Models In Epidemiology And Biostatistics
Gordon Hilton Fick

Session 17 :
Exact Methods With Discrete Outcomes

With Linear Regression, the p-values and confidence intervals are exact when one makes the assumption that the errors are [exactly] Normally distributed. The assumption of Normal errors then implicitly means one is assuming the outcome is continuous.

If the errors are assumed to symmetrically distributed [but not Normally distributed], then the p-values and confidence intervals are approximate and are based on asymptotics [so called 'large sample' mathematics]

We have also seen exact methods with outcomes from matched pair studies. Here the exact method is based on the binomial distribution. The approximate Chi Squared test is again making use of asymptotics [the Normal approximation to the Binomial]

Exact methods for discrete outcomes are now available for a number of models. Logistic regression [exlogistic] and Poisson regression [expoisson] can now be done with the exact methods in Stata [and R]. It is anticipated that other regression models will [eventually] be implemented for Stata and R [in the near future?]. These methods are often labelled as computationally intensive but even this label is fading [for these methods] as computing speeds increase and the algorithms get better and better. A group from Harvard [Mehta and Patel] have developed a considerable body of software [StatXact is now in release 11]. A book by them [1996] and also a book by Hirji [2006] are devoted entirely to these methods.

Two By Two Tables

Fisher's Exact Test involves the family of hypergeometric distributions. The central hypergeometric distribution is needed for p-values. The non-central hypergeometric distribution is the basis behind the confidence intervals for the odds ratio. The probability function for the central hypergeometric distribution is:

$$p(k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

where N is overall total, K is the number of cases [row 1 total],
n is the number of exposed [column 1 total]

	Exposed	Unexposed	
Cases	a	b	K
Controls	c	d	
	n		N

Once N, K and n are specified, the number [a] in the upper left cell determines the other entries [b,c,d]

As an example, we will take the data from a study included in Matthews & Farewell [2007, 4th edition] page 23 [Table 3.2]. The cases are those without remission, the controls are those in remission. The exposed are those receiving 6MP, the unexposed are those receiving Methyl GAG. [data is in Matthews_Farewell_3.2.dta]

```
. cc case drug,exact
                                                      Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |         7           3  |         10       0.7000
        Controls |         2           7  |          9       0.2222
-----------------+------------------------+------------------------
           Total |         9          10  |         19       0.4737
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |       8.166667        |    .7520602    113.4441 (exact)
                 +-------------------------------------------------
                               1-sided Fisher's exact P = 0.0513
                               2-sided Fisher's exact P = 0.0698
```

So we can see that N=19, K=10, n=9, a=7. The observed value for k is a=7. One can then see that the possible values for k range from 0 to 9. Each of the 10 values of k determines a 2 by 2 table.

The calculations of the 2 p-values can be done directly using Stata:

```
set obs 10
gen k=_n-1
gen pk=hypergeometricp(19,10,9,k)
gen Fk=hypergeometric(19,10,9,k)
gen Sk=1-Fk

. list

     +-----------------------------------+
     | k         pk         Fk         Sk |
     |-----------------------------------|
  1. | 0   .0000108   .0000108   .9999892 |
  2. | 1   .0009743   .0009851   .9990149 |
  3. | 2   .0175366   .0185217   .9814783 |
  4. | 3   .1091169   .1276386   .8723614 |
  5. | 4   .2864318   .4140705   .5859295 |
     |-----------------------------------|
  6. | 5   .3437182   .7577887   .2422113 |
  7. | 6   .1909546   .9487432   .0512568 |
  8. | 7   .0467644   .9955076   .0044924 |
  9. | 8   .0043842   .9998918   .0001082 |
 10. | 9   .0001083          1          0 |
     +-----------------------------------+
```

It is worth noting that the two-sided p-value is not twice the one-sided p-value. The probability distribution here is not symmetrical. The probability of the observed [k=a=7] is 0.0467644. Therefore, the p-value is probability of k=7 or more [0.0512568] plus the probability of k=2 or less [0.0185217] which equals .0697785.

The approximate Chi Squared is not appropriate here and is VERY misleading here. The incorrect p-value is 0.0373 [below the mighty 0.05] printed next.

```
cc case drug
                                                              Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |         7           3  |         10       0.7000
        Controls |         2           7  |          9       0.2222
-----------------+------------------------+------------------------
           Total |         9          10  |         19       0.4737
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         8.166667       |    .7520602    113.4441 (exact)
                 +-------------------------------------------------
                               chi2(1) =     4.34  Pr>chi2 = 0.0373
```

It is crucial to use the exact method here. In general, one cannot anticipate whether the approximate method will be OK. In years gone by, there were so called 'rules of thumb' that are now obsolete. Many attempts to 'correct' the Chi Squared are also obsolete [including Yates' correction!]

Notice that Stata gives the 'exact' confidence interval [it is now the default]. Unfortunately, Stata's output here gives the MLE of the Odds Ratio based on the approximate methods and not the correct MLE based on exact methods.

We can get the correct exact analysis now using exlogistic:

```
exlogistic case drug,coef test(prob) nolog

Exact logistic regression                       Number of obs =         19
                                                Model prob.   =   .0467644
                                                Pr <= prob.   =     0.0698
------------------------------------------------------------------------------
        case |      Coef.       Prob.    Pr<=Prob.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |    1.96939    .0467644      0.0698     -.2849994    4.731313
------------------------------------------------------------------------------

exlogistic case drug,test(prob) nolog

Exact logistic regression                       Number of obs =         19
                                                Model prob.   =   .0467644
                                                Pr <= prob.   =     0.0698
------------------------------------------------------------------------------
        case | Odds Ratio       Prob.    Pr<=Prob.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        drug |   7.166306    .0467644      0.0698      .7520147    113.4444
------------------------------------------------------------------------------
```

The correct exact odds ratio estimate is 7.166306. The correct two sided p-value is 0.0698

Lets now consider a project with two 2 by 2 tables. For this study, the investigators are very interested in modification. We have a case-control study of lung cancer by Caporaso et al. (1989) Here, a genetic factor is collapsed into two levels: PMM = poor or moderate metabolizer; and EXM = extensive metabolizer. The observed odds ratio among PMM group is .2619048, and among the EXM group, it is 1.938144 Does the genetic factor modify the impact of the environmental exposure [asbestos] on the etiology of lung cancer? Here are the two 2 by 2 tables:

```
cc case expo [fw=ct1] if metab==0
                                                              Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        47          97  |        144       0.3264
        Controls |        17          68  |         85       0.2000
-----------------+------------------------+------------------------
           Total |        64         165  |        229       0.2795
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         1.938144       |    .9899637    3.908095 (exact)
 Attr. frac. ex. |         .4840426       |   -.0101381    .7441208 (exact)
 Attr. frac. pop |         .1579861       |
                 +-------------------------------------------------
                           chi2(1) =     4.24  Pr>chi2 = 0.0395

cc case expo [fw=ct1] if metab==1
                                                              Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |         1          14  |         15       0.0667
        Controls |        15          55  |         70       0.2143
-----------------+------------------------+------------------------
           Total |        16          69  |         85       0.1882
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         .2619048       |    .0058123    2.029571 (exact)
 Prev. frac. ex. |         .7380952       |   -1.029571    .9941877 (exact)
 Prev. frac. pop |         .1581633       |
                 +-------------------------------------------------
                           chi2(1) =     1.76  Pr>chi2 = 0.1844

cc case expo [fw=ct1],by(metab)

           metab |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
               0 |   1.938144      .9899637   3.908095     7.200873 (exact)
               1 |   .2619048      .0058123   2.029571     2.470588 (exact)
-----------------+-------------------------------------------------
           Crude |   1.662162      .9628919   2.886143              (exact)
    M-H combined |   1.509947      .8525815   2.674159
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     3.25  Pr>chi2 = 0.0715

                  Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      1.91
                                                Pr>chi2 =    0.1667
```

The test for modification has a p-value of 0.0715. Interesting. But one of the cells equals one. “Small”, right? Can we use the approximate methods?

Lets try the exact logistic regression. Our attention is with modification. We can request that particular regression coefficient estimate using the exlogistic command. The model is specified in a different way from all our previous usages. We include the desired coefficient after the outcome. Then include all the other coefficients to be included in the model as arguments in the cond option. The example should make this clear(er).

$$p = P(E) \quad \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 C + \beta_2 M + \beta_3 CM$$

```
gen cm=case*metab

exlogistic expo case metab cm [fw=ct1], test(prob) coef memory(200m) nolog

Exact logistic regression                               Number of obs =        314
                                                        Model prob.   =   .0000606
                                                        Pr <= prob.   =     0.0379
------------------------------------------------------------------------------
        expo |      Coef.       Prob.    Pr<=Prob.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        case |   .6589393    .0145364      0.0475     -.0100997    1.363209
       metab |   .0864486    .1537656      0.8444     -.7738123    .9391256
          cm |  -1.939003     .046246      0.0684     -5.844093    .2480317
------------------------------------------------------------------------------

estat se,coef

-------------------------------------
        expo |      Coef.   Std. Err.
-------------+-----------------------
        case |   .6589393   .3235531
       metab |   .0864486    .396675
          cm |  -1.939003   1.112783
-------------------------------------

exlogistic expo cm [fw=ct1],coef test(prob) nolog cond(case metab)

Exact logistic regression
                                                        Number of obs =        314
------------------------------------------------------------------------------
        expo |      Coef.       Prob.    Pr<=Prob.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          cm |  -1.939003     .046246      0.0684     -5.844093    .2480317
------------------------------------------------------------------------------

estat se,coef

-------------------------------------
        expo |      Coef.   Std. Err.
-------------+-----------------------
          cm |  -1.939003   1.112783
-------------------------------------
```

We get the crucial regression coefficient b_3 using this approach. Our software developers have given us the ability to see one coefficient at a time [without memory issues and computing times].

We get an exact test for modification.

We can ask for the exact with metab==0 by trying out:

```
exlogistic expo case [fw=ct1],test(prob) nolog cond(cm metab)

Exact logistic regression
                                                        Number of obs =        314
------------------------------------------------------------------------------
        expo | Odds Ratio       Prob.    Pr<=Prob.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        case |   1.932741    .0145364      0.0475      .9899512    3.908717
------------------------------------------------------------------------------
```

```
cc case expo [fw=ct1] if metab==0,exact
                                                              Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |        47          97  |        144       0.3264
        Controls |        17          68  |         85       0.2000
-----------------+------------------------+------------------------
           Total |        64         165  |        229       0.2795
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         1.938144       |    .9899637    3.908095 (exact)
                 +-----------------------------------------------
                                  1-sided Fisher's exact P = 0.0270
                                  2-sided Fisher's exact P = 0.0475
```

Reverse coding gives us the other group.

```
gen omm=1-metab
gen co=case*omm

exlogistic expo case [fw=ct1],test(prob) nolog cond(co omm)

Exact logistic regression
                                                Number of obs =        314
------------------------------------------------------------------------------
        expo | Odds Ratio       Prob.    Pr<=Prob.     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        case |   .2649492    .1375434      0.2831      .0058123    2.029204
------------------------------------------------------------------------------

cc case expo [fw=ct1] if metab==1,exact
                                                              Proportion
                 |   Exposed   Unexposed  |      Total     Exposed
-----------------+------------------------+------------------------
           Cases |         1          14  |         15       0.0667
        Controls |        15          55  |         70       0.2143
-----------------+------------------------+------------------------
           Total |        16          69  |         85       0.1882
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
      Odds ratio |         .2619048       |    .0058123    2.029571 (exact)
                 +-----------------------------------------------
                                  1-sided Fisher's exact P = 0.1691
                                  2-sided Fisher's exact P = 0.2831
```