APPLICATIONS OF MULTIDIMENSIONAL CONTINGENCY TABLES TO THE ANALYSIS OF TERMINATION COUNTS IN DISABILITY INCOME CLAIM DATA

by

Edward J. Seligman

ABSTRACT

Log-linear analysis is applied to disability terminations, linking them to variables such as sex, age, region and maximum benefit period. Application to risk classification and underwriting are outlines.

APPLICATIONS OF MULTIDIMENSIONAL CONTINGENCY TABLES TO THE ANALYSIS OF TERMINATION COUNTS IN DISABILITY INCOME CLAIM DATA

Edward J. Seligman

Introduction

A contingency table is a set of counts or frequencies obtained by classifying observations in two or more different ways. The classes are called 'categories'. Within each classification the observations must be partitioned such that the categories are:

1) Exhaustive - each observation must fall into some category

2) Mutually Exclusive - an observation cannot fall into more than one category

In other words, an observation must fall into one and only one category of each classification. The number of classifications is called the 'dimension' of the contingency table. An n dimensional table with c_i categories within each classification is said to be a $c_1 \times c_2 \times \ldots \times c_n$ table ($n \geq 2$, $c_i \geq 2$)

Here is an example of a 2 dimensional (2 x 2) contingency table. The classifications are:

1. Claim Status after 1 month duration

   Category 1 - Off claim
   Category 2 - On claim

2. Sex

   Category 1 - Male
   Category 2 - Female

TABLE 1

	OFF CLAIM	ON CLAIM
MALE	43304	76596
FEMALE	13262	21415

CLAIM EXPERIENCE
AFTER ONE MONTH

The next example has 4 rows and 2 columns

TABLE 2

	ON CLAIM	OFF CLAIM
CALIF.	267	1796
N.J.	242	2048
N.Y.	514	4833
REST OF U.S.	2693	30308

CLAIM EXPERIENCE AFTER 12 MONTHS

Here is a 3 x 4 table giving counts of insured individuals on claim, classified by geographical location and by loss year

TABLE 3

	CALIF.	N.J.	N.Y.	REST OF U.S.
1977	474	617	1417	8467
1976	653	747	1795	10820
1975	936	926	2135	13714

INCIDENCE

Finally we have a 3 dimensional (2 x 4 x 2) table

TABLE 4

	MALE		FEMALE	
	ON	OFF	ON	OFF
CALIF.	40	312	12	110
N.J.	44	389	22	162
N.Y.	136	939	50	292
REST OF U.S.	653	6018	146	1650

CLAIM EXPERIENCE AFTER 12 MONTHS LOSS YEAR 1977

Testing the Hypothesis of Independence of Classifications

Within a hypothesis testing framework, we can test any contingency table for independence of classifications. The following definitions will be used, assuming a 2 dimensional table, with an obvious extension to the n dimensional case.

x_{ij} = Count in i^{th} row, j^{th} column

$x_{i.}$ = Total count in i^{th} row (row i marginal)

$x_{.j}$ = Total count in j^{th} column (column j marginal)

p_{ij} = Probability that an observation falls into the cell in row i, column j

$p_{i.}$ = Probability that an observation falls into row i

$p_{.j}$ = Probability that an observation falls into column j

$x_{..}$ = Total number of observations in the table

The following relations are based on the foregoing definitions.

$$x_{i.} = \sum_j x_{ij}$$

$$x_{.j} = \sum_i x_{ij}$$

$$x_{..} = \sum_i x_{i.} = \sum_j x_{.j} = \sum_i \sum_j x_{ij}$$

$$p_{i.} = \frac{x_{i.}}{x_{..}}$$

$$p_{.j} = \frac{x_{.j}}{x_{..}}$$

$$p_{ij} = \frac{x_{ij}}{x_{..}}$$

Now consider the null hypothesis (H_0) that row and column classifications are independent. H_0 implies the following

$$p_{ij} = p_{i.} \cdot p_{.j}$$

The maximum likelihood estimator of p_{ij} under H_0 can be obtained by maximizing L as a function of p_{ij} in the maximum likelihood equation

$$L = \prod_i \prod_j p_{ij}^{x_{ij}}$$

with fixed row and column marginals. The estimator of x_{ij} under H_0 is then computed by multiplying the estimated p_{ij} by $x_{..}$.

We can also estimate x_{ij} under H_0 for any 2-dimensional table by general reasoning. Consider again Table 1. If the classifications of sex and claim status are independent, then the knowledge of a claimant's sex give no information about claim status after one month on claim

TABLE 5

	OFF	ON	
MALE	43304 (44170.2)	76596 (75729.8)	119900
FEMALE	13862 (12995.8)	21415 (22281.2)	35277
	57166	98011	

$$\left(\frac{57166}{155177}\right) \cdot 119900 = 44170.2$$

The best estimate, then, of the probability that a male will be off claim before one month is 57166/155177, or the proportion of all claimants off claim before one month. Since the total number of males is 119900, the best estimate under H_0 of the number of males going off claim before one month is $\frac{57166}{155177} \cdot 119900 = 44170.2$. Looking at this result in another way, we see that the desired quantity can also be written as the row marginal for males times the column marginal for 'off claim' divided by the total observations. This result is the same as that obtained by using the maximum likelihood estimator approach. In general, the estimated count for the cell in row i, column j is given by

$$e_{ij} = \frac{x_{i.} \cdot x_{.j}}{x_{..}}$$

Approximations to the χ^2 variable

The quantity

$$\chi_c^2 \sim \sum_i \sum_j \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$$

is called the 'Pearson χ^2 statistic', and is asymptotically (as each $x_{ij} \rightarrow \infty$) distributed as χ^2 with (r-1) · (c-1) degrees of freedom, where r and c are the number of rows and columns in the contingency table.

In recent years, the 'minimum discrimination information statistic' (sometimes called the 'likelihood ratio statistic') has begun to replace the Pearson χ^2 statistic, because of its application to the construction of an appropriate log-linear model for multidimensional contingency tables. This statistic is also asymptotically distributed as χ^2 with (r-1) · (c-1) degrees of freedom, and is given by

$$\chi_L^2 \sim 2 \cdot \sum_i \sum_j x_{ij} \cdot \ln (x_{ij}/e_{ij})$$

The definitions for χ_c^2 and χ_L^2 given here are for 2-dimensional tables, but the extension to a multidimensional table is quite easy; the summation is performed over all dimensions of the table and the degrees of freedom are calculated by extending the 2-dimensional product chain. In most cases, χ_c^2 and χ_L^2 are close in value.

Partitioning the χ^2 Statistic

For reasons to be given later, hypothesis testing has application mostly to the analysis of 2-dimensional tables. For such tables which are bigger than 2 x 2, we can often proceed beyond the simple test of independence of the two classifications by partitioning the χ^2 statistic. Consider the following 3 x 2 table.

TABLE 6

	ON	OFF
1977	1103	9872
1976	1190	12825
1975	1423	16288

$\chi^2 \sim 35.8$

The χ_e^2 for this table is 35.8 (2d.f.), which is significant at the 1% level, implying that the 12 month termination pattern is dependent on loss year. Suppose that we want to dissect the χ_c^2 statistic into components which will show the source of the dependency of termination pattern on loss year

TABLE 6a

	ON	OFF
1976	1190	12825
1975	1423	16288

$\chi^2 \sim 2.1$

TABLE 6b

	ON	OFF
1977	1103	9872
1975 & 1976	2613	29113

$\chi^2 \sim 33.7$

The χ_c^2 of the original table has been partitioned into 2 components, each with 1 degree of freedom. Each component is associated with a 2 x 2 table. In general, an r x c table with (r-1)·(c-1) degrees of freedom can be partitioned into (r-1)·(c-1) 2 x 2 tables, each with 1 degree of freedom. The justification and methods for partitioning are given by Lancaster (1949, 1950), Irwin (1949), and Kimball (1954). In our numerical example, we see that the dependence of termination pattern on loss year is almost entirely due to a different termination pattern emerging in 1977, compared with the combined 1975 and 1976 experience.

Drawbacks of Hypothesis Testing

We can apply our tests of independence to contingency tables of any dimension. However, the tests for independence between pairs of classifications within n-dimensional tables are cumbersome to use when $n > 2$. Further, the estimates of expected cell counts (e_{ij}) under H_0, the hypothesis of inter-classification independence, can lead to maximum likelihood estimation equations with implicitly defined estimators.

There is a temptation, when faced with a multidimensional table, to "collapse" it over one or more classifications to obtain a 2-dimensional table which can then be easily analyzed. That this can be a dangerous practice is illustrated in the following hypothetical example

TABLE 7

	MALE		FEMALE	
	ON	OFF	ON	OFF
SICKNESS	5	10	20	5
ACCIDENT	200	400	100	25

Inspection shows that for each of the two sexes, sickness and accident claims have the same termination pattern. Thus the classifications of claim status and type of disablement are independent. But if we had regarded the classification of sex as unimportant, and collapsed (summed) the male and female categories, the resulting 2 x 2 table would be

TABLE 7a

	ON	OFF
SICKNESS	25	15
ACCIDENT	300	425

Now we have reason to reject (but incorrectly) our null hypothesis of independence of classifications if we look only at the collapsed table. It has been shown (Bishop, Fienberg, and Holland (1975)), that we can collapse over one or more classifications only if those classifications are independent of at least one of the remaining classifications. This condition is not met in our hypothetical example, hence the false rejection of H_0.

Another limitation of the hypothesis testing analysis of contingency tables is inherent in hypothesis testing itself. Consider the 2 x 2 table

TABLE 8

	ON	OFF
MALE	2561	5970
FEMALE	751	1693

$p_M = 0.300$

$p_F = 0.307$

$\chi^2 \sim 0.5$

The χ_c^2 statistic is not large enough to reject the null hypothesis of independence of classifications. We can also consider another H_0, i.e., that the probability of termination is the same for males and females. Then the test is that of the equality of the parameters of two binomial populations, based on samples from each population. It is easy to show that the two tests are algebraically equivalent; thus this alternate test will also fail to reject its null hypothesis.

Now look at this 'augmented' table

TABLE 8a

	OFF	ON
MALE	256123	597034
FEMALE	75123	171413

$p_M = 0.300$

$p_F = 0.305$

$\chi^2 \sim 18.5$

Both tests now reject their H_0's, but the difference between the sample probabilities of going off claim is closer than what it was in the first table. The reason is that the greatly increased sample size in the second table has enabled us to detect a much smaller difference between the parameters of the two parent binomial populations. It is important that the phrase "statistically significant difference" not be confused with the phrase "significant difference". A difference between sample statistics which is significant at the 1% or 5% level may imply a difference between parent population parameters which is too small to be of any consequence to the actuary or accountant. It must be remembered that the power of a statistical test to reject the null hypothesis almost always increases with the size of the sample.

Analysis Using the Log-Linear Model

We now look at the analysis of contingency tables as a problem in estimation. Instead of testing for the presence or absence of dependency among classifications, we now estimate the size of the dependency. In this way, we can rank the dependencies by their relative sizes.

Consider the 2 x 2 table:

TABLE 9

	ON	OFF	
MALE	x_{11}	x_{12}	$x_{1.}$
FEMALE	x_{21}	x_{22}	$x_{2.}$
	$x_{.1}$	$x_{.2}$	$x_{..}$

Under H_0, the expected value for the count in row i, column j is

$$e_{ij} = \frac{x_{i.} \cdot x_{.j}}{x_{..}}$$

Taking the log of both sides of the above gives

$$\ell n\ e_{ij} = \ell n\ x_{i.} + \ell n\ x_{.j} - \ell n\ x_{..}$$

The log of the expected value of any observation under H_0 is linear in the logs of the row and column marginals; hence the name 'log-linear model'. Now, ignoring H_0, we will write $\ell n\ e_{ij}$ in the following form:

$$\ell n\ e_{ij} = U + U_{1(i)} + U_{2(j)} + U_{12(ij)}$$

where
$$U = \frac{\sum_i \sum_j \ln e_{ij}}{r \cdot c}$$

$$U_{1(i)} = \frac{\sum_j \ln e_{ij}}{c} - U$$

$$U_{2(j)} = \frac{\sum_i \ln e_{ij}}{r} - U$$

$$U_{12(ij)} = (\ln e_{ij} - U) - (U_{1(i)} + U_{2(j)}) \quad .$$

We can relate each of the U's to a quantity of interest. Consider U; it is the overall mean of the log expected value of each cell. The quantity $U_{1(i)}$ is the difference between the mean of the log expected value of the cells in row i and the overall mean, while $U_{2(j)}$ has the same meaning for the cells in column j. Thus $U_{1(i)}$ and $U_{2(j)}$ are the contributions to $\ln e_{ij}$ of classification 1, category i; and classification 2, category j, respectively. The term $U_{12(ij)}$ is the one we are most interested in, since it measures the dependency between classification 1 and classification 2 for categories i and j. This may be explained by the following reasoning. If the deviation between $\ln e_{ij}$ and U were exactly equal to the sum of the contributions of $U_{1(i)}$ and $U_{2(j)}$, then $U_{12(ij)}$ would be zero. This would imply that $U_{1(i)}$ and $U_{2(j)}$ act additively to give the deviation between the log expected value of a cell ($\ln e_{ij}$) and the overall mean (U). The extent, then, to which $U_{12(ij)}$ is <u>not</u> equal to zero is a measure of the non-additivity of $U_{1(i)}$ and $U_{2(j)}$ in their effect on ($\ln e_{ij}$ - U).

The log-linear model is quite similar to the linear model of the analysis of variance (ANOVA). For this reason, ANOVA terminology is commonly used, e.g., U is the 'grand mean', $U_{1(i)}$ and $U_{2(j)}$ are 'main effects', and $U_{12(ij)}$ is an 'interaction'. One of the principal differences between log-linear contingency table analysis and ANOVA is that in the former we are interested only in the interaction term, since the main effects are an indication only of the relative sizes of the counts for individual classifications.

An Application of the Log-Linear Model

We will analyze a subset of the 3-dimensional contingency table already presented

TABLE 10 (BASED ON TABLE 4)

	MALE		FEMALE	
	ON	OFF	ON	OFF
N.Y.	136	939	50	292
REST OF U.S.	653	6018	146	1650

The following model was used:

$$\ell n\ e_{ijk} = U + U_{1(i)} + U_{2(j)} + U_{3(k)} + U_{12(ij)} + U_{13(ik)} + U_{23(jk)} + U_{123(ijk)}$$

where the subscripts 1, 2, and 3 represent sex, geographical region, and claim status. The terms in the model which are of interest to us are:

$U_{12(ij)}$ = interaction of sex and geographical region

$U_{13(ik)}$ = interaction of sex and claim status

$U_{23(jk)}$ = interaction of geographical region and claim status

$U_{123(ijk)}$ = interaction of $U_{12(ij)}$ with claim status

or interaction of $U_{13(ik)}$ with geographical region

or interaction of $U_{23(jk)}$ with sex

The computation of the above quantities for i = j = k = 1 was done by using a program (ECTA), available from the Statistics Department of the University of Chicago. The results were:

$U_{12(11)} = -0.078*$

$U_{13(11)} = 0.005$

$U_{23(11)} = 0.119*$

$U_{123(111)} = -0.046$

An asterisk (*) denotes significance of the statistic at the 1% level. From these numbers, we may conclude that:

1. New York males are underrepresented in the claim population (the sign of $U_{12(11)}$ is negative, thus implying underrepresentation.)
2. Sex has no influence on terminations (note that $U_{13(11)}$ is almost zero)
3. New York has higher persistence than the rest of the U.S. (since $U_{23(11)}$ is positive)

Work in Progress

We have been using individual disability income claim files to construct multidimensional contingency tables with classifications chosen from:

1. State or region
2. Elimination Class
3. Monthly Indemnity
4. Sex
5. Age on Claim
6. Sickness or Accident
7. Loss Year
8. Claim Status after 1, 3, 6, and 12 months duration

The analysis in progress has assessed the relative importance of classifications 1 through 7 on each other, and on classification 8. We have also begun work on a study of incidence; i.e., we are estimating the interactions of demographic classifications with the classification of active life/disabled life based on policy files.

References

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). 'Discrete Multivariate Analysis: Theory and Practice'. Cambridge, Mass., The MIT Press

Everitt, B.S. (1977). 'The Analysis of Contingency Tables' London, England, Chapman and Hall. New York, John Wiley

Fienberg, S.E. (1977). 'The Analysis of Cross-Classified Data' Cambridge, Mass., The MIT Press

Irwin, J.O. (1949). 'A Note on the Subdivision of χ^2 into Components'. Biometrika, Vol. 36. p. 130

Kimball, A.W. (1954) 'Short-Cut Formulas for the Exact Partition of χ^2 in Contingency Tables'. Biometrika, Vol 10, p. 452

Lancaster, H.O. (1949). 'The Derivation and Partition of χ^2 in Certain Discrete Distributions'. Biometrika, Vol. 36, p. 117

Lancaster, H.O. (1950). 'The Exact Partition of χ^2 and its Application to the Problem of Pooling of Small Expectations'. Biometrika, Vol. 37, p. 267.