Chapter 16 – Categorical Data Analysis

Imagine that we walk into our local casino and examine the suits of the 208 cards in a stack of 4 decks of cards and we find 43 ♠, 59 ♣, 60 ♥, and 46 ♦. Is this unusual? What if we walk to a neighboring Genetics lab where white (W) and yellow (Y) summer squash have been crossed to test the 12:3:1 genetic theory, and which produce the data: 1550 W, 400 Y and 100 G (green)? Does this theory apply for these data? Which test can be used to test the underlying hypotheses?

First, note that these are *count data*; in the first case, we have c = 4 categories and in the second case, we have 3. We therefore extend the Binomial distribution to the Multinomial distribution, which has pdf:

$$\mathbf{f(k_1,k_2 \ldots k_c)} = \frac{n!}{k_1!k_2!\ldots k_c!}\pi_1^{k_1}\pi_2^{k_2}\ldots\pi_c^{k_c}$$

In this expression, the π's are the probabilities of falling in the c respective categories (so they are non-negative and sum to one); also, the k's are the numbers falling in the respective categories (so they sum to n). Thus, the null hypotheses for the above examples are H_0: $\pi_1 = \pi_2 = \pi_3 = \pi_4 = ¼$ and H_0: $\pi_1 = 12/16$, $\pi_2 = 3/16$ and $\pi_3 = 1/16$ respectively. It turns out that for these tests the TS has the chi-square (χ^2) distribution with ν df (degrees of freedom), and the pdf is:

$$\mathbf{f(x)} = \frac{1}{\Gamma(\nu/2)2^{\nu/2}}x^{(\nu/2)-1}e^{-x/2}$$

Γ(*) is the usual gamma function, and graphs of the χ^2 pdf are given on p.449 (bottom) for df = ν = 5, 10 and 20. What distribution does this reduce to for df = ν = 2? Also, in a math-stats class, we prove that the χ^2 distribution with df = ν = 1 is identical to the square of the standard normal: $Z^2 = \chi_1^2$. From Table 6 (p.513), note $3.8415 = 1.96^2$. This only works for the first row of this table – i.e., for df = ν = 1.

Whereas the rejection region (RR) for the Normal distribution lies in 2 tails, for the chi-square distribution, the RR is only in the right tail.

Chi-square tests are in very widespread use, so – as for the Normal case – instead of using the above pdf, we use the chi-square table (Table 6) extensively. Students are often under the mistaken impression that there is *one* 'chi-square test' and this is incorrect. In Chap. 16 alone, the relevant TS has a chi-square distribution for both the Goodness of Fit (GOF) tests and for tests associated with contingency tables. Bear in mind that many other applications of chi-square tests lay beyond this introductory course material.

16.3. Testing Goodness of Fit – the relevant TS here is of the form:

$$\chi_s^2 = \sum_{k=1}^{c} \frac{(O_k - E_k)^2}{E_k}$$

where c is the number of categories, O_k and E_k are the observed and expected (assuming H_0) values for the k^{th} category, and the relevant df here is n = c - 1. Returning to our above casino example, note that since H_0 is $\pi_1 = \pi_2 = \pi_3 = \pi_4 = ¼$ (the casino deck is fair), we have: $\nu = 3$; $O_1 = 43$, $O_2 = 59$, $O_3 = 60$, $O_4 = 46$; $E_1 = E_2 = E_3 = E_4 = 52$, so the TS is

$$\chi_s^2 = \frac{(43-52)^2}{52} + \frac{(59-52)^2}{52} + \frac{(60-52)^2}{52} + \frac{(46-52)^2}{52} = 230/52 = 4.4231$$

Since from Table 6 the p-value is between 0.10 and 0.25 (actual p-value is 0.2193), we retain the claim that *the deck is fair*. For the summer squash example, we have: $\nu = 2$; $O_1 = 1550$, $O_2 = 400$, $O_3 = 100$; $E_1 = 1537.5$, $E_2 = 384.375$, $E_3 = 128.125$, so the TS is

$$\chi_s^2 = \frac{(1550-1537.5)^2}{1537.5} + \frac{(400-384.375)^2}{384.375} + \frac{(100-128.125)^2}{128.125} = 6.9106$$

Here, 0.025 < P < 0.05 (actual p-value is 0.0316), so at the 5% α level, we *reject the claim that the genetic theory applies to this variety of summer squash* – perhaps it is a mutant variety.

Note that for GOF tests, the values of the π's are specified by H_0 – that is, neither the π's nor anything else must be estimated. An interesting extension of GOF tests is provided next (see Section 16.6).

Example - Researchers gather the data in the following table related to the number of lifetime offspring produced by 1000 randomly selected F birds of the species *petit oiseau*, and wonder if the Poisson distribution would provide a reasonable model fit to these data.

Number of Offspring	Count	Estimated Probability	Expected Value	Contrib. to the TS
0	249	0.240028	240.028	0.33537
1	328	0.342520	342.520	0.61551
2	254	0.244388	244.388	0.37806
3	105	0.116247	116.247	1.08819
4	49	0.041471	41.471	1.36681
5	11	0.011836	11.836	0.05903
6	3	0.002815	2.815	0.01216
7	1	0.000574	0.574	0.31646

Assume that the Poisson distribution fits these data. We can estimate the mean (λ) with the (weighted) sample mean here – thus, $\hat{\lambda} = 1.427$. Next, let's use this value to obtain the 'estimated probabilities' via the Poisson pmf, then multiply these probabilities by 1000 to get the 'expected values', and finally calculate the TS. Here, $\chi_s^2 = 4.1716$; *what is the relevant 'df' here?* It's not c – 1 = 7 since we first had to estimate λ to perform the above calculations: rather, $\nu = c - 1 - 1 = 6$. Then, the p-value is between 0.25 and 0.90, and we retain the Poisson

fit to these data. This is indeed a variation on the GOF test – we're testing whether the Poisson distribution fits our data but we are first estimating a model parameter (and take account of that estimation by reducing our 'df') – had we estimated two parameters, we would have reduced our 'df' by two, etc. Also, there's nothing special about the Poisson distribution – we could fit any distribution to data and test for GOF, remembering to deduce the 'df' for all estimated parameter(s).

16.4. Testing Association via Contingency Tables

Now, we consider two variables measured on each individual, but – in contrast with Chapter 15 – here we simply count the number of individuals falling into the cells of a "contingency table".

For example, premature babies who suffer from respiratory disease have always been given the conventional 'CMT' treatment. There is now a new 'ECMO' treatment and our goal of the following study is to test whether the survival rates differ for the two treatments. The data are in the following table (ignore the (blue) numbers for now).

		Survival		
		Yes	No	Total
Treatment	ECMO	18 (14.8)	2 (5.2)	20
	CMT	19 (22.2)	11 (7.8)	30
	Total	37	13	50

Denote the true survival probability for the ECMO group by π_1 and the survival probability for the CMT group by π_2. Then, H_0 is $\pi_1 = \pi_2$ or equivalently that the variables 'Treatment' and 'Survival' are *independent*; also, H_A is $\pi_1 \neq \pi_2$ or equivalently that 'Treatment' and 'Survival' are *dependent*. Our sample estimates here are $p_1 = 18/20 = 0.90$ and $p_2 = 19/30 = 0.63$ – is this enough evidence to reject H_0? To perform the significance test, we have two options here: (A) the χ^2 test (Chap. 16, p.457) and (B) the z-test (Chap. 14, pp. 398-9).

A. The χ^2 test involves first finding the expected cell counts (E_{ij}) assuming independence and comparing these values with the observed cell counts (O_{ij}). The expected cell counts are easily obtained by multiplying the row and column totals and dividing by n – they're given above in (blue). We then form the TS:

$$\chi_s^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

This TS has the χ^2 distribution with $\nu = (r-1)\times(c-1)$. In the above 2x2 example, since r = c = 2, $\nu = 1$; also, the TS is:

$$\chi_s^2 = \frac{(18-14.8)^2}{14.8}+\frac{(2-5.2)^2}{5.2}+\frac{(19-22.2)^2}{22.2}+\frac{(11-7.8)^2}{7.8}=4.4352$$

In Minitab:

Chi-Square Test: survive, dont

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

        survive   dont  Total
    1        18      2     20
          14.80   5.20
          0.692  1.969

    2        19     11     30
          22.20   7.80
          0.461  1.313

Total        37     13     50

Chi-Sq = 4.435, DF = 1, P-Value = 0.035
```

Since p = 0.035 < 5%, we reject H_0 and *conclude that Treatment and Survival rate are associated or equivalently that the survival rates differ for the two treatments.*

B. The Z test approach works with the sample proportions ($p_1 = 0.90$ and $p_2 = 0.63$) and also with $p = 37/50 = 0.74$. Note that, assuming H_0 is true (so $\pi_1 = \pi_2 = \pi$), p is our estimate of π. Then the TS is (p. 399):

$$z_s = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.90 - 0.6333}{\sqrt{0.74*0.26*\left(\frac{1}{20}+\frac{1}{30}\right)}} = 2.10599$$

Then, the p-value is $2*Pr\{Z > 2.10599\} = 0.0352$ (same as above). Indeed, we will get the same results since $z_s^2 = 2.10599^2 = 4.435 = \chi_s^2$.

The above (Z) methodology also indicates how we can test for a one-sided alternative hypothesis for a 2x2 contingency table (Section 16.5): just use the Z approach and do not double the probability to get the p-value. Thus, in the above example, if we wanted to test whether the survival rate for the new ECMO treatment is *better* than the survival rate for the old CMT treatment (H_0: $\pi_1 = \pi_2$ versus H_A: $\pi_1 > \pi_2$), then we obtain the same z_s value (2.10599), and the p-value is $Pr\{Z > 2.10599\} = 0.0176$. Alternatively, we could do the χ^2 test (in A) and half the reported p-value to get the one-sided p-value (since the χ^2 test always tests for a *difference* in the alternative hypothesis).

Testing a one-sided alternative hypothesis is possible only in 2x2 tables – not in larger tables (see below). Before we leave 2x2 tables, we point out that important measures reported in biomedicine are the relative risk (RR), π_1/π_2, and the odds ratio (OR); estimation and testing methods for RR and OR are discussed in subsequent courses.

There's nothing sacred about 2x2 tables. For example, researchers have established a relationship between blood type and incidence of ulcers using the data on the next page (Witmer, p. 433). Here, the null hypothesis is that blood type and ulcer incidence are independent, and the alternative is that they are associated. Since calculation of the expected counts and TS is arduous, we use Minitab to obtain χ_s^2.

Blood Type	Ulcer Patients	Controls
O	911	4578
A	579	4219
B	124	890
AB	41	313
Total	1655	10000

Minitab χ^2 test results:

Chi-Square Test: ulcer, control

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

         ulcer  control  Total
    1      911     4578   5489
        779.43  4709.57
        22.208    3.675

    2      579     4219   4798
        681.31  4116.69
        15.364    2.543

    3      124      890   1014
        143.99   870.01
         2.774    0.459

    4       41      313    354
         50.27   303.73
         1.709    0.283

Total     1655    10000  11655

Chi-Sq = 49.016, DF = 3, P-Value = 0.000
```

Note that the TS here has (4-1)*(2-1) = 3 degrees of freedom. Since the p-value is tiny (p < 0.0001), we conclude that *blood group and ulcer status are related*. Here, 'related' is synonymous with 'dependent' or 'associated'. This analysis allows us to go one step beyond and notice that it's not the 'B' and 'AB' blood groups but rather the 'O' and 'A' blood groups that differ for ulcer status.