ViSta

The Visual Statistics System

Analysis of Variance

Forrest W. Young & Carla M. Bann

THE L.L. THURSTONE
PSYCHOMETRIC LABORATORY
UNIVERSITY OF NORTH CAROLINA
CB 3270 DAVIE HALL, CHAPEL HILL N.C., USA 27599-3270

VISUAL STATISTICS PROJECT
WWW.VISUALSTATS.ORG
REPORT NUMBER 2000-1
JANUARY, 2000

Analysis of Variance

ForrestYoung and Carla M. Bann

This report presents ViSta-ANOVA, the ViSta procedure for performing analysis of variance. This procedure is capable of analyzing balanced or unbalanced, complete (i.e., every cell must have at least 1 observation) n-way data for main effects and two-way interactions. ViSta-ANOVA's unique visualization includes a box, diamond, and dot plot; a frequency polygon and histogram; a profile plot; a residual plot; and a partial regression plot.

1 Introduction

Analysis of variance (ANOVA) allows researchers to test for differences in the means of several different groups or populations. ANOVA tests the null hypothesis that the means for all of the groups are equal. In order to test this hypothesis, an F statistic is calculated which compares the variation among the groups with the variation within the groups.

The analysis of variance model makes the following assumptions about the data: (1) the variances of the groups are equal (homogeneity of variance) and (2) the errors are normally distributed. If the assumptions of the model are not met, the results of the analysis may be inaccurate. The visualization for ANOVA is designed to provide information on the validity of these assumptions, as well as information about the significance of effects. It includes a residual plot which is useful for checking if the errors are normally distributed. Also, a violation of the homogeneity of variance may be detected by examining the visualization's box plot or residual plot. The partial regression plot provides information about the significance of effects, and the profile plot shows information about the specific levels of an effect.

We illustrate the various features of the ViSta-ANOVA program with an example of data on the hunting habits of three foxes and three coyotes. During each season of the year, the number of miles each animal wandered from its den was recorded. The dataset is displayed in Figure 1.

An analysis of variance with two-way interactions was performed on this dataset. The season of the year, the species of the animal, the specific subject animal, and the two-way interactions of these variables were used to predict the average number of miles each animal wandered from its den. The results of the analysis are presented in the visualization and report, discussed in the next paragraphs.

The visualization, shown in Figure,2 has five plots, plus a list of source names. Clicking on source names in the Sources window changes the plots in the other windows so that they display information about the chosen

Animals DataSheet (DataType: Classification)

☐Help ☐Save ■Lock ☐Restore

4 Vars / 24 Obs	Miles	Species	Subject	Season
	Numeric	Category	Category	Category
Coyote 1 Fall	4.00	Coyote	1	Fall
Coyote 1 Spring	7.00	Coyote	1	Spring
Coyote 1 Summer	8.00	Coyote	1	Summer
Coyote 1 Winter	2.00	Coyote	1	Winter
Coyote 2 Fall	5.00	Coyote	2	Fall
Coyote 2 Spring	6.00	Coyote	2	Spring
Coyote 2 Summer	6.00	Coyote	2	Summer
Coyote 2 Winter	4.00	Coyote	2	Winter
Coyote 3 Fall	7.00	Coyote	3	Fall
Coyote 3 Spring	8.00	Coyote	3	Spring
Coyote 3 Summer	9.00	Coyote	3	Summer
Coyote 3 Winter	5.00	Coyote	3	Winter
Fox 1 Fall	0.00	Fox	1	Fall
Fox 1 Spring	5.00	Fox	1	Spring
Fox 1 Summer	3.00	Fox	1	Summer
Fox 1 Winter	0.00	Fox	1	Winter
Fox 2 Fall	3.00	Fox	2	Fall
Fox 2 Spring	5.00	Fox	2	Spring
Fox 2 Summer	4.00	Fox	2	Summer
Fox 2 Winter	1.00	Fox	2	Winter
Fox 3 Fall	4.00	Fox	3	Fall
Fox 3 Spring	6.00	Fox	3	Spring
Fox 3 Summer	2.00	Fox	3	Summer
Fox 3 Winter	3.00	Fox	3	Winter

Figure 1: Animals Datasheet

source. As shown, they show information about “season”. Examining the box, diamond, and dot plot reveals that the observations for the summer season appear to have greater variability than those from the other seasons. On the other hand, spring seems to have the smallest variability. This suggests that the assumption of homogeneity of variance may be violated. The residual plot does not seem to indicate that there are any outlying data points, but it does suggest that residuals are non-homogeneous. The confidence curves in the partial regression plot cross the horizontal line, indicating that the effect for season is significant at the .05 level. The profile plot shows that spring has the highest mean, and winter the lowest. This means that the animals wander farthest from their dens during the spring, and stay closest to their dens during winter

The report, shown in Figure 3, displays numeric information about the ANOVA results. The top portion of the report presents details about the data being analyzed, including the ways and classes of the design. The middle portion of the report (which is optional and is not shown in the Figure) displays details about the regression analysis underlying the analysis of variance. The regression analysis involves a design matrix of “Terms” (dummy variables) which is constructed to fit the model to the data. The bottom portion of the report presents summary fit statistics, an overall model fit test, and the standard ANOVA table. The ANOVA table shows the significance of the main (and, optionally, interaction) effects. This is the most important portion of the report. For these data the report shows that the overall model fits significantly, and that the fitted model accounts for 96% of the variance in the response variable. It also shows that all main effects, but no two-way interactions, are significant

Finally, the interpretation, a portion of which is shown in Figure 4, presents a several paragraphs interpreting the results. The prose is constructed to reflect the actual results, and varies depending on what the results are like.

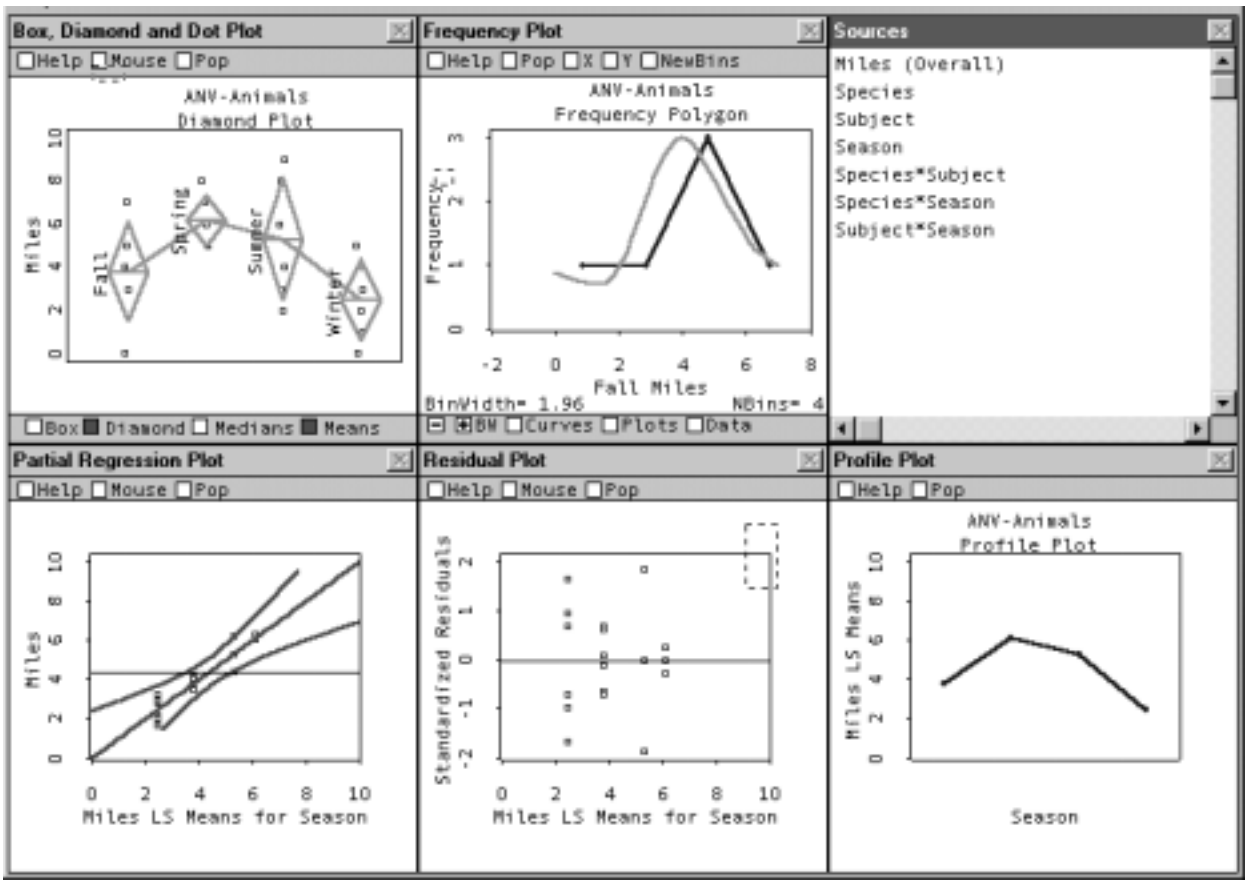

Figure 2: Animals Data SpreadPlot

ANALYSIS OF VARIANCE of the Animals data

Model:
Miles = linear function of Species + Subject + Season + Species*Subject +
Species*Season + Subject*Season

Response Variable:	Miles
Way Names:	(Species Subject Season)
Two-Way Interactions?	Yes
Number of Classes:	(2 3 4)
Class Names:	((Coyote Fox) (1 2 3) (Fall Spring Summer Winter))
Number of Observations:	24
Number of Cells:	24
All Cell Frequencies Are:	1
Design Type:	Balanced

RESULTS:

SUMMARY OF FIT:

R Squared (Total Effect Strength):	0.96
Adjusted R Squared:	0.85
Sigma Hat (RMS error):	0.95
Number of cases:	24
Degrees of freedom:	6

ANALYSIS OF VARIANCE: EFFECTS TESTS

Source	Sum-of-Squares	df	Mean-Square	F-Ratio	Significance F-Value	Strength R-Square
Species	51.04	1	51.04	56.54	0.0003	0.37
Subject	14.58	2	7.29	8.08	0.0199	0.11
Season	47.46	3	15.82	17.52	0.0025	0.34
Species*Subject	2.58	2	1.29	1.43	0.3104	0.02
Species*Season	7.46	3	2.49	2.75	0.1345	0.05
Subject*Season	9.42	6	1.57	1.74	0.2592	0.07
All Sources	132.54	17	7.80	8.64	0.0069	0.96
Error	5.42	6	0.90			
Total	137.96	23				

Figure 3: ANOVA Report for the Animals Data

F-TEST - Full Model:
For these data we performed an F-Test having 17 and 6 degrees of freedom. This test resulted in F=8.64, p=0.0069. This means that when the null hypothesis of no effect for any source in the model is true, the probability is p=0.0069 of obtaining a sample of data with an effect that is as extreme or is more extreme. Thus, of many repetitions of this sampling design, the probability is p=0.0069 that we would be wrong when we reject the null hypothesis of no effect in favor of the alternative hypothesis that at least one pair of treatment means is different.

STRENGTH OF RELATIONSHIP - Full Model:
Based on this sample of data, we estimate that 96% of the variability in Miles is related to variability in all sources. This estimate would vary over repeated samples of data.

Figure 4: Interpretation of the ANOVA of the Animal Data

2 Using the Analysis of Variance Procedure

ViSta-ANOVA performs analysis of variance on n-way data. The data may balanced or unbalanced, but they must be complete (no empty cells). The analysis procedure computes information about main effects and, optionally, two-way interactions. The procedure allows users to view the results of the analysis using a visualization containing five interactive plots. The visualization provide information about the validity of the assumptions of the analysis, as well as about the fit of the model to the data. In addition, a report displays parameter estimates, fit statistics and test statistics. Finally, an interpretation can be produced which consists of several paragraphs of prose the present the appropriate interpretation of the results. The following sections present details

2.1 Preparing Data for Analysis

ViSta-ANOVA analyzes classification data (data with one numeric variable and one or more category variables) and multivariate data (when the data contain one or more of each kind of variable). Existing data can be loaded into ViSta using the Open Data menu items. New data may be entered via the data editor using the New DataSheet menu item, or imported via the Import Data item. (Note that this is much simpler than in previous releases, in which the data had to be in the form of table data, which no longer exist in ViSta).

2.2 Analysis Option

You can perform analysis of variance when the current data are classification data or multivariate data. Simply select the Analysis of Variance item from the Analyze menu or click on the ANOVA button on the toolbar. When the data are more than one way (that is, where there is more than one category variable), a dialog box is displayed, giving you the choice of including two-way interaction terms in the model. If the data are only one way, the dialog box will not be displayed. The data must be complete, but do not have to be balanced.

2.3 Command Lin

You may also perform analysis of variance by typing `(analysis-of-variance)` in the listener. This gives you the default analysis, which is without interaction terms. There are several keywords that may be used when typing in the listener window. These keywords and their default values are given below:

`:data` Used to change the name of the dataset to be analyzed. The name is entered without quotes. Default is the current data.

`:title` Used to specify a title of the report. The title is entered with quotes. Default is "`Analysis of Variance`".

`:name` Used to control the name of the model object which is created. Default is "`ANV-`" concatenated in front of the current data's name.

`:interaction` Used to include (`t`) or exclude (`nil`) interaction terms. Default is `nil`.

`:dialog` The value `t` indicates that the parameter dialog box should be displayed, whereas `nil` indicates that the dialog box should not be displayed. Default is `nil`.

For example, for an analysis of variance of the current data, with interactions, type:

```
(analysis-of-variance :interaction t).
```

2.4 Report

You can see the report of the analysis by clicking on the report button on the WorkMap's toolbar; by clicking on the report icon attached to the model icon (if it is displayed); by selecting the Report Model item from the Model menu; by typing `(report-model)`; or by sending the model object the `:report-model` message. After clicking, you will see a dialog box that provides a choice of report options. By default, the report method displays two-tailed t-tests for each parameter, an F test for each effect being fit to the response, including the test's P-value (the significance of the effect); and the R-squared value for each effect (the size of the effect in terms of proporion of response variance fit by the model). It can optionally report the LS Means, the parameter estimates, additional fit indices, and model effects.

2.5 Interpretation

You can see the interpretation of the results by choosing the Interpret Model item from the Model menu. This provides you with a set of paragraphs which describe, in English prose, the results of the data. This information is appropriate for inclusion in written material about your analysis.

2.6 Statistical Visualization

You can see the visualization of the analysis results by clicking on the visualize button on the WorkMap's toolbar; by clicking on the visualize icon attached to the model icon (if it is displayed); by selecting the Visualize Model item from the Model menu, by typing `(visualize-model)` in the listener window; or by sending the model object the `:visualize-model` message. The visualization is a spreadplot of six interacting plots. This section describes the plots contained in the spreadplot.

The **Sources** window contains a list of the sources in the model that has been fit to the data. These are the overall model, the main effects and the two-way interactions. This window is used to control what is displayed in the other windows. When the user clicks on one of the source names in this window, the profile, residual, box, and leverage plots are changed so that they display the information for that source.

The **Box, Diamond, and Dot Plot** plots the classes of the selected source on the horizontal axis, and the values of the response variable on the vertical axis. The plot displays the data within a source course as dots plotted against their values. The plot allows the user to view box or diamond plots. The user may select whether boxes, diamonds, or both are displayed by clicking on the Box and Diamond buttons. Also, the user may use the Mean and Median buttons to display lines connecting means or medians (or both) of the data within a level of the source.

The box plot displays quantiles of the response values within a class of the source. The plot is used to examine the spread of the observations and to compare distributions of the data for the various classes of the source. A line is drawn through the box at the median (the 50th quantile) of the values in the class. The top of the box is at the 75th quantile and the bottom of the box is at the 25th quantile. The upper and lower lines are at the 90th and 10th quantiles, respectively. The sample size of each class is represented by the width of the box. Classes with larger sample sizes will have wider boxes.

The diamond plot is used to compare the means of each class. The distance from the upper to lower point of the diamond represents a 95% confidence interval. The sample size of each class is represented by the width of the diamond. Classes with larger sample sizes will have wider diamonds.

The **Frequency Polygon** / **Histogram Plot** presents schematic visual information about the shape of the distribution of the response variable for a given cell or categorization of your data. These distribution should be symmetric in shape. The Plots button controls whether you see a frequency polygon, a histogram or a hollow histogram. The frequency polygon usually presents a better schematic of the shape than the histogram. Note that both depend heavily on the binning chosen, and that it is recommended that you "play" with the bin controls to see many different polygons/histograms. This gives a better idea of the shape of the data. You can also use the Curves button to fit a Normal or Kernel Density curve to the distribution, if you wish, to help judge symmetry and normality. Finally, if you wish to save the current binning information, use the Data button.

The **Partial Regression Plot** allows you to visually determine whether an effect of a source is significant. The plot is a plot of the response variable versus the Least Squared (LS) Means for the selected ANOVA source. The LS Means are the values of the response variable that are predicted by the selected source. Since the LS Mean for a given level of the selected source is the same for all observations within that level, the plot shows vertical lines of dots. The dots in a line are the observations within a level of the source.

The plot shows the relationship between the response variable and the predictions of the response made by the selected source. This relationship is represented by the scatter of points, and it is summarized by the straight. 45 degree line. This line is the (partial) regression line. The slope and intercept of this line are based on the parameter estimates computed by the analysis. If the scatter of points displays a linear relationship, then the assumption of linearity is satisfied for the analysis. The strength of relationship is displayed by the scatter of points around the regression line.

The plot also shows a horizontal line and two curved lines. The horizontal line is drawn at the mean of the response variable. The two curved lines are the upper and lower 95% confidence boundaries for the (partial) regression. If these lines intersect with the horizontal line, then the ANOVA source is significant, at the 95% level, in predicting the response variable. If the confidence curves do not cross the horizontal line the effect is not significant at that level. This plot is the same as the leverage plot discussed by Sall (1990). We call it a partial regression plot because this name is more commonly used in the literature, and because "leverage plot" has a different commonly used meaning.

The **Residual Plot** displays the response variable on the horizontal axis versus the residuals from fit on the vertical axis. When the "Overall" source is selected the horizontal axis is raw data. When another source is selected the horizontal axis is least squares means for the classes of the selected source. The LS Means are the values of the response variable that are predicted by the selected source. Since the LS Mean for a given level of the selected source is the same for all observations within that level, the plot shows vertical lines of dots. The dots in a line are the observations within a level of the source. The Y-axis button may be used to change the type of residuals displayed on the vertical axis. The types include raw residuals, studentized residuals, or externally studentized residuals. These are explained in Tierney (1990).

The residuals plot is an ANOVA diagnostic plot: It helps diagnose the suitability of the assumptions underlying ANOVA for the data being analyzed. Residual plots may be used to detect nonnormal error distributions, nonconstant error variance (heteroscedasticity), nonlinearity and outliers.

- NORMALITY: The points in the plot should be normally distributed about the zero line within each source level. If they are not, then the assumption of normality has probably not been met.
- LINEARITY: Points that form a systematic pattern within a souce level suggest that the assumption of linearity has been violated.
- HETEROSCADASTICITY: The variance of the residuals should be about the same for all source levels. If the variance changes systematically across the levels, then the assumption of constant error variance has not been met.
- OUTLIERS: Outliers may be identified by examining observations which have residuals that are much larger than the rest of the residual values. There should be no outliers.

The **Profile Plot** displays the least squares means for each of the classes of the selected source versus the classes. The means are connected by a line to emphasize the relationship between classes. This plot is useful for determining if the means of the groups are different. When the selected source is an interaction term, there will be several profile lines. When these profile lines are roughly parallel, there is no significant interaction effect. If the lines are not parallel or are intersecting, there may be a significant interaction.

3 Algorithm

ViSta-ANOVA creates indicator variables using the same coding as is used in the SAS JMP software (SAS Institute, 1989). These indicator variables become predictor variables in a multiple regression in which the observed data are the response. The analysis is performed by an nway model object which does not require the remainder of the ViSta system. It creates the indicator variables and then has the XLispStat regression model object perfor the analysis.

4 References

Sall, J. (1990) *Leverage Plots for General Linear Hypothesis*. The American Statistician, 44(4). pp. 308-315.

SAS Institute (1989). *JMP User's Guide*. SAS Institute. Cary NC.

Tierney, L. (1990) *Lisp-Stat: An Object-Oriented Environment for Statistical Computing & Dynamic Graphics*. Addison-Wesley, Reading, Massachusetts.