into 4 components. The column labeled Inertia contains the χ squared / n value accounted for by each component. Of the total inertia, 65.972 / 796 or 0.0829,

© 2003 Minitab Inc.

47.2% is accounted for by the first component, 36.66% by the second component, and so on. Here, 65.972 is the χ squared statistic you would obtain if you performed a χ squared test of association with this contingency table.

Row Contributions. You can use the third table to interpret the different components. Since the number of components was not specified, Minitab calculates 2 components.

- The column labeled Qual, or quality, is the proportion of the row inertia represented by the two components. The rows Zoology and Geology, with quality = 0.928 and 0.916, respectively, are best represented among the rows by the two component breakdown, while Math has the poorest representation, with a quality value of 0.319.
- The column labeled Mass has the same meaning as in the Row Profiles table– the proportion of the class in the whole data set.
- The column labeled Inert is the proportion of the total inertia contributed by each row. Thus, Geology contributes 13.7% to the total χ squared statistic.

Next, Minitab displays information for each of the two components (axes).

- The column labeled Coord gives the principal coordinates of the rows.
- The column labeled Corr represents the contribution of the component to the inertia of the row. Thus, Component 1 accounts for most of the inertia of Zoology and Physics (Coor = 0.846 and 0.880, respectively), but explains little of the inertia of Microbiology (Coor = 0.009).
- Contr, the contribution of each row to the axis inertia, shows that Zoology and Physics contribute the most, with Botany contributing to a smaller degree, to Component 1. Geology, Biochemistry, and Engineering contribute the most to Component 2.

Supplementary rows. You can interpret this table in a similar fashion as the row contributions table.

Column Contributions. The fifth table shows that two components explain most of the variability in funding categories B, D, and E. The funded categories A, B, C, and D contribute most to component 1, while the unfunded category, E, contributes most to component 2.

Row Plot. This plot displays the row principal coordinates. Component 1, which best explains Zoology and Physics, shows these two classes well removed from the origin, but with opposite sign. Component 1 might be thought of as contrasting the biological sciences Zoology and Botany with Physics. Component 2 might be thought of as contrasting Biochemistry and Engineering with Geology.

Asymmetric Row Plot. Here, the rows are scaled in principal coordinates and the columns are scaled in standard coordinates. Among funding classes, Component 1 contrasts levels of funding, while Component 2 contrasts being funded (A to D) with not being funded (E). Among the disciplines, Physics tends to have the highest funding level and Zoology has the lowest. Biochemistry tends to be in the middle of the funding level, but highest among unfunded researchers. Museums tend to be funded, but at a lower level than academic researchers

Multiple Correspondence Analysis

Multiple Correspondence Analysis

Stat > Multivariate > Multiple Correspondence Analysis

Multiple correspondence analysis extends simple correspondence analysis to the case of three or more categorical variables. Multiple correspondence analysis performs a simple correspondence analysis on a matrix of indicator variables where each column of the matrix corresponds to a level of categorical variable. Rather than having the two-way table of simple correspondence analysis, here the multi-way table is collapsed into one dimension. By moving from the simple to multiple procedure, you gain information on a potentially larger number of variables, but you may lose information on how rows and columns relate to each other.

Dialog box items

Input Data

Categorical variables: Choose If your data are in raw form and then enter the columns containing the categorical variables.

Indicator variables: Choose if your data are arranged as indicator variables and then enter the columns containing the indicator in the text box. The entries in all columns must be either the integers 0 and 1.

Category names: Enter the column that contains the category names if you want to assign category names. The name column must be a text column whose length matches the number of categories on all categorical variables.

For example, suppose there are 3 categorical variables: Sex (male, female), Hair color (blond, brown, black), and Age (under 20, from 20 to 50, over 50), and no supplementary variables. You would assign 2 + 3 + 3 = 8 category names, so the name column would contain 8 rows.

Minitab only uses the first 8 characters of the names in printed tables, but uses all characters on graphs.

Number of components: Enter the number of components to calculate. The default number of components is 2.

<Supp Data>

 © 2003 Minitab Inc.

<Results>

<Graphs>

<Storage>

Data – Multiple Correspondence Analysis

Worksheet data may be arranged in two ways: raw or indicator variable form. See Arrangement of Input Data. Worksheet data arrangement determines acceptable data values.

- If your data are in **raw form**, you can have one or more classification columns with each row representing one observation. The data represent categories and may be numeric, text, or date/time. If you wish to change the order in which text categories are processed from their default alphabetized order, you can define your own order. See Ordering Text Categories. You must delete missing data before using this procedure.
- If your data are in **indicator variable form**, each row will also represent one observation. There will be one indicator column for each category level. You can use **Calc > Make Indicator Variables** to create indicator variables from raw data. You must delete missing data before using this procedure.

Supplementary data

When performing a multiple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or **supplementary data** in the same form as the main set, and you might want to see how this supplementary data are "scored" using the results from the main set. These supplementary data are typically a classification of your variables that can help you to interpret the results. Minitab does not include these data when calculating the components, but you can obtain a profile and display supplementary data in graphs.

Set up your supplementary data in your worksheet using the same form, either raw data or indicator variables, as you did for the input data. Because your supplementary data will provide additional information about your observations, your supplementary data column(s) must be the same length as your input data.

To perform a multiple correspondence analysis

1 Choose **Stat > Multivariate > Multiple Correspondence Analysis**.

2 To enter your data, do one of the following:
   - For raw data, enter the columns containing the raw data in **Categorical variables**.
   - For indicator variable data, enter the columns containing the indicator variable data in **Indicator variables**.

3 If you like, use any dialog box options, then click **OK**.

Multiple Correspondence Analysis – Supplementary Data

Stat > Multivariate > Multiple Correspondence Analysis > Supp Data

Dialog box items

Supplementary data (in same form as input data): Enter one or more columns containing supplementary column data. See What are Supplementary Data?

Category names: Enter a column containing a text name for each category of all the supplementary data, arranged by numerical order of the corresponding categories by variable.

What are Supplementary Data?

When performing a simple correspondence analysis, you have a main classification set of data on which you perform your analysis. However, you may also have additional or supplementary data in the same form as the main set, because you can see how these supplementary data are "scored" using the results from the main set. These supplementary data may be further information from the same study, information from other studies, or target profiles [4]. Minitab does not include these data when calculating the components, but you can obtain a profile and display supplementary data in graphs.

You can have row supplementary data or column supplementary data. Row supplementary data constitutes an additional row(s) of the contingency table, while column supplementary data constitutes an additional column(s) of the contingency table. Supplementary data must be entered in contingency table form. Therefore, each worksheet column of these data must contain c entries (where c is the number of contingency table columns) or r entries (where r is the number of contingency table rows).

© 2003 Minitab Inc.

Multiple Correspondence Analysis – Results

Stat > Multivariate > Multiple Correspondence Analysis > Results

Allows you to control displayed output.

Dialog box items

Indicator table: Check to display a table of indicator variables.

Burt table: Check to display the Burt table.

Multiple Correspondence Analysis – Graphs

Stat > Multivariate > Multiple Correspondence Analysis > Graphs

Allows you display column plots.

Points are plotted with blue squares--solid squares for regular points, and open squares for supplementary points.

The aspect ratio of the plots is one-to-one so that a unit on the x-axis is equal to a unit on the y-axis.

Dialog box items

Axis pairs for plots (Y then X): Enter one to 15 axis pairs to use for the column plots. The axes you list must be axes in the subspace you defined in the main dialog box. For example, if you entered 4 in number of components, you can only list axes 1, 2, 3, and 4.

The first axis in a pair will be the Y or vertical axis of the plot; the second axis will be the X or horizontal axis of the plot. For example, if you enter 2 1 3 1 plots component 2 versus component 1, and component 3 versus component 1.

Show supplementary points in all plots: Check to display supplementary points on all plots.

Display column plot: Check to display a plot that shows the column coordinates.

Multiple Correspondence Analysis – Storage

Stat > Multivariate > Multiple Correspondence Analysis > Storage

Allows you to store the column coordinates.

Dialog box item

Coordinates for the components: Check to store the column coordinates. Minitab stores the coordinate for the first component in the first listed column, the coordinate for the second component in the second listed column, etc. If there are supplementary points, their coordinates are stored at the ends of the columns.

Example of Multiple Correspondence Analysis

Automobile accidents are classified [8] (data from [3]) according to the type of accident (collision or rollover), severity of accident (not severe or severe), whether or not the driver was ejected, and the size of the car (small or standard). Multiple correspondence analysis was used to examine how the categories in this four-way table are related to each other.

1 Open the worksheet EXH_TABL.MTW.

2 Choose **Stat > Multivariate > Multiple Correspondence Analysis**.

3 Choose **Categorical variables**, and enter *CarWt DrEject AccType AccSever*.

4 In **Category names**, enter *AccNames*.

5 Click **Graphs**. Check **Display column plot**.

6 Click **OK** in each dialog box.

Session window output

Multiple Correspondence Analysis: CarWt, DrEject, AccType, AccSever

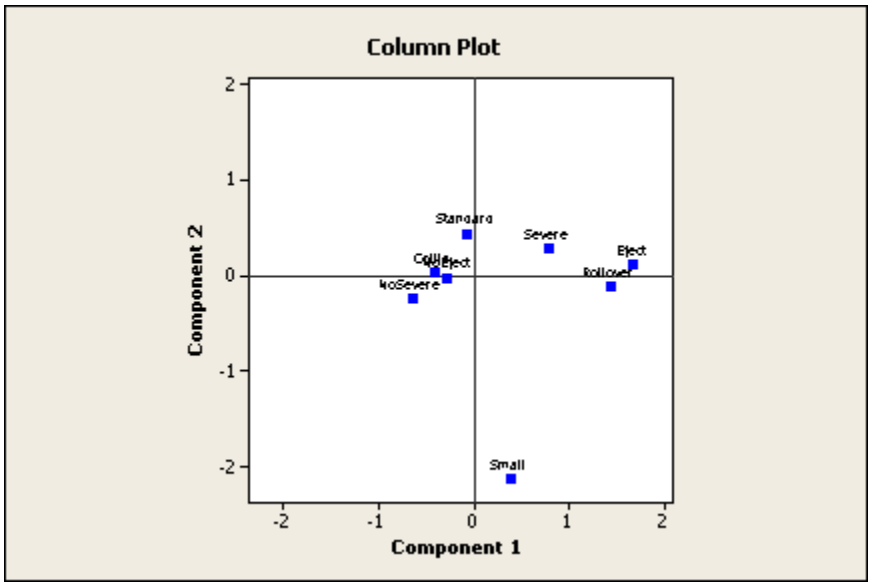
```
Analysis of Indicator Matrix

 Axis  Inertia  Proportion  Cumulative  Histogram
    1   0.4032      0.4032      0.4032  ******************************
    2   0.2520      0.2520      0.6552  ******************
    3   0.1899      0.1899      0.8451  **************
    4   0.1549      0.1549      1.0000  ***********
Total   1.0000
```


© 2003 Minitab Inc.

Column Contributions

					Component 1			Component 2	
ID	Name	Qual	Mass	Inert	Coord	Corr	Contr	Coord	Corr	Contr
1	Small	0.965	0.042	0.208	0.381	0.030	0.015	-2.139	0.936	0.771
2	Standard	0.965	0.208	0.042	-0.078	0.030	0.003	0.437	0.936	0.158
3	NoEject	0.474	0.213	0.037	-0.284	0.472	0.043	-0.020	0.002	0.000
4	Eject	0.474	0.037	0.213	1.659	0.472	0.250	0.115	0.002	0.002
5	Collis	0.613	0.193	0.057	-0.426	0.610	0.087	0.034	0.004	0.001
6	Rollover	0.613	0.057	0.193	1.429	0.610	0.291	-0.113	0.004	0.003
7	NoSevere	0.568	0.135	0.115	-0.652	0.502	0.143	-0.237	0.066	0.030
8	Severe	0.568	0.115	0.135	0.769	0.502	0.168	0.280	0.066	0.036

Graph window output



Interpreting the results

Analysis of Indicator Matrix. This table gives a summary of the decomposition of variables. The column labeled Inertia is the χ squared / n value accounted for by each component. Of the total inertia of 1, 40.3%, 25.2%, 19.0%, and, 15.5% are accounted for by the first through fourth components, respectively.

Column Contributions. Use the column contributions to interpret the different components. Since we did not specify the number of components, Minitab calculates 2 components.

- The column labeled Qual, or quality, is the proportion of the column inertia represented by the all calculated components. The car-size categories (Small, Standard) are best represented by the two component breakdown with Qual = 0.965, while the ejection categories are the least represented with Qual = 0.474. When there are only two categories for each class, each is represented equally well by any component, but this rule would not necessarily be true for more than two categories.
- The column labeled Mass is the proportion of the class in the whole data set. In this example, the CarWt, DrEject, AccType, and AccSever classes combine for a proportion of 0.25.
- The column labeled Inert is the proportion of inertia contributed by each column. The categories small cars, ejections, and collisions have the highest inertia, summing 61.4%, which indicates that these categories are more dissociated from the others.

Next, Minitab displays information for each of the two components (axes).

- The column labeled Coord gives the column coordinates. Eject and Rollover have the largest absolute coordinates for component 1 and Small has the largest absolute coordinate for component 2. The sign and relative size of the coordinates are useful in interpreting components.
- The column labeled Corr represents the contribution of the respective component to the inertia of the row. Here, Component 1 accounts for 47 to 61% of the inertia of the ejection, collision type, and accident severity categories, but explains only 3.0% of the inertia of car size.

© 2003 Minitab Inc.

- Contr, the contribution of the row to the axis inertia, shows Eject and Rollover contributing the most to Component 1 (Contr = 0.250 and 0.291, respectively). Component 2, on the other hand accounts for 93.6% of the inertia of the car size categories, with Small contributing 77.1% of the axis inertia.

Column Plot. As the contribution values for Component 1 indicate, Eject and Rollover are most distant from the origin. This component contrasts Eject and Rollover and to some extent Severe with NoSevere. Component 2 separates Small with the other categories. Two components may not adequately explain the variability of these data, however.

© 2003 Minitab Inc.

Index

© 2003 Minitab Inc.