Relate – Attributes and Counts

This procedure is designed to summarize data that classifies observations according to two categorical factors. The data may consist of either:

1. Two *Attribute* variables. In such cases, the procedure will identify all unique values in both variables and calculate

   $c_{i,j}$ = frequency of occurrence of category *i* for first variable and category *j* for second variable, $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$

   It will then construct an *r* by *c* contingency table containing the frequencies. The leftmost variable selected defines the rows, while the other variable defines the columns.

2. *c Count* columns, each containing *r* counts. These counts are used to create the entries in the table directly. In this case, the column names will be used as labels for each column of the two-way table. A *Labels* column may also be selected to provide labels for each row of the table.

The charts created are:

- **Clustered Barchart –** plots the frequencies as a clustered barchart, with one bar for each cell in the table.

- **Stacked Barchart –** plots the frequencies as a stacked barchart, with one bar for each column in the table.

- **Percentage Barchart –** plots the frequencies as a stacked barchart, subdividing each bar according to the percentages represented by each column.

- **Mosaic Plot –** plots the frequencies in each cell in a manner that makes the area of each bar proportional to the cell frequency.

Access

Highlight: two or more *Count* columns or two *Attributes* column. If *Count* columns are selected, a *Labels* column may also be selected to supply labels for the row categories.

Select: *Relate* from the main menu.

Output Page 1: A clustered barchart.

Output Page 2: A stacked barchart.

Output Page 3: A percentage barchart.

Output Page 4: A mosaic plot.

Sample Data #1: Untabulated Data

The file *opinion.sgm* contains information about 50 individuals who were each asked to sample a product and rate it as Excellent, Good, Fair or Poor. The gender of each individual was also recorded. A portion of the data is shown below:

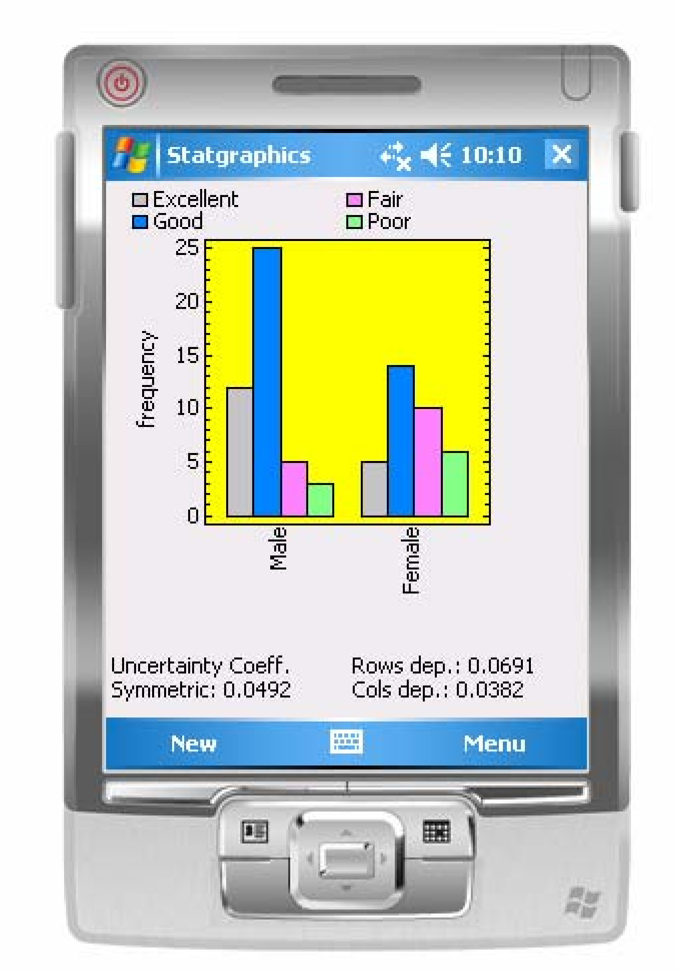
Row	*Gender*	*Rating*
1	Male	Excellent
2	Female	Good
3	Male	Fair
4	Male	Good
5	Female	Poor
6	Female	Good
7	Female	Excellent
8	Female	Fair

IMPORTANT NOTE:
If an *Attribute* variable is *Character*, its categories will be ordered according to their first occurrence in the data column. If order is important for a variable, be certain that the values appear in the desired order. Note that in the above table, the values for *Rating* appear in the order *Excellent*, *Good*, *Fair* and then *Poor*, which is the desired order. If an *Attribute* variable is numeric, the numeric values will be used to determine the order.

Clustered Barchart

If two *Attribute* columns are selected, the procedure begins by finding all unique values in both columns. It then calculates the frequency of occurrence of each pair of values. If there are *r* unique values in the first column and *c* unique values in the second column, the frequencies $c_{i,j}$ can be thought of as forming an *r* by *c* contingency table.

A clustered barchart plots the frequencies, grouping together all values for each value in the first variable.



```
Uncertainty Coeff.        Rows dep.: 0.0691
Symmetric: 0.0492         Cols dep.: 0.0382
```

At the bottom of the display are values of the uncertainty coefficient. This coefficient measures the proportional reduction in *entropy* (variance) of one classification variable if the value of the other is known. It ranges from a low of 0 (complete independence between rows and columns) to 1 (no conditional variation). Three values are calculated:

1. *Rows dep.:* calculated on the assumption that the first classification variable (*Gender*) is dependent on the second (*Rating*).

2. *Cols dep.*: calculated on the assumption that the second classification variable (*Rating*) is dependent on the first (*Gender*).

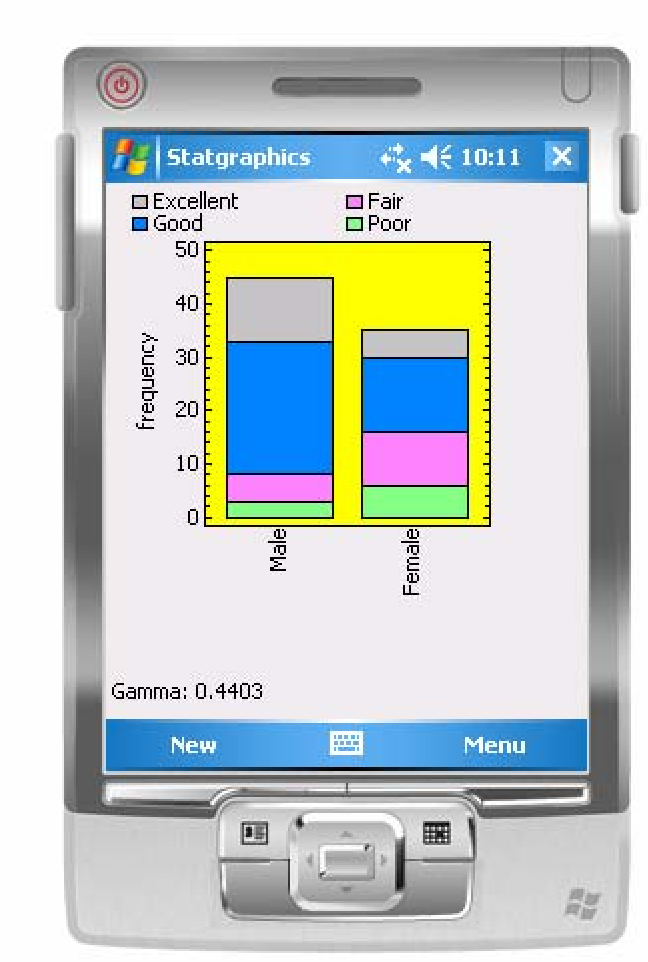
3. *Symmetric*: a combination of the two measures above.

In the current example, the *Cols dep.* coefficient would be most appropriate, since *Gender* might have an effect on *Rating*. The reduction in entropy, however, is only about 4%.

The uncertainty coefficient is typically used to provide a measure of association when the variables are nominal (have no natural ordering).

Stacked Barchart

The stacked barchart places the bars on top of each other, rather than in clusters.

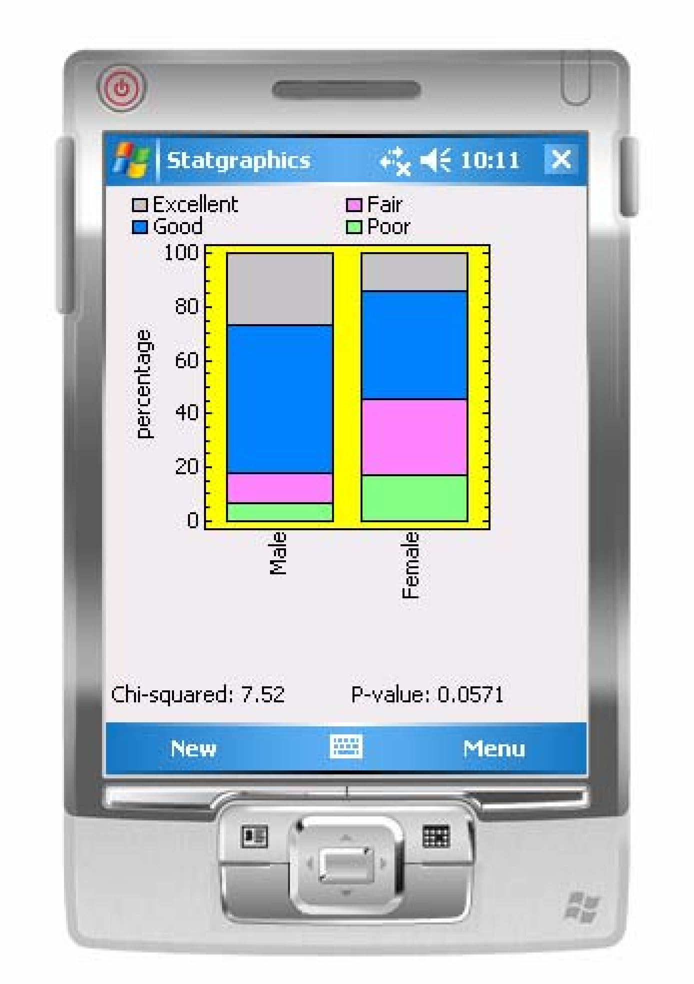


Gamma: 0.4403

The output also displays the value of a statistic called *Gamma*. Gamma is appropriate when both classification variables are ordinal (have a natural ordering). It is based on calculating the number of *discordant* and *concordant* pairs, where a pair of observations are concordant if a member of the pair has a higher rating on both classification variables and discordant if one member is higher on one variable and lower on the other. Gamma ranges between -1 (all pairs discordant) to +1 (all pairs concordant). A value close to 0 implies independence between the two classification factors.

Percentage Barchart

The percentage barchart is similar to a stacked barchart. However, the vertical axis runs from 0% to 100%. The height of the bars within each column represents the conditional distribution of the second variable within the first.



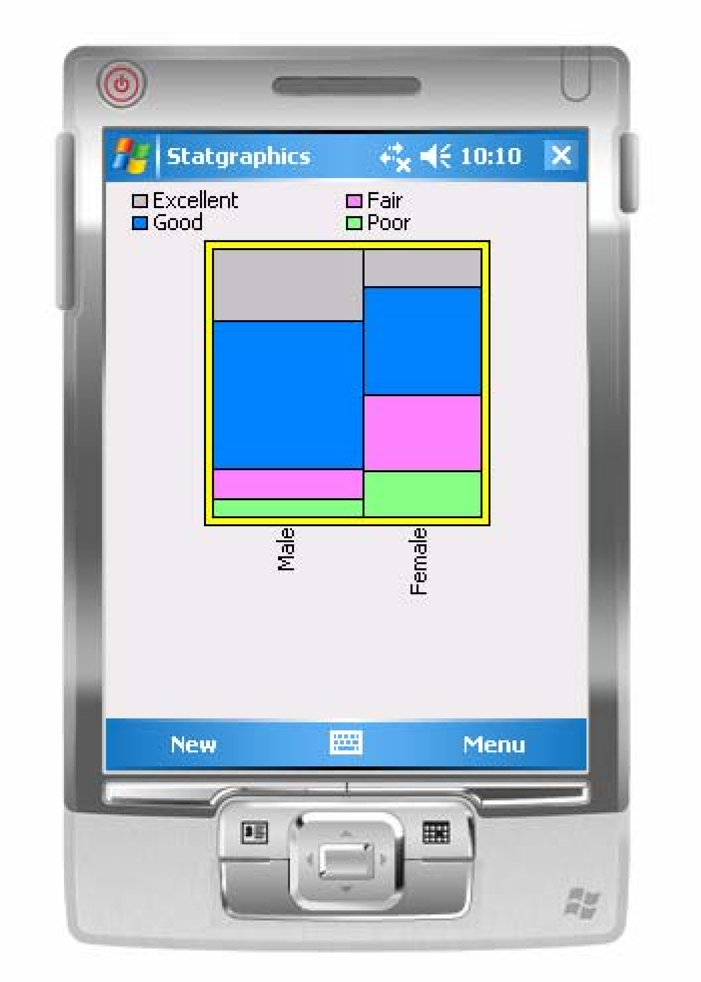
Chi-squared: 7.52 P-value: 0.0571

The results of a chi-squared test for independence between rows and columns is also shown. A small P-value implies that the two classification variables are not independent.

NOTE: The P-value may not be reliable for sparse data. i.e., data with small expected counts in many cells.

Mosaic Plot

The mosaic plot is similar to a percentage barchart. However, the width of each column of bars is scaled according to the relative percentage of data in each category of the first variable. The result is a display in which the area of each bar is proportional to the frequencies $c_{i,j}$.



Since there are more males than females in the sample data, the first bar is wider than the second.

Sample Data #2: Tabulated Data

When the data to be analyzed has already been tabulated, the two-way table of counts can be placed into adjacent columns of the datasheet. For example, the file *table.sgm* contains the same data as analyzed above, except it is placed in four *Count* variables.

Row	*Gender*	*Excellent*	*Good*	*Fair*	*Poor*
1	Male	12	25	5	3
2	Female	5	14	10	6

Select the columns containing the counts, and a *Labels* column containing the row labels. The select *Relate* from the main menu. The same output will be generated as in the first sample above.

Calculations

r = number of categories in first variable (rows)

c = number of categories in second variable (columns)

$c_{i,j}$ = frequency of occurrence of category *i* for first variable and category *j* for second variable, *i* = 1, 2, …, *r* and *j* = 1, 2, …, *c*

Row Totals

$$R_i = \sum_{j=1}^{c} c_{ij} \qquad (1)$$

Column Totals

$$C_j = \sum_{i=1}^{r} c_{ij} \qquad (2)$$

Total Count

$$n = \sum_{i=1}^{r} \sum_{j=1}^{c} c_{ij} \qquad (3)$$

Uncertainty Coefficient

Rows dependent: $U = \dfrac{U(X)+U(Y)-U(XY)}{U(X)}$ (4)

Columns dependent: $U = \dfrac{U(X)+U(Y)-U(XY)}{U(Y)}$ (5)

Symmetric: $U = 2\left[\dfrac{U(X)+U(Y)-U(XY)}{U(X)+U(Y)}\right]$ (6)

where

$$U(X) = -\sum_{i=1}^{r} \frac{R_i}{n} \log\left(\frac{R_i}{n}\right) \tag{7}$$

$$U(Y) = -\sum_{j=1}^{c} \frac{C_j}{n} \log\left(\frac{C_j}{n}\right) \tag{8}$$

$$U(XY) = -\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{c_{ij}}{n} \log\left(\frac{c_{ij}}{n}\right) \quad \text{for } c_{ij} > 0 \tag{9}$$

Gamma

$$\gamma = \frac{(P-Q)}{P+Q} \tag{10}$$

where P = number of concordant pairs, Q = number of discordant pairs

Chi-Squared

$$\mathrm{X}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(c_{ij}-e_{ij}\right)^2}{e_{ij}} \tag{11}$$

where

$$e_{ij} = \frac{R_i C_j}{n} \tag{12}$$

Degrees of freedom: $v = (r-1)(c-1)$ (13)