Berkeley Data

```
/************************** berkeley.sas ********************************/
options linesize=79 pagesize=35 noovp formdlim='_';
title 'Berkeley Graduate Admissions Data: ';

proc format;
     value sexfmt 1 = 'Female' 0 = 'Male';
     value ynfmt 1 = 'Yes'  0 = 'No';
data berkley;
     infile 'bogus.dat';
     input  line sex dept $ admit count;
     format sex sexfmt.; format admit ynfmt.;

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;
```

The first few data lines:

```
   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
```

Part of Berkeley.lst

```
The FREQ Procedure

                         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
```

```
                Statistics for Table of sex by admit

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

Table of dept by sex

dept	sex		
Frequency Row Pct	Male	Female	Total
A	825 88.42	108 11.58	933
B	560 95.73	25 4.27	585
C	325 35.40	593 64.60	918
D	417 52.65	375 47.35	792
E	191 32.71	393 67.29	584
F	373 52.24	341 47.76	714
Total	2691	1835	4526

The FREQ Procedure

Statistics for Table of dept by sex

Statistic	DF	Value	Prob
Chi-Square	5	1068.3717	<.0001

Table of dept by admit

dept admit

Frequency Row Pct	No	Yes	Total
A	332 35.58	601 64.42	933
B	215 36.75	370 63.25	585
C	596 64.92	322 35.08	918
D	523 66.04	269 33.96	792
E	437 74.83	147 25.17	584
F	668 93.56	46 6.44	714
Total	2771	1755	4526

The FREQ Procedure

Statistics for Table of dept by admit

Statistic	DF	Value	Prob
Chi-Square	5	778.9065	<.0001

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

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

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

Statistic	DF	Value	Prob
Chi-Square	1	0.2537	0.6145

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

Statistic	DF	Value	Prob
Chi-Square	1	0.7535	0.3854

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

Statistic	DF	Value	Prob
Chi-Square	1	0.2980	0.5852

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

Statistic	DF	Value	Prob
Chi-Square	1	1.0011	0.3171

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 3.1: 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

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 independent variables. Basically, if an observational study does not include every potential confounding variable you can think of, there is going to be trouble.

Now, the distinguishing feature of the "elementary" tests is that they all involve one independent variable and one dependent 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 independent 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

```
/************************** berkeley.sas ********************************/
options linesize=79 pagesize=35 noovp formdlim='_';
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;
               /*  Get 21.745 > 12.59159 */
     weight count;
```