Crosstabulation

The **Crosstabulation** procedure is designed to summarize two columns of attribute data. It constructs a two-way table showing the frequency of occurrence of all unique pairs of values in the two columns. Statistics are constructed to quantify the degree of association between the columns, and tests are run to determine whether or not there is a statistically significant dependence between the value in one column and the value in the second. The frequencies are displayed both in tabular form and graphically as a barchart, mosaic plot, or skychart.

Sample StatFolio: *crosstabulation.sgp*

Sample Data:

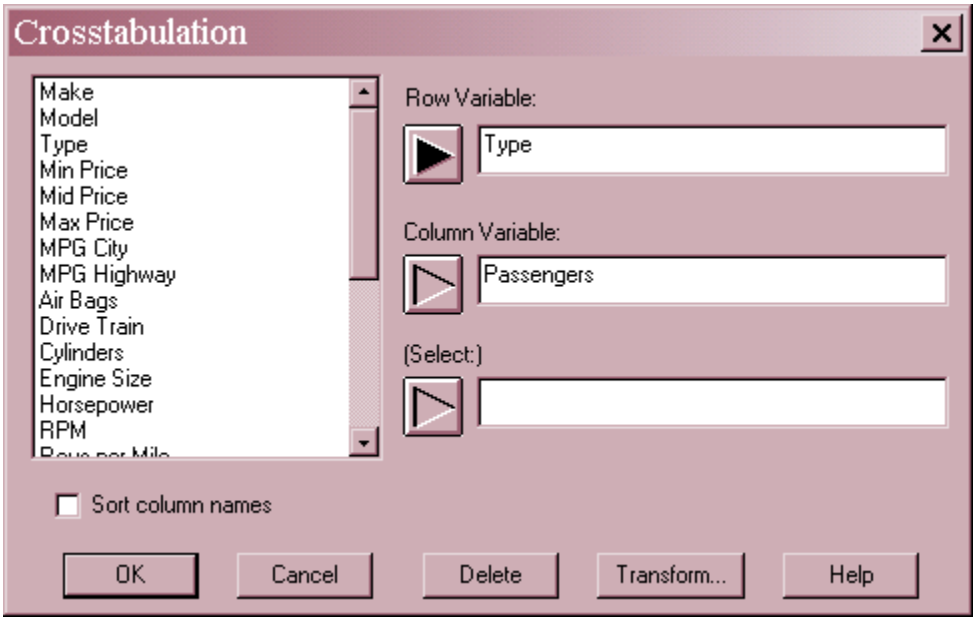
The file *93cars.sgd* contains information on 26 variables for $n = 93$ makes and models of automobiles, taken from Lock (1993). The table below shows a partial list of 4 columns from that file:

Make	*Model*	*Type*	*Passengers*
Acura	Integra	Small	5
Acura	Legend	Midsize	5
Audi	90	Compact	5
Audi	100	Midsize	6
BMW	535i	Midsize	4
Buick	Century	Midsize	6
Buick	LeSabre	Large	6
Buick	Roadmaster	Large	6
Buick	Riviera	Midsize	5
Cadillac	DeVille	Large	6
Cadillac	Seville	Midsize	5
Chevrolet	Cavalier	Compact	5

A crosstabulation will be performed between the vehicle type and the number of passengers it carries.

Data Input

The data input dialog box specifies the columns containing the data to be tabulated.



- **Row variable:** numeric or non-numeric column containing the attribute used to define the rows of the two-way table.

- **Column variable:** numeric or non-numeric column containing the attribute used to define the columns of the two-way table.

- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* shows the number of unique values in the row and column variables, as well as the number of observations (rows with non-missing data in both columns).

Crosstabulation - Type by Passengers
Row variable: Type
Column variable: Passengers

Number of observations: 93
Number of rows: 6
Number of columns: 6

Frequency Table

The *Frequency Table* shows the frequency of occurrence of each pair of values in the row and column variables, together with other information as defined on the *Pane Options* dialog box.

Frequency Table for Type by Passengers

	2	4	5	6	7	8	Row Total
Compact	0	1	13	2	0	0	16
Large	0	0	0	11	0	0	11
Midsize	0	2	15	5	0	0	22
Small	0	8	13	0	0	0	21
Sporty	2	12	0	0	0	0	14
Van	0	0	0	0	8	1	9
Column Total	2	23	41	18	8	1	93

Cell contents:
Observed frequency

The sample data consists of $r = 6$ different types of vehicles by $c = 6$ different numbers of passengers. Included in the table are:

- **Observed Frequency:** The cells in the main part of the table display O_{ij}, the number of times row i appeared together with column j in the data file.

- **Row totals**: The rightmost column of the table contains the row totals R_i:

$$R_i = \sum_{j=1}^{c} O_{ij} \tag{1}$$

- **Column totals**: The bottom row of the table contains the column totals C_j:

$$C_j = \sum_{i=1}^{r} O_{ij} \tag{2}$$

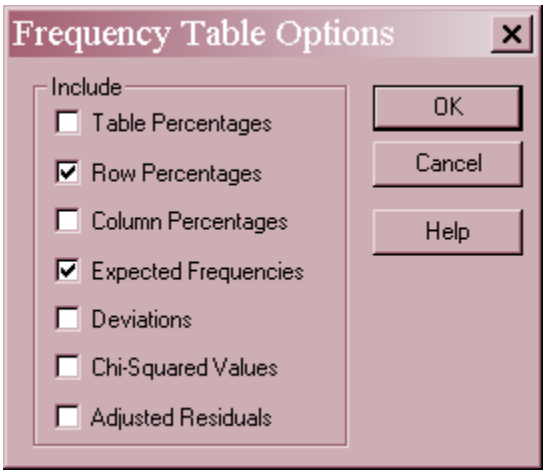
- **Table total**: the bottom right cell contains the total number of tabulated values:

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij} \tag{3}$$

For example, 14 of the 93 cars were classified as *Sporty*. Of these, 2 carry only two passengers while the other 12 carry four passengers.

Pane Options

Additional information may be added to each cell of the table using *Pane Options*:



- **Table Percentages:** the percentage each cell represents of the overall table, defined by

$$100\frac{O_{ij}}{n}\% \tag{4}$$

- **Row Percentages:** the percentage each cell represents of the total count in its row, defined by

$$100\frac{O_{ij}}{R_i}\% \tag{5}$$

- **Column Percentages:** the percentage each cell represents of the total count in its column, defined by

$$100\frac{O_{ij}}{C_j}\% \tag{6}$$

- **Expected frequency:** E_{ij}, the expected number of times row *i* would have appeared together with column *j* in the data file if the row and column classifications were independent:

$$E_{ij} = \frac{R_i C_j}{n} \tag{7}$$

- **Deviations**: the differences between the observed and expected frequencies:

$$O_{ij} - E_{ij} \tag{8}$$

- **Chi-Squared Values**: the contribution of each cell to a chi-squared statistic, used to test for independence between row and column classifications:

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (9)$$

- **Adjusted Residuals:** a form of standardized residuals computed by dividing each cell deviation by an estimate of its standard error:

$$\varepsilon_{ij} = \frac{(O_{ij} - E_{ij})}{\sqrt{E_{ij}\frac{(1-R_i)}{n}\frac{(1-C_j)}{n}}} \qquad (10)$$

Example – Additional Information on Sporty Cars

Frequency Table for Type by Passengers

	2	4	5	6	7	8	Row Total
Sporty	2	12	0	0	0	0	14
	2.15%	12.90%	0.00%	0.00%	0.00%	0.00%	15.05%
	14.29%	85.71%	0.00%	0.00%	0.00%	0.00%	
	100.00%	52.17%	0.00%	0.00%	0.00%	0.00%	
	0.30	3.46	6.17	2.71	1.20	0.15	
	1.70	8.54	-6.17	-2.71	-1.20	-0.15	
	9.59	21.05	6.17	2.71	1.20	0.15	
	3.40	5.74	-3.60	-1.99	-1.25	-0.42	
Column Total	2	23	41	18	8	1	93
	2.15%	24.73%	44.09%	19.35%	8.60%	1.08%	100.00%

Cell contents:
Observed frequency
Percentage of table
Percentage of row
Percentage of column
Expected frequency
Observed - expected frequency
Contribution to chi-squared
Adjusted residual

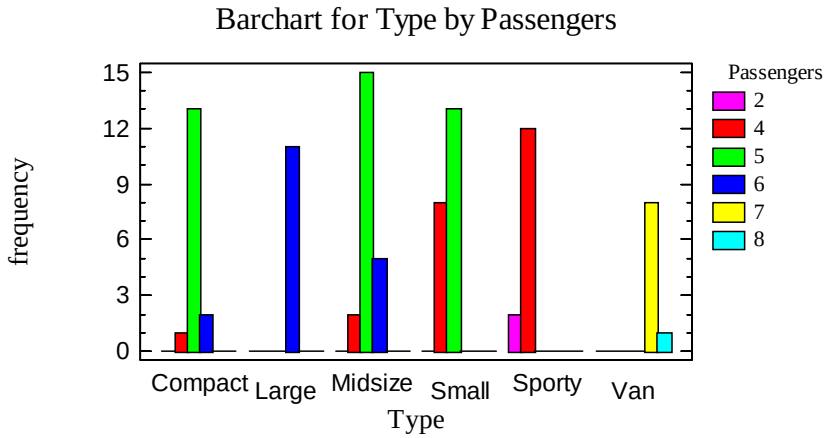
The 12 *Sporty* cars in the sample data file that carry 4 passengers represent:

12.90% of the total of n = 93 cars
85.71% of all 14 *Sporty* cars
52.17% of all 23 cars that carry 4 passengers

Were row and column classifications independent, the expected number of cars that should be both *Sporty* and carry four passengers is only 3.46, for a deviation of 8.54. In the computation of the Chi-squared test statistic, described below, this cell adds a total of 21.05 to that statistic. The adjusted residual indicates that the observed number of cars in this cell is 5.74 standard deviations above its expected value.

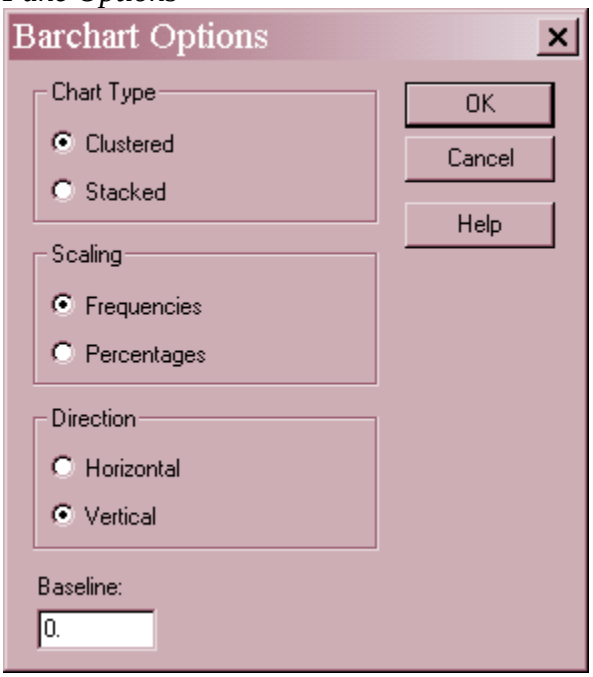
Barchart

A common way to display the data in a two-way table is by using a multiple barchart.



The height of each bar in the above chart represents the number of cars of each type that carried each number of passengers.

Pane Options



- **Chart Type:** The bars may be clustered side by side as shown in the example or stacked one upon the other.

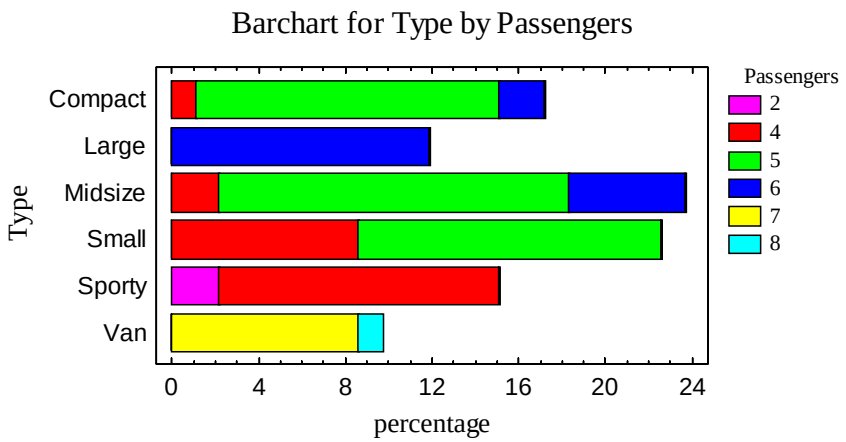
- **Scaling**: whether the axis scale shows the frequencies O_{ij} or the percentages given by

$$p_{ij} = 100\frac{O_{ij}}{n}\% \qquad (11)$$

- **Direction**: whether the bars extend horizontally or vertically.

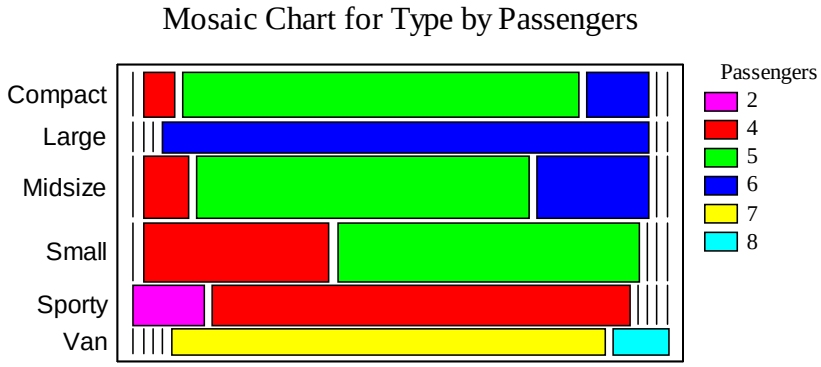
- **Baseline**: the value from which the bars extend.

Example – Stacked Horizontal Barchart Scaled by Percentage



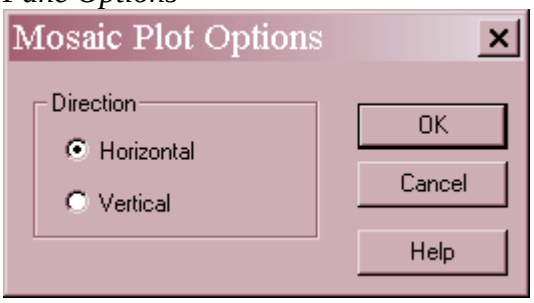
Mosaic Chart

An interesting variation of the Barchart occurs if the both the length and width of each bar are scaled to represent the frequency of the corresponding cell in the two-way table.



In this chart, the height of each row is proportional to its row total R_i. The width of the bars within each row is proportional to the frequency of each cell *within that row*. This results in bars whose *area* is proportional to the frequency in a particular cell. In the sample data, the largest bar corresponds to *Midsize* automobiles that carry 5 passengers.

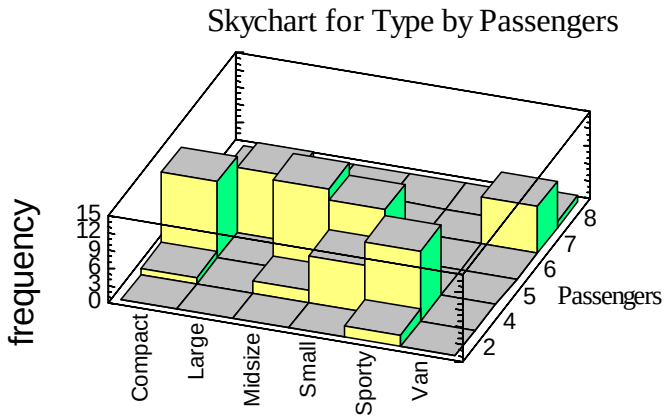
Pane Options



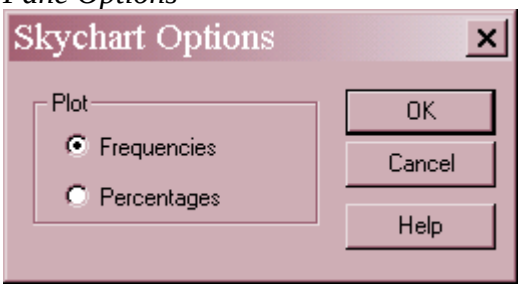
- **Direction:** the orientation of the bars.

SkyChart

The cell frequencies can also be represented using vertical bars.



Pane Options



- **Plot:** scaling for the vertical axis.

Tests of Independence

A common question asked of the data in two-way tables is whether or not the row and column classifications are independent. If rows and columns are independent, then the fact that an item falls in a particular row does not affect the probability of its falling in a given column. In the current example, independence would imply that the type of vehicle had no relationship to the number of passengers it carries.

The most common test for independence in a two-way table is the chi-squared test. This test compares the observed and expected frequencies by calculating:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (12)$$

STATGRAPHICS displays the results of this test and its corresponding P-value:

Tests of Independence

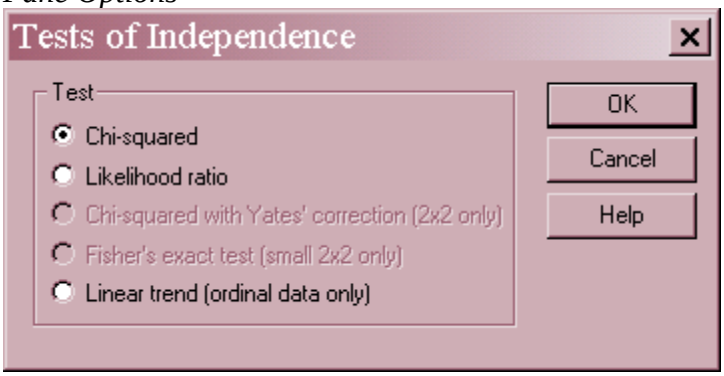
Test	*Statistic*	*Df*	*P-Value*
Chi-Squared	197.595	25	0.0000

Warning: some cell counts < 5.

The P-value is calculated by comparing the test statistic to a chi-squared distribution with *(r-1)(c-1)* degrees of freedom. Small P-values (less than 0.05 if operating at the 5% significance level) indicate a significant dependence between the row and column classifications. The P-Value in the above table clearly shows that the type of automobile and number of passengers it carries are not independent.

If the expected value E_{ij} in any cell of the table is less than 5, a warning will be displayed. In such cases, the calculated Chi-squared statistic may not be well represented by a chi-squared distribution. This is particularly serious if any expected values are less than 2, which is the case in the current example. When this occurs, consideration should be given to combining classes that do not contain much data, such as the 7 and 8 passenger automobiles.

Pane Options



- **Test** – the type of test to be performed.

Rather than performing the Chi-squared test, alternative tests can be run. Details about these tests are contained in the documentation for the *Contingency Tables* procedure.

Summary Statistics

Various statistics can also be computed to measure the degree of association between the rows and columns in the table.

Summary Statistics

		With Rows	*With Columns*
Statistic	*Symmetric*	*Dependent*	*Dependent*
Lambda	0.4715	0.3803	0.5962
Uncertainty Coeff.	0.5303	0.4730	0.6034
Somer's D	-0.2022	-0.2193	-0.1876
Eta		0.7767	0.8810

Statistic	*Value*	*P-Value*	*Df*
Contingency Coeff.	0.8246		
Cramer's V	0.6519		
Conditional Gamma	-0.2428		
Pearson's R	-0.0766	0.4657	91
Kendall's Tau b	-0.2028	0.0174	
Kendall's Tau c	-0.1840		

As an example, Cramer's V is a statistic that ranges between 0 and 1, calculated from the chi-squared test statistic. The stronger the association between rows and columns, the closer this statistic will be to 1.

Details about each statistic are contained in the documentation for the *Contingency Tables* procedure.

Odds Ratios

The *Odds Ratios* pane provides special information about cases where there are exactly 2 rows and 2 columns. For an example of their use, see the documentation for the *Contingency Tables* procedure.

Save Results

The following results may be saved to the datasheet:

1. *Cell Frequencies (single column)* – the cell frequencies O_{ij} in a single column, one row after another.

2. *Row Labels* – the identifiers for each row of the two-way table.

3. *Column Labels* – the identifiers for each column of the two-way table.

4. *Cell Frequencies (matrix)* - the cell frequencies O_{ij} in multiple columns, paralleling the layout of the two-way table.

