Cross Tabulation Analysis[1]

Cross-tabulation is one of the most useful analytical tools and is a main-stay of the market research industry. One estimate is that single variable frequency analysis and cross-tabulation analysis account for more than 90% of all research analyses.

Cross-tabulation analysis, also known as contingency table analysis, is most often used to analyze categorical (nominal measurement scale) data. A cross-tabulation is a two (or more) dimensional table that records the number (frequency) of respondents that have the specific characteristics described in the cells of the table. Cross-tabulation tables provide a wealth of information about the relationship between the variables.

Cross-tabulation analysis has its own unique language, using terms such as "banners", "stubs", "Chi-Square Statistic" and "Expected Values."

A typical cross-tabulation table comparing the two hypothetical variables "City of Residence" with "Favorite Baseball Team" is shown below. Are city of residence and being a fan of that city independent? The cells of the table report the frequency counts and percentages for the number of respondents in each cell.

Cross tabulation Frequency Percent		What is Your Favorite Baseball Team?			
		Toronto Blue Jays	Boston Red Socks	New York Yankees	Row Totals
In What City Do You Reside?	Boston, MA	11	33	7	51
	Row Percent	21.57%	64.71%	13.73%	34.93%
	Montreal, Canada	23	14	9	46
	Row Percent	50.00%	30.43%	19.57%	31.51%
	Montpellier, VT	22	13	14	49
	Row Percent	44.90%	26.53%	28.57%	33.56%
	Column totals	56	60	30	146
	Column Percent	38.36%	41.10%	20.55%	100.00%

In the above table, the text legend describes the row and column variables. You can create and analyze multiple tables in a side-by-side or sequential format. Tabulation Professionals call the column variables in these multiple tables "Banners" and row variables "Stubs".

[1] Version November 1, 2011

CROSS-TABULATION WITH CHI-SQUARE ANALYSIS

The Chi-square statistic is the primary statistic used for testing the statistical significance of the cross-tabulation table. Chi-square tests whether or not the two variables are independent. If the variables are independent (have no relationship), then the results of the statistical test will be "non-significant" and we "are not able to reject the null hypothesis", meaning that we believe there is no relationship between the variables.

If the variables are related, then the results of the statistical test will be "statistically significant" and we "are able to reject the null hypothesis", meaning that we can state that there is some relationship between the variables.

The chi-square statistic, along with the associated probability of chance observation, may be computed for any table. If the variables are related (i.e. the observed table relationships would occur with very low probability, say only 5%) then we say that the results are "statistically significant" at the ".05 or 5% level". This means that the variables have a low chance of being independent.

Students of statistics will recall that the probability values (.05 or .01) reflect the researcher's willingness to accept a type I error, or the probability of rejecting a true null hypothesis (meaning that we thought there was a relationship between the variables when there really wasn't). Furthermore these probabilities are cumulative, meaning that if 20 tables are tested, the researcher can be almost assured that one of the tables is incorrectly found to have a relationship (20 x .05 = 100% chance). Depending on the cost of making mistakes, the researcher may apply more stringent criteria for declaring "significance" such as .01 or .005.

COMPUTATION OF THE CHI-SQUARE STATISTIC FOR CROSS-TABULATION TABLES

The chi-square statistic is computed by first computing a chi-square value for each individual cell of the table and then summing them up to form a total Chi-square value for the table. The chi-square value for the cell is computed as:

(Observed Value − Expected Value)^2 / (Expected Value)

The Chi-Square computations are highlighted in gray.

In this example table, we observe that the chi-square value for the table is 19.35, and has an associated probability of occurring by chance less than one time in 1000. We therefore reject the null hypothesis of no difference and conclude that there must be a relationship between the variables.

We can observe the relationship in two places in the table. The most obvious is in the chi-square value computed for each cell. We observe that the cells "Red Socks and Boston", "Blue Jays and Montreal" and "Red Socks and Montpellier, Vermont" were the three cells where the number of observed respondents was greater than expected. We further note that when we examine the expected and observed frequencies, the "Yankees and Montreal", "Red Socks and Montpellier, Vermont", and "Red Socks and Montreal" frequencies were fewer than expected.

Cross tabulation Frequency Percent		What is Your Favorite Baseball Team?			Row Totals
		Toronto Blue Jays	Boston Red Socks	New York Yankees	
	Boston, MA	11	33	7	51
	Row Percent	21.57%	64.71%	13.73%	34.93%
	Expected Value	19.56	20.96	10.48	
	Cell Chi-Square	3.75	6.92	1.16	
	Computations:	Column total percent	Column total percent	Column total percent	
	Expected Value	(.3836*51)	(.4110*51)	(.2055*51)	
	Cell Chi-Square	(11-19.56)^2/19.56	(33-20.96)^2/20.96	(7-10.48)^2/10.48	
	Montreal, Canada	23	14	9	46
	Row Percent	50.00%	30.43%	19.57%	31.51%
	Expected Value	17.64	18.90	9.45	
	Cell Chi-Square	1.63	1.27	0.02	
In What City Do You Reside?	Computation of Cell	Column total percent	Column total percent	Column total percent	
	Expected Value	(.3836*46)	(.4110*46)	(.2055*46)	
	Cell Chi-Square	(23-17.64)^2/17.64	(14-18.90)^2/18.90	(9-9.45)^2/9.45	
	Montpilliar, VT	22	13	14	49
	Row Percent	44.90%	26.53%	28.57%	33.56%
	Expected Value	18.79	20.14	10.07	
	Cell Chi-Square	0.55	2.53	1.54	
	Computation of Cell	Column total percent	Column total percent	Column total percent	
	Expected Value	(.3836*49)	(.4110*49)	(.2055*49)	
	Cell Chi-Square	(22-18.79)^2/18.79	(13-20.14)^2/20.14	(14-10.07)^2/10.07	
	Column totals	56	60	30	146
	Column Percent	38.36%	41.10%	20.55%	100.00%

The sum of all cell Chi-Square Values = Table Chi-Square:	19.35
Degrees of Freedom = (# Rows -1)*(# Columns -1) = (3-1)*(3-1) =	4
Chi-Square Probability of Independence is not computed here, but can be looked up from a Chi-Square probability table for	
19.35 and 4 df = .00067	0.00067

Because the cell chi-square and the expected values are often not displayed, these same relationships can be observed by comparing the column total percent to the cell percent (of the row total). In cell "Red Socks and Boston" we would compare 41.10% with 64.71% and observe that more Red Socks fans liked Boston than expected.

Caution is urged when interpreting relationships found in any statistical analysis. We often desire to "explain" or conclude "causality" from analyses when data either is not designed to, or does not have the power to support such conclusions. In the current table we observe that "Red Socks and Boston" had the greatest delta between the number of observed and expected respondents, for any team preference and city of residence. However we must be careful in concluding that the Red Socks caused respondents to move to Boston, or that Boston as a city of residence causes fan loyalty. Red Socks and Boston are the most observed fan and city relationship, but are most likely totally independent when considering other concepts or relationships.

We're Here to Help!

Qualtrics.com provides the most advanced online survey building, data collection (via panels or corporate / personal contacts), real-time view of survey results, and advanced "dashboard reporting tools".

If you are interested in learning more about how the Qualtrics professional services team can help you with a conjoint analysis research project, contact us at research@qualtrics.com.