**Section 11.2 (Part 1) - Inference for Relationships**

**Comparing Distributions of a Categorical Variable** - In this section, we will compare more than two samples or groups. More generally, we are going to compare distributions of a single categorical variable across several populations or treatments.

**Example** - Market researchers suspect that background music may affect mood and buying behavior of customers. One study in a supermarket compared three randomly assigned treatments: no music, French accordion music, and Italian string music. Under each condition, the researchers recorded the numbers of bottles of French, Italian, and other wine purchased. Here is a table that summarizes the data



a) Calculate the conditional distribution of the type of wine sold for each treatment.

b) Given the graphs below, comparing the distributions of wine purchases under the different conditions, are the distributions similar or different?



**Chi-Square Test for Homogeneity** - When we are trying to determine if the distribution of a categorical variable is the same for several populations or treatments, we conduct the chi-square test for homogeneity. For our example, this will involve testing the hypotheses:

*H0: There is no difference in the distributions of wine purchases at this store when no music, French accordion music, or Italian string music is played.*

*Ha: There is a difference in the distributions of wine purchases at this store when no music, French accordion music, or Italian string music is played.*

**Computing Expected Counts** - Before we formally define the chi-square test for homogeneity, we will explore how to compute the expected counts.

|  |  |
| --- | --- |
| tableun_11_28.jpg | To find the expected counts, we start by assuming that *H0* is true. From the two way table, we can see that 99 of the 243 bottles of wine bought during the study were French wines. If the specific type of music that is playing had no effect on wine purchase, the proportion of French wine sold under each |

Condition should be 99/243 = 0.407. For instance, there were 84 bottles of wine bought when no music was playing. Then on average, how many bottles would we expect to be French wine?

The expected counts of the French wine bought under the other two conditions can be found in a similar way:

 French music: Italian music:

We repeat the procedure to find the expected counts for the other two types of wine.

 Italian wine

 No music: French music: Italian music:

 Other wine

 No music: French music: Italian music:

|  |  |
| --- | --- |
| tableun_11_29.jpg | tableun_11_30.jpg |

Using the numbers marked in the table above, let’s try to determine a general formula for determining the expected counts.

**Test Statistic** - The chi-square test statistic is computed the same way it was computed in the chi-square goodness-of-fit test.
 $χ^{2}=\sum\_{}^{}\frac{\left(Observed-Expected\right)^{2}}{Expected}$

Continuing with the music example, for French wine with no music, the observed count is 30 bottles and the expected count is 34.22. The contribution for the chi-square statistic for this cell is:

This would be repeated for all nine cells and summed to determine the chi-square statistic.

**The Chi-Square Test for Homogeneity**

Suppose the Random, Large Sample Size, and Independent conditions are met. You can use the chi-square test for homogeneity to test

*H0: There is no difference in the distributions of a categorical variable for several populations or treatments.
Ha: There is a difference in the distributions of a categorical variable for several populations or treatments.*

Start by finding the expected counts. Then calculate the chi-square statistic

$$χ^{2}=\sum\_{}^{}\frac{\left(Observed-Expected\right)^{2}}{Expected}$$

where the sum is over all cells (not including totals) in the two-way table. If *H0* is true, the chi-square statistic has approximately a chi-square distribution with degrees of freedom =
(number of rows - 1)(number of columns - 1). The *P*-value is the area to the right of the chi-square statistic under the corresponding chi-square density curve.

*Conditions*: Use this test when:

* **Random** - The data come from separate random samples from each population of interest or from the groups in a randomized experiment.
* **Large Sample Size** - All expected counts are at least 5.
* **Independent** - Both the samples or groups themselves and the individual observations in each sample or group are independent. When sampling without replacement, check that the individual populations are at least 10 times as large as the corresponding samples (10% condition).

**Example (cont)** - In the wine example, the chi-square test statistic was 18.28. The degrees of freedom will be (3 - 1)(3 - 1) = 4. Using the calculator, the *P*-value is 0.0011.

Conclusion:

**Technology** - To conduct this test with a calculator, press [2nd] [x-1] (MATRIX) <EDIT> and choose 1: A. Enter the dimensions of the contingency table. Enter the observed counts. Press [STAT] <TESTS> and choose C: $χ^{2}$-Test.

**Application** - Complete CYU question 2 on p. 708.



**Follow-up Analysis**



HW: Read Sec 11.2; problems 27, 29, 31, 37, 39.