Contingency Tests

This menu item will allow tests based upon chi-square probabilities. These tests use principles of comparing observed frequencies, ie the experimental results, against expected frequencies. The latter are obtained using an hypothesis relevant to the task at hand. It usually relates to a distribution of some sort for that data but may be any form of coherent postulate in the Goodness of Fit subtest (where you supply the values that you expected to obtain).

The data should be entered into the spreadsheet as the actual counts (called frequencies), not as percentages or proportions. The calculations use these integers to calculate the appropriate comparisons. You should enter the data in the format of a table, either 2 or more columns depending upon the analysis required. You should then select the cells in the spreadsheet containing all the table values by clicking and dragging, though this is not essential, since this will automatically enter these row and column numbers in the dialog box when you select the Contingency option. This procedure will therefore save you time.

Contingency table

This option is selected by clicking the Contingency table button. This requires at least 2 rows and 2 columns of observed frequencies to do the comparison. The matrix can be much larger however (R rows and C columns). This test computes the expected frequencies in a similar matrix and then compares these with the observed frequencies. It is a test of the independence of the factors used to form the table based on the frequencies of occurrence in each of the cells, whose individual contributing counts are assumed to be independent.

Notation:

R x C table with R rows and C columns
N the total number of individuals
Oij the observed frequency in row i, column j
Eij the expected frequency in row i, column j
Ho : the row and column factors are independent.
H_1 : Ho is not true.

If Ho is true, the expected frequencies are given by

Eij = (row i total) x (column j total) / N

The chi-sq test statistic is given by

$X2 = \sum (O - E)^2/E$ with the sum taken over all (R x C) cells

For a 2 x 2 table a Yates' continuity correction is used in which (O-E) is replaced by (|O-E| - 1/2) in the above formula.

The value of the test statistic X2 is then compared with the chi-square tables at (R-1)(C-1) degrees of freedom. The larger X2 is, the less likely it is that Ho is true.

There are limitations which must be borne in mind when using this Contingency table test:

1. For a 2 x 2 table, either E must be > 5 in each cell, or if $N \geq 40$, only one of the E's is allowed to be as low as 1. If these conditions are not satisfied the chi-square test is invalid, and Fisher's Exact test must be used. These conditions are in fact detected by SchoolStat™ and if found, Fisher's Exact test is automatically applied and this is documented in the Results.

2. For larger tables, all E must be > 1 and 80% of the E values must be > 5. These conditions are also checked, and if found, you are informed that data catagories (rows) must be combined for valid comparisons. This will be left to you.

Chi-square Goodness of Fit test

This is enabled by clicking the Goodness of Fit button. You should have already entered your observed and expected frequencies in 2 columns in the spreadsheet. Make sure that there are exactly the same number of data in each. You can only have 2 columns, but as many rows as you require.

This test will measure how closely given observed frequencies agree with the expected frequencies calculated according to your hypothesis or model (eg. gene frequencies, success in dice throwing or specific card games, etc). Recall that you must enter the actual numbers and not proportions or percentages. The results detail the one- and two-sided probabilities for Ho.

Specifically:

This test is quite specific in its requirements:

1. Data needs to have been categorized, ie split into groups qualitatively or quantitatively. Each must be mutually exclusive AND exhaustive (ie each unit counted in each category must only have been counted ONCE and ONCE ONLY).

2. The expected proportion of all cases falling into each category must be calculated by reference to some hypothesis, then an integer frequency is obtained by multiplying this proportion by the total sample size N to produce the set of expected frequencies.

3. There must be actual numbers occurring in each category (the proportions alone will not work correctly).

4. You must decide your degrees of freedom. These are actually v = k - c - 1, where k is the number of classes or categories (ie rows), and c is the number of parameters or entities obtained or calculated directly from the observed frequencies which are used to obtain the expected frequencies. (This is quite tricky at times and is best illustrated by an example: to test whether a set of grouped data could have come from a normal population, the expected frequencies are obtained by multiplying N (the total sample size), by the probability that an observation falls in a particular group with probabilities obtained from a normal distribution with mean xbar and standard deviation s, the mean and standard deviation of the observed data. Here the observed and expected frequencies have three entities in common (N, xbar and s), hence c would be 3.)

The formula used is:

X2 = ∑ (O - E)^2/E with the sum taken over the k categories.

The result is then compared with the chi-square table at v degrees of freedom. (When v = 1 Yates' correction is used and this is noted in the Results). There are limitations to be understood when using this test:

1. All expected values must be > 1; and
2. 80% of expected values must be > 5, otherwise categories must be combined to give adequate frequencies.

If the first condition is found, then SchoolStat™ will still perform the calculation but with a warning that the results are probably invalid. If Yates' continuity correction (see above) is needed, then this will also be noted in the results.