Regression Analysis for the Social Sciences

The Association Between Categorical Variables

July 18, 2011

Often in the social sciences we are interested in relationships among categorical variables. Categorical variables consist of several—often unordered—discrete categories. Examples of categorical variables include marital status (never married, married, divorced, or widowed) or race (black, white, Asian, native American, and Inuit). Today we examine how to study associations between variables that may have more than two categories.

A classic area of application is the study of social mobility. In this case we are interested in how socio-economic characteristics of parents are related to the socio-economic characteristics of children. Social mobility research is motivated to assess the degree to which inequality is reproduced across generations.

Introducing Contingency Tables

Table 1 reproduces a classic *contingency table* from a pioneering work on *Social Mobility in Industrial Society*, by Seymour Martin Lipset and Reinhard Bendix. Contingency tables show the frequency or percentage distribution of two categorical variables. The tables are also sometimes called, cross-classifications, cross-tabulations, or crosstabs, for short. This table is called a two-way table, because it shows the relationship between two variables—father's occupation and son's occupation. We could also call this a 4×4 table, because each variable in the table consists of four categories.

Lipset and Bendix examine mobility from fathers to sons using a four-category measure of occupation: (1) professionals and managerial, (2) clerical and sales, (3) manual, and (4) farm. Here, we're looking at a mobility table for the urban residents of Indianapolis. Each of the cell entries in this table is a frequency from the survey of Indianapolis men. So, for example, 1791 men in the survey reported that their fathers were professionals or managers.

Table 1. Inter-generational occupational mobility in Indianapolis, 1940, Lipset and Bendix (1956, 31).

Son's	Father's Occupation			
Occupation	I	II	III	IV
I. Professional, managerial	591	229	537	180
II. Clerical and sales	519	459	913	245
III. Manual	681	404	3868	1145
IV. Farm	0	0	54	66
Total	1791	1092	5372	1635

Table 2. Percentage distribution of son's occupation by father's occupation, Indianapolis, 1940, Lipset and Bendix (1956, 31).

Son's	Father's Occupation			
Occupation	I	II	III	IV
I. Professional, managerial	33	21	10	11
II. Clerical and sales	29	42	17	15
III. Manual	38	37	72	70
IV. Farm	0	0	1	4
Total	100	100	100	100
N	1791	1092	5372	1635

Presented in this way, the table is not terribly informative. Ideally we would want to be able to read the probability of having a professional occupation, given that your father had a professional occupation. How could such *conditional probabilities* be written? To write the son's probability of being in a certain occupation, conditional on the father's occupation, we percentage down the columns of the table. This percentaged version of the table is shown in Table 2.

Once the table is written this way, there are two striking results. First, there is remarkably little mobility out of the manual occupations. In the 1940s in urban America, if you were a man who was a manual worker, your son would be very likely to be a manual worker, too. In fact, we can say that the probability of being a manual worker is 72%. Second, farmers are different. There is virtually no mobility into farming (not surprising for an urban sample), and most of those who leave a farm life, go into a manual occupation. In this case, if your father worked on a farm, there is a 70%

chance that you would arrive in a manual occupation.

Also note that the final row of the table shows the raw frequencies from the survey data. Reporting the raw frequencies allows us to reproduce the frequencies for all the cells of the table. The row of frequencies at the bottom of the table is also called the *marginal distribution*. The marginal distribution of father's occupation shows that most Indianapolis families originated from manual occupations. In a two-way table there are two marginal distributions. In this example, we have the marginal distribution for father's occupation and the marginal distribution for son's occupation. (What is the marginal distribution for sons?)

There are several more general lessons we can take away from this example. Frequencies of contingency tables should always be reported so the raw frequencies can be reconstructed. In forming percentages for a contingency table, you should always percentage across the categories of the dependent variable. In the mobility example, we were interested in the probability of being in a certain occupation, that given that your father was in a given occupation. Father's occupation was thus the explanatory variable, trying to explain son's occupation. With son's occupation as the dependent variable, we formed percentages across the categories of that variable.

Inference for Contingency Tables

Typically in the analysis of contingency tables we are interested to see if there is an association between the two variables of the table. We can say that there is an *association* between a predictor and a dependent variable if the conditional distribution of the dependent variable changes with values of the predictor.

Figure 1 shows a barchart that describes the percentage distribution of son's occupation, given fathers occupation. In this figure, just a portion of the full mobility table is shown. We see the percentage distribution of son's occupation given that father is either in occupation I (a professional or managerial occupation) or occupation III (a manual occupation). Clearly the two distributions are shaped quite differently. In particular, the likelihood of entering occupation I is much lower if your father was a manual worker, and your likelihood of entering occupation III is relatively low if your father had a professional or managerial occupation. In this case, because the conditional distribution of the dependent variable, changes across values of the independent variable, we can say that there is an association between the

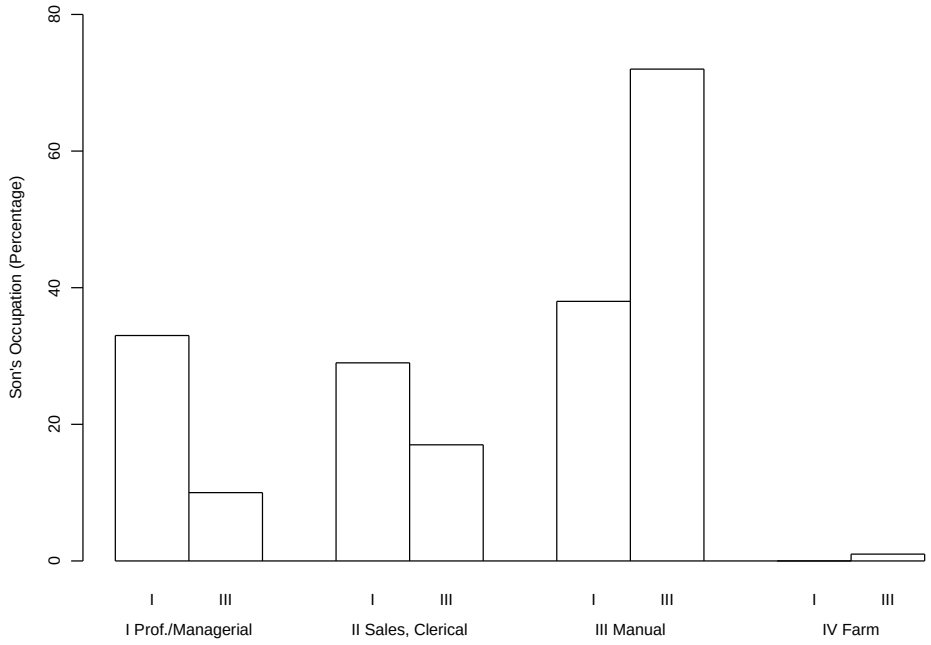

Figure 1. Barchart showing the percentage distribution of son's occupation (I–IV), for respondents with fathers in occupations I (professional/managerial) and III (manual), Indianapolis, 1940.

predictor and the dependent variable.

How do we test the hypothesis that there is an association between two variables of a contingency table? First we must form a null hypothesis that describes what we would observe if the hypothesis were true. In this case, if there was no association between our predictor and our dependent variable (the two variables are *independent*), the conditional distribution of the dependent variable will be relatively similar across all values of the predictor.

Let's introduce another example to examine the inference for a contingency table. For this example we are analyzing an extract of the US Current Population Survey, the large monthly labor force survey used to measure unemployment and demographic characteristics of the noninstitutional population. In this example, we're looking at two categorical variables, race/ethnicity and marital status.

```
> tab0 <- table(x$reth, x$mstat)
> tab0

                    Divorced/separated Married Never married Widowed
  Black                            821    1199          1533      99
  Hispanic                         527    1541          1079      30
  Non-Hispanic White              1402    3263          2500     121
  Other                            140     575           318      19
> round(prop.table(tab0,1),3)

                    Divorced/separated Married Never married Widowed
  Black                          0.225   0.328         0.420   0.027
  Hispanic                       0.166   0.485         0.340   0.009
  Non-Hispanic White             0.192   0.448         0.343   0.017
  Other                          0.133   0.547         0.302   0.018
```

The output looks rather unpleasant and a nicer presentation is shown in Table 3. The table shows that African Americans have a relatively high probability of being divorced or never married. Hispanics are relatively unlikely to be divorced and more likely to be married.

Is there a statistically significant association between race and marital status? If race and marital status were independent we would expect that the distribution of marital status would look the same for each racial group, and would follow the marginal distribution of marital status. Under this null hypothesis—that race and marital status are independent—we can construct a table of *expected frequencies*. These expected frequencies are those we would

Table 3. Marital status by race and ethnicity, 1999 US Current Population Survey.

	Marital Status					
	Divorced/ Separated	Married	Never married	Widowed	Total	N
Black	22.5	32.8	42.0	2.7	100	3652
Hispanic	16.6	48.5	34.0	0.9	100	3177
Non-Hispanic White	19.2	44.8	34.3	1.7	100	7286
Other	13.3	54.7	30.2	1.8	100	1052
All	19.1	43.4	35.8	1.8	100	

Table 4. Observed and expected frequencies of the race by marital status, 1999 Current Population Survey.

	Divorced/ separated	Married	Never Married	Widowed
Expected frequencies under null hypotheses				
Black	695.9	1583.9	1307.4675	64.8
Hispanic	605.4	1377.9	1137.4108	56.4
Non-Hispanic White	1388.3	3160.0	2608.4908	129.3
Other	200.5	456.3	376.6308	18.7
Expected frequencies under null hypotheses				
Black	821	1199	1533	99
Hispanic	527	1541	1079	30
Non-Hispanic White	1402	3263	2500	121
Other	140	575	318	19

expect to see, if the null were true. The first two panels of Table 4 report the observed and expected frequencies. For the first cell in the table, divorced blacks, we have 3652 black respondents, and 19.05% of *all* respondents are divorced or separated. Multiplying $3652 \times .1905$ yields the expected count, 695.9. An easy formula says that the expected frequency under the null hypothesis for a particular cell is given by the row count times column count divided N.

With observed, f_o, and expected frequencies, f_e, we can construct a test statistic. We can use these observed and expected frequencies to construct a test statistic. The statistic is called a "chi-square statistic" and it is often denoted in Greek, χ^2. Like our earlier test statistics, the χ^2 is formed by comparing the observed value of the statistic (the cell counts in this case) to the expected value under the null:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}.$$

We sum over all the cells in the table, so in the marital status example we sum 16 numbers in our 4×4 table. Notice how the χ^2 statistic behaves. When the null is true, the observed and expected frequencies will be close together and χ^2 will tend to be small. If the data are inconsistent with the null, the difference between the observed and the expected frequencies will be large, and χ^2 will be large. As with our one and two samples tests, a large test statistic will lead us to reject the null.

If the null hypothesis is true, the χ^2 statistic will a have special probability distribution, called a chi-squared distribution. Just as the t distribution has a degree of freedom parameter that affects its shape, so does the chi-squared distribution. The degrees of freedom for our χ^2 statistic is given by:

$$df = (r-1)(c-1),$$

where r is the number of rows in the contingency tables, and c is the number of columns. In the marital status example we have 4 rows and 4 columns so $df = (4-1)(4-1) = 9$. Knowing the degrees of freedom, we can then look up a p-value for our χ^2 statistic.

In our data, the chi-square statistics of 284.6 on 9 degrees of freedom, which is highly significant. We can thus reject the null that marital status is independent of race.

In `R` the chi-square statistic does not have to be calculated, instead we can just pass the table object to the summary function:

```
> summary(tab0)
Number of cases in table: 15167
Number of factors: 2
Test for independence of all factors:
Chisq = 284.59, df = 9, p-value = 4.817e-56
>
```

The function prints to screen the chi-square statistic, the degrees of freedom, and the corresponding p-values. The results indicate that it would be very unlikely to obtain a chi-square statistic this large, if the null hypothesis of independence were true.