CROSS TABULATE

PURPOSE

Generates a cross tabulation of a response variable for two independent variables.

DESCRIPTION

The independent variables are mutually exclusive categories which form a two-way table. The response variable must fall into exactly one row and column of this table. By default, this command calculates the counts for each row and column combination. Alternatively, it can calculate the mean, standard deviation, sum, or range of each row and column combination. In addition, it can perform a chi-square test for the independence of two qualitative variables.

The response (i.e., dependent) variable, if specified, is typically a quantitative variable while the 2 independent variables are typically categorical variables that are coded as integer values (e.g., 1 for male, 2 for female).

SYNTAX 1

CROSS TABULATE <y1> <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
CROSS TABULATE <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
CROSS TABULATE COUNTS <y1> <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
CROSS TABULATE COUNTS <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
where <y1> is a response variable;
<tag1> is the first group identifier variable;
<tag2> is the second group identifier variable;
and where the <SUBSET/EXCEPT/FOR qualification> is optional.

This syntax generates a count of the number of elements in each row and column combination. Specifying the response variable (<y1>) is optional (and usually omitted) since it is not used in the calculation for the counts.

SYNTAX 2

CROSS TABULATE MEANS <y1> <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
where <y1> is a response variable;
<tag1> is the first group identifier variable;
<tag2> is the second group identifier variable;
and where the <SUBSET/EXCEPT/FOR qualification> is optional.

This syntax computes the mean of the elements in the response variable (<y1>) for each row and column combination.

SYNTAX 3

CROSS TABULATE RANGE <y1> <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
where <y1> is a response variable;
<tag1> is the first group identifier variable;
<tag2> is the second group identifier variable;
and where the <SUBSET/EXCEPT/FOR qualification> is optional.

This syntax computes the range of the elements in the response variable (<y1>) for each row and column combination.

SYNTAX 4

CROSS TABULATE SD <y1> <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
where <y1> is a response variable;
<tag1> is the first group identifier variable;
<tag2> is the second group identifier variable;
and where the <SUBSET/EXCEPT/FOR qualification> is optional.

This syntax computes the standard deviation of the elements in the response variable (<y1>) for each row and column combination.

SYNTAX 5

CROSS TABULATE SUM <y1> <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
where <y1> is a response variable;
<tag1> is the first group identifier variable;
<tag2> is the second group identifier variable;
and where the <SUBSET/EXCEPT/FOR qualification> is optional.

This syntax computes the sum of the elements in the response variable (<y1>) for each row and column combination.

SYNTAX 6

CROSS TABULATE CHI-SQUARE <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
CROSS TABULATE CHI-SQUARE <y1> <tag1> <tag2> <SUBSET/EXCEPT/FOR qualification>
where <y1> is a response variable;
<tag1> is the first group identifier variable;
<tag2> is the second group identifier variable;
and where the <SUBSET/EXCEPT/FOR qualification> is optional.

This syntax performs a chi-square test for independence of two variables. This test is described in most introductory statistics books. Specifying the response variable (<y1>) is optional (and usually omitted) since it is not used in the calculation for the counts. If all the counts are 1, it assumes that <tag1> and <tag2> contain previously computed frequencies. If at least one of the counts is greater than 1, it assumes the analyst is providing raw data and it calculates the frequencies before calculating the chi-square statistic.

EXAMPLES

CROSS TABULATE TAG1 TAG2
CROSS TABULATE TAG1 TAG2 SUBSET TAG2 = 2 TO 4

NOTE 1

The formula for the chi-square test is:

$$T = \sum_i \frac{(o_i - e_i)^2}{e_i} \qquad \textbf{(EQ 3-41)}$$

where o_i is the observed frequency for a given cell and e_i is the expected frequency for a given cell. The expected frequency is the row total times the column total divided by the grand total. The test statistic is compared to a chi-square distribution with (r-1)(c-1) degrees of freedom where r is the number of rows and c is the number of columns.

For 2x2 tables, the test statistic is also computed with the Yates continuity correction applied. This correction is recommended for 2x2 tables with low expected frequencies (between 5 and 10). Exact tests are recommended for frequencies less than 5.

NOTE 2

In some cases, it may be desirable to use the generated cross tabulation in further analysis. Although DATAPLOT does not save these values in internal variables or automatically write them to a file, they can be retrieved. Enter the CAPTURE <file> command before the CROSS TABULATE command (and END OF CAPTURE after) to write the values to a file.

For the counts, mean, standard deviation, range, and sum cases, enter the following commands (assume the capture file is JUNK.DAT):

```
SET READ FORMAT F15.6,F15.6,3X,F15.6
SKIP 3
READ FORMAT JUNK.DAT GROUP1 GROUP2 STAT
SKIP 0; SET READ FORMAT
```

The variable STAT will contain the value of the desired statistic for each row and column combination stored in the variables GROUP1 and GROUP2.

Since the chi-square case contains the test output, you will need to determine the last row that contains the count information (the NLIST command can be used for this purpose). In the following, NLAST should be set to the line number of the last line in the table. Then enter the following commands:

```
SET READ FORMAT F15.6,F15.6,3X,3F10.0,F15.6
LET NLAST = <value>
ROW LIMITS 4 NLAST
READ FORMAT JUNK.DAT GROUP1 GROUP2 COUNTS ROWTOT COLTOT EXPECTED
SKIP 0; SET READ FORMAT
```

The variable COUNTS will contain the frequency for each row and column combination stored in the variables GROUP1 and GROUP2. The variables ROWTOT, COLTOT, and EXPECTED will contain the corresponding row totals, column totals, and expected number of observations, respectively.

NOTE 3

DATAPLOT does not currently support log-linear analysis of contingency tables.

DEFAULT

None

SYNONYMS

None

RELATED COMMANDS

TABULATE	=	Generate a tabulation for a one-way table.
ANOVA	=	Performs an ANOVA.

APPLICATIONS

Exploratory Data Analysis

IMPLEMENTATION DATE

The output format for the CROSS TABULATE CHI-SQUARE was modified and the CROSS TABULATE SUM command was fixed 94/2.

PROGRAM

```
. THIS IS DATAPLOT DATA FILE RIPKEN.DAT
. CAL RIPKEN BATTING AVERAGE SENSITIVITY
. RESPONSE VARIABLE = BATTING AVERAGE
.   1. BATTING AVERAGE = RESPONSE VARIABLE
.   2. PITCH TYPE (1 = FASTBALL, 2 = CURVE)
.   3. PITCHING HAND (1 = LEFT, 2 = RIGHT)
. SOURCE--
.   SECONDARY--THE BALTIMORE SUN, APRIL 5, 1992
.   PRIMARY  --BASEBALL ANALYSIS & REPORTING SYSTEM
READ BA TYPE HAND
 0.400    1    1
 0.354    1    1
 0.388    1    1
 0.380    1    1
 0.409    1    1
 0.391    1    1
 0.136    1    1
 0.322    1    1
 0.304    1    1
 0.166    2    1
 0.333    2    1
 0.000    2    1
 0.300    2    1
 0.875    2    1
 1.000    2    1
 0.166    2    1
 0.090    2    1
 0.333    2    1
 0.382    1    2
 0.373    1    2
 0.223    1    2
 0.333    1    2
 0.350    1    2
 0.304    1    2
 0.086    1    2
 0.280    1    2
 0.218    1    2
 0.166    2    2
 0.363    2    2
 0.333    2    2
 0.300    2    2
 0.454    2    2
 0.285    2    2
 0.162    2    2
 0.211    2    2
 0.000    2    2
END OF DATA
CAPTURE CROSSTAB.DAT
CROSS TABULATE TYPE HAND
CROSS TABULATE MEANS BA TYPE HAND
CROSS TABULATE SD BA TYPE HAND
CROSS TABULATE RANGE BA TYPE HAND
CROSS TABULATE CHI-SQUARE TYPE HAND
```

This command generates the following output.

```
                                *        COUNTS
************************************************
       1.000000       1.000000 *        9.000000
       1.000000       2.000000 *        9.000000
       2.000000       1.000000 *        9.000000
       2.000000       2.000000 *        9.000000

                                *        MEAN
************************************************
       1.000000       1.000000 *        0.342667
       1.000000       2.000000 *        0.283222
       2.000000       1.000000 *        0.362556
       2.000000       2.000000 *        0.252667

                                *        STAN. DEV.
************************************************
       1.000000       1.000000 *        0.085319
       1.000000       2.000000 *        0.094851
       2.000000       1.000000 *        0.345976
       2.000000       2.000000 *        0.134050

                                *        RANGE
************************************************
       1.000000       1.000000 *        0.273000
       1.000000       2.000000 *        0.296000
       2.000000       1.000000 *        1.000000
       2.000000       2.000000 *        0.454000

                                                    ROW     COLUMN
                                *     COUNTS      TOTAL      TOTAL        EXPECTED
************************************************************************************
       1.000000       1.000000 *          9         18         18        9.000000
       1.000000       2.000000 *          9         18         18        9.000000
       2.000000       1.000000 *          9         18         18        9.000000
       2.000000       2.000000 *          9         18         18        9.000000


                CHI-SQUARED TEST FOR INDEPENDENCE
TEST:
   CHI-SQUARED STATISTIC                 =          0.
      WITH YATES CONTINUITY CORRECTION   =   0.1111111
   DEGREES OF FREEDOM                    =    1.000000
   CHI-SQUARED CDF VALUE                 =          0.
      WITH YATES CONTINUITY CORRECTION   =   0.2611172

   HYPOTHESIS    ACCEPTANCE INTERVAL    CONCLUSION
INDEPENDENT        (0.000,0.950)        ACCEPT

   WITH YATES CONTINUITY CORRECTION
   HYPOTHESIS    ACCEPTANCE INTERVAL    CONCLUSION
INDEPENDENT        (0.000,0.950)        ACCEPT
```