Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

In addition to chi-square and Fisher exact analysis of contingency tables, using generalized linear models (GLM) to perform loglinear analysis is often preferred as it provides more insight into the structure of the table, and can be extended to contingency tables with more than two dimensions.

From the main SIMFIT menu select [Statistics], [Generalized linear models], [Contingency table analysis], then observe the format of the default data file `loglin.tf1` which contains the following contingency table.

	c_1	c_2	c_3	c_4	c_5
r_1	141	67	114	79	39
r_2	131	66	143	72	35
r_3	36	14	38	28	16

When these data are analyzed, SIMFIT creates a temporary data file formatted for GLM analysis using Poisson error with a log link then applies the constraint that the sum of row coefficients and also the sum of column coefficients add to zero to output the next tables of results.

No. rows = 3, No. columns = 5
Deviance (D) = 9.03788E+00, Deg. freedom = 8
$P(\chi^2 \geq D)$ = 0.3391

Parameter	Value	Lower95%cl	Upper95%cl	Std. error	p
Constant	3.98308	0.0395833	3.89180	4.07435	0.0000
Row 1	0.39606	0.0458291	0.29038	0.50175	0.0000
Row 2	0.41185	0.0456995	0.30646	0.51723	0.0000
Row 3	-0.80791	0.0621905	-0.95132	-0.66450	0.0000
Col 1	0.51116	0.0561557	0.38166	0.64065	0.0000
Col 2	-0.22851	0.0727114	-0.39618	-0.06084	0.0137 *
Col 3	0.46804	0.0569148	0.33679	0.59933	0.0000
Col 4	-0.03156	0.0675080	-0.18723	0.12412	0.6527 ***
Col 5	-0.71913	0.0887225	-0.92373	-0.51454	0.0000

Data	Model	Delta	Dev-resid	Leverage
141	132.9931	8.0069	0.6875	0.6035
67	63.4740	3.5260	0.4386	0.5138
114	127.3798	-13.3798	-1.2072	0.5963
79	77.2915	1.7085	0.1936	0.5316
39	38.8616	0.1384	0.0222	0.4820
131	135.1089	-4.1089	-0.3553	0.6083
66	64.4838	1.5162	0.1881	0.5196
143	129.4063	13.5937	1.1749	0.6012
72	78.5211	-6.5211	-0.7465	0.5373
35	39.4799	-4.4799	-0.7271	0.4882
36	39.8979	-3.8979	-0.6276	0.3926
14	19.0422	-5.0422	-1.2131	0.2551
38	38.2139	-0.2139	-0.0346	0.3815
28	23.1874	4.8126	0.9675	0.2825
16	11.6585	4.3415	1.2028	0.2064

Theory

A contingency table is an array of nonnegative frequencies with n rows and m columns, such as this table contained in SimFIT test file `chisqd.tf4`, for 15 observations carried out on two populations to test for equal probabilities of success.

	Success	Failure	
Sample 1	3	3	6
Sample 2	7	2	9
	10	5	15

Here, the cell frequencies f_{ij} are $(3, 3, 7, 2)$, the sum of row frequencies known as row marginals are $(6, 9)$, the sum of column frequencies known as column marginals are $(10, 5)$, and obviously the row and column marginals must separately both add up to the total number of frequencies (15). The null hypothesis is usually to test for homogeneity or independence, which is the condition that the f_{ij} only depend on row i and column j, and there are no additional influences affecting frequencies in special cells.

To be precise, in the general case there will be frequencies f_{ij} where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$, and it is wished to test for homogeneity, i.e. independence, or no association between the variables, which can be stated as the null hypothesis

$$H_0 : \mu_{ij} = \mu_{i+}\mu_{+j}, \text{ for } i = 1, 2, \ldots, n, \text{ and } j = 1, 2, \ldots, m$$

where each cell probability μ_{ij} is completely determined by the corresponding row marginal μ_{i+}, and the column marginal μ_{+j} probabilities.

To do this, SimFIT defines dummy indicator variables for the rows and columns, then fits a generalized linear model assuming a Poisson error distribution and log link, but imposing the constraints that the sum of row coefficients is zero and the sum of column coefficients is zero, to avoid fitting an over-determined model, and to be consistent with an assumed loglinear model.

The advantage of this approach is that the deviance, predicted frequencies, deviance residuals, and leverages can be calculated for the model

$$\log(\mu_{ij}) = \theta + \alpha_i + \beta_j,$$

where μ_{ij} are the expected cell frequencies expressed as functions of an overall mean θ, row coefficients α_i, and column coefficients β_j. The row and column coefficients reflect the main effects of the categories, according to the above model, where

$$\sum_{i=1}^{n} \alpha_i = \sum_{j=1}^{m} \beta_j = 0$$

and the deviance, which is a likelihood ratio test statistic, can be used to test the justification for a mixed term γ_{ij} in the saturated model

$$\log(\mu_{ij}) = \theta + \alpha_i + \beta_j + \gamma_{ij},$$

which fits exactly, i.e., with zero deviance.

SimFIT performs a chi-square test on the deviance to test the null hypotheses of homogeneity, which is the same as testing that all γ_{ij} are zero, the effect of individual cells can be assessed from the leverages, and various deviance residuals plots can be done to estimate goodness of fit of the assumed loglinear model.

Clearly, the chi-square test for data in test file `loglin.tf1` presented in the previous table does not support rejection of the null hypothesis of homogeneity.