15.5: Contingency Tables for Two Variables

The goodness-of-fit test is a useful tool for assessing a single categorical variable. However, what is more common is wanting to know if two categorical variables are related to one another. This type of analysis is similar to a correlation, the only difference being that we are working with nominal data, which violates the assumptions of traditional correlation coefficients. This is where the χ^2 test for independence comes in handy.

As noted above, our only description for nominal data is frequency, so we will again present our observations in a frequency table. When we have two categorical variables, our frequency table is crossed. That is, each combination of levels from each categorical variable are presented. This type of frequency table is called a contingency table because it shows the frequency of each category in one variable, contingent upon the specific level of the other variable.

An example contingency table is shown in Table 15.5.1, which displays whether or not 168 college students watched college sports growing up (Yes/No) and whether the students' final choice of which college to attend was influenced by the college's sports teams (Yes – Primary, Yes – Somewhat, No):

Table 15.5.1: Contingency table of college sports and decision making

College Sports		Affected Decision			
		Primary	Somewhat	No	Total
Watched	Yes	47	26	14	87
	No	21	23	37	81
	Total	68	49	51	168

In contrast to the frequency table for our goodness-of-fit test, our contingency table does not contain expected values, only observed data. Within our table, wherever our rows and columns cross, we have a cell. A cell contains the frequency of observing it's corresponding specific levels of each variable at the same time. The top left cell in Table 15.5.1 shows us that 47 people in our study watched college sports as a child AND had college sports as their primary deciding factor in which college to attend.

Cells are numbered based on which row they are in (rows are numbered top to bottom) and which column they are in (columns are numbered left to right). We always name the cell using (R,C), with the row first and the column second. A quick and easy way to remember the order is that R/C Cola exists but C/R Cola does not. Based on this convention, the top left cell containing our 47 participants who watched college sports as a child and had sports as a primary criteria is cell (1,1). Next to it, which has 26 people who watched college sports as a child but had sports only somewhat affect their decision, is cell (1,2), and so on. We only number the cells where our categories cross. We do not number our total cells, which have their own special name: marginal values.

Marginal values are the total values for a single category of one variable, added up across levels of the other variable. In table 3, these marginal values have been italicized for ease of explanation, though this is not normally the case. We can see that, in total, 87 of our participants (47+26+14) watched college sports growing up and 81 (21+23+37) did not. The total of these two marginal values is 168, the total number of people in our study. Likewise, 68 people used sports as a primary criteria for deciding which college to attend, 49 considered it somewhat, and 51 did not use it as criteria at all. The total of these marginal values is also 168, our total number of people. The marginal values for rows and columns will always both add up to the total number of participants, N, in the study. If they do not, then a calculation error was made and you must go back and check your work.

Expected Values of Contingency Tables

Our expected values for contingency tables are based on the same logic as they were for frequency tables, but now we must incorporate information about how frequently each row and column was observed (the marginal values) and how many people were in the sample overall (N) to find what random chance would have made the frequencies out to be. Specifically:

$$E_{ij} = \frac{R_i C_j}{N} \tag{15.5.1}$$

The subscripts i and j indicate which row and column, respectively, correspond to the cell we are calculating the expected

frequency for, and the R_i and C_j are the row and column marginal values, respectively. N is still the total sample size. Using the data from Table 15.5.1, we can calculate the expected frequency for cell (1,1), the college sport watchers who used sports at their primary criteria, to be:

$$E_{1,1} = \frac{87 * 68}{168} = 35.21$$

We can follow the same math to find all the expected values for this table:

Table 15.5.2: Contingency table of college sports and decision making

College Sports		Affected Decision			
		Primary	Somewhat	No	Total
Watched	Yes	35.21	25.38	26.41	87
	No	32.79	23.62	24.59	81
	Total	68	49	51	

Notice that the marginal values still add up to the same totals as before. This is because the expected frequencies are just row and column averages simultaneously. Our total N will also add up to the same value.

The observed and expected frequencies can be used to calculate the same χ^2 statistic as we did for the goodness-of-fit test. Before we get there, though, we should look at the hypotheses and degrees of freedom used for contingency tables.
