Chapter 3 Notes

Simple Inference for Categorical Data

Categorical data is based on groups or categories (as the name would imply). Looking at a single categorical variable, we might want to know how many observations there are for each group. We could look at the counts for each group. It may be useful to look at descriptive charts such as pie charts or bar charts to visualize either the absolute number of observations in each group or the proportion of the total number of observations obtained from each group (e.g. what percentage were from group A, from group B, group C, etc.). From these counts or percentages we can infer the proportions in the broader population the data came from.

In this chapter we focus on analyzing counts for **cross-classified** data-- data classified based on multiple categorical variables-- and checking for significant **associations** between categorical variables. Cross-classified data is often given in a contingency table (also called a **cross-tabulation** or **cross-tabs**). Associations will result in higher than expected counts in some **cells** of the table. A **cell** is defined by the values of the categorical variables, and we will usually look at counts or percentages in cells of a contingency table. We mainly will focus on the case of two categorical variables, but categorical data analysis can extend to more than two categorical variables.

As a conceptual example, suppose we are looking at a particular political issue. We might look at political party (e.g. Democratic or Republican) and how an individual feels about that issue (e.g. in favor or opposed). We may then want to know if there is a significant association between party affiliation and stance on the particular issue.

Assuming there is no significant association, we would expect the counts for each combination of categories (a political affiliation and stance on the issue in this case) to be equal to the total number of observations times the product of proportions

$$\boldsymbol{E}(\boldsymbol{N}_{\boldsymbol{Democrat\ in\ Favor}}) = \boldsymbol{n}_{\boldsymbol{total}} \frac{\boldsymbol{n}_{\boldsymbol{Democrat}}}{\boldsymbol{n}_{\boldsymbol{total}}} \frac{\boldsymbol{n}_{\boldsymbol{in\ Favor}}}{\boldsymbol{n}_{\boldsymbol{total}}}$$

where $\boldsymbol{E}$ is the expectation operator, $\boldsymbol{n}_{\boldsymbol{total}}$ is the total number of observation, $\boldsymbol{n}_{\boldsymbol{Democrat}}$ is number of Democrats in the sample, and $\boldsymbol{n}_{\boldsymbol{in\ Favor}}$ is the total number in favor of the issue in the sample. So $\boldsymbol{E}(\boldsymbol{N}_{\boldsymbol{Democrat\ in\ Favor}})$ is the expected value for the count in the cell for **Democrats** with stance **"In Favor"** on the particular issue. This is the proportion of Democrats in the sample (which estimates the probability of being Democrat in the broader population) multiplied by the portion in favor in the sample (which estimates the probability of being in favor in the broader population) times the total number of observations. If the party and stance are independent, we can multiple the marginal probability to estimate the probability of being a Democrat in favor, and the proportion of the total is what we would expect.

To test for a significant association between categorical variables, we would want to test if the actual counts in our contingency table deviate significantly from the expected counts. This idea will be the basis for measures of association (roughly similar in concept to correlation measures for continuous data) and tests for significant association for categorical data.

proc freq

The **proc freq** procedure can be used to create and analyze contingency tables. Data can either be given as raw data or as counts (via the **weight** statement), and we add options to perform various tests of association or dependence, see expected counts, and contributions to chi-squared statistics when we wish to see those results.

Measures and Tests of Association

Association is somewhat similar in concept to correlations we've seen for numeric data (e.g. Pearson correlation) and ordinal data (e.g. Spearman's rank correlation; note that ranks are based on ordering, so we only need for our data to have a meaningful ordering to use it… though we've only used it for numeric data).

Under the null hypothesis of no association, we assume no relationship between the categorical variables. The marginal distributions are thus independent of each other and so the expected count for the $(i,j)^{th}$ cell is the total number of observations $\boldsymbol{n}$ multiplied by the estimated proportions for the i^{th} row category and the j^{th} column category. The estimate for the i^{th} row category proportion is $\boldsymbol{n_{i.}}/\boldsymbol{n}$, where $\boldsymbol{n_{i.}}$ is the i^{th} row sum. The estimate for the j^{th} column category proportion is $\boldsymbol{n_{.j}}/\boldsymbol{n}$, where $\boldsymbol{n_{.j}}$ is the j^{th} column sum. Thus the expected value for the cell count $\boldsymbol{n_{ij}}$ is

$$\boldsymbol{E(n_{ij}) = e_{ij} = n\,\frac{n_{i.}}{n}\frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n}}$$

Deviations from these expectations will be an indication of association between the categorical variables.

For more complete details, we will refer to the **proc freq** documentation in **Details>Statistical Computations>Chi-Square Tests and Statistics**, but we will summarize some of the results here. The results discussed in this section are obtained using the **chisq** option to **proc freq**.

Pearson Chi-Square

We start with Pearson's Chi-square which is the basis for several of the results and is defined as follows

$$\boldsymbol{X^2 = \sum_i \sum_j \frac{(n_{ij} - e_{ij})^2}{e_{ij}}}$$

which asymptotically has a chi-square distribution with $(I-1)(J-1)$ degrees of freedom where I and J, respectively, are the number of rows and columns in the table. So for a 2x2 table, the chi-square statistic would have 1 degree of freedom.

The test will be one-sided. Large values of $\boldsymbol{X^2}$ will provide evidence that the categorical variables are not independent and we would conclude there is an association.

This is a large sample result, so it may not be a good approximation for small samples. The general rule of thumb is that as long as the individual expected cell counts are greater than 5, the approximation is OK.

Likelihood-Ratio Chi-Square Test (G^2 test)

This test statistic is based on maximized likelihoods. **Likelihood ratio** statistics are a fairly general class of statistics which also approach a chi-square distribution as the sample size gets large.

First recall that the **likelihood** given some data set $x_1, x_2, \dots, x_n$ and some distribution (or model) with unknown parameters $\theta_1, \theta_2, \dots, \theta_p$ is

$$L(\theta_1, \theta_2, \dots, \theta_p | x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n | \theta_1, \theta_2, \dots, \theta_p)$$

Where $f(x_1, x_2, \dots, x_n | \theta_1, \theta_2, \dots, \theta_p)$ is the pdf (or pmf if the distribution is discrete) for the assumed distribution (or model), so we know the data but don't know the parameters. The parameter values that maximize the likelihood given the data are the **maximum likelihood estimates**.

Maximum likelihood estimates have lots of nice properties, and if we take a ratio of likelihoods for two competing models in general we can get a test statistic known to be asymptotically chi-square distributed (under some fairly broad conditions). The general result is that

$$-2\ln\left(\frac{\textbf{maximized likelihood under the null model}}{\textbf{maximized likelihood under the alternative model}}\right)$$

is asymptotically chi-square distributed, and if its value is large that will provide evidence against the null model and in favor of the alternative.

For the specific case of two dimensional contingency tables where our null hypothesis is independence of the two categorical variables and the alternative is that they are not independent, the likelihood-ratio statistic becomes

$$\boldsymbol{G^2 = 2\sum_i \sum_j n_{ij} ln\left(\frac{n_{ij}}{e_{ij}}\right)}$$

The asymptotic distribution is again chi-square with $(I-1)(J-1)$ degrees of freedom. A very crude rule of thumb here is that the total count divided by the number of cells should be greater than 5.

People are still studying properties of these two statistics under various conditions to see which better matches a chi-square distribution when. Generally speaking though, Pearson and the likelihood ratio test will tend to agree for reasonably large samples (and reasonably large can be pretty small as we can see from the crude rules of thumb) and for small samples we can use small sample methods like Fisher's exact test.

Chi-Square Based Measures of Association

The **phi coefficient**, **contingency coefficient**, and **Cramér's V** are chi-square based measures of association. We can think of these as analogues of correlations.

Phi and **V** have the nice property of going from -1 to 1 for 2x2 tables. In all other cases, these measures are non-negative (direction specified by + or – doesn't really make sense if we do not have a defined order, but magnitude of association still can). **V** is always bounded above by 1. The upper bounds for the other two measures depend on the dimensions of the table.

Mantel-Haenszel Chi-Square Test

While the other two tests are only interested in finding indications of association, this test is interested in existence of a linear trend as we move across the rows and down the columns. For this to be meaningful, our categories need to be ordinal and they need to be given in order (e.g. age groups listed from youngest to oldest, level of agreement ordered from strongly disagree to strongly agree, etc.)

Mantel-Haenszel asymptotically follows a chi-square distribution with 1 degree of freedom. The statistic is based on a correlation coefficient for contingency tables (Pearson's by default). This correlation is effectively looking at the degree to which counts increase or decrease as we go across and up/down the table, so the impact of moving through the table is akin to the impact of increasing/decreasing x and y variables for continuous data (e.g. do counts tend to increase or decrease as both ordinal variables increase, and how strong is that relationship).

Fisher's Exact Test

While it's not a chi-square based test, Fisher's exact test is included with the **chisq** option for 2x2 tables and is an exact alternative to chi-square tests. For small samples when asymptotic chi-square results may not be trustworthy, we can use this exact test.

Fisher's exact test for 2x2 tables follows directly from the hypergeometric distribution. Given that we know the total number of observations, and the number in each row and column, the count $\boldsymbol{n_{11}}$ in the upper left cell determines the other cell counts. Under the null hypothesis, this count follows a hypergeometric distribution exactly (hence the name exact test…).

To discuss the hypotheses being tested (particularly the one-sided alternatives), it may help to construct a general 2x2 table and define the concept of **risk**. Remember that our table is as follows.

$\boldsymbol{n_{11}}$	$\boldsymbol{n_{12}}$	$\boldsymbol{n_{1.}}$
$\boldsymbol{n_{21}}$	$\boldsymbol{n_{22}}$	$\boldsymbol{n_{2.}}$
$\boldsymbol{n_{.1}}$	$\boldsymbol{n_{.2}}$	$\boldsymbol{n}$

The risks are just marginal proportions. The risk of column j for row i is the count in cell (i, j) divided by the sum of counts for row i. The risk of row i for column j is the count in cell (i, j) divided by the sum of counts for column j. Thus, $\boldsymbol{n_{11}/n_{.1}}$ is the row 1 risk for column 1 and $\boldsymbol{n_{11}/n_{1.}}$ is the column 1 risk for row 1.

The test statistic for Fisher's exact test (in the 2x2 case) is just $\boldsymbol{n_{11}}$. The null hypothesis is still independence of the two categorical variables.

The left-tailed alternative would be that the row 1 risk for column 1 and the column 1 risk for row 1 are both *smaller* than expected under independence. Small values of $\boldsymbol{n_{11}}$ would provide evidence against the null and in favor of this alternative, and the p-value of interest is $\boldsymbol{P(\mathbf{F} \leq n_{11})}$.

A right-tailed alternative would be that the row 1 risk for column 1 and the column 1 risk for row 1 are both *greater* than expected under independence. Large values of $\boldsymbol{n_{11}}$ would provide evidence against the null and in favor of this alternative, and the p-value of interest is $\boldsymbol{P(\mathbf{F} \geq n_{11})}$.

Note that because the row and column sums can't change, if $\boldsymbol{n_{11}}$ is larger than expect, $\boldsymbol{n_{22}}$ will also be larger than expected and $\boldsymbol{n_{12}}$ and $\boldsymbol{n_{21}}$ will be smaller than expected. There is a similar relationship if $\boldsymbol{n_{11}}$ is smaller than expected.

A two-sided alternative would be that there is some sort of association (the probability of being in cell (1,1) is significantly greater or significantly less than would have been expected under independence). In the 2x2 case, the p-value would be the sum of the probabilities for all possible cell (1,1) count values which are less likely than the $\boldsymbol{n_{11}}$ value we actually observed.

The two-sided p-value will give the probability of obtaining a table at least as unlikely as the one we observed. This allows for an extension to RxC tables with R>2 or C>2. Given the row sums and column sums, it would be possible to list out all the possible tables that could occur and the probability with which they would be expected to occur (you wouldn't want to do that by hand, but it could be done…). So, given that we observed a particular RxC table and we know the row and column sums, we (or more likely the software we're using) can sum up the probabilities for the tables which are at least as unlikely as the one we observed.

Confidence Intervals for Risk and Risk Difference

From each cell in a contingency table and the associated row sum we can obtain a risk estimate (just divide the cell count by the row sum). We could do the same for columns. These estimates are estimated proportions, and each is a proportion parameter from a binomial (if we have more than 2 possibilities, we would be dealing with multinomials, but let's stick with binomials).

The binomial distribution goes to a normal as the number of trials gets large, so we can get approximate confidence intervals for the risks (which are just binomial proportions) based on a normal approximation. We could also use exact formulas for binomials to construct the intervals. This is similar in concept to obtaining confidence intervals for a mean like we did with t tests, but now we're obtaining a confidence interval for a proportion instead (so the distributional properties come from a binomial instead).

Just like we considered location differences for two populations (e.g. difference between north and south mortality rates) and obtained confidence intervals for those differences, we could consider differences of risks. For instance, for a chosen row we could look at the difference of risks for two

columns to see if there is a significant difference between the column risks for the group represented by that row value. So returning to the idea of political parties and political stance, we could look at the difference of proportions of Republicans in favor and Democrats in favor to see if Republicans have a significantly higher probability of being in favor than Democrats do.

Example

For the grouped **water** data set, we will construct the frequency table with rows being the mortality groups and columns the calcium concentration groups. For the low calcium group column, the sample proportion of low mortality observations which had low calcium concentrations will be the estimate of the risk of having low water calcium given that the mortality rate is low. The sample proportion of high mortality observations which had low calcium concentrations will be the estimate of the risk of having low water calcium given that the mortality rate is high.

We get confidence intervals for these two proportions either by using the asymptotic approximation of binomials as normal, or by using the interval (labeled as exact in the SAS table) based on the binomial distribution. The approximate result will typically be fine if samples aren't too small, and we can obtain intervals for risk differences based on asymptotic approximation. Risk difference intervals containing 0 will tell us that the risk is not significantly different for the two row categories. A completely negative interval would say that the second row category has a significantly higher risk of being in that column category. A completely positive interval would say that the first row category is at greater risk of being in that column category.

SAS will give us the risk estimates and intervals for each column (so estimates and intervals for being in a particular calcium concentration group given the mortality groups as our tables were set up). Switching the rows and columns will allow us to estimate the other risks and differences.

Exercises

Sandflies data

For the sandflies data set we could do the same type of analysis.

- Obtain row and column percentages.
- Comment on significance of association.
- Do there appear to be any differences within rows or columns?
- Obtain confidence intervals for row and column risks and comment on significant differences.

Acacia Ants data

For this data set (the **ants** data set from the text), we want to know if there is a greater risk of ants invading one species of acacia tree. We have two species of trees and counts for trees of each species that were invaded by ants.

- Obtain a contingency table including expected counts. Does independence seem reasonable based on the counts? If not, what type of relationship might we expect based on the counts?
- Test for association to check your intuition from the previous part.
- Obtain risk estimates to see if there are significant differences.

Oral Cancer data

This example is a table where we have more than 2 rows and more than 2 columns. We will use the code for defining the **lesions** data set given in Chapter 3 of the text. Note that Fisher's exact test is only generated by default for 2x2 tables. We need to use the **exact** (or **fisher**) option if we want that result for larger tables. In this data we have geographic regions of India (3 regions) and types of oral cancer (9 types).

- Get a contingency table with just frequencies.
- Do counts suggest association?
- Obtain measures and test results for association.
- Do measures suggest association?
- Do tests reject a hypothesis of independence?
- Which results should be trust and why?