www.statsoft.com

- Products
- Solutions
- Buy
- Trials
- Support

What can we help you find? Search

TextbookLog-Linear Analysis Frequency Tables

Looking for info about statistics?
We wrote the book on it.
And you can read it for free!

Elementary Concepts
Statistics Glossary
Basic Statistics
ANOVA / MANOVA
Association Rules
Boosting Trees
Canonical Analysis
CHAID Analysis
C & R Trees
Classification Trees
Cluster Analysis
Correspondence Analysis
Data Mining Techniques
Discriminant Analysis
Distribution Fitting
Experimental Design
Factor Analysis
General Discrim. Analysis
General Linear Models
Generalized Additive Mod.
Generalized Linear Mod.
General Regression Mod.
Graphical Techniques
Ind.Components Analysis
Linear Regression
Log-Linear Analysis
MARSplines
Machine Learning
Multidimensional Scaling
Neural Networks
Nonlinear Estimation
Nonparametric Statistics
Partial Least Squares
Power Analysis
Process Analysis
Quality Control Charts
Reliability / Item Analysis
SEPATH (Structural eq.)
Survival Analysis
Text Mining
Time Series / Forecasting

Log-Linear Analysis of Frequency Tables

General Purpose

One basic and straightforward method for analyzing data is via crosstabulation. For example, a medical researcher may tabulate the frequency of different symptoms by patients' age and gender; an educational researcher may tabulate the number of high school drop-outs by age, gender, and ethnic background; an economist may tabulate the number of business failures by industry, region, and initial capitalization; a market researcher may tabulate consumer preferences by product, age, and gender; etc. In all of these cases, the major results of interest can be summarized in a multi-way frequency table, that is, in a crosstabulation table with two or more factors.

Log-Linear provides a more "sophisticated" way of looking at crosstabulation tables. Specifically, you can test the different factors that are used in the crosstabulation (e.g., gender, region, etc.) and their interactions for statistical significance (see Elementary Concepts for a discussion of statistical significance testing). The following text will present a brief introduction to these methods, their logic, and interpretation.

Correspondence analysis is a descriptive/exploratory technique designed to analyze two-way and multi-way tables containing some measure of correspondence between the rows and columns. The results provide information which is similar in nature to those produced by Factor Analysis techniques, and they allow one to explore the structure of the categorical variables included in the table.

To index

Two-way Frequency Tables

Variance Components
Statistical Advisor
Distribution Tables
References Cited
Send Comments
Business Solutions
Free Resources
About Textbook

Let us begin with the simplest possible crosstabulation, the 2 by 2 table. Suppose we were interested in the relationship between age and the graying of people's hair. We took a sample of 100 subjects, and determined who does and does not have gray hair. We also recorded the approximate age of the subjects. The results of this study may be summarized as follows:

Gray	Age		
Hair	Below 40	40 or older	Total
No	40	5	45
Yes	20	35	55
Total	60	40	100

While interpreting the results of our little study, let us introduce the terminology that will allow us to generalize to complex tables more easily.

Design variables and response variables. In multiple regression (Multiple Regression) or analysis of variance (ANOVA/MANOVA) one customarily distinguishes between independent and dependent variables. Dependent variables are those that we are trying to explain, that is, that we hypothesize to depend on the independent variables. We could classify the factors in the 2 by 2 table accordingly: we may think of hair color (gray, not gray) as the dependent variable, and age as the independent variable. Alternative terms that are often used in the context of frequency tables are response variables and design variables, respectively. Response variables are those that vary in response to the design variables. Thus, in the example table above, hair color can be considered to be the response variable, and age the design variable.

Fitting marginal frequencies. Let us now turn to the analysis of our example table. We could ask ourselves what the frequencies would look like if there were no relationship between variables (the null hypothesis). Without going into details, intuitively one could expect that the frequencies in each cell would proportionately reflect the marginal frequencies (Totals). For example, consider the following table:

Gray	Age		
Hair	Below 40	40 or older	Total
No	27	18	45
Yes	33	22	55
Total	60	40	100

In this table, the proportions of the marginal frequencies are reflected in the individual cells. Thus, 27/33=18/22=45/55 and 27/18=33/22=60/40. Given the marginal frequencies, these are the cell frequencies that we would expect if there were no relationship between age and graying. If you compare this table with the previous one you will see that the previous table does reflect a relationship between the two variables: There are more than expected (under the null hypothesis) cases below age 40 without gray hair, and more cases above age 40 with gray hair.

This example illustrates the general principle on which the log-linear analysis is based: Given the marginal totals for two (or more) factors, we can compute the cell frequencies that would be expected if the two (or more) factors are unrelated. Significant deviations of the observed frequencies from those expected frequencies reflect a relationship between the two (or more) variables.

Model fitting approach. Let us now rephrase our discussion of the 2 by 2 table so far. We can say that fitting the model of two variables that are not related (age and hair color) amounts to computing the cell frequencies in the table based on the respective marginal frequencies (totals). Significant deviations of the observed table from those fitted frequencies reflect the lack of fit of the independence (between two variables) model. In that case we would reject that model for our data, and instead accept the model that allows for a relationship or association between age and hair color.

To index

Multi-way Frequency Tables

The reasoning presented for the analysis of the 2 by 2 table can be generalized to more complex tables. For example, suppose we had a third variable in our study, namely whether or not the individuals in our sample experience stress at work. Because we are interested in the effect of stress on graying, we will consider Stress as another design variable. (Note that, if our study were concerned with the effect of gray hair on subsequent stress, variable stress would be the response variable, and hair color would be the design variable.). The resultant table is a three-way frequency table.

Fitting models. We can apply our previous reasoning to analyze this table. Specifically, we could fit different models that reflect different hypotheses about the data. For example, we could begin with a model that hypothesizes independence between all factors. As before, the expected frequencies in that case would reflect the respective marginal frequencies. If any significant deviations occur, we would reject this model.

Interaction effects. Another conceivable model would be that age is related to hair color, and stress is related to hair color, but the two (age and stress) factors do not interact in their effect. In that case, we would need to simultaneously fit the marginal totals for the two-way table of age by hair color collapsed across levels of stress, and the two-way table of stress by hair color collapsed across the levels of age. If this model does not fit the data, we would have to conclude that age, stress, and hair color all are interrelated. Put another way, we would conclude that age and stress interact in their effect on graying.

The concept of interaction here is analogous to that used in analysis of variance (ANOVA /MANOVA). For example, the age by stress interaction could be interpreted such that the relationship of age to hair color is modified by stress. While age brings about only little graying in the absence of stress, age is highly related when stress is present. Put another way, the effects of age and stress on graying are not additive, but interactive.

If you are not familiar with the concept of interaction, we recommend that you read the Introductory Overview to ANOVA/MANOVA. Many aspects of the interpretation of results from a log-linear analysis of a multi-way frequency table are very similar to ANOVA.

Iterative proportional fitting. The computation of expected frequencies becomes increasingly complex when there are more than two factors in the table. However, they can be computed, and, therefore, we can easily apply the reasoning developed for the 2 by 2 table to complex tables. The commonly used method for computing the expected frequencies is the so-called iterative proportional fitting procedure.

The Log-Linear Model

The term log-linear derives from the fact that one can, through logarithmic transformations, restate the problem of analyzing multi-way frequency tables in terms that are very similar to ANOVA. Specifically, one may think of the multi-way frequency table to reflect various main effects and interaction effects that add together in a linear fashion to bring about the observed table of frequencies. Bishop, Fienberg, and Holland (1974) provide details on how to derive log-linear equations to express the relationship between factors in a multi-way frequency table. Chi-square test. You can compute two types of Chi-squares, the traditional Pearson Chi-square statistic and the maximum likelihood ratio Chi-square statistic (the term likelihood ratio was first introduced by Neyman and Pearson, 1931; the term maximum likelihood was first used by Fisher, 1922a). In practice, the interpretation and magnitude of those two Chi-square statistics are essentially identical. Both tests evaluate whether the expected cell frequencies under the respective model are significantly different from the observed cell frequencies. If so, the respective model for the table is rejected.

Goodness-of-Fit

In the previous discussion we have repeatedly made reference to the "significance" of deviations of the observed frequencies from the expected frequencies. One can evaluate the statistical significance of the goodness-of-fit of a particular model via a

Reviewing and plotting residual frequencies. After one has chosen a model for the observed table, it is always a good idea to inspect the residual frequencies, that is, the observed minus the expected frequencies. If the model is appropriate for the table, then all residual frequencies should be "random noise," that is, consist of positive and negative values of approximately equal magnitudes that are distributed evenly across the cells of the table.

Statistical significance of effects. The Chi-squares of models that are hierarchically related to each other can be directly compared. For example, if we first fit a model with the age by hair color interaction and the stress by hair color interaction, and then fit a model with the age by stress by hair color (three-way) interaction, then the second model is a superset of the previous model. We could evaluate the difference in the Chi-square statistics, based on the difference in the degrees of freedom; if the differential Chi-square statistic is significant, then we would conclude that the three-way interaction model provides a significantly better fit to the observed table than the model without this interaction. Therefore, the three-way interaction is statistically significant.

In general, two models are hierarchically related to each other if one can be produced from the other by either adding terms (variables or interactions) or deleting terms (but not both at the same time).

Automatic Model Fitting

When analyzing four- or higher-way tables, finding the best fitting model can become increasingly difficult. You can use automatic model fitting options to facilitate the search for a "good model" that fits the data. The general logic of this algorithm is as follows. First, fit a model with no relationships between factors; if that model does not fit (i.e., the respective Chi- square statistic is significant), then it will fit a model with all two-way interactions. If that model does not fit either, then the program will fit all three-way interactions, and so on. Let's assume that this process found the model with all two-way interactions to fit the data. The program will then proceed to eliminate all two-way interactions that are not statistically significant. The resulting model will be the one that includes the least number of interactions necessary to fit the observed table.

To index