Using Mantel-Haenszel Statistics

Xiaoxun Sun 4 October 2012

ConQuest 3 is able to report Mantel-Haenszel Statistics to assist differential item functioning (DIF) analysis, which was investigated using log-odds estimators. This note is concerned with the implementation of Mantel-Haenszel Statistics (Mantel and Haenszel, 1959) for dichotomous and polytomous variables in ConQuest 3.

Mantel-Haenszel Method

The Mantel-Haenszel method, introduced in 1959 by Mantel and Haenszel, is suitable for testing the *null* hypothesis of independence between two dichotomous variables using data from a population subdivided into *K* groups: it is, therefore, a method for analysing a 2×2×*K* contingency table. The Mantel-Haenszel method is used for estimating and testing a common two-factor association parameter in a 2×2×*K* table.

The Mantel-Haenszel method is used to detect DIF. Suppose the interest is in examining whether the dichotomously scored item ***i*** shows DIF for a focal group and a reference group. Typically, the sample is divided into *K* matched groups based on the raw scores. In ConQuest 3, the classification of the population is based on population ability estimations EAPs, WLEs, MLEs or PVs, and the estimation type can be specified by the statement `estimate`. This approach better deals with incomplete designs.

	Correct on i^{th} item	Incorrect on i^{th} item	Total
Reference group	A_{ik}	B_{ik}	$A_{ik}+B_{ik}$
Focal group	C_{ik}	D_{ik}	$C_{ik}+D_{ik}$
Total	$A_{ik}+C_{ik}$	$B_{ik}+D_{ik}$	$A_{ik}+B_{ik}+C_{ik}+D_{ik}$

Table 1 Example of a 2×2 contingency table for Mantel-Haenszel statistics computation

As shown in Table 1, $A_{ik}, B_{ik}, C_{ik}, D_{ik}$ represent the number of observations in the k^{th} matched group who belong to the reference group and answered the i^{th} item correctly, who belong to the reference group and answered the i^{th} item incorrectly, who belong to the focal group and answered the i^{th} item correctly and who belong to the focal group and answered the i^{th} item incorrectly, respectively. The *K*, 2x2 tables (one for each matched group) form a 2x2x*K* table. Under the hypothesis of no DIF, the proportion of correct answers in both reference and focal groups should be the same for all *K*. The formula for computing Mantel-Haenszel statistics for the i^{th} item is shown as follows:

$$MH\_DIF = -2.35 \times log_e \frac{\sum_K A_{ik}D_{ik}/(A_{ik}+B_{ik}+C_{ik}+D_{ik})}{\sum_K B_{ik}C_{ik}/(A_{ik}+B_{ik}+C_{ik}+D_{ik})}$$

According to the definition, the Mantel-Haenszel statistics is defined to be negative when the item is more difficult for members in the focal group than for the reference group. If there is no DIF, the Mantel-Haenszel statistics is 0. Further, a statistic that follows an approximate χ^2 distribution with (*K*-1) degrees of freedom can be computed by as:

$$MH\_CHISQ = \frac{(|\sum_k A_{ik} - \sum_k E(A_{ik})| - 0.5)^2}{\sum_k V(A_{ik})}$$

In which, $E(A_{ik})$ and $V(A_{ik})$ are the expected value and variance of A_{ik}, where

$$E(A_{ik}) = \frac{(A_{ik}+C_{ik})*(A_{ik}+B_{ik})}{(A_{ik}+B_{ik}+C_{ik}+D_{ik})},$$

$$V(A_{ik}) = \frac{(A_{ik}+C_{ik})*(C_{ik}+D_{ik})*(B_{ik}+D_{ik})*(A_{ik}+B_{ik})}{(A_{ik}+B_{ik}+C_{ik}+D_{ik})^2*(A_{ik}+B_{ik}+C_{ik}+D_{ik}-1)}$$

and 0.5 is the Yates' correction for continuity (Yates, 1934).

In addition to reporting the Mantel-Haenszel statistics, the suggested ETS DIF category is also included in the output of ConQuest 3. The ETS DIF category (Zwick, Thayer and Lewis, 1999) is shown in Table 2.

ETS DIF Category	P(MH_CHISQ)≤0.05	P(MH_CHISQ)>0.05
|MH_DIF|≥1.5	C = Moderate to Large	A = Negligible
1<|MH_DIF|<1.5	B = Slight to Moderate	A = Negligible
|MH_DIF|≤1	A = Negligible	A = Negligible

Table 2 ETS DIF category

Extension of Mantel-Haenszel method in ConQuest 3

The implementation of Mantel-Haenszel statistics in ConQuest 3 extends the method to support multiple focal groups and multiple scoring categories.

In order to support multiple focal groups, all focal groups are compared with the specified reference group. The Mantel- Haenszel statistics are reported between each focal and reference groups.

In order to support multiple scoring categories, the comparison is made for all pairs of adjacent scoring categories. For example, if a partial credit item has scoring categories 0, 1, and 2; the Mantel-Haenszel statistics is reported between scoring categories 0 and 1, and scoring categories 1 and 2. Please note that this is not the unique way to extend the Mantel-Haenszel statistics to multiple scoring categories, but this approach is consistent with the construction of partial credit model which relies upon the application of the simple logistic model to sequential pairs of response categories.

Examples and Display of Results

There are two ways in ConQuest 3 to request Mantel-Haenszel procedure. The first way is through `mh` command. The other way is through the `plot` command.

Some examples of requesting Mantel-Haenszel procedure are as follows:

Example 1:

```
mh! gins=2,bins=5,estimates=latent,group=gender,reference=F;
```

In this example, the Mantel-Haenszel procedure is requested for item 2, in which the group is `gender` and the reference group is specified to be `F`. the grouping variable is gender and the reference group is specified to be F. The population has been divided into 5 matched groups based on `PVs`. The ConQuest 3 output of *Example 1* is shown as follows:

```
================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=F
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
============================================================Build: Apr 22 2012===
item:2 (2)
               Bins                        OBSERVED   EXPECTED
-----------------------------------          GROUP     GROUP      CELL       CELL
     Low         High        Mean      N    AVERAGE   AVERAGE     CHISQ       P
--------------------------------------------------------------------------------
      <-      -0.19819    -0.59158    198   1.000      0.970      6.223      0.013
  -0.19819     0.19032     0.00169    202   1.000      0.983      3.508      0.061
   0.19032     0.54286     0.37718    193   1.000      0.988      2.302      0.129
   0.54286     0.96489     0.75860    211   0.995      0.992      0.294      0.588
   0.96489        ->       1.30421    196   1.000      0.995      0.925      0.336
--------------------------------------------------------------------------------
Between ability group fit:   13.252  df=4   p=0.010

================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=M
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
============================================================Build: Apr 22 2012===
item:2 (2)
               Bins                        OBSERVED   EXPECTED
-----------------------------------          GROUP     GROUP      CELL       CELL
     Low         High        Mean      N    AVERAGE   AVERAGE     CHISQ       P
--------------------------------------------------------------------------------
      <-      -0.19819    -0.58012    204   1.000      0.970      6.339      0.012
  -0.19819     0.19032     0.00935    197   1.000      0.983      3.395      0.065
   0.19032     0.54286     0.38162    207   1.000      0.988      2.458      0.117
   0.54286     0.96489     0.73823    189   1.000      0.992      1.571      0.210
   0.96489        ->       1.32252    203   0.995      0.995      0.004      0.948
--------------------------------------------------------------------------------
Between ability group fit:   13.768  df=4   p=0.008

/* There are 2 scoring categories */
Mantel-Haenszel Statistics (between categories 0 and 1) is -0.000; CHISQ=0.001; df=4;
p=1.00000
Suggested DIF Category is A: Negligible
The reference group is :F
```

Example 2 presents the Mantel-Haenszel procedure in ConQuest 3 supporting the multiple groups.

Example 2:

```
mh! gins=3,bins=3,estimates=wle,group=country,reference=AUS;
```

In this example, the Mantel-Haenszel procedure is requested for item 3, the grouping variable is country and the reference group is `AUS`. The population has been divided into 3 matched groups based on `WLEs`. The ConQuest 3 output of *Example 2* is shown as follows:

```
================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: country=AUS
DATA TABLE FOR EXPECTED SCORE
Estimator type is weighted maximum likelihood
============================================================Build: Apr 22 2012===
item:3 (3)
               Bins                        OBSERVED   EXPECTED
-----------------------------------          GROUP     GROUP      CELL       CELL
     Low         High        Mean      N    AVERAGE   AVERAGE     CHISQ       P
--------------------------------------------------------------------------------
      <-       0.32868    -0.37637    398   0.349      0.292      6.418      0.011
   0.32868     1.11308     0.88123    201   0.791      0.591     33.161      0.000
   1.11308        ->       1.65531     67   0.970      0.758     16.399      0.000
--------------------------------------------------------------------------------
Between ability group fit:   55.979  df=2   p=0.000
================================================================================
```

```
ConQuest: Generalised Item Response Modelling Software
Group: country=CHN
DATA TABLE FOR EXPECTED SCORE
Estimator type is weighted maximum likelihood
============================================================Build: Apr 22 2012===
item:3 (3)
              Bins                        OBSERVED   EXPECTED
---------------------------------          GROUP     GROUP      CELL       CELL
     Low         High        Mean      N   AVERAGE   AVERAGE    CHISQ        P
-------------------------------------------------------------------------------
      <-       0.32868    -0.08544    248  0.298      0.355      3.475      0.062
   0.32868     1.11308     0.89307    267  0.640      0.594      2.365      0.124
   1.11308       ->        1.84492    152  0.875      0.791      6.436      0.011
-------------------------------------------------------------------------------
Between ability group fit:   12.275  df=2   p=0.002

/* There are 2 scoring categories */
Mantel-Haenszel Statistics (between categories 0 and 1) is -0.400; CHISQ=2.929; df=2;
p=0.56975
Suggested DIF Category is A: Negligible
The reference group is :AUS
===============================================================================
ConQuest: Generalised Item Response Modelling Software
Group: country=USA
DATA TABLE FOR EXPECTED SCORE
Estimator type is weighted maximum likelihood
============================================================Build: Apr 22 2012===
item:3 (3)
              Bins                        OBSERVED   EXPECTED
---------------------------------          GROUP     GROUP      CELL       CELL
     Low         High        Mean      N   AVERAGE   AVERAGE    CHISQ        P
-------------------------------------------------------------------------------
      <-       0.32868     0.10757    168  0.119      0.400     55.375      0.000
   0.32868     1.11308     0.92625    218  0.518      0.602      6.400      0.011
   1.11308       ->        2.07120    281  0.786      0.826      3.105      0.078
-------------------------------------------------------------------------------
Between ability group fit:   64.880  df=2   p=0.000

/* There are 2 scoring categories */
Mantel-Haenszel Statistics (between categories 0 and 1) is -1.183; CHISQ=_BIG_; df=2;
p=0.00035
Suggested DIF Category is B: Slight to moderate
The reference group is :AUS
```

Example 3 demonstrates the Mantel-Haenszel procedure in ConQuest 3 supports the multiple scoring categories as well.

Example 3:

```
mh! gins=1,bins=5,estimates=latent,group=gender,reference=F;
```

In this example, the Mantel-Haenszel procedure is requested for item 1 with three scoring categories, the grouping variable is gender and the reference group is specified to be `F`. The population has been divided into 5 matched groups based on `PVs`. The ConQuest 3 output of the *Example 3* is shown as follows:

```
===============================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=F
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
============================================================Build: Apr 22 2012===
item:1 (1)
              Bins                        OBSERVED   EXPECTED
---------------------------------          GROUP     GROUP      CELL       CELL
     Low         High        Mean      N   AVERAGE   AVERAGE    CHISQ        P
-------------------------------------------------------------------------------
      <-       0.16097    -0.24979    182  0.599      0.437     10.948      0.001
   0.16097     0.53659     0.37551    210  0.814      0.774      0.543      0.461
   0.53659     0.89430     0.72553    204  1.103      1.003      3.065      0.080
   0.89430     1.28020     1.07357    206  1.112      1.230      4.614      0.032
   1.28020       ->        1.62958    198  1.343      1.535     15.894      0.000
-------------------------------------------------------------------------------
```

```
Between ability group fit:   35.062  df=4   p=0.000
================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=M
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
============================================================Build: Apr 22 2012===
item:1 (1)
               Bins                       OBSERVED   EXPECTED
---------------------------------          GROUP      GROUP       CELL        CELL
     Low         High        Mean      N   AVERAGE    AVERAGE     CHISQ        P
--------------------------------------------------------------------------------
      <-       0.16097    -0.24694    219  0.603       0.438      13.569       0.000
   0.16097     0.53659     0.36543    190  0.932       0.768       8.150       0.004
   0.53659     0.89430     0.74078    197  1.020       1.013       0.016       0.901
   0.89430     1.28020     1.07338    195  1.138       1.230       2.608       0.106
   1.28020       ->        1.67917    199  1.352       1.557      19.101       0.000
--------------------------------------------------------------------------------
Between ability group fit:   43.444  df=4   p=0.000
```

```
/* There are 3 scoring categories */
Mantel-Haenszel Statistics (between categories 0 and 1) is 0.017; CHISQ=0.001; df=4; p=1.00000
Suggested DIF Category is A: Negligible
The reference group is :F

Mantel-Haenszel Statistics (between categories 1 and 2) is 0.039; CHISQ=0.042; df=4; p=0.99978
Suggested DIF Category is A: Negligible
The reference group is :F
```

Example 4 shows the other way to request Mantel-Haenszel statistics, which is to use by `plot` command

Example 4:

```
plot expected! gins=3,bins=5,table=yes,estimates=latent,group=gender,mh=F;
```

In this example, the Mantel-Haenszel procedure is requested from the plot command, and it calculates the Mantel-Haenszel statistics for item 3, the grouping variable is gender and the reference group is specified to be `F`. The population has been divided into 5 matched groups based on `PVs`. The ConQuest 3 output of E*xample 4* is as follows:

```
================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=F
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
============================================================Build: Apr 22 2012===
item:3 (3)
               Bins                       OBSERVED   EXPECTED
---------------------------------          GROUP      GROUP       CELL        CELL
     Low         High        Mean      N   AVERAGE    AVERAGE     CHISQ        P
--------------------------------------------------------------------------------
      <-       0.12583    -0.30123    208  0.298       0.307       0.079       0.779
   0.12583     0.50174     0.34824    184  0.484       0.459       0.452       0.501
   0.50174     0.85885     0.68975    206  0.583       0.544       1.222       0.269
   0.85885     1.30104     1.06090    202  0.673       0.634       1.360       0.243
   1.30104       ->        1.70622    200  0.730       0.767       1.566       0.211
--------------------------------------------------------------------------------
Between ability group fit:   4.679  df=4   p=0.322
```

```
================================================================================
ConQuest: Generalised Item Response Modelling Software
Group: gender=M
DATA TABLE FOR EXPECTED SCORE
Estimator type is plausible value
============================================================Build: Apr 22 2012===
item:3 (3)
               Bins                        OBSERVED   EXPECTED
-----------------------------------          GROUP     GROUP      CELL       CELL
     Low          High         Mean       N  AVERAGE   AVERAGE    CHISQ       P
--------------------------------------------------------------------------------
      <-        0.12583    -0.29723     193  0.295      0.308      0.143      0.705
   0.12583      0.50174     0.32425     216  0.463      0.453      0.086      0.770
   0.50174      0.85885     0.69555     196  0.536      0.546      0.077      0.781
   0.85885      1.30104     1.06600     196  0.633      0.635      0.004      0.948
   1.30104        ->        1.68689     199  0.784      0.764      0.442      0.506
--------------------------------------------------------------------------------
Between ability group fit:    0.752  df=4   p=0.945

/* There are 2 scoring categories */
Mantel-Haenszel Statistics (between categories 0 and 1) is -0.054; CHISQ=0.071; df=4;
p=0.99939
Suggested DIF Category is A: Negligible
The reference group is :F
```

By using the `plot` command, the item plot across the reference and focal groups will also be shown on the screen.

References

Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22, 719-748.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

Zwick, R., Thayer, D.T., Lewis, C. (1999) An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis. *Journal of Educational Measurement*, 36, 1, 1-28.

Yates, F (1934). Contingency table involving small numbers and the χ^2 test. *Supplement to the Journal of the Royal Statistical Society 1(2)*: 217–235