Categorical Data: Relationships

In this chapter we extend our study of categorical data to several populations. We will

- discuss independence and association for categorical variables.
- describe a chi-square test to assess the independence between two categorical variables.
- discuss confidence intervals for difference between proportions.

The Chi-Square Test for the 2×2 Contingency Table

Example: migraine headache

Patients who suffered from moderate to severe migraine headache took part in a double-blind clinical trial to assess an experimental surgery. A group of 75 patients were randomly assigned to receive either the real surgery on migraine trigger sites ($n = 49$) or a sham surgery ($n = 26$) in which an incision was made but no further procedure was performed. The surgeons hoped that patients would experience "a substantial reduction in migraine headaches," which we will label as "success." The table below shows the results of the experiment, which is called a **contingency table**.

The focus of interest in a contingency table is *the dependence or association between the column variable and the row variable* (between treatment and response in the table below). In particular, the table below is called 2×2 ("two-by-two") contingency tables, because it consists of two rows (excluding the "total" row) and two columns. Each category in the contingency table is called a cell; thus, a 2×2 contingency table has four cells.

Table 10.1.1 Response to migraine surgery

		Surgery	
		Real	Sham
Substantial reduction in migraine headaches?	Success	41	15
	No success	8	11
	Total	49	26

We often want to test whether there is a significant relationship between the column variable and the row variable. The null hypothesis is

$$H_0: \text{ Surgery and substantial reduction in migraine headaches are independent.}$$

and the alternative hypothesis is

$$H_A: \text{ Surgery and substantial reduction in migraine headaches are dependent.}$$

Note that surgery and substantial reduction in migraine headaches are independent if and only if

$$P(\text{Success}|\text{Real}) = P(\text{Success}|\text{Sham}).$$

To formulate an appropriate test, recall the test statistic introduced in Chapter 9 for chi-square goodness-of-fit test,

$$T = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}.$$

Here we will use the same test statistic with $k = 4$, where the sum is taken over all four cells in the contingency table.

The first step in determining the e's for a contingency table is to calculate the row and column total frequencies (these are called the marginal frequencies) and also the grand total of all the cell frequencies; see the following table.

Table 10.2.1 Observed frequencies for migraine study

	Surgery		
	Real	Sham	Total
Success	41	15	56
No success	8	11	19
Total	49	26	75

The e's should agree exactly with the null hypothesis. Under the null, one has

$$P(\text{Success}) = \frac{56}{75}, P(\text{No success}) = \frac{19}{75}$$

no matter what surgery was performed (real or sham). The expected frequency for the top left cell is thus

$$\frac{56}{75} \times 49 = 36.59.$$

Similarly we can find the expected frequencies for the other three cells; see the following table with expected frequencies shown in parentheses.

Table 10.2.2 Observed and expected frequencies for migraine study

	Surgery		
	Real	Sham	Total
Success	41 (36.59)	15 (19.41)	56
No success	8 (12.41)	11 (6.59)	19
Total	49	26	75

General formula for expected frequencies:

$$e = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

- Other than the differences noted previously when computing expected counts, the chi-square test for a contingency table is carried out similarly to the chi-square goodness-of-fit test.
- Large values of T indicate evidence against H_0. Critical values are determined from χ^2 Table; the number of degrees of freedom for a 2×2 contingency table is $\text{df} = 1$.
- The chi-square test for a 2×2 table has 1 degree of freedom because, in a sense, there only is one free cell in the table.
- H_0 is rejected at the α level of significance if

$$p\text{-value } = P(\chi^2_1 > T) < \alpha \text{ or } T > \chi^2_1(\alpha).$$

For the migraine experiment, the test statistic is

$$\begin{aligned} T &= \frac{(41 - 36.59)^2}{36.59} + \frac{(15 - 19.41)^2}{19.41} + \frac{(8 - 12.41)^2}{12.41} + \frac{(11 - 6.59)^2}{6.59} \\ &= 6.06. \end{aligned}$$

From χ^2 Table with $\text{df} = 1$, we find that $\chi^2_1(0.02) = 5.41$ and $\chi^2_1(0.01) = 6.63$, and so we have $0.01 < p\text{-value} < 0.02$. We reject H_0 and find that the data provide sufficient evidence to conclude that surgery and substantial reduction in migraine headaches are dependent. In other words, the real surgery is different from the sham surgery for reducing migraine headache.

Confidence interval for difference between proportions

- When we discussed constructing a confidence interval for a single proportion, p, in Chapter 9, we defined an estimate $\tilde{p}$, based on the idea of "adding 2 successes and 2 failures to the data." Making this adjustment to the data resulted in a confidence interval procedure that has good coverage properties.
- Likewise, when constructing a confidence interval for the difference in two proportions, we will define new estimates that are based on the idea of adding 1 observation to each cell of the table (so that a total of 2 successes and 2 failures are added to the data).

Consider a 2×2 contingency table that can be viewed as a comparison of two samples, of sizes n_1 and n_2, with respect to a dichotomous response variable. Let the 2×2 table be given as

Sample 1	Sample 2
y_1	y_2
$n_1 - y_1$	$n_2 - y_2$
n_1	n_2

We define

$$\tilde{p}_1 = \frac{Y_1 + 1}{n_1 + 2}, \quad \tilde{p}_2 = \frac{Y_2 + 1}{n_2 + 2}.$$

We will use the difference in the new values, $\tilde{p}_1 - \tilde{p}_2$, to construct a confidence interval for $p_1 - p_2$.

The standard error of $\tilde{p}_1 - \tilde{p}_2$ is

$$\text{SE}_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}.$$

An approximate confidence interval can be based on $\text{SE}_{\tilde{p}_1 - \tilde{p}_2}$; for instance, a 95% confidence interval is

$$(\tilde{p}_1 - \tilde{p}_2) \pm 1.96 \times \text{SE}_{\tilde{p}_1 - \tilde{p}_2}.$$

For the migraine headache data, the sample sizes are $n_1 = 49, n_2 = 26$, and the estimated probabilities of substantial reduction in migraines are

$$\tilde{p}_1 = \frac{41 + 1}{49 + 2} = \frac{42}{51} = 0.824, \quad \tilde{p}_2 = \frac{15 + 1}{26 + 2} = \frac{16}{28} = 0.571.$$

The difference between these is

$$\tilde{p}_1 - \tilde{p}_2 = 0.824 - 0.571 = 0.253.$$

Thus, we estimate that the real surgery increases the probability of substantial reduction in migraines by 0.253, compared to the sham surgery. To set confidence limits on this estimate, we calculate the standard error as

$$\text{SE}_{\tilde{p}_1 - \tilde{p}_2} = \sqrt{\frac{0.824(1 - 0.824)}{49 + 2} + \frac{0.571(1 - 0.571)}{26 + 2}} = 0.1077.$$

The 95% confidence interval is

$$0.253 \pm 1.96 \times 0.1077$$

or $(0.042, 0.464)$. We are 95% confident that the probability of substantial reduction in migraines is between 0.042 and 0.464 higher with the real surgery than with the sham surgery.

Relationship to test

The chi-square test of independence for a 2×2 contingency table is approximately, but not exactly, equivalent to checking whether a confidence interval for $p_1 - p_2$ includes zero.

The $r \times k$ Contingency Table

The ideas of chi-square test of independence extend readily to contingency tables that are larger than 2×2. We now consider a contingency table with r rows and k columns, which is termed an $r \times k$ contingency table.

Example: plover nesting

Wildlife ecologists monitored the breeding habitats of mountain plovers for 3 years and made note of where the plovers nested. They found 66 nests on agricultural fields (AF), 67 nests in shortgrass prairie dog habitat (PD), and 20 nests on other grassland (G). The nesting choices varied across the years for these 153 sampled plover broods; the table below shows the data.

Table 10.5.1 Plover nest locations across 3 years

	Year			
Location	2004	2005	2006	Total
Agricultural field (AF)	21	19	26	66
Prairie dog habitat (PD)	17	38	12	67
Grassland (G)	5	6	9	20
Total	43	63	47	153

The goal of statistical analysis of an $r \times k$ contingency table is to investigate the relationship between the row variable and the column variable. Consider the following hypotheses,

$$H_0\text{: Year and location are independent v.s. } H_A\text{: Year and location are dependent.}$$

Similar to 2×2 contingency tables, the chi-square statistic is calculated from the familiar formula

$$T = \sum_{i=1}^{r \times k} \frac{(o_i - e_i)^2}{e_i},$$

where the sum is over all $r \times k$ cells of the contingency table, and the expected frequencies are calculated as

$$e = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}.$$

The null distribution of the test statistic T is $T \overset{H_0}{\sim} \chi^2_{(r-1)(k-1)}$. H_0 is rejected at the α level of significance if

$$p\text{-value } = P(\chi^2_{(r-1)(k-1)} > T) < \alpha \text{ or } T > \chi^2_{(r-1)(k-1)}(\alpha).$$

The expected frequencies are shown in parentheses in the table below.

Table 10.5.3 Observed and expected frequencies of plover nests

	Year			
Location	2004	2005	2006	Total
Agricultural field (AF)	21 (18.55)	19 (21.18)	26 (20.27)	66
Prairie dog habitat (PD)	17 (18.83)	38 (27.59)	12 (20.58)	67
Grassland (G)	5 (5.62)	6 (8.24)	9 (6.14)	20
Total	43	63	47	153

We can calculate the test statistic as

$$\begin{aligned} T &= \frac{(21-18.55)^2}{18.55} + \frac{(19-21.18)^2}{21.18} + \cdots + \frac{(19-6.14)^2}{6.14} \\ &= 14.09. \end{aligned}$$

For these data, $r = 3, k = 3$, so $\text{df} = (3-1)(3-1) = 4$. From χ^2 Table with $\text{df} = 4$, we find that $\chi^2_4(0.01) = 13.28$ and $\chi^2_4(0.001) = 18.47$, and so we have $0.001 < p\text{-value} < 0.01$. Thus, the chi-square test shows that there is significant evidence that the nesting location preferences differed across the 3 years.

Summary of Chi-Square Test of Independence

- Null hypothesis:

$$H_0: \text{ Row variable and column variable are independent.}$$

- Test statistic:

$$T = \sum_{i=1}^{r\times k} \frac{(o_i - e_i)^2}{e_i}.$$

- Null distribution (approximate):

$$\chi^2 \text{ distribution with } \text{df} = (r-1)(k-1),$$

where r is the number of rows and k is the number of columns in the contingency table. This approximation is adequate if $e_i \geq 5$ for every cell.

- Expected frequencies:

$$e = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}.$$

- The observations must be independent of one another.