Two-way Contingency Tables

Considering two categorical variables at a time leads to a cross-classification called a contingency table. An example from the survey data might be a 2 x 2 table of gender by degree. When a cross classification occurs, two common questions arise. The first considers whether the two factors are associated or independent. In the example above this would address whether degree is influenced by gender (this is analogous to an interaction in continuous data). The second question examines the homogeneity of frequency distributions across a factor. For example, the question may be asked: is the distribution of genders consistent for each degree? In order to answer this last question the data should meet an assumption that each gender was sampled with equal precision (i.e. an equal number of females and males were sampled). Although these questions differ in objectives and assumptions, the respective tests of null hypotheses for independence or homogeneity of distributions turn out to be the same. A P^2 statistic is calculated similar to that of the one-way test (handout #1) except the expected value is the product of the marginal totals divided by the overall total. In this case the Procedure FREQ will be employed. The decision of which hypothesis is actually being tested is left to the discretion of the user.

Statements

TABLES

The essential statement of PROC FREQ is the TABLES statement. This is the place where the user specifies the cross classifications to be considered. The general form is:

TABLES factor1*factor2/ *options*;

Although more than two factors can be included in the TABLES statement, PROC FREQ will produce tables and associated tests for only the last two factors at fixed levels of the preceding factors. Thus, a TABLES statement with factor1*factor2*factor3 will produce a factor2*factor3 table for each level of factor1.

PROC FREQ can output a bewildering array of information. By default each two factor table will have cell frequencies, cell percentages, percentage of row total for a cell, percentage of column total for a cell, and similar frequencies and percentages for rows and columns. If these are not of interest, they may be turned off with the appropriate options in the TABLES statement. For example the options NOPERCENT and NOFREQ would exclude cell frequencies

and percentages from the output. Refer to the SAS STAT help system for a complete list of options.

Other TABLES statement options will control the hypothesis testing procedures. For the tests mentioned above the option CHISQ is used. This will produce the Pearson P^2 statistic as described, along with other P^2 statistics produced by different statistical methods. These additional statistics may or may not be appropriate. The Pearson chi-square, likelihood ratio chi-square, and continuity adj. chi-square are used for the hypotheses outlined earlier. Special hypotheses involving ordinal variables can also be specified (Mantel-Haenszel, gamma, Kendall's tau, Stuart's tau, etc). For the tests of independence or homogeneity, the options CELLCHI2, DEVIATION, and EXPECT may also be of interest because they examine how each cell is influencing the overall value of the P^2 statistic.

WEIGHT

By default PROC FREQ expects the input data to be in the binary form as described earlier. If the data is in the summarized count form, a WEIGHT statement is required. The variable in the WEIGHT statement represents the number of times each observation occurs. An example would be:

```
PROC FREQ;
     WEIGHT COUNTS;
     TABLES factor1*factor2/CHISQ;
```

The WEIGHT statement occurs before the TABLES statement and lists only one variable. [Note the WEIGHT statement for PROC FREQ does **not** preform the same function as in other procedures such as GLM and REG].

Example.

Test for independence of DEGREE and FIELD.

```
PROC FREQ DATA=SURVEY;
     WEIGHT COUNT;
     TABLES DEGREE*FIELD/CHISQ DEVIATION CELLCHI2;

RUN;
```

Other Features

Output

It is sometimes useful to output the results of PROC FREQ to a new SAS data set. The TABLES statement includes an option to do this:

TABLES factor1*factor2 / OUT = *NEW DATASET NAME*;

The new data set name will contain the factor1 and factor2 variables in addition to two new variables COUNT and PERCENT. Thus, this is another method of summarizing a binary data form to the count form.

Example.

Summarizing data from binary to count form.

```
DATA BINARY;
     INPUT VAR1 VAR2;
     CARDS;
     1    1
     1    1
     1    1
     1    2
     2    1
     2    1
     2    2
     2    2
     ;
PROC FREQ DATA=BINARY;
     TABLES VAR1*VAR2/OUT=CNTS NOPRINT;
```

Output from example.

VAR1	VAR2	COUNT	PERCENT
1	1	3	37.50
1	2	1	12.50
2	1	2	25.00
2	2	2	25.00