III. Test for independence of two categorical variables with a contingency table

A. The **one category variable** problem on page 120 tested one variable (sales) against some hypothesized frequency to determine if there was a good fit between the hypothesized frequency and the observed frequency.

B. Here, a contingency table is used to determine if there is a relationship (statistical dependency) between two variables. That is, does knowledge of variable A's value provide knowledge of variable B's value. If so, variables are statistically dependent. Otherwise, they are independent.

C. The idea of statistical dependency was first encountered in the probability chapter on page 46. At that time, advertising expenditures and sales revenue were said to be dependent. To be sure sales increased enough when advertising increased to indicate dependency, a statistical proof for dependency could be conducted. We must adjust the monthly data on page 46 to weekly data because cell values (f_e) must be ≥ 5. Fifty weeks of data will be studied. The null hypothesis will again proclaim no difference (sales and advertising are independent). The test will measure whether the difference is large and did not happen by chance. That is, the variables are dependent.

Weekly Advertising and Sales

Sales / Advertising	Less than or equal to $12,000	Greater than $12,000	Totals
Less than or equal to $1,000	20	5	25
Greater than $1,000	5	20	25
Totals	25	25	50

D. The 5-step approach to hypothesis testing

1. H_0 : advertising and sales are independent
   H_1 : advertising and sales are dependent
2. The significance level is .01.
3. Chi-square is the test statistic.

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right] \quad \text{where } f_e = \frac{f_r \times f_c}{n}$$

df = (r - 1) (c - 1)
r is the number of rows,
and c the number of columns

Contingency Table of Weekly Advertising and Sales

Sales / Advertising	Less than or equal to $12,000		Greater than $12,000		Totals	
	f_o	f_e	f_o	f_e	f_o	f_e
Less than or equal to $1,000	20	12.5	5	12.5	25	25
Greater than $1,000	5	12.5	20	12.5	25	25
Totals	25	25	25	25	50	50

4. If χ^2 from the test statistic is beyond the critical value, reject the null hypothesis.
5. Applying the decision rule for this one-tail test.
   a. Imagine the above contingency table has only f_o data and the f_e data cells and totals are blank.
   b. Row and column totals for f_e are equal to those of f_o.
   c. A table cell is completed by multiplying its row total by its column total and dividing by the grand total. For example, the first f_e cell has been calculated in the frame to the right.

Note: If 2 variables are independent, their cell values are in proportion. This formula is used to determine expected row and column cell values.

$$f_e = \frac{f_r \times f_c}{n} = \frac{25 \times 25}{50} = 12.5$$

df = (r - 1) (c - 1) = (2 - 1) (2 - 1) = 1 → χ^2 = 6.64 (see chart page 120)

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right] = \Sigma\left[\frac{(20-12.5)^2}{12.5} + \frac{(5-12.5)^2}{12.5} + \frac{(5-12.5)^2}{12.5} + \frac{(20-12.5)^2}{12.5}\right]$$

$$= 4.5 + 4.5 + 4.5 + 4.5 = 18$$

The null hypothesis is rejected because 18 > 6.64. Advertising expenditures affect sales revenue. These variables are dependent at the .01 level of significance.

Note: Chi-square analysis is used to test interesting relationships such as level of income (low, medium, and high) and frequency of purchase (often and not often).

Note: As demonstrated with this advertising/sales data, it is often necessary to regroup data to assure that f_e is ≥ 5. Classes with a low frequency are combined until the requirement is observed.