Visualizing Categorical Data in ViSta

Pedro M. Valero-Mora
Universitat de València-EG
Email: valerop@uv.es

Forrest W. Young
University of North Carolina
Email: forrest@unc.edu

Michael Friendly
York University
Email: friendly@yorku.ca

This paper presents the modules in the statistical package ViSta related to categorical data analysis. These modules are: visualization of frequency data with mosaic and bar plots, correspondence analysis, multiple correspondence analysis and loglinear analysis. All these methods are implemented in ViSta with a big emphasis on plots and graphical representations of data, as well as interactivity for the user with the system. These provide a system that has shown to be easy, useful, and powerful, both for novice and experienced users.

Keywords: Categorical data, visualization, statistical packages, programming languages.

Introduction

Young (Young and Bann, 1997; Young, Faldowsky and MacFarlane, 1993) has developed ViSta, a statistical package based on Lisp-Stat. ViSta incorporates the object-oriented approach as part of its internal and external functioning. In particular, it extends Lisp-Stat with additional graphical, statistical and data objects; it provides objects for mapping the process of data analyses and it has objects that guide novices through their early attempts to carry out analyses. All these characteristics shape a system that has been shown to be appropriate for students and teachers of statistics as well as for researchers and developers in computational and graphical statistics.

ViSta is focused on techniques for visualizing data. Therefore, traditional statistical methods are considered from a graphical, dynamic, linked and interactive way. One of the most innovative characteristics for visualization of statistical methods in ViSta is the idea of spreadplots: coherently linked displays that simultaneously provide alternative views of data or model objects, and interactive manipulation and analysis (Young et al., in press). At the time this is being written, ViSta integrates about 20 different kinds of spreadplots. These include spreadplots for exploring raw data, numeric (univariate, bivariate and multivariate), category, classification, frequency classification (one-way and n-way), frequency table, crosstabulation and data simulation. ViSta also has spreadplots for data transformations such as the Box-Cox, folded power and missing data imputation. Finally, there are spreadplots for visualizing statistical models, including analysis of variance, correspondence analysis, multiple correspondence analysis, multidimensional scaling, multivariate regression, principal components, regression analysis, univariate analysis, cluster analysis and frequency analysis.

This paper will focus on spreadplots for categorical data. As remarked by Friendly (2000), there is an almost paradoxical disparity between availability and use of methods for visualization of quantitative data and categorical data. While it is habitual that an analyst carrying out, for example, analysis of regression uses plots for exploration and model fitting on a routine basis, it is certainly unusual to see the same practice with categorical data.

The plan of this paper will be the following: First, we will describe the different types or representations of categorical data that are built into ViSta. This information is important, because ViSta decides which analysis are acceptable as a function of the type of data selected and, consequently, the analyst will have to know how to transform the data to the right type when necessary. Second, we will describe the visualization for raw frequency data available in ViSta. This visualization provides a preliminary examination of the data that can help to decide the type of analysis to be carried out. Third, we will describe correspondence analysis and homogeneity analysis (also known as multiple correspondence analysis). Fourth, we will describe loglinear analysis, using an extended example to illustrate the dynamic and interactive facilities for model building in ViSta. A general discussion with the trends and improvements planned for ViSta will close the paper.

Data representations of categorical data in ViSta.

Categorical data can be represented in different ways in ViSta. As statistical analysis methods will sometimes expect data of a particular type and many of them are equivalent, ViSta provides transformations that change data from one representation to another. Hence, it is important to know the different representations and the possible actions that can be pursued with each one and the way to carry out the transformations.

We will use as example the data of survival from the sinking of the Titanic that were introduced by Dawson (1995). This data presents the cross-classification of 2201 passengers and crew by Age (Child, Adult), Gender (Female, Male), Class (1st, 2nd, 3rd, Crew) and Survival (Lived, Died). The analysis of this dataset is usually concerned with analyzing the relationship among the background variables and how these are related to survival. Data are shown in Table 1.

Gender	Age	Survived	Class			
			1st	*2nd*	*3rd*	*Crew*
Male	*Adult*	*Died*	118	154	387	670
Female			4	13	89	3
Male	*Child*		0	0	35	0
Female			0	0	17	0
Male	*Adult*	*Lived*	57	14	75	192
Female			140	80	76	20
Male	*Child*		5	11	13	0
Female			1	13	14	0

Table 1: Data of survival from the sinking of the Titanic (from Dawson, 1995)

This dataset will be used throughout the paper to illustrate the different visualizations in ViSta. One of the most appealing questions to be examined in this data is whether the phrase "Women and children first" often heard in relation with these events actually is a good description of what happened this day. We will investigate this fact and others using the different capabilities existing in ViSta.

ViSta distinguishes three types of data:

1. **Table data** displays the categories of one variable as rows and the categories of another as columns. The frequencies of the cross-tabulation are shown in the cells. Table data type is only appropriate for bivariate data (although three- and higher-way data can be accommodated by stacking two or more

variables along the rows and/or columns). The analysis that can be carried out directly with this type of data in ViSta are basic frequency analysis and correspondence analysis.

2. **Frequency classification data** has a row for each combination of categories of the categorical variables in the dataset. A column variable lists the frequency associated to the row. Frequency classification data is appropriate for n-way data. Basic frequency analysis and loglinear analysis can be applied directly to this type of data.
3. **Category data** has a row for each individual case in the data. This representation is quite wasteful in terms of disk space but it has the advantage of being compatible with multivariate data, the most general data type in ViSta. Frequency analysis and homogeneity analysis is applicable directly to this type of data.

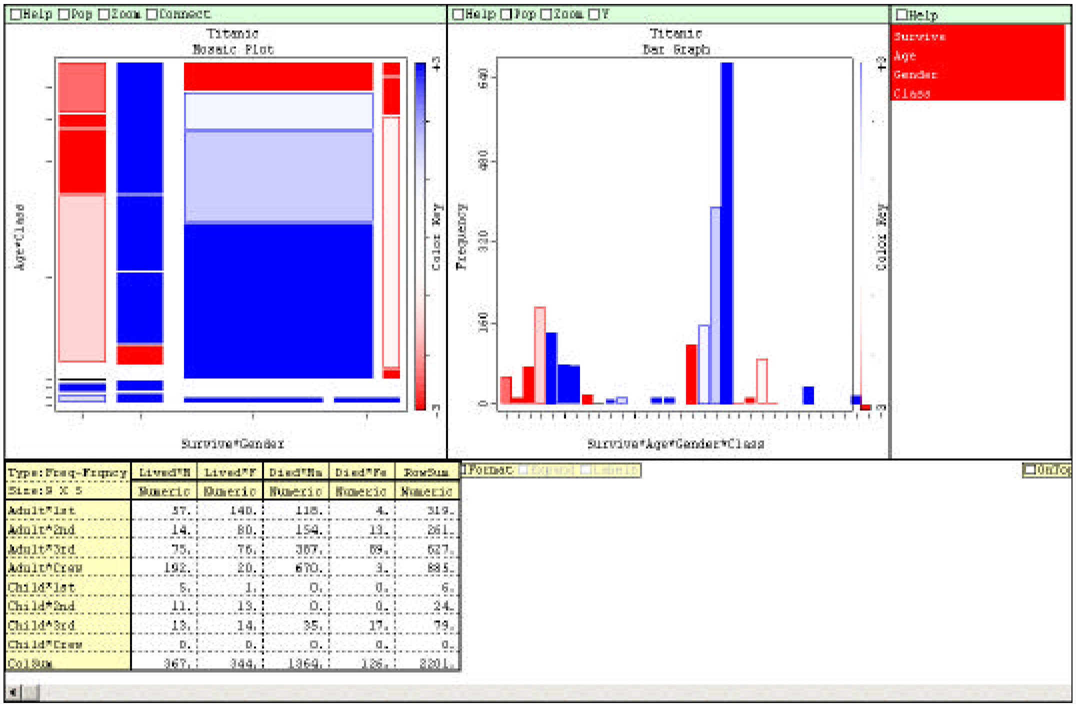
ViSta can transform data from a representation to another so the user can change easily the data before applying an analysis.

Basic visualization of categorical data

Visualization of raw frequency data is a topic that has received considerable attention in recent times. We illustrate ViSta's frequency data spreadplot with data from the sinking of the Titanic. Figure 1 shows a spreadplot of the cross-classification of the four variables in this dataset.

This spreadplot includes the following plots for the combinations of up to four classification variables (more variables may be used, but four is the maximum that can be combined)

1. A mosaic plot of frequencies. This plot visualizes an n-way contingency table by portraying the frequencies as "tiles" whose size (area) is proportional to the table's frequencies. The colors encode the Chi Square residuals of the cell with respect to the model of mutual independence (in this model the probabilities in a cell are products of the one way marginal probabilities). Blue means a positive residual. Red indicates a negative residual. The names of the categories, the value of the observed frequency and the value of the residual of the associated cell in the table data are shown when the pointer of the mouse is moved across the mosaic plot.
2. A stacked bar graph of the frequencies. This plot shows bars that have a length proportional to the frequencies in the cells. A variable can be crossed with these bars so they are split into stacks of rectangles that are proportional to the value of the corresponding cell. Colors of the bars are the same as the colors of the Mosaic plot's tiles. Mouse movement has a similar effect on the bars as on the tiles.
3. A table of frequencies for the cross-classification of the categories of the variables included in the visualization.
4. A list of ways of the table (classification variable names). These allow the user to determine which variables are visualized in the other cells of the spreadplot.



Type:Freq-Frqncy	Lived*M	Lived*F	Died*Ma	Died*Fe	RowSum
Size:8 X 5	Numeric	Numeric	Numeric	Numeric	Numeric
Adult*1st	57.	140.	118.	4.	319.
Adult*2nd	14.	80.	154.	13.	261.
Adult*3rd	75.	76.	387.	89.	627.
Adult*Crew	192.	20.	670.	3.	885.
Child*1st	5.	1.	0.	0.	6.
Child*2nd	11.	13.	0.	0.	24.
Child*3rd	13.	14.	35.	17.	79.
Child*Crew	0.	0.	0.	0.	0.
ColSum	367.	344.	1364.	126.	2201.

Figure 1: Spreadplot for raw frequency analysis

The spreadplot for frequency data is designed to make it easy to explore the effect of interactively adding new variables to it. This is controlled by the "Change Plots" window in the upper right corner, a window that displays the names of the ways of the table. Clicking on a name produces a "one-way" spreadplot for the variable selected. A ctrl-click on another name adds that name to those already selected and changes the spreadplot to a "two-way" spreadplot. Three-way and four-way spreadplots can be created by additional ctrl-clicks. Similar effects can be obtained by dragging the mouse across adjacent names. Mosaic plots and bar plots provide complementary views of the data, because the former reports proportional sizes of cells and the later displays absolute values.

The mosaic plot in Figure 1 reveals patterns of residuals running from top to bottom in the upper part of the plot. The columns in this part correspond, respectively, to Males and Females that Lived and to Males and Females that Died. The rows within these columns are for the four categories of Class (1st, 2nd, 3rd and Crew), first for Adults and then for Children. The Survival rates for Adult Males and Females are sorted according to the Class. For example, while the Survival for Adult Males in 1st Class was lower than expected (top leftmost cell in the mosaic plot) it is better than the Survival for 2nd Class (lower cell) and 3rd Class (next lower cell). A similar pattern is shown for Females in the second column. 1st and 2nd Class attain better Survival rates than 3rd Class for Adults. However, Females had higher Survival rates than Males. This suggests an interaction between Class, Age and Survival. We will have the opportunity to corroborate this observation later in the paper.

Correspondence Analysis

Correspondence analysis is an exploratory technique that displays the rows and columns of a contingency table as points in a graph. This technique is regarded as a dimensionality reduction method because it tries to fit the data in typically two or three dimensions accounting with the maximum of variance.

The Titanic data does not lend itself to correspondence analysis directly because this technique usually focuses on two-way tables of frequencies. An alternative approach (van der Heijden and de Leeuw, 1985; Friendly, 2000) consists of constructing two-way tables by stacking two or more variables along the rows and/or columns. For example, for a four-way table, with variables A,B,C,D combining A and B along the rows and B and C along the columns is equivalent to fitting the loglinear model [AB][CD]. In our case, we concatenated the Survival and Gender variables on one hand, and the Age and Class variables on the other hand. The table so obtained is shown in Table 2. A correspondence analysis of this table analyzes the residuals from the loglinear model [SG][AC]. This analysis ignores the association between the Survival and Gender variables and between the Age and Class. However, the rest of the interactions will be decomposed in the correspondence analysis of the data in Table 2.

	Adult-1st	Adult-2nd	Adult-3rd	Adut-Crew	Child-1st	Child-2nd	Child-3rd	Total
Lived-Male	57	14	75	192	5	11	13	367
Lived-Female	140	80	76	20	1	13	14	344
Died-Male	118	154	387	670	1	0	35	1364
Died-Female	4	13	89	3	0	0	17	126
Total	319	261	627	885	6	24	79	2201

Table 2: Cross tabulation of the Gender-Survive and Age-Class variables

The visualization for Correspondence Analysis for this table is shown in Figure 2. This visualization includes five plots and a list. We will describe them from left to right and top to down.

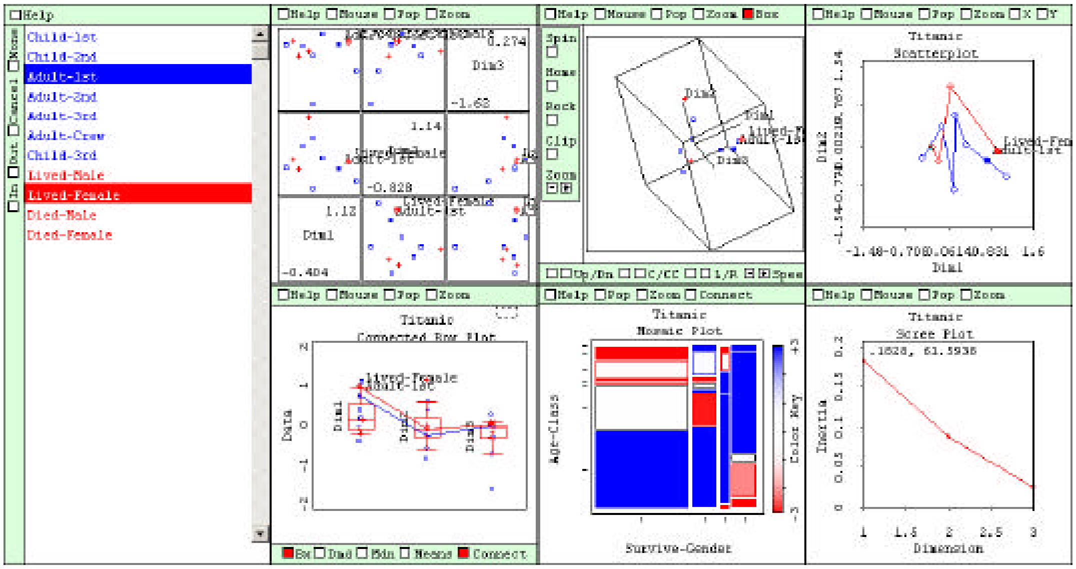


Figure 2. Spreadplot for correspondence analysis

1. List of categories for variables analyzed. This list is linked to the spin plot, the scatterplot, the scatterplot-matrix and the boxplot of categories of variables. Colors are used to distinguish between the column variable (Survival-Gender) and the row variable (Age-Class).
2. Scatterplot-matrix. This plot is linked with the spin-plot, the scatterplot and the boxplot. Clicking on a plot cell selects the dimensions shown in other plots of the spreadplot.

3. Spin-plot of row and column-points. This plot portrays three dimensions of the correspondence analysis solution. The usual controls for spin-plots are provided.

4. Scatterplot of row and column points. This plot shows initially the first two dimensions of the correspondence analysis solution. Bubbles proportional to the quality of the representation of the points in the dimensions chosen for the analysis are drawn over the points. Lines connect the points in each category variable to improve interpretation.

5. Boxplot of row and column points. This plot shows the coordinates of the correspondence analysis solution. It can show as many dimensions as computed in the analysis.

6. Mosaic plot of the data. The colors of the cells symbolize the standardized residuals from the model of independence of the stacked table as in the basic visualization of the data described previously. The rows and columns are sorted according to the scores of the objects in the first dimension of the correspondence analysis. This makes those cells with positive residuals fall approximately on the diagonal running from left down to right top. Negative residuals will fall on the diagonal running from left top to right down. This `effect ordering' (Friendly and Kwan, 2002) of the row and column variables is designed to facilitate perception of patterns, trends and anomalies.

7. Scree plot. The line in the plot represents the inertia of the dimensions of the result. Brushing the spreadplot or selecting a point will show the value of the inertia and the percentage of total inertia of each dimension.

The spreadplot for correspondence offers many alternative views for visualizing the results of an analysis. The user can select the plot that more closely agrees with the number of dimensions suggested by the plot of inertia. We may focus on the 2D scatterplot for the data because two dimensions seem to be enough for interpretation of the results. The circles drawn are quite similar so the quality of the representation for the points seems appropriate for them. We have selected the points for Adults-1st and Lived-Female, showing that these two categories are in the same direction.

Homogeneity Analysis-Multiple Correspondence Analysis

ViSta performs multiple correspondence analysis (MCA) using the homogeneity analysis by alternating least squares (HOMALS) method (Gifi, 1990). MCA provides a graphic portrayal of the bivariate relationships between categorical variables, and can be useful for understanding large, multivariate categorical datasets. MCA is a technique that has received many different names and that has been derived in different ways in different disciplines and contexts (Tenenhaus & Young, 1985).

The visualization for homogeneity analysis is shown in Figure 3. There are four elements in this visualization.

1. List of the objects or rows in the dataset and the categories of the variables or columns. This list is linked with both the spin-plot and the scatterplot matrix. Colors are used to distinguish between variables. In this case, the categories of the variable Survive are coded in red, Age in green, Gender in grey and Class in purple.

2. Spin-plot for the scores of the categories and the objects in the dataset. Categories are represented using a square symbol while objects are shown by

a dot. Notice that when there are several objects with the same combination of categories they will be represented with exactly the same score in the plot, so each visible dot will correspond to many objects.

3. A scatterplot matrix of the first 5 dimensions of the result. This plot is also a control for the spin-plot because clicking on any combination of variables will cause it to show these variables.
4. A plot of the discrimination measures for each variable and, simultaneously, the inertia by each dimension. The discrimination measures indicate how well a variable is represented by each dimension (Gifi, 1990). They are the sum of the squares of the distances of the categories of every variable with respect to the origin. The inertia of each dimension is the average of the discrimination measures of all the variables for that dimension and is represented by the green line in the plot. The labels of the variables and the profiles along the dimensions can be examined by using the brush tool in the scatterplot.

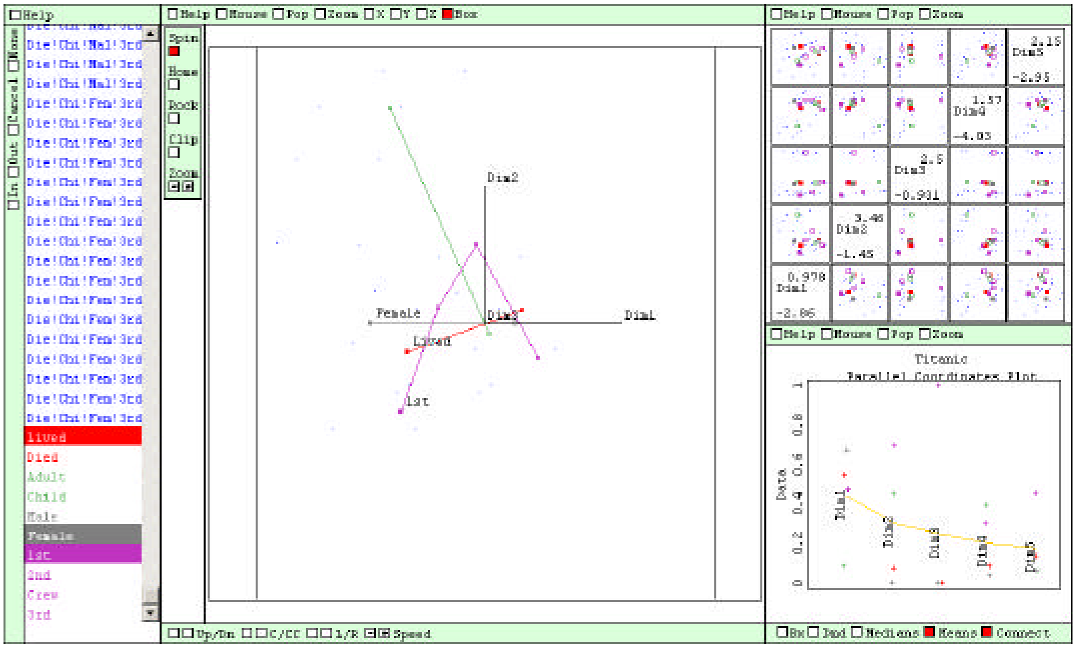
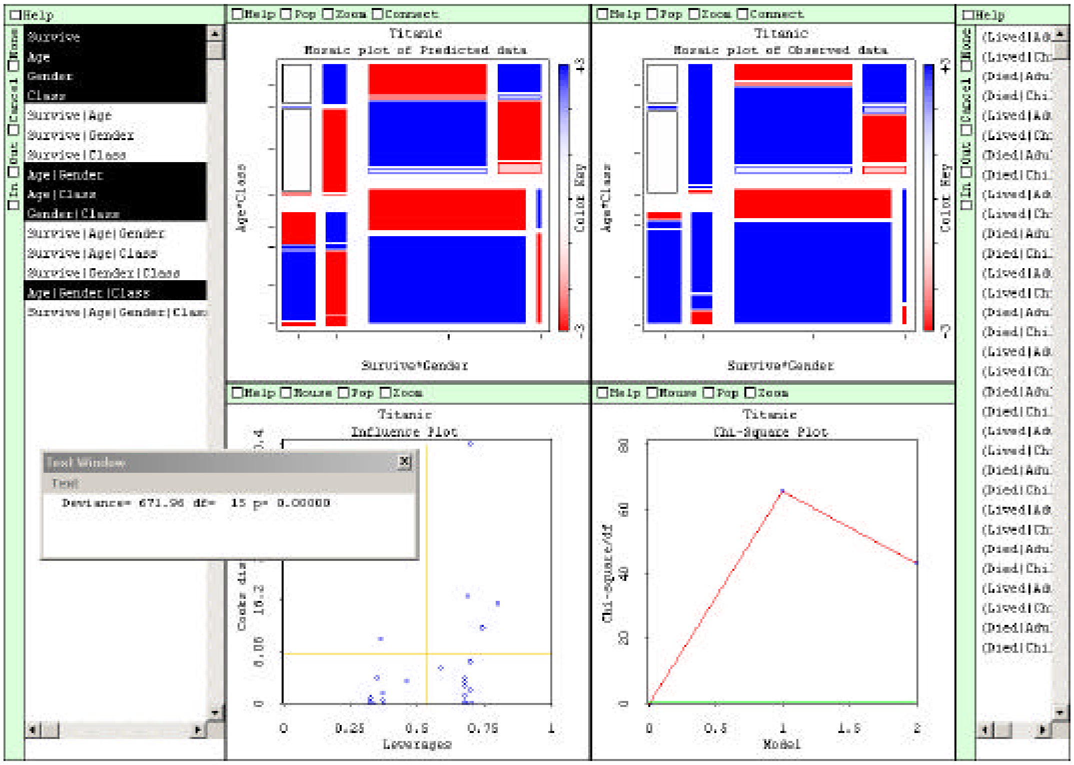


Figure 3: Spreadplot for homogeneity analysis

The spin-plot in Figure 3 has been moved to a position where only the first and second dimensions are visible. We have selected the points corresponding to the categories Lived, Female and 1st. The visualization shows that Females traveling in 1st Class had an advantage with regard to surviving.

Loglinear analysis

Loglinear models provide a method for analyzing associations between two or more categorical variables and they have become widely accepted as a tool for researchers during the last two decades. There is therefore a range of excellent textbooks which give overviews of categorical data analysis by means of loglinear analysis (Agresti, 1990; Ato and López 1996; Andersen 1996; Christensen 1990) and almost every major statistical package includes capabilities for computing them. However, the usual computer programs for loglinear analysis present the results in terms of tables of parameter estimates. This makes it difficult to understand the nature of the associations, particularly for large tables. Friendly (2000) has emphasized the use of graphics and plots for helping to understand more easily the results of loglinear models, arguing that graphics should play the same roles for data exploration and model building with categorical data that they do with numeric data. To illustrate the interactive capabilities for model building and diagnosis in Loglin-ViSta, we show a few steps in the analysis of the Titanic data.



Survive
Age
Gender
Class
Survive|Age
Survive|Gender
Survive|Class
Age|Gender
Age|Class
Gender|Class
Survive|Age|Gender
Survive|Age|Class
Survive|Gender|Class
Age|Gender|Class
Survive|Age|Gender|Class

```
Deviance= 671.96 df=  15 p= 0.00000
```

Figure 4: Spreadplot for loglinear analysis

We first describe the organization of the spreadplot for loglinear analyisis in ViSta, shown in Figure 4. This spreadplot includes two list windows, four plots and a floating text output window. The user can fit loglinear models, visualize them and carry out model comparisons using the spreadplot by selecting the terms to be included in the model in the Terms window on the left of the spreadplot. This window lists all the possible terms for the data analyzed, including every possible interaction. In the Titanic dataset, this window has one term for each of the four variables in the dataset (Survival, Gender, Class and Age), six terms for the two-way interactions, four terms for the three-way interactions, and one additional term for the four-way interaction.

The user selects or deselects the terms required for a model by using Ctrl-mouse_click or Ctrl-mouse_drag. This list has two working modes. When the list is in the Hierarchical mode, including an interaction term involves automatically adding all the terms hierarchically below the (added) one. On the other hand, when the list is in the non-Hierarchical mode, each term is added or removed individually. Note that LoginViSta uses the same computational method for the fitting of hierarchical or non-hierarchical models (Tierney, 1991). From left to right, top to bottom, the plots in the spreadplot are:

1. A mosaic plot of Predicted values using the current loglinear model selected in the terms window. This mosaic plot has colors that portray the adjusted

chi square residuals for the model (blue means positive residual, red means negative).

2. The standard mosaic plot of observed values using the adjusted chi square residuals for the model to color the tiles.
3. A diagnostic scatterplot of leverages versus Cook's distances, useful for identifying badly-fitted cells. Limits for points considered inadequate are drawn in the plot using reference lines (leverage: #parameters / #cells; Cook's D: 4/#cells). Cells may be identified by selecting points in this plot or their labels in the list window at the left.
4. A “model history” line plot for the chi Square divided by the degrees of freedom of the successive models computed for all models within an interactive session. A rule of thumb for a model to be considered adequate using this measure is X^{2} /df˜ 1. This line plot allows for past models review and for model comparison. More significantly, clicking on a point corresponding to a past model will make the spreadplot show the terms included, the plots and the numerical information of that model. Selecting two points will make the floating window display a test for the difference of the models in case that the models are nested.

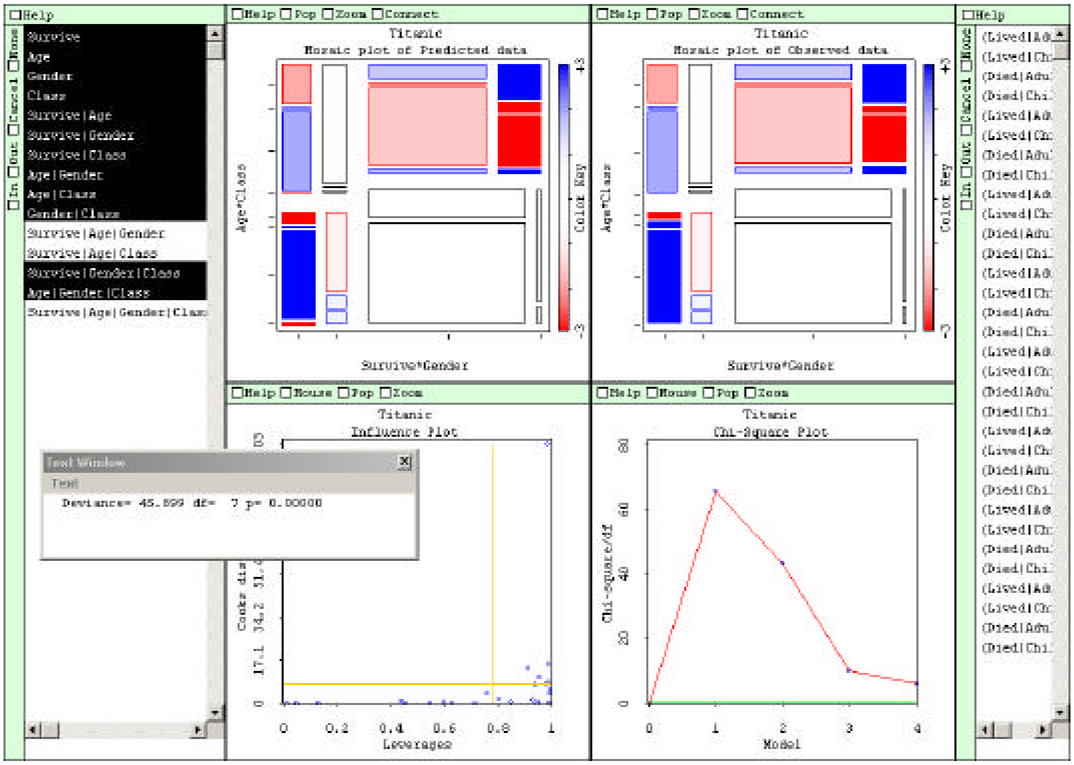
The spreadplot also includes a floating window that shows numerical information about the fit (deviance X^2 , *df, p*-value) of the current model to the data.

The combination of all these plots together with the smoothness of user interaction provides the possibility of quickly exploring and testing hypothesis in the data as well as comparing the fit of the different models considered. Thus, Figure 4 shows the spreadplot after three models have been fit. Model 0 is the saturated model. Model 1 is the independence model, with only the four one-way terms (Survival, Gender, Class and Age) selected. The independence model can be considered as a “null” model because no relation between any variables is specified. This corresponds to the basic visualization for frequency data in Figure 1 (except that Loglin-ViSta uses adjusted rather than simple Pearson residuals). Because we are mainly interested in the relation between the explanatory variables and Survival, Model 2 is the “baseline” model, which includes all posible interactions among Gender, Class and Age, but none with Survival. This model controls for these background variables, but pretends that Survival is independent of them all jointly.

However, the shading in the mosaic plots in Figure 4 reveals that there is considerable association between the background variables and Survival. This is confirmed with the Deviance lack-of-fit measure displayed in the Deviance plot and printed in the floating window (Deviance= 671.96 df= 15 p< 0.001). A second visual check of the fit Model 2 is provided by visual comparison of the mosaic plot of predicted data versus the mosaic plot of observed data. Finally, the diagnostic plot of leverage vs. Cook’s D highlights some badly-fitted and influential cells. We believe that this combination of multiple visualizations of model fit, combined with the model history line plot and directly-manipulable model specification provides an effective tool for analysis of loglinear models, and illustrates the power of the spreadplot approach of ViSta.

The natural next step is to include associations of each of the background variables with Survival. In a general way, these associations were shown in the homogeneity analysis of Figure 3. In the loglinear analysis, individual testing carried out after sequential inclusion of each term showed that the reduction of Deviance obtained was always significant, giving Model 3, [AGC] [AS] [GS] [CS] (not shown to conserve space). This model says that Survival depends on Age, Gender, and Class ("women, children and first class first"), but with none of their interactions. Sadly, this simple model does not fit well, as we saw in all the panels of the spreadplot.

Yet, the mosaic plots in that spreadplot *immediately* revealed an interaction of Gender and Class on Survival (requiring the three-way term [GCS])--- for men the association between Class and Survival is much stronger than for women. Adding this term gives Model 4, [AGC] [AS] [GCS], whose spreadplot is shown in Figure 5.



```
Deviance= 45.899 df=  7 p= 0.00000
```

Figure 5: Spreadplot for loglinear model [AGC] [AS] [GCS]

This spreadplot reveals a potential difficulty with mosaic plots when there are low (or zero) frequency cells or high-order margins. In the Titanic data, there are a small total number of children. Tables of Age by other factors can have, therefore, a number of low frequency cells, which appear in mosaics as tiles with small area, so it is hard to see the residuals for these cells. One simple solution is to provide an alternative mosaic plot, where frequencies are shown as square roots, rather than as raw values, as shown in Figure 6.

Comparing the bottom part of the mosaic plots of Figure 6 (Age=Child) with the top part (Age=Adult) we can observe the interaction between Class and Age on Survival. This is done examining the residuals for the cells for the same category in Gender, Class and Survival for children and adults, where we see that the residuals have an opposite pattern over the Class categories.

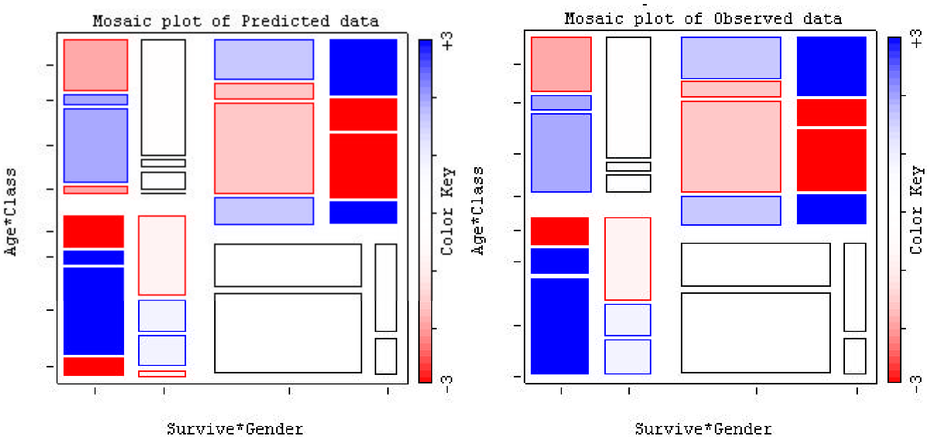
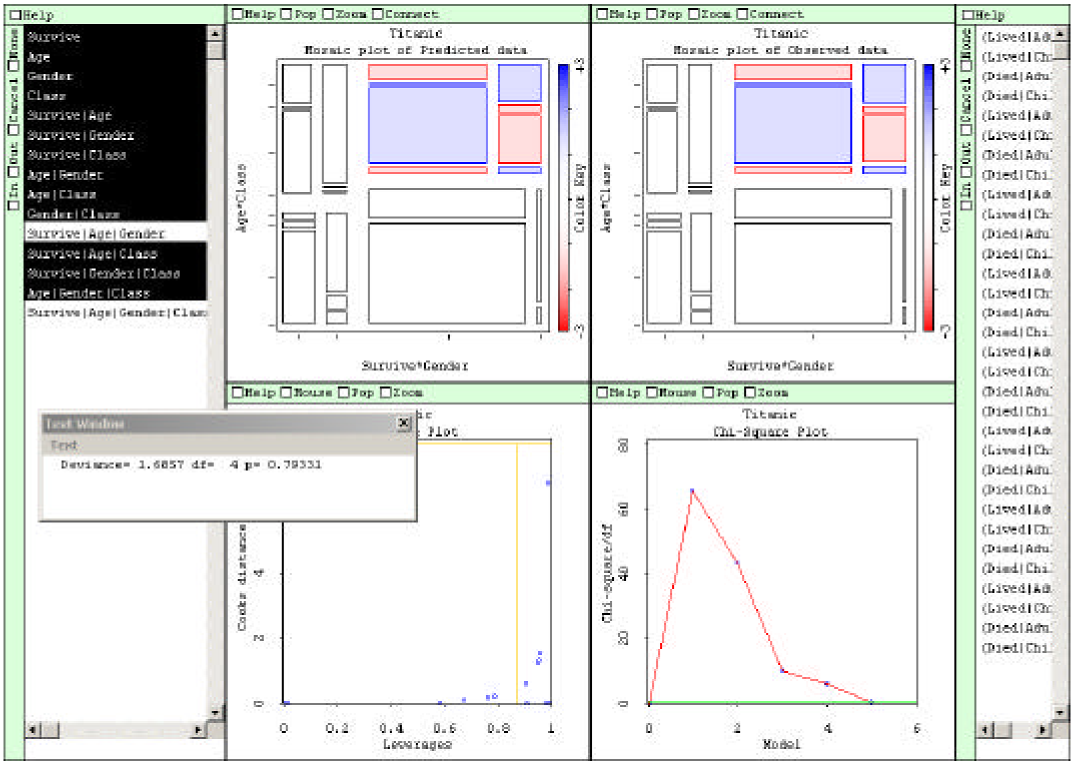


Figure 6: Mosaic plots for model [AGC] [AS] [GCS] showing square roots

A model that includes the interaction for Age and Class on Survival is shown as Model 5 in Figure 7. This model fits the data very well. In this figure, neither the shading of the mosaic plots nor the diagnostic plot reveal any residuals of importance and the deviance is now non-significant. (The comparative test of the difference between Models 4 and 5, obtained by selecting both models in the X^2 plot, show that this interaction is significant (+Age | Class | Survival: ? D=44.214, ? d.f.=3, p<0.0001)).



```
Survive
Age
Gender
Class
Survive|Age
Survive|Gender
Survive|Class
Age|Gender
Age|Class
Gender|Class
Survive|Age|Gender
Survive|Age|Class
Survive|Gender|Class
Age|Gender|Class
Survive|Age|Gender|Class
```

```
Deviance= 1.6857 df=  4 p= 0.79331
```

```
(Lived|Ad
(Lived|Ch
(Died|Adu
(Died|Chi
(Lived|Ad
(Lived|Ch
(Died|Adu
(Died|Chi
(Lived|Ad
(Lived|Ch
(Died|Adu
(Died|Chi
(Lived|Ad
(Lived|Ch
(Died|Adu
(Died|Chi
(Lived|Ad
(Lived|Ch
(Died|Adu
(Died|Chi
(Lived|Ad
(Lived|Ch
(Died|Adu
(Died|Chi
(Lived|Ad
(Lived|Ch
(Died|Adu
(Died|Chi
(Lived|Ad
(Lived|Ch
(Died|Adu
(Died|Chi
```

Figure 7: Spreadplot for model [AGC][ACS][GCS]

It is difficult to convey the ease and power of a truly dynamic, interactive system for loglinear models in words and a few static screen-shots. Yet, we hope that this example has shown how the various views of the data, fitted model and residuals presented in the spreadplot combine and work together to provide the user with an effective tool for model building, diagnosis and comparison.

Discussion

The present paper has focused on the application of interactive graphics to categorical data and to statistical analysis techniques for categorical data. The methods for visualization of raw frequency data, correspondence analysis, multiple correspondence analysis and loglinear analysis presented here allow the user to interact with the graphics to obtain a deeper and sounder insight about the results. Also, spreadplots, a way for laying out and controlling several graphics simultaneously, have been used extensively, making available to the user a variety of visualizations and displays that provide alternative or complementary views of the information for each method.

The present paper describes visualizations that are ready to be used by researchers and practitioners interested in graphics for categorical data. However, the process of developing these visualizations has led us to think of new possibilities. In particular, it seems to be interesting to add data manipulation features to mosaic plots in the same way that scatterplots or other plots for numerical variables already have in some systems (such as excluding or modifying properties of the cases). Thus, operations like collapsing categories, sorting, transforming, and linking are features that we plan to develop in the near future.

REFERENCES

Agresti, A., *Categorical Data Analysis* (Wiley, New York, 1990).

Andersen, E.B., *Introduction to the Statistical Analysis of Categorical Data* (Springer-Verlag, New York, 1996).

Ato M. and López, J.J., *Análisis estadísticos para datos categóricos* (Síntesis, Madrid, 1996).

Christensen, R., *Log-Linear Models* (Springer-Verlag, New York, 1990).

Dawson, R. J. M., The "unusual episode" data revisited, *Journal of Statistics Education*, **3** (3) (1995).

Friendly, M. and Kwan, E., Effect order for data displays. In press, *Computational Statistics and Data Analysis*, (2002).

Friendly, M., Conceptual and Visual models for categorical data, *The American Statistician*, **49** (1995) 153-160.

Friendly, M., Mosaic displays for multi-way contingency tables, *Journal of the American Statistical Association*, **89** (1994) 190-200.

Friendly, M., *Visualizing Categorical Data* (SAS Institute, Cary NC, 2000)

Gifi, A., *Nonlinear Multivariate Analysis* (Wiley, Chichester, 1990).

Snee, R., Graphical Display of Two-Way Contingency Tables. *The American Statistician*, **28** (1974) 9-12.

Tenenhaus, M. and Young, F. W., An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, **50**(1) (1985) 91-119.

Tierney, L. Generalized Linear Models in Lisp-Stat (Technical Report Nº 557, School of Statistics, University of Minnesota, 1991).

Tierney, L., *Lisp-Stat. An Object Oriented Environment for Statistical Computing and Dynamic Graphics* (Wiley, New York, 1990).

van der Heijden, P. G. M. and de Leeuw, J., Correspondence analysis used complementary to loglinear analysis, *Psychometrika*, **50** (1985) 429-447.

Young, F. W. and Bann, C., ViSta: A Visual Statistics System, in R. Stine and J. Fox (Eds.), *Statistical Computing Environments for Social Research* (Sage, Thousand Oaks, 1997) 207-235.

Young, F. W., Faldowski, R. A. and McFarlane, M. M., Multivariate Statistical Visualization, in C. R. Rao, (Ed.) *Handbook of Statistics*, Vol. 9 (Elsevier, Amsterdam, 1993) 959-998.

Young, F. W., Valero-Mora, P. M., Faldowski, R. A. and Bann, C. Gossip: The Architecture of Spreadplots. *Journal of Computational and Graphical Statistics*, In press.