

## Contingency Tables

### Prerequisites

prerequisites

This section shows how to use Chi Square to test the relationship between nominal variables for significance. For example, Table 1 shows the data from the [Mediterranean Diet and Health](#) case study.

Table 1. Frequencies for Diet and Health Study

| Diet          | Outcome |                     |                         |         |       |
|---------------|---------|---------------------|-------------------------|---------|-------|
|               | Cancers | Fatal Heart Disease | Non-Fatal Heart Disease | Healthy | Total |
| AHA           | 15      | 24                  | 25                      | 239     | 303   |
| Mediterranean | 7       | 14                  | 8                       | 273     | 302   |
| Total         | 22      | 38                  | 33                      | 512     | 605   |

The question is whether there is a [significant relationship](#) between diet and outcome. The first step is to compute the expected frequency for each cell based on the assumption that there is no relationship. These expected frequencies are computed from the totals as follows. We begin by computing the expected frequency for the AHA Diet/Cancers combination. Note that 22/605 subjects developed cancer. The proportion who developed cancer is therefore 0.0364. If there were no relationship between diet and outcome, then we would expect 0.0364 of those on the AHA diet to develop cancer. Since 303 subjects were on the AHA diet, we would expect  $(0.0364)(303) = 11.02$  cancers on the AHA diet. Similarly, we would expect  $(0.0364)(302) = 10.98$  cancers on the Mediterranean diet. In general, the expected frequency for a cell in the  $i$ th row and the  $j$ th column is equal to

$$E_{i,j} = \frac{T_i T_j}{T}$$

where  $E_{i,j}$  is the expected frequency for cell  $i,j$ ,  $T_i$  is the total  $i$ th row,  $T_j$  is the total for the  $j$ th column, and  $T$  is the total number of observations. For the AHA Diet/Cancers cell,  $i = 1$ ,  $j = 1$ ,  $T_i = 303$ ,  $T_j = 22$ , and  $T = 605$ . Table 2 shows the expected frequencies (in parenthesis) for each cell in the experiment. Table

2 shows the expected frequencies (in parenthesis) for each cell in the experiment.

Table 2. Observed and Expected Frequencies for Diet and Health Study

| Diet          | Outcome       |                     |                         |                 |       |
|---------------|---------------|---------------------|-------------------------|-----------------|-------|
|               | Cancers       | Fatal Heart Disease | Non-Fatal Heart Disease | Healthy         | Total |
| AHA           | 15<br>(11.02) | 24<br>(19.03)       | 25<br>(16.53)           | 239<br>(256.42) | 303   |
| Mediterranean | 7<br>(10.98)  | 14<br>(18.97)       | 8<br>(16.47)            | 273<br>(255.58) | 302   |
| Total         | 22            | 38                  | 33                      | 512             | 605   |

The significance test is conducted by computing Chi Square as follows.

$$\chi^2_3 = \sum \frac{(E-O)^2}{E} = 16.55$$

The degrees of freedom is equal to  $(r-1)(c-1)$  where  $r$  is the number of rows and  $c$  is the number of columns. For this example, the degrees of freedom is  $(2-1)(4-1) = 3$ . The [Chi Square calculator](#) can be used to determine that the probability value for a Chi Square of 16.55 with three degrees of freedom is less 0.0009. Therefore, the null hypothesis of no relationship between diet and outcome can be rejected.

A key assumption of the Chi Square test of independence is that each subject contributes data to only one cell. Therefore the sum of all cell frequencies in the table must be the same as the number of subjects in the experiment. Consider an experiment in which each of 16 subjects each attempted two anagram problems. The data are shown in Table 3.

Table 3. Anagram Problem Data

|               | Anagram 1 | Anagram 2 |
|---------------|-----------|-----------|
| Solved        | 10        | 4         |
| Did not Solve | 6         | 12        |

It would not be valid to use the Chi Square test on these data since each subject contributed data to two cells: one cell based on their performance on Anagram 1 and one cell based on their performance on Anagram 2. The total of the cell frequencies in the table is 32 but the total number of subjects is only 16.

The formula for Chi Square yields a statistic that is only approximately a Chi Square distribution. In order for the approximation to be adequate, the total number of subjects should be at least 20. Some authors claim that the correction for continuity should be used whenever an expected cell frequency is below 5. Research in statistics has shown that this practice is not advisable. For example, see:

Bradley, D. R., Bradley, T. D., McGrath, S. G., & Cutcomb, S. D. (1979) Type I error rate of the chi square test of independence in  $r \times c$  tables that have small expected frequencies. *Psychological Bulletin*, 86, 1200-1297.

The correction for continuity when applied to  $2 \times 2$  contingency tables is called the Yates correction. The simulation [2 x 2 tables](#) lets you explore the value of this correction.