Contingency Tables Overview

Suppose you have the following 2x2 table:

	Exposed	Not exposed	
Disease	a	b	a+b
No Disease	c	d	c+d
	a+c	b+d	N

In a *prospective* study you start by collecting a bunch of people who either were or were not exposed and follow them forward to see if they get the disease. Then you compare the proportion of people who get the disease in those two exposure groups.

So you want to compare the following probabilities:

θ_1 = P(disease| exposed)
θ_2 = P(disease| not exposed)

where $\hat{\theta}_1 = \frac{a}{a+c}$ *and* $\hat{\theta}_2 = \frac{b}{b+d}$

The estimated odds ratio of disease is given by

$$\frac{\hat{\theta}_1/(1-\hat{\theta}_1)}{\hat{\theta}_2/(1-\hat{\theta}_2)} = \frac{\hat{\theta}_1(1-\hat{\theta}_2)}{\hat{\theta}_2(1-\hat{\theta}_1)} = \frac{ad}{bc}$$

And the estimated relative risk is given by

$$\frac{\hat{\theta}_1}{\hat{\theta}_2} = \frac{a(b+d)}{b(a+c)}$$

Now in a retrospective study, you start by collecting a bunch of people who either have or do not have the disease and you look back into the past to see if they were exposed.

But now you can not compare the conditional probabilities above, because the number of people who have the disease is fixed by the study design.

So you have the same table, but you want to compare the following probabilities:

θ_3 = P(exposed| disease)
θ_4 = P(exposed| no disease)

where $\hat{\theta}_3 = \frac{a}{a+b}$ *and* $\hat{\theta}_4 = \frac{c}{c+d}$

Luckily, the estimated odds ratio for exposure is the same as the odds ratio for disease before

$$\frac{\hat{\theta}_3/(1-\hat{\theta}_3)}{\hat{\theta}_4/(1-\hat{\theta}_4)} = \frac{\hat{\theta}_3(1-\hat{\theta}_4)}{\hat{\theta}_4(1-\hat{\theta}_3)} = \frac{ad}{bc}$$

Here the associated relative risk is:

$$\frac{\hat{\theta}_3}{\hat{\theta}_4} = \frac{a(c+d)}{c(a+b)}$$

Here is the example we discussed in class. Suppose we collected data on the relationship between some disease and exposure, say cancer and living by power lines. And we collect the following data from 3500 females.

Females	Live close to power lines	Do not liver close to power lines	
Cancer	110	380	490
No Cancer	390	2620	3010
	500	3000	3500

The analysis yields:

```
. csi 110 380 390 2620, or w

                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |       110         380  |       490
        Noncases |       390        2620  |      3010
-----------------+------------------------+----------
           Total |       500        3000  |      3500
                 |                        |
            Risk |       .22    .1266667  |       .14
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .0933333       |  .0551229    .1315438
      Risk ratio |         1.736842       |  1.436418    2.100099
 Attr. frac. ex. |         .4242424       |  .3038239     .523832
 Attr. frac. pop |         .0952381       |
      Odds ratio |         1.944669       |   1.53375    2.465682  (Woolf)
                 +-------------------------------------------------
                            chi2(1) =    31.01  Pr>chi2 = 0.0000
```

And we would get similar results from a hypothesis test that the two probabilities are equal:

```
. prtesti 500 .22 3000 .1267

Two-sample test of proportion                     x: Number of obs =      500
                                                  y: Number of obs =     3000

------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |        .22   .0185257                      .1836904    .2563096
           y |      .1267   .0060731                       .114797     .138603
-------------+----------------------------------------------------------------
        diff |      .0933   .0194957                      .0550891    .1315109
             |  under Ho:   .0167625     5.57   0.000
------------------------------------------------------------------------------

                    Ho: proportion(x) - proportion(y) = diff = 0

            Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
             z =  5.566                   z =  5.566                   z =  5.566
         P < z =  1.0000           P > |z| =  0.0000           P > z =  0.0000
```

Notice that observed Chi-Square test statistic, 31.01, is the square of the observed z-statistic, 5.56.

Indeed these tests are testing the same hypothesis (that the proportions are equal).

Also, I can get the exact p-values (Fisher exact test), by specifying ‘exact’ after the comma in the ‘csi’ command.

Side note: Notice that the confidence interval for the difference between two proportions uses one SE, while the hypothesis tests that the two proportions are equal uses the SE that pools the data from the two groups.

This can lead to situations where the duality between confidence intervals and hypothesis tests breaks down (e.g., you reject the null, but zero is still in the interval – or the reverse).

For example:

```
. prtesti 50 .24 300 .1267

Two-sample test of proportion                      x: Number of obs =       50
                                                   y: Number of obs =      300
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |        .24   .0603987                      .1216208    .3583792
           y |      .1267   .0192048                      .0890593    .1643407
-------------+----------------------------------------------------------------
        diff |      .1133   .0633784                     -.0109194    .2375194
             |  under Ho:   .0534567     2.12   0.034
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                      z =   2.1195
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.9830         Pr(|Z| < |z|) = 0.0341          Pr(Z > z) = 0.0170
```

This is why it is critical to understand the formulae behind the tools. Here the hypothesis test suggests there is a difference, where the confidence interval suggests there is not.

Also, instead of using the ‘csi’ command, we can use the ‘tabi’ command:

```
. tabi 110 380 \ 390 2620

           |          col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |       110        380 |       490
         2 |       390      2,620 |     3,010
-----------+----------------------+----------
     Total |       500      3,000 |     3,500

           Fisher's exact =                 0.000
   1-sided Fisher's exact =                 0.000


. tabi 110 380 \ 390 2620, col row cell
+--------------------+
| Key                |
|--------------------|
|     frequency      |
|  row percentage    |
| column percentage  |
|  cell percentage   |
+--------------------+
           |          col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |       110        380 |       490
           |     22.45      77.55 |    100.00
           |     22.00      12.67 |     14.00
           |      3.14      10.86 |     14.00
-----------+----------------------+----------
         2 |       390      2,620 |     3,010
           |     12.96      87.04 |    100.00
           |     78.00      87.33 |     86.00
           |     11.14      74.86 |     86.00
-----------+----------------------+----------
     Total |       500      3,000 |     3,500
           |     14.29      85.71 |    100.00
           |    100.00     100.00 |    100.00
           |     14.29      85.71 |    100.00
```

```
. tabi 110 380 \ 390 2620, chi2 cchi2 exp

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
| chi2 contribution  |
+--------------------+

           |          col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |       110        380 |       490
           |      70.0      420.0 |     490.0
           |      22.9        3.8 |      26.7
-----------+----------------------+----------
         2 |       390      2,620 |     3,010
           |     430.0    2,580.0 |   3,010.0
           |       3.7        0.6 |       4.3
-----------+----------------------+----------
     Total |       500      3,000 |     3,500
           |     500.0    3,000.0 |   3,500.0
           |      26.6        4.4 |      31.0

          Pearson chi2(1) =  31.0078   Pr = 0.000
```

Expected counts in each cell – under null hypothesis (=row total*col total/total total)

```
. display 490*500/3500
70
```

Chi-square contribution of each cell ($=(O-E)^2/E$)

```
. display ((110-70)^2)/70
22.857143
```

Now back to the original example:
Notice also, that if I reverse the table, I get the same odds ratio and inference (confidence interval and test statistics on the odds ratio):

```
. ***********************************
. csi 110 380 390 2620, or w

                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |       110         380  |       490
        Noncases |       390        2620  |      3010
-----------------+------------------------+----------
           Total |       500        3000  |      3500
                 |                        |
            Risk |       .22    .1266667  |       .14
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .0933333       |  .0551229    .1315438
      Risk ratio |         1.736842       |  1.436418    2.100099
 Attr. frac. ex. |         .4242424       |  .3038239     .523832
 Attr. frac. pop |         .0952381       |
      Odds ratio |         1.944669       |   1.53375    2.465682  (Woolf)
                 +-----------------------------------------------
                             chi2(1) =    31.01  Pr>chi2 = 0.0000

. csi 110 390 380 2620, or w

                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |       110         390  |       500
        Noncases |       380        2620  |      3000
-----------------+------------------------+----------
           Total |       490        3010  |      3500
                 |                        |
            Risk |  .2244898    .1295681  |  .1428571
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |         .0949217       |  .0560787    .1337647
      Risk ratio |         1.732601       |  1.434469    2.092694
 Attr. frac. ex. |          .422833       |   .302878    .5221471
 Attr. frac. pop |         .0930233       |
      Odds ratio |         1.944669       |   1.53375    2.465682  (Woolf)
                 +-----------------------------------------------
                             chi2(1) =    31.01  Pr>chi2 = 0.0000

. ***********************************
```

Finally, we took a look at Simpson's paradox by considering the same experiment in males and then looking at the combined table.

Here are the data for the males:

```
. csi 90 20 1410 980, or w

                 |   Exposed   Unexposed  |     Total
-----------------+------------------------+----------
           Cases |        90          20  |       110
        Noncases |      1410         980  |      2390
-----------------+------------------------+----------
           Total |      1500        1000  |      2500
                 |                        |
            Risk |       .06         .02  |      .044
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |               .04      |  .0251767    .0548233
      Risk ratio |                 3      |  1.860321    4.837875
 Attr. frac. ex. |         .6666667       |  .4624583    .7932977
 Attr. frac. pop |         .5454545       |
      Odds ratio |          3.12766       |   1.91355    5.112098
(Woolf)
                 +-----------------------------------------------
                             chi2(1) =    22.82  Pr>chi2 = 0.0000
```

For males, we see a similar association. But notice that the proportion of exposed males 1500/2500 is much greater than that for females (500/3500).

But we when we looked at the overall table (male and females combined) we see something totally different.

```
. csi 200 400 1800 3600, or w

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+-----------
           Cases |       200         400  |       600
        Noncases |      1800        3600  |      5400
-----------------+------------------------+-----------
           Total |      2000        4000  |      6000
                 |                        |
            Risk |        .1          .1  |        .1
                 |                        |
                 |      Point estimate    |  [95% Conf. Interval]
                 |------------------------+----------------------
 Risk difference |                 0      | -.0161027    .0161027
      Risk ratio |                 1      |  .8512687    1.174717
 Attr. frac. ex. |                 0      | -.1747172    .1487313
 Attr. frac. pop |                 0      |
      Odds ratio |                 1      |  .8361733    1.195924
(Woolf)
                 +-------------------------------------------------
                            chi2(1) =     0.00  Pr>chi2 = 1.0000
```

Here the association has disappeared. There are several reasons for this, which I won't go into here, but the point is that you must be very careful when combining tables or splitting them up. You might find things that are artifacts of the way you split you data.

This is one good reason why subgroup analyses should be specified before you see the data.

The reasons we get this weird result is not really that weird after all. It is the same reason why we can not compare death crude rates in populations where the age distribution is different.

The overall odds ratio is, in some sense, a weighted average of the strata specific odds ratios.

Specifically, the combined odds ratio is derived from the overall probabilities (i.e., the overall probability of disease given exposure) and the overall probabilities are nothing more than the weighted averages of the male and female probability of diseases given exposure.

But the weights are determined by the margins of the table, so when the table changes (i.e., two tables are combined), so do the weights. And the weights act as confounders here.

Hence you can get a combined odds ratio in a different direction than in the two original tables, simply because the weights have substantially changed.

It is possible to calculate an “Adjusted odds ratio”, but for that you’ll have to find a statistician.