

Topic 10. Contingency Tables (Ch. 15)

1) Chi-Square Test

- A contingency table is a tabular arrangement of nominal data from multiple populations.

1.1 2x2 tables

- One way to analyze such data is the chi-square test. Suppose one takes a random sample of n units, and then categorizes the units on the basis of 2 (or more) categorical variables. For simplicity, let's consider first a 2x2 table, where there are two categorical variables, each with two possible outcomes.

- As an example, consider a random sample of $n = 793$ people involved in bicycle accidents. The accident report specifies whether or not each person 1) was wearing a helmet and 2) suffered a head injury. Suppose 147 were wearing helmets, and 646 were not. In the group with helmets, 17 (or $p_1 = .116$) had a head injury and 130 (or $1 - p_1 = .884$) did not; whereas in the group without helmets, 218 (or $p_2 = .337$) had a head injury and 428 (or $1 - p_2 = .663$) did not. The data in the form of a 2x2 contingency table are:

		Helmet		Total
		Yes	No	
Head Injury	Yes	17	218	235
	No	130	428	558
Total		147	646	793

(1)

Arranged as proportions, the data are

		Helmet		Total
		Yes	No	
Head Injury	Yes	.116	.337	.296
	No	.884	.663	.704
Total		1.0	1.0	1.0

- An obvious question is whether there is any association between the incidence of head injury and the use of helmets among those involved in bicycle accidents. The generic null hypothesis is

H_0 : variable A is independent of variable B ,

or in the context of this problem;

H_0 : the incidence of head injuries is independent
of (or has no association with) the use of helmets.

- One alternative way to state the equivalent hypothesis (as given in the text) is

$$H_0 : p_1 = p_2 \quad \text{vs.}$$

$$H_A : p_1 \neq p_2$$

where p_1 and p_2 are the proportions of head injuries for those with helmets and those without helmets, respectively. Note that if $p_1 = p_2$, there is no association between the two variables.

- Recall that we had a test for $p_1 = p_2$ in the previous chapter. A chi-square test gives us an alternative way to solve the problem, a method that will generalize to more than 2 categories for one or both of the variables.

- To carry out the test, consider the following nomenclature. Let O_{ij} denote the observed count in row i and column j ; $O_{i\cdot}$ the total in row i for $i = 1, \dots, r$; $O_{\cdot j}$ the total in column j for $j = 1, \dots, c$; and n the grand total of all observations. This gives table

		Column				Total
		1	2	...	c	
Row	1	O_{11}	O_{12}	...	O_{1c}	$O_{1\cdot}$
	2	O_{21}	O_{22}	...	O_{2c}	$O_{2\cdot}$
	\vdots	\vdots	\vdots		\vdots	\vdots
	r	O_{r1}	O_{r2}	...	O_{rc}	$O_{r\cdot}$
	Total	$O_{\cdot 1}$	$O_{\cdot 2}$...	$O_{\cdot c}$	n

(2)

- To calculate the test statistic, one must first calculate the expected counts assuming H_0 is true, e.g. if the use of helmets has no association with the incidence of head injuries. If true, and hence $p_1 = p_2$, then the best estimate of the common injury rate is the total number of head injuries, or $O_{1\cdot} = O_{11} + O_{12}$, divided by the total sample size n . For these data, that is $O_{1\cdot}/n = 235/793 = 0.296$. Recall that $O_{\cdot 1} = 147$ people wore helmets. Therefore, assuming independence, the expected number of people wearing helmets that would sustain a head injury is $O_{\cdot 1} \times (O_{1\cdot}/n)$. Labeling the expected number in row 1 and column 1 as E_{11} , one has

$$E_{11} = O_{\cdot 1} O_{1\cdot} / n.$$

For these data,

$$E_{11} = 235 \times 147 / 793 = 43.6$$

Note that E_{21} , which is the expected number of helmet users not sustaining a head injury is, by similar reasoning,

$$E_{21} = 558 \times 147 / 793 = 103.4.$$

This could also be found by subtraction, i.e.

$$E_{21} = O_{\cdot 1} - E_{11}.$$

In general, one can find the expected counts as

$$E_{ij} = O_i \cdot O_j / n. \quad (3)$$

Using formula (3) for the data in (1), the table of expected counts is

		Helmet		
		Yes	No	Total
Head Injury	Yes	43.6	191.4	235
	No	103.4	454.6	558
	Total	147	646	793

(4)

Note that the column and row totals in (1) and (4) are the same, but in (4) the counts are redistributed to give expected values under H_0 .

• The test statistic is

$$\chi^2_{(r-1)(c-1)} = \sum_{ij} (O_{ij} - E_{ij})^2 / E_{ij} \quad (5)$$

Under H_0 , this is a chi-square statistic with $(r-1)(c-1)$ df provided

- 1) all $E_{ij} > 1$
- 2) no more than 20% of $E_{ij} < 5$.

• The χ^2 is illustrated in Figure 15.1.

• As an example, for the present data, one has

$$\begin{aligned} \chi^2_1 &= \frac{(17 - 43.6)^2}{43.6} + \frac{(130 - 103.4)^2}{103.4} + \frac{(218 - 191.4)^2}{191.4} \\ &\quad + \frac{(428 - 454.6)^2}{454.6} = 28.3. \end{aligned}$$

To determine whether this is large under H_0 , one can find the RR in Table A.8. This is a one-sided test, with critical value

$$\chi^2_{1,0.05} = 3.84.$$

Therefore, for this data, one would reject H_0 , with $p < 0.001$. How would you interpret the result?

• The hypothesis testing framework is

- 1) H_0 : variables are independent
- 2) H_A : variables have some association
- 3) TS $\chi^2_{(r-1)(c-1)}$ in (5)
- 4) RR $\chi^2 > \chi^2_\alpha$
- 5) Calculations. Find E_{ij} in (3), and substitute data into (5).

- For the case of a 2x2 table with small n , the use of the continuity correction improves the approximation. The test statistic with the correction is

$$\chi^2 = \sum_{ij} \left[\frac{|O_{ij} - E_{ij}| - 0.5}{E_{ij}} \right]^2$$

Clearly, it would always reduce the χ^2 statistic. When used with the bicycle helmet data, the new value of the statistic is

$$\chi_1^2 = 27.3.$$

- Another test procedure for 2x2 tables is called Fisher's exact test. It is computationally intensive. Though used in many statistical software packages, we will not develop it.

1.2 $r \times c$ tables

- Consider a table in which r and/or c exceeds 2. This is called an $r \times c$ table. The general table layout is given in (2).
- As an example, consider the text example. There are 575 death certificates which are investigated and classified according to 2 variables. One variable is type of hospital, with outcomes A or B denoting community and university, respectively. The other is death certificate accuracy, with 3 possible outcomes. The data are

		Certificate Status			Total
		Accurate	Incomplete	Needs Change	
Hospital	A	157	18	54	229
	B	268	44	34	346
	Total	425	62	88	575

- The general hypothesis which one could test is
 H_0 : death certificate status is independent of hospital type.
- This hypothesis is sometimes expressed in an equivalent but more technical way. Let p_{ij} denote the proportion of certificates from hospital i with certificate status j . This would give the table.

		Certificate Status			Total
		1	2	3	
Hospital	1	p_{11}	p_{12}	p_{13}	1.0
	2	p_{21}	p_{22}	p_{23}	1.0

The null hypothesis is that the proportions in each certificate status are the same for both hospitals, i.e.

$$H_0 : p_{11} = p_{21}, p_{12} = p_{22} \text{ and } p_{13} = p_{23}$$

H_A : the proportions in some categories are different.

- To test the hypothesis, one would again find the expected counts using (3). For example,

$$E_{11} = 229 \times 425 / 575 = 169.3$$

$$E_{21} = 346 \times 425 / 575 = 255.7$$

The table of expected counts under H_0 is

		Certificate Status			
		1	2	3	
Hospital	A	169.3	24.7	35.0	229
	B	255.7	37.3	53.0	346
		425	62	88	575

- Note that the E_{ij} satisfy the conditions in (6). Therefore, one can calculate the χ^2 as

$$\begin{aligned} \chi^2 &= (157 - 169.3)^2 / 169.3 + \dots + (34 - 53)^2 / 53 \\ &= 24.62. \end{aligned}$$

Because $\chi_{2,.05}^2 = 5.99$, we reject H_0 , indeed $p < .001$. We conclude that there is an association between hospital and death certificate status. Calculating the observed proportions in the table,

		Certificate Status			
		1	2	3	
Hospital	A	.686	.079	.236	1.0
	B	.775	.127	.098	1.0

it is clear that the difference between the 23.6% of certificates at A needing a charge vs. only 9.8% at B is a major contributor to rejecting H_0 .

2) McNemar's Test

- The assumption for the chi-square test is that one has a random sample of n observations. If the data come from matched pairs instead, one would use McNemar's test. We will defer the test for now.

3) Odds Ratio

- The chi-square statistic tests whether there is association between two variables, but does not quantify the strength of the association. A measure of the strength of the association would be useful to determine whether an effect which might be "statistically significant" is also of practical significance.

- One such measure for a 2x2 table is called the odds ratio, OR . It is related to another concept, relative risk (RR) which is discussed in Chapter 6 but which we will not cover.

- Consider the following example to illustrate the concept. Suppose one has 111 mice chosen from a population. Of these 57 were given a bacteria and an antiserum and 54 were given the bacteria only. The survival data for this experiment are given below.

		Antiserum		Total
		Yes	No	
Survival	Yes	44	29	73
	No	13	25	38
	Total	57	54	111

The χ^2 statistic is 6.80, which implies an association between survival and the presence of the antiserum.

- Letting p_1 and p_2 denote the two survival rates, the data could be displayed as

		Antiserum	
		Yes	No
Survival	Yes	$p_1 = 0.772$	$p_2 = 0.537$
	No	$q_1 = 0.228$	$q_2 = 0.463$
	Total	1.0	1.0

One question is how to compare the survival rates, i.e. 0.77 vs. 0.54. One could find the difference, but a difference, e.g. of 23%, depends on the baseline value, e.g. 54%. An alternative way is to calculate the odds ratio. Note that with the serum, the “odds” that a mouse will survive is $p_1/q_1 = 0.77/0.23 = 3.38$ to 1. That is, a mouse is more than three times more likely to survive than to die with the antiserum. The survival rate without the serum is $p_2/q_2 = 1.16$ to 1.

- The odds ratio, OR , is just the ratio of these two odds. An odds ratio of 1 would imply that there is no difference in the odds of the two treatments. In this case, one has the estimate

$$\widehat{OR} = 3.38/1.16 = 2.92.$$

The interpretation of this is that the odds on survival improve almost threefold with the antiserum.

- One can simplify the calculations with a little algebra. Let the data be represented by the numbers a , b , c , and d such that

		Antiserum		Total
		Yes	No	
Survival	Yes	a	b	$a+b$
	No	c	d	$c+d$
	Total	$a+c$	$b+d$	

One can show

$$\widehat{OR} = \frac{ad}{bc},$$

e.g.

$$\widehat{OR} = \frac{(44)(25)}{(13)(29)} = 2.92.$$

- The distribution of \widehat{OR} is skewed, but its log is approximately normal. A confidence interval for log OR is

$$\log \widehat{OR} \pm z_{\alpha/2} s_{\log OR}, \quad (7)$$

where

$$s_{\log OR} = \sqrt{a^{-1} + b^{-1} + c^{-1} + d^{-1}}. \quad (8)$$

(If any of a , b , c , or d are 0, add 0.5 to each). Note that log denotes the natural log.

- A confidence interval on OR is found by taking the antilogs, i.e.

$$(e^L, e^U)$$

where L and U are the lower and upper limits in (7).

- For example, for our data one has

$$\log \widehat{OR} = \log 2.92 = 1.072$$

$$s_{\log OR} = (44^{-1} + 25^{-1} + 13^{-1} + 29^{-1})^{1/2} = 0.417,$$

hence a 95% confidence interval for log OR is

$$1.072 \pm 1.96(0.417)$$

$$= 1.072 \pm 0.817$$

$$= (0.255, 1.889).$$

Therefore, a 95% confidence interval for OR is

$$(e^{0.255}, e^{1.889})$$

$$= (1.29, 6.61).$$

Obviously, this interval does not contain the value 1, hence one may reject the

$$H_0 : OR = 1.$$

We have statistical evidence that the antiserum increased the survival odds, at least by 29% and perhaps as much as six fold.

- It is instructive to consider the example in the book, as the numbers are larger and hence the standard error is smaller. The data concern the relationship between electronic fetal monitoring (EFM) and the incidence of c-section delivery. The data are

		Monitoring	
		Yes	No
C-Section	Yes	358	229
	No	2492	2745

with proportions

		Monitoring	
		Yes	No
C-Section	Yes	.126	.077
	No	.874	.923
		1.0	1.0

- The calculations are

$$\widehat{OR} = \frac{(358)(2745)}{(2492)(229)} = 1.72$$

$$\log \widehat{OR} = 0.542$$

$$s_{\log OR} = (358^{-1} + \dots + 229^{-1})^{1/2} = 0.089$$

95% confidence interval on $\log OR$ is

$$0.542 \pm 1.96(0.089)$$

$$= (0.368, 0.716)$$

95% confidence interval on OR is

$$(e^{0.368}, e^{0.716})$$

$$= (1.44, 2.05).$$

In other words, we are 95% confident that the odds of a c-section are between 1.44 and 2.05 times higher with monitoring as compared to deliveries without monitoring.

4) Berkson's Fallacy (skip)

5) χ^2 Goodness-of-Fit for Multinomial Data (added to book)

- Previously we had the binomial experiment, where an experiment with 2 possible outcomes is replicated n times. Consider an experiment with $k \geq 2$ possible outcomes. It is multinomial provided:

- 1) there are n independent trials,
- 2) each trial results in only one of $k \geq 2$ outcomes,
- 3) the probability of outcome $i, i = 1, \dots, k$ is p_i which is the same for each trial.

Note $\sum p_i = 1$.

- Let O_i denote the number of trials with outcome i . Then

- 1) $\sum O_i = n$ and

- 2) the expected value of O_i is

$$E_i = np_i. \tag{9}$$

- For example, suppose one rolls a fair die $n = 100$ times. The probability of each outcome is

$$p_i = 1/6,$$

and the expected numbers with each outcome are

$$E_i = 100/6 = 16.67.$$

- One question is how one could test some hypothesis about the p_i 's, e.g.

$$H_0 : p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0} \tag{10}$$

An example would be to test that the die is fair

e.g. $H_0 : p_i = 1/6$ for all i ,

from observed data.

- Another important example in genetics is to test phenotype ratios such as 9:3:3:1. This is equivalent to

$$H_0 : p_1 = 9/16, p_2 = 3/16, p_3 = 3/16, p_4 = 1/16.$$

- One may test hypotheses of the form in (10) with a χ^2 statistic with $k - 1$ df, formed as

$$\chi_{k-1}^2 = \sum (O_i - E_i)^2 / E_i$$

where the E_i are given by (9) and must satisfy the χ^2 conditions in (6).

- For example, suppose one has 1301 crossings of heterozygous maize. The results are

Characteristic	Number
green, smooth	773
golden, smooth	231
green striped	238
golden, striped	59

The corresponding expected values are

$$E_1 = 1301(9/16) = 731.9$$

$$E_2 = E_3 = 1301(3/16) = 243.9$$

$$E_4 = 1301/16 = 81.3.$$

Hence the test statistic is

$$\begin{aligned}\chi_3^2 &= \frac{(773-731.9)^2}{731.9} + \dots + \frac{(59-81.3)^2}{81.3} \\ &= 9.25.\end{aligned}$$

Because $\chi_{3,0.01}^2 = 7.81$, one has $p < .01$ and would reject H_0 . How would you interpret the result?

- This χ^2 goodness-of-fit test is one way also to test whether the observed data follow a normal, Poisson, or any other given distribution. There are more powerful tests for normality, but this test is a general one in widespread usage.

Homework: 8, 16, 20