Frequency Tables

The **Frequency Tables** procedure analyzes a single column containing counts. It displays the counts using either a barchart or piechart. Statistical tests may also be performed to determine whether the data conform to a set of multinomial probabilities.

Sample StatFolio: *freqtable.sgp*

Sample Data:

The file *mendel.sgd* contains the results of an experiment performed to test Mendel's theory of natural inheritance (see the discussion in Agresti, 2002). Mendel crossed pea plants of pure green and yellow strains. According to his theory, 75% of all second-generation plants should show the dominant yellow color. The file shows the observed strain of $n = 8023$ seeds:

Color	*Theory*	*Seeds*
Yellow	0.75	6022
Green	0.25	2001

The *Frequency Table* procedure will be used to determine whether the observed seeds are in conformance with that expected from the theory.

Data Input

The data input dialog box specifies the column containing the counts, optional labels, and optional expected values:
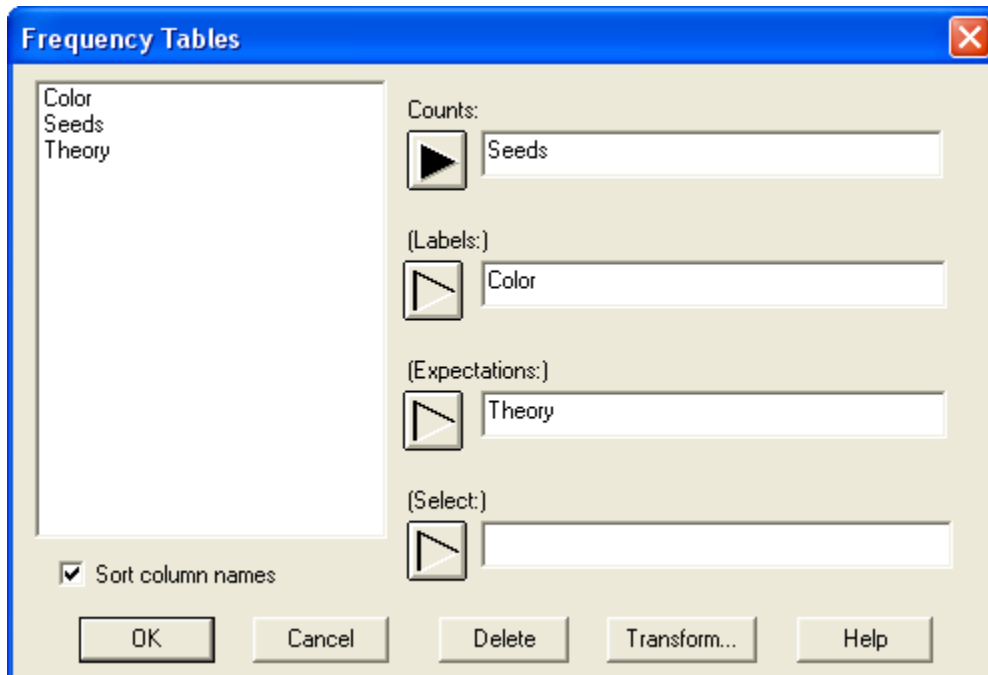
- **Counts:** numeric column containing *m* counts.

- **Labels:** column containing identifiers for each row of the table.

- **Expectations:** numeric column from which the expected distribution of the counts can be determined. The column may contain either probabilities π_j or expected frequencies. The values in the column will be adjusted if necessary to sum to 1. NOTE: if expectations are not specified, equal probabilities will be assumed for all rows.

- **Select:** subset selection.

Analysis Summary

The *Analysis Summary* shows the number of rows with non-missing data and the sum of the data in the *Counts* column.

Frequency Tables - Seeds
Counts: Seeds
Sum of 2 values: 8023.0

Frequency Table

This pane presents the data in tabular form.

Frequency Table for Seeds

Class	Value	Frequency	Relative Frequency	Cumulative Frequency	Cum. Rel. Frequency
1	Yellow	6022	0.7506	6022	0.7506
2	Green	2001	0.2494	8023	1.0000

This table displays the following information for each row of the file:

- **Value** - the label identifying the row.

- **Frequency** - the frequency f_j obtained from the *Counts* column.

- **Relative Frequency** - the proportion of times each value occurred, given by f_j/n.

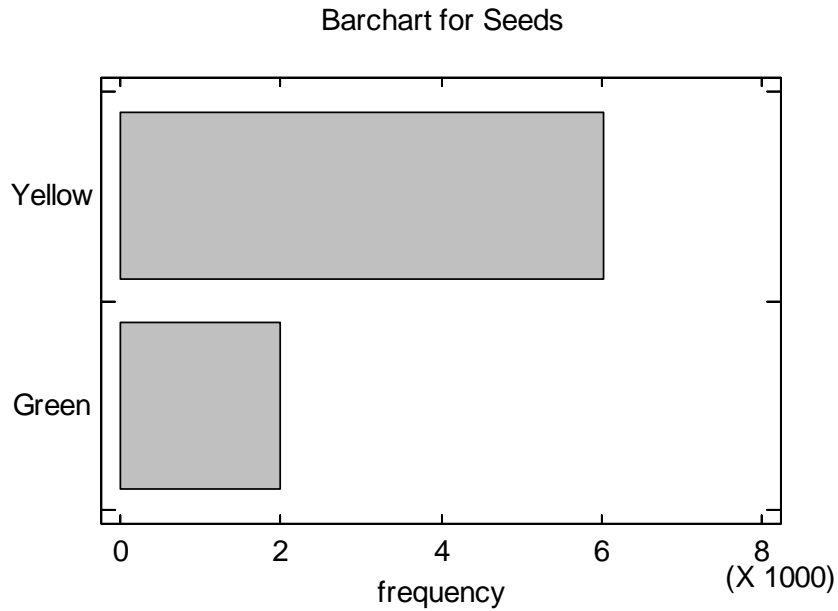- **Cumulative Frequency** - the sum of the counts in the current and previous rows:

$$\sum_{i=1}^{j} f_i \qquad (1)$$

- **Cumulative Relative Frequency** - the proportion of counts in the current and previous rows:
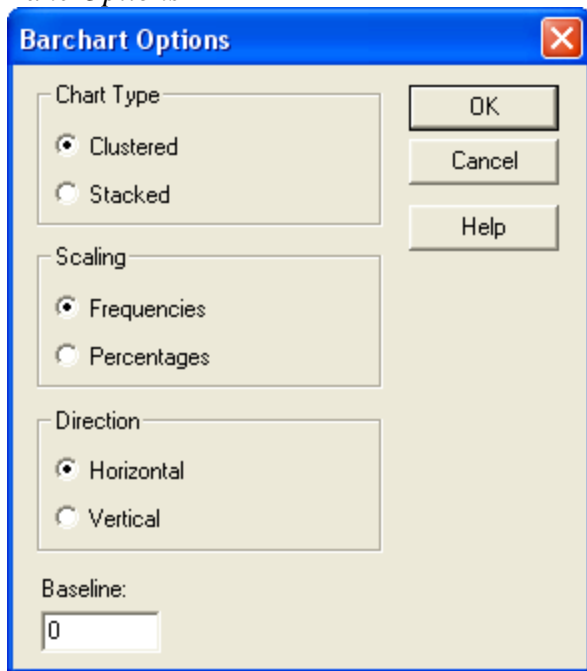
$$\frac{\sum_{i=1}^{j} f_i}{n} \qquad (2)$$

Barchart

The *Barchart* plots each count as a vertical or horizontal bar.

Barchart for Seeds

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 f_j or the percentages given by
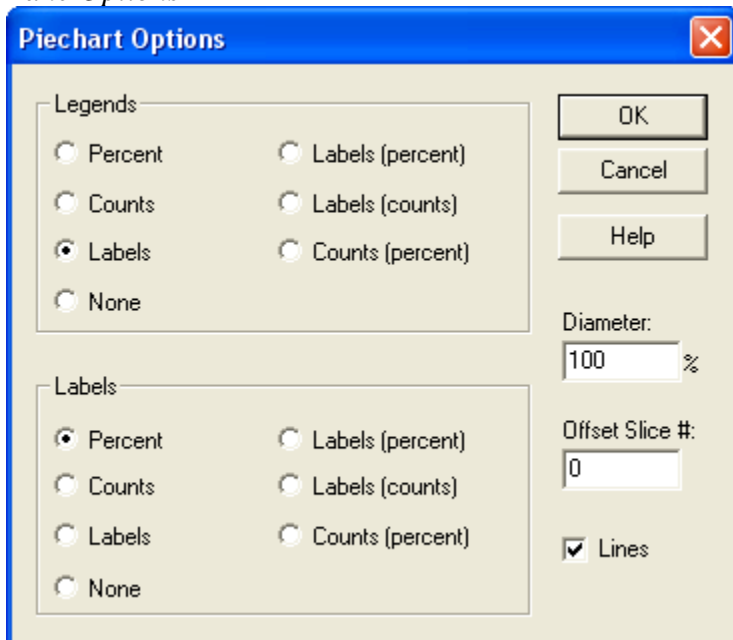
$$p_j = 100\frac{f_j}{n}\%$$ (3)

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

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

Piechart

The *Piechart* plots each unique value using slices of a pie to represent their relative magnitude.

Piechart for Seeds
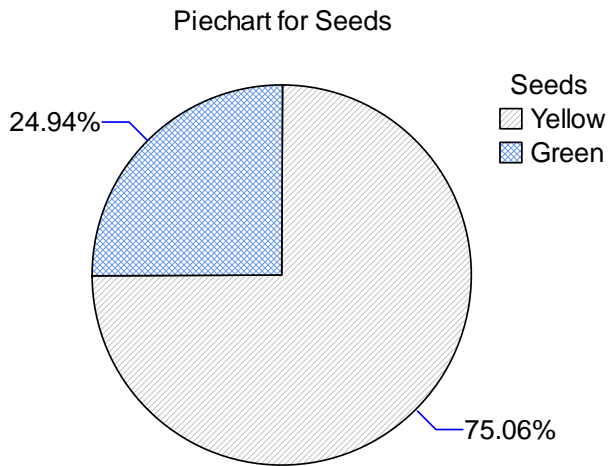
Pane Options

- **Legends:** the desired information to be placed in the legend block to the right of the pie.

- **Labels**: the desired information to be placed next to each slice of the pie.

- **Diameter**: the size of the pie relative to the size of the plotting area. Reducing this value makes the pie smaller.

- **Offset #**: number of a slice between 1 and k to offset from the rest of the pie.

- **Lines**: whether to extend lines from each slice to its label.

Example: Piechart with Offset Slice

Piechart for Seeds

24.94%

Seeds
☐ Yellow
☒ Green

75.06%
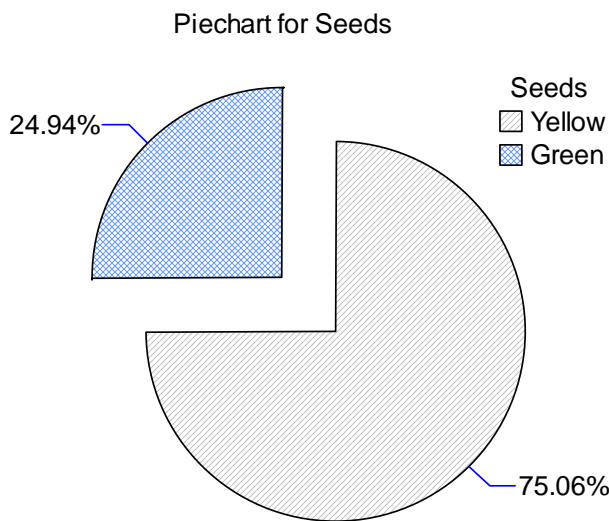
Goodness-Of-Fit Tests

The distribution of the counts can be compared to a set of specified class probabilities π_1, π_2, …, π_m to determine whether they could reasonably have come from a multinomial distribution with those probabilities. *Pane Options* allows the user to select from three tests:

Goodness-of-Fit Tests

Test
⦿ Chi-squared
○ Likelihood ratio
○ Exact (2 categories only)

OK
Cancel
Help
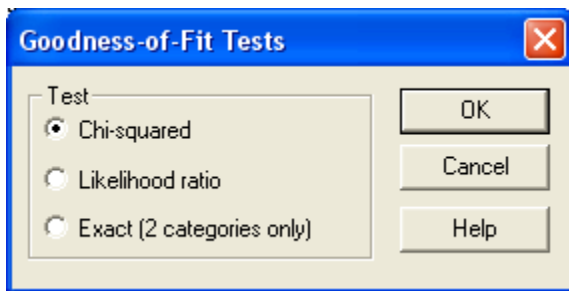
1. *Chi-squared test*: compares the observed frequencies in each row with the expected values using a statistic calculated from:

$$X^2 = \sum_{i=1}^{m} \frac{(f_i - n\pi_i)^2}{n\pi_i} \tag{4}$$

This statistic is compared to a chi-squared distribution with *m-1* degrees of freedom.

2. *Likelihood ratio test*: compares the observed frequencies in each row with the expected values using a statistic calculated from:

$$G^2 = 2\sum_{i=1}^{m} f_i \log\left(\frac{f_i}{n\pi_i}\right) \tag{5}$$

This statistic is also compared to a chi-squared distribution with *m-1* degrees of freedom.

3. *Exact*: for the case of 2 rows only, calculates the exact probability of obtaining data as extreme or more extreme than that observed, calculated from the cumulative binomial distribution.

The table below shows the results of the chi-squared test:

Goodness-of-Fit Tests for Seeds

Class	Value	Observed Frequency	Expected Frequency	Observed Proportion	Expected Proportion	Contribution to Test Statistic
1	Yellow	6022	6017.25	0.750592	0.75	0.00374964
2	Green	2001	2005.75	0.249408	0.25	0.0112489

Chi-Squared = 0.0149985 with 1 d.f.   P-Value = 0.9025

A small P-Value would indicate that the observed data did not conform to the specified probabilities. Since the P-Value in the above table is well above 0.10, there is no discrepancy between the data and Mendel's theory at the 10% or higher significance level.
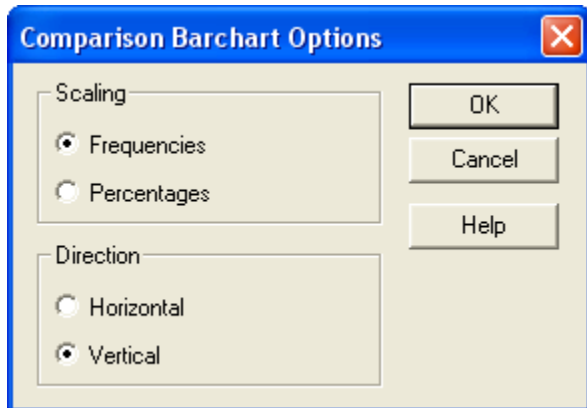
Comparison Barchart

The comparison between observed and expected counts may also be displayed using a barchart:

Barchart for Seeds
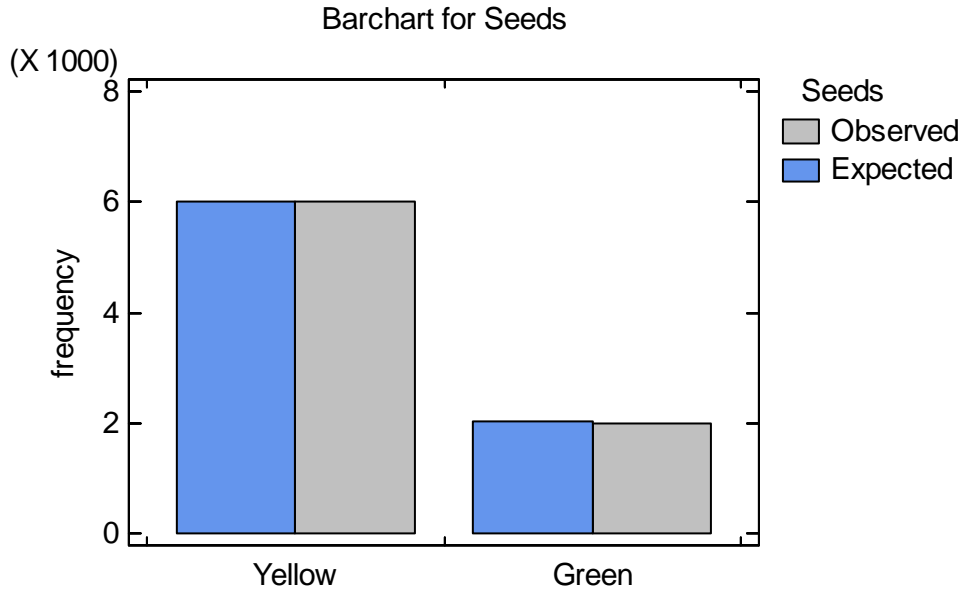
In this case, the observed and expected counts are almost identical.

Pane Options

The direction and scaling of the barchart can be specified using *Pane Options*:

- **Scaling**: whether the axis scale shows the observed and expected frequencies or the observed and expected percentages in each row.

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