Chapter 10 Class Notes – Categorical Data Analysis

We'll skip §10.4 (Fisher's Exact Test), §10.8 (Paired methods for 2x2 tables), and §10.9 (RR & OR) and cover the rest.

§10.1 Introduction: Here, we consider count data as in the last chapter, but here cross-classified by two or more variables; the data can result from experiments or observational studies.

§10.2. The χ^2_s Test for 2x2 Contingency Tables

In Ex. 10.2.5 on p.371, biologists are again stressing (cotton) plants, now to answer the question: does mite infestation induce resistance to (subsequent) wilt disease? Here are the data:

		Response: Wilt Disease?		
		Yes	No	Total
Treatment (Mites)	Mites	11	15	26
	No Mites	17	4	21
	Total	28	19	47

Before we analyze these data, let's define: a table of counts with for example 2 rows and 4 columns as at the bottom of p.390 is called a 2x4 contingency table (CT) made up of 8 cells. The numbers in the center of the table are joint frequencies and those in the row and column totals are the marginal totals. We are looking to assess whether the two variables are associated or independent. Only in the special case of a 2x2 CT (as above), can we test a directional alternative. For all other tables, the alternative is non-directional.

Now back to the cotton plant example, let p_1 = Pr{WD|Mites} = the probability that a plant gets wilt disease given the Mite treatment

and p_2 = Pr{WD|no Mites} = the probability that a plant gets wilt disease given the No Mite treatment. Here, we have:

H_0: Mites do not induce resistance to wilt, or $p_1 = p_2$

H_A: Mites do induce resistance to wilt, or $p_1 < p_2$

Before we compute the TS, we have: $\hat{p}_1 = \widehat{Pr}\{WD|Mites\} = \frac{11}{26} = 0.4231$ and $\hat{p}_2 = \widehat{Pr}\{WD|no\ Mites\} = \frac{17}{21} = 0.8095$. Let's test the directionality of the data – here, $\hat{p}_1 < \hat{p}_2$, so we proceed with the test:

$$\chi_s^2 = \frac{(11-15.49)^2}{15.49} + \frac{(15-10.51)^2}{10.51} + \frac{(17-12.51)^2}{12.51} + \frac{(4-8.49)^2}{8.49}$$
$$= (-1.141)^2 + (1.385)^2 + (1.269)^2 + (-1.541)^2 = 7.20$$

Here, df = 1, so 0.0005 < p-value < 0.005. Minitab can only be used to perform the two-tailed test, and gives:

Chi-Square Test: WD yes, WD no

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below
expected counts

        WD yes  WD no  Total
    1       11     15     26
         15.49  10.51
         1.301  1.918

    2       17      4     21
         12.51   8.49
         1.611  2.374

Total       28     19     47

Chi-Sq = 7.204, DF = 1, P-Value = 0.007
```

So the correct p-value is 0.007 / 2 = 0.0035. These data suggest that there is sufficient evidence to conclude that mites do indeed induce resistance to wilt disease.

Note that here we use the same χ_s^2 test statistic as in the previous chapter, but here we get the expected values (E_k) by the formula:

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

§10.3. Independence and Association for 2x2 Contingency Tables

2x2 Contingency Tables address two contexts:

- **Two independent samples with a dichotomous observed variable. One example is the following HIV study (p.364, 370):**

	Had HIV test	Didn't have HIV test	Total
Females	9	52	61
Males	8	51	59

- **One sample with two dichotomous variables. From p.414:**

		Hair Color		
		Dark	Light	Total
Eye Color	Dark	726	131	857
	Light	3129	2814	5943
	Total	3855	2945	6800

In the HIV testing example, let

p_F = Pr{had HIV test | Female} and p_M = Pr{had HIV test | Male}

Then $\hat{p}_F = \frac{9}{61} = 0.1475$, and $\hat{p}_M = \frac{8}{59} = 0.1356$.

For the German Hair/Eye example, we have:

- $\widehat{Pr}\{Dark\ Eyes|Dark\ Hair\} = \frac{726}{3855} = 0.1883$
- $\widehat{Pr}\{Dark\ Eyes|Light\ Hair\} = \frac{131}{2945} = 0.0445$
- $\widehat{Pr}\{Dark\ Hair|Dark\ Eyes\} = \frac{726}{857} = 0.8471$
- $\widehat{Pr}\{Dark\ Hair|Light\ Eyes\} = \frac{3129}{5943} = 0.5265$

That the first and second probabilities differ so much tells us that Hair and Eye color are probability not independent (the same conclusion is reached by comparing the 3rd and 4th probabilities).

For this example, the null hypothesis can be expressed either

- **column-wise - H_0: P{D Eyes | D Hair} = P{D Eyes | L Hair}**
- **row-wise - H_0: P{D Hair | D Eyes} = P{D Hair | L Eyes}**

Also, for tests of statistical independence (H_0) versus dependence or association (H_A), we can write in this case:

H_0: Eye color is independent of hair color, or
H_0: Hair color is independent of eye color, or better yet
H_0: Eye color and hair color are independent.

For the Hair/Eye German men data, $\chi_s^2 = 314$, so with a near zero p-value we reject the null (independence) and state that these data suggest that there is sufficient evidence to conclude that dark-haired German men have a greater tendency to be dark-eyed than do light-haired German men Remember to finish the sentence – without the underlined portion, the reader might think than to be light-haired.

10.5. The r×k Contingency Table

The number of rows is r and the number of columns is k. We are only considering two-way contingency tables (two factors or variables) in this chapter, but we are allowing the number of levels to be r ≥ 2 and k ≥ 2 respectively. Here is an r×k example:

pp.432-3 ex.10.49 – Case-control study linking ulcers with blood type: 1,655 peptic ulcer patients and 10,000 controls.

		Patient Status		
		Ulcer	Control	Total
Blood Group	O	911 (55.0%)	4,578 (45.8%)	5,489
	A	579 (35.0%)	4,219 (42.2%)	4,798
	B	124 (7.5%)	890 (8.9%)	1,014
	AB	41 (2.5%)	313 (3.1%)	354
	Total	1,655	10,000	11,655

The Null here states that Blood Group and Patient Status are independent, and the Alternative states that they are dependent or associated. Alternatively, we could state:

H_0: Pr{O | Ulcer} = Pr{O | Control} & Pr{A | Ulcer} = Pr{A | Control} & Pr{B | Ulcer} = Pr{B | Control} & Pr{AB | Ulcer} = Pr{AB | Control}

H_A: The blood distributions are not the same for the two groups

The test statistic is: $\chi_s^2 = \frac{(911-779.4)^2}{779.4} + \cdots + \frac{(313-313.7)^2}{313.7} = 49.0$**.**
Here, df = (4-1) × (2-1) = 3, so p-value < 0.0001 since $\chi_s^2 = 49.0 > 21.11 = \chi^2_{3,0.0001}$**. Thus we reject the null (since p < α = 0.01) and conclude the blood type and patient status are associated (or that the distributions of blood type for the two groups differ).**

The percentages associated with the estimated conditional probabilities are given above, and it looks like the real differences between the two groups occur in the Type O & Type A blood groups.

10.6. Application of Methods

This section addresses requirements/assumptions for χ^2 tests. It is okay to use these tests when:

- **The data results from a random sample regarding two nominal variables**
- **Need independent measurements (unlike flower example on p.393)**
- **The sample size is 'large' – this translates into each one of the expected cell counts must be at least five (each $e_{ij} \geq 5$)**

For contingency tables, the generic null hypothesis is that the row variable and the column variable are independent.

Here is a violation of one of the above requirements:

	Response to Pain Medication – Pain Relief				
	None	Some	Substantial	Complete	Total
Drug A	3	7	10	5	25
Drug B	7	11	5	2	25

Since the 'Pain Relief' response variable here is ordinal – not nominal – the χ^2 test lacks power, and should not be used here; see also http://webpages.math.luc.edu/~tobrien/research/OB2009.pdf

10.7. Confidence Intervals for $p_1 - p_2$

Data are given below related to angina pectoris, a chronic heart condition, and application of either Timolol or Placebo to patients to see if survival rates differ for these two drugs. 160 patients were randomized to the Timolol regime and 147 to Placebo, and the resulting count table is given below (row percentages are given in parentheses). Let p_1 denote the probability that a Timolol patient is angina-free (at the end of the 28-week trial); also, p_2 denotes the probability that a Placebo patient is angina-free.

		Week 28 Angina Status		
		Angina Free	Not Angina Free	Total
Treatment	Timolol	44 (0.2750)	116	160
	Placebo	19 (0.1293)	128	147
	Total	63	244	307

Here, the hypotheses are:

$$H_0: p_1 = p_2 \qquad \text{versus} \qquad H_A: p_1 \neq p_2$$

The chi-square TS here is χ^2_s = 9.978, df = 1, p-value = 0.0016 so the angina-free rates appear to differ for the two drugs.

An (almost) equivalent testing method is to use the *Z-approach for the difference of two independent Binomial proportions*. For these data and using the methods on p.396, we get $\tilde{p}_1 = \frac{44+1}{160+2} = 0.2778$ and $\tilde{p}_1 = \frac{19+1}{147+2} = 0.1342$, and both of these estimate p = p_1 = p_2 when the proportions are equal. The standard error associated with the difference of these estimated proportions is

$$SE_{(\tilde{P}_1-\tilde{P}_2)} = \sqrt{\frac{(0.2778)(0.7222)}{162} + \frac{(0.1342)(0.8658)}{149}} = 0.0449$$

In general, the 95% CI for $(p_1 - p_2)$ is

$$(\tilde{p}_1 - \tilde{p}_2) \pm 1.96 \times SE_{(\tilde{P}_1-\tilde{P}_2)}$$

For the Angina data, the 95% CI is:

(0.2778 - 0.1342) $\pm$ 1.96 $\times$ 0.044926

= 0.1435 $\pm$ 0.0881

= (0.0555 , 0.2316)

Based on these data, we conclude that with 95% confidence Timolol increases the probability of angina-free response by between 5.6% and 23.2% compared to Placebo.

Note: Except in very rare cases, the above χ^2 test will reject $H_0: p_1 = p_2$ and accept $H_A: p_1 \neq p_2$ at the α = 5% level if and only if the 95% CI for $(p_1 - p_2)$ excludes zero (and vice versa).