DigitalCommons
@WayneState

Journal of Modern Applied Statistical Methods

Volume 13 | Issue 1 Article 7

5-1-2014

An Exploratory Graphical Method for Identifying Associations in r x c Contingency Tables

Martin L. Lesser
Feinstein Institute for Medical Research, mlesser@nshs.edu

Meredith B. Akerman
Feinstein Institute for Medical Research, makerman@nshs.edu

Recommended Citation

Lesser, Martin L. and Akerman, Meredith B. (2014) "An Exploratory Graphical Method for Identifying Associations in r x c Contingency Tables," *Journal of Modern Applied Statistical Methods*: Vol. 13 : Iss. 1 , Article 7.
DOI: 10.22237/jmasm/1398917160
Available at: http://digitalcommons.wayne.edu/jmasm/vol13/iss1/7

Journal of Modern Applied Statistical Methods
May 2014, Vol. 13, No. 1, 91-109.

ISSN 1538 − 9472

An Exploratory Graphical Method for Identifying Associations in r x c Contingency Tables

Martin L. Lesser
Feinstein Institute for Medical Research
Manhasset, NY

Meredith B. Akerman
Feinstein Institute for Medical Research
Manhasset, NY

On finding a significant association between rows and columns of an r x c contingency table, the next step is to study the nature of the association in more detail. The use of a scree plot to visualize the largest contributions to X^2 among all cells in the table in order to determine the nature of the association in more detail is proposed.

Keywords: contingency table; graphical method; exploratory analysis; scree plot; contribution to chi-square

Introduction

A graphical method is proposed for exploring associations between rows and columns in an r x c contingency table. Typically, the Pearson chi-square test (or alternatively, the Fisher exact test) is used to test for independence of two categorical variables arranged in an r x c contingency table. (When one or both categories are ordinal, other procedures more suited to test for ordinal associations are available but the method being proposed here can be applied to both ordinal and non-ordinal data.)

On finding a significant association between rows and columns of an r x c table, the next step is to study the nature of the association (i.e., lack of independence) in more detail. One approach is to partition the r x c table and to use principles of chi-square partitioning to compare various groupings of rows and columns in order to make sense of the association (Agresti, 1990). Another method is to "collapse" the r x c table into some meaningful 2 x 2 table, the results for which are much easier to interpret (Feinstein, 2002). The advantage of

Dr. Lesser is Director, Investigator, and Professor of Biostatistics in the Biostatistics Unit. Email him at: MLesser@nshs.edu. Meredith Akerman is a Statistician in the Biostatistics Unit. Email her at: MAkerman@nshs.edu.

the first approach is that it is truly inferential, but the choice of how to partition the table may be impractical for very large r x c tables. The second method, while appealing due to its simplicity, may result in combining categories that have no appropriate justification or interpretation with respect to the subject matter being studied.

Consider the situation where the data analyst is interested more in exploration of the association rather than formal inference, in which case an exploratory graphical approach might be appropriate. There is the method known as Correspondence Analysis (CA) with applications in areas of social science, psychology, market research, and, to some extent, biomedical research (Greenacre, 1984; Greenacre, 1992). This graphical approach is based on linear algebraic techniques, which project the rows and columns of a data matrix in points onto a graph in Euclidean space, from which a better understanding of the data may be derived.

A simpler, yet intuitive method is proposed: exploratory graphical approach based on a method suggested by Snedecor and Cochran (1989), in which the data analyst identifies the cell entries providing the largest percentage contributions to X^2 because those will suggest departure from the null hypothesis of independence, and will be row-column combinations of interest. Some drawbacks of this approach are that searching an r x c table for the "largest" contributions to X^2 can be tedious (especially for large tables), inefficient, and prone to error (i.e., failing to identify all the cells that are "large" contributors). Given these potential problems, a graphical approach to summarizing these contributions would be helpful, especially when there are many cells to analyze.

The graphical approach used herein is to use an adaptation of the scree plot to visualize the largest contributions to X^2 among all of the cells of the r x c table. (The scree plot is commonly used in principal components analysis to help choose the most important principal components [Khattree and Naik, 2000]).

As an example, Table 1 (hypothetical data for illustrative purposes) is a 6 x 5 cross-tabulation of a patient's primary hospital admitting diagnosis according to the patient's race. There is a highly significant association between diagnosis and race ($X^2 = 326.4$, $p < 0.0001$). The common interpretation of this significance is that diagnosis is not independent of race or, alternatively, that there are at least two races for which the distributions of diagnosis differ. Which two (or more) columns differ from one another?

Table 1. Cross-tabulation of a patient's race according to patient's primary hospital admitting diagnosis

	Primary hospital admitting diagnosis						
	DM	**Chest pain**	**CVA**	**Fever**	**GI distress**	**Other**	**Total**
White	39 23.49	18 10.84	51 30.72	22 13.25	16 9.64	20 12.05	166
Black	11 6.25	15 8.52	8 4.55	2 1.14	92 52.27	48 27.27	176
Hispanic	90 40.54	56 25.23	19 8.56	15 6.76	13 5.86	29 13.06	222
Asian	13 26.53	0 0	14 28.57	7 14.29	15 30.61	0 0	49
Other	44 46.32	18 18.95	10 10.53	11 11.58	9 9.47	3 3.16	95
Total	197	107	102	57	145	100	708

Note. The top entry in each cell is the frequency count; the lower entry is the "row percent," which is the percentage based on the row total.

To answer that question, two methods are commonly used. The first is simply to inspect the many so-called "column proportions" and informally, based on subjective visualization, make a judgment as to which columns differ. The second is to more formally perform all 10 pairwise comparisons of the columns using a X^2 test with 5 degrees of freedom and to declare two columns as different if the associated p-value is less than some critical value that is appropriately adjusted for multiple comparisons. (In general there would be *c!/(2!(c-2)!)* each with *r-1* degrees of freedom.)

The first method is deficient because it is highly subjective and requires simultaneous visual processing of all of the column percentages. The second method has the advantage of being truly inferential, but, in finding two columns that differ, it fails to identify the row locations of those differences.

The graphical method proposed is computationally objective and reproducible and can be easily programmed in most statistical software packages, including SAS® for which a publically available macro has been written.

Methodology

Suppose data are arranged in an r x c contingency table. The individual entries in the r x c table represent the frequency, or, number, of observations of a given row-column combination (e.g. race and diagnosis as in Table 1.)

Using standard statistical notation, let O_{ij} represent the observed entry in row i, column j, $O_i.$ the total of all entries in row i, $O._j$ the total of all entries in column j, and E_{ij} the expected entry in row i, column j. Letting n denote the sum total of all frequencies entered in the table, the expected frequency of row i, column j, E_{ij}, is calculated as the product of the total frequency in row i multiplied by the total frequency in column j, divided by n (i.e., $E_{ij} = \left(O_i.\, x\ O._j\right)\ /\ n$).

Using this notation, the standard Pearson X^2 statistic is calculated as

$$X^2 = \Sigma_i \Sigma_j \left[\left(O_{ij} - E_{ij}\right)^2 / \ E_{ij}\right],$$

where the summations correspond to $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$. Snedecor and Cochran (1989) denote the contribution of the ij^{th} entry to the X^2 statistic as

$$X^2_{ij} = \left(O_{ij} - E_{ij}\right)^2 / \ E_{ij}$$

Compute all values of X^2_{ij} for i=1, 2,…, r and j=1,2,…,c. Then compute P_{ij} = 100* X^2_{ij} / X^2 = percentage of overall X^2 contributed by the ij^{th} entry. Snedecor and Cochran (1989) propose that the entries providing the largest percentage contributions to X^2 are those that will suggest departure from the null hypothesis of independence. Note that "contribution to X^2" is sometimes referred to as the square of the "standardized residuals" (Agresti, 1990).

The general idea of the proposed graphical method is to compute each table entry's P_{ij}, order the P_{ij}s from largest to smallest, and to find the first P_{ij} for which the remaining ordered P_{ij}s remain relatively constant. This ordering can be visually displayed in a graph, known as a "scree plot". The algorithm for constructing the scree plot is given in the following steps:

Step 1

Order the values of P_{ij} from largest to smallest and denote the ordered values (i.e. "order statistics") as $P_{(1)} \geq P_{(2)}, \geq \ldots, \geq P_{(rc)}$.

Step 2

Plot $P_{(i)}$ against i to form a scree plot, analogous to what is done with eigenvalues in principal components analysis (PCA) (Khattree and Naik, 2000).

Step 3

Find the cells in the r x c table that significantly contribute to the departures from independence. This can be done using any of the following three criteria.

Cumulative Percent Method Find the left-most point on the horizontal axis that corresponds to a cumulative sum of percent contributions to chi-square that totals as close to, but does not exceed some pre-specified percentage, π. For example, π might be set to 50%. It should be noted that π is often chosen arbitrarily with no formal justification of its utility. Using $\pi = 50\%$ is “middle of the road”. Increasing π would result in a more “liberal” rule, allowing more cells to be implicated in the departure from independence, possibly increasing the false positive rate with respect to identifying the number of such cells. Decreasing π would restrict the number of cells, possibly increasing the false negative rate. (Note that in PCA, π, which would be the cumulative variance explained, is often set to 90% [Khattree and Naik, 2000])

Subjective Elbow Method Find the “bend of the elbow” or “turning point” of the scree plot to determine which cells in the r x c table contribute substantially to the X^2 statistic. Typically, the bend in the elbow would be defined as the point on the plot for which all points to the left of it will have a much steeper downward slope than those to the right. The idea behind this choice of a bending point is that the number of cells to be selected is such that the differences between consecutive contributions to chi-square are becoming increasing smaller (Khattree and Naik, 2000). This subjective method is based only on visual inspection of the scree plot. This approach may be useful when there is a fairly clear elbow. The primary shortcoming is that this method is subjective and may not be reproducible between data analysts.

Objective Elbow Method Because the determination of the bend in the elbow using the Subjective Elbow Method is not necessarily reproducible, it is proposed to systematize the identification of the elbow by finding the ordered pair $(i,P_{(i)})$ which is closest to the origin (0,0). This can be done by computing the squared-Euclidean distances of each point on the scree plot,

$(i-0)^2 + (P_{(i)}-0)^2 = i^2 + P_{(i)}^2$ and finding the ordered pair, (i^*, P^*), corresponding to the minimum value of those distances (i.e. (i^*, P^*) is the point closest to the origin). All cells that are represented on the plot with $i \leq i^*$ would then be implicated in the departure from independence. In the context of a scree plot, which is a plot of a non-increasing concave function, the "ideal" elbow would be two straight line segments connected at a "pivot" point forming an angle of 90º to less than 180º between the segments. For such a function, the bend of the elbow would correspond to the point with minimum distance to the origin. An example of an ideal elbow would be a perfect "L" shape curve with its vertical and horizontal components parallel to the vertical and horizontal axes of the scree plot, respectively.

It should be emphasized that while the proposed method relies on the use of the chi-square statistic, as an exploratory tool, it can be used even when the r x c table does not meet the criteria for the use of the Pearson chi-square test and a Fisher's exact test would be more appropriate.

For this manuscript, the authors used the PROC FREQ procedure in SAS Version 9.3 (SAS Institute, Cary, NC).

Results and Examples

The proposed method is illustrated using data from the Asia-Pacific Quality of Life Study (APQOL) in Lung Cancer. (The data are provided courtesy of Drs. Richard Gralla and Patricia Hollen [Gralla, 2013; Thongprassert, 2013]). This data consists of, among other variables, country of diagnosis (China, Korea, Thailand, Taiwan), Karnofsky Performance Status at diagnosis (KPS=50, 60, 70, 80, 90, 100), lung cancer T stage (T0, T1, T2, T3, T4, and TX), node status (N0, N1, N2, N3, NX), and metastasis (M0, M1, MX). [The so-called "TNM staging system" for cancer classifies cancers according to tumor size (T), lymph node involvement (N), and presence or absence of metastatic disease (M). The KPS is a measure of a patient's general well-being and activities of daily life.] Analyses investigated whether there was any association between any of these variables and country of diagnosis. Standard Pearson chi-square analysis for r x c contingency tables was carried out. Four examples were chosen to illustrate variation in the way that the location of the elbow might be visually and subjectively judged.

Example 1

Table 2a is the contingency table of Country vs. KPS and displays, respectively, each cell's frequency, deviation from expected ($O_{ij}–E_{ij}$), cell chi-square ($X^2_{ij} =[O_{ij}–E_{ij}]^2/E_{ij}$), and row percent (frequency relative to the row total). As shown in the footnote to Table 2a, $X^2 = 97.72$, $df = 15$, $p < 0.0001$ and the Fisher exact test yields $p < 0.0001$.

Table 2a. Country vs. KPS, including frequency, deviation, cell chi-square and row percent.

	50	**60**	**70**	**80**	**90**	**100**	**Total**
China	0	0	8.0000	24.0000	52.0000	15.0000	99
	-0.1920	-0.1920	0.5174	-2.2850	0.7733	1.3779	
	0.1919	0.1919	0.0358	0.1986	0.0117	0.1394	
	0	0	8.0800	24.2400	52.5300	15.1500	
Korea	0	0	8.0000	51.0000	111.0000	8.0000	178
	-0.3450	-0.3450	-5.4530	3.7403	18.8950	-16.4900	
	0.3450	0.3450	2.2106	0.2960	3.8764	11.1050	
	0	0	4.4900	28.6500	62.3600	4.4900	
Thailand	1.0000	0	19.0000	48.0000	41.0000	9.0000	118
	0.7713	-0.2290	10.0810	16.6710	-20.0600	-7.2360	
	2.6016	0.2287	11.3960	8.8705	6.5893	3.2252	
	0.8500	0	16.1000	40.6800	34.7500	7.6300	
Taiwan	0	1.0000	4.0000	14.0000	63.0000	39.0000	121
	-0.2340	0.7655	-5.1450	-18.1300	0.3895	22.3510	
	0.2345	2.4990	2.8949	10.2270	0.0024	30.0050	
	0	0.8300	3.3100	11.5700	52.0700	32.2300	
Total	1	1	39	137	267	71	516

Note. X^2=97.72, *df*=15, *p*<0.0001 and Fisher exact test *p*<0.0001. The top entry in each cell is the frequency count; the second entry is the cell deviation (*O*−*E*); the third entry is the cell contribution to chi-square [$(O−E)^2$ / *E*]; the last entry is the "row percent," which is the cell percentage based on the row total.

Table 2b contains the same information as Table 2a (in a list format), where the percent contribution to chi-square of each cell has been computed ($P_{ij}= 100*X^2_{ij} / X^2$), the table has been sorted by decreasing P_{ij}, and the cumulative percent contributions have been computed.

Table 2b. Country vs. KPS, including frequency, deviation, cell chi-square and row percent, in list format, sorted by decreasing P_{ij}

Rank	Country	KPS	Cell Chi-Square	Deviation (*O-E*)	% Row Frequency	% contrib. to chi sq.	Cumulative % contribution
1	Taiwan	100	30.0048	22.3508	32.2314	30.7046	30.7046
2	Thailand	70	11.3958	10.0814	16.1017	11.6616	42.3661
3	Korea	100	11.1053	-16.4922	4.4944	11.3643	53.7305
4	Taiwan	80	10.2270	-18.1260	11.5702	10.4655	64.1959
5	Thailand	80	8.8705	16.6705	40.6780	9.0773	73.2733
6	Thailand	90	6.5893	-20.0581	34.7458	6.7429	80.0162
7	Korea	90	3.8764	18.8953	62.3596	3.9668	83.9830
8	Thailand	100	3.2252	-7.2364	7.6271	3.3004	87.2834
9	Taiwan	70	2.8949	-5.1453	3.3058	2.9624	90.2458
10	Thailand	50	2.6016	0.7713	0.8475	2.6622	92.9081
11	Taiwan	60	2.4990	0.7655	0.8264	2.5572	95.4653
12	Korea	70	2.2106	-5.4535	4.4944	2.2622	97.7275
13	Korea	50	0.3450	-0.3450	0	0.3530	98.0805
14	Korea	60	0.3450	-0.3450	0	0.3530	98.4335
15	Korea	80	0.2960	3.7403	28.6517	0.3029	98.7364
16	Taiwan	50	0.2345	-0.2345	0	0.2400	98.9764
17	Thailand	60	0.2287	-0.2287	0	0.2340	99.2104
18	China	80	0.1986	-2.2849	24.2424	0.2033	99.4136
19	China	50	0.1919	-0.1919	0	0.1963	99.6100
20	China	60	0.1919	-0.1919	0	0.1963	99.8063
21	China	100	0.1394	1.3779	15.1515	0.1426	99.9489
22	China	70	0.0358	0.5174	8.0808	0.0366	99.9856
23	China	90	0.0117	0.7733	52.5253	0.0119	99.9975
24	Taiwan	90	0.0024	0.3895	52.0661	0.0025	100.0000

Figure 1 displays the corresponding scree plot where each $P_{(i)}$ is plotted on the vertical axis against its rank order and the plot is further annotated with the respective cumulative cell percentages. Visual inspection of the scree plot (Figure 1) does not reveal a clear cut turning point. Depending on the observer's perspective, rank 2, 7, or 13 could be considered the turning point. Based on the more objective Euclidean distance method, the turning point corresponds to rank 7. (The calculation of each cell's Euclidean distance was deliberately omitted from each table in order to let the reader better appreciate the shortcomings of the visual process of finding the elbow, without being biased by knowing the corresponding distances. For the record, the squared distances for the first 10 ordered cells were 943.8, 140.0, 138.1, 125.5, 107.4, 81.5, 64.7, 74.9, 89.8, and 107.1, with the minimum (64.7) occurring at rank 7.)

Referring back to Table 2b one can examine the ranks of the cells corresponding to ranks 1 through 7 to identify those cells in the table that deviate the most from their expected values, as well as the direction of their deviation under the null hypothesis of independence, in order to better understand the nature of the association. Taiwan appears to have an overrepresentation of patients with KPS 100, while Korea's frequency is less than expected. Patients with KPS 80 tend to be underrepresented in Taiwan, but overrepresented in Thailand. Patients with KPS 90 tend to be underrepresented in Thailand and overrepresented in Korea. Finally, patients with KPS 70 tend to be overrepresented in Thailand.

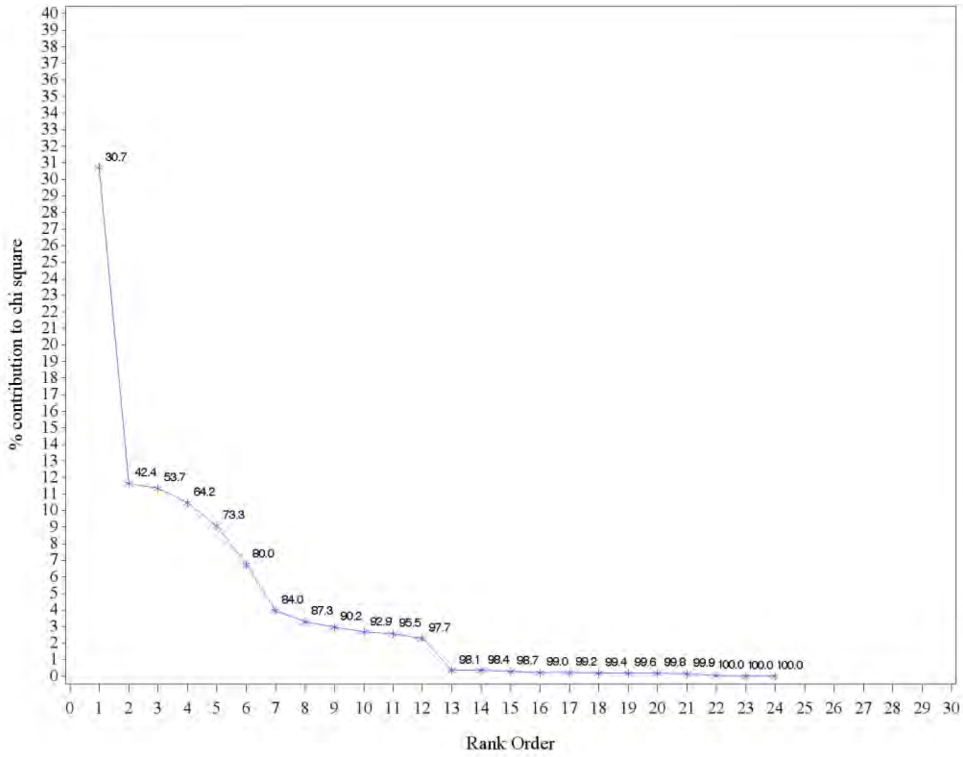

Figure 1. Scree plot of Country vs. KPS data in Table 2. $P_{(i)}$ is plotted on the vertical axis against its rank order; the plot is annotated with the respective cumulative cell percentages (rounded up).

Example 2

Tables 3a and 3b show the relevant calculations for the association between Country and T stage. In this example, the association is not significant (X^2=22.29, df=15, p=0.10, and the Fisher exact test yields p=0.085.) Although not significant and the general shape of the curve is similar to that in Figure 1, consider this example to show that it may still be of interest to apply the proposed method to discover patterns in the data.

Table 3a. Country vs. Tumor stage, including frequency, deviation, cell chi-square and row percent.

	T0	T1	T2	T3	T4	TX	Total
China	0	3	26	19	43	9	100
	-1.758	-1.297	-0.563	-1.508	1.3984	3.7266	
	1.7578	0.3914	0.0119	0.1109	0.047	2.6334	
	0	3	26	19	43	9	
Korea	4	9	47	31	73	8	172
	0.9766	1.6094	1.3125	-4.273	1.4453	-1.07	
	0.3154	0.3505	0.0377	0.5177	0.0292	0.1263	
	2.33	5.23	27.33	18.02	42.44	4.65	
Thailand	5	6	28	22	48	9	118
	2.9258	0.9297	-3.344	-2.199	-1.09	2.7773	
	4.1269	0.1705	0.3567	0.1999	0.0242	1.2396	
	4.24	5.08	23.73	18.64	40.68	7.63	
Taiwan	0	4	35	33	49	1	122
	-2.145	-1.242	2.5938	7.9805	-1.754	-5.434	
	2.1445	0.2943	0.2076	2.5455	0.0606	4.589	
	0	3.28	28.69	27.05	40.16	0.82	
Total	9	22	136	105	213	27	512

Note. X^2=22.29, *df*=15, *p*=0.10; Fisher exact test *p*=0.085. The top entry in each cell is the frequency count; the second entry is the cell deviation (*O−E*); the third entry is the cell contribution to chi-square [$(O-E)^2 / E$]; the last entry is the “row percent,” which is the cell percentage based on the row total.

Table 3b. Country vs. Tumor stage, including frequency, deviation, cell chi-square and row percent, in list format, sorted by decreasing P_{ij}

Rank	Country	Tumor	Cell Chi-Square	Deviation (*O-E*)	% of Row Frequency	% contrib. to chi sq.	Cumulative % contribution
1	Taiwan	TX	4.58903	-5.43359	0.8197	20.5890	20.589
2	Thailand	T0	4.12695	2.92578	4.2373	18.5159	39.105
3	China	TX	2.63344	3.72656	9.0000	11.8151	50.920
4	Taiwan	T3	2.54553	7.98047	27.0492	11.4207	62.341
5	Taiwan	T0	2.14453	-2.14453	0	9.6216	71.962
6	China	T0	1.75781	-1.75781	0	7.8866	79.849
7	Thailand	TX	1.23961	2.77734	7.6271	5.5616	85.411
8	Korea	T3	0.51773	-4.27344	18.0233	2.3229	87.733
9	China	T1	0.39142	-1.29688	3.0000	1.7561	89.489
10	Thailand	T2	0.35671	-3.34375	23.7288	1.6004	91.090
11	Korea	T1	0.35046	1.60938	5.2326	1.5723	92.662
12	Korea	T0	0.31543	0.97656	2.3256	1.4152	94.077
13	Taiwan	T1	0.29435	-1.24219	3.2787	1.3206	95.398
14	Taiwan	T2	0.20760	2.59375	28.6885	0.9314	96.329
15	Thailand	T3	0.19986	-2.19922	18.6441	0.8967	97.226
16	Thailand	T1	0.17047	0.92969	5.0847	0.7648	97.991
17	Korea	TX	0.12630	-1.07031	4.6512	0.5666	98.558
18	China	T3	0.11086	-1.50781	19.0000	0.4974	99.055
19	Taiwan	T4	0.06061	-1.75391	40.1639	0.2719	99.327
20	China	T4	0.04701	1.39844	43.0000	0.2109	99.538
21	Korea	T2	0.03771	1.31250	27.3256	0.1692	99.707
22	Korea	T4	0.02919	1.44531	42.4419	0.1310	99.838
23	Thailand	T4	0.02420	-1.08984	40.6780	0.1086	99.947
24	China	T2	0.01191	-0.56250	26.0000	0.0534	100.000

In the scree plot for this example (Figure 2), the bend in the elbow is more obvious than in Figure 1 and appears to be at rank 8. This is confirmed using the Euclidean distance method.

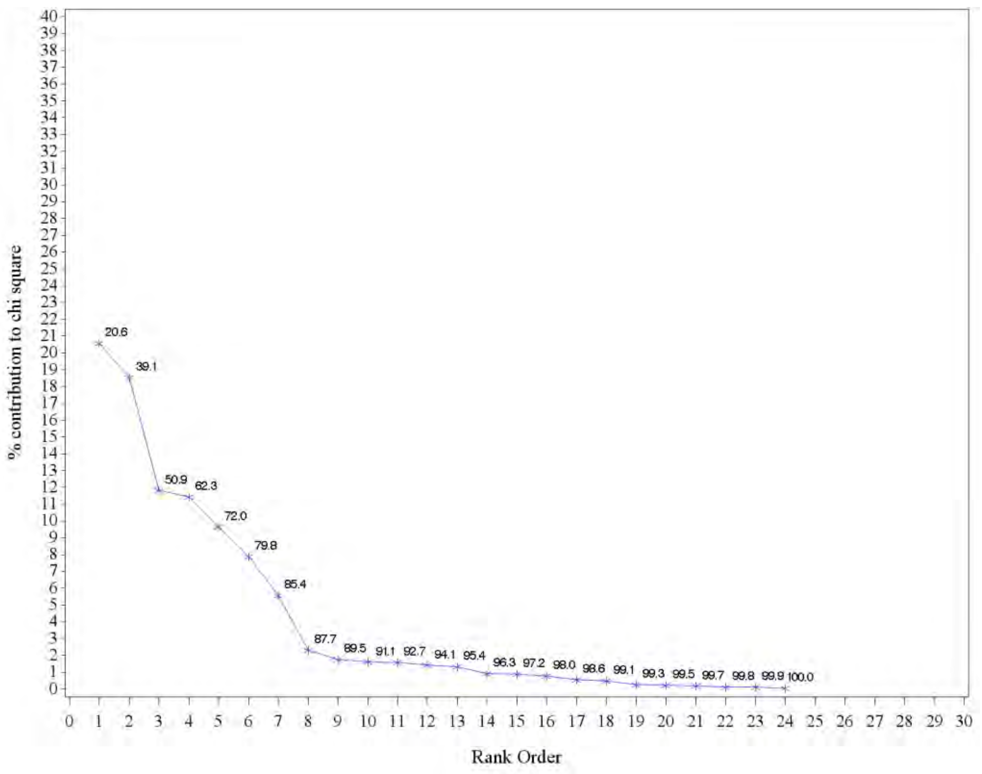

Figure 2. Scree plot of Country vs. Tumor stage data in Table 3.

Referring back to Table 3b, it appears that the departures are explained by the frequency distribution of unclassified (TX) and *in situ* (T0) tumors primarily among China, Thailand, and Taiwan. Furthermore, the direction of the deviation from each country can be seen in the column labeled Deviation. China and Thailand appear to have more TX tumors than expected, while Taiwan's frequency is decreased. T0 tumors tend to be underrepresented in China and Taiwan, but overrepresented in Thailand.

Even though the observed association between Country and T stage was not significant (Fisher's p=0.085), the observed pattern may still be of clinical interest.

Example 3

Tables 4a and 4b show the relevant calculations for the association between Country and N stage. In this example, the association is significant (X^2=33.96, df=12, p=0.0007.)

Table 4a. Country vs. Node stage, including frequency, deviation, cell chi-square and row percent.

	N0	N1	N2	N3	NX	Total
China	10.0000	8.0000	29.0000	43.0000	10.0000	100
	-3.0860	-0.5940	0.6797	4.5234	-1.5230	
	0.7277	0.0410	0.0163	0.5318	0.2014	
	10.0000	8.0000	29.0000	43.0000	10.0000	
Korea	23.0000	22.0000	44.0000	74.0000	9.0000	172
	0.4922	7.2188	-4.7110	7.8203	-10.8200	
	0.0108	3.5254	0.4556	0.9241	5.9070	
	13.3700	12.7900	25.5800	43.0200	5.2300	
Thailand	19.0000	6.0000	25.0000	43.0000	25.0000	118
	3.5586	-4.1410	-8.4180	-2.4020	11.4020	
	0.8201	1.6907	2.1205	0.1271	9.5615	
	16.1000	5.0800	21.1900	36.4400	21.1900	
Taiwan	15.0000	8.0000	47.0000	37.0000	15.0000	122
	-0.9650	-2.4840	12.4490	-9.9410	0.9414	
	0.0583	0.5887	4.4857	2.1054	0.0630	
	12.3000	6.5600	38.5200	30.3300	12.3000	
Total	67	44	145	197	59	512

Note. *X^2*=33.96, *df*=12, *p*=0.0007. The top entry in each cell is the frequency count; the second entry is the cell deviation (*O*−*E*); the third entry is the cell contribution to chi-square [(*O*−*E*)^2 / *E*]; the last entry is the "row percent," which is the cell percentage based on the row total.

Table 4b. Country vs. Node stage, including frequency, deviation, cell chi-square and row percent, in list format, sorted by decreasing P_{ij}

Rank	Country	Nodes	Cell Chi-Square	Deviation (*O-E*)	% of Row Frequency	% contrib. to chi sq.	Cumulative % contribution
1	Thailand	NX	9.56146	11.4023	21.1864	28.1532	28.1532
2	Korea	NX	5.90703	-10.8203	5.2326	17.3930	45.5462
3	Taiwan	N2	4.48566	12.4492	38.5246	13.2078	58.7540
4	Korea	N1	3.52544	7.2188	12.7907	10.3805	69.1345
5	Thailand	N2	2.12048	-8.4180	21.1864	6.2437	75.3781
6	Taiwan	N3	2.10542	-9.9414	30.3279	6.1993	81.5774
7	Thailand	N1	1.69070	-4.1406	5.0847	4.9782	86.5556
8	Korea	N3	0.92411	7.8203	43.0233	2.7210	89.2766
9	Thailand	N0	0.82011	3.5586	16.1017	2.4148	91.6914
10	China	N0	0.72773	-3.0859	10.0000	2.1428	93.8341
11	Taiwan	N1	0.58870	-2.4844	6.5574	1.7334	95.5675
12	China	N3	0.53179	4.5234	43.0000	1.5658	97.1334
13	Korea	N2	0.45560	-4.7109	25.5814	1.3415	98.4749
14	China	NX	0.20140	-1.5234	10.0000	0.5930	99.0679
15	Thailand	N3	0.12711	-2.4023	36.4407	0.3743	99.4422
16	Taiwan	NX	0.06304	0.9414	12.2951	0.1856	99.6278
17	Taiwan	N0	0.05831	-0.9648	12.2951	0.1717	99.7995
18	China	N1	0.04102	-0.5938	8.0000	0.1208	99.9203
19	China	N2	0.01631	0.6797	29.0000	0.0480	99.9683
20	Korea	N0	0.01076	0.4922	13.3721	0.0317	100.0000

Visual inspection of the scree plot (Figure 3) reveals a much smoother curve then those shown in Figures 1 and 2 and does not reveal a clear cut bending point. Using the Euclidean distance method, the turning point corresponds to rank 5. Referring back to Table 4b, it appears that the departures are explained by the frequency distribution of unclassified (NX) and N2 nodes primarily among Korea, Thailand, and Taiwan. Thailand appears to have an excess of NX nodes, while Korea's frequency is decreased. N2 nodes tend to be underrepresented in Thailand, but overrepresented in Taiwan.

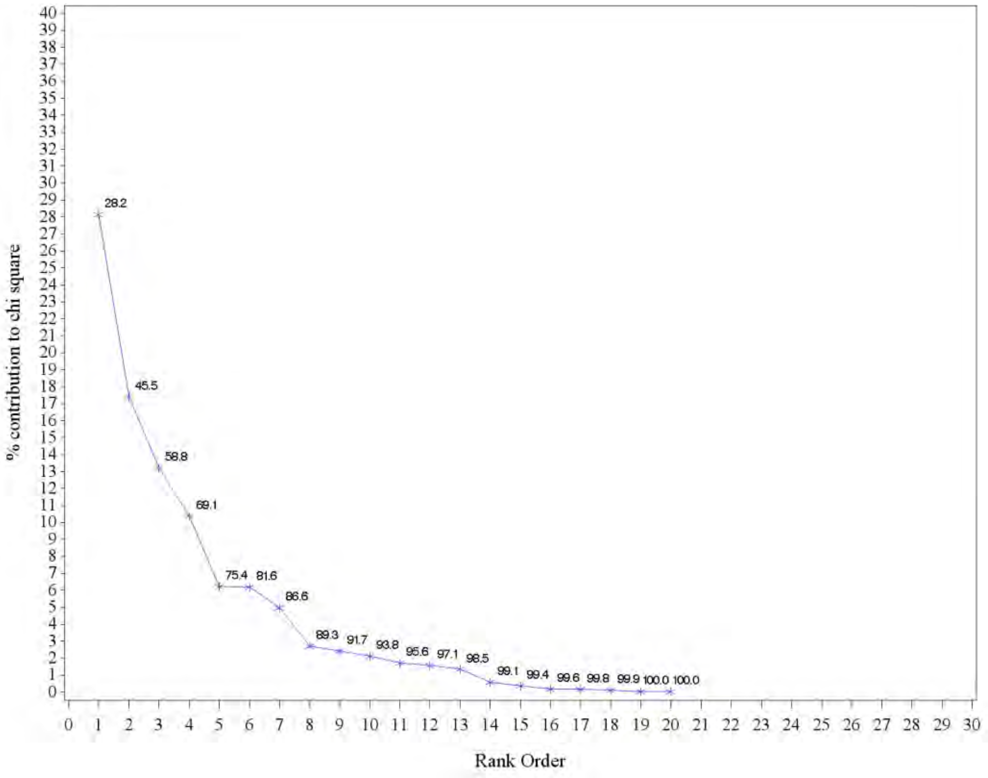

Figure 3. Scree plot of Country vs. Node stage data in Table 4.

Example 4

Tables 5a and 5b show the relevant calculations for the association between Country and M stage. In this example, the association is also significant (X^2=30.64, *df*=6, *p*<0.0001.)

Table 5a. Country vs. Metastasis stage, including frequency, deviation, cell chi-square and row percent.

	M0	M1	MX	Total
China	26.0000	71.0000	3.0000	100
	6.4688	-2.6330	-3.8360	
	2.1425	0.0941	2.1525	
	26.0000	71.0000	3.0000	
Korea	19.0000	147.0000	6.0000	172
	-14.5900	20.3520	-5.7580	
	6.3398	3.2704	2.8196	
	11.0500	85.4700	3.4900	
Thailand	31.0000	71.0000	16.0000	118
	7.9531	-15.8900	7.9336	
	2.7445	2.9048	7.8030	
	26.2700	60.1700	13.5600	
Taiwan	24.0000	88.0000	10.0000	122
	0.1719	-1.8320	1.6602	
	0.0012	0.0374	0.3305	
	19.6700	72.1300	8.2000	
Total	100	377	35	512

Note. X^2=30.64, *df*=6, *p*=0.0001. The top entry in each cell is the frequency count; the second entry is the cell deviation (*O*−*E*); the third entry is the cell contribution to chi-square [(*O*−*E*)^2 / *E*]; the last entry is the "row percent," which is the cell percentage based on the row total.

Table 5b. Country vs. Metastasis stage, including frequency, deviation, cell chi-square and row percent, in list format, sorted by decreasing P_{ij}

Rank	Country	Metastasis	Cell Chi-Square	Deviation (*O-E*)	% of Row Frequency	% contrib. to chi sq.	Cumulative % contribution
1	Thailand	MX	7.80297	7.9336	13.5593	25.4664	25.466
2	Korea	M0	6.33980	-14.5938	11.0465	20.6911	46.158
3	Korea	M1	3.27036	20.3516	85.4651	10.6734	56.831
4	Thailand	M1	2.90479	-15.8867	60.1695	9.4803	66.311
5	Korea	MX	2.81961	-5.7578	3.4884	9.2023	75.514
6	Thailand	M0	2.74450	7.9531	26.2712	8.9572	84.471
7	China	MX	2.15251	-3.8359	3.0000	7.0251	91.496
8	China	M0	2.14245	6.4688	26.0000	6.9923	98.488
9	Taiwan	MX	0.33048	1.6602	8.1967	1.0786	99.567
10	China	M1	0.09414	-2.6328	71.0000	0.3072	99.874
11	Taiwan	M1	0.03736	-1.8320	72.1311	0.1219	99.996
12	Taiwan	M0	0.00124	0.1719	19.6721	0.0040	100.000

The scree plot (Figure 4) does not reveal a clear cut bending point. Either rank 3 or rank 9 could be judged as the turning points. However, using the Euclidean distance method, the turning point corresponds to rank 9. Thailand and Taiwan appear to have an excess of unknown metastases (MX), while Korea and China's frequencies are decreased. Patients with no distant metastases (M0) tend to be underrepresented in Korea, but overrepresented in Thailand and China. Patients with metastases to distant organs (M1) tend to be overrepresented in Korea but underrepresented in Thailand.

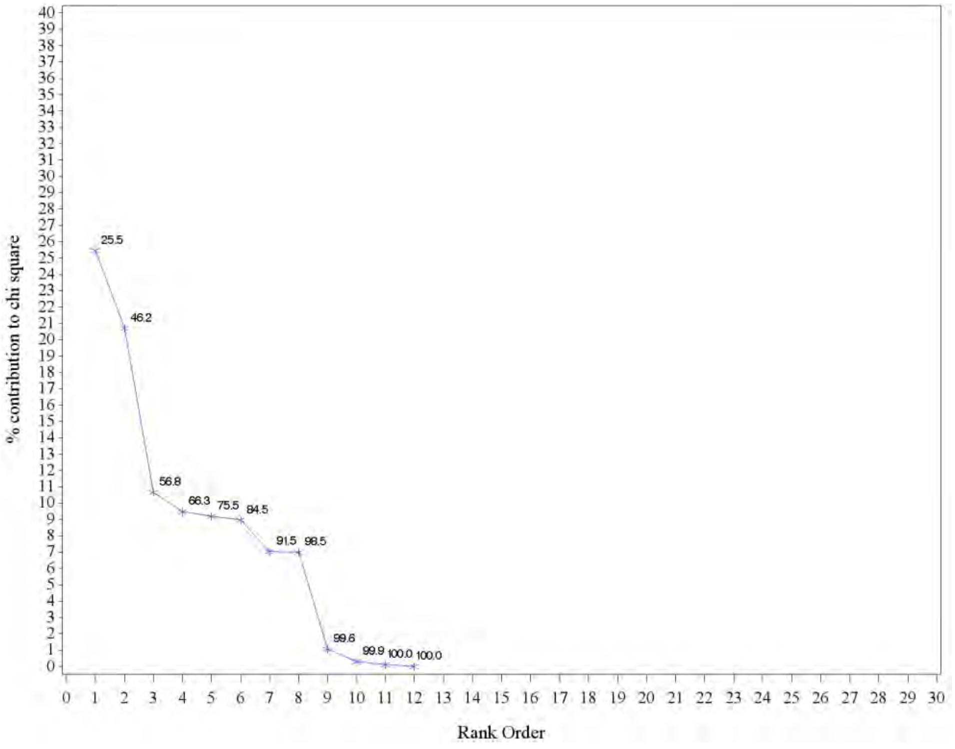

Figure 4. Scree plot of Country vs. Metastasis stage data in Table 5.

Conclusion

In statistical problems involving the cross-classification of frequency counts, it is common to test for an association between one variable and another using the well-known Pearson chi-square test (or, alternatively, the Fisher exact test, particularly for sparse tables). Upon finding a significant association, it is of interest to identify the cells in the table that are "responsible" for the lack of independence. As the dimension of the table gets larger (i.e., the number of rows and/or columns grows larger), it becomes more difficult to identify these row-column combinations.

An exploratory, graphical method of discovering those cells that account for the observed association was proposed. This method is computationally objective and completely reproducible.

The method is based on two frequently used techniques: assessment of contribution to chi-square in contingency tables and construction of scree plots as in principal components analysis. All of the computations required for applying this method are available in virtually all commonly used statistical software packages.

Several examples of r x c tables were provided that exemplify the use of this method both when the observed associations are statistically significant and when they are not. The examples illustrate how the use of a cutoff point for the cumulative percent contribution to chi-square ("Cumulative Percent Method" as described above) is purely arbitrary. Of course, most statistical procedures include some elements of arbitrariness – most notably the use of "$p < 0.05$" or "95%" for constructing confidence intervals. The examples further show that visual appraisal of the scree plot ("Subjective Elbow Method") can be highly subjective and might, therefore, vary from one observer to another.

In order to address these shortcomings, it has been shown how the proposed Objective Elbow Method for exploring contingency tables parallels the currently accepted approach to identifying important principal components in PCA with the addition of an objective and reproducible calculation (Euclidean distance) that identifies the bend in the scree plot that constitutes the "elbow".

As discussed in the introduction, Correspondence Analysis has been used in the current r x c setting. While CA is a useful and powerful method, it requires somewhat specialized, albeit, readily available software (e.g., PROC CORRESP in SAS, CORRESPONDENCE module in SPSS). The proposed method, while not providing the level of detail contained in CA, is much simpler to execute, intuitively appealing to the non-statistician, and requires no more than the ability to perform standard contingency table analysis.

The use of graphical methodology as a complement to inferential analysis is widespread in statistical practice – even in the absence of statistical significance. Common examples include the already cited scree plots in PCA, scatterplots, side-by-side boxplots, receiver operating characteristic (ROC) curves, survival and hazard function curves, ANOVA interaction plots, heat maps in genetics problems, to name only a few.

This method could be readily adopted by investigators in many fields of research involving r x c contingency tables because the ability to perform these calculations is readily available in commonly used statistical software packages.

For this manuscript, the PROC FREQ procedure in SAS Version 9.3 (SAS Institute, Cary, NC) was used. The following list shows the availability of the components of the proposed calculation in various software packages.

- SAS (SAS Institute, Cary, NC): PROC FREQ, “cellchi2” TABLE option.
- JMP (SAS Institute, Cary, NC): Contingency Table, choose the drop down labeled “Cell Chi Square”.
- Minitab (Minitab, Inc., State College, PA), Stat: Tables: Cross Tabulation and Chi-Square, check the box labeled “Each cell’s contribution to the Chi-Square statistic”
- Stata (StataCorp LP, College Station, TX): “tabulate” with the cchi2 option
- R (R Foundation for Statistical Computing, r-project.org): chisq.detail
- Excel (Microsoft Corp., Redmond, WA): programmed and calculated by user
- SPSS (IBM Inc., Armonk, NY): Crosstabs, Cells subcommand, check the box labeled “Standardized” under Residuals; contribution to cell chi-square must be programmed and calculated from these Residuals by the user

Finally, it is not proposed that the Objective Elbow Method be rigidly obeyed. This method simply provides a reproducible guidance as to which cells may be responsible for the observed association. Upon finding i^*, corresponding to the point closest to the origin, the data analyst might also want to consider points to the right of i^* but very close to it, as other potential cells of interest. Based on study results, the proposed method is believed to be potentially useful to data analysts using large r x c tables.

References

Agresti, A. (1990). *Categorical data analysis*. New York: John Wiley and Sons.

Feinstein, A. R. (2002). *Principles of medical statistics*. Boca Raton, FL: Chapman and Hall/CRC.

Gralla, R. J., Hollen, P., Thongprassert, S., Kim, H. K., Hsia, T. C., Yuankai, S., Kohn, N., & Lesser, M. (2013). *Accurate prediction of survival outcomes in nsclc using a new pro index from the lcss (lung cancer symptom scale): results of a 622 patient prospective trial*. Presented at ASCO Annual Meeting, Chicago, IL, May 31 – June 4, 2013.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Greenacre, M. J. (1992). Correspondence analysis in medical research, *Stat Methods Med Res*, *1*(1): 97-117.

Huber, W., Li, X., & Gentleman, R. (2005). Visualizing Data. In R. Gentleman, V. Carey, et al. (Eds.), *Bioinformatics and computational biology solutions using r and bioconductor* (pp. 161-179). New York: Springer Verlag.

Khattree, R., & Naik, D. N. (2000). *Multivariate data reduction and discrimination with SAS software*. Cary, NC: SAS Institute Inc.

Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., & Zhao, Y. (2004). Class Discovery. In *Design and analysis of DNA microarray investigations* (pp. 121-155). New York: Springer Verlag.

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical methods* (Eighth Ed.). Ames, IA: Iowa State University Press.

Thongprassert, S., Gralla, R. J., Hollen, P., Kim, H. K., Hsia, T. C., Yuankai, S., Kohn, N., & Lesser, M. (2013). *Overcoming barriers in incorporating evaluation of quality of life (QL) and symptoms by using the EPRO version of the LCSS (ELCSS-QL) in a large-scale multinational NSCLC trial (AP-QL trial).* Presented at ASCO Annual Meeting, Chicago, IL, May 31 – June 4, 2013.