PharmaSUG 2014 - Paper CC19

Let Chi-Square Pass Decision to Fisher's Programmatically

Linga Reddy Baddam, inVentiv Health Clinical, Hyderabad, India
Sudarshan Reddy Shabadu, inVentiv Health Clinical, Hyderabad, India

ABSTRACT

There is often the question of "when" and "which" statistical test should be used against any kind of clinical data in wide variety of situations as there are many significant cases when insufficient information can lead a statistician to choose an inappropriate analysis. It is common to test the hypothesis of independence of events across different treatment arms to identify the appropriate statistical test based on certain clinical reporting criterion and the clinical programmer has to develop code to carry out the subsequent statistical tests based on the nature of data and assumptions which have been made. In order to optimize efforts of clinical programmers, automating such code to find the appropriate statistical test is key-job and the features of such an automated %fc_pval macro are discussed in further detail in this paper.

INTRODUCTION

In the pharmaceutical industry, challenge for clinical trials is to carry them out ethically and fairly as it is always a challenge to investigate the relationships that exist between particular variables of interest. To reveal this statistical-affair there is often the question of "when" and "what" statistical test should be used against any kind of clinical data in wide variety of situations as there are many significant cases when insufficient information can lead a statistician to choose a less appropriate analysis. With respect to nominal data, it is common to test the hypothesis of independence of events (m >= 2, e.g. mild, moderate, severe, etc.) across different treatment arms (n >= 2, e.g. T1, T2, T3, etc.) and then identify the appropriate statistical test based on certain clinical reporting criteria.
For example, if at least 80% of cells have expected counts >= 5 we should use a chi-squared test; otherwise we should use a Fisher`s exact test to compute the p-value.
Now it is extremely important to understand this criterion and assess this goal accurately while keeping in mind the statistical consequences regarding the magnitude of the randomized study sample size.
Since clinical data is conservative, in order to better understand the situation we typically test the statistical assumptions on the data to see what kind of patterns develop thus leading the statistician to make a more informed decision.

In a scenario when insufficient data exists meaning expected cell counts are too small, the chi-squared test may not be the best option and to deal with this case the Fisher's exact test is more appropriate as it calculates the p-values based on exact reliability of data rather than relying on an approximation.

To make this process smoother the programmers should be able to construct a logical algorithm in such a way as to test the chi-squared assumptions first, and then Fisher's exact test.

BASIC ASSUMPTIONS OF THE CHI-SQUARED TEST

- Total frequency (N) should be reasonably large, i.e. greater than 50. Using the Chi squared test on small samples might result in a Type II error.
- The sample observations should be independent, this implies that no individual item should be included twice or more in the sample and it cannot be used to test correlated data like matched pairs.
- Any constraints on the cell frequencies should be linear (i.e. they should not involve square and higher powers of the frequencies) such as $\sum O = \sum E = N$.
- Expected cell counts should be >10 but must be >= 5, if any expected cell frequency is < 5 then we cannot use a Chi-Squared test; this situation will lead to use of the Fisher's Exact test.

In good conscience, the chi-squared test is appropriate for large samples or well-balanced data but the approximation is inadequate when sample sizes are small or the data is unequally distributed. Fortunately Fisher's exact test is suitable for small, spare and unbalanced data regardless of sample characteristics.

PROGRAMMING CONSIDERATIONS

If at least 80 % of the expected cells counts >= 5 then obtain p-value from the chi-square test,
Else use Fisher's exact test.

To do this, the first step of programming is calculating each expected cell frequency and testing the percentage rule against it.

This step then drives the subsequent steps, which are detailed later in the macro section of this paper.

MACRO OVERVIEW

While the primary goal of this macro is to check the clinical reporting criteria, there are the times when study requirements demand either chi-squared or Fisher's exact test by itself. To facilitate use of this feature by the programmers, the macro includes test_option parameter.

The %fc_pval macro call takes the following form:

```
%fc_pval ( indata =,
           in_var =,
           res_var =,
           cnt_var =,
           test_option =,
           outdata =);
```

The %fc_pval macro call is comprised of six keyword parameters,

indata:
  Required - name of the SAS dataset containing the observed frequencies of the response variable per group.
in_var:
  Required - name of the group variable
res_var:
  Required - name of the response variable
cnt_var:
  Required - name of the variable that contains observed frequencies
test_option:
  Optional- valid values = (FISHER_CHISQ, CHISQ, FISHER)
    - FISHER_CHISQ- performs a test based on the clinical reporting criteria, and it is the default value.
    - FISHER- requests fisher's exact test p-value in m x n contingency table.
    - CHISQ- requests chi-squared test p-value in m x n contingency table.
outdata:
  Required- name of the SAS output dataset containing p-value.

```
%MACRO fc_pval(  indata =,
                 in_var =,
                 res_var =,
                 cnt_var =,
                 test_option = fisher_chisq,
                 outdata = );

%IF (%SYSFUNC(EXIST(&indata))) > 0 %THEN %DO;
  %IF %UPCASE(&test_option.)=FISHER_CHISQ %THEN %DO;
    *Getting expected cell frequencies;
    PROC FREQ DATA = &indata. ;
      WEIGHT &cnt_var.;
         TABLES &in_var.*&res_var./EXPECTED OUT=count OUTEXPECT NOROW NOCOL NOPERCENT;
      RUN;

      PROC SQL NOPRINT;
        SELECT COUNT(DISTINCT &res_var.) INTO : row
          FROM &indata.
             WHERE NOT MISSING(&res_var.);
        SELECT COUNT(DISTINCT &in_var.) INTO : col
             FROM &indata.
             WHERE NOT MISSING(&in_var.);
        SELECT COUNT(expected) INTO : expect
             FROM count
             WHERE expected ge 5;
      QUIT;
```

```
      %LET row = &row.;
      %LET col = &col.;
      %LET expect = &expect.;
      %LET percent = %SYSEVALF(&expect./(&row.*&col.)*100);
  %END;
  %ELSE %LET percent = ;  /*assigning percent value to missing*/

  %IF %UPCASE(&test_option.) = CHISQ or &percent. ge 80 %THEN %DO;
      *Obtaining the p-value from chi-square test;
      PROC FREQ DATA = &indata.;
        WEIGHT &cnt_var.;
        TABLES &in_var.*&res_var./ CHISQ;
        ODS OUTPUT CHISQ = pval_c;
      RUN;

      DATA &outdata.(KEEP= test pvalue);
        SET pval_c (WHERE=(statistic="Chi-Square"));
        LENGTH test $ 10;
        pvalue = prob;
        test = "Chi-Square";
        FORMAT pvalue pvalue6.4;
      RUN;
  %END;
  %ELSE %IF %UPCASE (&test_option.) = FISHER or &percent. lt 80 %THEN %DO;
      *Obtaining the p-value from Fisher exact test;
      PROC FREQ DATA = &indata.;
        WEIGHT &cnt_var.;
        TABLES &in_var.*&res_var./ FISHER;
        ODS OUTPUT FISHERSEXACT = pval_f;
      RUN;

      DATA &outdata. (KEEP= test pvalue);
        SET pval_f (WHERE=(name1="XP2_FISH"));
        LENGTH test $ 10;
        pvalue = nvalue1;
        test = "Fisher";
        FORMAT pvalue pvalue6.4;
      RUN;
      %END;
%END;
%MEND fc_pval;
```

USING THE MACRO

Example

Extracorporeal membrane oxygenation (EcMO) is a potentially life-saving procedure that is used to treat newborn babies who suffer from severe respiratory failure. An experiment was conducted in which 29 babies were treated with EcMO and 10 babies were treated with conventional medical therapy (CMT). Among the 5 babies who died, 4 were treated with CMT and among the 34 babies that survived 6 were treated with CMT.

In order to test the independency between CMT (Treatment 1), EcMO (Treatment 2) the following dataset has been created where the procedures are the randomized treatments tested upon subjects, the result variable has 2 categories (DIE, LIVE) and the count variable cell frequencies of the 2 x 2 (treatment x result) contingency table.

```
DATA ecmo;
INPUT treatment result $ cnt @@;
DATALINES;
1 DIE 4 1 LIVE 6
2 DIE 1 2 LIVE 28
;
RUN;
/* treatment 1 = CMT, treatment 2 = EcMO */
```

Invoking the %fc_pval macro produces the following PROC FREQ outputs.

```
%fc_pval(     indata = ecmo,
              in_var = treatment,
              res_var = result,
              cnt_var = cnt,
              test_option = FISHER_CHISQ,
              outdata = pvalue);
```

The FREQ Procedure

Table of treatment by result

treatment result

Frequency Expected	DIE	LIVE	Total
1	4 1.2821	6 8.7179	10
2	1 3.7179	28 25.282	29
Total	5	34	39

Observed frequencies

Expected frequencies

Output 1. Output from a FREQ procedure shows the observed and expected cell frequencies.

The FREQ Procedure

Statistics for Table of treatment by result

Statistic	DF	Value	Prob
Chi-Square	1	8.8885	0.0029
Likelihood Ratio Chi-Square	1	7.7110	0.0055
Continuity Adj. Chi-Square	1	5.9190	0.0150
Mantel-Haenszel Chi-Square	1	8.6606	0.0033
Phi Coefficient		0.4774	
Contingency Coefficient		0.4308	
Cramer's V		0.4774	

WARNING: 50% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

Fisher's Exact Test

Cell (1,1) Frequency (F)	4
Left-sided Pr <= F	0.9996
Right-sided Pr >= F	0.0110
Table Probability (P)	0.0106
Two-sided Pr <= P	0.0110

Sample Size = 39

Output 2. Output from PROC FREQ shows the chi-squared and Fisher's exact test p-values.

It is clear from the output 1 and output 2 that 50% of the cells (treatment vs result) have expected frequencies (counts) less than 5 so the criteria for using the chi-squared test are not satisfied in this case.

Also notice that in output 2, SAS prints a warning about using the chi-squared test when expected cell frequencies are too small, however this will automatically lead to use of Fisher's exact test as demonstrated above which will be taken care of by the %fc_pval macro.

INTERPRETATION

The p-value for the Fisher's exact test is 0.011 for a non-directional (two sided) test, so the experiment provided strong evidence that EcMO really is statistically significant from CMT.

CONCLUSION

The %fc_pval macro certainly streamlines programming activity in accordance with clinical business needs since clinical programmers may not be sure of which statistical analysis to use for testing the independence of the events that can lead to inefficient and inadequate code.

To educate the programmer and improve the quality of the reports, this macro has been automated. As such, it will allow the programmer to work efficiently which has a positive impact on business needs in terms of good quality, efficiency and productivity.

REFERENCES

Book: S.C. Gupta. Fundamentals of Statistics. Page No: 18.6, Delhi, Himalaya Publishers.
SAS Institute Inc. 2011. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.

ACKNOWLEDGMENTS

We take this opportunity to thank Chunxue Shi, Manager of Biostatistics at inVentiv Health Clinical who has helped us in understanding these statistical concepts and provided useful review comments on our paper.
We are grateful to Nitin Pawar, Manager of Statistical Programming and Prayankotveettil Ranjith, Associate Director at inVentiv Health Clinical, whose support and guidance encouraged us to write this paper.
A special thanks to Nancy Brucken and Heather Murphy who gave us helping hand in meticulously organizing our words.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Linga Reddy Baddam
inVentiv Health Clinical
9th Floor, B-Block, Laxmi Cyber City, Kondapur-500084
Hyderabad, Andhra Pradesh, India.
+91-9652303033
linga.reddy@inventivhealth.com or lingareddy.baddam@gmail.com
http://www.inVentivHealthclinical.com

Sudarshan Reddy Shabadu
inVentiv Health Clinical
9th Floor, B-Block, Laxmi Cyber City, Kondapur-500084
Hyderabad, Andhra Pradesh, India.
+91-9703137367
sudarshan.reddy@inventivhealth.com or sudarshansas@gmail.com
http://www.inVentivHealthclinical.com