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

Martin L. Lesser*, PhD, Meredith B. Akerman*, MS, Anne Davidson*, MD
*Feinstein Institute for Medical Research, Manhasset, NY

HOFSTRA NORTH SHORE-LIJ SCHOOL of MEDICINE AT HOFSTRA UNIVERSITY

The Feinstein Institute for Medical Research — North Shore-Long Island Jewish Health System

Introduction

Specific Aim: To propose an exploratory graphical, statistical method for identifying which rows, columns, or cells contribute to the association.

Background:

- Upon 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** (i.e., lack of independence) in more detail.
- This can be particularly difficult for "sparse" tables, where the cell sample sizes are small.
- Snedecor and Cochran (1989) propose that, in r x c tables, the entries providing the largest **percentage contributions to X^2** are those that will suggest departure from the null hypothesis of independence.
- The "**scree plot**" is often used in principal component analysis (PCA) to visually identify the PCs that contain most of the variance (as expressed through the eigenvalues). In PCA, the eigenvalues are plotted against the corresponding PC to produce a scree plot that illustrates the rate of change in the magnitude of the eigenvalues for the PC. This plot is **monotonic non-increasing** and the rate of decline tends to be fast first, eventually leveling off. The '**elbow**', or the point at which the curve bends, is considered to indicate the maximum number of PCs to extract.
- We propose an **adaptation of the scree plot** to visualizing the **largest contributions to X^2** among all of the cells of the r x c table. **Graphical visualization** should reveal a sequence of non-increasing contributions that will level off in such a way as to indicate **which cells are primarily responsible for the departure from the null hypothesis.**

Methods

- Suppose categorical data are cross-classified in a standard r x c contingency table.
- Let O_{ij} = the observed entry in row i, column j, $O_{i\cdot}$ = total of all entries in row i, and $O_{\cdot j}$ = total of all entries in column j.
- Let n denote the sum total of all frequencies. Then, the expected frequency of row i, column j is calculated as $E_{ij} = (O_{i\cdot} \times O_{\cdot j}) / n$.
- The standard Pearson X^2 statistic is calculated as: $X^2 = \sum_i \sum_j [(O_{ij} - E_{ij})^2 / E_{ij}]$, where i = 1, 2, …, r and j = 1, 2, …, c.
- Snedecor and Cochran denote the contribution of the ij^{th} entry to the X^2 statistic as:

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

- The percentage contribution to X^2 from cell $_{i,j}$ is: $P_{ij} = 100 * X^2_{ij} / X^2$.

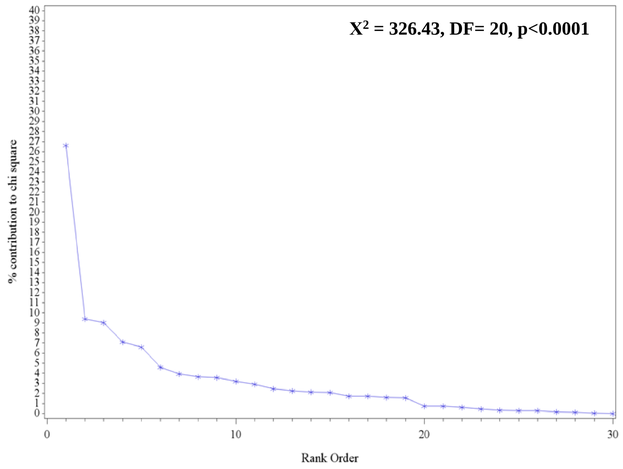
Results

Example 1. Association between Admitting Diagnosis and Race – Large Sample Size (Asymptotic chi-square applies)

Primary Diagnosis on Admission	Race					
Frequency / Cell Chi-Square / Col Pct	White	Black	Hispanic	Asian	Other	Total
DM	39 1.119 23.49	11 29.443 6.25	90 12.9 40.54	13 0.0295 26.53	44 11.674 46.32	197
Chest pain	18 2.0023 10.84	15 5.0579 8.52	56 15.021 25.23	0 7.4054 0.00	18 0.9242 18.95	107
CVA	51 30.674 30.72	8 11.88 4.55	19 5.2703 8.56	14 6.824 28.57	10 0.9929 10.53	102
Fever	22 5.58 13.25	2 10.452 1.14	15 0.4618 6.76	7 2.366 14.29	11 1.4688 11.58	57
GI distress	16 9.5272 9.64	92 86.861 52.27	13 23.183 5.86	15 2.4561 30.61	9 5.6194 9.47	145
Other	20 0.5066 12.05	48 21.542 27.27	29 0.177 13.06	0 6.9209 0.00	3 8.0888 3.16	100
Total	166	176	222	49	95	708

Primary Diagnosis on Admission	Race	Contribution of entry to Chi-Square	O-E	Percent of Column Frequency	% contribution to chi square
GI distress	Black	86.8615	55.9548	52.2727	26.6096
CVA	White	30.6743	27.0847	30.7229	9.3969
DM	Black	29.4426	-37.9718	6.2500	9.0196
GI distress	Hispanic	23.1832	-32.4661	5.8559	7.1021
Other	Black	21.5424	23.1412	27.2727	6.5994
Chest pain	Hispanic	15.0209	22.4492	25.2252	4.6016
DM	Hispanic	12.9003	28.2288	40.5405	3.9519

Additional rows in the table not shown

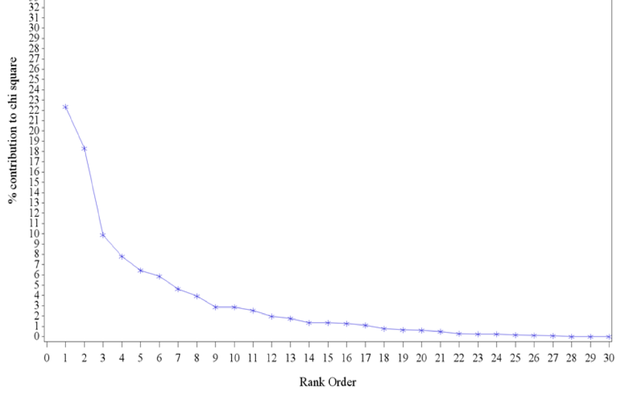


Example 2. Association between Admitting Diagnosis and Race – Small Sample Size (Sparse Table, Exact Test required)

Primary Diagnosis on Admission	Race					
Frequency / Cell Chi-Square / Col Pct	White	Black	Hispanic	Asian	Other	Total
DM	4 0.4356 20.00	10 0.4693 34.48	3 0.983 15.79	2 0.1 22.22	6 1.5803 46.15	25
Chest pain	1 2.6694 5.00	8 0.3755 27.59	8 3.3801 42.11	0 2 0.00	3 0.0043 23.08	20
CVA	9 6.25 45.00	3 1.3517 10.34	3 0.1684 15.79	2 0.0222 22.22	1 0.9846 7.69	18
Fever	3 0.2722 15.00	2 0.4636 6.90	2 0.0058 10.53	1 0 11.11	2 0.2137 15.38	10
GI distress	2 0.0808 10.00	2 0.673 6.90	2 0.0447 10.53	4 7.6455 44.44	1 0.2183 7.69	11
Other	1 0.0833 5.00	4 2.2092 13.79	1 0.0561 5.26	0 0.6 0.00	0 0.8667 0.00	6
Total	20	29	19	9	13	90

Primary Diagnosis on Admission	Race	Contribution of entry to Chi-Square	O-E	Percent of Column Frequency	% contribution to chi square
GI distress	Asian	7.64545	2.90000	44.4444	22.3502
CVA	White	6.25000	5.00000	45.0000	18.2708
Chest pain	Hispanic	3.38012	3.77778	42.1053	9.8812
Chest pain	White	2.66944	-3.44444	5.0000	7.8037
Other	Black	2.20920	2.06667	13.7931	6.4582
Chest pain	Asian	2.00000	-2.00000	0.0000	5.8467
DM	Other	1.58034	2.38889	46.1538	4.6199

Additional rows in the table not shown

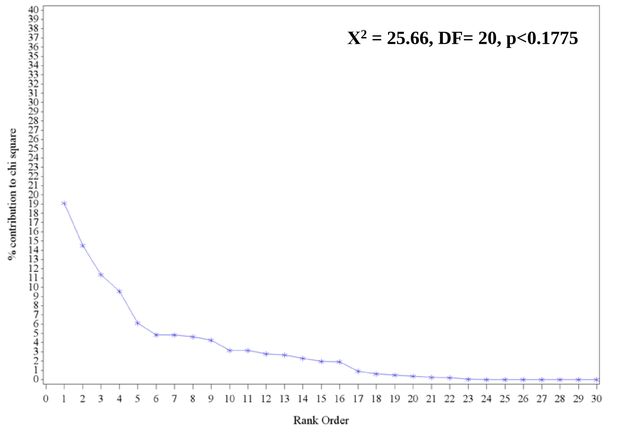


Example 3. Association between Admitting Diagnosis and Race – Large Sample Size (Asymptotic chi-square applies)
NOTE: not significant: use scree to explore

Primary Diagnosis on Admission	Race					
Frequency / Cell Chi-Square / Col Pct	White	Black	Hispanic	Asian	Other	Total
DM	20 0.001 20.00	46 0.8105 23.00	18 0.228 18.00	40 0.002 20.00	17 0.4904 17.00	141
Chest pain	15 0.5904 15.00	36 0.0089 18.00	17 0.0904 17.00	42 0.8058 21.00	18 0.0045 18.00	128
CVA	5 0.7202 5.00	4 1.1905 2.00	7 3.7202 7.00	5 0.503 2.50	3 0.0536 3.00	24
Fever	18 0.6807 18.00	44 0.0019 22.00	16 1.5696 16.00	55 2.9136 27.50	20 0.1578 20.00	153
GI distress	32 1.2422 32.00	50 0.1258 25.00	32 1.2422 32.00	45 1.0905 22.50	25 0.0629 25.00	184
Other	10 0 10.00	20 0 10.00	10 0 10.00	13 2.45 6.50	17 4.9 17.00	70
Total	100	200	100	200	100	700

Primary Diagnosis on Admission	Race	Contribution of entry to Chi-Square	O-E	Percent of Column Frequency	% contribution to chi square
Other	Other	4.90000	7.0000	17.0	19.0985
CVA	Hispanic	3.72024	3.5714	7.0	14.5002
Fever	Asian	2.91363	11.2857	27.5	11.3563
Other	Asian	2.45000	-7.0000	6.5	9.5492
Fever	Hispanic	1.56956	-5.8571	16.0	6.1176
GI distress	White	1.24224	5.7143	32.0	4.8418
GI distress	Hispanic	1.24224	5.7143	32.0	4.8418

Additional rows in the table not shown



Discussion

- We have proposed an exploratory graphical method for discovering the cells that contribute to the departure from independence in an r x c contingency table.
- Techniques used: 1) Assessment of contribution to chi-square in contingency tables, 2) Construction of a scree plot.
- This method is useful in both large sample tables where the chi-square statistic is a valid method, as well as in sparse tables where chi-square is not valid and is replaced by the Fisher's exact test.

A SAS macro that performs the analysis, tabulation of contribution to chi-square, and construction of the screen plot, is available upon request (MLesser@nshs.edu, Makerman@nshs.edu)