Two-Way Tables and the Chi-Square Test

When analysis of categorical data is concerned with more than one variable, two-way tables (also known as *contingency tables*) are employed. These tables provide a foundation for statistical inference, where statistical tests question the relationship between the variables on the basis of the data observed.

Example

In the dataset "Popular Kids," students in grades 4-6 were asked whether good grades, athletic ability, or popularity was most important to them. A two-way table separating the students by grade and by choice of most important factor is shown below:

	Grade			
Goals	4	5	6	Total
Grades	49	50	69	168
Popular	24	36	38	98
Sports	19	22	28	69
Total	92	108	135	335

To investigate possible differences among the students' choices by grade, it is useful to compute the column percentages for each choice, as follows:

	Grade		
Goals	4	5	6
Grades	53	46	51
Popular	26	33	28
Sports	21	20	21
Total	100	100	100

There is error in the second column (the percentages sum to 99, not 100) due to rounding. From the appearance of the column percentages, it does not appear that there is much of a variation in preference across the three grades.

Data source: Chase, M.A and Dummer, G.M. (1992), "The Role of Sports as a Social Determinant for Children," Research Quarterly for Exercise and Sport, 63, 418-424. Dataset available through the Statlib Data and Story Library (DASL).

The ***chi-square test*** provides a method for testing the association between the row and column variables in a two-way table. The null hypothesis H_0 assumes that there is no association between the variables (in other words, one variable does not vary according to the other variable), while the alternative hypothesis H_a claims that some association does exist. The alternative hypothesis does not specify the *type* of association, so close attention to the data is required to interpret the information provided by the test.

The chi-square test is based on a test statistic that measures the divergence of the observed data from the values that would be *expected* under the null hypothesis of no association. This requires calculation of the expected values based on the data. The expected value for each cell in a two-way table is equal to *(row total*column total)/n*, where *n* is the total number of observations included in the table.

Example

Continuing from the above example with the two-way table for students choice of grades, athletic ability, or popularity by grade, the expected values are calculated as shown below:

Original Table

	Grade			
Goals	4	5	6	Total
Grades	49	50	69	168
Popular	24	36	38	98
Sports	19	22	28	69
Total	92	108	135	335

Expected Values

	Grade		
Goals	4	5	6
Grades	46.1	54.2	67.7
Popular	26.9	31.6	39.5
Sports	18.9	22.2	27.8

The first cell in the expected values table, Grade 4 with "grades" chosen to be most important, is calculated to be 168*92/335 = 46.1, for example.

Once the expected values have been computed (done automatically in most software packages), the ***chi-square test statistic*** is computed as

$$X^2 = \sum \frac{(\text{observed - expected})^2}{\text{expected}}$$

where the square of the differences between the observed and expected values in each cell, divided by the expected value, are added across all of the cells in the table.

The distribution of the statistic X^2 is ***chi-square*** with (r-1)(c-1) degrees of freedom, where r represents the number of rows in the two-way table and c represents the number of columns. The distribution is denoted χ^2(df), where df is the number of degrees of freedom.

The chi-square distribution is defined for all positive values. The *P-value* for the chi-square test is *P(*χ^2 *≥*X^2*)*, the probability of observing a value at least as extreme as the test statistic for a chi-square distribution with (r-1)(c-1) degrees of freedom.

Example

The chi-square statistic for the above example is computed as follows:
X^2 = (49 - 46.1)²/46.1 + (50 - 54.2)²/54.2 + (69 - 67.7)²/67.7 + ... + (28 - 27.8)²/27.8
= 0.18 + 0.33 + 0.03 + ... + 0.01
= 1.51
The degrees of freedom are equal to (3-1)(3-1) = 2*2 = 4, so we are interested in the probability *P(*χ^2 *≥ 1.51)* = 0.8244 on 4 degrees of freedom. This indicates that there is no association between the choice of most important factor and the grade of the student -- the difference between observed and expected values under the null hypothesis is negligible.

Example

The "Popular Kids" dataset also divided the students' responses into "Urban," "Suburban," and "Rural" school areas. Is there an association between the type of school area and the students' choice of good grades, athletic ability, or popularity as most important?

A two-way table for student goals and school area appears as follows:

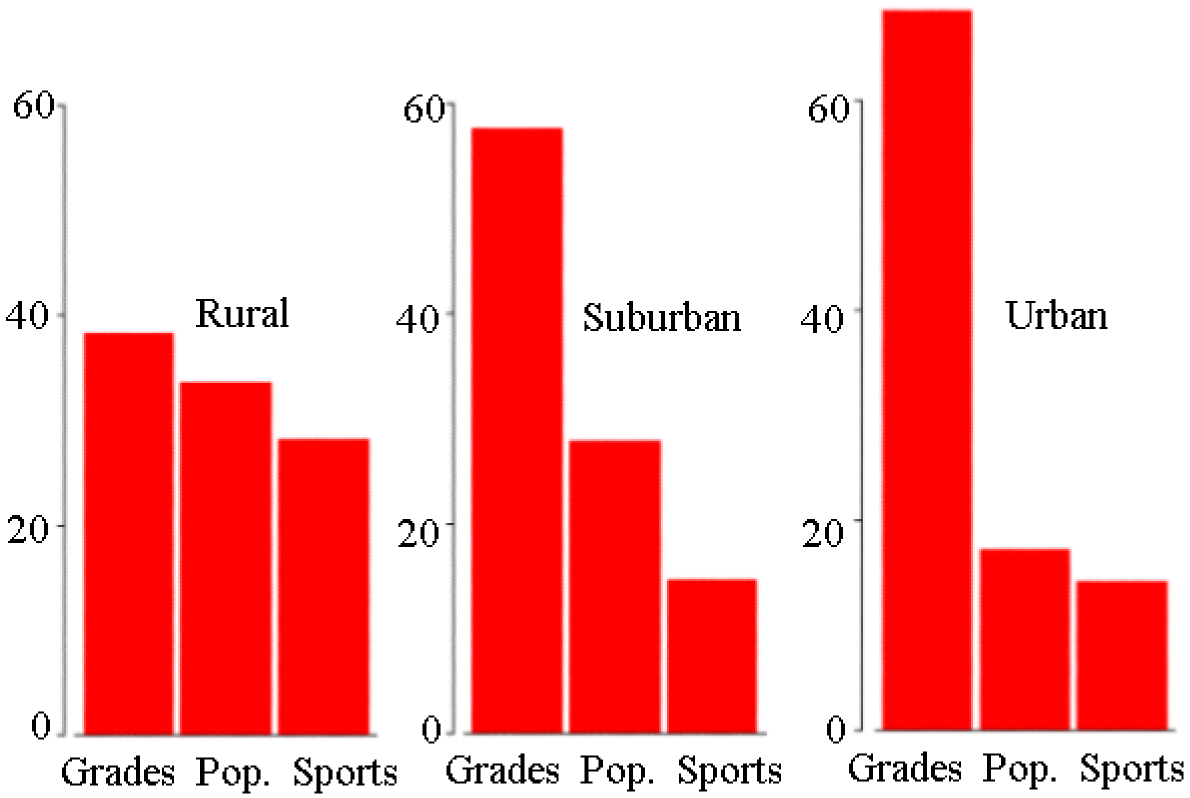
	School Area			
Goals	Rural	Suburban	Urban	Total
Grades	57	87	24	168

```
Popular   |  50       42        6       98
 Sports   |  42       22        5       69
--------------------------------------------
Total     |  149      151       35      335
```

The corresponding column percentages are the following:

```
            School Area
Goals     | Rural  Suburban   Urban
-----------------------------------
 Grades   |  38       58       69
Popular   |  34       28       17
 Sports   |  28       14       14
-----------------------------------
Total     | 100      100      100
```

Barplots comparing the percentages of students' choices by school area appear below:



From the table and corresponding graphs, it appears that the emphasis on grades increases as the school areas become more urban, while the emphasis on popularity decreases. Is this association significant?

Using the MINITAB "CHIS" command to perform a chi-square test on the tabular data gives the following results:

```
Chi-Square Test

Expected counts are printed below observed counts

          Rural Suburban    Urban    Total
     1       57       87       24      168
          74.72    75.73    17.55

     2       50       42        6       98
          43.59    44.17    10.24

     3       42       22        5       69
          30.69    31.10     7.21
```

```
Total        149      151       35      335

Chi-Sq =  4.203 +  1.679 +  2.369 +
          0.943 +  0.107 +  1.755 +
          4.168 +  2.663 +  0.677 =  18.564
DF = 4, P-Value = 0.001
```

The *P-value* is highly significant, indicating that some association between the variables is present. We can conclude that the urban students' increased emphasis on grades is not due to random variation.

Data source: Chase, M.A and Dummer, G.M. (1992), "The Role of Sports as a Social Determinant for Children," Research Quarterly for Exercise and Sport, 63, 418-424. Dataset available through the Statlib Data and Story Library (DASL).

The chi-square index in the Statlib Data and Story Library (DASL) provides several other examples of the use of the chi-square test in categorical data analysis.