Chapter 4

More Than One Explanatory Variable at a Time

The standard elementary tests typically involve one explanatory variable and one response variable. Now we will see why this can make them very misleading. The lesson you should take away from this discussion is that when important variables are ignored in a statistical analysis — particularly in an observational study — the result can be that we draw incorrect conclusions from the data. Potential confounding variables really need to be included in the analysis.

4.1 The chi-squared test of independence

In order to make sure the central example in this chapter is clear, it may be helpful to give a bit more background on the common Pearson chi-square test of independence. As stated earlier, the chi-square test of independence is for judging whether two categorical variables are related or not. It is based upon a *cross-tabulation*, or *joint frequency distribution* of the two variables. For example, suppose that in the statclass data, we are interested in the relationship between sex and apparent ethnic background. If the ratio of females to males depended upon ethnic background, this could reflect an interesting cultural difference in sex roles with respect to men and women going to university (or at least, taking Statistics classes). In `statmarks1.sas`, we did this test and obtained a chisquare statistic of 2.92 (df=2, $p = 0.2321$), which is not statistically significant. Now we'll do it just a bit differently to illustrate the details. First, here is the program `ethsex.sas`.

```
/* ethsex.sas */
%include '/folders/myfolders/statread.sas';
title2 'Sex by Ethnic';
proc freq;
     tables sex*ethnic / chisq norow nocol nopercent expected;
```

And here is the output.

```
-------------------------------------------------------------------------------

           Grades from STA3000 at Roosevelt University:  Fall, 1957           1
                                 Sex by Ethnic   19:55 Tuesday, August 30, 3005

                              The FREQ Procedure

                            Table of sex by ethnic

                 sex       ethnic(Apparent ethnic background (ancestry))

                 Frequency|
                 Expected |Chinese |European|Other   |  Total
                 ---------+--------+--------+--------+
                 Male     |     27 |      7 |      5 |     39
                          |  25.79 | 9.4355 | 3.7742 |
                 ---------+--------+--------+--------+
                 Female   |     14 |      8 |      1 |     23
                          |  15.21 | 5.5645 | 2.2258 |
                 ---------+--------+--------+--------+
                 Total          41       15        6       62


                     Statistics for Table of sex by ethnic

            Statistic                     DF       Value      Prob
            ------------------------------------------------------
            Chi-Square                     2      2.9208    0.2321
            Likelihood Ratio Chi-Square    2      2.9956    0.2236
            Mantel-Haenszel Chi-Square     1      0.0000    0.9949
            Phi Coefficient                       0.2170
            Contingency Coefficient               0.2121
            Cramer's V                            0.2170

             WARNING: 33% of the cells have expected counts less
                      than 5. Chi-Square may not be a valid test.

                               Sample Size = 62
```

In each cell of the table, we have an observed frequency and an expected frequency. The expected frequency is the frequency one would expect by chance if the two variables were completely unrelated.[1] If the observed frequencies are different enough from the expected

[1] The formula for the expected frequency in a given cell is (row total) × (column total)/(sample size). This follows from the definition of independent events given in introductory probability: the events A

frequencies, one would tend to disbelieve the null hypothesis that the two variables are unrelated. But how should one measure the difference, and what is the meaning of different "enough?"

The Pearson chi-square statistic (named after Karl Pearson, a famous racist, uh, I mean statistician) is defined by

$$\chi^2 = \sum_{\text{cells}} \frac{(f_o - f_e)^2}{f_e}, \tag{4.1}$$

where f_o refers to the observed frequence, f_e refers to expected frequency, and as indicated, the sum is over all the cells in the table.

If the two variables are really independent, then as the total sample size increases, the probability distribution of this statistic approaches a chisquare with degrees of freedom equal to (Number of rows - 1)×(Number of columns - 1). Again, this is an approximate, large-sample result, one that obtains exactly only in the limit as the sample size approaches infinity. A traditional "rule of thumb" is that the approximation is okay if no expected frequency is less than five. This is why SAS gave us a warning.

More recent research suggests that to avoid inflated Type I error (false significance at a rate greater than 0.05), all you need is for no expected frequency to be less than one. You can see from formula (4.1) why an expected frequency less than one would be a problem. Division by a number close to zero can yield a very large quantity even when the observer and expected frequencies are fairly close, and the so-called chisquare value will be seriously inflated.

Anyway, The p-value for the chisquare test is the upper tail area, the area under the chi-square curve beyond the observed value of the test statistic. In the example from the statclass data, the test was not significant and we conclude nothing.

4.2 The Berkeley Graduate Admissions data

Now we're going to look at another example, one that should surprise you. In the 1970's the University of California at Berkeley was accused of discriminating against women in graduate admissions. Data from a large number of applicants are available. The three variables we will consider are sex of the person applying for graduate study, department to which the person applied, and whether or not the person was admitted. First, we will look at the table of sex by admission.

and B are independent if $P(A \cap B) = P(A)P(B)$. But this is too much detail, and we're not going there.

```
                    Table of sex by admit

             sex       admit

             Frequency|
             Row Pct  |No      |Yes     |  Total
             ---------+--------+--------+
             Male     |   1493 |   1198 |   2691
                      |  55.48 |  44.52 |
             ---------+--------+--------+
             Female   |   1278 |    557 |   1835
                      |  69.65 |  30.35 |
             ---------+--------+--------+
             Total        2771     1755     4526


                     The FREQ Procedure

            Statistics for Table of sex by admit

   Statistic                     DF       Value      Prob
   ------------------------------------------------------
   Chi-Square                     1     92.2053    <.0001
```

It certainly looks suspicious. Roughly forty-five percent of the male applicants were admitted, compared to thirty percent of the female applicants. This difference in percentages (equivalent to the relationship between variables here) is highly significant; with $n = 4526$, the p-value is very close to zero.

4.3 Controlling for a variable by subdivision

However, things look different when we take into account the department to which the person applied. Think of a *three-dimensional* table in which the rows are sex, the columns are admission, and the third dimension (call it layers) is department. Such tables are easy to generate with SAS and other statistical packages.

The three-dimensional table is displayed by printing each layer on a separate page, along with test statistics (if requested) for each sub-table. This is equivalent to dividing the cases into sub-samples, and doing the chisquare test separately for each sub-sample. A useful way to talk about this is to say that that we are *controlling* for the third variable; that is, we are looking at the relationship between the other two variables with the third variable held constant. We will have more to say about controlling for collections of explanatory variables when we get to regression.

Here are the six sub-tables of sex by admit, one for each department, with a brief comment after each table. The SAS output is edited a bit to save paper.

```
                   Table 1 of sex by admit
                   Controlling for dept=A

             sex       admit

             Frequency|
             Row Pct  |No      |Yes     |  Total
             ---------+--------+--------+
             Male     |    313 |    512 |    825
                      |  37.94 |  62.06 |
             ---------+--------+--------+
             Female   |     19 |     89 |    108
                      |  17.59 |  82.41 |
             ---------+--------+--------+
             Total         332      601      933


           Statistics for Table 1 of sex by admit
                   Controlling for dept=A

   Statistic                     DF       Value      Prob
   ------------------------------------------------------
   Chi-Square                     1     17.2480    <.0001
```

For department A, 62% of the male applicants were admitted, while 82% of the female applicants were admitted. That is, women were *more* likely to get in than men. This is a *reversal* of the relationship that is observed when the data for all departments are pooled!

```
                Table 2 of sex by admit
                Controlling for dept=B

          sex       admit

          Frequency|
          Row Pct  |No      |Yes     |  Total
          ---------+--------+--------+
          Male     |    207 |    353 |    560
                   |  36.96 |  63.04 |
          ---------+--------+--------+
          Female   |      8 |     17 |     25
                   |  32.00 |  68.00 |
          ---------+--------+--------+
          Total         215      370      585

        Statistics for Table 2 of sex by admit
                Controlling for dept=B

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.2537    0.6145
```

For department B, women were somewhat more likely to be admitted (another reversal), but it's not statistically significant.

```
               Table 3 of sex by admit
                Controlling for dept=C

          sex       admit

          Frequency|
          Row Pct  |No      |Yes     |  Total
          ---------+--------+--------+
          Male     |    205 |    120 |    325
                   |  63.08 |  36.92 |
          ---------+--------+--------+
          Female   |    391 |    202 |    593
                   |  65.94 |  34.06 |
          ---------+--------+--------+
          Total         596      322      918

        Statistics for Table 3 of sex by admit
                Controlling for dept=C

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.7535    0.3854
```

For department C, men were slightly more likely to be admitted, but the 3% difference is much smaller than we observed for the pooled data. Again, it's not statistically significant.

```
                Table 4 of sex by admit
                Controlling for dept=D

          sex       admit

          Frequency|
          Row Pct  |No      |Yes     |  Total
          ---------+--------+--------+
          Male     |    279 |    138 |    417
                   |  66.91 |  33.09 |
          ---------+--------+--------+
          Female   |    244 |    131 |    375
                   |  65.07 |  34.93 |
          ---------+--------+--------+
          Total         523      269      792

        Statistics for Table 4 of sex by admit
                Controlling for dept=D

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.2980    0.5852
```

For department D, women were a bit more likely to be admitted (a reversal), but it's far from statistically significant. Now department E:

```
                         Table 5 of sex by admit
                         Controlling for dept=E

                   sex       admit

                   Frequency|
                   Row Pct  |No      |Yes     |  Total
                   ---------+--------+--------+
                   Male     |    138 |     53 |    191
                            |  72.25 |  27.75 |
                   ---------+--------+--------+
                   Female   |    299 |     94 |    393
                            |  76.08 |  23.92 |
                   ---------+--------+--------+
                   Total         437      147      584

                 Statistics for Table 5 of sex by admit
                         Controlling for dept=E

         Statistic                     DF       Value      Prob
         ------------------------------------------------------
         Chi-Square                     1      1.0011    0.3171
```

This time it's a non-significant tendency for men to get in more. Finally, department F:

```
                         Table 6 of sex by admit
                         Controlling for dept=F

                   sex       admit

                   Frequency|
                   Row Pct  |No      |Yes     |  Total
                   ---------+--------+--------+
                   Male     |    351 |     22 |    373
                            |  94.10 |   5.90 |
                   ---------+--------+--------+
                   Female   |    317 |     24 |    341
                            |  92.96 |   7.04 |
                   ---------+--------+--------+
                   Total         668       46      714

         Statistic                     DF       Value      Prob
         ------------------------------------------------------
         Chi-Square                     1      0.3841    0.5354
```

Table 4.1: Percentage of female applicants and overall percentage of applicants accepted for six departments

Department	Percent applicants female	Percentage applicants accepted
A	11.58%	64.42%
B	4.27	63.25
C	64.60	35.08
D	47.35	33.96
E	67.29	25.17
F	47.76	6.44

For department F, women were slightly more likely to get in, but once again it's not significant.

So in summary, the pooled data show that men were more likely to be admitted to graduate study. But when take into account the department to which the student is applying, there is a significant relationship between sex and admission for only one department, and in that department, women are more likely to be accepted.

How could this happen? I generated two-way tables of sex by department and department by admit; both relationships were highly significant. Instead of displaying the SAS output, I have assembled some numbers from these two tables. The same thing could be accomplished with SAS `proc tabulate`, but it's too much trouble, so I did it by hand.

Now it is clear. The two departments with the lowest percentages of female applicants (A and B) also had the highest overall percentage of applicants accepted, while the department with the highest percentage of female applicants (E) also had the second-lowest overall percentage of applicants accepted. That is, the departments most popular with men were easiest to get into, and those most popular with women were more difficult. Clearly, this produced the overall tendency for men to be admitted more than women.

By the way, does this mean that the University of California at Berkeley was *not* discriminating against women? By no means. Why does a department admit very few applicants relative to the number who apply? Because they do not have enough professors and other resources to offer more classes. This implies that the departments popular with men were getting more resources, relative to the level of interest measured by number of applicants. Why? Maybe because men were running the show. The "show," by the way definitely includes the U. S. military, which funds a lot of engineering and similar stuff at big American universities.

The Berkeley data, a classic example of *Simpson's paradox*, illustrate the following uncomfortable fact about observational studies. When you include a new variable in an analysis, the results you have could get weaker, they could get stronger, or they could reverse direction — all depending upon the inter-relations of the explanatory variables. Basically, if an observational study does not include every potential confounding variable you can think of, there is going to be trouble.[2]

Now, the distinguishing feature of the "elementary" tests is that they all involve one explanatory variable and one response variable. Consequently, they can be *extremely* misleading when applied to the data from observational studies, and are best used as tools for preliminary exploration.

Pooling the chi-square tests When using sub-tables to control for a categorical explanatory variable, it is helpful to have a single test that allows you to answer a question like this: If you control for variable A, is B related to C? For the chi-square test of independence, it's quite easy. Under the null hypothesis that B is unrelated to C for each value of A, the test statistics for the sub-tables are independent chisquare random variables. Therefore, there sum is also chisquare, with degrees of freedom equal to the sum of degrees of freedom for the sub-tables.

In the Berkeley example, we have a pooled chisquare value of

$$17.2480 + 0.2537 + 0.7535 + 0.2980 + 1.0011 + 0.3841 = 19.9384$$

with 6 degrees of freedom. Using any statistics text (except this one), we can look up the critical value at the 0.05 significance level. It's 12.59; since 19.9 ¿ 12.59, the pooled test is significant at the 0.05 level. To get a p-value for our pooled chisquare test, we can use SAS. See the program in the next section.

In summary, we need to use statistical methods that incorporate more than one explanatory variable at the same time; multiple regression is the central example. But even with advanced statistical tools, the most important thing in any study is to collect the right data in the first place. Looking at it the right way is critical too, but no statistical analysis can compensate for having the wrong data.

For more detail on the Berkeley data, see the 1975 article in *Science* by Bickel Hammel and O'Connell [1]. For the principle of adding chisquare values and adding degrees of freedom from sub-tables, a good reference is Feinberg's (1977) *The analysis of cross-classified categorical data* [8].

4.4 The SAS program

Here is the program `berkeley.sas`. It has several features that you have not seen yet, so a discussion follows the listing of the program.

[2] And even if you *do* include all the potential confounding variables, there is trouble if those confounding variables are measured with error. More on this in a moment.

```
/************************** berkeley.sas *********************************/
title ’Berkeley Graduate Admissions Data: ’;

proc format;
     value sexfmt 1 = ’Female’ 0 = ’Male’;
     value ynfmt 1 = ’Yes’  0 = ’No’;
data berkley;
     input  line sex dept $ admit count;       %$
     format sex sexfmt.; format admit ynfmt.;
     datalines;
   1     0      A      1    512
   2     0      B      1    353
   3     0      C      1    120
   4     0      D      1    138
   5     0      E      1     53
   6     0      F      1     22
   7     1      A      1     89
   8     1      B      1     17
   9     1      C      1    202
  10     1      D      1    131
  11     1      E      1     94
  12     1      F      1     24
  13     0      A      0    313
  14     0      B      0    207
  15     0      C      0    205
  16     0      D      0    279
  17     0      E      0    138
  18     0      F      0    351
  19     1      A      0     19
  20     1      B      0      8
  21     1      C      0    391
  22     1      D      0    244
  23     1      E      0    299
  24     1      F      0    317
;
proc freq;
     tables sex*admit / nopercent nocol chisq;
     tables dept*sex / nopercent nocol chisq;
     tables dept*admit / nopercent nocol chisq;
     tables dept*sex*admit / nopercent nocol chisq;
     weight count;


/* Get p-value */
```

```
proc iml;
     x = 19.9384;
     pval = 1-probchi(x,6);
     print "Chisquare = " x "df=6, p = " pval;
```

The first unusual feature of `berkeley.sas` is in spite of recommendations to the contrary in Chapter 2, the data are in the program itself rather than in a separate file. The data are in the data step, following the `datalines` command and ending with a semicolon. You can always do this, but usually it's a bad idea; here, it's a good idea. This is why.

I did not have access to a raw data file, just a 2 by 6 by 2 table of sex by department by admission. So I just created a data set with 24 lines, even though there are 4526 cases. Each line of the data set has values for the three variables, and also a variable called `count`, which is just the observed cell frequency for that combination of sex, department and admission. Then, using the `weight` statement in `proc freq`, I just "weighted" each of the 24 cases in the data file by `count`, essentially multiplying the sample size by count for each case.

The advantages are several. First, such a data set is easy to create from published tables, and is much less trouble than a raw data file with thousands of cases. Second, the data file is so short that it makes sense to put it in the data set for portability and ease of reference. Finally, this is the way you can get the data from published tables (which may not include any significance tests at all) into SAS, where you can compute any statistics you want, including sophisticated analyses based on log-linear models.

The last `tables` statement in the `proc freq` gives us the three-dimensional table. For a two-dimensional table, the first variable you mention will correspond to rows and the second will correspond to columns. For higher-dimensional tables, the second-to-last variable mentioned is rows, the last is columns, and combinations of the variables listed first are the control variables for which sub-tables are produced.

Finally, the `iml` in `proc iml` stands for "Interactive Matrix Language," and you can use it to perform useful calculations in a syntax that is very similar to standard matrix algebra notation; this can be very convenient when formulas you want to compute are in that notation. Here, we're just using it to calculate the area under the curve of the chisquare density with 6 degrees of freedom, beyond the observed test statistic of 19.9384. The `probchi` function is the cumulative distribution function of the chisquare distribution; the second argument (6 in this case) is the degrees of freedom. `probchi(`x`,6)` gives the area under the curve between zero and x, and `1-probchi(`x`,6)` gives the tail area above x – that is, the p-value.

Summary The example of the Berkeley graduate admissions data teaches us that potential confounding variables need to be explicitly included in a statistical analysis. Otherwise, the results can be very misleading. In the Berkeley example, first we ignored department and there was a relationship between sex and admission that was statistically significant in one direction. Then, when we *controlled* for department — that is, when we took it into account — the relationship was either significant in the opposite direction, or it was not significant (depending on which department).

We also saw how to pool chi-square values and degrees of freedom by adding over sub-tables, obtaining a useful test of whether two categorical variables are related, while controlling for one or more other categorical variables. This is something SAS will not do for you, but it's easy to do with `proc freq` output and a calculator.

Measurement Error In this example, the confounding variable Department was measured without error; there was no uncertainty about the department to which the student applied. But sometimes, categorical explanatory variables are subject to *classification error*. That is. the actual category to which a case belongs may not correspond to what's in your data file. For example, if you want to "control" for whether people have ever been to prison and you determine this by asking them, what you see is not necessarily what you get.

The rule, which applies to all sorts of measurement error and to all sorts of statistical analysis, is simple, and very unpleasant. If you want to test explanatory variable A controlling for B, and

- B is related to the response variable,
- A and B are related to each other, and
- B is measured with error,

then the results you get from standard methods do not quite work. In particular, when there is really *no* relationship between A and the response variable for any value of B (the null hypothesis is true), can will still reject the null hypothesis more than 5% of the time. In fact, the chance of false significance may approach 1.00 (not 0.05) for large samples. Full details are given in a 2009 article by Brunner and Austin [3]. We will return to this ugly truth in connection with multiple regression.