16.5: Introduction to Test of Independence

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 qualitative data (nominal scale of measurement), which violates the assumptions of traditional correlation analylses. Although we learned in the correlation chapter that there is a type of correlation (Phi correlation) when there are two binary variables (two variables that each only have two options), the χ^2 test for independence comes in handy when the variables have more than two categories or levels. The χ^2 test performed when there are two variables is known as the Test of Independence. In this analysis, we are looking to see if the values of each qualitative variable (that is, the frequency of their levels) is related to or independent of the values of the other qualitative variable.

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

Example 16.5.1

How are contingency tables different from factorial design tables?

Solution

The difference is what's in the cells. In factorial design squares, the variables are crossed and the means for each cell (and the margins) are what we're interested in. For contingency tables, what's in the cells is the actual frequencies (not means).

An example contingency table is shown in Table 16.5.1, 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 16.5.1: Contingency Table of College Sports and College Decision

College Sports	Primary Affect	Somewhat Affected	Did Not Affect Decision	*Total*
Watched	47	26	14	$\sum_{Row} = 87$
Did Not Watch	21	23	37	$\sum_{Row} = 81$
Total	$\sum_{Column} = 68$	$\sum_{Column} = 49$	$\sum_{Column} = 51$	$\sum_{Column} = 168$

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 16.5.1 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 for Rows and C for Columns, 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 or R1,C1. Next to it, which has 26 people who watched college sports as a child but had sports only somewhat affect their decision, cell is 1,2 or R1,C2, 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. Ooh, that sounds familiar! Marginal values are the total values for a single category of one variable, added up across levels of the other variable. In Table 16.5.1, 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, 50 considered it somewhat, and 50 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. Yay for calculation checks!

Expected Values of Contingency Tables

Because these crossed contingency tables have more data, we don't usually start out with the Expected values in any row or column like we included in the Goodness of Fit tables. However, the 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:

$$E_{EachCell} = \frac{RT * CT}{N}$$

in which RT stands for Row Total and CT stands for Column Total. This allows for calculating an Expected frequency for each cell. Using the data from Table 16.5.1, we can calculate the expected frequency for cell R1,C1, the college sport watchers who used sports at their primary criteria, to be:

$$E_{R1,C1} = \frac{87 * 68}{168} = 35.21$$

We can follow the same math to find all the expected values for all of Expected frequencies in each cell.

Example 16.5.1

Use the Observed frequencies from Table 16.5.1 to find the Expected frequencies for each combination of the two qualitative variables.

Solution

Table 16.5.2: Table of EXPECTED Frequences of College Sports and College Decision

College Sports	Primary Affect	Somewhat Affected	Did Not Affect Decision	*Total*
Watched	$E_{R1,C1} = \frac{87 \times 68}{168} = 35.21$	$E_{R1,C2} = \frac{87 \times 49}{168} = 25.38$	$E_{R1,C1} = \frac{87 \times 51}{168} = 26.41$	$\sum_{Row} = 87$
Did Not Watch	32.79	23.63	24.59	$\sum_{Row} = 81$
Total	$\sum_{Column} = 68$	$\sum_{Column} = 49$	$\sum_{Column} = 51$	$\sum_{Row} = 168$

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

These 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 can do that, though, we should look at the hypotheses and the critical value table.
