

Statistical Analysis with WebStat, a Java applet for the World Wide Web

R. Webster West

Department of Statistics
University of South Carolina
Columbia, SC 29208

R. Todd Ogden

Department of Statistics
University of South Carolina
Columbia, SC 29208

Abstract

The Java programming language has added a new tool for delivering computing applications over the World Wide Web (WWW). WebStat is a new computing environment for basic statistical analysis which is delivered in the form of a Java applet. Anyone with WWW access and a Java capable browser can access this new analysis environment. Along with an overall introduction of the environment, the main features of this package are illustrated, and the prospect of using basic WebStat components for more advanced applications is discussed.

1 Introduction

The establishment of the World Wide Web (WWW) has added a new avenue for the delivery of statistical applications and software. West et. al. (1997) discuss a forms-based interface for statistical procedures which utilizes the Common Gateway Interface (CGI) protocol. For this type of interface, a user simply inputs information into a WWW form and then submits the information for processing on the serving machine. This approach can be burdensome to both developers and users as the computational load is placed squarely on the server. An overloaded server means a developer using the server may be slowed down and results may be provided to the user at the remote site at a very slow rate. Readers interested in this type of interface are referred to the Globally Accessible Statistical Procedures (GASP) initiative at <http://www.stat.sc.edu/rsrch/gasp/>.

The development of the Java computing language has added a new alternative to the forms-based interface. Actual applications in the form of Java applets may be transferred over the World Wide Web which then run locally on the user's machine. The quality of application in terms of the user's ability to interact is far superior for this new approach than with the forms-based interface. The only user requirement is a Java capable WWW browser such as Netscape Navigator 2.0 or higher, Microsoft's Internet Explorer, or Sun Microsystems' Hotjava. This is not much of a restriction, however, as Java capable browsers are used by roughly 95% of the people browsing the WWW.

Utilizing the new approach, the authors have developed WebStat, a basic statistical package for delivery over the WWW. The WWW address for WebStat is

`http://www.stat.sc.edu/~west/webstat.`

This site also contains extensive documentation on using WebStat. A screen image of the WebStat home page is given in Figure 1. The current version of WebStat is 1.0 Beta.



Figure 1: The WebStat home page

2 Using WebStat

When a user clicks on the button to download WebStat, three new windows will appear on the users machine, the WebStat main window, the Graphics window, and the Statistics window. WebStat's main window has the look and feel of some commercial data analysis packages such as Minitab and Systat. It is made up of a menubar and a data table. The menubar contains three menus for importing data and conducting numerical and graphical statistical procedures.

2.1 Data input

Data sets are stored as columns of variables; that is, a bivariate data set should have an “x” column and a “y” column. Users may use the default *var1*, *var2*, etc., or specify their own variable names

by clicking on the default variable name. Missing values are allowed, as is exponential notation. WebStat currently allows for three modes of data entry:

- Typing data into the data table
- “Pasting” data into the data table
- Importing data from a WWW accessible location
- Loading WebStat’s sample data.

The data table allows the use of scrollbars, arrow keys, and mouse clicks to move about. Numbers may be typed in from the keyboard and edited using the backspace and delete keys. Due to the current implementation of Java in some browsers, all of these navigation tools may not be available under all platforms.

An alternative (and simple) method for entering data into WebStat is to highlight the data in another window and “paste” it into the data table (see the “Paste data” option in the “Data” menu). Users may still edit the data set after pasting.

Also under the “Data” menu on the main window is the option, “Get data,” in which users need only specify the URL of a WWW-accessible data file (see Figure 2). The applet will fetch the data and insert it into the data table, again allowing the user to make any changes afterwards. It is possible (though not required) to include a header with the data file that will specify the variable names and the number of observations per variable. See the WebStat documentation for more information on file headers.

It is possible to add the WWW address of a data file as a parameter to the applet tag within the HTML page containing the applet. This data set will then be loaded into the data table as WebStat starts up. This option should be useful in an educational setting when several students will each be accessing the same data set.



Figure 2: Loading the Iris data into WebStat by supplying the URL

Finally, some sample data sets, such as the famous Fisher iris data (Fisher, 1936) and the Old Faithful eruption time data (Weisburg, 1985), are available for users to use when familiarizing themselves with WebStat. These data sets are listed under the “Sample data sets” option under the data menu. Figure 3 displays the data table after the Fisher Iris data has been loaded.

Due to browser restrictions placed on applets in terms of reading and writing local files, loading a local file into WebStat is not possible for most browsers at this time. An anticipated move to a trusted applet framework in the future should make this task easier to perform. At this point, Hotjava is the only browser which will allow user’s to read and write local files.

Data Stat Graphics				
	sepal	sepalw	petal	petalw
1	5.1	3.5	1.4	0.2
2	4.9	3	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3	1.4	0.1
14	4.3	3	1.1	0.1
15	5.8	4	1.2	0.2
16	5.7	4.4	1.5	0.4
17	5.4	3.9	1.3	0.4
18	5.1	3.5	1.4	0.3
19	5.7	3.8	1.7	0.3
20	5.1	3.8	1.5	0.3

Unsigned Java Applet Window

Figure 3: The Iris data loaded into WebStat

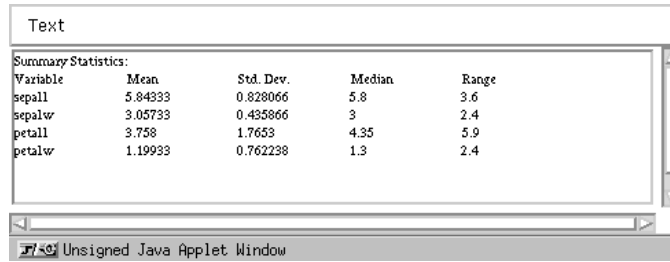
2.2 Numerical procedures

The second menu on the main window is the “Stat” menu, under which users may select any of the various numerical procedures available. These include most of the procedures that are commonly used in introductory statistical methods class. Specifically, they are:

- Summary statistics
- Z tests (one and two sample) for population means
- t tests (one and two sample) for population means
- A chi-square test for population variance
- An F test for comparing population variances
- Linear regression (parameter estimates and tests for significance)

Upon selecting one of these menu items, a dialog window will appear, allowing the user to specify which variable(s) are to be included in the analysis and requesting any additional information (such as a null-hypothesized value of a parameter). Clicking the “Okay” button will begin the procedure.

Results of these procedures are displayed in textual form in the Statistics Window. The resulting plain text can then be selected and “pasted” into another document if desired. Summary statistics for the Iris data appear in Figure 4.



Text				
Summary Statistics:				
Variable	Mean	Std. Dev.	Median	Range
sepal	5.84333	0.828066	5.8	3.6
sepalw	3.05733	0.435866	3	2.4
petal	3.758	1.7653	4.35	5.9
petalw	1.19933	0.762238	1.3	2.4

Figure 4: Example summary statistics for the Iris data

2.3 Graphical procedures

WebStat is also equipped with a wide array of graphical capabilities, which appear under the Graphics menu on the main window. Again, the graphical data analysis procedures typically discussed in introductory classes are all included. Currently available are the following:

- Histograms
- Stem and Leaf plots
- Boxplots
- Dotplots
- Parallel Coordinate plots
- Means plots
- Scatter plots
- QQ plots
- Time Series Plots

Again, upon choosing a menu item, a dialog window appears so that the user can specify the variable(s) to be included in the plot. If more than one variable is selected, then the procedure will be applied to each variable, with the results plotted on the same scale. The resulting plot (with the exception of the stem-and-leaf plot, which goes to the Statistics window) is displayed in the Graphics window.

The Graphics window can be subdivided so that several plots can all be displayed. This feature is accessed by selecting the Multiple Plots option on the Graphics menu (which appears on the Graphics window) — see Figure 5.

WebStat allows the user some flexibility in customizing graphics. By placing the mouse over the graphic of interest and double clicking, the user will cause a dialog window to appear, which gives the option of changing the graphics title and adding (or removing) a bounding box. The changes are made to the graphic by clicking the “Okay” button.

One difference between the Graphics window and the Statistics window is that when a new graphic is created, the old graphics are replaced, while results of new numerical procedures are

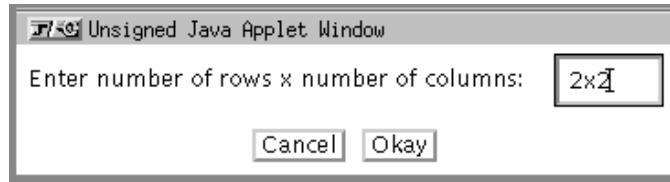


Figure 5: Using the Multiple Plots option

simply appended onto the results already in the Statistics window. Several plots for the Iris data, in a two-by-two array of plots, are displayed in Figure 6.

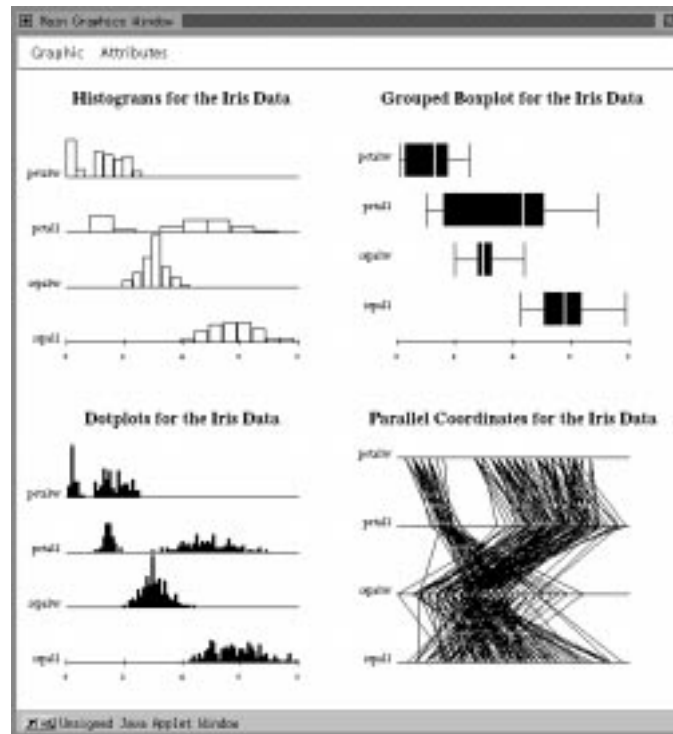


Figure 6: Some of the graphical capabilities of WebStat

2.4 Saving results

Due to the restrictions placed on applets by most browsers, saving statistical results in WebStat is not as easy as it should be. In the future, a move by the browser designers to a trusted applet framework should help matters considerably. As mentioned before, output in the Statistics window can be selected and pasted into another document. To save graphics one must often resort to screen dumps. The Hotjava browser allows graphics results to be saved directly to a GIF file by selecting the “Save Graphic” option under the “Graphic” menu.

3 Discussion

The current version of WebStat illustrates the power of Java-based software as a new approach to statistical network computing. While at present only basic analysis procedures are available, the authors are expanding WebStat to include more sophisticated procedures. The authors are working on providing procedures for multiple regression, analysis of variance, time series analysis, basic nonparametric inference and graphical and numerical routines for quality control. Since WebStat is built around an established core set of Java components, adding these additional routines requires only a relatively small amount of programming time. It is important to note that the inclusion of large number of analysis options will not significantly affect the download time for the WebStat applet. This is because the individual analysis procedures are downloaded the first time the user requests the procedure.

The WebStat package in its current form gets around many of the problems that often arise when introducing computer applications to an introductory statistics class. Aside from some browser differences, there are no platform differences or problems that could arise when students are using different versions of the software. It is anticipated that browser implementations will become more standard in future versions, so that the minor browser differences will be eliminated. Furthermore, when students move on to jobs or graduate school, they can still access the same software package that they used in their classes.

The authors are also working to develop a complete set of Java classes for building a graphical user interface and performing basic numerical and graphical operations. With this set of components, it will be possible to construct customized statistical applications which can be tailored according to individual preferences. Interested readers should visit the WebStat homepage periodically for information on when these classes will be publicly available.

References

- Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, Vol. 91, pp. 179–188.
- Weisburg, S. (1985), *Applied Linear Regression, Second Edition*, Wiley.
- West, R. W., Ogden, R. T., and Rossini, A. J. (1995), "The GASP Initiative: Interactive statistics via the World Wide Web, University of South Carolina, Department of Statistics, Technical Report No. 187.