PLSC 502: "Statistical Methods for Political Research"

Measures of Association: Nominal Variables
October 27, 2016

Introduction

Today we'll discuss inference for nominal-level variates. After talking about contingency tables, we'll introduce the chi-square statistic as a general way of conducting inference on nominal-level variates. We'll briefly describe inference for one-way tables, and spend the rest of the time going over chi-square tests for the statistical independence of two nominal-level variates. At the end, we'll very briefly mention a few alternatives to the chi-square approach.

One-Way and Two-Way Crosstabs

Crosstabs (more correctly referred to as *crosstables*, or *contingency tables*) are tabular representations of data. We discussed frequency tables and crosstabs briefly when we reviewed summary statistics, but a review is probably in order.

One-Way Frequency Tables

A *one-way* (or *frequency*) table is simply a table listing the categories of Y and the number of observations in each of those categories. Frequency tables also often contain category (cell) proportions or percentages; if n_y is the observed frequency of observations in Y's category $Y = y$, then the category proportion is just

$$P_y = \frac{n_y}{N}.$$

Thus, for example, in our Africa data, we have:

Category	Frequency	Proportion
No Civil War	30	0.70
Civil War	13	0.30
Total	43	1.00

Two-Way Crosstables

Crosstabs are cross-classified frequency tables. We typically define the main variable of interest Y and place it on the "Y" (vertical) side of the table, while the covariate ("independent," or "X") variable's categories are listed on the horizontal axis of the table. Each cell is then defined as n_{yx}, the number of observations in the data for which $Y = y$ *and* $X = x$.

In addition to the cell frequencies, we can calculate a number of other proportions in two-way tables:

- *Row proportions* (or percentages) are the proportion of observations in that row of the table (that is, with $Y = y$) falling into the column defined by $X = x$. They sum to 1.0 across columns.
- *Column proportions* (or percentages) are the proportion of observations in that column of the table (that is, with $X = x$) falling into the row defined by $Y = y$. They sum to 1.0 down rows.
- *Cell proportions* (or percentages) are the proportion of the total number of observations in that cell of the table. They sum to 1.0 overall columns and rows (cells).

Of these, we are usually most interested in column proportions, since (as we'll discuss) they allow us the most intuitive examination of whether Y and X are *independent*. More on this below.

A typical two-way table might look like this:

	Sub-Saharan?		
Civil War?	**No**	**Yes**	**Total**
No	5	25	30
(Row)	(0.17)	(0.83)	(1.00)
[Column]	[0.83]	[0.68]	[0.70]
{Cell}	{0.12}	{0.58}	{0.70}
Yes	1	12	13
(Row)	(0.08)	(0.92)	(1.00)
[Column]	[0.17]	[0.32]	[0.30]
{Cell}	{0.02}	{0.28}	{0.30}
Total	6	37	43
	(0.14)	(0.86)	(1.00)
	[1.00]	[1.00]	[1.00]
	{0.14}	{0.86}	{1.00}

Note that we'd *never* actually put all three types of proportions in a table – typically we focus on the column proportions. Moreover, many software packages (e.g., Stata) express the cell proportions in terms of percentages rather than proportions.

Statistical Independence

If we believe there is a relationship between two (here, nominal) variables (say, Y and X), the direct implication is that the distribution of Y is different for different values of X. Formally, then, we're interested in $f(Y|X)$, the distribution (say, density, or whatever) of outcomes on Y once we "condition" on values of X. For example, in the table above, the conditional distribution of civil wars given that a country is in sub-Saharan Africa is 12 countries with

wars and 25 without them (for a total of 37). This is in contrast to the unconditional distribution of Y, $f(Y)$, which is the distribution of Y for *all* values of X (above, that's 13 wars and 30 non-wars).

Considered in this way, there's an obvious "null hypothesis": that the distribution of Y is the same across all values of X. We write this:

$$H_0 : f(Y|X) = f(Y). \tag{1}$$

In other words, if two variables are unrelated, then the conditional distribution of each on the other is the same as its unconditional ("marginal") distribution. If H_0 is true, we say X and Y are statistically *independent*.

In practice, and particularly when examining nominal-level variables, there will almost inevitably be *some* departure from strict independence; that is, the conditional distributions of Y will almost never be exactly equal to its marginal distribution. The relevant question is whether those differences are sufficiently small that we can/should attribute them to sampling error, or whether they are reflective of a "real" difference in the population. Knowing something about the sampling error of our variables allows us to make this assessment.

The Chi-Square Statistic

Consider the cells of a one- or two-way frequency table containing a total of N observations on two nominal-level variables Y and X. Define k_Y and k_X as the number of different categories of Y and X, respectively. As above, let n_{yx} be the observed frequency of observations in the cell corresponding to category y of ("dependent") variable Y and, if it is present, category x of ("independent") variable X. The "marginals" of Y and X are defined as

$$R_y = \sum_{k_X} n_{yx}$$

and

$$C_x = \sum_{k_Y} n_{yx},$$

respectively. That is, the total number of observations in category y is the sum of all observations with $Y = y$ across all the categories (columns) of X, and the total number of observations in category x is the (row) sum of all observations with $X = x$.

For a particular cell defined by y (and x), the *expected* number of observations in that cell under the assumption that Y and X are independent is thus equal to:

$$E_{yx} = \frac{R_y \times C_x}{N}. \tag{2}$$

In the case of a one-way table, this reduces to simply $N \times \pi$, where π is $1/k_Y$, the proportion defined as the reciprocal of the number of categories.

Under the "null" hypothesis of the independence of Y and X, we would expect two things:

1. On average $n_{yx} = E_{yx}$. That is, we should expect our cell count to be equal to the expected number of observations in that cell, a defined by the marginals for each variable.

2. The difference between n_{yx} and E_{yx} should be small.

These two points lead directly to the *chi-square test* for the independence of two nominal variates. The test statistic is defined as:

$$\chi^2 = \sum_{k_Y k_x} \frac{(n_{yx} - E_{yx})^2}{E_{yx}}. \tag{3}$$

Under the null hypothesis, this test statistic has a sampling distribution that is chi-squared with degrees of freedom equal to $(k_Y - 1)(k_X - 1)$.[1] The reason this is chi-square is straightforward. By the Law of Large Numbers, we would expect that, under the null hypothesis of independence,

$$n_{yx} - E_{yx} \sim \mathcal{N}(0, \sigma^2_E) \tag{4}$$

where σ^2_E is the sampling variability of the difference between n_{yx} and E_{yx} and is roughly proportional to $\sqrt{E}$. Thus, each squared difference, divided by the expected cell frequency, is χ^2_1, and the sum of $(k_Y - 1) \times (k_X - 1)$ independent χ^2_1 variates is $\chi^2_{(k_Y-1)\times(k_X-1)}$.

If one were to do a chi-square test "by hand," then, it would proceed in four steps:

1. Calculate E_{yx} for each cell,

2. Calculate $\frac{(n_{yx}-E_{yx})^2}{E_{yx}}$ for each cell,

3. Sum these values across all cells, and

4. Compare the resulting statistic to a chi-square distribution with $(k_Y - 1) \times (k_X - 1)$ degrees of freedom

Of course, we have computers for such things these days...

Large values of χ^2 are evidence against the null hypothesis. In addition, the (normed) differences between observed and expected cell counts are often referred to as *Pearson residuals*; these will turn out to be useful later on.

[1] The reason for this number of degrees of freedom is straightforward: it represents the amount of "free" variation in the data once the marginals are accounted for.

A Few Pointers...

1. Note that while the chi-square test as it is implemented in nearly every default I know tests for the independence of two variables (or, alternatively, for the equiprobability of each of the k_Y possible outcomes), it's possible to plug in any values of E_{yx} you care to, so long as they sum to N. This means that one can test for things that are not (necessarily) based upon the data marginals; one common hypothesis, for example, is that all cross-categories are equally likely (that is, that $E_{yx} = \frac{N}{k_Y k_X \, \forall \, x,y}$).

2. A good rule of thumb is that, if your chi-squared statistic is equal to or less than the number of degrees of freedom, you will fail to reject the null at any really viable level of significance.

3. In instances where there are relatively "sparse" data (that is, where there are more than one or two instances where $E_{yx} < 5$), the chi-square test is not recommended; see below.

Other Alternatives: Fisher's "Exact" Test

The chi-square test requires the Law of Large Numbers to be operative in order to work properly. In situations where one or more values of E_{yx} are less than five (and particularly if $E_{yx} < 1$), the chi-square distribution will be a very poor fit to the actual distribution of the test statistic; in those cases, we're not close enough to "asymptopia" . In such circumstances, it is unwise to use a chi-square test, and instead we typically rely on an alternative test based on combinatorics, known as *Fisher's Exact Test*.

The formula for Fisher's test in the case of a $k_y \times k_X$ contingency table is:

$$P = \frac{(R_1!R_2!...R_{k_Y}!)(C_1!C_2!...C_{k_X}!)}{N! \prod_{k_Y,k_X} n_{yx}!}. \tag{5}$$

Here, R, C, and n denote the row, column, and cell frequencies, respectively, and N is once again the total number of observations. The logic behind (5) (which is in fact a multivariate generalization of the probability function for a hypergeometric distribution) is that the denominator represents the possible ways in which one could arrange the data on N observations in a $k_y \times k_X$ contingency table, while the numerator reflects the possible orderings with the marginals determined by the values of R and C (the marginals).

Fisher developed his test for 2×2 tables; it is computationally challenging for larger tables. Moreover, for tables where cell counts are all sufficiently large (that is, with $E_{yx} > 5$ or so for all cells), it is asymptotically the same as a (much easier to calculate) chi-square test. However, it is superior in instances where we have relatively small cell frequencies.

An Example: Feminism as an Insult

In the aforementioned September 1997 CBS/NYT Poll (the one with the questions on the designated hitter, with $N \approx 1000$), the pollsters asked a somewhat different question as well:

> Do you consider calling someone a feminist to be a compliment, an insult, or a neutral description?

The first was coded "1," the second "2," and the third "3." The slides illustrate three different examples of using frequency and contingency tables, and chi-square tests, on these data.