MATH 1070 - Chapter 25 Notes

Two Categorical Variables: The Chi-Square Test

So far we have been studying inference in the context of variables that take on continuous values. Ex:

What if we are interested in a categorical variable?

Recall: A **categorical variable** is one which takes on a finite number of values or categories.

Example: Quality of recycled products:

	Higher	Same	Lower
Buyers	20	7	9
Non - Buyers	29	25	43

We may be interested in asking questions such as:

1. Presenting Categorical Data

- When there are two categorical variables, the data are summarized in a two-way table.
- The number of observations falling into each combination of the two categorical variables is entered into each cell of the table
- Relationships between categorical variables are described by calculating appropriate percents from the counts given in the table

Example: (Health Care - Canada and U.S.) Consider the data from patients? own assessment of their quality of life relative to what it had been before their heart attack (data from patients who survived at least a year after the attack.) Data presented for the US and Canada.

Quality of life	**Canada**	**US**
Much better	75	541
Somewhat better	71	498
About the same	96	779
Somewhat worse	50	282
Much worse	19	65
Total	311	2165

Compare the Canadian and the US patients in terms of feeling much better.

Now consider the following table:

Quality of life	**Canada**	**US**
Much better	24%	25%
Somewhat better	23%	23%
About the same	31%	36%
Somewhat worse	16%	13%
Much worse	6%	3%
Total	100%	100%

What do we see?

Potential question of interest: Is there a relationship between the explanatory variable (Country) and the response variable (Quality of life)?

2. Hypothesis Test

Just as we have seen in previous chapters, with tests for two categorical variables, we are interested in whether a relationship observed in a single sample reflects a real relationship in the population.

Hypotheses:
H_0: The percentages for one variable are the same for every level of the other variable.
H_A: The percentages for one variable vary over levels of the other variable.

Which are the two variables in this example?

What are their levels?

For instance, could look at differences in percentages between Canada and U.S. for each level of 'Quality of life':
24% vs. 25% for those who felt 'Much better'.
23% vs. 23% for 'Somewhat better', etc.
We are making **multiple comparisons**.

This boils down to the problem of how to make many comparisons at the same time with some overall measure of confidence in all the conclusions.

Steps involved:

(1) **Overall test** to test for any differences.
(2) If there is sufficient evidence of differences with the overall test, then do a **follow-up analysis** to decide which parameters (or groups) differ and how large the differences are.

In this chapter, we are concerned with the **overall test.**

3. χ^2 Test of Overall Significance.

What are the hypotheses involved in this over all test?

How do we test this hypothesis?

Note: If the observed counts are far from the expected counts, that is considered evidence against H_0.

Expected Counts: (Health Care - Canada and U.S.) Find the expected value in each cell.

First, Compute the totals for each level of the variables.

Quality of life	**Canada**	**US**	**Total**
Much better	75	541	
Somewhat better	71	498	
About the same	96	779	
Somewhat worse	50	282	
Much worse	19	65	
Total	311	2165	

The **expected count** of Canadians who felt 'Much Better' (Row 1, Column 1) is given by:

Similarly, the **expected count** of Americans who felt 'Much Worse' (Row 5, Column 2) is given by:

So, in general, the **expected count** of any cell in a two-way table under the H_0 is given by:

The table of all observed and expected counts for the example is then given by:

Quality of life	Canada	US	Total
OBSERVED COUNTS	**Much better**	75	541
	Somewhat better	71	498
	About the same	96	779
	Somewhat worse	50	282
	Much worse	19	65
EXPECTED COUNTS	**Much better**	77.37	538.63
	Somewhat better	71.47	497.53
	About the same	109.91	765.09
	Somewhat worse	41.70	290.30
	Much worse	10.55	73.45

We now compare to see if the data supports the H_0.

To determine if the differences between the observed counts and expected counts are statistically significant (to show a real relationship between the two categorical variables), we use the following **chi-squared statistic**:

Notation: The **chi-squared statistic** is represented by the Greek letter:

Properties of the chi-squared statistic: The chi-squared statistic is a measure of the distance of the observed counts from the expected counts.

(1) Range of values:

(2) Exactly equal to 0 when

(3) Large values

(4) One-sided or two-sided?

Example: (Health Care - Canada and U.S.) Compute the χ^2 statistic value:

Quality of life	Canada	US	Total
OBSERVED COUNTS	**Much better**	75	541
	Somewhat better	71	498
	About the same	96	779
	Somewhat worse	50	282
	Much worse	19	65
EXPECTED COUNTS	**Much better**	77.37	538.63
	Somewhat better	71.47	497.53
	About the same	109.91	765.09
	Somewhat worse	41.70	290.30
	Much worse	10.55	73.45

p-value: Use the χ^2 distribution's table from textbook **(Table D)** to find the p-value:
p-value =

If the p-value is small, then we reject H_0 (recall H_0: no significant relationship exists.)

- Compare appropriate percents in data table.
- Compare individual observed and expected cell counts.
- Look at individual terms in the chi-square statistic.

Properties of the χ^2 distribution:

(1)

(2) The only parameter of the distribution:

(3) A χ^2 test for a two-way table with r rows and c columns uses critical values from a χ^2 distribution with:

(4) The p-value is the area to the right of the observed χ^2 statistic value under the density curve.

It answers the question:

Example: (Health Care - Canada and U.S.) Compute the degrees of freedom and the p-value associated with the observed χ^2 statistic value.

Assumptions/Requirements of the χ^2 test:

(1) The chi-square test is an approximate method, and becomes more accurate as the counts in the cells of the table get larger.
(2) The following must be satisfied for the approximation to be accurate:
    - No more than 20% of the expected counts are less than 5
    - All individual expected counts are 1 or greater
(3) If these requirements fail, then two or more groups must be combined to form a new ('smaller') two-way table.
(4) The data for each categorical variable is collected as a simple random sample.
(5) The samples across variables are independently collected.

4. χ^2 Goodness of Fit Test

A variation of the Chi-square statistic can also be used to test the H_0:

The H_0 specifies the probabilities (p_i)

The chi-square goodness of fit test compares the observed counts for each category with the expected counts under the null hypothesis.

Formally, the steps involved are:

(1) : Write hypotheses:

H_0:
H_A:

(2) For a sample of n subjects, observe how many subjects fall in each category.

(3) Calculate the expected number of subjects in each category under the null hypothesis:

(4) We then calculate the chi-square statistic (in a manner identical to the previous test):

(5) The degrees of freedom for this statistic are $k-1$ (the number of possible categories minus one).

(6) Find p-value using Table D.

Example: The manufacturer of M&Ms reports that on average their candies contain 13 percent of each of browns and reds, 14 percent yellows, 16 percent greens, 20 percent oranges, and 24 percent blues. The colors in a sample bag of M&Ms are counted and presented below:

Color	Blue	Orange	Green	Yellow	Red	Brown	Total
Count	9	8	12	15	10	6	60

Is the data consistent with what is reported by the manufacturer?

Example: A random sample of 140 births from local records was collected to show that there are fewer births on Saturdays and Sundays than there are on weekdays. Here is the data:

Day	Sun.	Mon	Tue	Wed	Thu	Fri	Sat
Births	13	23	24	20	27	18	15

Do these data give significant evidence that local births are not equally likely on all days of the week?