Chapter 10: Analyzing the Association Between Categorical Variables

Section 10.2: How Can We Test Whether Categorical Variables Are Independent?

Learning Objectives

1. A Significance Test for Categorical Variables
2. What Do We Expect for Cell Counts if the Variables Are Independent?
3. How Do We Find the Expected Cell Counts?
4. The Chi-Squared Test Statistic
5. The Chi-Squared Distribution
6. The Five Steps of the Chi^2 Test of Independence
7. Chi-Squared and the Test Comparing Proportions in 2x2 Tables
8. Limitations of the Chi-Squared Test

Learning Objective 1: A Significance Test for Categorical Variables

- **Create a table of frequencies divided into the categories of the two variables**
  - **The hypotheses for the test are:**

    H_0: The two variables are *independent*

    H_a: The two variables are *dependent (associated)*

  - **The test assumes random sampling and a large sample size (cell counts in the frequency table of at least 5)**

Learning Objective 2: What Do We Expect for Cell Counts if the Variables Are Independent?

- **The *count* in any particular cell is a random variable**
  - **Different samples have different count values**
- **The *mean* of its distribution is called an expected cell count**
  - **This is found under the presumption that H_0 is true**

Learning Objective 3: How Do We Find the Expected Cell Counts?

- **Expected Cell Count**:
  - **For a particular cell,**

$$\text{Expected cell count} = \frac{(\text{Row total}) \times (\text{Column total})}{\text{Total sample size}}$$

- **The expected frequencies are values that have the same row and column totals as the observed counts, but for which the conditional distributions are identical (this is the assumption of the null hypothesis).**

5

Learning Objective 3: How Do We Find the Expected Cell Counts? Example

TABLE 11.5: Happiness by Family Income, Showing Observed and Expected Cell Counts

We use the highlighted totals to get the expected count of 35.8 = (290 × 168)/1362 in the first cell.

	Happiness			
INCOME	Not too Happy	Pretty Happy	Very Happy	Total
Above average	21	159	110	290
	35.8	166.1	88.1	
Average	53	372	221	646
	79.7	370.0	196.4	
Below average	94	249	83	426
	52.5	244.0	129.5	
Total	168 (12.3%)	780 (57.3%)	414 (30.4%)	1362 (100%)

6

Learning Objective 4: The Chi-Squared Test Statistic

- **The chi-squared statistic summarizes how far the observed cell counts in a contingency table fall from the expected cell counts for a null hypothesis**

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

7

Learning Objective 4: Example: Happiness and Family Income

- **State the null and alternative hypotheses for this test**
- **H_0: Happiness and family income are independent**
- **H_a: Happiness and family income are dependent (associated)**

8

Learning Objective 4:
Example: Happiness and Family Income

- **Report the χ^2 statistic and explain how it was calculated:**
- **To calculate the χ^2 statistic, for each cell, calculate:**

$$\frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

- **Sum the values for all the cells**
- **The χ^2 value is 73.4**

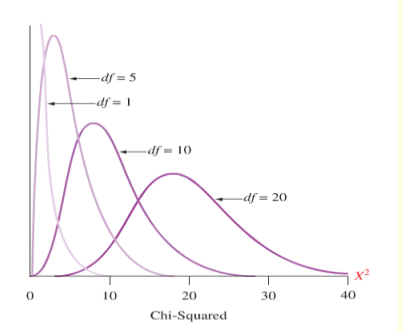
Learning Objective 4:
The Chi-Squared Test Statistic

- **The larger the χ^2 value, the greater the evidence against the null hypothesis of independence and in support of the alternative hypothesis that happiness and income are associated**

Learning Objective 5:
The Chi-Squared Distribution

- **To convert the χ^2 test statistic to a P-value, we use the sampling distribution of the statistic χ^2**
- **For large sample sizes, this sampling distribution is well approximated by the *chi-squared probability distribution***

Learning Objective 5:
The Chi-Squared Distribution



Learning Objective 5:
The Chi-Squared Distribution

- **Main properties of the chi-squared distribution:**
  - **It falls on the positive part of the real number line**
  - **The precise shape of the distribution depends on the degrees of freedom:**

    df = (r-1)(c-1)

13

Learning Objective 5:
The Chi-Squared Distribution

- **Main properties of the chi-squared distribution:**
  - **The mean of the distribution equals the *df* value**
  - **It is skewed to the right**
  - **The larger the χ^2 value, the greater the evidence against H_0: independence**

14

Learning Objective 5:
The Chi-Squared Distribution

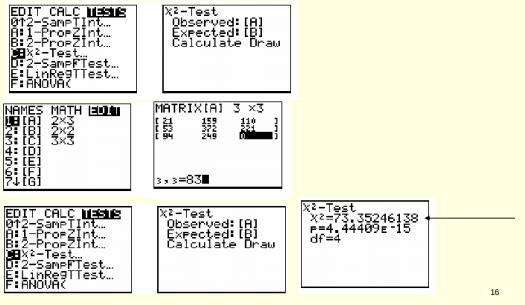
TABLE 11.7: Some Rows of Table C Displaying Chi-Squared Values

The values have right-tail probabilities between 0.250 and 0.001. For a table with $r = 3$ rows and $c = 3$ columns, $df = (r - 1)(c - 1) = 4$, and 9.49 is the chi-squared value with a right-tail probability of 0.05.

	Right-Tail Probability						
df	.250	.100	.050	.025	.010	.005	.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.81	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47

15

Learning Objective 4:
Example: Happiness and Family Income



16

Learning Objective 6:
The Five Steps of the Chi-Squared Test of Independence

1. Assumptions:

- **Two categorical variables**
- **Randomization**
- **Expected counts ≥ 5 in all cells**

Learning Objective 6:
The Five Steps of the Chi-Squared Test of Independence

2. Hypotheses:

- **H_0: The two variables are independent**
- **H_a: The two variables are dependent (associated)**

Learning Objective 6:
The Five Steps of the Chi-Squared Test of Independence

3. **Test Statistic:**

$$\chi^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

Learning Objective 6:
The Five Steps of the Chi-Squared Test of Independence

4. P-value: Right-tail probability above the observed χ^2 value, for the chi-squared distribution with *df* = (r-1)(c-1)

5. Conclusion: Report P-value and interpret in context

- **If a decision is needed, reject H_0 when P-value ≤ significance level**

Learning Objective 7:
Chi-Squared and the Test Comparing Proportions in 2x2 Tables

- In practice, contingency tables of size 2x2 are very common. They often occur in summarizing the responses of two groups on a binary response variable.
  - Denote the population proportion of success by p_1 in group 1 and p_2 in group 2
  - If the response variable is independent of the group, $p_1=p_2$, so the conditional distributions are equal
    - H_0: $p_1=p_2$ is equivalent to H_0: independence
    - $$z^2 = \chi^2 \quad \text{where}$$
      $$z = (\hat{p}_1 - \hat{p}_2)/se_0$$

Learning Objective 7:
Example: Aspirin and Heart Attacks Revisited

The same P-value results as with a two-sided z test comparing the two population proportions.

```
Rows: group   Columns: heart
                yes      no
placebo         189   10845      <-- Cell counts
                146   10888      <-- Expected cell counts
aspirin         104   10933
                147   10890
Cell Contents:      Count
                    Expected count
Pearson Chi-Square = 25.01, DF = 1, P-Value = 0.000
```


Learning Objective 7:
Example: Aspirin and Heart Attacks Revisited

- **What are the hypotheses for the chi-squared test for these data?**
- **The *null hypothesis* is that whether a doctor has a heart attack is *independent* of whether he takes placebo or aspirin**
- **The *alternative hypothesis* is that there's *an association***

Learning Objective 7:
Example: Aspirin and Heart Attacks Revisited

- **Report the test statistic and P-value for the chi-squared test:**
  - **The test statistic is 25.01 with a P-value of 0.000**
- **This is very strong evidence that the population proportion of heart attacks differed for those taking aspirin and for those taking placebo**

Learning Objective 7:
Example: Aspirin and Heart Attacks Revisited

- **The sample proportions indicate that the aspirin group had a lower rate of heart attacks than the placebo group**

Learning Objective 8:
Limitations of the Chi-Squared Test

- **If the P-value is very small, strong evidence exists against the null hypothesis of independence**

But…

- **The chi-squared statistic and the P-value tell us nothing about the nature of the strength of the association**

Learning Objective 8:
Limitations of the Chi-Squared Test

- **We know that there is *statistical* significance, but the test alone does not indicate whether there is *practical* significance as well**

Learning Objective 8:
Limitations of the Chi-Squared Test

- The chi-squared test is often misused. Some examples are:
  - when some of the expected frequencies are too small
  - when separate rows or columns are dependent samples
  - data are not random
  - quantitative data are classified into categories - results in loss of information