A Web-based Statistical Analysis Framework

By

David Leis Chodos

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

David Chodos

Abstract

Statistical software packages have been used for decades to perform statistical analyses. Recently, the emergence of the Internet has expanded the potential for these packages. However, none of the existing packages have fully realized the collaborative potential of the Internet. This medium, which is beginning to gain acceptance as a software development platform, allows people who might otherwise be separated by organizational or geographic barriers to come together and tackle complex issues using commonly available data sets, analysis tools and communications tools. Interestingly, there has been little work towards solving this problem in a generally applicable way. Rather, systems in this area have tended to focus on particular data sets, industries, or user groups.

The Web-based statistical analysis modeldescribed in this thesis fills this gap. It includes a statistical analysis engine, data set management tools, an analysis storage framework and a communication component to facilitate information dissemination. Furthermore, its focus on enabling users with little statistical training to perform basic data analysis means that users of all skill levels will be able to take advantage of its capabilities.

The value of the system is shown both through a rigorous analysis of the system's structure and through a detailed case study conducted with the tobacco control community.

Acknowledgements

I am happy to thank the many people who helped me in my research. First, I would like to thank my supervisors, Don Cowan and Daniel Berry, and the members of my thesis committee, Dominic Covvey, Steve Manske, and Anne Banks Pidduck, for their thoughtful feedback, editorial suggestions, and guidance. I would also like to thank all of the interview participants for their valuable input, patience, and willingness to help. I would like to thank the statistics professors I consulted, Mary Thompson and Steven Brown, for their expertise and assistance. I am grateful for the help that the researchers working in the Computer Systems Group, Doug Mulholland, Kyle Young, and Anthony Robins, provided with the technical details of the WIDE system. I would like to thank the people working on the SHAPES project, Heather McGrath, Janice Tiessen, who helped with accessing a SHAPES data set for use in the WebStats prototype. As well, I appreciate the advice and assistance provided by my fellow students, Rosanna Morales, William Malyk, Eduardo Barrenechea, and Marcilio Mendonca. Finally, thanks to my friends and family for all of their help and support.

Financial support for the research was provided by the Strategic Training Program for Tobacco Research, which is conducted by the Canadian Institute for Health Research.

Table of Contents

Table of Figures

Chapter 1

Introduction

1.1 Statistical Software

For decades, computers have been employed to manage large sets of data and perform statistical analyses on these data using what are commonly referred to as statistical packages. These packages have matured over the years, offering users better interfaces and an ever-increasing array of statistical operations, visualization techniques, and auxiliary analysis tools. The emergence of the Internet as a software development environment represents a potential leap forward, as it allows data sets and analyses to be shared and colleagues to communicate with each other quickly and conveniently. Indeed, over the past decade, the major statistics package providers have introduced Web-based components and add-ons, and professionals from a variety of groups ranging from epidemiologists to geologists have created their own Web-based data analysis tools.

1.2 Weaknesses of Current Systems

Despite recent advances in statistical software, there are still several key areas for improvement. First, most statistical packages – both Web-based and conventional – are based on the assumption that they will be used within a single organization or field of study, which severely limits their utility in solving complex problems addressed by multiple organizations. In commercial packages this assumption is understandable, as data sharing among competing corporations is not expected. In an academic context, this is usually because tools are created for use within specific research areas. Second, the sorts of tools that might be brought together to provide a collaborative data analysis environment are spread out across a wide range of sources. Some are offered as commercial data analysis tools and others are offered as freeware collaboration tools, and still others are offered as experimental, research-oriented tools. Moreover, there is a marked lack of integration of analysis and communication tools. Third, many Web-based tools are created for use with a pre-determined collection of data sets, obviously limiting the analysis capabilities of the tools to those data sets. Fourth, despite the fact that graphical user interfaces (GUIs) have been in use for over two decades, many of the most widely used statistical packages, such as SAS, R and, to a lesser extent, SPSS, still use command-based interfaces, which makes them very difficult to learn for many computer users. Finally, many packages assume that the user has a certain amount of statistical training. This alienates many a potential user who would like to be able to perform simple analyses on data sets, but is overwhelmed by a package's wealth of options and unhelpful interface.

1.3 Thesis Statement

There is a need for a Web-based statistical toolkit which is usable across organizations, can integrate new data sets, can be learned quickly by novice users, and can enable collaboration. The system developed in this thesis, called WebStats, and the model upon which it is based, fulfill these requirements. This is demonstrated through a case study conducted within the field of tobacco control and via a detailed analysis of the system.

1.4 Overview of Method

1.4.1 Development of Model

First, by conducting a review of existing statistical packages, examining current academic literature and conducting informal interviews with selected tobacco control stakeholders, the nature of the user population, its relevant uses of statistical processing systems, and a set of basic requirements was established for a Web-based statistical toolkit. Next, based on these requirements, a responsive system model was developed that was congruent with the user-uses model, met these initial requirements and addressed the issues outlined previously. A prototype system was then created based on this system model.

1.4.2 Requirements Engineering

This prototype was then used in a series of deeper requirements-gathering interviews with potential stakeholders in order to generate a more precise, context-specific set of requirements that reflected the needs of a specific group of users – in this case, those working in the area of tobacco control. The interviewed stakeholders represented a variety of roles within the tobacco control community. Academic researchers, public-health nurses, professionals working for non-profit groups, and civil servants were interviewed from across Canada in order to get a comprehensive understanding of the requirements for WebStats within the context of tobacco control.

These requirements were used to guide the development of the alpha version of WebStats and, through this process, a beta version of the system was developed. The participants in the initial requirements-gathering interviews were then asked to provide feedback on the improved version of the prototype via a second round of interviews. The responses from these interviews were used to refine WebStats further, with the result being a more usable and robust final version of the system.

1.4.3 Assessment

The final stage in the development process was the assessment of WebStats. To this end, an assessment framework was defined and the stakeholders were asked to comment on the final version of WebStats, thus providing real-world assessment of the system's strengths and weaknesses. As well, WebStats was assessed using the same criteria that were used in analyzing existing statistical packages, thereby providing a means of comparing the system with other statistical systems described in this thesis.

1.5 Contributions

1.5.1 Model for Web-based Statistical Computation

The Web-based statistical computation model developed in this thesis is useful because it provides a framework for the creation of collaborative Web-based statistical applications. The model includes a statistical calculation engine, a graphical presentation engine, input and output components, a communications component, a statistical advisor, and a framework for storing data sets, information about the data sets, analyses, and information about these analyses. The model's generality is particularly important for researchers or Web developers, as it allows them to use the structural aspects of the model regardless of their particular development environment.

The WebStats model includes several components – such as a statistical advice tool, an analysis storage framework, and a communication component – that increase usability and facilitate collaboration across organizations and work environments. Developing a system guided by our model has resulted in a system that supports statistical analysis by professionals with varying levels of statistical training, and sharing of analysis results with colleagues in other organizations. Thus, it allows a broad population of users to perform basic analyses, and allows those analyses to be disseminated to a similarly broad population of recipients.

The extensible nature of the WebStats model, that is, its ability to incorporate additional components, also adds to the model's significance. The model's extensibility means that it can be used as a starting point or basis on which further statistical, communications, or data management tools may be developed, perhaps to meet the needs of different user groups or environments. Armed with a basic understanding of the components of the model and the interactions between the components, software developers will be able to add custom components to the system quickly and easily.

1.5.2 Results of Requirements Analysis

The requirements analysis, which was performed in order to determine the stakeholders' needs and thus shape the development of the system, is useful on several levels. First, by taking a systematic look at the way statistical analysis is performed in the field of tobacco control, the requirements-analysis process identified important areas for improvement.

Second, in conducting this process with the goal of creating a piece of software, solutions were identified to the problems mentioned in the previous paragraph. Some of these solutions were technically oriented, but others involved identifying and mitigating organizational or cultural factors. On the whole the requirements analysis took a positive approach to the issues that were discovered and sought to go beyond identifying the problems to mitigating and, ultimately, solving them.

Lastly, it should be noted that the field of tobacco control has a lot in common both with other healthcare contexts – with regards to the related organizations and their place within the Canadian healthcare system – and with other work environments that are dispersed across multiple geographic locations or organizations. Thus, the lessons learned through the requirements analysis process are applicable to a wide range of other environments. For example, many of the issues in tobacco control related to shaping youth behaviour and creating effective youth-centred population-based intervention are also faced by groups dealing with issues like physical activity, nutrition and alcohol use.

1.5.3 Literature Review

The review of existing statistical software packages – with an emphasis on usability and Web-based systems – makes several contributions. First, it establishes a set of criteria against which statistical software may be evaluated. These criteria include factors such as mathematical accuracy, usability concerns, and licensing fees. Second, it provides a survey of the current state of statistical software across a wide range of disciplines, platforms and contexts. The review includes both commercial and open-source applications and both Web-based and conventional systems, and it investigates niche user groups such as scientists analyzing gene sequences.

1.5.4 Requirements Engineering Experience

The process of eliciting initial requirements from a stakeholder population, designing a system based on those requirements, refining the design and implementation through further interviews and system walk-throughs, and finally presenting the finished product for evaluation provides further validation of the requirements-engineering-based approach to software development. Specifically, the application of requirements-engineering methods to the context of developing statistical software for tobacco control professionals helps inform future requirements-engineering research, as this is a new context with unique characteristics.

1.6 Outline of Thesis

Chapter 2 gives an overview of the statistics software that exists currently, analyzing various systems with respect to the characteristics outlined at the beginning of this section. The chapter looks at spreadsheet and statistical analysis systems, software distributed using both open-source and commercial business models, and applications which run on both Web-based and conventional platforms. The chapter includes also a review of the academic literature related to Web-based statistics, including areas such as Web-based system architecture, statistical accuracy and survey analysis guidelines.

Chapter 3 gives a description of the Web-based statistical analysis model. The overall architecture is presented, the model's major components are described, and some of the details of the model's implementation within the WIDE framework are discussed. The design is briefly evaluated in terms of the characteristics set out at the beginning of the thesis.

Chapter 4 describes the requirements-elicitation process used. This process consists of the formal validation of the model through technical analysis, user surveys, evaluation of WebStats's use in real-life situations, and interviews with potential users.

Chapters 5 and 6 are concerned with the case study, tobacco control, that was used to validate the model. Chapter 5 begins by describing the goals of the tobacco control community, and the principles behind the programs being implemented to achieve these goals. Next, the importance of program evaluation in promoting effective tobacco control programs is discussed, establishing the importance of effective data analysis tools for this context. This is followed by an analysis of the current tools and techniques being used to perform program evaluation within the tobacco control field. Chapter 6 describes the system requirements that were specified by professionals from the tobacco control community. This chapter describes in detail also the features of WebStats that relate to the toolkit's suitability for the context of tobacco control. The case study finishes with a summary of the stakeholders' assessment of the final product.

Finally, the thesis explores several areas for future work and draws conclusions about the project as a whole.

1.7 Motivation

The need for information is universal; individuals and organizations use it to form opinions and make decisions and, assuming that they have the means to interpret it properly, the more information individuals have at their disposal the more informed, and hopefully wiser, their decisions will be. At first, the information-sharing potential of the Internet was seen as a panacea for these issues, and it was predicted that giving people access to more information would cause their level of knowledge and their decision-making capabilities to improve dramatically. However, as the popularity of the Internet has grown over the past decade, it has become clear that just enabling access to information is not sufficient, for several reasons. First, information overload has emerged; people are faced with an overwhelming amount of information: too many search results, an overflowing email inbox, and so on. Simply giving people *more* information will not make them any better informed. Second, the deluge of mostly irrelevant information has, perversely, made it harder to find relevant, high-quality information, much of which is tucked away in relatively inaccessible academic or private repositories. Finally, an important distinction must be made between data – numbers and text with little or no associated context – and information, which is data along with the context required to interpret the data. For example, a table of numbers may be considered data, while that table, along with the meaning of the rows and columns, the situation in which the data were collected, and the method with which the numbers were compiled, would be considered information. Thus, in order to realize the potential of the Internet, there must be an improvement not just in people's access to data but, more importantly, in people's access to information that can actually be used to assist evaluation and decision-making.

What is needed, then, is a way to take the vast amount of information that is available and allow users to collect related data in a single place, identify relevant information, and then share that information with interested colleagues. The Internet has the potential to foster communication between disparate groups, improve access to information, and – via Web-based software – give users the ability to manage and disseminate that information. Indeed, progress has been made in various areas towards these goals.

However, that some of these issues have been solved separately does not mean that the problem as a whole has been solved. First, many groups that have a strong need for the sort of data analysis and information management tool described above cannot afford the licensing fees for commercial statistical analysis packages. Second, these packages are intended to be used within an organization, and are not suitable for tasks which are spread across organizations. Third, the sorts of tools that have been created to solve parts of the problem are spread out across a wide range of sources. This thesis presents a model for creating this kind of tool, and shows the utility that such a tool can have for a particular user group.

Chapter 2

Literature Review

One of the first questions that must be asked when developing a new system is: "What else is out there?" Before putting forth the effort to develop a Web-based statistical toolkit, it is certainly worthwhile to investigate the statistical tools that are currently available. Given the depth and maturity of the statistical software field, it is particularly important to ensure that the particular collection of features offered by WebStats are not already available in another statistical software system. An issue that must be addressed before this question may be answered, however, is establishing the criteria by which statistical software systems will be evaluated. A few criteria – flexibility, ease of use, accessibility, collaboration potential – were set out at the beginning of the thesis. The reasons for choosing these criteria will be addressed, with reference to the experience gained through decades of statistical computing.

This section analyzes a broad range of statistical software packages according to these criteria in order to determine which functionality already exists, and how accessible and appropriate the software is for users with varying levels of technical expertise, statistical training, and limited budgets. It is impractical to analyze every one of the hundreds of statistical packages currently in use; thus, a representative sample has been chosen which encompasses three main types of software – commercial software, research-based software, and open-source software – and several target user groups – business users, academic users, and statistics students. This section will also review the current academic literature to assess the state of research in Web-based statistical systems. This research, which ranges in scope from data warehousing to statistics education, offers a wide variety of ideas that will prove useful in developing WebStats.

2.1 Determining Evaluation Criteria

The issue of evaluating statistical software is not a new one. More than thirty years ago, Slysz performed a comparative analysis of the statistical software packages that were available for use in the social sciences [Slysz, 1974]. While many of these programs, such as DATA-TEXT and TSAR, are no longer in use, some of the criteria that were examined, such as efficiency on data sets of various sizes, output quality, and interface ease-of-use, are quite relevant today. Thirteen years later, Stutz conducted a similar survey of microcomputer-based statistical software that was available to the Instruction and Research Computer Center (IRCC) at Ohio State University [Stutz, 1987]. This survey, which was conducted within the context of the needs of the IRCC, identified additional criteria such as the cost of the software and the details of the software's licensing agreement, the needs and statistical training of the software's users, the system resources – memory, disk types, monitor – needed to run the software, and the quality of the documentation provided with the software.

McCaskell *et al* undertook a similar effort at the University of Guelph in 1989. One of the historical problems identified in this analysis was that the user needed to understand "the 'mechanics' of [mainframe] computing" before they could perform "even the simplest statistical procedure" [McCaskell, 1989]. The details of the technological barrier have changed – in the analysis by McCaskell, users were intimidated by keypunches and understanding resource scheduling algorithms, while today's users may well be stymied by driver conflicts and insufficient virtual memory. The issue of users dealing with technological barriers, however, is as relevant today as it was two decades ago. Some other issues identified by McCaskell were the difficulties inherent in supporting a growing population of statistical software users with "widely varying computer skills" and the "significant outgoing cash flow for the purchase of individual copies of software" [McCaskell, 1989]. McCaskell sought to combat these issues by providing more flexible consulting services and purchasing software on a site-license basis rather than an individual copy basis as part of a five-point strategy. However, as the following analysis will show, these technological and financial issues are still being confronted by users and administrators today.

More recently, Suchan undertook a usability study of the geovisualization software being used in government agencies, academia and business, analyzing organizations such as the Census Bureau, Rice University and AT&T [Suchan, 2002]. Among Suchan's findings were the importance of including users in the software design process, the desire on the part of potential users of a software system to be able to use their own data to evaluate a system rather than default or canned data, and the importance of considering confidentiality issues when managing, analyzing, and disseminating data [Suchan, 2002].

2.2 Statistical Software Packages

Excel

Excel, which is the spreadsheet component of Microsoft Office suite, is a widely adopted spreadsheet program. One researcher wrote that "it is quite possible that more basic statistical calculations are done worldwide in Excel than in all statistical packages combined" [Wilkinson, 1994], which gives an indication of Microsoft Office's hegemony in the work environment. Excel, like other spreadsheet programs such as OpenOffice's chart program or Corel's Quattro Pro, has a number of features that make it very useful for a wide range of applications. However, it also has some drawbacks, described below, that make it less than ideal as a statistical application.

In terms of ease of use, Excel has its pros and cons. Its Chart Wizard allows users, by navigating through a series of dialog boxes, to create a wide variety of pie charts, bar charts and line graphs – several dozen different types, in all. Users can see a preview of a chart before they create it, and can edit it afterwards through a point-and-click interface. The charts that users create can then be pasted into any other Office program, such as Microsoft Word, through the Clipboard.

Excel also has a robust statistical calculation package that allows the user to do everything from simple descriptive calculations, such as calculating the mean of a set of numbers, to advanced inferential statistics, like the Chi-squared test [Donnelly, 2004]. The main drawback of Excel's approach to statistical calculation, though, is that it relies on the user remembering the function names and syntax required for each statistical operation. In this way, it is much more like using a programming language than an end-user friendly application, and thus may be quite intimidating to less computer-literate users.

Another major drawback of Excel in terms of its support for collaboration is that it has virtually no capacity for integration with a Web-based framework, a database management system (DBMS), or any other application outside of the Microsoft Office suite. Any data to be passed into Excel must be exported from the data source as a delimited text file (e.g., in comma-separated value format) and then imported into a local copy of Excel, and may be manipulated only after it is in a worksheet within Excel. Thus, anyone wishing to use Excel must have a copy of the program installed on his computer, which may be too expensive. Since Microsoft licenses its software on a per-computer basis, installing Excel on every computer in a public health unit, for example, may be prohibitively expensive. The suggested retail price for a single copy of Excel as of December 2006 was $309, which gives a sense of the expense involved in using the software on a large scale, although it should be noted that substantial discounts are available for educational or government organizations [Microsoft Canada, 2006].

Another drawback of Excel is that a spreadsheet may contain a maximum of 65,536 rows [Hertel, 2004]. When working with very large data sets, this row limit can be critical.

Finally, there is no capacity for data management on any scale beyond a collection of worksheets. As mentioned earlier, one must import data into Excel from an outside source in order to use it, and these data are then accessible only within the spreadsheet. There is no ability to compare data across spreadsheets, for example, or to take the average of a particular value from several spreadsheets and combine it into a single graph or measurement.

In summary, Excel is very useful for local, small-scale data management and chart creation, and its statistical operations make it a powerful analysis tool for advanced computer users. However, its high cost, minimal data management abilities and lack of Web accessibility leave much to be desired from a large-scale, collaborative statistical analysis point of view.

SAS

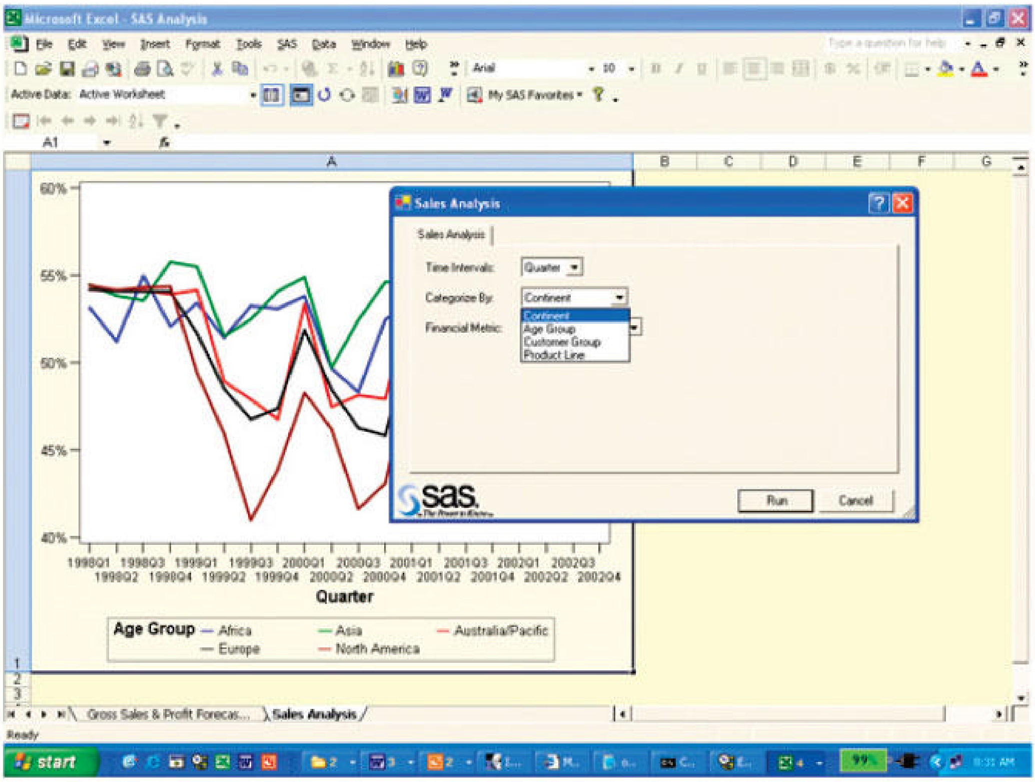
SAS is a command-based statistics package that is commonly used by statisticians and academics for performing a wide range of statistical calculations. Its strengths are the broad range of statistical operations available, the flexibility it gives to advanced users, and the extensions available through various add-on packages. It is, however, not an easy application to learn, especially for users who are more comfortable with a point-and-click interface. As well, licensing costs may make it unaffordable for those without significant government, business, or academic funding.

Statistical operations in SAS are executed via series of SAS commands, referred to as programs, entered by the user. These programs are divided into two major sections: a data section, where the data for the program, including the definition of variables to be used by the program, are loaded and defined and a procedure section, which contains the actions to be performed on the data loaded into the program. These actions range in complexity from sorting on a particular column to performing complex statistical calculations. One must know the syntax for each procedure; the number and type of its arguments, the expected return type, and so forth. The output, by default, is text-based. Graphs are generated using asterisks, dashes, and vertical bars, and contain a fair bit of plain-text information. There are, however, additional packages that can be added to SAS to allow the user to create high-quality graphs, among a multitude of other functions.

From a power and flexibility point of view, a command-based interface is very appealing. It gives users a great deal of control, allowing them to assemble commands, subroutines and the like to perform very complex queries. Indeed, for this very reason SAS is commonly used by epidemiologists, biostatisticians and others who have a strong background in statistics and a high level of computer literacy [Leatherdale, 2005; Seliske, 2005]. The combination of a powerful, versatile analysis system and a robust add-on graphing package means that SAS has the potential to create very detailed, highly customized graphs.

However, this power and flexibility comes with a cost: a user interface that is quite difficult to learn. While a command-based interface may be fine for users with a high comfort level with computers and, in particular, users who started out using command-line operating systems like DOS or Unix, it is quite challenging for less computer-literate users [Thompson, 2006]. In requirements analysis interviews, users expressed frustration in needing to memorize commands and syntax, adhere to the program's "picky" formatting rules, and compared using SAS to learning a foreign language [Pathammavong, 2006; Silverman, 2006; Taylor, 2006]. Overall, these are not challenges that one wants to deal with, and they are especially frustrating to users with little familiarity with computers.

SAS offers an Add-In for Microsoft Office that "enables users to transparently leverage the power of SAS data access… from Microsoft Office through integrated menus and toolbars." [Hertel, 2004]. By using a mechanism referred to as "stored processes", the user can execute SAS programs from within Excel, thus achieving the statistical power of SAS with a much more familiar and accessible user interface. As well, the user can make use of sixty common analyses, referred to as "Tasks", in order to analyze data quickly and easily via SAS from within Word or Excel. An added benefit of using the Add-In is that it gets around Excel's 65,536 row limitation, mentioned in the previous section [Hertel, 2004]. See Figure 1 for a screenshot of the SAS Add-In.


Microsoft Excel - SAS Analysis
File Edit View Insert Format Tools SAS Data Window Help
Active Data: Active Worksheet
My SAS Favorites
A1
A B C D E F G
60%
55%
50%
45%
40%
1998Q1 1998Q3 1999Q1 1999Q3 2000Q1 2000Q3 2001Q1 2001Q3 2002Q1 2002Q3
1998Q2 1998Q4 1999Q2 1999Q4 2000Q2 2000Q4 2001Q2 2001Q4 2002Q2 2002Q4
Quarter
Age Group
Africa
Asia
Australia/Pacific
Europe
North America
Sales Analysis
Sales Analysis
Continent
Age Group
Customer Group
Product Line
Run
Cancel
Gross Sales & Profit Forecas...
Sales Analysis
Ready
start

Time Intervals: Quarter

Categorize By: Continent

Financial Metric:

Figure 1: SAS Add-In for Microsoft Office

Like Microsoft's suite of programs, SAS is commercial software and, as such, entails licensing fees for its use. For example, Executive Information Systems (a company that acts as a reseller for SAS for the U.S. government) charges $5,895 (USD) per workstation for its "Base SAS" PC bundle [Executive Information Systems, 2006]. Thus, only organizations with a large enough budget are able to run SAS. Of course, using the Microsoft Office Add-in package means that the organization must pay two licensing fees, one for SAS and another for Microsoft Office, to combine the analysis power of SAS with the document creation abilities of Microsoft Office.

Finally, SAS does not have any data-sharing functionality integrated into the software, making it difficult to share analysis results using SAS. Two ways of getting around this are embedding SAS results in a word processing document, such as a Word document or Excel spreadsheet, or providing the recipient with the entire SAS dataset or program. Embedding SAS results has the advantage of a wide potential audience, since many more people have access to word processing software than to SAS. However, one loses the ability to do further statistical analysis on the data. Sending a SAS dataset allows the recipient to see exactly what was done in analyzing the data and then continue the analysis on their own if they so choose; however, it requires the recipient to have a compatible version of SAS, and have the necessary expertise in using the program [Thompson, 2006].

R

R is a freely available language and environment for statistical computing and graphics that provides a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, statistical tests, time series analysis, classification and clustering [Comprehensive R Archive Network, 2006].

Like SAS, R uses a command-based interface; R is built on a "well-developed, simple and effective programming language" called, unsurprisingly, S. Thus, for users with a strong programming background, it is quite easy to pick up the syntax and commands of this language and quickly use R to perform tasks such as managing data, performing analyses and creating graphs. In addition, the programming-based nature of the system means that advanced users can even create their own statistical models, graph types and data structures with relative ease [Venables, 2006]. For people without this background, however, learning R is quite daunting. To perform simple tasks such as loading data or creating graphs one must learn the proper commands, syntax and parameters [Thompson, 2006].

One of the advantages of R, as mentioned earlier, is that it is an open-source project, and thus does not pose the same financial hurdles as its commercial counterparts. Specifically, it is licensed under the GNU General Public License (GPL), which permits the software to be used without charge [GNU Project, 2006]. However, this also means that some of the benefits of commercial software – a help desk, for example, or official documentation – are not available. R does, however, provide help from within the program and an online manual, in addition to a newsgroup and an extensive list of frequently asked questions (FAQ) [Hornik, 2006].

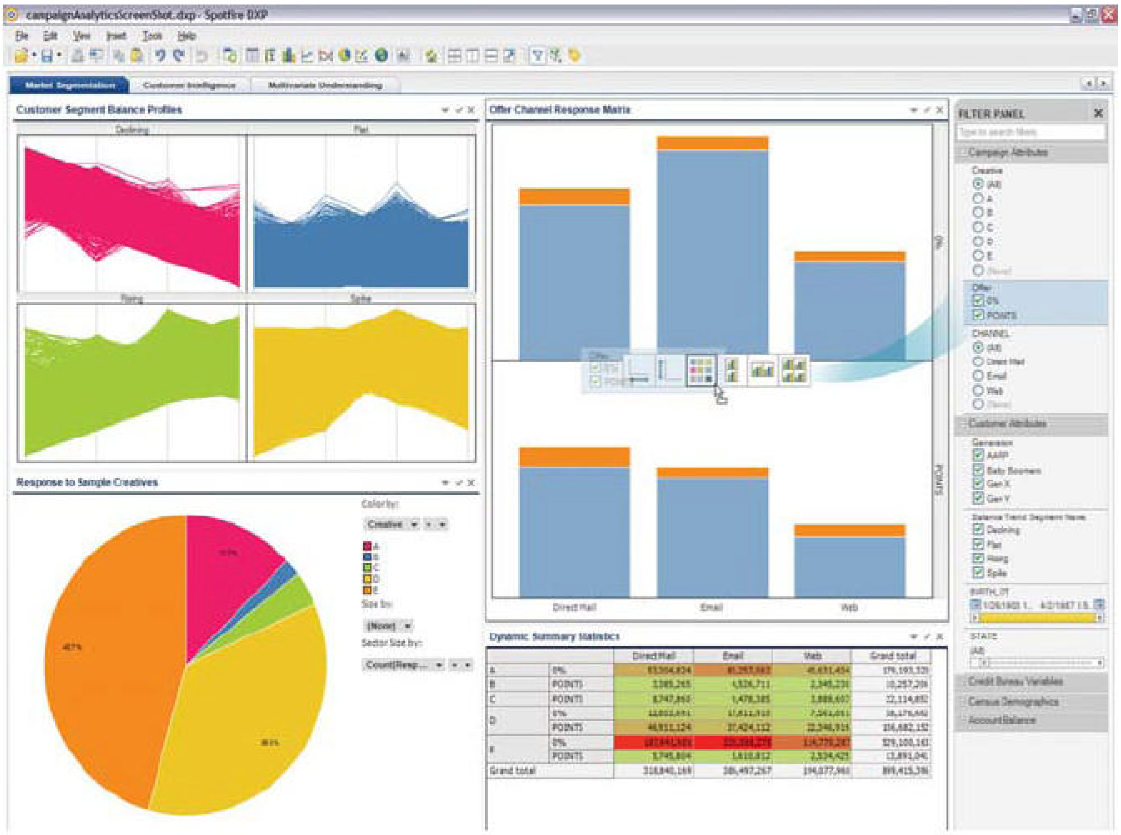
SPSS

SPSS offers a similarly broad range of statistical operations as SAS and R, but uses a comparatively user-friendly point-and-click interface. This removes the burden from users of learning a set of commands, keywords and parameters, and instead asks them to associate statistical concepts such as frequencies, central tendencies and standard deviations with buttons and icons. While this is certainly an improvement, from a user interface perspective, it is not ideal. There are still certain operations that can only be performed through a scripting language, similar to those used by R or SAS. As well, given the abstract nature of some of the ideas being conveyed graphically, some of the icons function more as mnemonics than instructions. That is, one might need to remember that a certain icon is intended to represent a standard deviation calculation, for example, rather than being able to look at that icon and understand its meaning. On the whole, however, the learning curve is much less steep for SPSS than it is for SAS and other command-based packages [Thompson, 2006].

Another drawback to SPSS is that, like Excel and SAS, it is commercial software, and charges users licensing fees. Thus, it may well be prohibitively expensive for small businesses, government agencies and, particularly, non-governmental organizations (NGOs). According to the Commercial Pricing page on the SPSS site, purchasing SPSS Base 15.0 for Windows costs US $1,599, while the Survey Research Analyst Bundle is priced at US $7,452 [SPSS website].

Spotfire

Spotfire is an information visualization tool that grew out of research by Ben Schneiderman *et al* at the University of Maryland. It is now a well-established piece of commercial software known as Spotfire DXP, whose customers include industry leaders among the Global 2000, and boasting powerful features such as dynamic filtering and zooming, a guided analysis tool, role-based configuration options, and a commonly accessible library which enables cross-functional discussions [Spotfire, 2006]. See Figure 2 for a screenshot of Spotfire. However, it is marketed towards large corporations, competing with other enterprise analysis tools like Viz, Cognos and Visual Analytics, and is an expensive piece of software; a license costs approximately US $1,000 per person, per year [Schneiderman, 2006]. Thus, for both cost and application scope reasons, it may not be appropriate for not-for-profit organizations, or for projects spanning multiple organizations or organizational contexts.



Dynamic Summary Statistics

		Direct Mail	Email	Web	Grand total
A	0%	53,304,824	81,257,882	45,631,454	179,193,320
B	POINTS	3,385,265	4,526,711	2,345,230	10,257,206
C	POINTS	8,747,860	1,478,385	3,888,607	22,114,852
D	0%	14,002,651	17,844,510	7,264,051	38,479,446
	POINTS	46,911,124	37,424,112	22,346,915	106,682,152
E	0%	182,644,[illegible]	221,358,278	114,779,287	529,100,161
	POINTS	3,745,804	1,810,812	2,324,425	13,891,041
Grand total		318,840,168	385,497,267	194,077,961	899,415,396

Figure 2: Spotfire DXP Screenshot

2.3 Web-based Statistical Tools

Analysis Tools -- StatCrunch

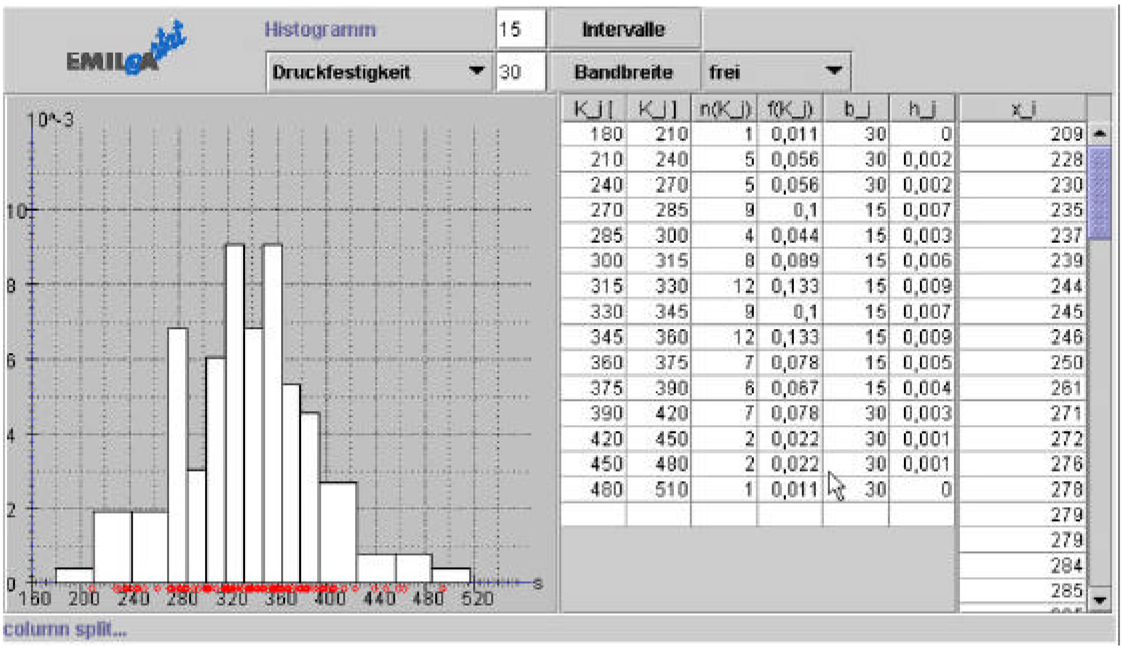
One well-established Web-based statistical analysis tool is StatCrunch. It allows the user to import and manipulate data, and create graphs via an intuitive Web-based interface. Thus, the tool is well suited for quick, easy analyses and graph generation [West, 2004]. Specifically, data may be imported from several different kinds of sources – Excel or text files, a URL, or pasted from the clipboard – and then viewed or modified via a data management component, which also supports statistical concepts such as binning, sampling, and the generation of simulated data conforming to a particular distribution. The graphing component of StatCrunch offers a wide variety of graph types – from standard bar charts and pie graphs to more advanced graphs such as 3-D rotating graphs – as well as customization options such as the ability to create a custom colour scheme [StatCrunch, 2006]. It has been criticized, however, for its lack of precision when working with large, complex data sets [Kitchen, 2003]. As well, it lacks integrated communication tools, which limits its usefulness as a collaboration or data-sharing tool.

Survey Tools

Computer-Aided Survey Information Collection (CASIC) tools have been used by researchers and data collectors for years to help conduct and manage survey data. As the Internet emerged as a viable means of exchanging information, researchers began exploring its potential to improve CASIC tools. In one case, researchers at the U.S. Department of Agriculture, in conjunction with the Iowa State University Statistical Laboratory, used Web-based CASIC tools to distribute training information to data collectors, help data managers oversee the data collection process and assist the data editors in ensuring that the data gathered was valid and accurate [Nusser, 1998]. Although the project included a utility that generated summary reports (in PDF format) on the progress of the data collection effort, there was no capability for a user to generate his own reports containing only the information he desires, which severely limits the tool's utility as a general-purpose analysis tool.

Statistical Education Systems

Another area in which researchers have been exploring the potential of Web-based statistical analysis tools is statistical education. Several analytical tools have emerged over the past few years, including SurfStat, StatSoft, and HyperStat [Dinov, 2006]. E-stat, another educational tool, is "a multimedia, Web-based, and interactive learning and teaching environment in applied statistics" developed at the University of Oldenburg [Cramer, 2002]. While, as a teaching tool, it does not offer the full range of analysis and dissemination capabilities offered by other systems, it does have several intriguing features. One is the ability to present material at one of three levels (elementary, basic or advanced), depending on the user's familiarity with statistics. As well, the system's designers acknowledge that "different types of users have different needs," and offer "user-specific views and scenarios" in order to ensure that the system is usable – and useful – for as wide an audience as possible [Cramer, 2002]. Another important feature area is the system's descriptive statistics visualization capabilities. The system uses Java applets to provide interactive histograms, measures of the mean and median, and more advanced measures such as the Lorenz Curve, which is a graphical representation of the cumulative distribution function of a probability distribution, and is often used to represent income distribution. See Figure 3 for a screenshot of the histogram component. Note that the button labels are in German.



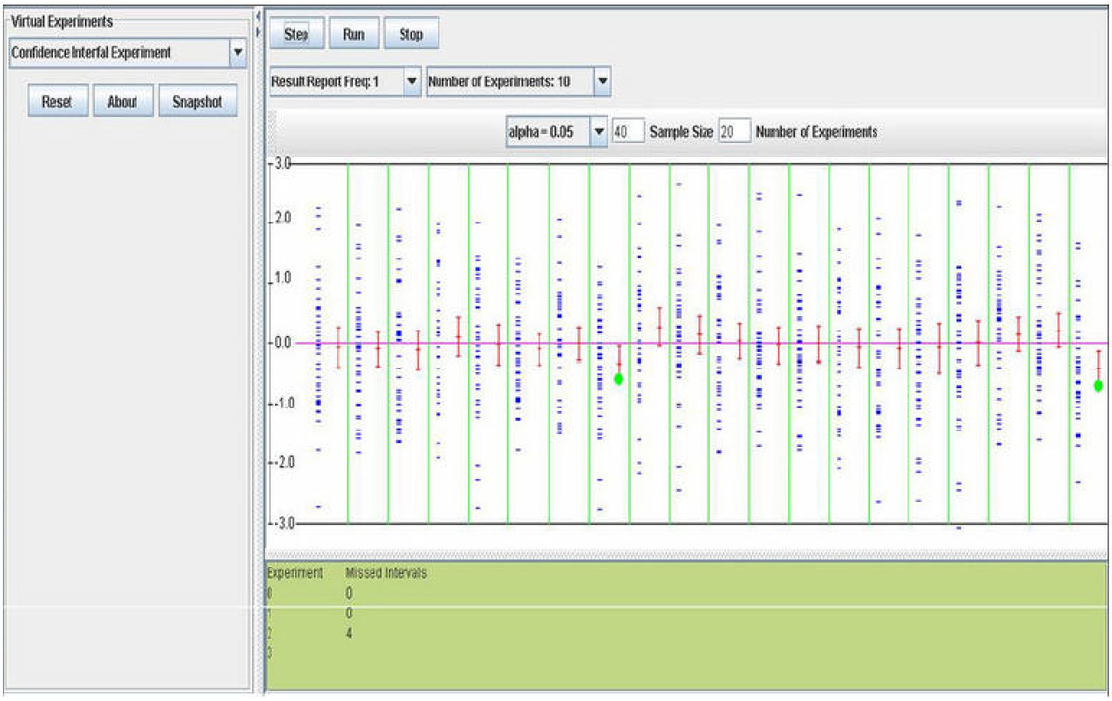
Histogramm 15 Intervalle

Druckfestigkeit 30 Bandbreite frei

K_j [	K_j]	n(K_j)	f(K_j)	b_j	h_j	x_i
180	210	1	0,011	30	0	209
210	240	5	0,056	30	0,002	228
240	270	5	0,056	30	0,002	230
270	285	9	0,1	15	0,007	235
285	300	4	0,044	15	0,003	237
300	315	8	0,089	15	0,006	239
315	330	12	0,133	15	0,009	244
330	345	9	0,1	15	0,007	245
345	360	12	0,133	15	0,009	246
360	375	7	0,078	15	0,005	250
375	390	6	0,067	15	0,004	261
390	420	7	0,078	30	0,003	271
420	450	2	0,022	30	0,001	272
450	480	2	0,022	30	0,001	276
480	510	1	0,011	30	0	278
						279
						279
						284
						285

Figure 3: E-stat Screenshot

Another Web-based statistical education package is SOCR, which was developed by Ivo Dinov at UCLA. This package consists of "a collection of Java applets useful for interactive learning and for motivation of various probability and statistics concepts" [Dinov, 2006]. Among these applets are a virtual experiment component, designed to "build intuition and confidence in understanding" expe riments which teach basic probability concepts, a confidence interval experiment component which illustrates the relationship between the abstract and procedural definitions of confidence intervals, and a suite of distribution modeling aids, which provide a "foundation for sampling/resampling demonstrations, hypothesis tests, statistical inference, model-fitting and critical value estimation" [Dinov, 2006]. See Figure 4 for a screenshot of the confidence interval experiment component.



Confidence Interfal Experiment

Result Report Freq: 1

Number of Experiments: 10

alpha = 0.05 40 Sample Size 20 Number of Experiments

Experiment	Missed Intervals
0	0
1	0
2	4
3	

Figure 4: SOCR Screenshot

While statistical education tools such as E-stat and SOCR have covered a lot of ground in terms of their Web-based implementation of statistical operations and presentation of graphs , their *raison d'etre* is pedagogical rather than analytical; they are aimed at a different group of users from general purpose analytical tools and seek to educate rather than to allow the user to perform meaningful analyses of data sets.

Specialized Analysis Tools

Some researchers have created Web-based statistical analysis tools to meet the needs of specific groups of users. Two such examples are BASE, a system that analyzes micro-array data, and B-Course, a tool for finding probabilistic dependencies between variables in a multivariate data set.

According to an overview of open-source micro-array analysis software, BASE is "a Web-accessible system that uses standard browsers to interact with a central microarray database and appropriate data tools" [Dudoit, 2003]. Once the data set has been analyzed, the system allows the results to be displayed "as scatterplots, displayed in histograms or viewed as tables" [Dudoit, 2003]. BASE also boasts a "well-designed [annotation] system that allows complex ancillary annotation", which would be quite useful for many contexts beyond the biological, DNA microarray analysis context described in Dudoit's paper. For users within the bioinformatics and computational biology communities, this certainly seems to be a useful, robust tool. However, its focus on this particular user group limits its usefulness for a wider audience.

B-Course is a “free Web-based online data analysis tool which allows users to analyze their data for multivariate probabilistic dependencies” [Myllymäki, 2002]. B-Course is quite accessible, not only in that it is delivered over the Internet, but also in its usability with most Web browsers, and it “requires no downloading or installing of software” [Myllymäki, 2002]. The system recognizes the issue of users who “are not experts in data analysis” trying to use statistics packages, relying on the system’s default parameters, and thus ending up with “conclusions derived from an analysis [which] are frequently far from the intended plausible reasoning” [Myllymäki, 2002]. The system attempts to solve this problem via a “tutorial-style” user interface, which “intertwines… data analysis with support material which gives an informal introduction” to the statistical concepts used in the system. Thus, the user is not only given the tools needed to analyze the data via a clear, well-designed user interface, but is also presented with the background knowledge required to use the tools properly. See Figure 5 for a screenshot of B-Course. Much of the work done on B-Course overlaps with the goals of WebStats with regards to its accessibility, usability, and guidance-based analysis process. However, B-Course is limited to analyzing data sets for pairwise conditional dependencies between variables, which is a small subset of the kinds of statistical analysis that may be performed. Indeed, the functional requirements elicited in the process of designing WebStats found that among novice users, there is a strong demand for the ability to perform descriptive statistics, with a much lesser demand for inferential statistical operations. Furthermore, even within that small area of analysis, B-Course considers only models for discrete data and considers only dependency models in which the list of dependencies can be represented in a graphical format using Bayesian network structures [Myllymäki, 2002]. However, this limitation pales in comparison to the system’s restricting the user to finding dependencies between variables, which effectively negates its usefulness as a general-purpose statistical system.

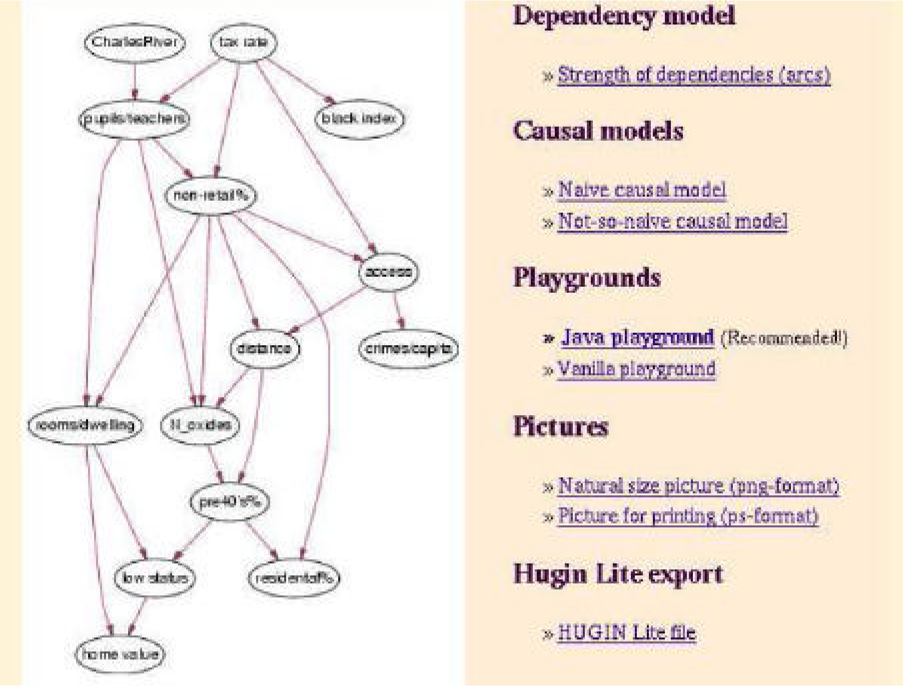


Figure 5: B-Course Screenshot

Web-based Interfaces for Statistics Packages

Another category of Web-based tools consists of Web-based versions of conventional software systems. Both SAS and SPSS have Web-based analysis tools that accompany their main statistical analysis packages. In addition, open-source software developers have created several Web-based extensions to R; among them are Rweb and Rpad. Rweb offers a command-line interface to R. Rpad, on the other hand, provides a graphical, point-and-click interface for many of the more common data management and analyses functions provided by R [Banfield, 2006].

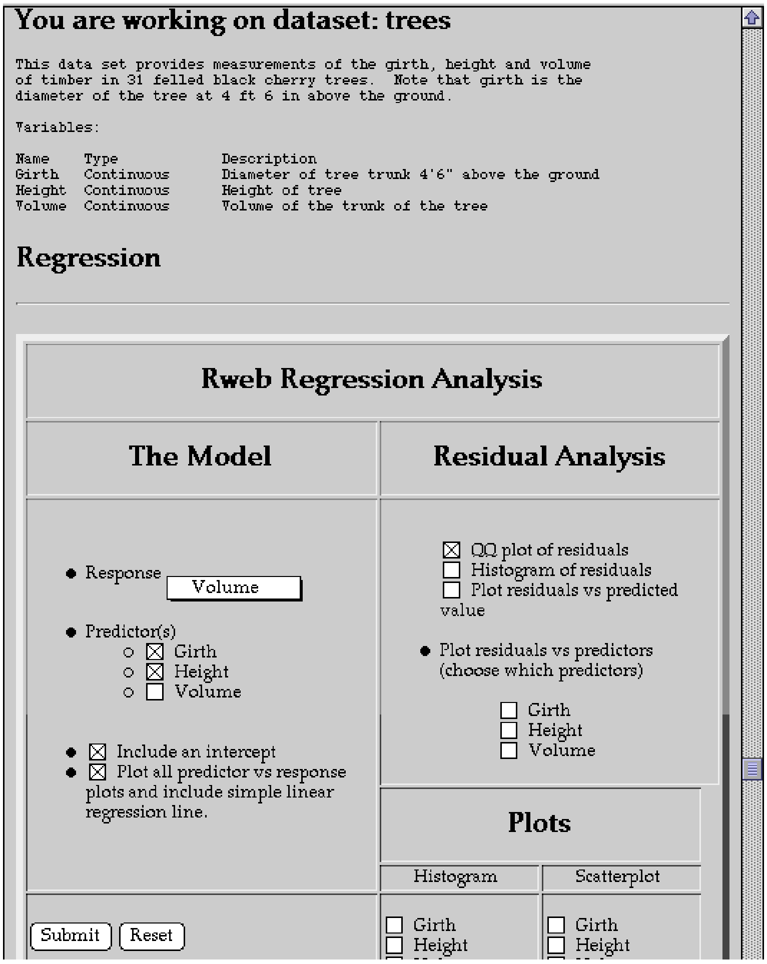
In addition to its main statistical analysis package and Office add-in, SAS has recently developed a Web-based query and reporting tool. It seeks to address a growing need in the business world for query, reporting and analysis tools which are usable outside of the IT department, and allow users at all organizational levels to be more self-sufficient in fulfilling their reporting needs [Fuchs, 2006]. This Web-based tool has a number of features, including scalability to thousands of users, a user-friendly interface, an ad hoc query engine and a wide variety of pre-formatted canned reports, flexibility in the kind of data source used, and report bursting (the mass distribution of reports). As well, the fact that it is Web-based means that avoids problems of "installing and maintaining desktop tools on thousands of computers", which is a serious IT burden [Fuchs, 2006]. As well, reports can be stored in commonly accessible areas, which allows them to be disseminated across an entire company.

While these features are very attractive, and address many of the concerns with the main SAS statistical package outlined previously, such as the command-based interface and the lack of data sharing potential, there are still some drawbacks that are inherent in a software package aimed at a business environment. One, as mentioned earlier, is the cost. Another is the fact that the data sharing capability of the Web-based tool is designed on an enterprise-wide scale – that is, within a particular company. While this is fine for a corporate environment, it is not appropriate for a work environment where data must be shared across organizations and organizational contexts. Please see Sections 3.1 and 5.4 for further discussion of this issue.

Like SAS, SPSS has also responded to the business community's need for analytical tools that can be accessed across an organization. Its WebApp Framework gives "employees, clients and partners access to sophisticated analytics and graphical reports from any computer with access to the Internet" [SPSS press release, 2002]. The WebApp Framework's features include smooth handling of large data sets, allowing system administrators to create flexible security policies, and allowing users to "drill down" in a data set by month, year, or other hierarchies in order to perform an analysis.

However, the drawbacks that applied to SAS's Web-based query and reporting package apply equally to the WebApp Framework. Pricing for the system starts at US $100,000, although special pricing programs are available for government and academic institutions [SPSS press release, 2002]. As well, SPSS is designed to work on an enterprise-wide scale, and is ill-suited to projects involving multiple organizations.

While Rweb does indeed offer the user a Web-based statistical package, it still requires that the user learn how to use R and, specifically, the programming language S, and is thus subject to the same drawbacks as R that were discussed earlier in this section. The Rweb site also describes a series of modules that offer point-and-click forms based interface to R that allow the user to choose a data set, the type of analysis, and the options for the analysis. The author describes several available analysis types, including summary statistics, linear regression, one and two-way analysis of variance and a probability calculator [Rweb, 2006]. See Figure 6 for a screenshot of one of the modules.



You are working on dataset: trees

```
This data set provides measurements of the girth, height and volume
of timber in 31 felled black cherry trees.  Note that girth is the
diameter of the tree at 4 ft 6 in above the ground.

Variables:

Name    Type          Description
Girth   Continuous    Diameter of tree trunk 4'6" above the ground
Height  Continuous    Height of tree
Volume  Continuous    Volume of the trunk of the tree
```

Regression

Rweb Regression Analysis

The Model

- Response: Volume
- Predictor(s)
  - ☒ Girth
  - ☒ Height
  - ☐ Volume
- ☒ Include an intercept
- ☒ Plot all predictor vs response plots and include simple linear regression line.

Residual Analysis

☒ QQ plot of residuals
☐ Histogram of residuals
☐ Plot residuals vs predicted value

- Plot residuals vs predictors (choose which predictors)

☐ Girth
☐ Height
☐ Volume

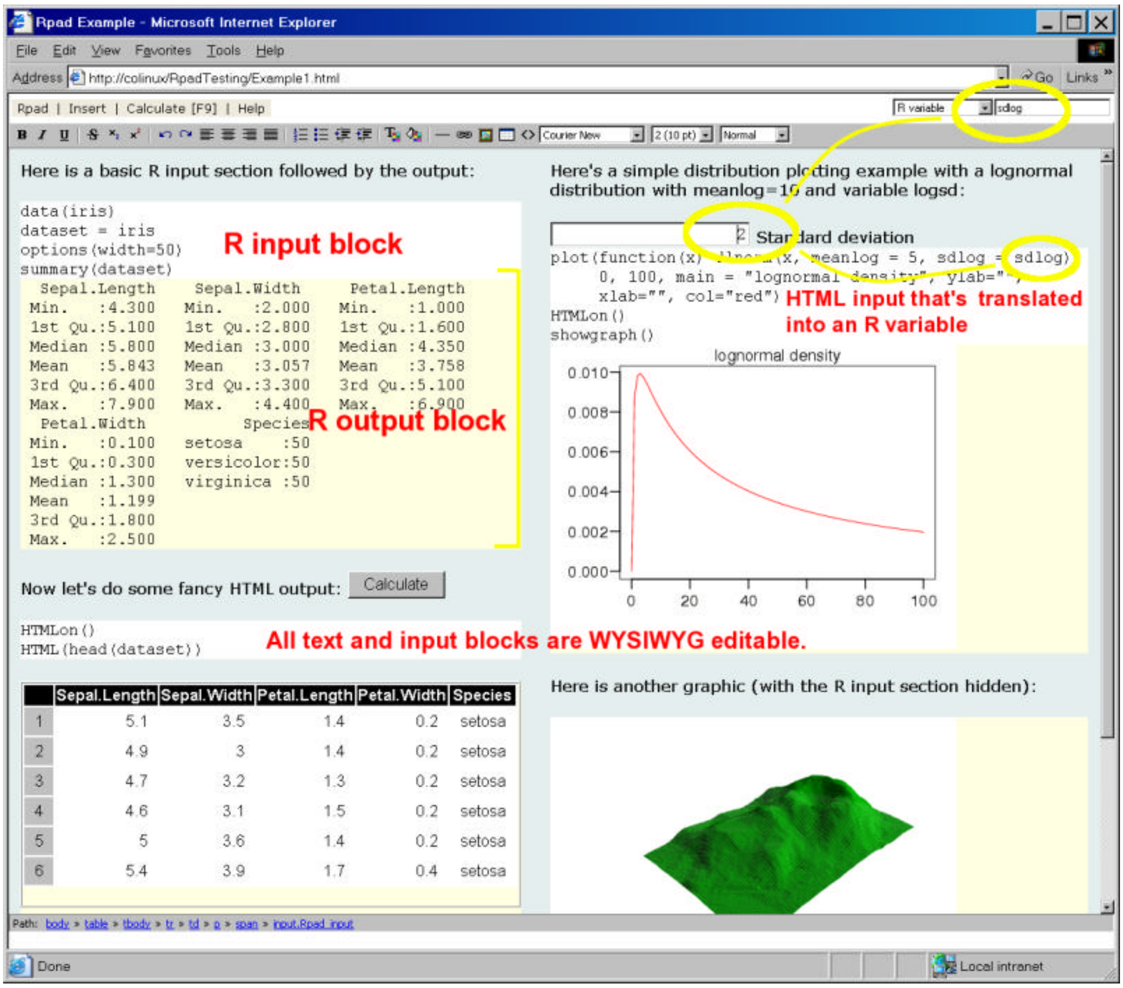
Plots

Histogram	Scatterplot
☐ Girth ☐ Height	☐ Girth ☐ Height

Figure 6: Screenshot of an Rweb Module

While this would seem to overcome the problems with the command-based version of Rweb, there are two major drawbacks. One is that the author freely admits that “I’m making up the modules as I go (both content and design)” which indicates that the Rweb modules have not been rigorously – or necessarily coherently – designed, and thus may not be a significant improvement over a command-line based interface [Rweb, 2006]. Second, when the Rweb modules were accessed, a “Page not found” error was received, which means that any functionality must be considered on a theoretical rather than practical level [Rweb, 2006].

Rpad, meanwhile, offers some of the graphical interactivity promised by the Rweb modules, but still expects the user to have a working knowledge of R in order to use the system. See Figure 7 for an annotated screenshot of Rpad.


Rpad Example - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://colinux/RpadTesting/Example1.html
Go Links
Rpad | Insert | Calculate [F9] | Help
R variable
sdlog
Courier New
2 (10 pt)
Normal
Calculate
lognormal density
0.010
0.008
0.006
0.004
0.002
0.000
0
20
40
60
80
100
Path: body » table » tbody » tr » td » p » span » input.Rpad_input
Done
Local intranet

Here is a basic R input section followed by the output:

```
data(iris)
dataset = iris
options(width=50)
summary(dataset)
```

R input block

```
  Sepal.Length    Sepal.Width     Petal.Length
 Min.   :4.300   Min.   :2.000   Min.   :1.000
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600
 Median :5.800   Median :3.000   Median :4.350
 Mean   :5.843   Mean   :3.057   Mean   :3.758
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100
 Max.   :7.900   Max.   :4.400   Max.   :6.900
  Petal.Width          Species
 Min.   :0.100   setosa    :50
 1st Qu.:0.300   versicolor:50
 Median :1.300   virginica :50
 Mean   :1.199
 3rd Qu.:1.800
 Max.   :2.500
```

R output block

Now let's do some fancy HTML output:

```
HTMLon()
HTML(head(dataset))
```

All text and input blocks are WYSIWYG editable.

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa

Here's a simple distribution plotting example with a lognormal distribution with meanlog=10 and variable logsd:

2 Standard deviation

```
plot(function(x) dlnorm(x, meanlog = 5, sdlog = sdlog),
     0, 100, main = "lognormal density", ylab="",
     xlab="", col="red")
HTMLon()
showgraph()
```

HTML input that's translated into an R variable

Here is another graphic (with the R input section hidden):

Figure 7: Rpad Screenshot (annotated)

2.4 Review of Academic Literature

Aside from the variety of existing statistical packages, the analysis of which informs the creation of the system described herein, there is also a diverse body of academic literature related to the development of a Web-based statistical toolkit. This literature can be broken down into several categories: descriptions of the architecture for Web-based statistical systems; numerical accuracy in statistical software packages; and guidelines for analyzing survey data, characteristics of potential users, and implementation-related issues. Each of these categories will be discussed in detail in the sections that follow.

2.4.1 Web-based System Architecture

Fielding and Taylor published a thorough description of the proper method for designing Web-based systems. One of their insights is that the stateless nature of the Internet forces the client – usually a

Web browser – to maintain all state-related information and pass that information along to the server. As well, they draw attention the possibility of the user entering “malformed or maliciously constructed data”, and emphasize the necessity of separating user interface concerns from data storage concerns [Fielding, 2002]. On the whole, the article contains a number of excellent principles and guidelines for designing Web-based applications.

Fernández took a slightly more specific look at the issue, focusing on what she calls “data-intensive web sites”, that “integrate information from multiple data sources, often have complex structure, and present increasingly detailed views of data” [Fernández, 1999]. Building these sites, she claimed, involves three tasks: accessing and integrating data, building the site’s structures, and generating the HTML representation of the pages. Fernández and her colleagues have developed a framework called STRUDEL to facilitate the development of this kind of site.

Shih and Lee, in presenting the architecture for the WWW CALS system, convey a common architectural structure; they use a statistical calculation engine as a central component that sits between a browser at the front end and data sets stored in a database at the back end [Shih, 1998]. This three tier architecture is described also by Huang in his JSIM Database project [Huang, 2000], and may be augmented with further components, as in DHLAS [Thriskos, 2006] or with graphics visualization tools proposed by Holmberg *et al* [Holmberg, 2006].

W^{3}MCSim, a Web-based Monte Carlo simulator for interactive probabilistic and statistical modeling developed by Bulis and DiStefano, takes a slightly different approach. It takes advantage of Microsoft’s Component Object Model (COM) technology to download the content and application logic to the client – in this case, the Internet Explorer Web browser – and executes the required calculations locally rather than on the server [Bulis, 2005]. However, this system does not incorporate a data storage component, and thus does not have the same level of data analysis capabilities as the system proposed in this thesis. This COM-based approach is also used by Moon *et al* in their Web-based sample size and power estimation system [Moon, 2002]. However, Moon’s system was limited to a very specific set of calculations, and thus is not suitable for general-purpose analysis.

At the other end of the complexity spectrum, Chiu recently proposed a “hypermedia-enabled and Web-based data analysis framework” which would allow users to “utilize historical data to discover useful information and improve the process of business decisions” [Chiu, 2004]. This framework is composed of eight logical components: operational and external data, transformation tools, data warehouse and other databases, data analysis tools, pattern discovery, hypermedia engine, Web servers, and browsers.

Günther *et al* took yet another approach to Web-based statistical software; they sought to create a platform which would allow users to make use of several statistical packages in performing an analysis. Their system, MMM, “is a collection of middleware services to … facilitate the sharing of software modules across heterogeneous networks” [Günther, 1997]. For example, a user might use a Mathematica script to access a database, a MatLab routine to compute the time series of estimated residuals, and then use those residuals in a XploRe macro in order to analyze the conditional second moments. To this end, they used a design based on a distributed services paradigm – the user communicates through a broker, which “mediates transactions between services and the user” [Günther, 1997]. This approach is appealing in its focus on allowing users from heterogeneous environments to perform statistical analysis using a common platform, but it requires the user to be familiar with all of the component tools (such as Mathematica and MatLab), which poses a daunting challenge to users who do not have a strong background in using statistical software.

2.4.2 Statistical Accuracy

In 1994, Wilkinson published a set of principles to guide the assessment of the accuracy and quality of the algorithms used in statistical software. Among these principles were that borrowed code should be avoided, the importance of verifying both simple and complex operations, and the necessity of rigor in conducting tests. This article included also a few simple data sets which illustrated these points; one contained nasty data of various kinds: very large numbers, very small numbers, and numbers which were very close to a whole number, to list a few of the columns. The other data set contained data that, although quite simple, was organized in a way that made analysis quite difficult, since it ran counter to the standard row and column organization of most statistical packages. Sawitzki ran the first data set through a number of commonly used statistical packages, and found serious deficiencies in many of them, including Microsoft Excel [Wilkinson, 1994]. Wilkinson attempted to use the second data set to perform a simple analysis, and found that few statistical packages were able to perform the analysis [Wilkinson, 1994].

McCullough, in a pair of papers published in 1998 and 1999, proposed a set of benchmarks that could be used to test the accuracy of statistical software. The first paper outlines some of the underlying reasons for inaccuracy in statistical software, such as infinite binary representation of finite decimal numbers (e.g., $0.1 = 000110011\underline{0011}$), the limited number of bits available to represent numbers, referred to as "precision", and errors due to truncation. It includes also descriptions of tests for univariate analyses (i.e., calculating the mean and standard deviation), linear regression, non-linear regression, analysis of variance (ANOVA), random number generation and statistical distributions [McCullough, 1998]. The second paper uses these tests to evaluate SAS, SPSS and S-Plus. In summary, while all three programs performed well on univariate analyses, McCullough uncovered flaws in estimation, random number generation and statistical distributions in each of the packages [McCullough, 1999]. Thus, the article highlights the importance of assessing the reliability of statistical software, both in order to inform users of the quality of the software that they are using, and also to motivate statistical software developers to write more accurate statistical calculations.

2.4.3 Data Analysis

Alongside issues of statistical accuracy, which were discussed in the previous section, issues surrounding the analysis of data have also been explored. In 1999, Wilkinson et al published a set of guidelines for experimental design and the analysis of the resulting data. Some of the guidelines that relate to the analysis of data are that one should look at a graphical representation of the data to identify invalid data, one should use common sense when using statistical software, and that assumptions, effect sizes, and limitations should be clearly stated. As well, claims of causality should always be supported by both the data and the assumptions that went into the experiment [Wilkinson, 1999].

In 2002, Whitley and Ball presented a review article that gave advice on summarizing and presenting data. The article describes the appropriate usage of measures of central tendency, such as mean, median and mode, and of variance, such as interquartile range, standard deviation and variance, and describes common distributions and the logarithmic transformation [Whitley, 2002].

In 2003, Kitchenham and Pfleeger published recommendations for the analysis of survey data, and focused specifically on software engineering surveys. They recommended, as did Wilkinson, that the data be validated before it is analyzed, for several reasons: to identify erroneous values and also to identify potential analysis issues related to missing values. Finally, an analysis being performed on a pilot survey can identify questions that need revision before being used in the main survey. They cautioned against treating coded ordinal and nominal data as if it were numerical data, and discussed the increased probability of spurious results when analyzing long questionnaires [Kitchenham, 2003].

2.4.4 Characteristics of Potential Users

One important issue in identifying the requirements for the users of statistical systems is the level of familiarity with statistics in general and specific statistical software packages. The existence of established statistical packages such as SAS and SPSS does not mean that all users will find these packages usable or – even if they do – that the analyses that they perform are valid. In a recent survey of a broad cross-section of tobacco control professionals, Morales found that 55.8% of respondents rated their ability to conduct data analysis using statistical software such as SAS or SPSS as "Poor", the lowest rating; for respondents from the health sector, as opposed to research or education, this percentage increased to 78% [Morales, 2006].

One response to the potential for inexperienced users to make mistakes when performing advanced statistical procedures is to prevent them from performing these procedures in the first place. Justifying the limited set of statistical operations available in MedCalc, a computer program for medical statistics, Schoonjans *et al* stated that "more complex statistical analyses are not implemented in the software. If researchers with limited statistical training require more sophisticated statistical analysis, they should refer to a statistician, not to a more complete statistical package" [Schoonjans, 1995]. Furthermore, "when unqualified users refer to a software package designed for statisticians, there is a danger of misinterpretation and error" [Schoonjans, 1995]. To a certain extent, this issue can be addressed with a statistical advisory component; see Section 3.4 for further discussion of this issue.

2.4.5 Implementation Issues

Jan de Leeuw, in a paper on statistics scripting in PHP, discussed the issues surrounding server-side versus client-side implementation of Web-based systems [de Leeuw, 1997]. Client-side systems, while putting fewer computational demands on the server, often require the user to download additional software or plug-ins. If a client-side system is implemented in JavaScript, it will tend to be quite slow. Programming on the server, on the other hand, allows the developer to configure the system themselves, rather than relying on the user's browser, and all the user has to download are HTML pages and graphics. While server-side programs do not offer the same interactivity as client-side programs, they are able to take advantage of the security features offered by the server.

While work has been done in various fields on using SQL for computational tasks –examples include Wolfram's research on using SQL for performing informetric queries and work by Warren and Johnson on implementing RS1 using SQL – the most relevant to this research is Choobineh's SQLSAM project [Wolfram, 2006; Warren, 2000; Choobineh, 1995]. SQLSAM is "an extension to the standard database language SQL for statistical modeling and analysis", which includes descriptive statistics, inferential statistics, probability distributions and regression analysis [Choobineh, 1995].

2.5 Summary

To sum up the preceding survey of statistical tools, there is a wide range of programs available commercially and on an open-source or academic basis. Major commercial packages like SAS and SPSS offer a comprehensive set of statistical calculations, but may be prohibitively expensive for many potential user groups, are not geared towards novice users – especially those with minimal statistical training – and offer Web-accessibility only via add-on packages, if at all. The open-source equivalent to these packages, R, while computationally comprehensive and freely available, is quite difficult for a user without programming experience to learn. While Web-accessibility is possible through programs like Rpad or R-php, it is left up to keen third-party developers to create, distribute and maintain these add-on components. In general, packages such as SAS, SPSS and R assume a greater knowledge of statistics than is actually possessed by many users, and thus alienate these users with a bewildering array of incomprehensible options.

In investigating academically oriented Web-based statistical tools, a few trends were discovered. First, many of these packages tend to be aimed towards very specific user groups, such as biotechnologists or geologists, and thus shut out the general user population. Second, these tools tend to either assume a working knowledge of the underlying statistical engine, often SAS or R, which then makes them inaccessible to many computer users, or were developed for pedagogical purposes, and thus are not robust enough for day-to-day real world analysis.

In short, while statistical analysis programs may be found which address some of the criteria described at the start of the chapter, none fulfills all or, indeed, most of them. Thus, there is a need for a statistical tool that not only enables online analysis, but also does so in a way that is accessible to users with little statistical training.

Chapter 3

Description of WebStats Model

This chapter describes and analyses the WebStats model on several levels. First, some questions about the system's basic design choices are addressed. Next, the overall structure of the system is presented, with an emphasis on the high-level interaction between the various components of the system. Next, the major components are described in detail. Finally, the system's user interface and implementation within the WIDE framework is discussed.

3.1 Overview of Model

Before discussing the particulars of the model's design, it is worthwhile to address some broader questions about the basic design choices that were made. Why develop a Web-based application? Why write statistical calculations in SQL instead of any other language, such as C++ or Java? Why write them at all, instead of relying on a third-party statistical application, such as R or SAS, to perform the calculations? These questions are answered in detail in the paragraphs that follow.

First, why should WebStats be Web-based, as opposed to a stand-alone program, or running on a client-server networked environment, which is quite common in university computer labs and large corporations? In short, Web-based software offers a "standard interface, multimedia presentation, multiple platform support, point-and-click capability, server extensibility, and low-cost global access", all of which "have made it a powerful and popular platform for information access" [Chiu, 2004]. Web-based software does, however, have disadvantages, as well; two of the most important are its dependence on the browser for the user interface and the stateless nature of the Internet, which was discussed in Section 2.4.1. By depending on a browser to present the user interface, the application is not in complete control over how the user interacts with the software, as the developer is in writing a piece of conventional software. As well, some of the most powerful browser features are only available on specific browsers, such as Internet Explorer's COM system, which further reduces a developer's options. Second, as discussed by Fielding, the Internet is designed to be stateless; that is, every time a server fulfills a client's request to load a web page, that request is made independently of any other requests that may have occurred previously [Fielding, 2002]. To get around this statelessness, the client must keep and send to the server all of the relevant state information on every request, which is cumbersome, but feasible.

Developing a client-server based application – that is, one that can run over a network – would seem to solve many of these problems. Client-server applications have been around since the days of mainframes in the late 1960's and early 1970's, and thus the issues related to developing applications for this environment are well understood. A client-server application, like a Web-based application, allows many people to use the same software, and even share the same data sets. Thus, it allows users to collaborate on projects, even if the users are dispersed across multiple locations. Moreover, a client-server application does not suffer the drawbacks – such as browser idiosyncrasies and a stateless medium – that Web-based applications must deal with, since it runs on a mainstream network-enabled operating system such as Windows, MacOS or Unix.

Given these factors, why have we chosen to develop a Web-based application? First of all, the drawbacks to Web-based applications do indeed make the process of writing the application code more painstaking for the developer. However, with enough thoroughness and an awareness of established workarounds, these drawbacks are not insurmountable. For an illustration of how this sort of issue can be resolved, see Section 6.5.2. Second, while a client-server application meets most of the collaboration needs for many users, it is not universally appropriate for all collaborative projects. One particular challenge that was found within the tobacco control community was that projects were distributed not only geographically, but also among organizations, and even across organizational contexts. For example, the evaluation of a particular smoking prevention program might involve a school board, a research group at a university, a regional public health unit, and a non-profit agency. These four contexts – educational, academic, government healthcare and non-profit – are quite different, and there will not necessarily be any co-operation or synchronization between the organizations, aside from their work on the tobacco control program. Thus, setting up a network for these disparate groups to use for data analysis would be prohibitively difficult, from both a financial and an administrative perspective. With a Web-based system, on the other hand, all that each organization needs to take full advantage of the system is an Internet connection, the URL of the site, and login information.

Another important issue is the way in which the statistical operations are calculated. A common design choice among several of the Web-based systems reviewed in the Chapter 2 is to use R or, equivalently, a commercial package such as SAS, to perform the statistical calculations. This choice is appealing for a couple of reasons. First of all, any of these packages is quite well established, and thus the statistical calculations have been written and checked by dozens of experts over many years of use. Second, there is a broad range of operations available, so users will not feel limited by the statistical calculations provided by the system. However, there are also some drawbacks. First of all, many of the Web-based systems mentioned in Chapter 2 require each user to be familiar with the package being used to perform statistical calculations, which reduces the system's usability to the level of the statistical package. Second, the data being analyzed must be in a format that the statistical package accepts. Depending on the type of data being analyzed – numerical vs. coded data, discrete vs. continuous values – this dependency could pose an insurmountable problem. Third, given that the system is not intended to be a replacement for the statistical breadth and depth offered by major packages, offering the user the full range of calculations available in R or SAS is unnecessary at best, and may be overwhelming or intimidating to the novice user. The calculations required for the tobacco control context are basic enough that encoding them was not felt to be a major deterrent for the application developer. If it is determined at a later date that this approach is impractical and employing a SQL-based library of statistical functions makes more sense, then this change can be made with minimal disruption to the system as a whole. For more on this, please see Section 3.7.

One important related issue, however, is statistical accuracy. As was mentioned previously, one of the advantages of using an established statistical package is that statisticians have verified its calculations. To address this issue, the calculation results in WebStats were checked against Excel, R and SAS using a randomly generated set of 100 numbers to verify their accuracy. The only case in which the results did not agree was in the calculation of quartiles. This was expected, since quartile calculation varies slightly between various established packages such as Excel and MiniTab. WebStats uses the definition of quartiles proposed by Moore and McCabe [Moore; McCabe, 2002].

Given, then, that the statistical operations are to be written by hand, why write them in SQL, which is usually used for querying a database for particular sets of records? Why not use a language like Java, which is commonly used in Web-based applications, or C, which is often used for scientific or mathematical applications? While SQL is perhaps not the most intuitive language to work with, it is Turing complete, which means that it can be used to express any programming construct or calculation. As was shown by Choobineh's SQLSAM project, SQL can be used to create a comprehensive library of statistical operations [Choobineh, 1995]. Another factor is the close relationship between the data storage and analysis components. Both are important parts of the system as a whole, and to favour data analysis concerns over data storage concerns may have lead to statistical calculation code that, while familiar to many programmers, was very inefficient in terms of memory usage.

Having addressed these initial questions, we turn now to the Web-based statistical analysis model itself. The model, when viewed at a high level, is composed of a few engines and components, linked together by data sources, as shown in Figure 8. We developed this architecture in the early stages of the system development process, which is described in Chapter 4, and modified it slightly as we progressed in implementing and refining the system. The architecture is similar to the commonly used three tier model described in Section 2.4.1. This model, with its separation of data storage, calculation and presentation concerns, was a logical starting point. However, this model differs significantly in its placement of the analysis storage component at the core of the system. In doing so, the model allows the system's functionality to expand beyond that of a statistical analysis tool to support analysis retrieval and dissemination. An alternative design is the COM-based design used by W^{3}MCSim, a Web-based statistical modeling tool [Bulis, 2005]. However, COM is a proprietary Microsoft technology that is viewable only via Microsoft's Internet Explorer browser, which presents a barrier for people using other browsers, such as Safari, which is commonly used on the Apple platform, or Mozilla Firefox.

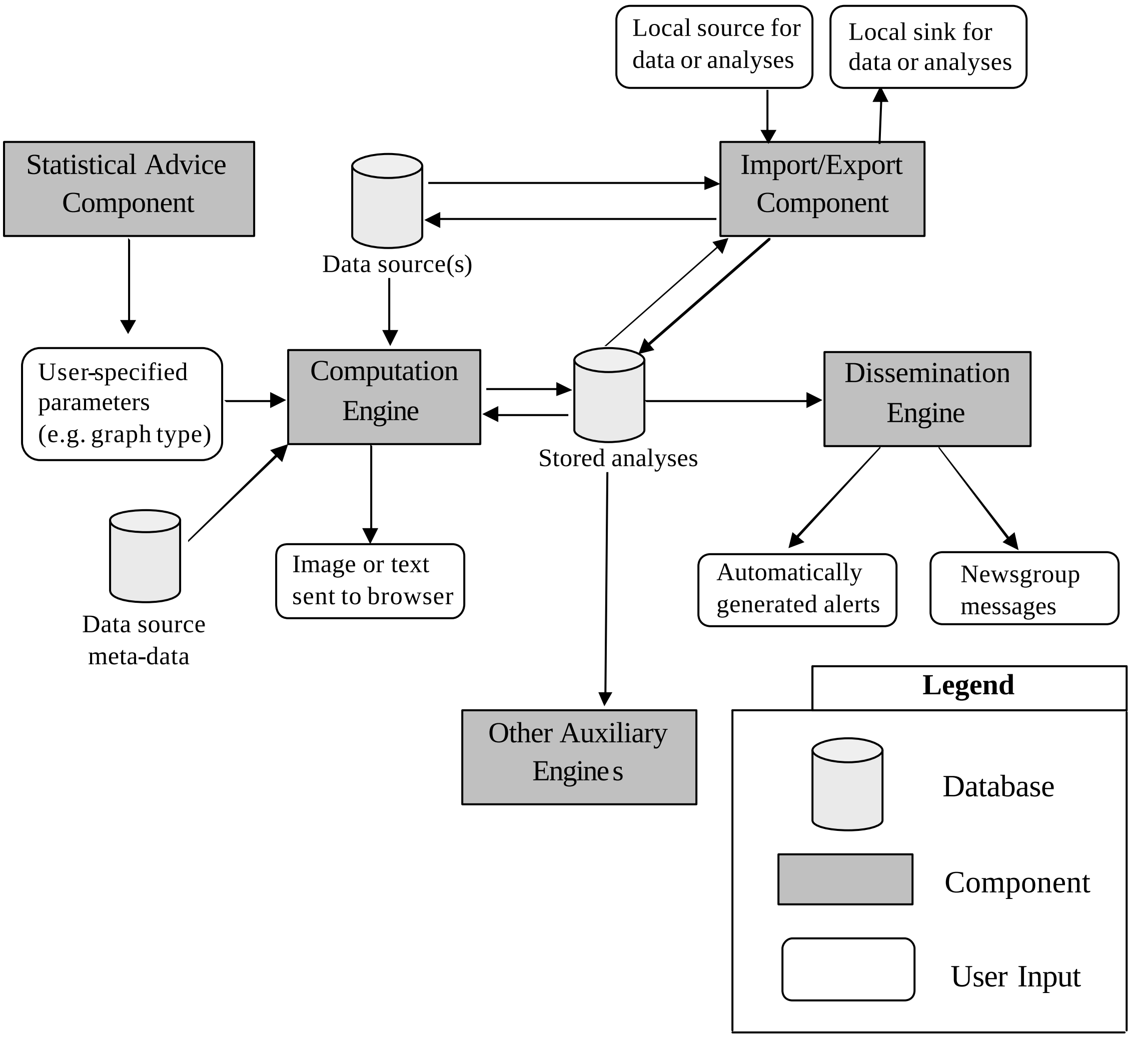

Local source for data or analyses
Local sink for data or analyses
Statistical Advice Component
Import/Export Component
Data source(s)
User-specified parameters (e.g. graph type)
Computation Engine
Stored analyses
Dissemination Engine
Data source meta-data
Image or text sent to browser
Automatically generated alerts
Newsgroup messages
Other Auxiliary Engines
Legend
Database
Component
User Input

Figure 8: High-level View of WebStats System

The computation engine is responsible for performing statistical calculations and creating graphs based on stored data, user parameters, and information about the stored data. The results of these calculations are stored in a separate set of tables, for use by other parts of the system. For example, the results are used by the dissemination engine, which keeps users informed about relevant analyses. The statistical advisor component, meanwhile, presents the user with advice to assist them in performing statistically valid analyses. Finally, the import and export component allows the data and analyses used by the system to be augmented by resources that are available locally, as files stored on a user's hard drive, for example, or saved to a user's computer. Each component is discussed in detail in the following sections.

3.2 Statistical Computation Model

The model used by WebStats to perform statistical calculations is composed of three parts: a statistical operation engine, a graph generation engine, and a compilation engine. Figure 9 shows how these components fit together in the system. They are described in detail below.

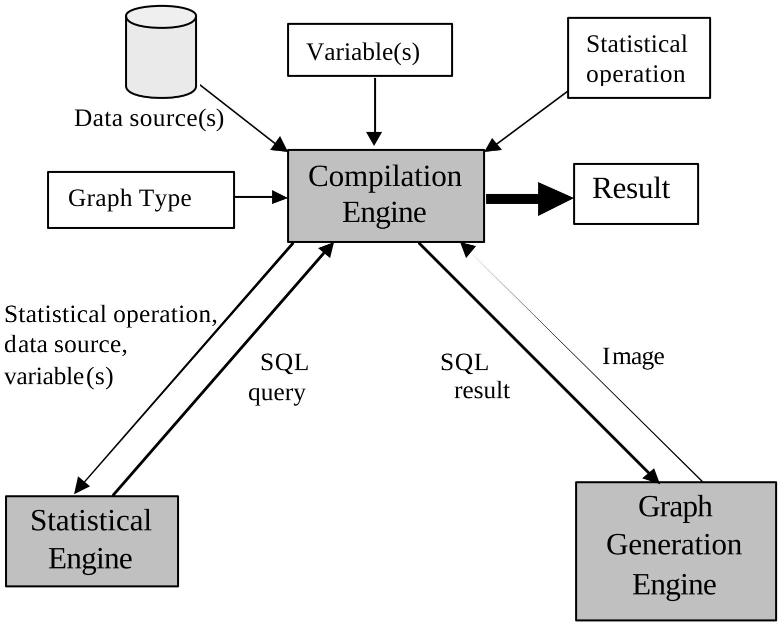


Figure 9: Statistical Computation Model

The statistical operation engine takes the user's parameters, such as the tables from which to retrieve data, the type of operation to perform, and the variables used, and creates an appropriate SQL query. In some cases, this query will use aggregate methods built into SQL such as `STDDEV` and `VAR` for standard deviation and variance, respectively. For more complex queries, however, SQL's mathematical operators, such as `SUM` and basic arithmetic operators, are used to create custom SQL methods that execute the requested operation. The results of these queries are placed into a result table with a standard format, to enable processing by other components of the system. For example, when the system processes a frequency table operation for a data table with coded answers, the result table stores the variable name, the possible values that the variable can take, labels associated with these values, and the result of the calculation. The labels are quite useful when working with survey results and other data sets where the value may not convey very much meaning. See Figure 10 for a sample result table after calculating a frequency table for an "Age" variable:

Name	Value	Column Label	Result
Age	1	14 or younger	10
Age	2	15	18
Age	3	16	32
Age	4	17	27
Age	5	18 or older	13

Figure 10: Sample Result Table

The graph generation engine takes the standard result table generated by the SQL query and turns it into the requested kind of graph, such as a line graph or a bar chart. This output is produced using the GD graphics engine, which offers clean, crisp graphics and TrueType fonts. It should be noted that since this component is decoupled from the rest of the system, the graphics format can easily be changed to use the SVG format, for example, if necessary.

The compilation engine ties the statistical operation and the graph generation engines together and handles input from and output to the user interface. While dealing with the input and output are fairly simple tasks, consisting of processing HTTP content headers and outputting appropriate HTML tags, the first task becomes somewhat complex when dealing with anything beyond simple statistical operations. If the operation involves a single operation on a single table, then the compilation engine simply converts the user's request into an SQL query, passes the results into the graph generation engine, and shows the graph generation engine's output using an `IMG SRC` tag.

However, it is desirable to allow the user to perform a sequence of statistical calculations on a variable, such as a frequency table followed by a standard deviation calculation, or to perform a calculation over several tables, such as calculating the average of a value in two data sets and showing the results side-by-side. To do so, the compilation engine needs to perform a series of statistical operations. In addition, it may need to create multiple graphs, which will be superimposed on top of each other. It is in facilitating this sort of complex calculation that the compilation engine goes beyond acting as a simple conversion mechanism, since it must ensure that the data are managed correctly and the results are passed correctly from one calculation to the next or from a calculation to the appropriate graph.

To achieve this kind of computing power, slightly more complex data structures are required. Specifically, if a single statistical computation, referred to as an *operation*, is the most basic unit of computation, the operations that are to be presented on the same graph are grouped together in an *analysis*, as shown in Figure 11 and discussed in greater detail in Section 6.3.

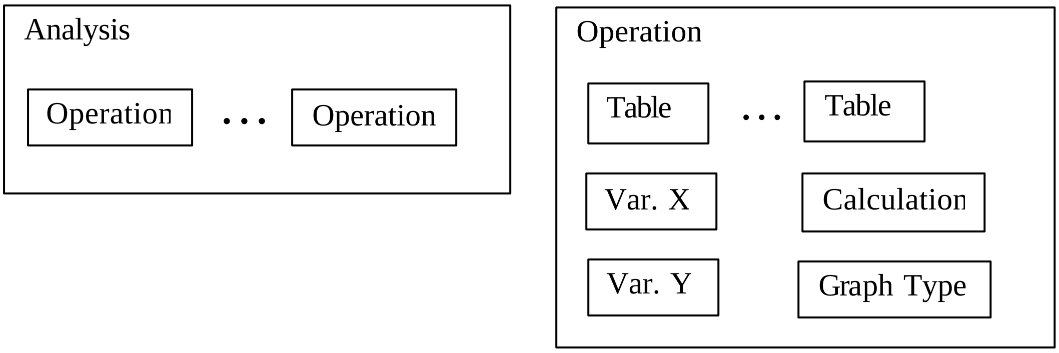


Figure 11: Scheme for Managing Complex Analyses

Thus, by making use of these data structures, the compilation engine can ensure that arbitrarily complex analyses are processed and shown properly.

3.3 Dissemination Engine

WebStats takes the storage of analyses as its starting point, and adds the concept of labelled analyses. WebStats allows any analysis, data set or user to be labelled with keywords. A keyword, which can be applied manually by a user or can be determined automatically by an intelligent classifying agent, helps to categorize the data set, analysis or user of the system. Keywords may be used to present the user with analyses or data sets which they might find interesting, to send the user automatic updates about new, relevant analyses that have been performed, and even help the user connect with other colleagues with similar interests who are using the system. These keywords can also be used to tie together quantitative data – such as the statistical analyses performed through WebStats – and qualitative data – for example, anecdotal evidence, or interview transcripts. Thus, the simple concept of labelling, when applied to a variety of components (e.g., users, analyses, anecdotes) can become quite powerful, and can lend structure to what might otherwise be a chaotic collection of disconnected data.

Another important feature of the dissemination engine is the range of options available to the user once they view the results of their analysis. Typically, this is limited to the user's viewing the resulting image in his Web browser, and then perhaps printing or saving the image. One of the goals of WebStats, however, is to facilitate the sharing of analyses among a large number of people. To this end, WebStats stores its analyses in a database that users can search via a Web-based form and also offers the user the capability to post the results of his analyses on a built-in message board for others to view.

3.4 Statistical Advisor

WebStats has been designed so that it may be used effectively by anyone who does not have a strong background in statistics. One of the implications of this design choice is that such a user will require some guidance in order to generate statistically valid analyses from data sets. This guidance is necessary throughout the process of performing an analysis, whether it uses descriptive or inferential statistics. As was mentioned in Section 2.4.3, there is the potential for misinterpretation and error if a novice user performs analyses without proper training or guidance [Schoonjans *et al*, 1995].

By including in WebStats an intelligent component that is encoded with some basic rules of statistical analysis and inference, the system can help the user interpret the generated results and warn the user if he attempts to perform an operation which is not supported by the available data. This assistance could be extended to having WebStats guide the user in creating his experiments, entering data, performing the necessary statistical calculations and, finally, interpreting the results. It should be noted, however, that this amount of assistance goes well beyond the system's current scope of analyzing stored data.

There are a number of factors that may make a particular statistical calculation or graph type inappropriate for a data set. Among these are the data set's modality and the number of outliers present in the data set.

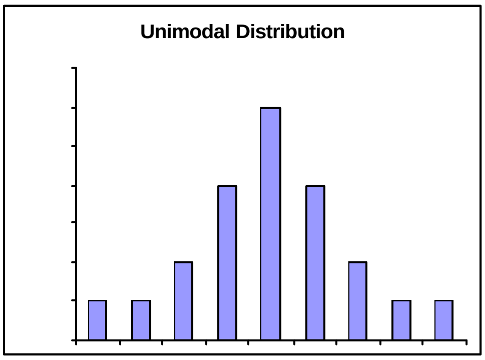


Figure 12: Unimodal Distribution

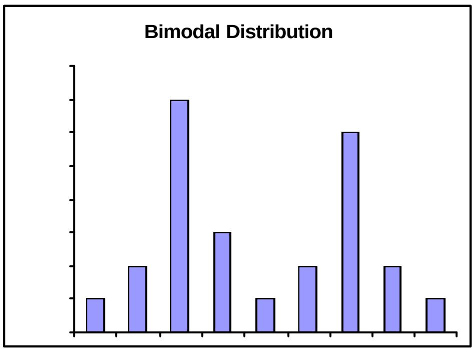


Figure 13: Bimodal Distribution

One crucial characteristic of a data set that must be taken into account is its modality. Simply put, the modality of a data set may be thought of as the number of central points it has. A unimodal data set has one central point, as shown in Figure 12, and standard measures of central tendency, such as mean, median and mode, will indeed find the central point of the data set.

A bimodal data set, however, has two central points, as shown in Figure 13, and thus some measures of central tendency will not give the desired results. The mean of a bimodal distribution, for example, will tend to fall between the two central points, and thus will not convey the location of either of the data set's central points. One way to perform an analysis of bimodal data, suggested by Jones and James in the context of analyzing geologic data, is to estimate parameters related to each mode separately; however, this strategy does not work very well if the modes are close to each other, the proportion of the dominant mode is near 1.0 (that is, the modes have similar values), or coarsely grouped data are used [Jones, 1969].

Another factor that is important to consider is the number of outliers in a data set, as measures of central tendency vary in the degree to which they are affected strongly by the presence of outliers. Specifically, a median is affected more strongly by outliers than a mode and thus, when using a data set with a large number of outliers, it may be more appropriate to use the mode than the mean in order to find the centre of the data [Whitley, 2002].

The statistical advisor is programmed with these heuristics and several others and offers advice based on these heuristics to the user before he submits an analysis to the system. The advice is presented in a pop-up window, so that the user must, at the very least, read the message before proceeding with his analysis.

It should be noted that these guidelines are not always applicable, and thus none is encoded into the system as an inviolable rule. Rather, a collection of guidelines appropriate to the current analysis is presented as suggestions to help the user achieve a statistically valid analysis.

3.5 Import and Export Components

Another important set of system components is the import and export tools. One of the potential pitfalls of developing a new system to solve an existing technical problem is that users of existing, incompatible solutions are discouraged from adopting the new system, particularly because of the wide variety of existing solutions and varying technical skillls of the user population. Some users may find that adopting a new system is a technologically daunting task. Others may find that the system does not meet all of their needs. Thus, these users may prefer to perform some tasks, such as complex statistical calculations, with their own software. To address these concerns, WebStats allows users to import data from common statistical programs, such as those mentioned in Chapter 2, into the system in order to create graphs and reports, or to export data from WebStats to a variety of formats. The ability to import and export allows advanced users to use WebStats as a component in a larger system, and makes up for WebStats's deficiencies compared to specialized statistical packages. Basic users, meanwhile, can import and export data from programs with which they are familiar, such as Excel. Thus, they will be able to learn the components of WebStats on an as-needed basis, reducing the potential barrier posed by learning a new system in its entirety. It should be noted that the import and export capabilities are, of course, subject to information privacy conditions, which are discussed in detail in Section 6.3.8.

3.6 User Interface Features

In order for WebStats to be usable by a wide range of people, the user profile is expanded so that it controls both access to the system – that is, who can view the website – and the system's appearance and behaviour. A user can specify his familiarity with statistics, his level of comfort with computers, and the degree of customization he would like to have in creating analyses. WebStats, in turn, will adapt the options it presents to the users accordingly. A beginner might be shown a basic subset of the operations available, such as creating a bar chart or finding the average of a set of values, while a more advanced user might be given the option to perform more complex calculations involving regressions and confidence intervals.

This concept of dynamically customizing the system may be extended beyond the scope of an individual user to that of an entire group of users. For example, people in tobacco control might have a very different set of needs – and thus require a different set of functionality – from those of people in environmental assessment or public policy.

3.7 Analysis of Model

One of the primary strengths of the model is the decoupling of the statistics-operation, compilation and graph-generation components. This decoupling allows each component to be expanded, modified, or even completely replaced with little or no impact on the other components. For instance, to add a new statistical operation, the statistics engine must be expanded, obviously, but – aside from the slight modification of the compilation engine to allow it to call the appropriate method in the statistics engine – the other components are essentially unaffected.

This decoupling allows the compilation engine to handle also quite complex queries in a logical, intuitive fashion. In executing an operation composed of several statistical operations, the compilation engine simply needs to call each operation, *o*, in turn, using the previous operation's output as *o*'s input, following a standard pipeline design. Similarly, in creating a graph composed of several layered components, the compilation engine iterates through the graph components, adding each component to the composite graph, one by one.

As mentioned in the beginning of this chapter, having the analysis storage component be a core component of the model offers several advantages. First, by enabling WebStats to store analyses for future retrieval and dissemination, WebStats gains added functionality as a communication and collaboration tool for users and their colleagues. Second, it allows additional components to take full advantage of the stored analysis information – for example, a report generation component (discussed in Section 6.3.1) could use the analysis information to create well-formatted reports in PDF and Word formats.

3.8 Implementation in WIDE

The model has been implemented using the WIDE (Web-based Informatics Development Environment) framework. By building the system on top of WIDE, WebStats can make use of the various data management and software tools already implemented in the WIDE framework. For more details about the WIDE framework, please see Appendix A.

One of the most important features of WIDE used by WebStats is its database storage capability, which underpins the entire WIDE system. All of the information stored within WIDE, from the content of each Web-based system to the code that assembles each system, is stored within a collection of Web-accessible databases [Cowan, 2006]. This means that incorporating access to data sets for statistical analysis is made very straightforward by WIDE's structure, since WIDE is built on a paradigm of accessing and editing database tables via Web-based forms.

Another feature of WIDE that is quite useful for WebStats as a piece of Web-based software is WIDE's user permission system. WIDE contains a user authentication system which is used by WebStats not only to permit or deny access to the system as a whole, but also to tailor the appearance and behaviour of the system to any user's profile.

One other aspect of WIDE that WebStats uses is WIDE's information dissemination facilities. One of the goals of WebStats, as mentioned earlier, is to facilitate the sharing of information, in particular, of statistical analyses, among a large audience. WIDE already contains two components that enable this sort of information exchange. One is an e-mail sending tool which allows people to receive information via e-mail directly from the system. Another is WIDE's ability – which is a key feature of the WIDE system as a whole – to take information stored in a database and displa y it online clearly and simply.

3.9 Summary

The Web-based statistical analysis model presented in this chapter consists of several distinct modules, which are responsible for storing data sets and related information, analyzing data sets and presenting the results, and storing and disseminating information about these analyses. The data analysis component is itself made up of a statistical calculation, a graphical presentation and a compilation engine, subdividing the functionality of the system into manageable pieces. The model specifies also a data storage scheme for storing data sets, analyses, and auxiliary information about both the data sets and the analyses performed on the data sets.

The organization of WebStats has several advantages. First, the compartmentalization of the WebStats's features allows for painless replacement or improvement of a single component. For example, if chart drawing were to be performed using a library other than GD, the only component that would need to be changed is the graph generation engine. Second, putting the analysis storage component at the core of the system allows additional modules to take full advantage of the rich set of information stored about the analyses.

Implementing the system within the WIDE framework offers several advantages, as well. The data storage and information dissemination capabilities offered by WIDE fit well with WebStats's analysis storage and collaboration tools. In addition, WIDE's user permission system offers a starting point for WebStats's concepts of analysis and data set permissions.

Chapter 4

Requirements Engineering Process

This chapter describes the process that was used to validate the model described in the previous chapter. At the highest level, stakeholders from one user group, in this case, people in the tobacco control field, were consulted in order to better understand the requirements for the system. These requirements were used to develop a robust prototype, which was tested and evaluated by the stakeholders in an iterative development process. Finally, follow-up work was done to assess the effectiveness of the model for the tobacco control context. The first section of this chapter describes the user population. The second section of this chapter outlines the reasons for choosing the tobacco control field for the case study. The third section goes into detail about the process that was used to develop and validate the model.

4.1 Description of User Population

Tobacco control is a diverse field and includes professionals from a wide variety of disciplines working in numerous organizations and environments. Broadly speaking, they can be divided into three main groups: public health professionals, managers, researchers. Each of these groups will be described briefly in the following sections. For more on how these groups interact in the field of program evaluation, see Section 5.3. For a description of the user population in terms of a User-Uses model, see Section 6.2.

4.1.1 Public Health Professionals

A public health professional is a front-line worker who creates, delivers, and collects data on tobacco control programs. He often works at public health units, although there are also non-governmental organizations active in this area. He works at the regional scope, and his goals include creating effective tobacco control programs for his region, delivering those programs to the relevant populations in his region, and keeping track of measures of the programs' effectiveness. He is usually trained in fields such as public health, communications and program delivery, although this training can vary widely.

4.1.2 Managers

A manager guides broad tobacco control strategies, usually on a provincial level. A part of this job is managing up to several dozen tobacco control programs. She is responsible for determining which programs are meeting benchmarks, ensuring that funding is allocated appropriately, and dealing with related government agencies and other relevant groups. She usually works within a government agency, such as the provincial Ministry of Health, and has a substantial amount of training in data analysis, program management, and other related fields.

4.1.3 Researchers

A researcher, working within academic institutions and research groups, analyzes tobacco control programs in order to answer broad research questions. The scope of these questions will vary depending on the question being tackled, and may deal with anywhere from a few hundred to hundreds of thousands of people. A researcher typically has at least one graduate degree in statistics, mathematics or a related field. Thus, he has a high level of statistical training, and is quite comfortable using both standard statistical packages and more advanced tools appropriate for specific types of analyses. A related group of research users is that of epidemiologists, each of whom also has a strong background in statistics, a similar skill set to academic researchers, and tends to answer broad questions. One important difference between an epidemiologist and an academic researcher, however, is that an epidemiologist tends to work in a public health context, rather than in an academic context.

4.1.4 Summary

Please see Figure 14 for a summary of the characteristics of the groups described in the previous sections.

Group	Organization	Scope	Goals
Public Health Professionals	Public health units, NGOs	Regional	Create and deliver programs
Managers	Health ministries	Provincial	Guide and evaluate broad strategies
Researchers, Epidemiologists	Universities, research groups, public health units	Varied	Answer theoretical questions, help inform policy, identify trends

Figure 14: Table of User Characteristics

4.2 Choice of Stakeholder Group

When validating a model, it is important to ensure that the method of validation is appropriate for the situation being modeled. There were several factors that made the context of tobacco control particularly well-suited to testing the model developed in this thesis. These factors are discussed in detail below.

4.2.1 Statistical needs

Given the variety of responsibilities within the tobacco control field, statistical analysis, program development and evaluation, and public education campaigns, to list a few examples, it is not surprising that the range of statistical calculations being performed is equally diverse. From simple histograms and pie charts to inferential calculations across multiple data sets, the kinds of calculations used by professionals in tobacco control provided a good test of the system's statistical robustness.

4.2.2 Data needs

The data requirements of tobacco control professionals are rigorous and varied. First, the data must be easily accessible by people in government, health, and education-based organizations, and thus cannot be confined to a private storage medium or proprietary format. As well, these data must be as up to date as possible to ensure that the analysis performed on these data is meaningful and accurate. Since the data being analyzed may be the results of a survey given to thousands of students or the comparison of a handful of schools, the system must be able to present both large and small data sets quickly and clearly. Finally, data privacy and security concerns must be taken into account, given the potentially sensitive nature of the information being stored, analyzed, and disseminated.

4.2.3 Heterogeneity of user population

As mentioned previously, the field of tobacco control encompasses a wide variety of roles and organizations. Thus, not only do the potential users have a broad range of statistical needs, but they have also varying degrees of training in computer usage and differing goals in analyzing data, and they work in a variety of organizational settings. Thus, the user interface, functionality, system help and maintenance features, and options for distributing the analysis results must exhibit the flexibility necessary to accommodate these demands.

4.2.4 Communication and collaboration needs

One of the consequences of bringing together a diverse group of professionals to solve a complex problem, such as reducing the smoking rate, is that there needs to be a means for the various parties to communicate easily with each other and share information. Thus, WebStats must provide not just analysis tools, but also collaboration tools to allow these analyses to be shared quickly and efficiently among members of the tobacco control community.

4.3 The Development Process

WebStats was developed using an iterative, consultation-based approach which adhered to principles of requirements engineering. Requirements engineering has been defined by one researcher as: "a systematic process of developing requirements through an iterative cooperative process of analyzing the problem, documenting the resulting observations in a variety of representation formats, and checking the accuracy of the understanding gained." [Loucopoulos, 1995]. Specifically, we followed the Boehm's Spiral Model while developing the system [Boehm, 1988]. See Figure 15 for a diagram of the model.

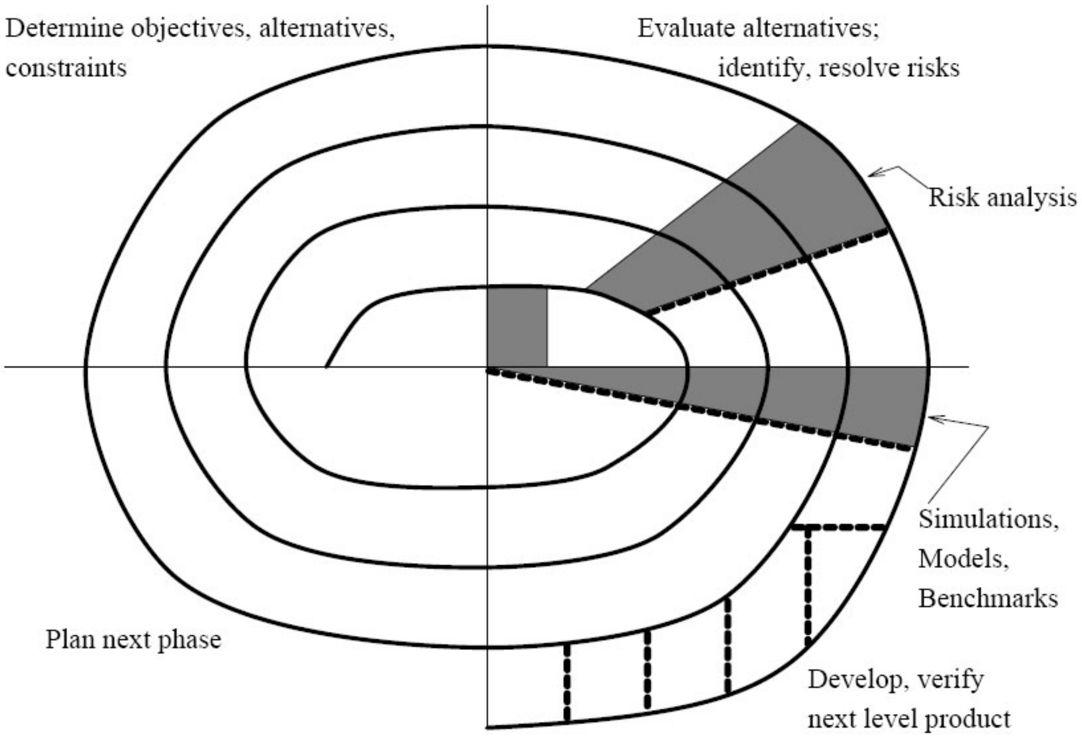

Determine objectives, alternatives, constraints
Evaluate alternatives; identify, resolve risks
Risk analysis
Simulations, Models, Benchmarks
Develop, verify next level product
Plan next phase

Figure 15: Spiral Model (Boehm, 1988)

Thus, based on this model, a broad view of the necessary steps in a requirements engineering-based development process are as follows:

1. Consult with stakeholders, elicit requirements.
2. Develop/refine requirements through analysis of stakeholder feedback.
3. Develop/refine a prototype based on these requirements.
4. Gather feedback on the prototype from stakeholders.
5. Using the feedback, assess the validity of the requirements.
6. If the requirements need refinement, return to Step 1.

These steps correlated with the four phases of the spiral model as follows: Step 1 falls in the "Determine objectives" quadrant, Step 2 in the "Assessing alternatives" quadrant, Steps 3 and 4 in the "Develop next level product" quadrant, and steps 5 and 6 in the "Plan next phase" quadrant.

The first step in the requirements elicitation process, in apparent contradiction to the Spiral Model, was the creation of a rough prototype system to be used in initial conversations. Because of the unfamiliar nature of the toolkit, the interviewees would benefit from having a concrete piece of software that could provide a sense of the possible features, appearance, and behaviour that WebStats might possess. According to one researcher, "most people (especially non-technically oriented) learn while doing; they've got to see some kind of prototype… to discover what they want" [Dameron]. When conducting these interviews, it was emphasized that this prototype was simply an *example* of the sort of tool that was being designed, rather than the implementation of the tool. That is, the prototype was used to help initiate the interview process, rather than to try and restrict or guide the interviewee's responses. The rest of the requirements engineering is described in terms of Boehm's model.

First iteration

Determine objectives: Once the prototype was created, the next step was to conduct initial, informal conversations with people at the University of Waterloo using this prototype. These conversations were held with readily available stakeholders from a variety of roles and user groups.

Evaluate alternatives: Through these conversations, we gained a greater understanding of the tobacco control context, and began to get a sense of what the core requirements for the system would be. As well, by conducting the interviews across a wide variety of roles, we saw how different users might have contrasting or even contradictory requirements. Finally, through these conversations, we made connections with a broad potential user base. These connections would become quite useful in the later stages of the project.

Develop product: Based on this analysis, development of the prototype proceeded with the implementation of several core areas of functionality. As well, the system was designed with the proper level of flexibility to accommodate the conflicting demands of the varied user population.

Plan next phase: An essential goal for the next iteration of the process was to broaden the pool of stakeholders to be interviewed. To this end, we collaborated with Rosanna Morales who was, at the time, a Masters Degree candidate in Applied Health Sciences at the University of Waterloo, in developing a survey to elicit requirements for a national information system to support youth tobacco control, with the understanding that the researcher would be able to contact any participant who agreed to be contacted for a follow-up survey. This arrangement served two purposes. First, data were obtained from people across the country about their requirements for a system quite similar to the one developed in this thesis. Second, it added both breadth and depth to the potential user base.

In assisting with the survey conducted by Morales, and making connections through events such as the 2005 and 2006 Symposia on Tobacco Control in Toronto, we developed a pool of stakeholders from which potential participants were selected. In choosing participants, the primary goal was to get feedback from a diverse group of people. Thus, feedback was sought from professionals from the three user groups, public health professionals, managers, and researchers, and the various organizational settings, universities, research institutions, public health units, government agencies, and non-governmental organizations, described in Section 4.1, and from as many parts of Canada as possible. Please see Appendix D for a summary of the demographic characteristics of the participants.

Second iteration

Determine objectives: The next step was to carry out more detailed interviews with interested stakeholders, i.e., public health nurses, professors, epidemiologists, statisticians, program developers, and a knowledge exchange officer. The interviewees included approximately the same number of women and men, and were from a wide range of ages. Please see Appendix D for a list of the questions that were used to guide the direction of these interviews.

Evaluate alternatives: Through these interviews, feedback was obtained from a wide variety of people on what the proposed system should do, who would be interested in using it, and how its user interface should appear.

Develop product: Based on this analysis, prototype development proceeded with the extension of the system's functionality, refinement of the user interface, and a great deal of testing to ensure that the input to the system was valid.

Plan next phase: After prototype development had progressed significantly, stakeholders were contacted once again and asked to provide feedback on the improved version of the prototype.

Third iteration

The third iteration of the development process proceeded in much the same manner as the second iteration: stakeholders were interviewed, requirements were analyzed, and the prototype was modified accordingly. By the end of the third iteration, we felt that the requirements had stabilized to the point where the requirements elicitation process could be considered complete, and the prototype could be evaluated in terms of its success in meeting the users` requirements.

Final Stage

In the final stage of the development process, interviews were conducted with potential users to evaluate the prototype. It should be noted that these interviews included both stakeholders who had participated in previous phases of the study, in order to determine whether the process as a whole had been successful, and people who had not seen previous versions of the system, in order to get a fresh perspective on the system. The feedback from these interviews indicated that WebStats did indeed meet the requirements that had been established and that people in a wide variety of roles within the field of tobacco control would find WebStats beneficial.

Chapter 5

Description of Tobacco Control Context

5.1 Smoking: A National Health Issue

Smoking is an issue that has a serious, wide-reaching impact on Canadians' health. Smoking greatly increases the risk of lung cancer and heart disease, which are the leading causes of cancer death and death from disease, respectively, in Canada [Canadian Cancer Society, 2006; Heart and Stroke Foundation, 2006]. It is estimated that smoking is responsible for over 85% of lung cancer deaths, of which there were nearly 19,000 in 2004 [Canadian Cancer Society, 2006]. A smoker has a 70% greater chance of dying from coronary heart disease, which accounted for over 74,000 deaths in 2002, than a non-smoker [Canadian Cancer Society, 2006; Heart and Stroke Foundation, 2006]. Overall, smoking kills over 47,500 people in Canada each year [Canadian Cancer Society, 2006]. On a global scale, smoking causes four times as many deaths as car accidents, suicide, homicide and AIDS combined [Manske, 2006].

The good news is that steady progress has been made in reducing the smoking rate in Canada over the last few years. The overall smoking rate has dropped from 25% in 1999 to 21% in 2003, to an all-time low of 20% in 2004 [Canadian Cancer Society, 2006]. One particularly important segment of the population is youth, people between 15 and 24 years old. 90% of smokers take up smoking before the age of 19 [U.S. Centers for Disease Control, 1994]; if this smoking uptake is prevented, then the smoking population can be dramatically decreased. The smoking rate for youth has also been decreasing recently, and at an even faster rate than the overall smoking rate. The youth smoking rate has declined from 28% in 1999 to 21% in 2003 to 20% in 2004 [Canadian Tobacco Usage Monitoring survey, 2006]. See Figure 16 for a detailed chart of youth smoking rates over the last twenty years.

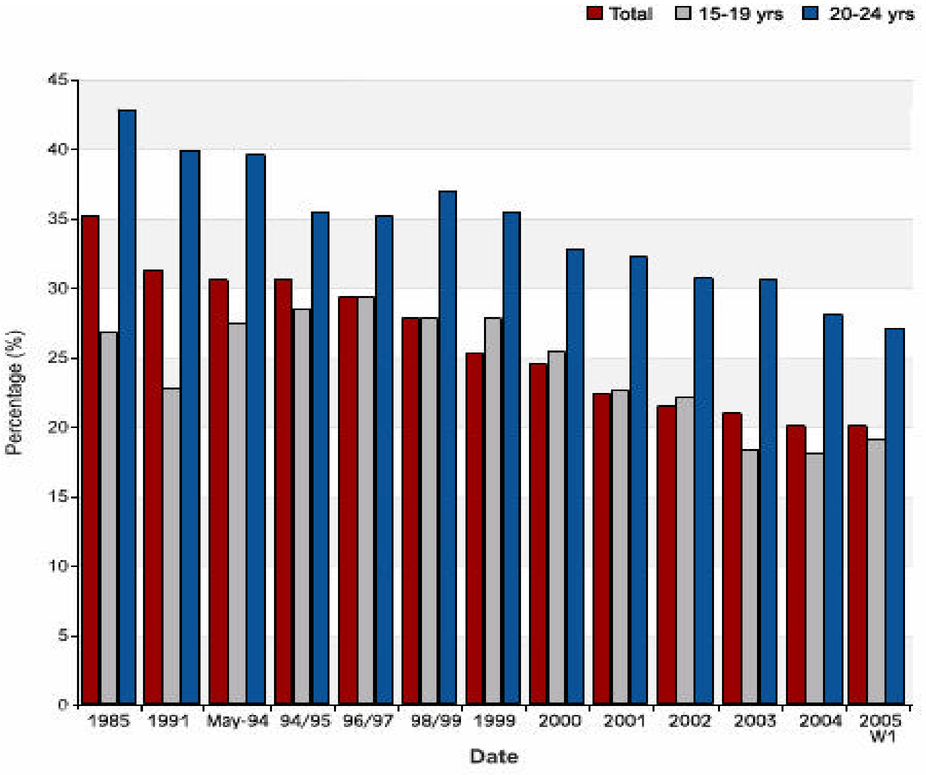


Figure 16: Youth Smoking Rates in Canada, 1985-2005 (CTUMS, 2006)

5.2 Population-Based Intervention

As mentioned earlier, the vast majority of smokers take up smoking before they turn 19; if this smoking uptake is prevented, then the smoking population will likely be dramatically decreased. That is, smoking is a solvable problem, if the proper populations can be effectively targeted with the proper measures. In the past, tobacco control policies tended to take one of two forms: one was a macro-level approach; that is, applying the same broad strategy, such as a nation-wide advertising campaign, or an increase in the tax on cigarettes, across an entire province or across the country. However, this strategy, although easy to implement, was not particularly effective, since it applied the same strategy to every region, regardless of its demographics and smoking health. That is, if the strategy was intended to reduce teen tobacco uptake, it would be applied the same way regardless of the number of teens who were likely to take up smoking – indeed, without considering the teenage population of a region at all [Morales, 2006]. The other approach, referred to as a clinical approach, involved targeting individual people who wanted to quit smoking, and providing them with materials to help them quit [Manske, 2006].

Clearly, it would be more effective for teen smoking prevention strategies, for example, to target populations with high levels of teens who are likely to start smoking, and to work with the entire population rather than individuals. Thus, over the past few decades, there has been a shift away from individual-based clinical interventions, and towards population-level interventions designed for a particular demographic [Cowan, 2005]. For these population-level interventions to be successful, however, large amounts of data, on a region's demographic profile, for example, or on a program's success rate, need to be shared among large numbers of people quickly and efficiently [Manske, 2005]. Ideally, public health programmers would like to implement a program known to be effective in a region with the appropriate demographics. However, choosing such a program requires data to show that the program has indeed been effective elsewhere under similar conditions, and current demographic data to ensure that the population being targeted does indeed fit with the program being implemented. This demand has lead to the emergence of program evaluation as an important sub-field within tobacco control [Manske, 2006].

Taking a population-based approach to intervention has also helped identify populations, such as aboriginal women and schizophrenics, with very high smoking rates for various reasons; carefully targeted programs will help these populations, as well [McDonald, 2003].

5.3 Data Analysis and Program Evaluation

Population-based tobacco control programs must be evaluated for funding to be allocated effectively and programs to evolve appropriately. Some programs may not be ideal for certain contexts or populations or may not be effective at all. Typically, scientific studies are evaluated using randomized-controlled trials (RCTs). RCTs are seen as the gold standard in science, but are inappropriate for many population health interventions. RCTs are based on being able to perform a causative experiment which controls for all but the variables being tested (which are referred to as experimental variables), thus ensuring the test is repeatable and has a clear explanation. However, populations are unique, and one cannot say that one population is exactly like another in terms of each population's non-experimental variables, so controlling for all non-experimental variables is impossible. As well, performing the same test on multiple populations, using one program with one set of populations and another program with the other set, is both practically infeasible and morally questionable, as one population is identified as benefiting from a program or treatment which is deliberately withheld [Cowan, 2005].

The practice that has emerged in population health has been to conduct descriptive tests; that is, observing the changes in a single population which is using a particular program, and measuring if this change matches what was expected at the outset. Time series and comparison group designs may be possible. This approach, obviously, requires that the outcome of a program be evaluated by a neutral party – often, a researcher interested in the program. It also requires a great deal of data – baseline data, data gathered before the start of the program, data collected after it has finished, etc. The field of program evaluation, as it relates to tobacco control, deals with the issues surrounding the gathering, analysis and dissemination of this data. One researcher summed up the issue as follows: "Effective public health practice requires timely, accurate, and authoritative information from a wide variety of sources" [Yasnoff, 2000].

5.4 Evaluation Crosses Role Boundaries

In general, public health workers, researchers and public policy people play quite different roles in the field of tobacco research. Public health workers are responsible for creating and implementing programs in a regional or community context. They have day-to-day contact with the people who are using the tobacco control programs, and thus have a very hands-on approach to tobacco control

[Taylor,2006; Zimmerman, 2006]. Researchers, meanwhile, tend to ask questions about the long-term performance of a program, or the relative performance of one program measured against another. Thus, they are usually not closely connected to a particular community or program, but have looser connections with a broad range of programs. Finally, public policy professionals work with government agencies at provincial and national levels in order to determine which programs to fund, and to advocate for more funding of tobacco control programs in general [Seliske, 2005].

Program evaluation is a huge field that cuts across many areas of tobacco control. Professionals working in public health want to know which programs are doing well and which ones are not so that they can use their resources as effectively as possible. Establishing initial benchmarks is also crucial in determining gaps that need to be addressed by new programs – for instance, a high smoking rate among adolescent girls in a particular region. Finally, program evaluation is used by external entities, such as government agencies, to determine their level of funding, and thus has a major impact on people working in public health in that regard, as well [Manske, 2005; Jolin, 2005].

For researchers, program evaluation is essential, as it allows them to answer research questions about the effectiveness of a program or strategy with a particular population without relying on randomized controlled trials, which are inappropriate for this context. As was mentioned, researchers tend not to be directly involved with individual programs or regions. Rather, researchers use program evaluation data to find trends over time or across populations. Thus, they tend to perform much more complex statistical calculations than people in either public health or public policy, and have a great deal more statistical training. As such, they act as resources for those in public health, and are often asked to answer statistical questions that are beyond the expertise of people who do not have as much statistical training [Seliske, 2005; Leatherdale, 2005].

Finally, people working in public policy use program evaluation data to make funding decisions and to put forth proposals for further funding to higher levels of government. Specifically, program evaluation data are essential for policy analysts so that they can assess which programs are working – and thus should be promoted, given continued funding, and used as models for others – and which ones are not – and thus may need to be reviewed further, modified, or have their funding cut altogether [Garcia, 2005]. Assessing if a program is working may be a simple matter of comparing outcome measures such as participants' quit rate or uptake rate to established benchmarks, or it might involve measuring less tangible metrics such as behavioural change. In any case, a more challenging task is assessing which characteristics of a program are causing it to succeed or fail. These characteristics are often hard to pinpoint, but it is critical to perform this assessment accurately so that the correct characteristics are promoted in successful programs and changed in unsuccessful ones [Manske, 2005]. Another important way in which public policy people use analysis data is to summarize, translate, and share results among stakeholders. For example, someone at a provincial organization might be responsible for reporting on the progress of several dozen local tobacco control programs. They must have the proper data to assess each program's performance accurately and, at the same time, have the analysis tools to be able to summarize and present this data in a meaningful way to someone who might not have any familiarity with the field of tobacco control, to say nothing of a particular regional program [Silverman, 2006]. Indeed, presenting analysis results simply and clearly is quite important, since "…decision makers tend to be rather busy individuals who have little time to spend reading through long documents. …graphs, charts, or even pictograms are often effective" [Sandiford, 1992].

In summary, public health professionals, managers and researchers are all involved in the program evaluation process, each playing their own role. For a sense of how they interact, see Figure 17.

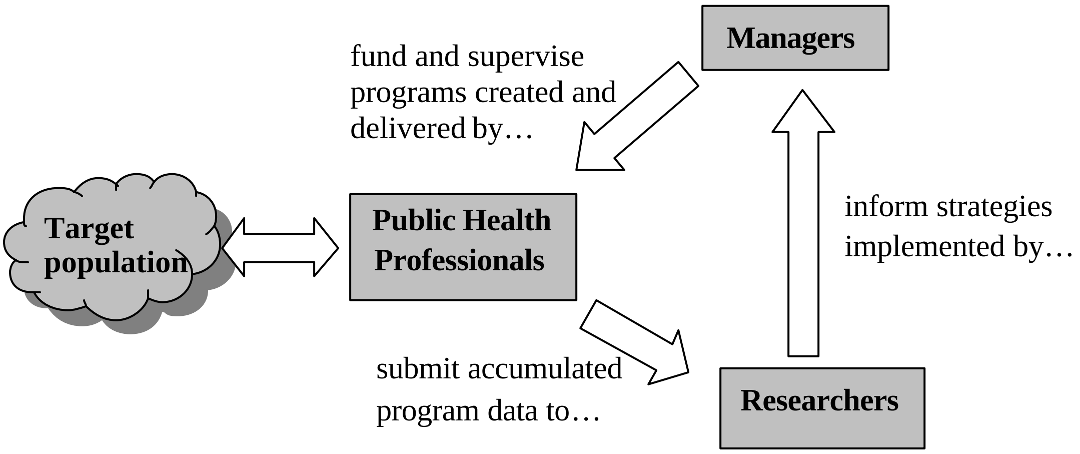

Managers
fund and supervise programs created and delivered by…
Target population
Public Health Professionals
inform strategies implemented by…
submit accumulated program data to…
Researchers

Figure 17: Interaction of Stakeholders

5.5 Impact of Improved Data Analysis

The fact that data analysis is an important part of program evaluation means that improvements in data analysis capabilities will have a strong impact on program evaluation, and thus on program delivery as a whole. Greater precision and efficiency in the analysis of public health program outcomes will result in improved assessment of those programs' fitness for their contexts, or other contexts. Furthermore, better information dissemination will help other communities use the information gathered in these evaluations to better assess the appropriateness of their own programs, and perhaps make changes accordingly [Morales, 2006]. In general, improved analysis capabilities will result in information which will be more likely to be relevant, while more widespread dissemination will result in more programs being implemented, and better targeting of those programs.

5.6 Current Solutions

Currently, in public health units, government departments, and university campuses, a wide range of technical and organizational solutions are being used to enable access to data and to the tools needed to analyze these data. These solutions vary widely depending on the needs of the users, the resources available and the emphasis placed on data analysis within the organization. This section describes two such solutions.

One pattern for data gathering and analysis is exhibited by research-driven surveys such as the Youth Smoking Survey (YSS). In this case, survey data are gathered from thousands of students at hundreds of schools across Canada and sent into a central processing office at the University of Waterloo, which hosts the study [Tiessen, 2005]. Data processing and analysis people on campus then take the raw data, encode it, analyze it using a statistical program (SAS), and produce individualized reports for each school using spreadsheet and word-processing programs, such as Microsoft Excel and Word. These reports are then sent out to the participating schools, as well as potentially interested public-health units and government agencies [Vandermeer, 2005].

Another means by which data is analyzed is through online tools such as the Non-Communicable Diseases Surveillance Infobase (NCDSI), offered by the Public Health Agency of Canada (PHAC), or the Health Indicator Query system (HIQ), offered by the Health Information for Ontario program, and the Rapid Risk Factor Surveillance System (RRFSS), which is a survey conducted by public health units in Ontario. Each of these tools offers a user-friendly Web-based interface for analyzing data sets, and thus offers some of the same functionality as WebStats [Puchtinger, 2007]. However, the tools have several serious limitations. One is that each tool contains a static collection of data sets, which limits both the user's ability to draw comparisons between data sets and the researcher's ability to share data sets. For example, the HIQ system uses only data collected from Ontario public health units between 1996 and 2003 [Health Indicator Query, 2007]. Similarly, the tobacco control-related variables that can be analyzed using each tool are quite limited, since these tools are intended for users working within the broad context of population health, rather than the specific area of tobacco control. For instance, the NCDSI offers only one variable, "Current daily or occasional smoker", under the data category "Smoking" [Non-Communicable Diseases Surveillance Infobase, 2007]. Finally, none of these tools have any capacity for storing or sharing analysis results, which means that the potential for collaboration is quite limited.

5.7 Summary

Tobacco control is a serious issue, both from a population health perspective and from an economic standpoint. While smoking rates have declined over the past decade, more work is needed in order to ensure that this trend continues, particularly in preventing teenagers from taking up smoking, and in high-risk communities and sub-populations. One of the most effective ways to effect change on a population scale is through carefully targeted population-based interventions.

Each intervention must be evaluated to assess its effectiveness, and thus to determine if it should be continued and attempted in other communities, or if it needs to be modified or cancelled entirely. The subfield of tobacco control which deals with this assessment, program evaluation, is highly dependent on accurate, timely data on the programs being evaluated, the characteristics of the communities in which the programs are being run, and comparable baseline data on similar communities and programs. Furthermore, program evaluation is undertaken by people from a variety of organizations, locations and roles; while these people share the broad principles outlined earlier, their specific goals may vary widely depending on their responsibilities and the context they are working within.

Currently, program evaluation is undertaken using a mix of spreadsheet programs and statistical packages, with communication between various stakeholders occurring on an ad hoc basis. As will be shown in the following section, there are several major problems with the status quo, and WebStats seeks to address these issues through technological and organizational improvements.

Chapter 6

Implementation of WebStats

This chapter describes the process of moving from a statistical analysis model, described in Chapter 3, and an understanding of a particular user population, described in Chapter 5, to a functional system. The chapter begins by identifying drawbacks in the statistical analysis tools and procedures currently used in the tobacco control field, which provides motivation for the creation of a new system for this context. Section 6.2 analyzes tobacco control users according to the User-type – Use-type Model, which serves two purposes. First, it offers a detailed look at the people who will be using the system, and the ways in which they will use it. Second, it identifies characteristics of these users which can then be applied to other contexts. For example, public health professionals, as a group, have limited statistical training; designing WebStats to account for this factor means that it will be usable by other user groups with a similarly limited background in statistics. Section 6.3 describes the functional requirements for WebStats that were identified through interviews with tobacco control professionals. Section 6.4 identifies several non-functional goals that were identified, such as improving the community of practice in tobacco control. Section 6.5 describes some of the changes that were made as a result of several rounds of interviews and demos of the prototype – that is, features which were added as a result of user feedback, or which users identified as needing modification. Section 6.6 describes in detail the implementation of several key system features, such as the statistical advisor and the user preferences system. Finally, the chapter concludes with a summary of the feedback users provided on the final version of the system.

6.1 Drawbacks of Current Solutions

6.1.1 Isolated, Out of Date Data Sets

One major problem with the current situation is that there is no unified source that interested parties, such as researchers, high school principals, and public health programmers, can access to find data or to post new data. One tobacco control researcher described the situation as trying to work with islands of data. That is, there were results from numerous surveys that were theoretically available to him, but it was difficult to make connections between the various survey results [Leatherdale, 2005]. Thus, it was difficult to establish trends beyond a single survey, that is, to establish trends across provinces or over a period of time. This issue of data isolation is also significant for people in tobacco control programming roles. Making connections between regions where particular programs have been successful and other regions in need of these sorts of programs would be quite valuable but requires easy and comprehensive access to a wide range of survey data [Cowan, 2005].

Another serious issue for public health programmers is analysis lag. Often, there will be a significant delay between when a survey is administered to a population and the raw data is sent to researchers to analyze and when the analysis is made available to the participating population. This lag is partially due to the time required for the researchers to perform the analysis, of course, but is extended still further by lengthy delays between a peer-reviewed paper being submitted, accepted, published and presented. This process, which can often take six to twelve months, is often seen as a prerequisite for releasing any kind of analysis of the data to the participants in the study. However, the kinds of analyses that the researchers are performing are often at a much higher level than what is needed by people at the regional or even provincial level. If people at the regional and provincial levels could have access to the raw data and tools to do basic analyses on the data, they could much more quickly realize the analytical value of the survey at the program implementation level [Jolin, 2005].

Another issue is that, because of the analysis lag mentioned above, researchers are commonly forced to work with data from older studies rather than being able to make comparisons within current report data. For example, a researcher working on a study performed this year will often have to use studies from several years ago as a baseline for comparing results. If users had access to up-to-date data, their comparisons would be more relevant, and thus more compelling, for policy analysts, public health researchers, and other interested parties.

6.1.2 Barriers to Statistical Packages

Another important issue is a lack of availability of data analysis tools. Researchers, typically, have access to high-powered statistical analysis tools such as SAS and SPSS through the universities or agencies with which they are affiliated, and have significant statistical training; thus, they are able to perform whatever analyses they see fit on their data. For people at the public health level, however, the situation is quite different. Tools are available based on the level of funding granted to a particular health agency which, in turn, is dependent on that agency's mandate and focus. In Ontario, for example, a handful of public health agencies across the province are designated as PHRED (Public Health Research, Education and Development) units, which means that they are granted an above-average level of funding for data management and analysis. These PHRED units, therefore, may be able to purchase licenses for advanced statistical analysis tools such as SPSS and SAS. However, most health units are not so fortunate and make do with spreadsheet programs such as Excel or Chart [Zimmerman, 2006]. As mentioned in Section 2, there are free alternatives such as R and Rweb. However, these tend to be less user-friendly than the point-and-click interface offered by spreadsheet programs, and are less well-known outside of the statistical and academic communities. Thus, for people with little formal training in computers or statistics, it usually makes sense to stick with a program, such as Excel, which is easy to use, well-understood and commonly used.

In either case, the tools at hand also require people with the proper training to use them, an issue which is actually more serious at PHRED units, since advanced statistical packages such as SAS and SPSS require a fairly high level of computer literacy and familiarity with statistical concepts. Often, questions requiring complex statistical analysis are passed along to epidemiologists for analysis, which solves the issue of access to the proper software and finding people with the proper training. However, this creates an analysis bottleneck, since there are often only one or two epidemiologists in each health unit, and they have their own analyses on which to work in addition to the assistance that they provide to other public health unit staff [Seliske, Leatherdale, 2005]. Thus, this is more of a band-aid fix to the issue of enabling analysis of information than a long-term, comprehensive solution.

6.1.3 Sociological Challenges

Even when the proper resources and people are available to process, analyze and disseminate evaluation data, there are organizational challenges to deal with, as well. Because of the close connection between the evaluation results and government funding decisions, evaluations are seen as pass or fail tests by the groups being evaluated, because evaluation results are used to make funding decisions. However, because of the separation between public health programmers, researchers and policymakers, people in public health see evaluation as something that is imposed by an external force, rather than a process that they are involved in.

6.2 User-type–Use-type Model

Covvey *et al* recently proposed the User-Uses-Effects model for defining EHR content [Covvey, 2003]. Part of this model is the User-type–Use-type model, which is used to define the users of a system and the ways in which they will use the system. This model may be applied to the requirements elicitation process for WebStats in order to gain a better theoretical understanding of its users and their uses of the system.

The users of any system will, of course, find myriad ways to manage, analyze, present and share data. However, users from each the three groups described in Section 4.1, public health professionals, managers and researchers, will tend to use the system in similar ways, according to the nature of their goals and responsibilities. According to the User-type–Use-type model, these groups can be modeled as User-types, and their uses of the system as Use-types. The following paragraphs describe the User-types and Use-types associated with each of these groups.

6.2.1 Public Health Professionals

A public health professional is mainly concerned with developing and running population-level interventions. Since a public health programmer's job is not analytical in nature, he typically does not have very much time or energy to devote to using statistical analysis tools. It should also be noted that his workplace, typically a public health unit, often does not have the financial or technical resources to supply him with statistical packages. However, his desire to create and deliver effective programs means that he has an interest in analyzing his programs in order to determine their efficacy for reducing the smoking rate in his region and to compare his programs to those being run in other regions. There is also a demand for basic analysis tools which can be used to create charts and simple calculations for inclusion in educational materials for the general public, schools, and workplaces.

In terms of a User-type, a public health professional may be characterized as a user with a minimal level of statistical training who performs mainly non-analytical tasks, and has little access to, or familiarity with, statistical packages. However, he occasionally performs simple analyses in his job, and has an interest in performing analyses more frequently, provided that he can use tools which match his skills and needs. As for his Use-type, a public health professional would use the system mainly to perform simple analyses, such as descriptive statistical calculations on a single data set. These analyses may be used internally, or published as part of materials distributed to the general public.

6.2.2 Managers

A manager is interested in analyzing data from tobacco control programs within the strategies she is guiding in order to determine the effectiveness of specific programs and entire strategies. She may then use these analyses in presentations to other government agencies, stakeholders and potential funding sources. While a manager typically has enough training in statistics and computers to make effective use of spreadsheet programs, such as Excel, and may have some degree of familiarity with standard statistical packages, she does not usually perform mathematically complex analyses. Rather, the complexity in her analyses will come from her desire to combine data sets, variables and results in a variety of ways to assess the effectiveness of programs and strategies relative to other programs, established baselines, and so forth.

More generally, a manager's Use-type is a user who has an intermediate level of training in analysis and data management, and uses analysis daily in her job. This user may have some experience with statistical packages, but is not an expert user of these systems. Her Use-type, like that of a public health professional, extends only as far as performing analyses. However, these analyses will be likely to incorporate multiple variables, multiple data sets, and more advanced statistical calculations.

6.2.3. Researchers

A researcher uses data from a range of sources to identify and address research questions or issues whose scope expands beyond a particular region, province or strategy. Since the researcher has his own set of advanced statistical tools that he is comfortable using, which may include standard packages such as SAS and SPSS, he would not need a system such as WebStats for performing analyses. However, the system could still be used by a researcher in other ways. One is as a communication tool to facilitate the fulfillment of requests for complex analyses from public health professionals, an issue discussed in Section 5.4. By providing a place where requests, data sets, and analyses can be posted, WebStats provides all the functionality necessary for a researcher to read an analysis request from a public health professional, perform the requested analysis, and make the result available to the person who made the request. Another is as a way to disseminate proprietary data sets after they have been used for the research purposes for which they were gathered. Often, a data set is compiled at the request of a researcher who would like to answer a particular question. This data set is the intellectual property of the researcher or, perhaps, the institution of which the researcher is a member, while he is performing the relevant analysis and going through the process of publishing the results. After publication, however, the data set has the potential to be made available in the public domain, where it can be used by professionals in the wider tobacco control community. The data set management component of the system has the ability to keep the researcher’s data sets private until he is ready to release them to the public, at which point they can be instantly accessed by anyone using the system.

As a general User-type, a researcher is a user whose job involves complex analysis on a daily basis; unsurprisingly, he has extensive training in statistics, and a high level of expertise using statistical packages. As an employee of a well-funded organization for which analysis is a priority, he has access to a variety of statistical packages. With regards to his Use-type, a researcher’s familiarity with, and access to, statistical packages precludes his use of a system that offers fewer statistical operations. However, his creation and frequent use of data sets means that he is likely to use WebStats as a tool for exchanging and disseminating data sets. His role as an expert resource person makes WebStats's ability to connect him with potential clients appealing, as well.

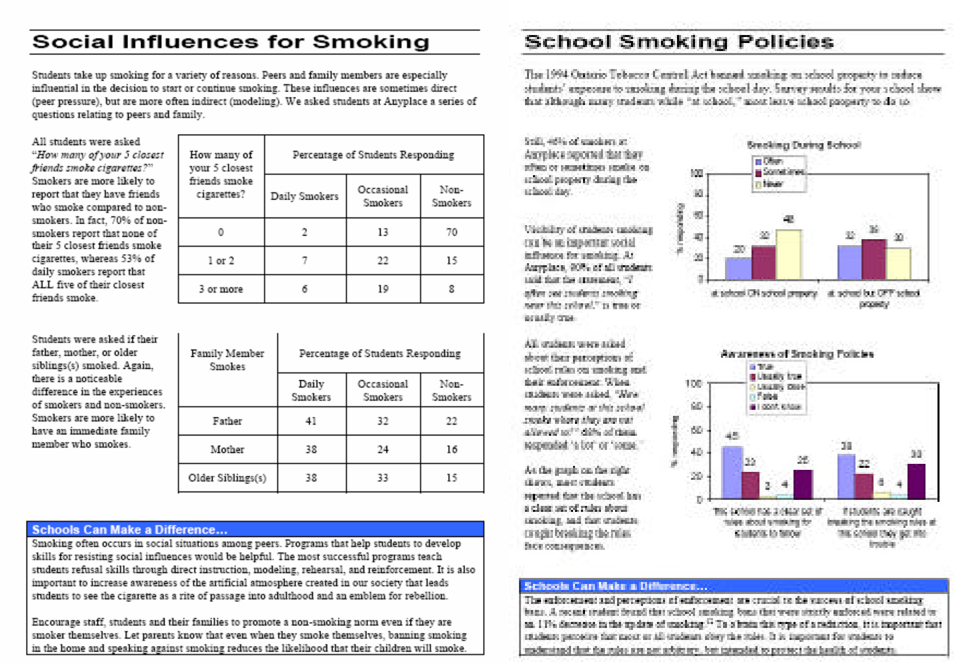
6.3 Requirements for WebStats

6.3.1 Basic Functional Requirements

At the most fundamental level, the system should allow the user to perform statistical analysis and visualization of results of raw data quickly and easily. Specifically, the system should provide the user with access to raw data from a variety of sources. The user should then be able to take that data and perform a statistical calculation, or a series of calculations, and then display the result either as text or in graphical form. These results may then be combined to create a composite graph or series of calculations.

Once these basic requirements had been determined, an issue that needed to be clarified was which operations, exactly, would be included in the system. Many of the stakeholders interviewed indicated that their statistical needs were mostly limited to creating charts and calculating descriptive statistics such as averages and standard deviations. These stakeholders performed inferential analyses rarely, if at all [Seliske, 2005; Barkley, 2006; Taylor, 2006; Zimmerman, 2006]. Thus, at first glance, it would seem to be sufficient to include just these operations. However, within the tobacco control community, the set of inferential statistical calculations that are performed vary from one role to another and, indeed, from one person to another. Thus, identifying a small set of common inferential operations was not feasible. As well, in the interest of making the tool useful to those with both basic and advanced statistical skills, it was worthwhile to present a representative sample of inferential statistical operations, so as not to exclude those with more extensive statistical training.

Another issue was the presentation of the analysis results. One option was to mimic the reports generated for each school that participates in the SHAPES program. These reports contain a mix of text and charts generated from data collected through the SHAPES program, with the proper analysis results, such as the percentage of students who smoke, inserted in the appropriate places. See Figure 18 for an example.



Social Influences for Smoking

Students take up smoking for a variety of reasons. Peers and family members are especially influential in the decision to start or continue smoking. These influences are sometimes direct (peer pressure), but are more often indirect (modeling). We asked students at Anyplace a series of questions relating to peers and family.

All students were asked *"How many of your 5 closest friends smoke cigarettes?"* Smokers are more likely to report that they have friends who smoke compared to non-smokers. In fact, 70% of non-smokers report that none of their 5 closest friends smoke cigarettes, whereas 53% of daily smokers report that ALL five of their closest friends smoke.

How many of your 5 closest friends smoke cigarettes?	Percentage of Students Responding		
	Daily Smokers	Occasional Smokers	Non-Smokers
0	2	13	70
1 or 2	7	22	15
3 or more	6	19	8

Students were asked if their father, mother, or older siblings(s) smoked. Again, there is a noticeable difference in the experiences of smokers and non-smokers. Smokers are more likely to have an immediate family member who smokes.

Family Member Smokes	Percentage of Students Responding		
	Daily Smokers	Occasional Smokers	Non-Smokers
Father	41	32	22
Mother	38	24	16
Older Siblings(s)	38	33	15

Schools Can Make a Difference…

Smoking often occurs in social situations among peers. Programs that help students to develop skills for resisting social influences would be helpful. The most successful programs teach students refusal skills through direct instruction, modeling, rehearsal, and reinforcement. It is also important to increase awareness of the artificial atmosphere created in our society that leads students to see the cigarette as a rite of passage into adulthood and an emblem for rebellion.

Encourage staff, students and their families to promote a non-smoking norm even if they are smoker themselves. Let parents know that even when they smoke themselves, banning smoking in the home and speaking against smoking reduces the likelihood that their children will smoke.

School Smoking Policies

The 1994 Ontario Tobacco Control Act banned smoking on school property to reduce students' exposure to smoking during the school day. Survey results for your school show that although many students while "at school," most leave school property to do so.

Still, 46% of smokers at Anyplace reported that they often or sometimes smoke on school property during the school day.

Visibility of students smoking can be an important social influence for smoking. At Anyplace, 80% of all students said that the statement, "I often see students smoking near this school." is true or usually true.

All students were asked about their perceptions of school rules on smoking and their enforcement. When students were asked, "How many students at this school smoke where they are not allowed to?" 46% of them responded 'a lot' or 'some.'

As the graph on the right shows, most students reported that the school has a clear set of rules about smoking, and that students caught breaking the rules face consequences.

Smoking During School
Often
Sometimes
Never
% responding
100
80
60
40
20
0
20
32
48
32
39
30
at school ON school property
at school but OFF school property
Awareness of Smoking Policies
True
Usually true
Usually false
False
I don't know
% responding
100
80
60
40
20
0
45
23
2
4
26
38
22
6
4
30
This school has a clear set of rules about smoking for students to follow
If students are caught breaking the smoking rules at this school they get into trouble

Schools Can Make a Difference…

The enforcement and perceptions of enforcement are crucial to the success of school smoking bans. A recent study found that school smoking bans that were strictly enforced were related to an 11% decrease in the uptake of smoking.[?] To attain this type of a reduction, it is important that students perceive that most or all students obey the rules. It is important for students to understand that the rules are not arbitrary, but intended to protect the health of students.

Figure 18: Sample SHAPES report

While this analysis result format is certainly appealing from the point of view of the person reading the resulting report, creating a Web-based document creation tool goes far beyond the scope of the project. At the other end of the complexity spectrum, the user could simply be presented with the results of the analysis in an image or table (in HTML format) and it would then be up to him to take that information and either incorporate it into an existing document, such as word processing file, or use it on its own. This would give users the greatest amount of flexibility, but would provide them with the least amount of assistance in turning the analysis into a coherent document. However, given the variety of stakeholder roles and responsibilities, the flexibility provided by this option is more important than the assistance, or lack thereof, provided by a more robust report generation tool. Thus, the version of the tool developed in this thesis uses the latter approach, while implementation of the former approach is left for future work.

6.3.2 Flexibility

The tobacco control community, as mentioned earlier, is quite heterogeneous, with members from a wide variety of disciplines, roles and backgrounds. Thus, it is crucial that the system be flexible enough to meet the needs of all of its users. For example, it should be simple enough so that someone with a limited background in statistics, little computer expertise, and not much time can log in and quickly do some simple calculations. On the other hand, it should also be robust enough so that a researcher, with years of experience, and a variety of statistical tools at his disposal, can find some utility in the system.

Specifically, there are three main areas in which users will vary widely, and the system should be flexible enough to meet each user's needs in each of these areas. One area is the user's training and resulting skill set. As alluded to earlier, users have different levels of training in statistics, computer literacy, program delivery, and program evaluation, to list some of the more readily identifiable skill sets. It should also be noted that expertise in one area does not imply expertise in another. For example, someone with extensive training in program delivery might be quite computer literate but have little statistical training. Thus, it is not sufficient to simply make simple and complex versions of the system and classify users as belonging to one group or the other. Another area is the environment in which the user is using the system, which is usually, but not always, the same as their work environment.

Two ways in which the environment can vary is in the computing and human resources available to the user, and the amount of time the user has available to use the system. Some users, such as public health nurses at PHRED units or epidemiologists, would have access to comprehensive statistical packages such as SAS and SPSS, while others would have access to only spreadsheet programs such as Excel. Human resources are similarly varied. At some public health units, there will be statistical experts on hand to help less knowledgeable co-workers, while at others workers will have to rely on their own, perhaps incomplete, statistical training. Finally, the amount of time the user spends on statistical analysis and, by association, using the system, may vary from a few minutes each day to several hours, depending on the nature of his job.

A user's goals will have a major impact on their use of the system, and will also vary between users. One kind will want to create public relations material, aimed at parents, students and the general public. Another kind will want to create budget presentations for meetings with managers and government officials. Thus, the level of statistical analysis, the way the analysis is presented, and even the formatting of the text will vary according to the user's goals and, more specifically, his intended audience.

To accommodate the extremes of the user spectrum, as well as all the users in between, WebStats uses a user preferences system which controls the user interface and the range of statistical operations offered to the user. This preference system is discussed in greater detail in Section 6.6.3.

6.3.3 Usability

WebStats's usability will be a major factor in determining whether or not it is actually adopted by its intended audience. To achieve usability, the most important user characteristic to consider is the user's familiarity with statistical concepts and packages.

For a novice user, WebStats should have an intuitive user interface, and should not require him to memorize a set of commands or specialized syntax. Thus, a command-based interface such as that used by R or SAS is unacceptable, as it has a steep learning curve, making it all but unusable for any novice computer user whose job involves statistical analysis on an occasional basis. Indeed, several of the stakeholders interviewed expressed a sense of intimidation and frustration when using command-based statistical analysis programs. Stakeholders also professed a general lack of training in statistics, which compounded the issue. Thus, the system should provide explanations of the user interface components and statistical operations that are made available, both to help users navigate the interface and also to facilitate their understanding of the statistical operations that they are performing. Thanks to WIDE's built-in HTML-based help widgets, these explanations can include sample graphs that each operation would generate, such as a graph with multiple sources as compared to a graph with a single source.

For an advanced user, the guidance provided by WebStats should not create a barrier for or otherwise hinder the user in being able to use WebStats quickly and effectively. To ensure that the appropriate level of help is provided, the tool allows the user to decide how much assistance the system should offer him as he navigates through the menus and performs statistical analyses. On a broader scale, the interface as a whole should be enabling rather than restrictive; that is, an advanced user should not feel that WebStats is preventing him from accomplishing his work because of simplifying assumptions made for novice users. Other features that an advanced user would find valuable include the ability to use WebStats alongside a more complex statistical analysis program such as Stata or R, the inclusion of complex inferential statistics, the ability to override the sorts of statistical safety checks that would be enabled for novice users, and the ability to add custom calculations and graph types.

There are also a few issues that should be considered independently of the user's expertise in statistics. Any users, no matter how extensive their training, needs to make use of WebStats's help system at some point, so some thought must be put into the documentation provided by this help system. This documentation should, first and foremost, be explanatory, and it should use appropriate terminology. Second, it should not assume any particular background in statistics or computers, to avoid alienating users who have little statistical training or poor computer skills. A user with advanced computer skills, meanwhile, might appreciate the ability to customize the user interface according to his preferences, or wish to integrate WebStats with other WIDE components, such as a workflow engine.

6.3.4 Availability

Having a single system that users from all areas of the field could use would be quite beneficial, given the multi-disciplinary nature of the tobacco control field. However, many people working in tobacco control are employed by non-profit organizations or small government agencies, such as public health units, which have limited budgets. Licensing fees, such as those charged by Microsoft or SAS, are prohibitive for these organizations. Thus, to ensure that people working in these organizations can use WebStats, it should be available for free or at a low cost. Furthermore, the dispersed nature of the organizations working in tobacco control means that there will not necessarily be any uniformity in the computing environment. Thus, WebStats should be usable across the three major platforms – Windows, Apple and Unix – and with a wide variety of computing systems.

6.3.5 Collaboration

Program evaluation involves professionals from a wide variety of disciplines. Thus, it was important that WebStats would not only be usable by a variety of people in isolation but would also help dispersed colleagues work together towards a common goal. The most basic requirement is to ensure that users can easily share information at every step of the evaluation process – raw data, analysis results, visualizations, and any other statistical results. However, because of the distributed nature of the work environment, it is quite likely that people from one organization will not know everyone – or, even, anyone at all – in other organizations. Thus, the system should also help users network with people in other disciplines, organizations, and jurisdictions.

As alluded to in Section 5.5, the more quickly and widely program evaluation results are disseminated, the greater effect these results will have on program implementation. Thus, in addition to enabling collaboration among colleagues in analyzing data, WebStats should also make information dissemination as efficient and effective as possible. Specifically, is desirable that the raw data collected by organizations such as public health departments and the Lung Association and assessment programs such as SHAPES and YSS be made available to researchers so that they can perform analyses and draw connections using the most relevant and recent data possible. These analyses, in turn, could be disseminated to public health programmers, program managers, and champions in the education system so that decisions may be made with up-to-date information. The reports based on these results, finally, could be sent out to parents, school principals, student leaders, and the general public.

6.3.6 Interpretation of Inferential Statistics

People working in the field of tobacco control come from a variety of backgrounds, such as nursing, epidemiology, and program management. Their level of training in statistics varies widely. For some, statistics is a crucial part of their jobs, and they have extensive training at the graduate level. Many others have taken one or two social sciences-oriented statistics courses in the course of earning their degrees, and they now deal with statistics occasionally. There are also those with no statistical training whatsoever who are asked to create promotional materials, information packages and the like, and they are forced to do their best with a bewildering array of statistical terminology and tools.

Because of the inherent uncertainty involved in inferential statistics, the novice user must also be given guidance after he has finished his statistical analyses. Without the proper understanding of the context within which an inferential statistical operation is calculated, it is quite easy to draw an incorrect conclusion from the calculation. Perhaps the most readily apparent example is the ubiquitous *p* value. The *p* value, roughly speaking, indicates the strength of the evidence in the data against the null hypothesis, which is the opposite of the hypothesis one is trying to prove, called the experimental hypothesis. Specifically, $p = 0.001$ indicates that there is strong evidence against the null hypothesis – and thus strong evidence *for* the experimental hypothesis. A *p* value = 0.10, meanwhile, indicates that there is little evidence against the null hypothesis, and thus little evidence for the experimental hypothesis. The *p* value is usually compared to the level of significance (denoted by the Greek letter a) to determine whether a statistical result is significant. When calculating inferential statistics, many packages simply give this *p* value with no indication of its meaning, which means that a novice user could easily overlook or misunderstand this indicator of a calculation's significance. Another example is the close relationship between the sample size, confidence level, and confidence interval. In short, the size of the sample has a direct impact on the width of the interval one can use with a high degree of confidence. As the sample size decreases, one must either shrink the width of the interval or lower the degree of confidence. Otherwise, the inference is invalid. Thus, WebStats should present not just the results of inferential calculations and the appropriate measures of confidence or accuracy but also an explanation of these measures, so that the user has the ability to correctly interpret the results. Finally, statistical significance does not always imply practical or, in a medical context, clinical significance. In a large sample, a difference may be large enough to be statistically significant but have very little practical impact. For example, research might show that a particular program has reduced the smoking rate in a population by 1% over ten years. While this change may be significant, statistically speaking, it represents a very small effect on the population.

6.3.7 Data Privacy and Ownership

An important set of considerations, which run counter to the goal of information dissemination, are privacy and ownership. In the tobacco control field, these considerations are quite important, since survey data is both private, as it may contain sensitive information about respondents, and subject to intellectual property concerns, as researchers may have spent a great deal of time and effort to perform statistical analysis on a set of data and would not want that information disseminated without their consent [Leatherdale, 2005]. However, it is not sufficient to simply restrict access to a data set to its owner. There will likely be cases in which a researcher will want to share a data set with his or her colleagues, for example, a professor and a group of graduate students, but will not want to make the data available to the general public. Thus, the owner of the data must have precise control over who is able to view, analyze, and disseminate his data.

Another related issue is that of identifiability. In many cases, such as anonymous surveys, the participants is assured that he will not be identified in the analysis results. As surveys are often administered anonymously, this would not seem to be an issue. However, if the person using the survey data is permitted to analyze the data in such a way as to drastically restrict the set of participants, then a participant's identity may be inferred. For example, there may be only one student at a small high school in a particular grade who is a member of a particular minority group, and thus restricting the analysis of that school's data set to that grade and minority group would, in effect, identify the student. To address respondent privacy issues, there should be safeguards on the presentation of data, so that in a situation in which a respondent's anonymity would be compromised, the system would only show aggregate data rather than individual survey responses. This concern may occur at an organizational level, too; schools or school boards that provide access to data on a specific population may not wish to be identified in a larger set of data. Thus, care must be taken to preserve confidentiality at this level, too [Manske, 2005; McDonald, 2006].

6.3.8 Data Set Information

When storing and managing a data set, there are many different kinds of information about that data set that must also be stored, aside from the data themselves. In addition to basic information such as which variables are used and how many items are in the data set, there is often specific contextual information about how data in a particular column have been collected, how a particular characteristic was measured, or what was done with invalid data. This contextual data can be essential in determining if two variables from different data sets may be compared with each other. For example, two surveys might use very different criteria to determine who is a regular smoker, making comparisons between the two surveys invalid.

Second, data within a set may be weighted to achieve the proper relationship between the sample and target populations. For example, if the target population for a survey is evenly divided between males and females, but the sample population has 25% males and 75% females, then the survey responses from males should be given a weighting of 2.0, while the survey responses from females should be given a weighting of 2/3. This example is, of course, very simple, and weightings can incorporate several different variables simultaneously. In general, the use, or lack thereof, of a weighting scheme can have a huge impact on the analysis of a data set and the interpretation of that analysis [McDonald, 2006].

Finally, a data set may use aggregate variables in addition to the variables determined by the raw data alone. For example, a survey of smoking habits among youth might include a range of questions about each respondent's likelihood of smoking in certain situations, social influences, his previous smoking behaviour, and the like. Based on his answers to these questions, the data set might include a smoking-risk variable, which would take all of the aforementioned questions and calculate an overall measure of the likelihood that the respondent would take up smoking within the next year.

6.4 Secondary Goals

WebStats hadalso a number of broader, not very quantifiable goals. First, it is hoped that by public health programmers' using WebStats, they will be more self-sufficient and will be able to answer more of their own questions, rather than having to rely on epidemiologists and other statistical experts for assistance. This self-sufficiency will be a benefit for both the programmers and the epidemiologists. The programmers will be able to do statistical analyses much more quickly, as they will be able to find the answers themselves in a matter of minutes, rather than waiting for someone else to do the analyses for them. The epidemiologists, meanwhile, will have fewer requests for simple analyses from programmers and will thus be able to spend more time on their own research, as well as handling more complex requests from programmers.

Second, making WebStats available to people in a variety of roles within the tobacco control community will allow research, policy, and public health groups to share data and, more importantly, to work together on projects rather than pursuing different goals, thus strengthening the community of practice.

Third, giving people who run tobacco control programs an easy-to-use, accessible analysis tool will allow the people at front lines to get past the evaluation stigma by doing their own evaluation and being a part of the process, rather than having the data evaluated by some external organization such as a government agency.

6.5 Iterative Design Process

WebStats's stakeholders were consulted throughout the design process. This helped ensure that WebStats met its requirements, and allowed the stakeholders to refine their requirements as WebStats progressed from a rough prototype to a functional system. There were a few areas in particular in which the feedback had a significant impact on the design and implementation of the system.

6.5.1 Presentation of Operations and Analyses

One area of the interface that saw a significant change was the menu for managing analyses and operations. Initially, WebStats presented the analysis management operations, such as creating a new analysis, or adding an operation to an analysis, as a part of the operation creation interface. However, this presentation format did not offer a complete range of operations – for example, there was no ability to delete an operation from an analysis – and users found the interface quite confusing. Many did not understand why they needed to create an analysis, and were not sure when they were working with an operation and when they were working with an analysis.

A revised version of WebStats separated analysis management operations from the operation creation interface entirely. When performing a complex analysis, the user is first asked to create or load an analysis, and is subsequently allowed to add, edit, or remove operations from the analysis, and may also edit or delete the entire analysis, or add constant values to the resulting graph or calculation result. This revision clarifies the distinction between an analysis and an individual operation and also allows the user to add, modify and delete both operations and entire analyses quickly and easily.

6.5.2 Wording of Help Text

Initially, the help text was written using mathematically precise language, which reflected the statistics texts from which the definitions of statistical operations were taken. Interviews with tobacco control professionals, however, revealed that this kind of language is not considered clear or understandable by those users. While the users have a great deal of experience with basic data analysis techniques, they are not as familiar with the mathematical terminology used by statisticians to define these analysis techniques. As a result of this feedback, much of the help text used in WebStats was re-worded so that it would be both mathematically correct and helpful for the WebStats`s users.

6.5.3 Analysis Visualization Options

One group of users which was particularly helpful in the refinement of WebStats was people working at the Ontario Tobacco Research Unit (OTRU), an organization which monitors programs conducted within the Ontario Tobacco Strategy and disseminates information for the research and public health communities [OTRU, 2007]. They saw the potential for WebStats to be integrated into a program evaluation management system that they were developing, and offered detailed feedback on WebStats`s functionality. In particular, they identified a number of customization options which could be integrated into WebStats in order to provide the user with more control over the presentation of graphical results. For example, the ability to label data points on a graph according to a variable value, as opposed to a default index label, was quite helpful in displaying certain types of charts.

6.5.4 Data Set Search Tool

Initially, the data set viewing component included in WebStats was quite simple: it presented all of the data items in a given data set. However, users identified a range of ways that data might be viewed, including viewing a subset of a large data set, viewing a specific set of columns, and searching the data set according to a particular set of criteria. These uses motivated the creation of a component which could both view and search a data set, which was used in place of the data viewing component included in the WIDE system.

6.6 Implementation Details

6.6.1 Storage of Operations and Analyses

As was discussed in Section 3.2, the statistical compilation engine allows the user to group multiple operations into a single analysis that will be presented on a graph. To achieve this grouping, a set of database tables was created to store complex operations. The diagram in Figure 19 shows the details of this set of tables.

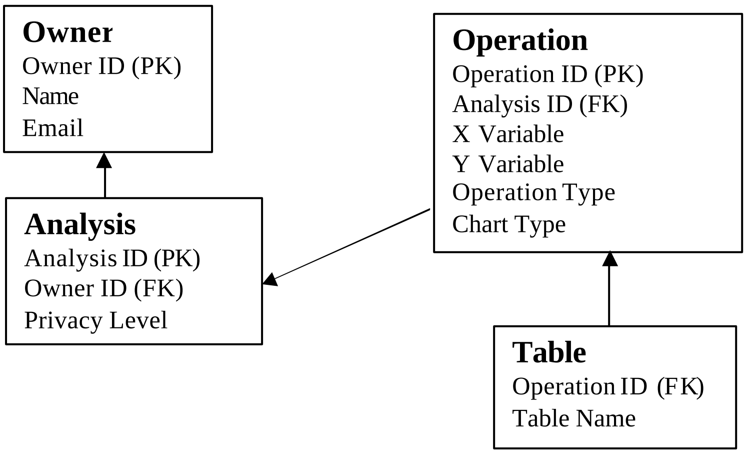


Figure 19: Database Schema

This scheme meets two important requirements: it allows all operations to be stored and retrieved quickly and easily, and it allows the user to create an analysis which consists of several operations presented on the same graph.

6.6.2 Dynamically Updating Menus

One implementation detail that proved to be somewhat complex was the dynamic updating of menu items to reflect currently available options, given other menu selections. For example, on the operation creation page, when a user selects his data set, the list of available variables should change to reflect those contained within the selected data set. At first, the implementation of this feature was attempted on the client side, with JavaScript and a set of hidden menus, which were used to dynamically populate the menus that needed to be changed. However, this approach did not prove flexible enough to handle all possibilities. Specifically, when the user selected a data set, the client-side JavaScript code needed to have all the variables for all available data sets pre-loaded on the operation creation page. This pre-loading was incompatible with allowing the user to add and remove data sets from the system. Instead, the operation creation page is automatically refreshed, that is, updated from the server side, whenever a menu option is changed, allowing WebStats to query the database for the information appropriate for the selected menu options. Thus, the page contains only the information relevant to the user's selections, and WebStats can take advantage of the organization implicit in its database. A flowchart of this update process is shown in Figure 20.

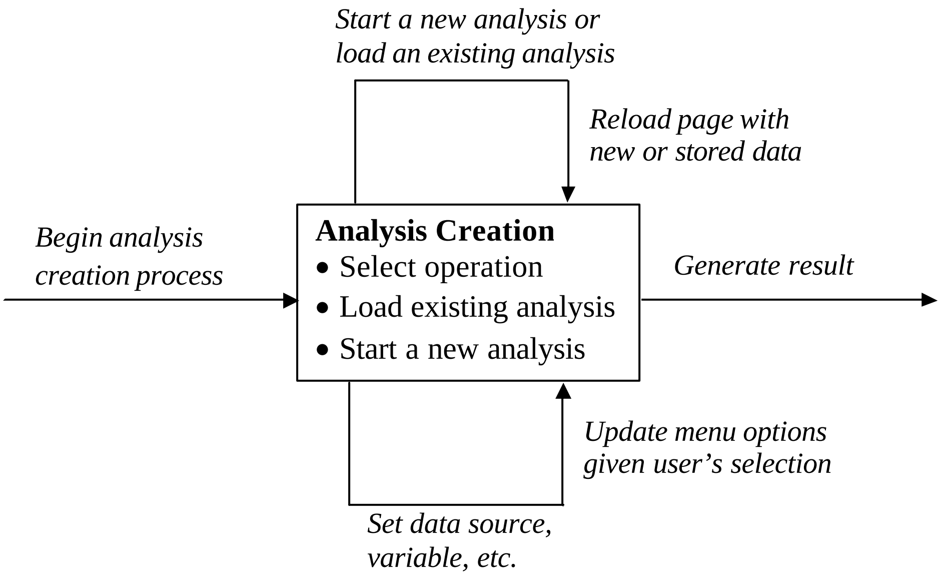

Start a new analysis or load an existing analysis
Reload page with new or stored data
Begin analysis creation process
Analysis Creation
• Select operation
• Load existing analysis
• Start a new analysis
Generate result
Update menu options given user's selection
Set data source, variable, etc.

Figure 20: User Interface Dynamic Update Method

6.6.3 User Preferences

As was mentioned at the end of Section 6.3.2, in order to improve WebStats's flexibility WebStats makes use of a set of user preferences, which are divided into three categories: statistical familiarity, computational complexity, and interface detail. Within each category, the user can choose a setting of "high", "medium", or "low". The user's selection will affect the range of statistical operations available, the adjustments which can be made to those operations, and the variety of customization options presented. The user can set these options either via a "Personal Settings" page or by using a guided "Create User Profile" wizard. Shown in Figure 21 is a chart that shows the effect that each setting has on the system.

Category	Setting	Effect
Statistical Familiarity	**Low**	Descriptive operations hidden Inferential operations hidden Help provided for calculation results
	Medium	Inferential operations hidden
	High	All operations shown
Computational Complexity	**Low**	Where clause hidden Second variable hidden Cumulative, absolute value graph options hidden
	Medium	Where clause hidden
	High	All operations shown
Interface Detail	**Low**	Collection options shown: name and privacy level
	Medium	Additional options shown: image, axis dimensions
	High	All options shown

Figure 21: Table of User Preferences

6.6.4 Keywords

One aspect of WebStats that is of particular value for groups of people dispersed across organizations or environments is WebStats's ability to organize and search for data based on concepts rather than on a particular naming convention or organizational standard. This ability is achieved in WebStats through the use of keywords which are applied to analyses, data sets, variables, and user profiles. To achieve this organization requires, however, that the user label items with the appropriate keywords. Labelling could become burdensome for the user if the responsibility for labelling was placed solely on the user. In an effort to avoid, or at least lessen this, the system uses cascading keywords in the following manner: a basic unit of information, such as a variable or data set, is given one or more keywords based on its areas of relevance. For instance, a survey given to high school students in Ontario might be labelled with the keywords "Youth smoking", "High school" and "Ontario", while a question in such a survey asking whether a student's father smokes might be labelled with "Family influences". Statistical analyses that use these variables and data sets, in turn, inherit these keywords. Thus, the user does not have to assign keywords to analyses, since they are transferred from the constituent parts of the analysis.

6.6.5 Information Privacy

Another important issue that was explored in Section 6.3.7 is privacy. The traditional concept of user permissions may be expanded slightly to create information source permissions. That is, a particular data set will have a privacy level which restricts users' access to and ability to modify that data. This is somewhat similar to the permissions system used in a standard database management system (DBMS): the owner of the data set is able to view and modify the data as he pleases, and can assign privacy or ownership level to the data, which will then control who else could view, modify, or analyze those data. One such scheme is shown in Figure 22:

Privacy Level	Summary Shown?	Details Shown?	Editable?
Public	Yes	Yes	Yes
Read-only	Yes	Yes	No
Private	Yes	No	No
Hidden	No	No	No

Figure 22: Permissions Scheme

This privacy scheme, however, is only one example of the kinds of privacy schemes that can be implemented. For a sense of the robustness possible with such a system, consider the following example.

Currently, survey data collected through the SHAPES program (See Appendix C for a description of the SHAPES program.) is kept private through a fairly rigorous protocol. A person outside the SHAPES program who wants to use the data must make a formal request to the Office of Research Ethics at the University of Waterloo. Once this request has been granted, he is permitted to view certain basic information about the data collected, such as the number of respondents, basic data trends, and so forth. Performing complex analyses and directly accessing raw data, however, are still prohibited. This policy could be implemented through the WIDE system by linking data access rights to the completion of a data access permission workflow, which might involve filling out forms online or on paper and waiting for official approval. Once this process is completed, the privacy level of the SHAPES data set would restrict the tools shown by WebStats to those that would allow the user to perform the approved analyses.

6.6.6 Help Text

Users need assistance in using WebStats, and this assistance is provided in three ways. First, help with navigating the user interface is provided via built-in help widgets that appear alongside most of the menus shown in WebStats. Clicking on any of these icons, which are speech bubbles with question marks in them, brings up a small pop-up window containing an explanation of how to use the given menu item. In particular, the help text provided with the statistical operation menu provides a definition of each operation, an example of how it might be used, and a description of the operation in plain English.

Second, if a user has selected a the low statistical familiarity level, then WebStats provides a link along with any calculation result that describes what the result means and gives an explanation of the operation used to generate that result. This information is included to ensure that a novice user, perhaps not noticing the help icons beside the menus, or assuming that he know what he’s doing, will not be confused by simply getting a numerical response to his query and will have the knowledge to put the result in context.

Finally, there is a PDF document available from within WebStats that is a complete reference guide, fully explaining every option, menu, and result provided by the system.

6.6.7 Statistical Advisor

An important component of WebStats is its ability to provide the user with statistical guidance and advice and thus help him perform analyses which not only are valid but also make sense statistically speaking. However, in seeking to implement such a component, several roadblocks were encountered. First, there is no simple way to determine the modality of a data set, that is, whether it has one central point or several of them. Although humans can easily spot central points, it is difficult to write an algorithm to do so. Thus, the initial idea of automatically determining a data set's modality and providing appropriate advice was discarded. An alternative, proposed by two professors in statistics, was to present a histogram for the data set alongside histograms for several different example modalities and let the user determine the data set's modality. The user would then be offered advice for the selected modality [Brown, 2006; Thompson, 2006].

Second, offering advice on the proper statistical procedures requires information not only about the data set's distribution, but also about the characteristics of the data themselves. One must determine whether the data being analyzed are interval, ordinal or nominal, continuous or discrete, as well, as fitting a host of other categorizations, before one can determine if a bar graph is more appropriate than a line graph, or a frequency table is better than a raw data table. Some of these characteristics may be determined automatically, but others must be entered manually by the person responsible for creating or maintaining the data set. Thus, the soundness of advice cannot be based on WebStats alone. It depends on WebStats's users, which means that the advice is influenced by the very people for whom it is intended.

However, assuming that the characteristics of the data set have been entered correctly, then there are established decision-making processes that may be used to determine the best way to view data. Figure 23 contains a flowchart for analyzing data taken from a medical-statistics context [Whitley, 2002]. Although the examples may not be familiar, the process of deciding how to analyze and present data is the same for any context, and thus this provides a good starting point for the advisor.

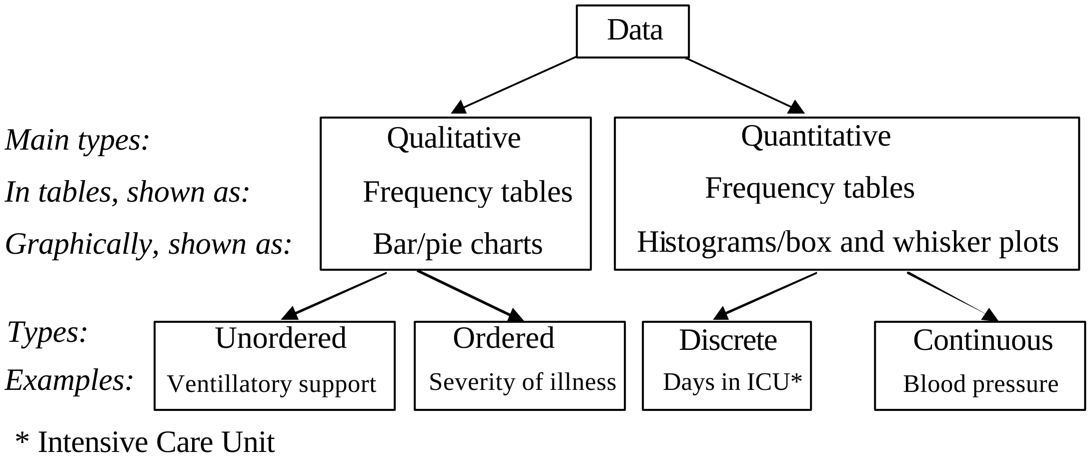


Figure 23: Flowchart for Analyzing Data

It is also important to recall a requirement, mentioned in Section 6.3.3, that the user interface not hinder the user. A prime example of a hindering user interface feature is Clippy, the helper added to

Microsoft's Office 97 suite. Although designed using artificial intelligence techniques such as Bayesian networks in an effort to provide the user with relevant help at appropriate times, users quickly grew irritated by Clippy's attempts to help them write letters and print documents. Clippy was retired by Microsoft in 2001 with a tongue-in-cheek advertising campaign that acknowledged the infamous reputation of the paperclip [Microsoft, 2001].

Finally, it must be emphasized that many of these guidelines, while often useful and reasonable, are rules of thumb rather than inviolable rules. There are situations in which the rules may not apply, and the user should be given enough flexibility to be able to ignore the advice.

6.7 Outcomes

After the implementation phase was finished, the stakeholders were once again asked for their feedback on WebStats, but this time to evaluate it. They were asked to identify the aspects of the system that they felt were particularly important, the reasons for or against WebStats's adoption by the tobacco control community, and what they saw as the advantages or disadvantages of the system as a whole.

In general, the feedback was quite positive, on several levels. With regards to the system's functionality, one person said that he was impressed with what the system could do, and felt that the backbone of the system was in place. Another was quite happy with the statistical advisor and message board components of the system, and particularly with the ability to associate analyses with message board postings. He felt that the concerns he expressed during the development process had been taken into account in the final version of the system. Another user commented that while having a broad range of options was certainly helpful, what I had done a particularly good job of was including explanations for these options, so that users could have an understanding of the tools that they are using [Garcia, 2007; Leatherdale, 2007; Pathammavong, 2007].

With regards to the tool's adoption by the tobacco control community, respondents were able to envision several groups making use of the tool. One user said that the tool would be "perfect for public health people who want some analysis functionality, including planners in public health." Another respondent talked about a particular group, an asthma education team at an NGO, who would find the tool useful in analyzing information about workshops they deliver. Instead of asking an outside statistical resource person to do their analysis, they could perform the analysis themselves, saving both time and resources. A project manager, similarly, could digest numbers and create his own charts "without having to worry about SPSS". A third user discussed a scenario in a high school context, where a principal could use the system in collaboration with a health promotion officer from a public health unit to identify problem areas at a school and come up with strategies to deal with the issues. Since a principal is not a statistician, the system's use of charts is quite helpful, and helps the principal talk about the issues at hand. This person also pointed out that a public health professional, who has expertise in program development and working collaboratively, often needs to portray data clearly and simply; thus, this tool would be an excellent fit for her. In general, the user was very supportive of WebStats's basic analysis tools, felt that the tools allowed many different people to participate in analysis, and believed that the system as a whole would be very helpful [Leatherdale, 2007; Loughead, 2007; Pathammavong, 2007].

In terms of the tool’s overall impact, respondents identified a few different areas where the tool would be beneficial One saw the tool as a big improvement over the status quo, in that more people can take part in analyzing data, and it reduces the ambiguity and scariness factor in program evaluation. Another saw the system as a solution to a problem of inefficient use of data; a lot of time and money goes into collecting raw data and, whether because of poor dissemination or privacy concerns, those data are often used only to answer a single research question. This tool allows the raw data to be used by many people to answer myriad questions. Another user raised a similar point, saying that the tool made raw data available to interested people and, in effect, provided analysis on demand. Yet another user talked about the tool’s potential for breaking down the statistical barrier faced by people with little statistical training. The tool’s emphasis on basic analysis meant that it could be used by people who felt that anything to do with statistics was out of their league. The next step will be to provide training and tutorials so that any user can see for himself that he is, in fact, capable of using the tool[Garcia, 2007; Leatherdale, 2007; Loughead, 2007; Pathammavong, 2007].

In addition to these positive comments, users also identified several areas for further work. Two people discussed the need for further pilot testing within a specific project such as OTRU or SHAPES, which would allow the system’s user interface to be better tailored to public health professionals. They felt that the user interface, while functional, was not clear enough for that user group. For example, one respondent commented that they would be likely to overlook functionality that was only available as a hypertext link, as opposed to a large, clear button. Similarly, another respondent recommended that browser-based functionality, such as the ability to print and save images, should be made available from within the system, as well. This would help users who, for example, might not realize that they can print an image via their browser get the maximum use out of the Web-based system [Garcia, 2007; Leatherdale, 2007].

Chapter 7

Conclusions and Future Work

7.1 Conclusions

The review of existing statistical packages given in Chapter 2 served two purposes. First, the evaluations of statistical software conducted over the last thirty years established a set of criteria to use in judging today's statistical packages. Second, the comprehensive review of statistical software, with a focus on collaboration, usability across organizations, and suitability for novice users, established that there is indeed an unfilled need in the general user community for the kind of system described herein.

The model for the statistical toolkit described in Chapter 3 has several important benefits. First, its integration of data storage, analysis, statistical advice, dissemination, and import and export components gives it a robustness not seen in other Web-based statistical models. Second, its modularity allows individual components to be modified or replaced entirely without affecting the rest of the system. Finally, the fact that it is built around a database of analysis information means that additional components can make full use of a database to enhance the capabilities of the system.

As was shown by the feedback throughout the development process and the response received from stakeholders at the end of this process, WebStats met the requirements of its users, filling a need for a usable, collaborative statistical analysis tool which was not met by existing tools and systems. In gathering the requirements for the system, implementing and refining the system, and evaluating the results, several important lessons were learned.

The first lesson was that even in a relatively small target user population, people working in tobacco control, there was a huge amount of variation in statistical training, reasons for using the system, and environments in which the system would be used. Thus, design decisions could not be made based on the characteristics of the average tobacco control user. Instead, flexibility became an essential requirement. To meet this need, WebStats includes a user-preferences system that allows a user to customize WebStats according to his needs, abilities, and level of statistical training.

The second lesson was that it was not sufficient to simply present the user with all possible statistical operations and hope that he would select the correct combination of data sets, variables, operations, and graph types. Some guidance is required to ensure that the user knows what he is doing, and does not create statistically invalid analyses. This guidance is achieved, in part, by restricting the available operations to mostly descriptive statistics. However, even within the realm of descriptive statistics, errors can be made. Thus, WebStats also includes a statistical advisor, a reference guide, help for every option, and explanations of calculation results.

The final lesson was that WebStats needed to be more than just another analysis tool. WebStats needed to recognize the importance of collaboration and information dissemination. That is, there was strong demand for a system that would make the results of any analysis available to colleagues on-demand, allowing colleagues to access any analysis without the creator of that analysis sending it to the colleague, or even knowing of the colleague's existence. Thus, the integration of the communications component with the rest of the system was an essential part of the system's success in meeting the needs of its users.

7.2 Future Work

There are several parts of WebStats which could be developed further, to better meet the needs of existing users or to broaden the user base.

First, the tobacco control case study undertaken for this thesis should be duplicated in other contexts. Some possible areas for expansion include aspects of population health such as physical activity and nutrition, areas of public concern such as environmentaldegradation, and public policy applications such as tracking the impact of social services in urban centres. By using WebStats in a variety of settings and with multiple populations of users, the tool's flexibility, usability, and general applicability will be tested more rigorously, and its features may be further refined to meet the needs of a broader user population.

The possibility of creating a report generation engine was explored while developing WebStats for the tobacco control community. During that development, it was found that the gain in flexibility offered by presenting the user with raw analysis results – that is, image files and text results – outweighed the convenience of allowing the user to create reports from within WebStats. However, because WebStats is used in a wide variety of settings, there may be some users for whom this convenience is an important feature. In any case, the possibility of creating presentation-quality analytical reports via a Web-based system is an appealing one.

The model presented in Section 3.1 has a great deal of potential for expansion, as any Web-based application can access any analysis results and associated semantic keywords stored by the system. The collaboration engine, for example, makes use of the keywords to create a user-interest-based communications platform, while the proposed report-generation engine would use the stored analysis results to create high-quality reports. Other tools that might be created include a mapping tool which could associate school survey results with geo-spatial data, thus creating a cartographic representation of schools' smoking health, which could then be linked to demographic data, for example, or to the location of stores which sell cigarettes.

Another enhancement is the inclusion of more statistical calculations in the system. Expansion in this direction is a virtually limitless, as there are many different formulae and models that may be used, depending on the user's statistical expertise, the sort of data he is analyzing, and the type of questions he is trying to answer. This enhancement could very well be investigated in co-ordination with the use of WebStats in other contexts. A new user group might identify specific sets of calculations or models that are particularly appropriate to its field of study, and these could then be included in WebStats on a context-by-context basis. A user from the tobacco-control field, for example, could be shown a set of statistical operations different from that shown to a user working within an environmental assessment context.

A generalization of this concept, which would be aimed at users with a high degree of computer literacy and a strong background in statistics, is to allow a user to create his own statistical models or formulae. This capability would be difficult from both a mathematical standpoint, in that the system should check to make sure the user's equations are, at the very least, syntactically correct, and a computational standpoint, in that WebStats would have to create SQL queries which would execute the specified calculations, but it would improve WebStats's flexibility and potential for expansion.

Appendix A

Description of WIDE tool

Members of the Computer Systems Group (CSG) at the University of Waterloo have been developing portal technologies and deploying Web-based portals based on these technologies since the early 1990s. Current portals, called Community Learning Spaces, are constructed using the Web-based Informatics Development Environment (WIDE).

The Community Learning Spaces are based on a service-oriented architecture and use the WIDE software toolkit to implement these services to enable the design, construction, deployment, maintenance and operation of complex Web-based systems. WIDE is primarily based on open source software technology and consists of a number of service and supporting frameworks. Applications can include input forms or reports containing extensive multimedia materials such as imaginative use of maps or any 2-dimensional diagram, Web sites, databases, indexing and searching methods, agents, and push technologies. WIDE also contains a knowledge management system that supports documentation of technical information and best practices.

Note: Adapted from "The Web-based Informatics Environment" by Cowan *et al.*

Appendix B

Statistics Reference

Boxplot This graph type summarizes the table's data, showing the minimum and maximum values, median (middle) value and the first and third quartiles (that is, the values such that 25% and 75% of the data is greater, respectively)

Central Tendency Calculations that measure where the centre of the data lies. Three common calculations of this type are mean, median and mode.

Confidence Level When working with inferential statistics, one cannot make any definitive statements, but rather one may consider results to be true with a certain level of confidence, usually considered in terms of a range of values; informally, one can say that one is 95% confident (for example) that the result falls within a certain range. Formally, if one is estimating a parameter s with a confidence level of 95%, this means that there is a 95% chance that a randomly selected sample from the target population will have a confidence interval which contains the true value of s for the target population.

Confidence Interval When estimating a parameter, s, in a sample population, one usually describes s as falling within a range of values with a certain confidence level. This range of values and the associated confidence level are referred to as the confidence interval.

Correlation Coefficient A numerical expression of the degree to which two variables in a data set are related to each other, and ranges from 1 to -1. Given an independent variable x and a dependent variable y, a correlation coefficient of 1 means x and y are positively correlated – that is, each unit of variation in x is associated with the same (proportional) size and direction variation in y. A correlation coefficient of 0 means that x and y are completely independent, and -1 means that x and y are negatively correlated.

Descriptive Statistics Calculations which convey information about the data set; two common types of descriptive statistics are central tendency and variance.

Frequency Table A table which records the number of times each value appears in a data set. When displayed as a chart (often a bar chart), it is also referred to as a histogram.

Inferential Statistics Calculations which are used to extrapolate information about a population from data about a sample from that population, expressed with a certain confidence level. More advanced inferential statistics may compare two or more populations based on samples drawn from each, again with a certain level of confidence.

Linear Regression When given a set of data points, one can use linear regression to find a "line of best fit" that describes these points. As the name implies, a "line of best fit" is a linear formula (that is, a formula of the form $y = mx + b$) which comes as close as possible to matching the points in the data set. As with other inferential calculations, this line is accompanied by a confidence interval, which indicates the range of values over which one can be confident that this formula is correct, and the associated degree of certainty.

Mean Often called the average, this is the sum of all the elements of a data set divided by the number of elements in the data set.

Median This is the value such that half the elements in a data set are above it, and half are below it.

Mode This is the element which occurs most frequently in a data set.

Quartile The first quartile is the value such that 25% of the elements in the data set are above it, and 75% are below it. Similarly, the third quartile is the value such that 75% of the elements in the data set are above it and 25% are below it.

Range The difference between the largest element and the smallest element in a data set.

Sample Population A subset of the target population which one can identify as being a potential subject for analysis. For example, if the target population is every Canadian, then the sample population might be everyone listed in a phone book.

Scatterplot When analyzing two variables, this graph type shows each value pair as a point on the (x,y) plane.

Standard Deviation This is the square root of the variance and, since it uses the same units as the data set, is used frequently in many inferential statistical operations.

Study Population A subset of the sample population which is small enough that it may be used for an experiment. If the target population is every Canadian and the sample population is everyone listed in a phone book, then the study population might be one hundred people randomly selected from the phone books of five large cities.

Target Population The population (often a group of people, but may also be a collection of items) which one would like to analyze. Often, this population is too large to be able to analyze directly, so sample or study populations are used, and then information about the target population is inferred from these smaller populations.

Variance This is the sum of the differences between each element in the data set and the mean of the data set, divided by the number of elements in the data set. Informally, it measures the amount of variation from the centre in a data set.

See: Bulmer (1967), Donnelly (2004), Fisher (1970), Klugh (1974), Richmond (1964).

Appendix C

Description of SHAPES

SHAPES (School Health Action, Planning and Evaluation System) is a modular local data collection system. It consists of a machine-readable questionnaire that can be administered to all students in a school, and a computer-generated feedback report for each school.

SHAPES was developed with several goals; these are to facilitate and stimulate the development, planning, and evaluation of interventions and policies related to health behaviours within schools; to enable high quality research to be conducted in real world settings; to minimize burden on school personnel and students; to maximize value to schools and stakeholders.

The SHAPES smoking behaviours module includes items on demographics, smoking behaviours (including amount, frequency and situation), attitudes, and the social and physical environment. The current smoking behaviours module is derived form the School Smoking Profile, which is already in widespread use. Since 2000, the smoking module has been administered in over 300 schools across Canada and data has been collected for over 110, 000 students.

Note: Adapted from “About SHAPES” (http://www.shapes.uwaterloo.ca/about/).

Appendix D

Details of Interview Process

Interview Questions

1. Could you give me a general overview of tobacco control programs you are involved with?
2. How is the data for these programs currently analyzed and used?
3. Ideally, how would you like to see the data used?
4. What kind of reports would you like to generate using the system?
5. How would you like to use statistics in your reports, ideally? What kinds of conclusions would you want to draw about the data you had gathered?
6. Have you used any computerized statistics packages (e.g. SPSS) in the past?
7. If so, what has your experience been (positive, negative, mixed)? Explain.
8. If so, what tools from that package would you most like to see in WebStats?
9. If not, have there been any barriers that have prevented you from using statistical analysis tools?
10. What is your familiarity with statistics? Practical? Mathematical? None?
11. What statistical calculations would you like to be able to perform in WebStats? Simple calculations? Data models? Confidence intervals?
12. What context would you use the system in? School? Office? Home?
13. How much time would you have to create your reports?
14. How would the reports be used by their target audience?

Note that this list of questions is not necessarily exhaustive, but rather gives a sense of the sorts of areas that the open-ended interview will explore.

Summary of Participants

Name	**Organization**	**Role**	**Location**
Andrew Loughead	Manitoba Health	Tobacco Control Coordinator	Winnipeg, MB
Caroline Silverman	Cancer Care Ontario	Knowledge Exchange Officer	Toronto, ON
Eleanor Taylor	Vancouver Island Hospital Authority	Public Health Nurse	Victoria, BC
Janice Tiessen	University of Waterloo	Data Manager	Waterloo, ON
John Garcia	Ontario Tobacco Research Unit	Director of Evaluation	Toronto, ON
Lynda Zimmerman	Ottawa Dept. of Public Health	Public Health Nurse	Ottawa, ON
Mari-Alice Jolin	University of Waterloo	Project Co-ordinator, PHR	Waterloo, ON
Mary Thompson	University of Waterloo	Professor, Dept. of Statistics	Waterloo, ON
Matt van der Meer	University of Waterloo	Data Analyst	Waterloo, ON
Patrick Seliske	Guelph Public Health Unit	Epidemiologist	Guelph, ON
Paul McDonald	University of Waterloo	Co-Director, PHR	Waterloo, ON
Phyllis Barkley	Ottawa Dept. of Public Health	Public Health Nurse	Ottawa, ON
Ratsamy Pathammavong	The Lung Association	Tobacco Control Manager	Toronto, ON
Scott Leatherdale	Cancer Care Ontario	Statistician	Toronto, ON
Steve Brown	University of Waterloo	Professor, Dept. of Statistics	Waterloo, ON
Steve Manske	University of Waterloo	Scientist, PHR	Waterloo, ON

Note: PHR is the Population Health Research group

Appendix E

WebStats Instruction Manual

Introduction

This is a guide to WebStats, a Web-based statistical analysis toolkit developed by the Computer Systems Group at the University of Waterloo. It describes all of the features available in the toolkit, and explains each screen and menu item.

The WebStats system is made up of several components, which work together to create a full-featured toolkit for storing data, doing statistical analyses, and sharing the results with others. These components are:

- A **data set viewer**, which allows you to browse all of the data sets stored by the system.
- A **data import tool**, which allows you to bring either numerical or coded data into the system from a plain-text file.
- Basic and advanced versions of a **statistical analysis tool**, which allow you to perform statistical calculations and create graphs from these data sets.
- A **message board**, which allows you to share your analyses with colleagues.
- **Data set management tools**, which allow you to modify the data sets stored by the system and change the characteristics of a particular data set.
- A **personal settings manager**, which allows you to modify the system to fit your needs.

These components make use of a few key concepts, which are important to understand when using the system. These concepts are:

- the use of **keywords** to describe your interests, and the topics covered by analyses, data sets and variables,
- the creation of an **analysis** by bringing together **operations** and **constants**,
- the distinction between **numerical** and **coded** data sets, and
- the use of **privacy settings** to control who may view and edit an analysis

The guide begins by explaining each of these concepts, and then goes on to describe each component in detail. There is also a brief statistical reference at the end of the guide.

Important Concepts

Keywords

Keywords are used to allow you to specify areas of interest, and to give meaning to variables, data sets and analyses. Specifically, you can express your interests via a set of descriptive keywords; for example, if you were interested in youth tobacco control programs, you might select keywords such as grade school, high school, or prevention.

Analysis Composition

An analysis is composed of operations and constants. Operations are statistical calculations or tables which are presented as either numerical results or as graphs. Constants, meanwhile, are fixed values which are either superimposed on graphs as horizontal lines or shown as a number alongside other numerical results. An analysis may include any number of operations and constants.

Numerical vs. Coded Data Sets

A numerical data set contains data which are meant to be interpreted literally. For example, if the data for a particular person has the value 20 in the "age" column, then it assumed that the person's age is equal to 20. A coded data set, meanwhile, contains data which are meant to be interpreted using a coding scheme. For example, if a particular person has the code 4 in the "age" column, you would look up the appropriate variable and code using a coding scheme like the one shown in Table 1.

Table 1: Sample Coding Scheme

Variable	Code	Meaning
age	1	14 years old and younger
	2	15-17 years old
	3	18-20 years old
	4	21 years old and older

Thus, Table 1 shows that the "age" code 4 corresponds to "21 years old and older".

Privacy Settings

Privacy settings allow you to control how much access other users will have to your data sets and analyses. Table 2 shows how the system's analysis privacy settings work.

Table 2: Privacy Settings

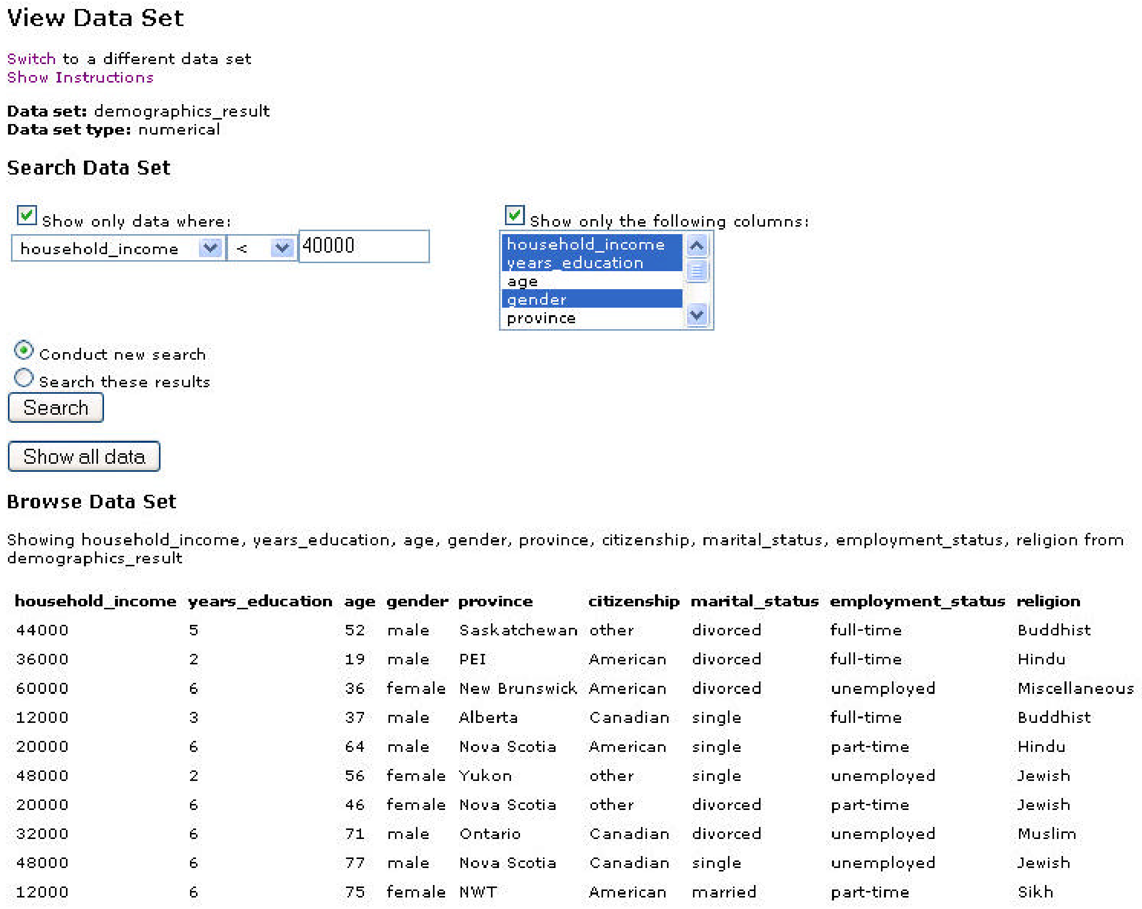
Privacy Level	Snapshot Shown?	Details Shown?	Editable?
Public	Yes	Yes	Yes
Read-only	Yes	Yes	No
Private	Yes	No	No
Hidden	No	No	No

Data Set Viewer

The data set viewer allows you to vie w the data contained in a data set which is stored by the system. To view a data set, click on the "View Data Sets" link on the sidebar, select the data set from the drop-down menu which will appear, and then click "View".

Once the data set has been loaded, you can filter the data shown using the "Search Data Set" menu at the top of the page. To show only data matching a certain criterion (e.g. age > 20), use the "Show only data where…" option. To show only certain columns, use the "Show only the following columns…" option. You can begin a new search by selecting the "Conduct new search" option, or further refine your current search by selecting "Search these results". You can view all the data at any time by clicking on the "Show all data" button.

If you are viewing a coded data set, you can switch between numerical and coded data by clicking on the "Switch to viewing numerical (or coded) data" link at the top of the page. If a value is missing, the system will show a "NULL" value.



View Data Set

Switch to a different data set
Show Instructions

Data set: demographics_result
Data set type: numerical

Search Data Set

☑ Show only data where:
household_income ▾ < ▾ 40000

☑ Show only the following columns:
household_income
years_education
age
gender
province

● Conduct new search
○ Search these results

Search

Show all data

Browse Data Set

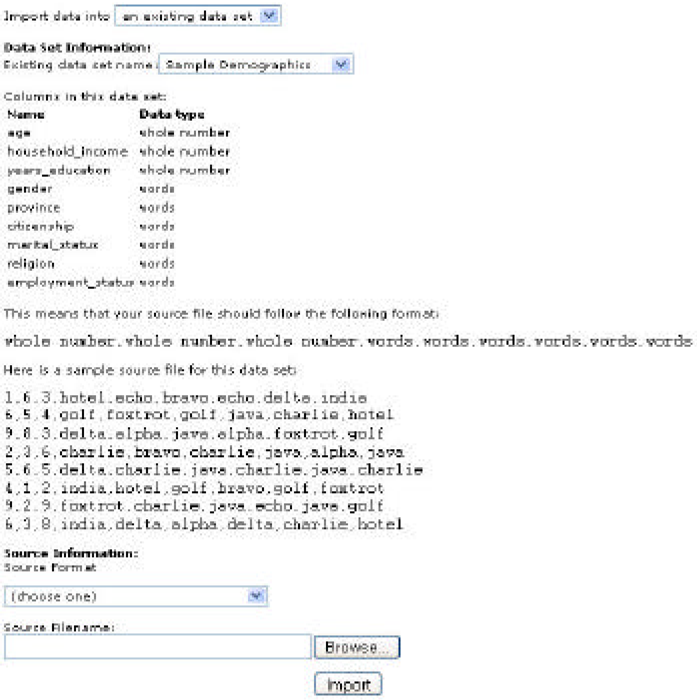
Showing household_income, years_education, age, gender, province, citizenship, marital_status, employment_status, religion from demographics_result

household_income	years_education	age	gender	province	citizenship	marital_status	employment_status	religion
44000	5	52	male	Saskatchewan	other	divorced	full-time	Buddhist
36000	2	19	male	PEI	American	divorced	full-time	Hindu
60000	6	36	female	New Brunswick	American	divorced	unemployed	Miscellaneous
12000	3	37	male	Alberta	Canadian	single	full-time	Buddhist
20000	6	64	male	Nova Scotia	American	single	part-time	Hindu
48000	2	56	female	Yukon	other	single	unemployed	Jewish
20000	6	46	female	Nova Scotia	other	divorced	part-time	Jewish
32000	6	71	male	Ontario	Canadian	divorced	unemployed	Muslim
48000	6	77	male	Nova Scotia	Canadian	single	unemployed	Jewish
12000	6	75	female	NWT	American	married	part-time	Sikh

Data Import Tool

The data import tool allows you to bring either numerical or coded data into the system from a plain-text file. You can either create a new data set, or add entries to an existing data set. There are three steps to importing data into the system.

1. Select the data set to import the data into. If you are creating a new data set, then you should import the data into a new data set. If you are adding to or replacing an existing data set, then you should import the data into an existing data set .

   Import data into: an existing data set

   Data Set Information:
   Existing data set name: Sample Demographics

   Columns in this data set:

Name	Data type
age	whole number
household_income	whole number
years_education	whole number
gender	words
province	words
citizenship	words
marital_status	words
religion	words
employment_status	words

   This means that your source file should follow the following format:

   ```
   whole number,whole number,whole number,words,words,words,words,words,words
   ```

   Here is a sample source file for this data set:

   ```
   1,6,3,hotel,echo,bravo,echo,delta,india
   6,5,4,golf,foxtrot,golf,java,charlie,hotel
   9,8,3,delta,alpha,java,alpha,foxtrot,golf
   2,3,6,charlie,bravo,charlie,java,alpha,java
   5,6,5,delta,charlie,java,charlie,java,charlie
   4,1,2,india,hotel,golf,bravo,golf,foxtrot
   9,2,9,foxtrot,charlie,java,echo,java,golf
   6,3,8,india,delta,alpha,delta,charlie,hotel
   ```

   Source Information:
   Source Format: (choose one)

   Source Filename:

   If you choose to import into a new data set, then you must give this data set a name, and specify the columns that will be in the data set – that is, describe the format of the data that is being imported. You must give each column a name describing the data stored in that column, set the data type to be either words or numbers, and specify the units that the data is measured in. For example, the units for a column containing temperatures might be "degrees Celsius".

   If you choose to import into an existing data set, then you simply need to select the data set to import into.

   In either case, the menu will show the columns in the destination data set and a sample data file made up of random data which conforms to the data types stored in the data set. This will help you ensure that the format of the raw data file being imported matches the format of the data set you are importing into.

2. Select the format of the data file being imported. This can either be "Plain text, separated by commas" or "Plain text, separated by tabs". In either format, each line of text corresponds to a row of data, and each column is separated by either a comma (",") or tab ("	") character.

3. Select the source filename. You can do this either by typing the filename (including the full path) into the text box, or by clicking on the "Browse" button, and navigating to the proper file. This should be a plain-text file, with the data separated by either commas or tab characters.

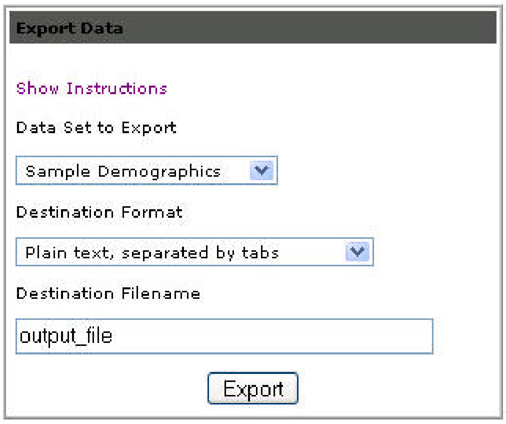
Data Export Tool

The data export tool allows you to extract data from the system into a plain-text file. There are three steps to exporting data from the system.

1. Select the data set to export the data from.

2. Select the format of the the file you are exporting into, which be referred to as the destination file. The file will be in plain text format, and the data in each row may be separated either by commas or tab characters.

3. Enter the name of the destination file. Note that this should just be the name of the file itself, and should not specify where the file will be saved on your computer. Also, since the file is in plain text format, you do not need to add a file extension, such as .txt, to the end of the filename.



Export Data

Show Instructions

Data Set to Export

Sample Demographics

Destination Format

Plain text, separated by tabs

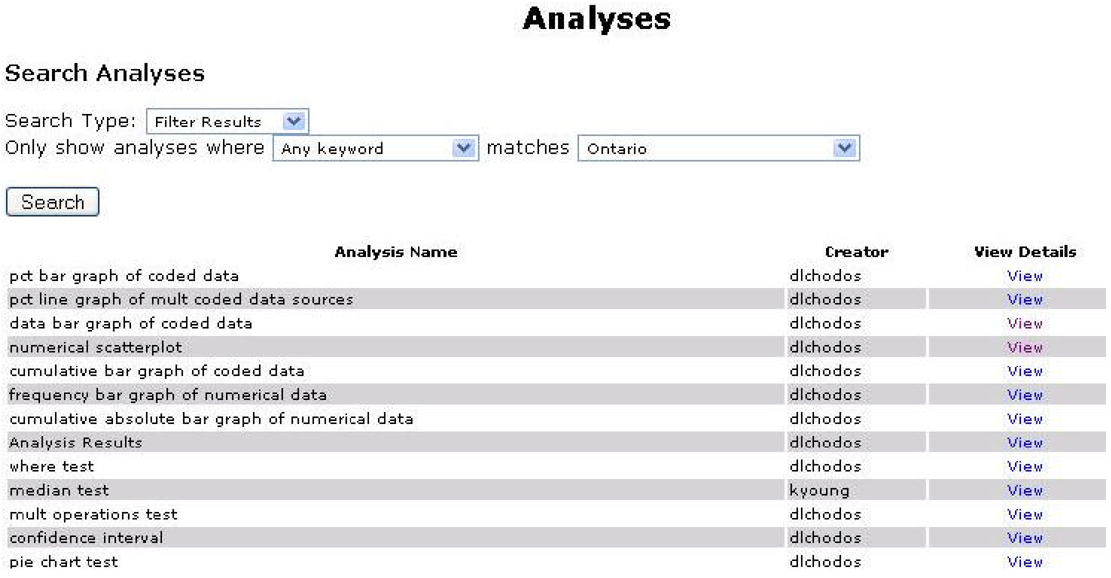
Destination Filename

output_file

Show Analyses

The show analyses tool allows you to search for and browse through analyses stored within the system. There are three different search types available: “Filter Results” shows analyses which match your search parameters. “User Interests” searches for analyses with keywords which match your interests – for more information on setting your interests, see the section on the “Personal Settings” tool. “Show All” shows all the analyses stored by the system.

After performing the search, the system shows the name of each analysis, the username of the person who created it, and might also show links to view more details and edit the analysis, depending on the analysis’s privacy settings.



Analyses

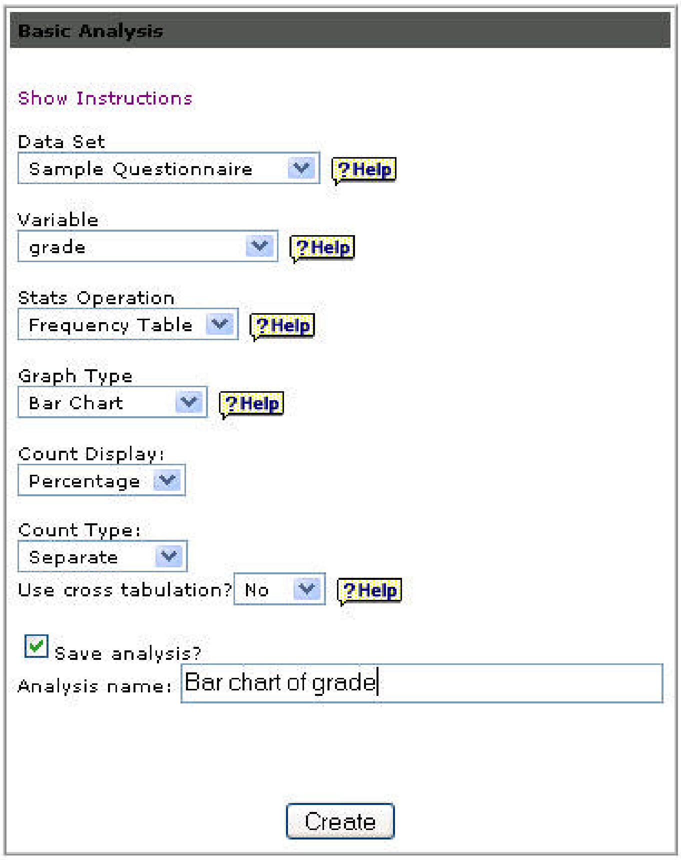
Search Analyses

Search Type: Filter Results
Only show analyses where Any keyword matches Ontario

Analysis Name	Creator	View Details
pct bar graph of coded data	dlchodos	View
pct line graph of mult coded data sources	dlchodos	View
data bar graph of coded data	dlchodos	View
numerical scatterplot	dlchodos	View
cumulative bar graph of coded data	dlchodos	View
frequency bar graph of numerical data	dlchodos	View
cumulative absolute bar graph of numerical data	dlchodos	View
Analysis Results	dlchodos	View
where test	dlchodos	View
median test	kyoung	View
mult operations test	dlchodos	View
confidence interval	dlchodos	View
pie chart test	dlchodos	View

Basic Analysis

When creating a basic analysis, there are a few parameters you need to specify:



Basic Analysis

Data Set: Sample Questionnaire

Variable: grade

Stats Operation: Frequency Table

Graph Type: Bar Chart

Count Display: Percentage

Count Type: Separate

Use cross tabulation? No

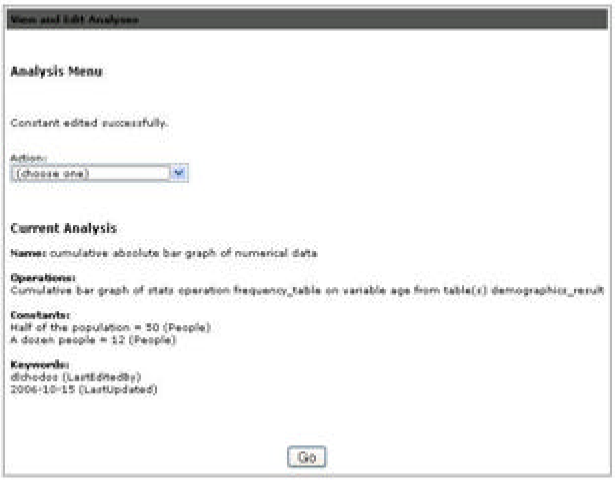
☑ Save analysis?

Analysis name: Bar chart of grade

- The data set that the data for the analysis will be drawn from.
- The variable that you are interested in analyzing; once you select a data set, the variables listed in this menu will correspond to the columns in the selected data set.
- The stats operation that you would like to perform on that variable; this may be either an operation which creates a table, or a descriptive operation. See the statistical reference at the end of this guide for more information on the operations available in this menu.
- The type of graph that you would like to generate; if you select a descriptive stats operation, then this menu will not be shown. If you select a stats operation which creates a table, then a number of different graph types will be available. See the statistical reference for more information.
- If you selected the frequency table stats operation, then the following menus will also be shown:
    - Count Display: There are two ways that the column values can be displayed in a frequency chart: either by percentage, where the graph will show the percentage of times that a column value occurs, or absolute, where the graph will show the number of times that column value occurs.
    - Count Type: There are two ways that the column values can be counted in a frequency chart: either separately, where each column in a bar graph is calculated separately, or cumulatively, where each column in a bar graph is calculated as the number of items less than or equal to that value.
    - Cross Tabulation: Selecting "Yes" in this menu allows you to split up the frequencies for each column according to the value of a second variable.
- If you want to save the analysis for later retrieval, viewing or editing, you need to check the “Save analysis?” box and specify a name in the analysis name field at the bottom of the menu.

Advanced Analysis

When performing an advanced analysis, you are first presented with a menu which allows you to manage the analysis as a whole; you can add, edit or remove operations or constant values from the analysis, create or delete an analysis, or change an analysis' settings, as shown in the screenshot on the right.



Analysis Menu

Constant edited successfully.

Action:
(choose one)

Current Analysis

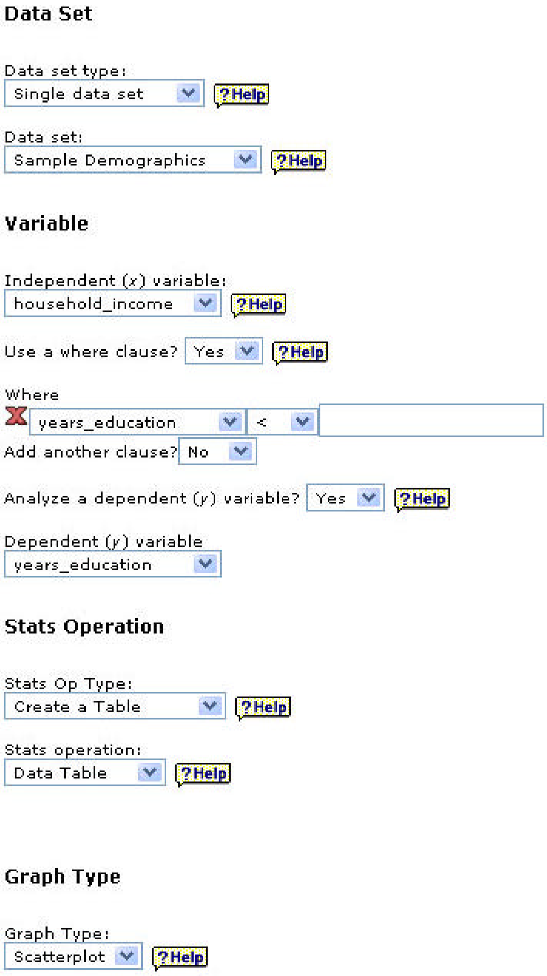
Name: cumulative absolute bar graph of numerical data

Operations:
Cumulative bar graph of stats operation frequency_table on variable age from table(s) demographics_result

Constants:
Half of the population = 50 (People)
A dozen people = 12 (People)

Keywords:
dichodos (LastEditedBy)
2006-10-15 (LastUpdated)

When creating or editing an operation, you have all of the basic options available, which are described in the previous section; however, you also have a number of more complex options and operations to work with:


Data Set

Data set type: Single data set

Data set: Sample Demographics

Variable

Independent (*x*) variable: household_income

Use a where clause? Yes

Where
years_education <
Add another clause? No

Analyze a dependent (*y*) variable? Yes

Dependent (*y*) variable: years_education

Stats Operation

Stats Op Type: Create a Table

Stats operation: Data Table

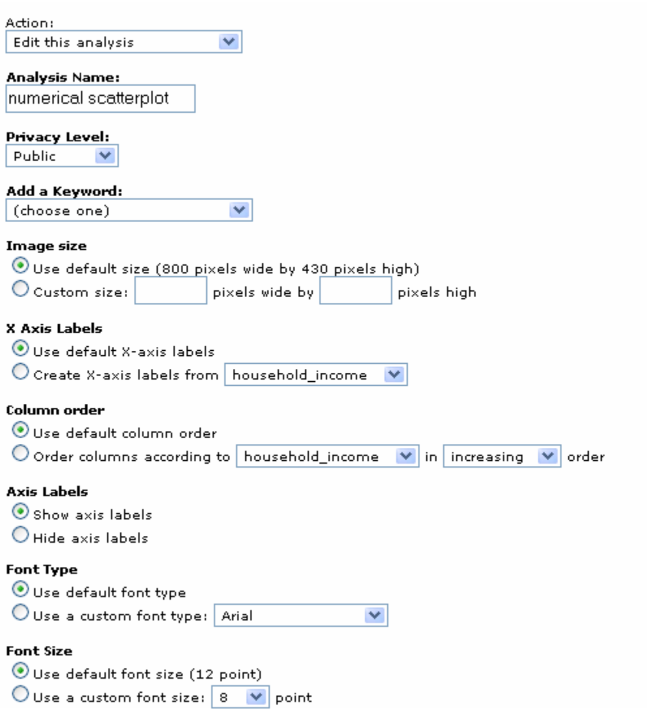
Graph Type

Graph Type: Scatterplot

- You can draw the data for the analysis from multiple data sets, in addition to being able to draw the data from a single data set (as in the basic analysis).
- You can analyze two variables to find relationships between variables in a data set. Specifically, you can select a dependent (*y*) variable and see if it is influenced by an independent (*x*) variable.
- You can use one or more where clauses to filter the results by a set of criteria. You can use the red X beside a clause to remove it from the set of criteria.
- You can perform inferential calculations such as linear regression or finding a confidence interval. Please see the statistical reference for more information.

Analysis Options

When editing an analysis's options, there are many ways in which you can customize the presentation of an analysis. These options are explained below:



Action:
Edit this analysis ▾

Analysis Name:
numerical scatterplot

Privacy Level:
Public ▾

Add a Keyword:
(choose one) ▾

Image size
● Use default size (800 pixels wide by 430 pixels high)
○ Custom size: ___ pixels wide by ___ pixels high

X Axis Labels
● Use default X-axis labels
○ Create X-axis labels from household_income ▾

Column order
● Use default column order
○ Order columns according to household_income ▾ in increasing ▾ order

Axis Labels
● Show axis labels
○ Hide axis labels

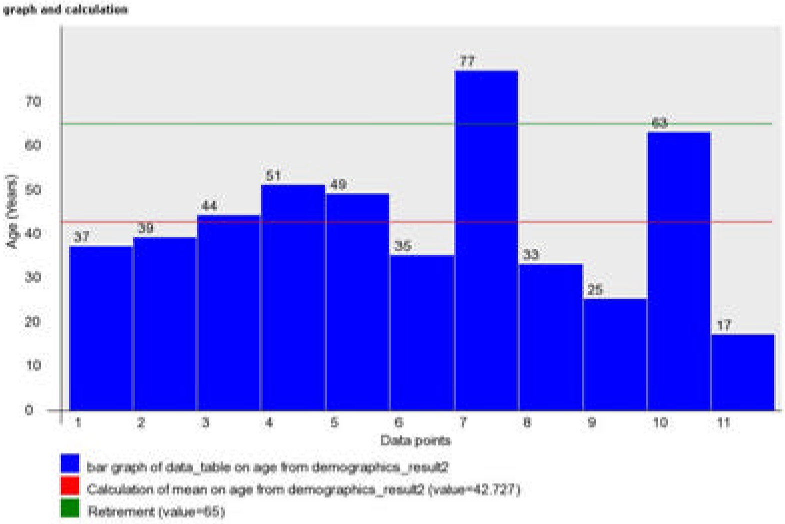
Font Type
● Use default font type
○ Use a custom font type: Arial ▾

Font Size
● Use default font size (12 point)
○ Use a custom font size: 8 ▾ point

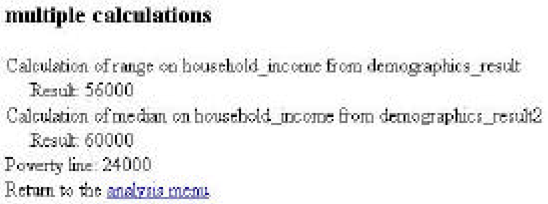
- **Privacy Level**: this controls who can view and edit the analysis
- **Add or Remove Keywords**: keywords are used to assign context or meaning to an analysis. Some keywords are assigned automatically based on the data sets and variables used in the analysis, while others may be added by users.
- **Image Size** : the dimensions of the image generated by the system. By default, the image is 800 pixels wide by 430 pixels high.
- **X Axis Labels**: the variable to use for X-axis labels when showing a raw data graph. By default, columns in a raw data graph are numbered according to the number of data points being graphed. That is, the first data point is labelled "1", the second is labelled "2", and so forth. However, by selecting a variable for the X axis label, the columns are labelled according to the value of that variable for each data point.
- **Column Order**: the name of the variable to be used in sorting the results, and the order in which the results are to be sorted. By default, the results are sorted by the variable being analyzed, in ascending order.
- **Axis Labels**: Controls whether or not the axis labels (the text describing the axes, as opposed to the labels on the axes themselves) are shown
- **Font Type** : the font to be used (the default is Arial)
- **Font Size**: the size of the font to be used, in points (the default is a 12-point font)

Analysis Results

When you are viewing an analysis, the results may be shown in one of two ways. If the analysis includes one or more graphs, then the results will be shown on a graph, as shown in the first screenshot. If the analysis consists entirely of calculations, then the results will be shown in text format, as shown in the second screenshot.



A set of analysis results, shown as a graph



multiple calculations

Calculation of range on household_income from demographics_result
Result: 56000
Calculation of median on household_income from demographics_result2
Result: 60000
Poverty line: 24000
Return to the analysis menu

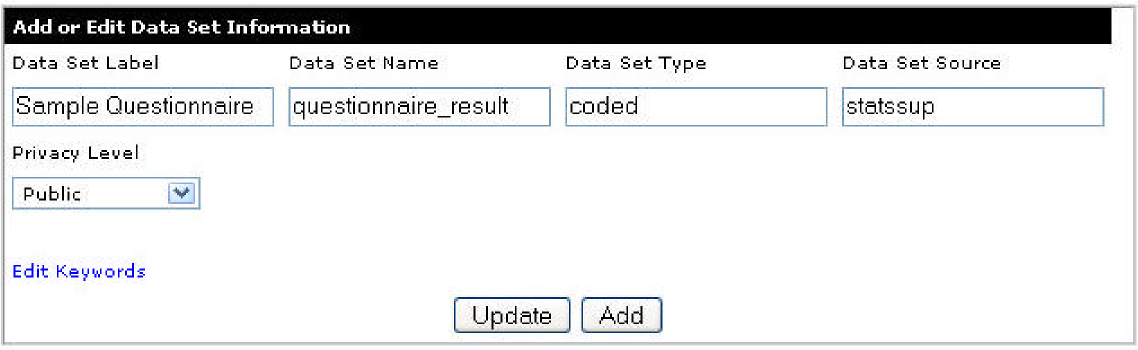
Another set of analysis results, shown in text format

Data Set Information

The data set information menu allows you to maintain information about the data sets that are used by the WebStats system. To edit the information for a particular data set, you can click on its name, which will bring up a new window. In maintaining information about the data set, the following fields are used:

- **Data Set Label**: the name to be used by the user interface for this data set
- **Data Set Name**: the name of the data source in the database
- **Data Set Type**: the kind of data stored by this data set; valid values are "numerical" or "coded"
- **Data Set Source** : the name of the database this data set is stored in

Additionally, you can click on the "Edit Keywords" link to edit the keywords which are associated with a particular data set.



Add or Edit Data Set Information

Data Set Label	Data Set Name	Data Set Type	Data Set Source
Sample Questionnaire	questionnaire_result	coded	statssup

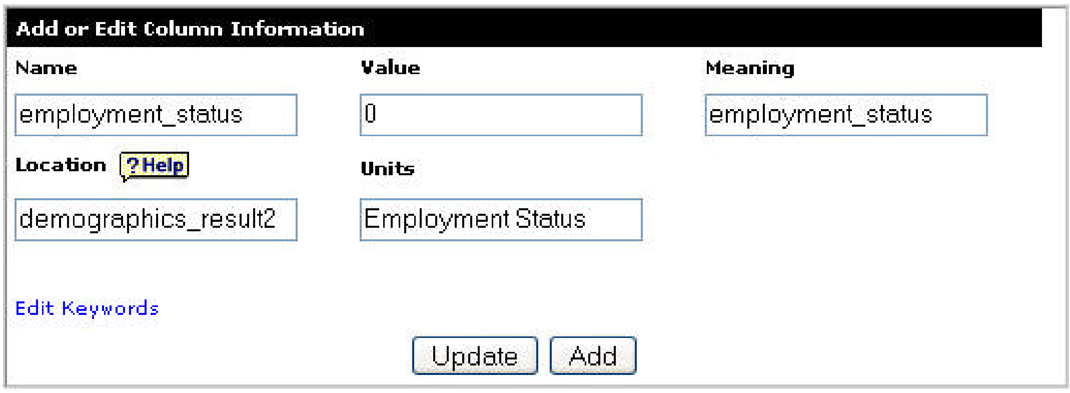
Privacy Level: Public

Column Information

The column information menu allows you to maintain information about the columns within a particular data set used by the WebStats system. To edit the information for a particular column, you can click on its name, which will bring up a new window. In maintaining information about a column, the following fields are used:

- **Name** : the name of the column in the data source
- **Value** : for coded data, a code value used for this column; for numerical data, this is set to 0
- **Meaning** : for coded data, the meaning corresponding to the code value.
  For numerical data, this is set to the column name
- **Location**: the name of the data set this column is stored in
- **Units** : what the column is measuring (e.g. dollars, people, or ages)

Additionally, you may click on the "Edit Keywords" link to edit the keywords which are associated with a particular column.



Add or Edit Column Information

Name: employment_status

Value: 0

Meaning: employment_status

Location: demographics_result2

Units: Employment Status

Personal Settings

To allow you to tailor the system to your needs, the system offers three categories of user preferences: statistical familiarity, computational complexity and interface detail. Within each category, the user can choose a setting of "high", "medium" or "low", which will affect the range of statistical operations available, the adjustments which can be made to those operations, and the variety of customization options presented, respectively. Shown below is a chart which explains the effect that each setting has on the system.

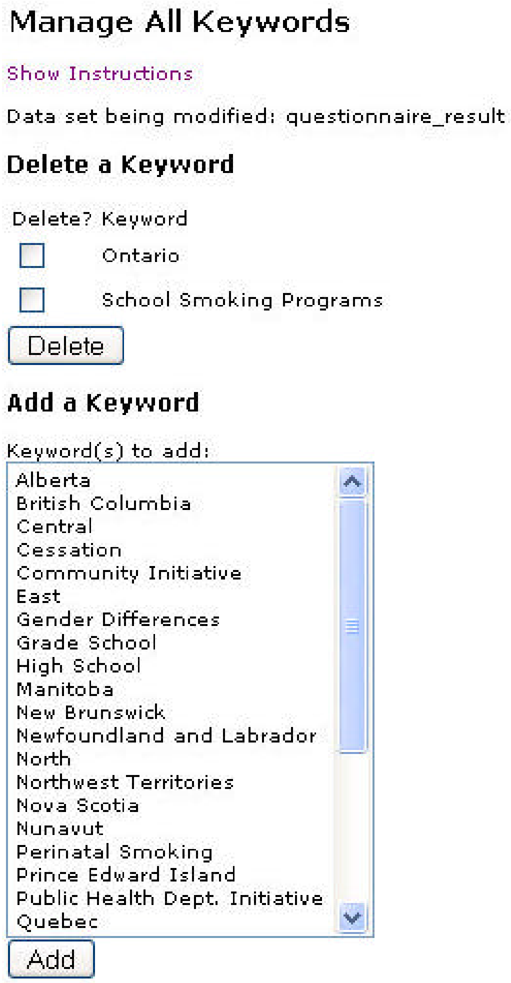
Category	Setting	Effect
Statistical Familiarity	**Low**	Descriptive operations hidden Inferential operations hidden Help provided for calculation results
	Medium	Inferential operations hidden
	High	All operations shown
Computational Complexity	**Low**	Where clause hidden Second variable hidden Cumulative, absolute value graph options hidden
	Medium	Where clause hidden
	High	All operations shown
Interface Detail	**Low**	Collection options shown: name and privacy level
	Medium	Additional options shown: image, axis dimensions
	High	All options shown

Keyword Management

Keywords are used to assign meaning to data sets, variables, and analyses, and also indicate a user's areas of interest. Management of these keywords is performed via the same basic interface, regardless of whether the keywords are associated with a variable, a data set, or a user profile. For the sake of explanation, the variable, data set or user profile with which a set of keywords is associated will be referred to as an "entity".

You may delete one or more keywords from an entity by going under the "Delete a Keyword" menu, clicking the checkbox next to each keyword you want to delete, and then clicking the "Delete" button. Note that this does not delete the keyword entirely, but merely removes its association with the entity.

Similarly, you may add one or more keywords to an entity by clicking on the keywords in the selection menu under the "Add a Keyword" menu, and then clicking the Add button.



Manage All Keywords

Show Instructions

Data set being modified: questionnaire_result

Delete a Keyword

Delete? Keyword

□ Ontario

□ School Smoking Programs

Delete

Add a Keyword

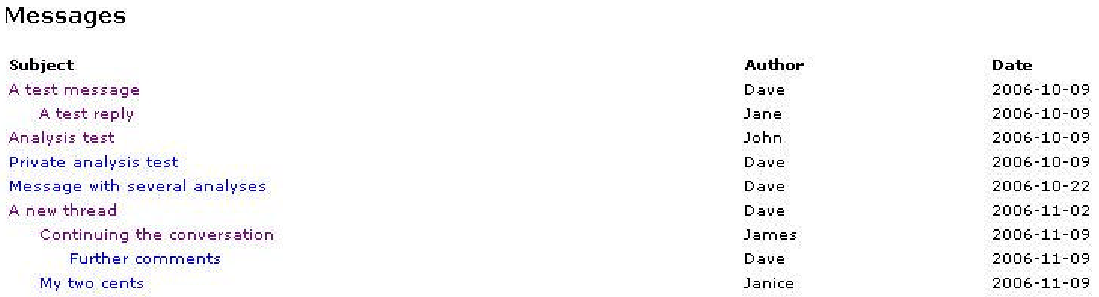
Keyword(s) to add:

Alberta
British Columbia
Central
Cessation
Community Initiative
East
Gender Differences
Grade School
High School
Manitoba
New Brunswick
Newfoundland and Labrador
North
Northwest Territories
Nova Scotia
Nunavut
Perinatal Smoking
Prince Edward Island
Public Health Dept. Initiative
Quebec

Add

Message Board

The message board is provided as a way for colleagues to share information about analyses, news about ongoing projects, or seek assistance in analyzing data from other users. The message board consists of three parts; the first part, shown in Figure 1, is a message board browser, which shows all of the messages on the board, organized hierarchically by conversation.

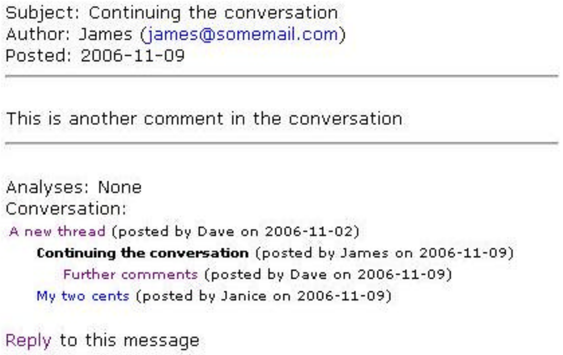


Messages

Subject	Author	Date
A test message	Dave	2006-10-09
A test reply	Jane	2006-10-09
Analysis test	John	2006-10-09
Private analysis test	Dave	2006-10-09
Message with several analyses	Dave	2006-10-22
A new thread	Dave	2006-11-02
Continuing the conversation	James	2006-11-09
Further comments	Dave	2006-11-09
My two cents	Janice	2006-11-09

Figure 1: Message Board Browser

The second part, shown in Figure 2, is the message view, which shows a single message and the conversation that it is a part of. You can click on another message within the conversation to jump to that message.



Subject: Continuing the conversation
Author: James (james@somemail.com)
Posted: 2006-11-09

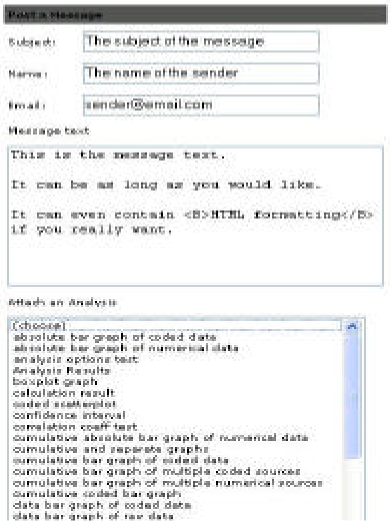
This is another comment in the conversation

Analyses: None
Conversation:

- A new thread (posted by Dave on 2006-11-02)
  - **Continuing the conversation** (posted by James on 2006-11-09)
    - Further comments (posted by Dave on 2006-11-09)
  - My two cents (posted by Janice on 2006-11-09)

Reply to this message

Figure 2: Message View



Post a Message

Subject: The subject of the message
Name: The name of the sender
Email: sender@email.com

Message text

```
This is the message text.

It can be as long as you would like.

It can even contain <B>HTML formatting</B>
if you really want.
```

Attach an Analysis

(choose)
absolute bar graph of coded data
absolute bar graph of numerical data
analysis options test
Analysis Results
boxplot graph
calculation result
coded scatterplot
confidence interval
correlation coeff test
cumulative absolute bar graph of numerical data
cumulative and separate graphs
cumulative bar graph of coded data
cumulative bar graph of multiple coded sources
cumulative bar graph of multiple numerical sources
cumulative coded bar graph
data bar graph of coded data
data bar graph of raw data

Figure 3: Message Composition Menu

Finally, there is a message composition menu, shown in Figure 3. Besides writing the message and specifying the message subject, author and email address, you can also attach one or more analyses to the message, which will cause a link to the analyses to be shown at the bottom of the posted message.

References

Banfield, Jeff. “Rweb: Web-based Statistical Analysis.” Website, accessed June 28, 2006. (http://www.jstatsoft.org/v04/i01/Rweb/Rweb.html)

Barkley, Phyllis (Public health nurse, Ottawa Department of Public Health) personal interview, July 5, 2006.

BASE website. Accessed December 19, 2006. (http://base.thep.lu.se/)

Boehm, B.W. “A Spiral Model of Software Development and Enhancement”, IEEE Computer, Volume 21, Issue 5, May 1988, pp. 61–72.

Brown, Steve; Thompson, Mary (Professors, Department of Statistics and Actuarial Science, University of Waterloo), personal interview, November 29, 2006.

Bulis, Kirk D.; DiStephano III, Joseph J. “W3MCSim: an online and reconfigurable Monte Carlo simulator for interactive probabilistic/statistical modeling.” Computer Methods and Programs in Biomedicine, Volume 77, 2005, pp. 71-79.

Bulmer, M.G. “Principles of Statistics.” Edinburgh: Oliver & Boyd. 1967.

Canadian Cancer Society website. Accessed on June 28, 2006. (http://www.cancer.ca/ccs/internet/standard/0,3182,3172_13163__langId-en,00.html)

Canadian Tobacco Usage Monitoring survey. Accessed online on June 28, 2006 (http://www.hc-sc.gc.ca/hl-vs/tobac-tabac/research-recherche/stat/ctums-esutc/2005/index_e.html)

Chiu, Chao-Min. “Towards a hypermedia-enabled and web-based data analysis Framework.” Journal of Information Sciences, Volume 30, Issue 1, 2004, pp. 60-72.

Choobineh, Joobin; Kini, Anil. “SQLSAM: SQL for Statistical Analysis and Modeling.” Proceedings of the 28th Annual Hawaii International Conference on Systems Sciences, 1995, pp. 418-427.

Comprehensive R Archive Network website. Accessed July 1, 2006. (http://cran.r-project.org/)

Covvey, H.D.; Zitner, D.; Berry, D.M.; Cowan, D.D.; Shepherd, M. “Formal structure for specifying the content and quality of the electronic health record.” Proceedings of the 11th IEEE International Requirements Engineering Conference, 2003. pp. 162-168.

Cowan, Don; Alencar, Paulo. “Software Support of Population Health Intervention.” WIHIR Research Seminar, presented October 12, 2005.

Cowan, Don; Fenton, Shirley; Mulholland, Doug. “The Web-based Informatics Development Environment (WIDE).” Computer Systems Group, University of Waterloo. Accessed December 8, 2006. (http://csg.uwaterloo.ca/wide.htm)

Cramer, Erhard; Cramer, Katharina; Kamps, Udo. “e-stat: A web-based learning environment in applied statistics.” Compstat 2002 (Proceedings in Computational Statistics), 2002,

de Leeuw, Jan. “Server-side Statistics Scripting in PHP.” Journal of Statistical Software, Volume 2, Issue 1, 1997.

Dameron, Ruth. Personal communication with Daniel Berry.

Dinov, Ivo D. “SOCR: Statistics Online Computational Resource.” Journal of Statistical Software, Volume 16, Issue 11, October 2006.

Donnelly Jr., Robert A. “The Complete Idiot’s Guide to Statistics.” New York: Alpha, 2004.

Dudoit, Sandrine; Gentleman, Robert C.; Quackenbush, John. “Open Source Software for the Analysis of Microarray Data.” BioTechniques 34:S45-S51 (March 2003)

Executive Information Systems. “SAS GSA Price List.” Accessed December 8, 2006. (http://www.execinfosys.com/SAS GSA PriceList.pdf)

Fernández, Mary; Suciu, Dan; Tatarinov, Igor. “Declarative Specification of Data-intensive Web sites.” In Proceedings of the USENIX Conference on Domain-Specific Languages, 1999, pp. 135-148.

Fielding, Roy, T.; Taylor, Richard N. “Principled Design of the Modern Web Architecture.” ACM Transactions on Internet Technology, Volume 2, Number 2, May 2002, pp. 115-150.

Fisher, Ronald A.“Statistical Methods for Research Workers.” Edinburgh: Oliver & Boyd. 1970.

Fuchs, Anton; Brenckmann, Ingo; Chitale, Anand. “SAS® Web-based Query and Reporting: An Introduction and Overview.” SAS, 2006.

Garcia, John (Director of Evaluation, Ontario Tobacco Research Unit), personal communication, 2006.

GNU Project. “GNU General Public License.” Accessed December 8, 2006. (http://www.gnu.org/copyleft/gpl.html)

Gordon, Scott V.; Bieman, James M. “Reported effects of rapid prototyping on industrial software quality.” Software Quality Journal, Volume 2, Number 2 (June, 1993), pp. 93-108.

Günther, Oliver; Müller, Rudolph; Schmidt, Peter; Bhargava, Hemant K.; Krishnan, Ramayya. “MMM: A Web-based System for Sharing Statistical Computing Modules.” IEEE Internet Computing, May/June 1997.

Harschbarger, Thad R. “Introductory Statistics: A Decision Map.” New York: Macmillan, 1971.

Health Indicator Query System website. Accessed March 20, 2007. (http://hip.on.ca/hipapp/IQ/data/index.php)

Heart and Stroke Foundation. “Incidence of Cardiovascular Disease.” Accessed February 22, 2007. (http://ww2.heartandstroke.ca/Page.asp?PageID=33&ArticleID=1077&Src=news&From=SubCategory)

Hertel, Ingrid; Clegg, Jennifer; McDaniel, Stephen. “SAS® Add-in for Microsoft Office: An Introduction and Overview.” SAS, 2004.

Holmberg, Nathan; Wünsche, Burkhard; Tempero, Ewan. “A Framework for Interactive Web-based Visualization.” Seventh Australasian Interface Conference, 2006.

Hornik, Kurt. “The R FAQ.” Accessed December 8, 2006. (http://cran.r-project.org/doc/FAQ/R-FAQ.html)

Huang, Xueqin; Tian, Hui; Xu, Xiangming. "JSIM Database: A Web-based Database Application Using XML." Proceedings of the ACM Southeast Regional Conference, 2000, pp. 171-178.

Jolin, Mari-Alice (Project Co-ordinator, Population Health Research, University of Waterloo), personal interview, July 4, 2005

Jones, Katherine. "An Introduction to Data Warehousing: What Are the Implications for the Network?" International Journal of Network Management, Volume 8, 1998, pp. 42-56.

Jones, Thomas A.; James, William R. "Analysis of Bimodal Orientation Data." Mathematical Geology, Volume 1, Number 2, 1969.

Kitchen, A.M.; Drachenberg, R.; Symanzik, J. "Assessing the reliability of web-based statistical software." Computational Statistics, 2003, pp. 1-18.

Kitchenham, Barbara; Pfleeger, Shari Lawrence. "Principles of Survey Research, Part 6: Data Analysis." Software Engineering Notes, Volume 28, Number 2, 2003, pp. 24-27.

Klugh, Henry E. "Statistics: The Essentials for Research." New York: Wiley and Sons, Inc. 1974.

Leatherdale, Scott (Researcher, Cancer Care Ontario), personal interview, July 7, 2005. Follow-up interview on June 2, 2006. Final feedback interview on February 9, 2007.

Loucopoulos, P.; Karakostas, V. "System Requirements Engineering." McGraw-Hill, New York, NY, 1995.

Loughead, Andrew (Tobacco Control Coordinator, Manitoba Health), personal interview, February 19, 2007.

Manske, Steve (Scientist, CBRPE, University of Waterloo), personal interview, July 4, 2005.

Manske, Steve. "2004-05 Youth Smoking Survey: Surveillance to Meet Local, Provincial and National Needs." Presentation given at the "Tobacco Reduction Together" conference in Edmonton, Alberta on May 26, 2006.

McCullough, B.D. “Assessing the Reliability of Statistical Software: Part I.” The American Statistician, Nov. 1998, Volume 52, Number 4, 1998, pp. 358-366.

McCullough, B.D. “Assessing the Reliability of Statistical Software: Part II.” The American Statistician, May 1999, Volume 53, Number 2, 1999, pp. 149-159.

McDonald, Paul. “Considerations and Rationale for a National Action Plan To Help Canadian Tobacco Users.” Published by the Population Health Research Group, University of Waterloo, September 29, 2003.

McDonald, Paul (Professor, Health Studies; Co-director, Population Health Research Group, University of Waterloo), personal interview, November 13, 2006.

McCaskell, Peter C. “Looking for Mr. X Bar: Supporting Statistical Computing in the Personal Computer Era.” ACM-SIGUCCS XVII, 1989. pp. 367-375.

Microsoft. “Farewell Clippy.” Press release, published April 11, 2006. Accessed February 20, 2007. (http://www.microsoft.com/presspass/features/2001/apr01/04-11clippy.mspx)

Microsoft Canada. “Microsoft Canada Pricing.” Accessed December 8, 2006. (http://www.microsoft.com/canada/pricelists/default.aspx)

Moon, Hojin; Lee, Jack J.; Ahn, Hongshik; Nikolova, Rumiana G. “A Web-based Simulator for Sample Size and Power Estimation in Animal Carcinogenicity Studies.” Journal of Statistical Software, Volume 7, Issue 13, 2002.

Moore, D. S.; McCabe, G. P. “Introduction to the Practice of Statistics, 4th ed.” New York: W. H. Freeman, 2002.

Morales, Rosanna. “Developing a Web Tool To Support Youth Tobacco Control.” Master’s Thesis (Health Studies and Gerontology), August 8, 2006.

Mullee, Mark A. “Web-based resources to assist the statistical analysis and preparation of data.” Pharmaceutical Statistics, Volume 4, 2005, pp. 129-139.

Myllymäki, Petri; Silander, Tomi; Tirri, Henry; Uronen, Pekka. "B-Course: A Web-Based Tool For Bayesian And Causal Data Analysis." International Journal on Artificial Intelligence Tools, Volume 11, Number 3, 2002, pp. 369–387.

Non-Communicable Diseases Surveillance Infobase website. Accessed March 20, 2007. (http://www.cvdinfobase.ca/surveillance/Mapdb/Infobase_e.htm)

Nusser, Sarah; Thompson, Dean. "Web-based Survey Tools." Proceedings of the Survey Research Methods Section, ASA, 1998, pp. 951-956.

OTRU website. "OTRU Overview." Accessed on March 18, 2007. (http://www.otru.org/about_otru.html)

Pathammavong, Ratsamy (Tobacco Control Manager, The Lung Assocation), personal interview, August 17, 2006. Follow-up interview, February 15, 2007.

Puchtinger, Rolf (Epidemiologist, Manitoba Health), personal communication, February 22, 2007.

Rapid Risk Factor Surveillance System website. Accessed March 20, 2007. (http://www.rrfss.on.ca/Prevalence.aspx)

Richmond, Samuel B. "Statistical Analysis." New York: The Ronald Press Company, 1964.

RPad website. Accessed unsuccessfully August 4, 2006. (http://www.rpad.org/Rpad/)

RWeb website. "Using Rweb." Accessed on December 13, 2006. (http://www.jstatsoft.org/v04/i01/Rweb/node2.html)

Sandiford, Peter; Annett, Hugh; Cibulskis, Richard. "What Can Information Systems Do For Primary Health Care? An International Perspective." Social Science and Medicine, Volume 34, Number 10, 1992, pp. 1077-1087.

Schneiderman, Ben. "The Thrill of Discovery: Information Visualization for High-Dimensional Spaces." Distinguished Lecture Series presentation given at the University of Waterloo, October 11, 2006.

Schoonjans, F.; Zalata, A.; Depuydt, E.; Comhaire, F.H. “MedCalc: a new computer program for medical statistics.” Computer Methods and Programs in Biomedicine, Volume 48, 1995, pp. 257-262.

Seliske, Patrick (Epidemiologist, Wellington-Dufferin-Guelph Department of Public Health), personal interview, July 21, 2005.

SHAPES website. “About SHAPES.” (http://www.shapes.uwaterloo.ca/about/)

Shih, Bih-Yaw; Lee, Wan-I. “A Web-Based Computer Assisted Statistical Software Learning Environment.” Frontiers in Education Conference, Volume 2, 1998, pg. 587.

Silverman, Carol (Knowledge Exchange Officer, Cancer Care Ontario), personal interview, June 23, 2006.

Solomon, David J. “Conducting web-based surveys.” Practical Assessment, Research & Evaluation, Volume 7, Issue 19, 2001.

Spotfire, Inc. “Spotfire DXP Product Overview.” Spotfire, Inc., 2006.

SPSS Press Release. “New version of SPSS WebApp Framework expands platform for building Web-based analytical applications.” May 7, 2002.

SPSS website. “Software (Commercial Pricing)”. Accessed December 12, 2006. (http://www.spss.com/stores/1/Software_Full_Version_C2.cfm)

StatCrunch website. “StatCrunch Help.” Accessed December 13, 2006. (http://www.statcrunch.com/4.0/help/helpContent.html)

Suchan, Trudy A. “Usability Studies of Geovisualization Software in the Workplace.” ACM International Conference Proceeding Series; Volume 129, 2002, pp. 1-6.

Slysz, William D. “An Evaluation of Statistical Software in the Social Sciences.” Communications of the ACM, June 1974, Volume 17, Number 6., 1974, pp. 326-332.

Stutz, Al. "Process for Selecting Microcomputer-Based Statistical Software." ACM-SIGUCCS XV, 1987, pp. 355-360.

Taylor, Eleanor (Tobacco Control Program, Vancouver Island Hospital Authority), personal interview, June 26, 2006.

Thompson, Mary (Professor, Department of Statistics and Actuarial Science, University of Waterloo), personal interview, March 8, 2006.

Thriskos, P.; Zintzaras, E.; Germenis, A. "DHLAS: A web-based information system for statistical genetic analysis of HLA population data." Computer Methods and Programs in Biomedicine, 2006.

Tiessen, Janice (Data Manager, SHAPES, University of Waterloo), personal interview, July 21 , 2005.

Tomz, Michael; Wittenberg, Jason; King, Gary. "Clarify: Software for Interpreting and Presenting Statistical Results." Journal of Statistical Software, Volume 8, Number 1, 2003. pp.

U.S. Centers for Disease Control. "Preventing Tobacco Use Among Young People: A Report of the Surgeon General." 1994.

van der Meer, Matt (Data Analyst, SHAPES, University of Waterloo), personal interview, July 4, 2005.

van Lamsweerde, Axel. "Requirements Engineering in the Year 00: A Research Perspective." Proceedings of the International Conference on Software Engineering, 2000.

Venables, W. N.; Smith, D. M.; R Development Core Team. "An Introduction to R." Version 2.3.1, published June 1, 2006.

West, Webster R.; Wu, Yuping; Heydt, Duane. "An Introduction to StatCrunch 3.0." Journal of Statistical Software, Volume 9, Issue 5, 2004 (http://www.jstatsoft.org/v09/i05/scjss/).

Whitley, Elise; Ball, Jonathan. "Statistics Review 1: Presenting and Summarizing Data." Critical Care, February 2002, Volume 6, Number 1, 2002.

Wilkinson, Leland. "Practical Guidelines for Testing Statistical Software." Computational Statistics, Physica-Verlag, 1994.

Wilkinson, Leland. "Statistical Methods in Psychology Journals: Guidelines and Explanations." American Psychologist, August 1999, Volume 54, pp. 594-604.

Wolfram, Dietmar. "Applications of SQL for informetric frequency distribution processing." Scientometrics, Volume 67, Number 2, 2006, pp. 301-313.

Yasnoff et al. "Public Health Informatics: Improving and Transforming Public Health in the Information Age." Public Health Management Practice, Volume 6, Issue 6, 2000, pp. 67-75.

Zimmerman, Lynda (Ottawa Department of Public Health), personal interview, March 13, 2006.