Критерий Хи-квадрат или точный критерий Фишера? Реши это с помощью SAS!

SAS (www.sas.com) – лидер на рынке приложений для статистического анализа и бизнес приложений нового поколения. Решения SAS используют более 40 000 компаний – 96 из них входят в 100 первых компаний списка FORTUNE Global 500®. Решения SAS используются для более доходных взаимоотношений между покупателями и поставщиками и для принятий наиболее правильных и обоснованных решений.

Исторически так сложилось, что критерий Хи-квадрат Пирсона был наиболее предпочтительней, чем точный критерий Фишера, вследствие того, что точный критерий Фишера требовал слишком много компьютерных вычислений, а следовательно ресурсов.

Когда Вы анализируете сопряженные таблицы размера 2x2, Вы можете использовать как критерий Хи-квадрат, так и точный критерий Фишера. Применение критерия Хи-квадрат правомерно, когда ожидаемая частота в любой клетке больше или равно 5, в противном случае нужно использовать точный критерий Фишера.

Но откуда узнать ожидаемые частоты содержатся в каждой клетке, в эпоху компьютеров, когда все вычисления за нас делает компьютер?

В этом нам поможет SAS!

Данный универсальный макрос предназначен для того, чтобы упростить процедуру выбора критерия. Разберем поподробнее:

```
/*****************************************************************
------------------+-----------------------------------------------
Переменные макроса        Описание
------------------+-----------------------------------------------
vrb1                      Variable 1 - Первая бинарная переменная
lbl1                      Label 1 - Метка для переменной 1
vrb2                      Variable 2 - Вторая бинарная переменная
lbl2                      Label 2 - Метка для переменной 2
tbl                       Table - Таблица в которой находятся данные
*****************************************************************/

%macro cmpr(vrb1, lbl1, vrb2, lbl2, tbl);

/* Выведем заголовок сравнения.      */

      title 'Сравнение ' &lbl2 ' и ' &lbl1;
      title2 'Исходные данные';

/* Присвоим метки переменным. */

      data &tbl;
            set &tbl;
            label &vrb2=&lbl2 &vrb1=&lbl1;
      run;

/* Выведем таблицу сопряженности 2x2. */

      proc tabulate data=&tbl format=8.1;     *Print Input Data;
            class &vrb1 &vrb2;
            freq count;
            keylabel n='N' pctn='%' all='Всего';
            table (&vrb1 all),(&vrb2 all)*(n*f=8.0 pctn<&vrb2 all>);
      run;
/* Рассчитаем значение p критерия Хи-квадрат, значение p точного    */
/* критерия Фишера, значения p других критериев. Поместим значения  */
/* p в таблицу tst. Ожидаемые частоты поместим в таблицу frq.       */
/* Результат процедуры на экран не выводим.                         */

      proc freq data=&tbl noprint;
            weight count;
            tables &vrb1*&vrb2/fisher nocol nopct out=frq outexpect;
            output out=tst chisq;
      run;

/* Оставляем только интересующие нас критерии: p_pchi - величина p  */
/* критерия Хи-квадрат, xp2_fish - величина p точного двустороннего */
```

```
/* критерия Фишера.                                                    */

      data tst;
            set tst;
            keep p_pchi xp2_fish;
      run;

/* Рассчитаем сколько ожидаемых частот меньших 5 */
/* содержится в таблице frq. */

      proc sql;
            create table exp as select count(expected) as cex
            from frq where expected<5;
      quit;

/* Совместим данные о критериях и величинах p
/* с данными об ожидаемых частотах.*/

      data tst;
            set tst;
            set exp;
      run;

/* Выберем какой критерий использовать, если переменная cex>0, */
/* тогда используем точный критерий Фишера, если cex=0, тогда  */
/* используем критерий Хи-квадрат.                             */

      data tst;
            set tst;
            if cex>0 then do;
                  tst=xp2_fish;
                  tstname='Точный двусторонний критерий Фишера';
            end;
            if cex=0 then do;
                  tst=p_pchi;
                  tstname='Критерий Хи-квадрат';
            end;
            label tst='Величина p' tstname='Критерий';
      run;

/* Выведем величину p и критерий, относящийся к ней */

      title2 'Сравнения';
      proc print data=tst label noobs;
            var tst tstname;
      run;
%mend cmpr;
```

Всего этого можно было не делать и рассчитать эти критерии, обычным способом, с помощью процедуры freq. Давайте посмотрим от чего мы избавились.

Возьмем абстрактные данные о различие 2х групп по полу:

```
data sex;
      input grp  $ sex $ count;
      cards;
Группа1      Женский      33
Группа1      Мужской      67
Группа2      Женский      8
Группа2      Мужской      42
;
run;
```

Сначала запустим процедуру freq без макроса:

```
proc freq data=sex;
      weight count;
      tables grp*sex/fisher nocol nopct;
run;
```

Что мы получим (приводится полная копия Output)?

```
The FREQ Procedure

Table of grp by sex

grp      sex(sex)

Frequency|
Row Pct  |Женский |Мужской |  Total
---------+--------+--------+
Группа1  |     33 |     67 |    100
         |  33.00 |  67.00 |
---------+--------+--------+
Группа2  |      8 |     42 |     50
         |  16.00 |  84.00 |
---------+--------+--------+
Total          41      109      150


Statistics for Table of grp by sex

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      4.8501    0.0276
Likelihood Ratio Chi-Square    1      5.1611    0.0231
Continuity Adj. Chi-Square     1      4.0319    0.0446
Mantel-Haenszel Chi-Square     1      4.8177    0.0282
Phi Coefficient                       0.1798
Contingency Coefficient               0.1770
Cramer's V                            0.1798


       Fisher's Exact Test
----------------------------------
Cell (1,1) Frequency (F)        33
Left-sided Pr <= F          0.9932
Right-sided Pr >= F         0.0202

Table Probability (P)       0.0134
Two-sided Pr <= P           0.0327

Sample Size = 150
```

Что выбрать? Вопросов нет, нужно найти и выбрать Chi-Square Test Prob. Минус - это неотформатированный отчет.

Что мы видим в Output, запустив наш макрос:

```
%cmpr(grp,'Группа',sex,'Пол',sex);
```

Output:

```
Сравнение Пол и Группа
Исходные данные
```

	Пол					
	Женский		Мужской		Всего	
	N	%	N	%	N	%
Группа						
Группа1	33	33.0	67	67.0	100	100.0
Группа2	8	16.0	42	84.0	50	100.0
Всего	41	27.3	109	72.7	150	100.0

```
Сравнение Пол и Группа
Сравнения

 Величина p      Критерий

0.027645        Критерий Хи-квадрат
```

Красивый отформатированный отчет, не требующий никаких усилий для выбора критерия и оформления.

Попробуем изменить данные так, чтобы появились ожидаемые частоты меньше 5, например:

```
data sex;
       input grp  $ sex $ count;
       cards;
Group1  Female  33
Group1  Male            67
Group2  Female  4
Group2  Male            5
;
run;
```

И проведем ту же процедуру:

```
proc freq data=sex;
       weight count;
       tables grp*sex/fisher nocol nopct;
run;
```

Output:

```
The FREQ Procedure

Table of grp by sex

grp     sex(sex)

Frequency|
Row Pct  |Женский |Мужской |  Total
---------+--------+--------+
Группа1  |     33 |     67 |    100
         |  33.00 |  67.00 |
---------+--------+--------+
Группа2  |      4 |      5 |      9
         |  44.44 |  55.56 |
---------+--------+--------+
Total          37       72      109


Statistics for Table of grp by sex

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      0.4823    0.4874
Likelihood Ratio Chi-Square    1      0.4650    0.4953
Continuity Adj. Chi-Square     1      0.1069    0.7437
Mantel-Haenszel Chi-Square     1      0.4779    0.4894
Phi Coefficient                      -0.0665
Contingency Coefficient               0.0664
Cramer's V                           -0.0665

WARNING: 25% 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)        33
Left-sided Pr <= F          0.3613
Right-sided Pr >= F         0.8554

Table Probability (P)       0.2167
Two-sided Pr <= P           0.4851

Sample Size = 109
```

Какой же теперь критерий выбрать? Под таблицей «Statistics for Table» мы видим предупреждение о невозможности применения критерия Хи-квадрат, следовательно, нужно найти и выбрать Fisher Exact Test Two-sided Pr<=P и опять отчет неотформатирован!

Попробуем наш макрос:

```
%cmpr(grp,'Группа',sex,'Пол',sex);
```

Output:

```
Сравнение Пол и Группа
Исходные данные
```

	Пол					
	Женский		Мужской		Всего	
	N	%	N	%	N	%
Группа						
Группа1	33	33.0	67	67.0	100	100.0
Группа2	4	44.4	5	55.6	9	100.0
Всего	37	33.9	72	66.1	109	100.0

Сравнение Пол и Группа
Сарвнения

Величина p	Критерий
0.48514	Точный двусторонний критерий Фишера

Выбор за Вами!

Если Вы хотите, чтобы мы провели статистический анализ Ваших данных звоните +7 (910) 720-2977 info@stathelp.ru

Иван Гудков.

Об авторе:

Иван Гудков, начальник отдела статистики компании «Статистическая помощь», занимающейся статистической обработкой данных различных исследований. В сферу деятельности компании входит помощь в написании протоколов исследований, помощь в составлении индивидуальных регистрационных карт, расчет оптимального размера выборки исследований, составление рандомизационного плана, создание баз данных на основе ИРК, двойной ввод карт в базу данных, описательная статистика данных, расчет существования статистической значимости различия данных, корреляция признаков, регрессия, подготовка отчетов статистического исследования (www.stathelp.ru). Он является профессионалом в статистической обработке данных, использует в своей работе SAS, MS SQL Server, Transact SQL. Связаться с Иваном Вы можете по телефону: +7 (910) 720-2977 или ivan.gudkov@stathelp.ru.