Chi-Square Tests in Stata (with post-hoc)

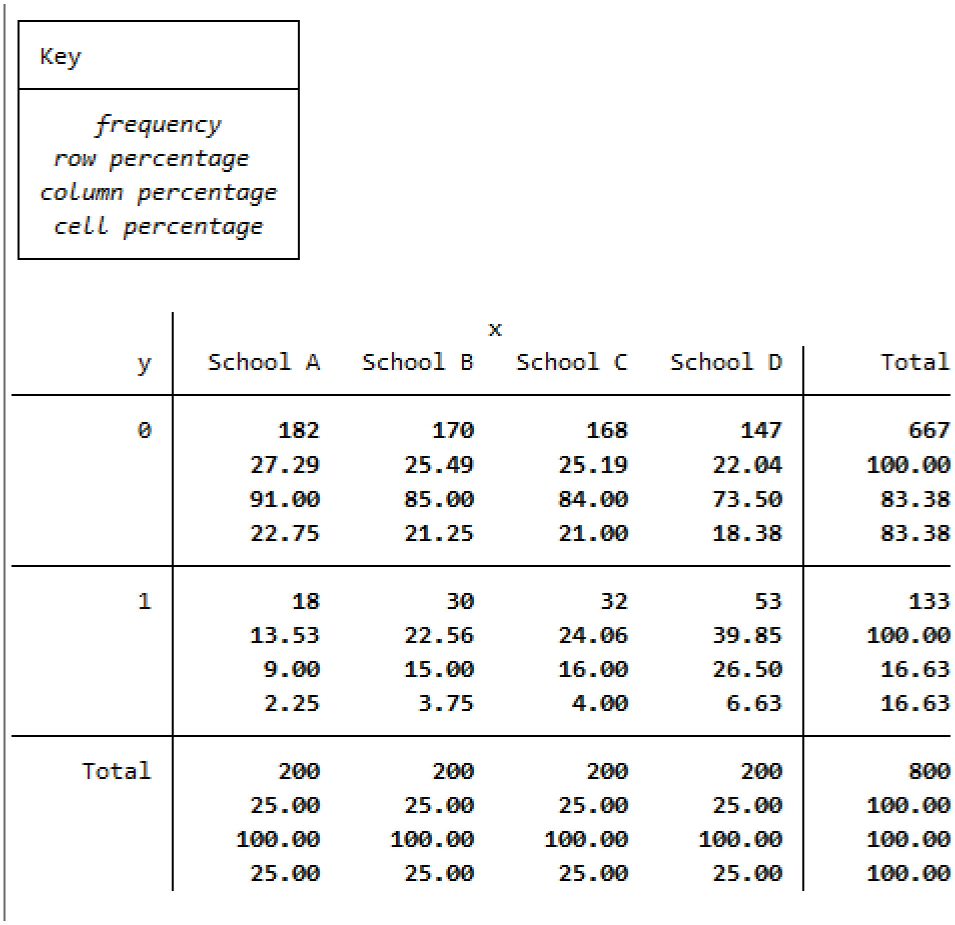
Suppose we are interested to know whether there is a difference between 4 different schools on the proportion of students who are missing at least one class each week.

Here is a small snippet of the data:

Identification	School	ClassMissed
ID001	School A	0
ID002	School A	0
ID003	School D	1
ID004	School C	1
…	…	…

This data might be organized with a contingency table, which can be obtained with the code:

tab ClassMissed School, row column cell

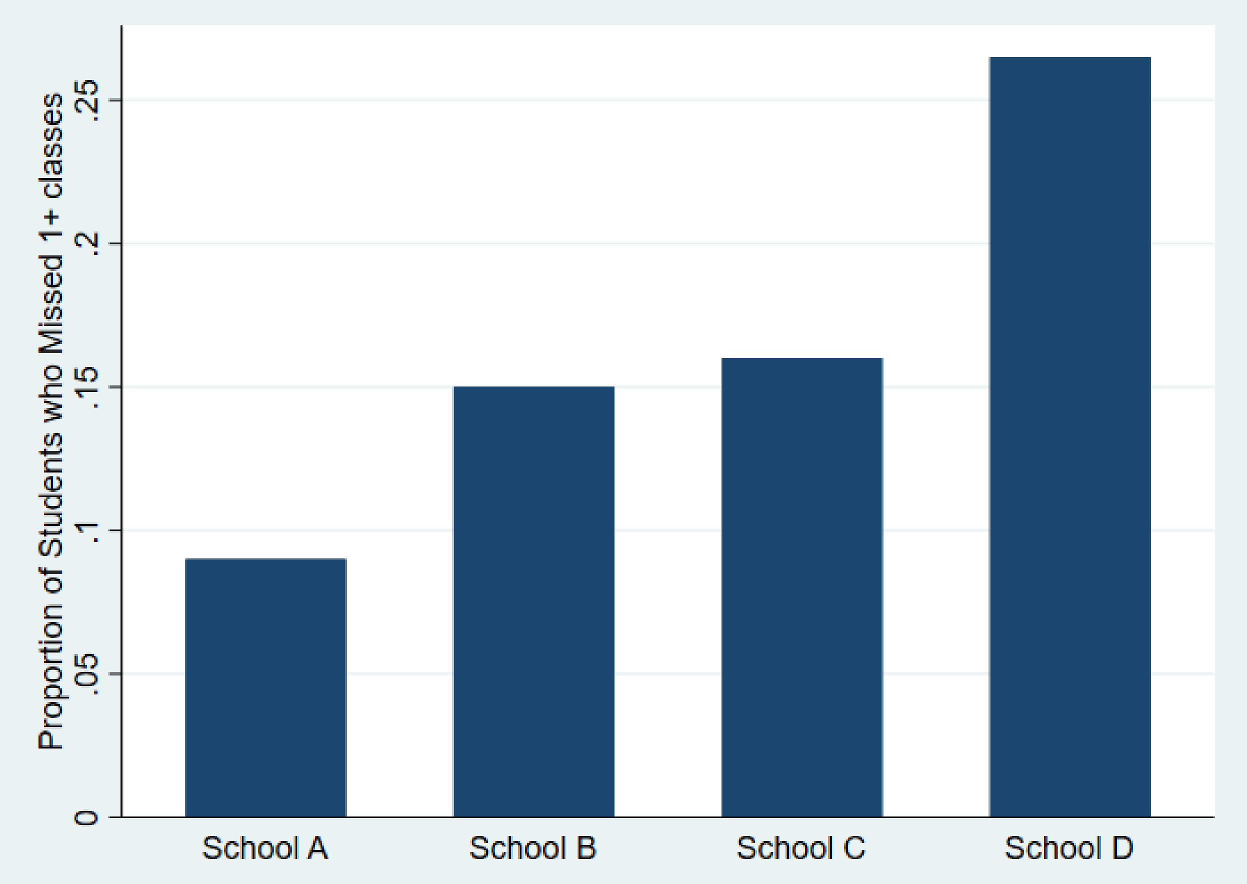


```
+-------------------+
| Key               |
|-------------------|
|     frequency     |
|  row percentage   |
| column percentage |
|  cell percentage  |
+-------------------+

           |                    x
         y |  School A   School B   School C   School D |     Total
-----------+--------------------------------------------+----------
         0 |       182        170        168        147 |       667 
           |     27.29      25.49      25.19      22.04 |    100.00 
           |     91.00      85.00      84.00      73.50 |     83.38 
           |     22.75      21.25      21.00      18.38 |     83.38 
-----------+--------------------------------------------+----------
         1 |        18         30         32         53 |       133 
           |     13.53      22.56      24.06      39.85 |    100.00 
           |      9.00      15.00      16.00      26.50 |     16.63 
           |      2.25       3.75       4.00       6.63 |     16.63 
-----------+--------------------------------------------+----------
     Total |       200        200        200        200 |       800 
           |     25.00      25.00      25.00      25.00 |    100.00 
           |    100.00     100.00     100.00     100.00 |    100.00 
           |     25.00      25.00      25.00      25.00 |    100.00 
```

Notice that the explanatory variable (School) represents the columns of this output and whether or not school was missed (ClassMiss) is the response variable. As a reminder, it is always useful to report the percentages that condition on the explanatory variable. Here that means the conditional column percentages.

We can see that 9% of students who are at School A miss class, 15% of students who are at School B miss class, 16% of students who are at School C miss class, and 26.5% students who are School D miss class. Students at School D seem to be much more likely to miss class than students at any of the others schools. We can also see this with a bivariate bar chart:

```
graph bar ClassMissed,  over(School)
```



Next, we will test the association

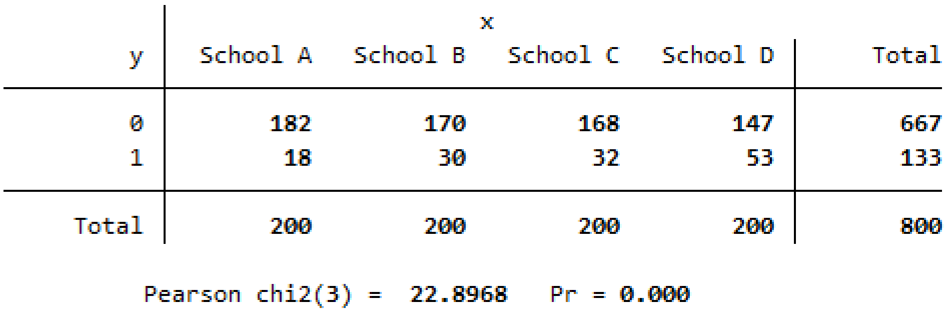
H0: There is no association between School and Classes Missed

HA: There is an association between School and Classes Missed

Since we have a categorical to categorical investigation, we will run a chi-square test.

Here is the code and corresponding output of the chi-square test in Stata:

tab ClassMissed School, chi2



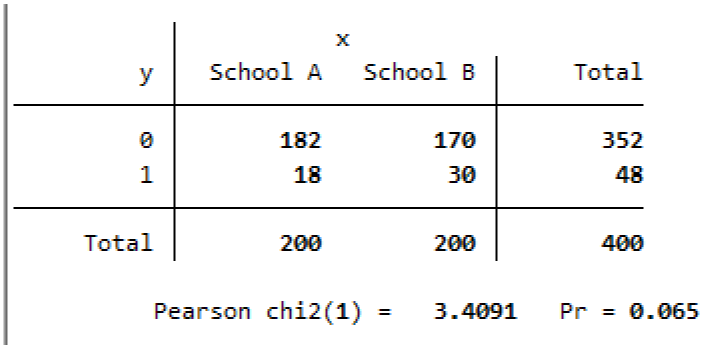
```
           |                       x
         y |  School A   School B   School C   School D |     Total
-----------+--------------------------------------------+----------
         0 |       182        170        168        147 |       667 
         1 |        18         30         32         53 |       133 
-----------+--------------------------------------------+----------
     Total |       200        200        200        200 |       800 

          Pearson chi2(3) =  22.8968   Pr = 0.000
```

- A chi-square test of independence revealed that missing class and school are significantly associated (X2=22.9, 3 df, p<0.001).

Great! Now our next logical step is to be able to decipher which schools are significantly different from one another. Since there are 4 schools, I want to compare A vs. B, A vs. C, A vs. D, B vs. C, B vs. D, and C vs. D. That is, I will need to compare 6 different combinations of schools.

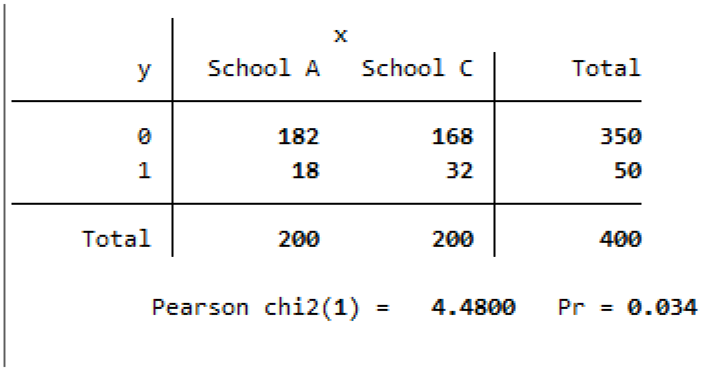
This part can be a little tedious!

tab ClassMissed School if School =="School A" | School =="School B", chi2



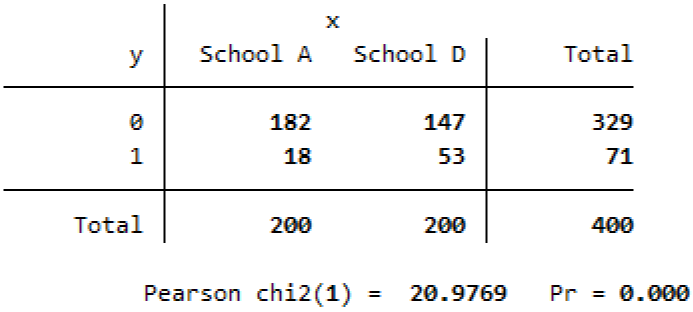
```
           |          x
         y |  School A   School B |     Total
-----------+----------------------+----------
         0 |       182        170 |       352 
         1 |        18         30 |        48 
-----------+----------------------+----------
     Total |       200        200 |       400 

          Pearson chi2(1) =   3.4091   Pr = 0.065
```

```
tab ClassMissed School if School =="School A" | School =="School C", chi2
```



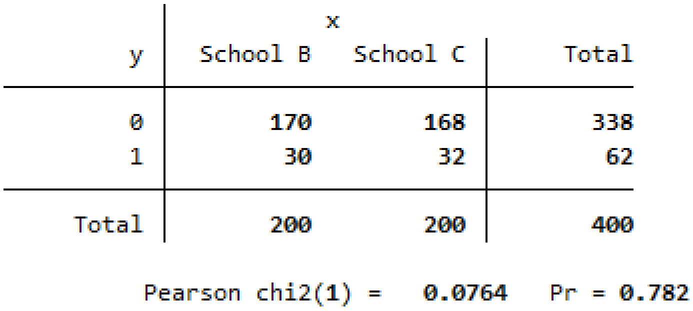
```
           |          x
         y |  School A   School C |     Total
-----------+----------------------+----------
         0 |       182        168 |       350 
         1 |        18         32 |        50 
-----------+----------------------+----------
     Total |       200        200 |       400 

          Pearson chi2(1) =   4.4800   Pr = 0.034
```

```
tab ClassMissed School if School =="School A" | School =="School D", chi2
```



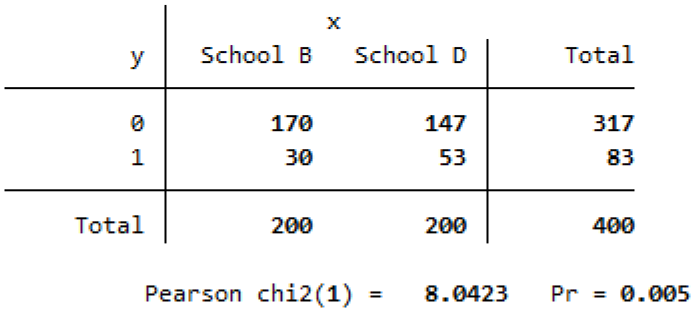
```
           |          x
         y |  School A   School D |     Total
-----------+----------------------+----------
         0 |       182        147 |       329 
         1 |        18         53 |        71 
-----------+----------------------+----------
     Total |       200        200 |       400 

          Pearson chi2(1) =  20.9769   Pr = 0.000
```

```
tab ClassMissed School if School =="School B" | School =="School C", chi2
```



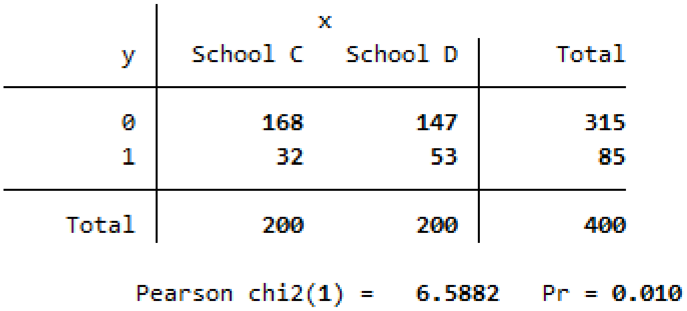
```
           |          x
         y |  School B   School C |     Total
-----------+----------------------+----------
         0 |       170        168 |       338 
         1 |        30         32 |        62 
-----------+----------------------+----------
     Total |       200        200 |       400 

          Pearson chi2(1) =   0.0764   Pr = 0.782
```

```
tab ClassMissed School if School =="School B" | School =="School D", chi2
```



```
           |          x
         y |  School B   School D |     Total
-----------+----------------------+----------
         0 |       170        147 |       317 
         1 |        30         53 |        83 
-----------+----------------------+----------
     Total |       200        200 |       400 

          Pearson chi2(1) =   8.0423   Pr = 0.005
```

tab ClassMissed School if School =="School C" | School =="School D", chi2



```
           |          x
         y |  School C   School D |     Total
-----------+----------------------+----------
         0 |       168        147 |       315 
         1 |        32         53 |        85 
-----------+----------------------+----------
     Total |       200        200 |       400 

          Pearson chi2(1) =   6.5882   Pr = 0.010
```

Since I ran a separate chi-square test on each pair – my overall Type I error across all of those tests will increase beyond my .05 limit. Therefore, I will need to implement a Bonferonni adjustment on the alpha level of .05 to control for this error!

The Bonferonni adjustment will require us to divide our alpha level by the total number of pairwise tests we ran. In this case, we ran 6 different tests for all possible comparisons. Therefore, I will be checking to see which tests have a p-value less than .0083.

- Post hoc comparisons of rates of missing class by pairs of schools revealed that School D had significantly higher rates of students missing class than School B and School A. All other schools were not found to differ significantly.

(Side note: When we run an ANOVA test and our explanatory variable has more than 2 levels, we may also need to do a post-hoc test. For ANOVA Stata has a command that automates this process entirely. It runs a separate test for each pair AND adjusts the p-values. This means that when we are running a post-hoc in ANOVA, we still compare it against the .05 level for each test when we use the “sidak” argument. Please note: You could alternatively run each pairwise test yourself and compare it against

.05/number of comparisons. That is, the Bonferonni adjustment can be used for ANOVA as well. It is just more work to do so).