MATH1. Part II

Probability and Statistics

Chapter 10

Categorical Data: Relationships

10.1 Introduction

Contingency Tables

- The focus of interest in a contingency table is the dependence or association (**relationship**) between the column variable and the row variable.
- **2 x 2 contingency tables**
  - Dependence or association between treatment and response
  - Two by two contingency table: two rows (excluding the "total" row) , two columns.
  - Each category in the contingency table is called a cell (a 2 x 2 contingency table has four cells).
- **r x k contingency tables**
  - a contingency table with r rows and k columns.

Example 10.1.2 HIV Testing

- A random sample of 120 college students found that 9 of the 61 women in the sample had taken an HIV test, compared to 8 of the 59 men.
- Draw a contingency tables for the data.

Table 10.1.3 HIV testing data

	Female	Male
HIV test	9	8
No HIV test	52	51
Total	61	59

10.2 The Chi-Square Test for the 2 x 2 Contingency Table

Null hypothesis

- H_0: the probability of an event E does not depend on whether the first condition, C, is present or the second condition, "not C," is present.

$$H_0 : Pr\{E|C\} = Pr\{E|\text{not } C\}$$

Example 10.2.2 Migraine Headache

- A group of 75 patients were randomly assigned to receive either the real surgery (n = 49) or a sham surgery (n = 26) in which an incision was made but no further procedure was performed.
- Patients experience "a substantial reduction* in migraine headaches" will label as "success."
- What is the H_0 ?

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

10.2 The Chi-Square Test for the 2 x 2 Contingency Table

Null hypothesis

Example 10.2.2 Migraine Headache

- A group of 75 patients were randomly assigned to receive either the real surgery (n = 49) or a sham surgery (n = 26) in which an incision was made but no further procedure was performed.
- Patients experience "a substantial reduction* in migraine headaches" will label as "success."
- What is the H_0 ?
  - H_0 : The real surgery is no better than the sham surgery for reducing migraine headache.
    $\Pr\{\text{Success}|\text{Real}\} \le \Pr\{\text{Success}|\text{Sham}\}$
  - H_A : The real surgery is better than the sham surgery for reducing migraine headache.
    $\Pr\{\text{Success}|\text{Real}\} > \Pr\{\text{Success}|\text{Sham}\}$

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

- H_0 is called **statistical independence** of the row variable and the column variable.
- Variables that are not independent are called dependent or associated.
- Thus, the **chi-square test** is sometimes called a "**test of independence**" or a "test for association."

10.2 The Chi-Square Test for the 2 x 2 Contingency Table

The Chi-square statistic

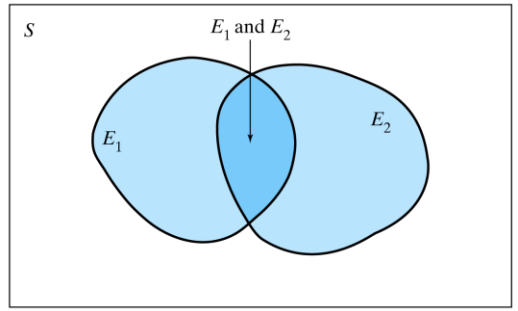
- 2 x 2 contingency tables

$$\chi^2_s = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i}$$

  - In the formula, the sum is taken over all four cells in the contingency table.
  - Each **o** represents an observed frequency, and
  - Each **e** represents the corresponding expected frequency according to H_0 .

- how to calculate the e's
  - calculate the row and column total frequencies (these are called the marginal frequencies)
  - calculate the grand total of all the cell frequencies.
  - **e = (Column total) x (Row total)/Grand total**
    - (Review of Chapter 3) Multiplication Rules:
      - If two events E_1 and E_2 are **independent**,

$$\Pr\{E_1 \text{ and } E_2\} = \Pr\{E_1\} \times \Pr\{E_2\}$$



10.2 The Chi-Square Test for the 2 x 2 Contingency Table

The Chi-square statistic

Example 10.2.2 Migraine Headache (continued)

- A group of 75 patients were randomly assigned to receive either the real surgery (n = 49) or a sham surgery (n = 26) in which an incision was made but no further procedure was performed.
- Patients experience “a substantial reduction* in migraine headaches” will label as “success.”
- What is the e’s?

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

10.2 The Chi-Square Test for the 2 x 2 Contingency Table

The Chi-square statistic

Example 10.2.2 Migraine Headache (continued)

- A group of 75 patients were randomly assigned to receive either the real surgery (n = 49) or a sham surgery (n = 26) in which an incision was made but no further procedure was performed.
- Patients experience “a substantial reduction* in migraine headaches” will label as “success.”
- What is the e’s?
  - e = (Column total) x (Row total)/Grand total
    - Real / Success : 56 x 49 /75= 36.59
    - Sham / Success : 56 x 26 /75= 19.41
    - Real / NO Success : 19 x 49 /75= 12.41
    - Sham / NO Success : 19x 26 /75 = 6.59

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

10.2 The Chi-Square Test for the 2 x 2 Contingency Table

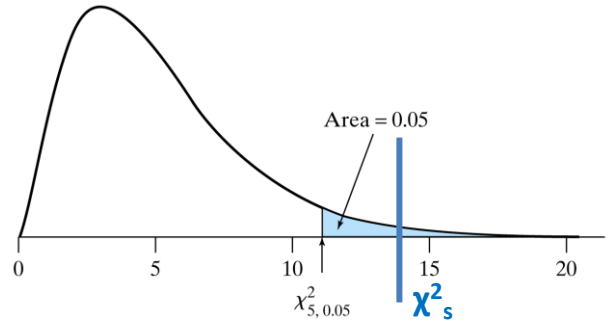
The Test Procedure

- 2 x 2 contingency tables

$$\chi^2_s = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i}$$

  - large values of χ^2_s indicate evidence against H_0 .
  - critical values are determined from **Table 9**.
  - df = 1 (there only is one free cell in the table).
  - alternative hypothesis can be directional or nondirectional.
  - Directional alternatives are handled by the familiar two-step procedure,
  - cutting the nondirectional P-value in half if the data deviate from H_0 in the direction specified by H_A

Figure 9.4.4 The χ^2 distribution with df = 5



$\chi^2_s > \chi^2_{df,\,0.05}$, P-value < 0.05, strong evidence against H_0 and in favor of H_A

10.2 The Chi-Square Test for the 2 x 2 Contingency Table

The Test Procedure

Example 10.2.2 Migraine Headache (continued)

- Is the real surgery better than the sham surgery for reducing migraine headache ($\alpha = 0.01$)?

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

10.2 The Chi-Square Test for the 2 x 2 Contingency Table

The Test Procedure

Example 10.2.2 Migraine Headache (continued)

- Is the real surgery better than the sham surgery for reducing migraine headache ($\alpha = 0.01$)?
  - H_0 : Pr {Success|Real} = Pr {Success|Sham}
  - H_A : Pr {Success|Real} > Pr {Success|Sham}

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

  - To check the directionality of the data
  - $\hat{P}_r${Success|Real} = 41/49 = 0.837
    > $\hat{P}_r$ {Success|Sham} = 15/26 = 0.577
  - Thus, the data do deviate from H_0 in the direction specified by H_A .
  - Then, we proceed to calculate the chi-square statistic from Table 10.2.2 as
  - $\chi^2_s = (41 - 36.59)^2/36.59 + (15 - 19.41)^2/19.41 + (8 - 12.41)^2/12.41 + (11 - 6.59)^2/6.59 = 6.06$
  - From Table 9 with df = 1, $\chi_{1,0.02}^2 = 5.41$ and $\chi_{1,0.01}^2 = 6.63$,
  - we have 0.005 < P-value < 0.01.
  - Thus, we reject H_0 and find that the data provide sufficient evidence to conclude that the real surgery is better than the sham surgery for reducing migraine headache.

10.5 The r x k Contingency Table

Contingency Tables

- **r x k contingency tables**
  - a contingency table with r rows and k columns.
  - The **Chi-square statistic**

$$\chi^2_s = \sum_{\text{all cells}} \frac{(o_i - e_i)^2}{e_i}$$

    - where the sum is over all I = r x k cells of the contingency table.
    - Each o represents an observed frequency, and
    - Each e represents the corresponding expected frequency according to H_0 .
    - df = (r - 1)(k - 1)

  - how to calculate the e's
    - calculate the row and column total frequencies (these are called the marginal frequencies)
    - calculate the grand total of all the cell frequencies.
    - e = (Column total) x (Row total)/Grand total

10.5 The r x k Contingency Table

The Test Procedure

Example 10.5.1 Plover Nesting

- The nesting choices of plover over 3 years
- Sampled 153 plover broods:
  - 66 nests on AF,
  - 67 nests in PD, and
  - 20 nests on G.
- Are the population distributions of nest locations the same in the 3 years?

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

10.5 The r x k Contingency Table

The Test Procedure

Example 10.5.1 Plover Nesting

- The nesting choices of plover over 3 years
- Sampled 153 plover broods:
  - 66 nests on AF,
  - 67 nests in PD, and
  - 20 nests on G.
- Are the population distributions of nest locations the same in the 3 years?

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

- H_0: The population distributions of nest locations are the same in the 3 years.
  - $Pr\{AF|2004\} = Pr\{AF|2005\} = Pr\{AF|2006\}$
  - $Pr\{PD|2004\} = Pr\{PD|2005\} = Pr\{PD|2006\}$
  - $Pr\{G|2004\} = Pr\{G|2005\} = Pr\{G|2006\}$

} **Compound null hypothesis**

- H_A : The population distributions of nest locations are NOT the same in all 3 years.

10.5 The r x k Contingency Table

The Test Procedure

Example 10.5.1 Plover Nesting

- The nesting choices of plover over 3 years
- Sampled 153 plover broods:
  - 66 nests on AF,
  - 67 nests in PD, and
  - 20 nests on G.
- Are the population distributions of nest locations the same in the 3 years?
  - $\chi^2_s = (21 - 18.55)^2/18.55 + (19 - 21.18)^2/21.18 + \ldots + (9 - 6.14)^2/6.14 = 14.09$
  - For these data, r = 3 and k = 3, so df = (3 - 1)(3 - 1) = 4
  - From Table 9 with df = 4, we find that $\chi^2_{4,\,0.01} = 13.28$ and $\chi^2_{4,\,0.001} = 18.47$,
  - so we have 0.001 < P-value < 0.01 (computer software gives a P-value = 0.0070).
  - Thus, the chi-square test shows that there is significant evidence that the nesting location preferences differed across the 3 years.

Table 10.5.1 Plover nest locations across 3 years

Location	Year 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

10.6 Applicability of Methods

Conditions for Validity

A chi-square test is valid under the following conditions:

- 1. Design conditions. For the contingency-table chi-square test, it must be appropriate to view the data in one of the following ways:
  - (a) As two or more independent random samples, observed with respect to a categorical variable; or
  - (b) As one random sample, observed with respect to two categorical variables.
  - For either type of chi-square test, the observations within a sample must be independent of each other.
- 2. Sample size conditions. The sample size must be large enough. The critical values given in Table 9 are only approximately correct for determining the P-value associated with χ_s^2 . As a rule of thumb, the approximation is considered adequate if each expected frequency (e) is at least equal to 5.*

10.6 Applicability of Methods

Conditions for Validity

A chi-square test is valid under the following conditions:

- 3. Form of H_0. A generic form of the null hypothesis for the contingency-table chi-square test may be stated as follows:
  - H_0: The row variable and the column variable are independent.
- 4. Scope of inference. As with other statistical tests, if the data arise from an experiment with random assignment of treatments, as in Example 10.1.1, then we can draw a causal inference; if the experimental units were drawn at random from a population, then we can extend the causal inference to that population. However, if the data arise from an observational study, as in Example 10.1.2, then a small P-value only allows us to infer that the observed association is not due to chance.

10.7 Confidence Interval for Difference Between Probabilities

Construct a confidence interval (p_1 - p_2)

- Consider a 2 x 2 contingency table that can be viewed as a comparison of two samples, of sizes n_1 and n_2.
- We define

$$\tilde{p}_1 = \frac{y_1 + 1}{n_1 + 2}; \quad \tilde{p}_2 = \frac{y_2 + 1}{n_2 + 2}$$

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

- The magnitude of the sampling error can be expressed by the standard error of $(\tilde{P}_1 - \tilde{P}_2)$

$$SE_{(\tilde{P}_1 - \tilde{P}_2)} = \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

- A 95% confidence interval is ($\tilde{p}_1$ - $\tilde{p}_2$) $\pm$ (1.96) $SE_{(\tilde{P}_1 - \tilde{P}_2)}$

10.7 Confidence Interval for Difference Between Probabilities

Construct a confidence interval (p_1 - p_2)

Example 10.7.1 Migraine Headache

- What is the 95% confidence interval of (p_1 - p_2)?

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

10.7 Confidence Interval for Difference Between Probabilities

Construct a confidence interval (p_1 - p_2)

Example 10.7.1 Migraine Headache

- What is the 95% confidence interval of (p_1 - p_2)?
  - The sample sizes are $n_1 = 49$ and $n_2 = 26$
  - The estimated probabilities of substantial reduction in migraines are

$$\tilde{p}_1 = \frac{y_1 + 1}{n_1 + 2} = 42/51 = 0.824; \quad \tilde{p}_2 = \frac{y_2 + 1}{n_2 + 2} = 16/28 = 0.571$$

  - The difference between these is ($\tilde{p}_1$ - $\tilde{p}_2$) = 0.824 - 0.571 = 0.253
  - The standard error is

$$SE_{(\tilde{P}_1 - \tilde{P}_2)} = \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}} = \sqrt{\frac{0.824(0.176)}{51} + \frac{0.571(0.429)}{28}} = 0.1077$$

  - A 95% confidence interval is

$$(\tilde{p}_1 - \tilde{p}_2) \pm (1.96)\, SE_{(\tilde{P}_1 - \tilde{P}_2)} \rightarrow 0.253 \pm (1.96)(0.1077) \rightarrow 0.042 < p_1 - p_2 < 0.464$$

  - 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.

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

10.10 Summary of Chi-Square Test

Summary of Chi-Square Test for a Contingency Table

Null hypothesis:

H_0: Row variable and column variable are independent

Calculation of expected frequencies:

$$e_i = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

Test statistic:

$$\chi_s^2 = \sum_{\text{all cells}} \frac{(o_i - e_i)^2}{e_i}$$

Null distribution (approximate):

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

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. If r and k are large, the condition that $e_i \geq 5$ is less critical and the χ^2 approximation is adequate if the average expected frequency is at least 5, and no expected frequency is less than 1.

The observations must be independent of one another. If paired data are collected for a 2×2 table, then McNemar's test is appropriate (Section 10.8).

Summary

Chapter 10. Categorical Data: Relationships

- 10.1 Introduction
- 10.2 The Chi-Square Test for the 2x2 Contingency Table
- 10.5 The r x k Contingency Table
- 10.6 Applicability of Methods
- 10.7 Confidence Interval for Difference between Probabilities
- 10.10 Summary of Chi-Square Test

Homework

Chapter 10

- 10.2.4
- 10.5.8
- 10.7.3