Cross Tabulation and Chi-square

The **CROSS TABULATION** command displays the joint distribution of two or more variables (*multivariate frequency distribution*) in a matrix format and thus allows to compare the relationship between the variables. Cross tabulation table (also known as contingency or crosstab table) is generated for each distinct value of a layer variable (optional) and contains counts and percentages. Chi-square test is used to check if the results of a cross tabulation are statistically significant. To run the chi-square test for tabulated data run the **CHI-SQUARE TEST (SUMMARIZED DATA)** command from the Nonparametric Statistics menu (v6.4+).

How To

- ✓ Run the **STATISTICS -> BASIC STATISTICS -> CROSS TABULATION AND CHI-SQUARE** command.
- ✓ Select a **ROW** variable (containing the categories that define the rows of the table) and a **COLUMN** variable (containing the categories that define the columns of the table).
- ✓ Optionally, select a **FREQUENCY** variable. Frequency variable specifies the number of observations that each row represents. When omitted, each row represents a single observation.
- ✓ Optionally, select a **LAYER** variable. Layer variable distinct levels (values) cause separate tables generated. The layer variable is also called the break variable, control variable or filter variable.
- ✓ Optionally, in the **ADVANCED OPTIONS** select the **PRINT TABLES** option value. This option allows to choose which tables are printed. Chi-square test summary and three tables (observed frequencies, expected frequencies and chi-squared values) are printed with any of these options. Available options are listed below.
    - **NONE**: No additional tables are printed.
    - **COMBINED FREQUENCY TABLE**: Contingency table (combined frequency table) with counts and cell percentages is printed
    - **SEPARATE PERCENTAGE TABLES:** Marginal proportion tables (row proportions, column proportions) and proportion of total table are printed in place of combined frequency table.
    - **ALL**: Three proportion tables and the contingency table are printed.
- ✓ **Casewise** deletion is used for missing values removal.

Results

In a two-way frequency table entries are frequency counts. Entries in the "*Total*" row and "*Total*" column are called marginal totals.

OBSERVED FREQUENCIES TABLE

Observed frequency is the number of times that a particular combination of categories occurred.

EXPECTED FREQUENCIES TABLE

Expected frequency is the number of observations that would be expected for a particular combination of categories if the null hypothesis were true (combination were to occur by chance). The formula for expected frequency in the i^{th} row and j^{th} column is:

$$E_{i,j} = \frac{T_i \cdot T_j}{N},$$

where T_i is the total in the ith row, T_j is the total in the jth column and *N* is the table grand total.

CROSS-TAB TABLE

Table entries consist of frequency, row and column percentages, and total percentage (denominator is the total number of observations in the table).

Row Variable	**Speaking**		**Combined Table Cell Contents**			
Column Variable	**Age**		*N (Observed)*			
Frequency Variable	**Frequency**		*N / Row totals*			
Layer Variable	**#N/A**		*N / Column totals*			
			N / Total			

Combined Frequency Table

	Column "20-39"	Column "40-59"	Column "60-79"	Column "80+"	Column "<20"	Row totals
Row "Not At All"	**557.**	**1,918.**	**895.**	**991.**	**24.**	**4,385.**
	0.1270	0.4374	0.2041	0.2260	0.0055	
	0.0188	0.0428	0.0280	0.0751	0.0022	
	0.0043	0.0147	0.0068	0.0076	0.0002	0.0335
Row "Not Well"	**10,089.**	**14,003.**	**15,417.**	**6,227.**	**2,862.**	**48,598.**
	0.2076	0.2881	0.3172	0.1281	0.0589	
	0.3399	0.3122	0.4826	0.4722	0.2578	
	0.0772	0.1071	0.1179	0.0476	0.0219	0.3717
Row "Well"	**19,036.**	**28,925.**	**15,634.**	**5,969.**	**8,214.**	**77,778.**
	0.2447	0.3719	0.2010	0.0767	0.1056	
	0.6413	0.6450	0.4894	0.4526	0.7400	
	0.1456	0.2212	0.1196	0.0456	0.0628	0.5948
Column totals	**29,682.**	**44,846.**	**31,946.**	**13,187.**	**11,100.**	**130,761.**
	0.2270	0.3430	0.2443	0.1008	0.0849	

CHI-SQUARE TEST SUMMARY

Chi-square statistic is a measure of how close the observed frequencies are to the expected frequencies. It is defined as $\chi = \sum \frac{(O-E)^2}{E}$, where *O* is an observed frequency, *E* is an expected frequency, sum is across all cells.

D.F. – degrees of freedom. The number of degrees of freedom is defined as: $d.f. = (r-1)(c-1)$, where *r* is the number of rows and *c* is the number of columns.

If the p-level (**P-LEVEL > X**) is less than selected α (0.05) the test is significant and null hypothesis is rejected, and it can be concluded that there is an association (dependence) between the row variable and the column variable. **NULL HYPOTHESIS H_0** states that the row and column variables are independent.

References

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (2007). Discrete multivariate analysis: Theory and practice. New York, NY: Springer-Verlag (Original work published in 1975).