Custom Tables

This document describes the algorithms used in the Custom Tables procedure.

A note on weights and multiple response sets

Case weights are always based on Counts, not Responses, even when one of the variables is a multiple response variable.

Pearson's Chi-Square

Notation

The following notation is used for the computation of Pearson's chi-square:

R	Number of rows in the sub-table.
C	Number of columns in the sub-table.
f_{ij}	Sum of case weights in cell (i,j).
r_i	Marginal case weights total in i-th row.
c_j	Marginal case weights total in j-th column.
W	Marginal case weights total in the sub-table.
E_{ij}	Expected cell counts.
χ_p^2	Pearson's Chi-Square statistic.
p_{ij}	Population proportion for cell (i,j).
$p_{i.}$	Marginal population proportion for i-th row.
$p_{.j}$	Marginal population proportion for j-th column.
df	Degrees of Freedom.
p	p-value of the chi-square test.

α Significance level supplied by the user.

Conditions and assumptions

- Tests will not be performed on Comperimeter tables.
- Chi-square tests are performed on each innermost sub-table of each layer.
- If a scale variable is in the layer, that layer will not be used in analysis.
- The row variable and column variable must be two different categorical variables or multiple response sets.
- The contingency table must have at least two non-empty rows and two non-empty columns.
- Non-empty rows and columns do not include subtotals and totals.
- Empty rows and columns are assumed to be structural zeros. Therefore, R and C are the numbers of non-empty rows and columns in the table.
- If weighting is on, cell statistics must include weighted cell counts or weighted simple row/column percents; the analysis will be performed using these weighted cell statistics. If weighting is off, cell statistics must include cell counts or simple row/column percents; the analysis will be unweighted.
- Tests are constructed by using all visible categories. Hiding of categories and showing of user-missing categories are respected.

Statistics

Hypothesis:

$$H_0 : p_{ij} = p_{i.}p_{.j} \ \ i = 1,\ldots,R \ \text{ and } \ j = 1,\ldots,C \ \text{ vs. not } H_0$$

Statistic:

$$\chi_p^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(f_{ij}-E_{ij})^2}{E_{ij}}, \text{ where } \mathrm{E_{ij}} = \frac{\mathrm{r_i c_j}}{\mathrm{W}}.$$

Under the null hypothesis, the statistic has a Chi-square distribution with df=(R-1)(C-1) degrees of freedom.

Alternatively, the chi-square statistics and degrees of freedom can be computed as the following,

$$\chi_\mathrm{p}^2 = \sum_{\mathrm{E_{ij}}>0}\frac{(\mathrm{f_{ij}}-\mathrm{E_{ij}})^2}{\mathrm{E_{ij}}},$$

R = #{ $\mathrm{r_i}$ >0} and C = #{ $\mathrm{c_j}$ >0}.

This avoids scanning for empty rows and columns before computations.

Categorical variable in rows and multiple response set in columns

$$\chi_p^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(f_{ij}-E_{ij})^2}{E_{ij}(1-\frac{c_j}{W})}$$

Under the null hypothesis, the statistic has an approximate Chi-square distribution with $df = (R-1)C$ degrees of freedom.

Multiple response set in rows and categorical variable in columns

$$\chi_p^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(f_{ij}-E_{ij})^2}{E_{ij}(1-\frac{r_i}{W})}$$

Under the null hypothesis, the statistic has an approximate Chi-square distribution with $df = R(C-1)$ degrees of freedom.

Multiple response sets in rows and columns

$$\chi_p^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(f_{ij}-E_{ij})^2}{E_{ij}(1-\frac{r_i}{W})(1-\frac{c_j}{W})}$$

Under the null hypothesis, the statistic has an approximate Chi-square distribution with $df = RC$ degrees of freedom.

P-value

$$\mathrm{p} = 1 - \mathrm{F}(\chi_\mathrm{p}^2;\mathrm{df}),$$

where $\mathrm{F}(\mathrm{x};\mathrm{df})$ is the cumulative distribution function of Chi-square distribution with df degrees of freedom.

The chi-square test is significant if the $\mathrm{p} < \alpha$.

Use of case weights:

The case weights (or frequency weights) are supposed to be integers representing number of replications of each case. In chi-square tests, we will only check if the aggregated cell counts f_{ij} are integers. If not, they will be rounded to nearest integer before computations.

Small sample validity of the test

Pearson's chi-square is a large sample test, it may not be valid when sample size is small. A rule of thumb is to check if there are more than 80% of cells have expected cell counts larger than 5 and expected cell counts are all larger than 1.

Test statistics for multiple response sets

The formulas above use a variation of the Pearson chi-square test statistics developed for a combination of categorical variable and a multiple response set as initially suggested by Agresti and Liu (1999). Formulas and properties of this test can be found in a comparative study by Bilder et al. (2000).

An extension of this approach when both variables are multiple response sets is given in the paper by Thomas and Decady (2004). It contains a study of the test properties as well as additional references.

References

Agresti, A. and Liu, I.-M. (1999), “Modeling responses to a categorical variable allowing arbitrarily many category choices”, Biometrics, 55, 936-943.

Bilder, C.R., Loughin, T.M. and Nettleton, D. (2000), “Multiple marginal independence testing for pick any/c variables”, Communications in Statistics: Simulation, 29, 1285-1316.

Thomas, D.R. and Decady, Y.J. (2004), “Testing for association using multiple response survey data: approximate procedures based on Rao-Scott Approach”, International Journal of Testing, 4, 43-59.

Column Proportions Test

Notation

The following notation is used for the computation of Column Proportions Tests:

R	Number of rows in the sub-table.
C	Number of columns in the sub-table.
A_i	i-th category of the row variable.

B_j	j-th category of the column variable.
f_{ij}	Total case weights in cell (i,j).
c_j	Marginal case weights total in j-th column.
$\tilde{c}_j$	Rounded marginal case weights total in j-th column.
z	z-statistic.
χ^2	Chi-Square statistic.
p_{ij}	Column proportion for cell (i,j).
$\hat{p}_{ij}$	Estimated column proportion for cell (i,j).
$\hat{p}_{ijk}$	Estimate of pooled column proportion of j and k-th column in i-th row.
p	p-value of a test.
p_B	Bonferroni corrected p-value.
α	The significance level supplied by the user.

Conditions and Assumptions

- Tests will not be performed on Comperimeter tables and tables with scale variables in the layer.

- Pairwise tests are performed on each row of all eligible innermost sub-tables within each layer.

- Sub-tables must have categorical variables or multiple response sets in both rows and columns.

- Number of rows and columns must be larger than or equal to two. i.e. $R \geq 2$ and $C \geq 2$.

- Tests are constructed by using all visible categories excluding totals and sub-totals. Hiding of categories and showing of user-missing categories are respected.

- If weighting is on, cell statistics must include weighted cell counts or weighted simple column percents; a weighted analysis will be performed. If weighting is off, cell statistics requested must include cell counts or simple column percents; an unweighted analysis will be performed.

- A proportion will be discarded if the proportion is equal to zero or one, or the sum of case weights in a category is less than 2, (i.e. $c_j < 2$). If less than two proportions are left after discarding proportions, test will not be performed.

Statistics

Table layout:

	B_1	**B_2**	**...**	**B_C**
A_1	P_{11}	p_{12}		p_{1C}
A_2	P_{21}	p_{22}		p_{2C}
...	...	...	...	...
A_R	p_{R1}	p_{R2}	...	p_{RC}

Hypothesis:

Without lost of generality, we will only look at the i-th row of the table. Let C* be the number of categories in the i-th row where the proportion is greater than zero and less than one, and where the sum of case weights in the corresponding column is at least 2. In the i-th row, C*(C*-1)/2 comparisons will be made among $p_{i1}, p_{i2}, \ldots, p_{iC}$. The (j,k)th hypothesis will be

$$H_{0jk} : p_{ij} = p_{ik} \textbf{ vs. } H_{1jk} : p_{ij} \neq p_{ik} .$$

Aggregated Statistics:

Column proportions tests are based on the aggregated proportions ($\hat{p}_{ij}$) and cell counts for each column (c_j). Column proportions are computed using the unrounded cell counts $\hat{p}_{ij} = \dfrac{f_{ij}}{c_j}$ which are equal to the proportions actually displayed in CTABLE.

Statistics for the (i,j)th comparisons:

Pooled proportion: $\hat{p}_{ijk} = \dfrac{\tilde{c}_j \hat{p}_{ij} + \tilde{c}_k \hat{p}_{ik}}{\tilde{c}_j + \tilde{c}_k}$.

z statistic with a categorical variable in the columns:

$$z = \frac{(\hat{p}_{ij} - \hat{p}_{ik})}{\sqrt{\hat{p}_{ijk}(1-\hat{p}_{ijk})(\frac{1}{\tilde{c}_j} + \frac{1}{\tilde{c}_k})}}.$$

When multiple response set defines columns there may exist cases that belong to both j-th and k-th columns. Let $\tilde{c}_{jk}$ be the rounded sum of weights for such cases.

z statistic with a multiple response set in the columns:

$$z = \frac{(\hat{p}_{ij} - \hat{p}_{ik})}{\sqrt{\hat{p}_{ijk}(1-\hat{p}_{ijk})(\frac{1}{\tilde{c}_j} + \frac{1}{\tilde{c}_k} - \frac{2\tilde{c}_{jk}}{\tilde{c}_j \tilde{c}_k})}}.$$

p-value: $p = 2[1 - \Phi(|z|)]$,

where $\Phi(z)$ is the CDF of standard normal distribution.

Alternatively, the statistics can be constructed as a chi-square statistic,

$$\chi^2 = z^2,$$

the p-value will now be given by $p = 1 - F(\chi^2;1)$, where $F(x;df)$ is the CDF of chi-square distribution with df degrees of freedom.

A comparison is significant if $p < \alpha$ (or $p_B < \alpha$, if Bonferroni adjusted).

Bonferroni adjustment:

If Bonferroni adjustment for multiple comparisons is requested, the p-value p will be adjusted by

$$p_B = \min(\frac{p * C * (C * -1)}{2}, 1)$$

Relationship to Pearson's chi-square tests:

With a categorical variable in the columns, the statistics used in column proportion tests is equivalent to the Pearson's chi-square test on a 2x2 table by taking j and k-th column and collapsing all rows except i-th row. Therefore performing column proportion tests on a 2x2 table will give you the same result as Pearson's chi-square test.

Use of case weights:

The case weights (or frequency weights) are supposed to be integers representing number of replications of each case. In column proportions tests, we will only check if the column marginal c_j's are integers. If not, they will be rounded to the nearest integer.

Column Means Tests

Notation

The following notation is used for the computation of Pairwise Comparisons of Column Means Tests:

k	Number of categories in the sub-table.

k^*	Number of categories with case weights greater than or equal to 2.
μ_i	Population mean of the i-th category, i=1,...,k.
x_{ij}	j-th observation in i-th group.
w_{ij}	Case weight of the j-th observation in i-th group.
w_i	Sum of case weights in category i, i=1,...,k.
$\tilde{w}_i$	Rounded sum of case weights in category i, i=1,...,k.
$\overline{x}_i$	Mean of category i, i=1,...,k.
s_i	Standard devation of category i, i=1,...,k.
s_{ij}	Pooled standard deviation from i-th and j-th group.
s_w	Pooled standard deviation of all categories.
W	Total case weights. Sum of rounded w_i 's.
p_B	p-value adjusted by using Bonferroni method.
α	Significance level supplied by the user.

Conditions and Assumptions

- Tests will not be performed for Comperimeter tables.
- Tests are performed on each innermost sub-tables for each layer.
- The row variable must be a scale variable, possibly nested under or over some categorical variables. The column variable must be categorical or a multiple response set.
- If weighting is on, cell statistics must include weighted means; a weighted analysis will be performed using the weighted statistics. If weighting is off, cell statistics must include means, an unweighted analysis will be performed.
- Tests are constructed by using all visible, non-empty categories excluding totals and sub-totals. Hiding of categories and showing of user-missing categories are respected.

- Total case weights in each category must be at least two. Categories not satisfying this assumption are not used. If number of categories satisfying this condition is less than two, no comparisons will be made.
- Variances of all categories are assumed to be equal.
- User and system missing values of scale variables are excluded.

Statistics

All Pairwise Comparisons

Hypotheses:

$H_{0ij}: \mu_i = \mu_j$, **vs.** $H_{1ij}: \mu_i \neq \mu_j$**, for all** $i > j$**.**

Total number of hypotheses: $\frac{k^*(k^*-1)}{2}$, (where $k^* = \sum_{i=1}^{k} I(w_i \geq 2)$).

Aggregated statistics:

The statistics in pairwise comparisons are computed from aggregated category means ($\bar{x}_i$), sample variances (s_i^2) and sample sizes (w_i), i=1,...,k. Various quantities used in the comparisons are shown below.

Total case weight (sample size): $W = \sum_{i=1}^{k} \text{round}(w_i) I(w_i \geq 2)$

Mean of i-th category: $\bar{x}_i = \frac{\sum_{j=1}^{n_i} w_{ij} x_{ij}}{w_i}$

Sample variance of i-th category: $s_i^2 = \dfrac{\sum_{j=1}^{n_i} w_{ij}(x_{ij} - \bar{x}_i)^2}{w_i - 1}$

Statisitics for (i,j)th comparisons:

Assuming $w_i \geq 2$ and $w_j \geq 2$,

Variance pooled from the two compared categories:

$$s_{ij}^2 = \frac{(\tilde{w}_i - 1)s_i^2 + (\tilde{w}_j - 1)s_j^2}{\tilde{w}_i + \tilde{w}_j - 2}$$

T-statistic, $t_{ij} = \dfrac{(\bar{x}_i - \bar{x}_j)}{s_{ij}\sqrt{\left(\dfrac{1}{\tilde{w}_i} + \dfrac{1}{\tilde{w}_j}\right)}}$,

P-value $p = 2[1 - F(|\, t_{ij} \,|; \tilde{w}_i + \tilde{w}_j - 2)]$,

where $F(t;n)$ is the distribution function of t-distribution with n degrees of freedom.

When multiple response set determines categories there may exist cases that belong to both i-th and j-th category. Let $\tilde{w}_{ij}$ be the rounded sum of weights for such cases.

T-statistic for comparing levels of a multiple response set

$$\mathrm{t}_{\mathrm{ij}} = \frac{(\bar{x}_i - \bar{x}_j)}{s_{ij}\sqrt{\left(\frac{1}{\tilde{w}_i} + \frac{1}{\tilde{w}_j} - \frac{2\tilde{w}_{ij}}{\tilde{w}_i \tilde{w}_j}\right)}},$$

P-value $p = 2[1 - F(| t_{ij} |; \tilde{w}_i + \tilde{w}_j - \tilde{w}_{ij} - 2)]$,

A comparison is significant if $\mathrm{p} < \alpha$ (or $\mathrm{p}_{\mathrm{B}} < \alpha$, if Bonferroni adjustment is used).

Statisitics for (i,j)th comparisons with variance pooled from all categories

Assume $\mathrm{w}_{\mathrm{i}} \geq 2$ and $\mathrm{w}_{\mathrm{j}} \geq 2$.

Within groups variance pooled from all the categories:

$$s_w^2 = \frac{\sum_{i=1}^{k} I(w_i \geq 2)(\tilde{w}_i - 1)s_i^2}{W - k^*}$$

T-statistic for levels of a categorical variable:

$$\mathrm{t}_{\mathrm{ij}} = \frac{(\bar{x}_i - \bar{x}_j)}{s_w\sqrt{\left(\frac{1}{\tilde{w}_i} + \frac{1}{\tilde{w}_j}\right)}}$$

P-value $\mathrm{p} = 2[1 - \mathrm{F}(| \mathrm{t}_{\mathrm{ij}} |; \mathrm{W} - \mathrm{k}^*)]$.

A comparison is significant if $\mathrm{p} < \alpha$ (or $\mathrm{p}_{\mathrm{B}} < \alpha$, if Bonferroni adjustment is used).

This test is available for categories defined by categorical variable only.

Bonferroni adjustment

If the Bonferroni adjustment for multiple comparisons is requested, the p-value p will be adjusted by

$$p_B = \min(\frac{pk^*(k^* - 1)}{2}, 1)$$

Possible computational problems:

From the formulas, we can see that comparison can be made as long as either s_{ij}^2 or s_w^2 is nonzero. If variances for both compared categories are zero, the first test cannot be conducted. If variances for all categories with cell count greater than or equal to two are zero, s_w^2 becomes zero and the second test conducted be conducted either.

Use of case weights:

The case weights (or frequency weights) are supposed to be integers representing number of replications of each case. If sum of case weights in each group (W_i,i=1,...,k) are not integers, they will be rounded to the nearest integers before calculations. Consequently, the total weight W will become the sum of rounded W_i's.