Communications of the Association for Information Systems

Volume 3 Article 12

May 2000

# User Profiles as a Basis for an Electronic Statistical Consulting System

Ayelet Berg
*Tel Aviv University*, berg@post.tau.ac.il

Israel Spiegler
*Tel Aviv University and Claremont Graduate University*, spiegler@post.tau.ac.il

Communications of the Association for Information Systems

Volume 3, Article 12
May 2000

# USER PROFILES AS A BASIS FOR AN ELECTRONIC STATISTICAL CONSULTING SYSTEM

Ayelet Berg
Information Systems Department
The Leon Recanati Graduate School of Business Administration,
Tel Aviv University

Israel Spiegler
The Leon Recanati Graduate School of Business Administration,
Tel Aviv University
and
School of Information Science
Claremont Graduate University

spiegler@post.tau.ac.il

**RESEARCH METHODOLOGY**

# USER PROFILES AS A BASIS FOR AN ELECTRONIC STATISTICAL CONSULTING SYSTEM

Ayelet Berg
Information Systems Department
The Leon Recanati Graduate School of Business Administration,
Tel Aviv University

Israel Spiegler
The Leon Recanati Graduate School of Business Administration,
Tel Aviv University
and
School of Information Science
Claremont Graduate University

spiegler@post.tau.ac.il

## ABSTRACT

User profiles, which personalize applications, are an important factor in designing on-line computer systems. This study targets user profiles to better understand the interactive process of users and consultants in social science research studies. Definition of such profiles should be a basis for designing an electronic consulting system. The results of an empirical survey of both users and statistical consultants are presented. Questionnaires and interviews were used to identify the different perceptions of using statistics, analysis tools, typical hypotheses, interpretation of results, and related topics.

The findings shed light on the users of statistical systems and the role of the consultant in supporting the conduct of a research study. We found that the majority of users possess only a basic course in statistics, find difficulty in stating their problem, and two-thirds of them use consultants to define and design their study. The implication for an electronic consulting system are that such systems should pay attention to the initial research problem definition and should use a

system-initiated question/answer approach to elicit problem information from the users.

Based on the findings, design guidelines are presented for an electronic consulting system which can bridge the gap between the user and the wide range of statistical methods available in commercial packages.

**Keywords:** User profiles, electronic consulting, surveys, statistical and research methods.

## 1. INTRODUCTION

User profiling is a major thrust in computer studies where the goal is to personalize and tailor-make applications and systems. Profiling ranges from consulting systems to one-to-one marketing and to Web browsing [Adomavicius and Tuzhilin, 1999; Quiroga and Mostafa, 1999]. Electronic consulting systems are developed in such services as computer support and use [Johnson et al, 1989; Sheddon and Schroeder, 1991], in human service [Bogue, 1989], on line surveys [Kay and Johnson, 1999], and on line systems [Coppeto et al, 1989]. These systems are a direct result of the rise in end user and computer interactions.

One area that receives less attention is the use of profiling in statistics. Statistics is a key tool for empirical research. Methods and models of this applied scientific discipline are widely used in most areas of inquiry from the social sciences through the physical and life sciences and in data mining [Chung and Gray, 1999; Ramakrishnan and Grama; 1999]. The surge and availability of hardware and software also affected statistics, bringing it closer to the end user. More and more untrained people are applying and using statistical packages to their data for analysis and research.

While statistical packages and programs provide a multitude of models and tools, their use requires advanced training and an understanding of statistical methods, the assumptions underlying such methods, and their correct application. Users soon find out that they need technical skills before they can do meaningful work in applying appropriate method to their data. Lacking this knowledge, the users turn to experts - consultants - to guide them in formulating the problem, in designing the data collection instrument, in selecting an appropriate tool for data analysis, and in interpreting computer results.

The goal of this article is to outline the key players' profiles of a statistical consulting process. We believe such profiles are a first step in the design and development of an electronic consulting system to support users who want to apply statistics to their data. To this end, we administered questionnaires to and conducted interviews with 50 users and statistical consultants. Findings of this survey are presented in this paper. We gained insights into areas such as the phases of a statistical study, profile of the naive user and his or her knowledge in statistics, utilization of consultants, interaction between these two players, and more.

These findings are outlined and discussed together with guidelines for an intelligent consulting system in line with research done in artificial intelligence and statistics [Berg and Speigler, 1992; Charniak and McDermott, 1985; Gale, 1989].

Section 2 describes the process of statistical studies and reviews related research in profiling. Sections 3 and 4 respectively provide profiles of users and of statistical consultants obtained in our survey. Section 5 discusses our findings in terms of the implications for a computer system design to support users. The concluding section describes the contribution of this study and direction for future research.

## II. STATISTICAL STUDIES

Statistics helps users give a scientific description of events or collective phenomena. Usually, the goal of using statistics is to develop a quantitative representation of a certain problem or reality. Such representation may then take the form of a mathematical formula, graph, diagram, or a number that describes the phenomena. The ultimate objective of statistical research is to gain insight and identify rules that govern a problem in order to make some predictions about its behavior, a similar and indeed related goal of data mining [Chung and Gray, 1999]. To gain such insight, users collect data (by applying instruments, or extracting them from large existing databases), which consist of the relevant ingredients of a problem. Such data are usually coded into machine format, a file or database, to enable subsequent analyses and processing.

Statistical data analysis differs from study to study but can be characterized by four phases [Hand, 1986]:

1. *Data Description*: An initial exploratory step which portrays the distribution of data, their dimensions and their appearance, in order to gain first impressions of the problem and its related factors.
2. *Data Manipulation* includes performing transformations on data, grouping and classifying, handling exceptions and deviant cases, and more.
3. *Confirmatory Analysis*: This step involves formulating hypotheses about the problem, based on the data, and performing tests to confirm or reject each study hypothesis.
4. *Conclusions:* Following the confirmatory analysis step, and often in conjunction with it, this step involves the interpretation of statistical output, translating it into problem terms and issues, reaching conclusions, and recording them in the form of summaries, papers, and communications.

While these phases are by and large chronological, they influence one another. Each phase may affect the next one, and may cause repeating of, or returning to, any previous steps as well. In short, conducting a statistical study is a *recursive* and complex process that is usually ill structured. Moreover, the structure of any study is generally unique to that study, and in many cases is dynamic, another reason why users resort to experts for advice and consultation on how to conduct the statistical part of their studies.

When a consultant enters the picture, a new dimension is added - an *interaction* that takes place between the two key players: user and statistical expert. The user has the relevant knowledge about the problem domain, area, meaning of data, and control over the resources needed to perform the study. The expert brings in knowledge in statistical methods and experience in data analysis and application of tools. The interaction, then, becomes a mutual, two-way interchange of information.

Interests hardly agree, and yet the common goal remains unchanged: to analyze the data stored in the database and to obtain results that address the research questions posed at the outset by the user, often with the aid of the consultant.

Previous studies examined the process of user profiling and statistical consulting that may affect the design of a computer-based system to support the research process. At the London Institute of Psychiatry, data on the consulting process were obtained by taping the sessions of users and consultants and by following the two parties during the study [Hand, 1984; Hand 1985]. The findings of that study are the following:

- Consulting is an ill-structured process.
- There is a communication gap between users and consultants.
- Users have a difficulty in defining their objectives.

- Studies vary from area to area; therefore consulting should vary too, adapting to the particular area and problem.
- Leading questions help users define their problem. This technique is preferred for knowledge elicitation.
- Repeated listening to a recording of a session brings "second thoughts" in designing the analysis approach.

Another study sought to learn the consulting process to help build a rule-based expert system for medical research [Clayden and Croft, 1989]. Here too, data were collected by means of audio-visual recordings of user/consultant sessions. The findings of this study too, have repercussions on the guidelines of a computerized expert system for statistical consulting. In particular:

- The system should be the one asking the questions.
- Interaction is important to reach a common base of statistical terms and understanding of the research process.
- Statistical literature is neither simple nor user friendly. Thus, people prefer turning to other sources for understanding statistics.

The recommended system design for solving these problems is a model using Hypertext for statistical terminology that is context-sensitive.

Profiling is found in marketing, where studies seek to model users via rule discovery of an explicit participation of a human expert [Adomavicius and Tuzhilin, 1999]. Computer practice support via electronic consulting and on-line surveys are reported in Coppeto et al, [1989], Kay and Johnson [1999], and Sheddon and Schroeder [1991]. Human services employ consulting system as shown in Bogue [1989], and user profiling takes place in data mining studies [Chung and Gray, 1999; Ramakrishnana and Grama; 1999]. Other studies in this area are cited in Butler and Corter [1986], Hand [1986], O'Keefe [1985], and Thisted [1986]. Electronic consulting increased with on-line system availability and the Internet where such systems are used in help desks, customer support, system training, and service.

## III. USER PROFILE

In the course of this study we administered questionnaires and interviewed 50 individuals, equally divided between users and statistical consultants. The 25 users come from two main groups: universities and commercial institutions. They work in several areas of the social sciences. Because they are generally naive statisticians, they hire consulting services to analyze research data. Statistics consultants come from three areas: universities, research centers, and independent consultants. We also joined consulting sessions to learn the diverse problems of this dynamic process. The study was done in Israel.

Users are surely different from one another in style and attitude towards the process and data analysis of their studies. The prevailing notion that the majority is interested in statistics only for the purpose of validating their initial feeling concerning a given problem had to be verified. Indeed, we found that many perform statistical analysis out of "academic necessity", and treat statistics as a black box of which only the output matters.

To understand and draft a profile of users, we administered a questionnaire (Appendix A) and conducted interviews. The questionnaire pursued various elements of the research process: the stages of study, use of consulting, knowledge of statistics, typical methods used, understanding of statistical results, and interactions with the consultant. Selecting users to participate in our study required careful consideration and followed these criteria:

1. The user requires statistical tools for data analysis in a research study
2. The user or researcher worked with a statistical consultant
3. The area of study comes from the social sciences.

*Academic Use.* This group includes 13 researchers and graduate students who were randomly selected from a university list in areas of psychology, management, social work, and sociology.

*Social and Commercial Users*. This group includes 12 users, three each from banking, retail, service, and government institutions.

Thus, about a half of the users perform their studies on a routine basis, for purposes such as commerce or business. The other half consists of academic users such as faculty or students who perform studies as part of academic requirements (articles, dissertations and the like). A summary of topics and findings obtained from the user questionnaire is shown in Table 1.

Table 1. Summary of User Profile

| **Topic** | **Findings** | **Number** | **%** |
|---|---|---|---|
| Use of consultant during design stage of study | Yes | 17 | 68 |
| | No | 8 | 32 |
| Type of study | Academic, research | 13 | 52 |
| | Business and routine | 12 | 48 |
| Consultant used in data coding | Yes | 12 | 48 |
| | No | 13 | 52 |
| Knowledge of Statistics | Low (response 1-3) | 6 | 24 |
| | Medium (response 4-5) | 19 | 76 |
| | High (response 6-7) | 0 | 0 |
| Acquisition of statistics knowledge | Advance course in statistics | 8 | 32 |
| | Basic course in statistics | 13 | 52 |
| | Experience with consultant | 4 | 16 |
| Knowledge of tests performed | Yes | 18 | 72 |
| | No | 7 | 28 |
| Data analyses performed by: | User | 8 | 32 |
| | User with consultant | 9 | 36 |
| | Consultant | 8 | 32 |
| Statistical tests and analyses performed – ranked | Tables and coefficients | Rank: | 1 |
| | Frequencies and distributions | | 2 |
| | Regression | | 3 |
| | T-test | | 4 |
| | ANOVA | | 5 |
| | Factor analysis, confidence, or other | | 6 |
| Interests in statistical considerations | Yes | 6 | 24 |
| | No | 19 | 76 |
| Level of understanding of statistical output | Good understanding | 6 | 24 |
| | General understanding | 17 | 68 |
| | Vague understanding | 2 | 8 |
| Use consultant to interpret output | Yes | 19 | 76 |
| | No | 6 | 24 |

*Steps of Study.* Analyzing the responses, we learned important aspects about the process of conducting a statistical study. The questionnaire asks users to state the different steps taken in carrying out their study. Here, a clear distinction is drawn between academic and commercial users. The different steps taken by user types in their studies are given in Table 2.

Table 2: Academic vs. Commercial Study Phases

| **Steps in Academic Study** | **Steps in Commercial Study** |
|---|---|
| 1. Literature survey and theoretical background | 1. Data definition |
| 2. Formulating an idea, problem definition | 2. Problem and hypothesis formulation |
| 3. Search for a target population | |
| 4. Instrument design | |
| 5. Preliminary test and instrument calibration | 3. Preliminary test, method selection |
| 6. Data collection with instrument, interviews | 4. Normally use existing data source |
| 7. Data analysis | 5. Data analysis |
| 8. Interpreting results and conclusions | 6. Interpreting results and conclusions |

The commercial study follows a different pattern due to availability of a database. This is evident in data mining studies [Ramakrishnan and Grama, 1999] that are usually done on existing and abundant data. When data for the study are not available, the commercial study follows the lines of the academic study steps with the exception of step 1 - the literature survey - which is usually omitted in commercial research.

*User Profile.* The survey shows that over two-thirds (68%) of users use consultants in their studies; 76% of the participants have a medium level of statistical knowledge; and 52% acquired that knowledge in a basic course. (Table 1). Nearly a third (32%) of the users let consultants perform the analysis for them. Such a profile is an essential consideration for designing an electronic consulting system.

*Statistical Methods.* A central part of the survey asked respondents to rank statistical tests and methods. The results obtained (see Table 1) are significant for understanding the research process and for the design of a computer consulting system. The statistical tools used by users in their studies are:

1. Tables and Coefficients. (Gamma, Chi-square, Pearson): 52% of the users ranked these tools in first place, and the other 48% ranked them in second place.
2. Frequencies and Distributions. 48% of the users ranked these tools in first place, 16% placed them second, and 36% did not mention them among tools they commonly use.
3. Regression. 12% ranked regression in first place, 24% put it second, 16% put it third, 16% fourth, 16% fifth, and 16% did not mention it among tools they use for data analysis.
4. t-test. This tool was ranked second in frequency of use by 16% of the surveyed users, 16% ranked it third, and 16% ranked it fifth. The rest of the users (52%) made no mention of the T-test at all.
5. Analysis of Variance (ANOVA). 16% placed it as their third most frequently used tool, and 16% put it fourth. The rest did not mention this form of data analysis.
6. Other tools. Other statistical tools mentioned infrequently in the responses are factor analysis, discriminant analysis, and confidence test.

These findings are checked and validated by way of another question in the instrument (Appendix A, number 12, not shown in Table 1) dealing with hypotheses formulation and the types of tests and methods used. The methods related to hypotheses testing types are:

- Data description, general impressions, scatters, averages, ranges, and frequencies (frequency and distribution tools)
- Relationships among variables, relationship between dependent and independent variables, strength and direction of relationships, common

frequencies, market shares, population identification (tables and coefficients tools)

- Regression – the influence of other variables on an index, explaining certain symptoms, explain dependent variable, forecast
- Compare populations and samples (T-test)
- Interactions among variable (discriminant analysis)
- Confidence testing (Spearman-Brown)
- Grouping and classification (factor analysis).

*Interest in Statistics.* Hardly surprising were the responses to the question about interest in statistics. The survey indicates that 76% of the users surveyed have no interest in statistical considerations, method or model. Despite the fact that 32% had an advance course in statistics, only 24% are interested in such considerations. This is consistent with the finding where 92% of the consultants replied that users have little interest in statistics (see Table 4 in Section 3).

*Understanding Results.* We also asked about interpreting computer output (Appendix A questions 14-16). While 68% indicated general understanding of output, still 76% turn to consultants for interpreting the computer output. We also asked which part of the output is of interest and the responses are summarized in Table 3.

Table 3: Computer Output Usage of Statistical Methods

| **Method/Test Applied** | **Elements Actually Used In Computer Output** |
|---|---|
| Tables | Percentages, Summaries, Pearson, Chi-square, Gamma |
| Frequencies and Distributions | Averages, standard deviation, range, percentages |
| Regression | Confidence, $R^2$, significant variables, regression coefficients, F values of variables |
| t-test | Averages, standard deviation, t, confidence intervals |
| Analysis of Variance | Averages, sizes, cumulative size, main effects, interactions |
| Factor analysis | Contribution of each variable to factor |

Only 24% of the users surveyed stated they have a good understanding of statistical results and a large majority (76%) use consultants to interpret the results (see Table 1). Regarding packages used by users, most use what is available or accessible to them at work or school. Most use PCs, though about 24% have access to mainframes for their databases.

## IV. CONSULTANT PROFILE

Our study also surveyed the statistical consultants. Consultants provide advisory services to researchers and users in performing their studies. They are obviously different from one another in style and work methods; at times the same consultant may use different methods, depending on the user, the study, or data available. Such a flexible approach is magnified by the behavioral characteristics of the user and the interaction between the two parties.

We administered a questionnaire to consultants (Appendix B) and interviewed them. All participants were told that the target of our work was to gain a better understanding of statistical consulting to help design an intelligent system that could act as a statistical consultant. While some consultants expressed satisfaction with the notion that work methods might be automated, others were taken back by the mere idea that a statistician might be replaced by a computer. They claimed that "such a complex and involved task can't be done satisfactorily without compromising on statistics".

The 25 statistical consultants who participated in our survey came from three populations:

*University Consultants.* This group consists of 15 consultants, who were sampled randomly from lists of statistical consultants obtained from three universities.

*Research Institutions*. This group of 4 statistical consultants were selected randomly from three commercial research institutions.

*Independent Consultants.* This group consists of 6 consultants who were selected randomly from a list of private consultants who advertise their services in various media.

The 25 consultants (the same number was used for users to obtain a balanced study) surveyed work in a variety of areas ranging from medicine, to the physical as well as the social sciences. Table 4 summarizes the consultant profile.

The first questions of the consultant questionnaire relate to the information discussed with the user in the first session. Those issues are vital as they establish direction and structure in designing the study. Our survey shows that most consultants seek the same types of information, which may be divided into five categories:

*Description of Subject Area*. The consultant asks users to describe the subject area that includes background on the problem, method of data collection, unit of sampling or measurement, information about the population, and possible tests or experiments.

*Research Problem.* The user is asked to state the question(s) to which he or she seeks answers by the proposed study. Consultants employ a variety of ways to elicit this information (the questionnaire provides for open-end wording). Some users come prepared with a series of well-defined study questions; others have a general idea of the research question; and still others have not defined the research problem for themselves. This observation is consistent with

Table 4. Summary of Consultant Profile

| Topic | Findings | Number | % |
|---|---|---|---|
| Principal purpose of consulting: | Description of subject area | 10 | 40 |
| | Definition of research problem | 4 | 16 |
| | Research objective | 4 | 16 |
| | Dependent and independent variables, links | 2 | 8 |
| | Going over study instrument | 5 | 20 |
| Types of problem: | Variables and relationships | 12 | 48 |
| | Understanding the data | 4 | 16 |
| | Comparing populations | 4 | 16 |
| | Prediction, models, time series | 5 | 20 |
| Database description: | Sample, units, time dimension | 4 | 16 |
| | Measurement levels | 4 | 16 |
| | Variables (what and how measured) | 4 | 16 |
| | Scales and ranges | 4 | 16 |
| | Study variables selection | 4 | 16 |
| | Missing, unknown variable, etc | 5 | 20 |
| Statistical tools used, ranked by frequency | Frequencies and distributions | Rank: | 1 |
| | Coefficients, Chi-square | | 2 |
| | Regressions | | 3 |
| | T-test and ANOVA | | 4 |
| | Time series | | 5 |
| | Recursive models, factor analysis | | 6 |
| Consultant role in leading research directions | Propose new directions | 14 | 56 |
| | Concentrate on original problem | 5 | 20 |
| | Change research problem based on results obtained | 6 | 24 |
| Consultant view on users interest | User interested only in operational issue | 23 | 92 |
| | User interested also in statistics | 2 | 8 |
| Participation in analysis process | User participate in process | 20 | 80 |
| | User participates by consultant initiation | 4 | 16 |
| | No user participation | 1 | 4 |

Campbell's [1982] that most people cannot describe clearly and directly what they want to do in their research.

*Objective of Study.* When asking for the objective of the study, some consultants expect answers that are similar to those given for the previous category. But some consultants are more specific in trying to get an idea whether the study is aimed at academic work, publishing a paper, commercial use, proof of a thesis, or just information.

*Variables and relationships.* These issues raised by the consultants are specific in nature and usually assume the research study seeks to test or explain a certain phenomenon.

*Instrument.* Often, the consultant's initial action is to go over the instrument used for data collection. This review involves examining the structure of the questionnaire, its internal consistency, the relation between the subject area and the questions as stated in the instrument, for accuracy, scale of measurement, and the like.

While the actual mode of the consultant and user interaction may vary, the meaning and intent are clear and we had little difficulty in categorizing it. We found that 48% of participants note the relationship between variables as the most typical problem raised by users in their initial meeting.

*Methods and Tools.* In asking consultants to rank the types of analyses they perform, it was clear that frequencies and distributions are the most common. Only a few placed them in second place. Similarly, tables and Chi-square were ranked in second place. The findings, as shown in Table 4, are indicative of processing demanded by users and are consistent with the user profile discussed in the previous Section.

More than half (56%) of the consultant said that they are the ones who propose new directions of research during the consulting process. The rest concentrate on the original problem and suggest changing it only when an unexpected outcome takes place.

Consultants point to a low interest in statistics on the part of users. A compelling majority (92%), indicate that users are interested mainly in operational results and hardly at all in the statistical theory behind them. In response to the same question, the majority of the users (76%) stated they have no interest in statistics beyond their immediate problem. As mentioned above, most users regard statistics as a black box of which only the output matters. Still, 80% of the statistical consultants said that users participate in the process of data analysis (Table 4).

The perception of statistics and the consulting process as emerging from our survey provides an additional dimension to the participant's profiles of a statistical oriented study.

## V. DISCUSSION

Comparing the different perceptions of users and consultants about a statistical study leads to the following observations.

*Commercial vs Academic Users.* Most users in commercial and business firms do their statistical analysis themselves, whereas academicians and students tend to rely on experts to perform their data analysis. This is evident from Table 2. and may be attributed to greater efficiency, attention to cost factors, and experience of the former group.

In interviews that accompanied the study we learned about the difference of the level and sophistication in statistical analyses used by commercial institutions and those used by academic users. The former tend to apply basic methods whereas the later apply more elaborate statistical tools. This difference changes with data mining tools such as clustering, classification, and neural nets that are mainly used today at the commercial level [Chung and Gray, 1999].

Another difference between commercial and academic users of statistics is in focus. The commercial consultant tends to stay within the realm of the original problem statement whereas the academic consultant is more eager to point to new directions and methods of analysis.

*Eliciting Information.* Using the findings in Table 4 and data from the consultant questionnaire (Question 11, Appendix B), we identified two methods and styles used by consultants in obtaining primary information from users in interviews:

1. Consultant asks the user to describe the problem, and

2. Consultant uses a question/answering technique to lead the user to define the problem, data, and other characteristic of the study.

*Selecting Statistical Tools.* While the process of selecting the statistical tools to be applied in a study is relatively structured, "tuning" is needed for each study according to the particular area, resources, experience, style, and intuition of the parties conducting the research. The majority of consultants (56%, Table 4) propose new directions of study.

*Mode of Presentation*. While all statistical studies obtain printed computer output, researchers indicate in interviews a desire for graphical representation of data either during analysis stage, or in presenting the results.

*Interpreting Results.* Here we found a wide gap between the normative approach, what the consultant has to offer and what the user expects to get in reality. Users generally prefer simple output which they can understand (Table 3). The information actually used by users is only a small part of the overall output produced by the statistical tool. Many users indicate in interviews that consultants are not aware of how little they know about interpreting statistical results. Others point out too that consultants show lack of patience in explaining the collection of figures and statistics contained in the computer output.

*Use of Statistical Literature.* A point that surfaced in interviews, and is not apparent from the data in questionnaires is the lack of support users obtain from the literature. The higher the user ranks her or his level of knowledge of statistics the less they tend to seek help from experts. Yet, most users surveyed did not turn to the literature for help. Refraining from the literature may be an indication of the gap between the exposition and demonstration in statistical books and the ability of naive users to comprehend the material. Closing this gap is at the heart of a successful computerized system to support users in statistics, which must include clear explanations of basic terms and methods.

*Interest in Statistics.* Most users (76%, Table 1) have little interest in statistics per se, beyond its use as a means to analyze their data. Users are, thus, hardly involved in the theoretical considerations of selecting the appropriate method or model, preferring to let the consultant perform the analysis for them. This attitude may result from many seeing statistics as a burden, or to their general lack of interest in the subject.

## IMPLICATIONS FOR ELECTRONIC CONSULTING SYSTEM

Several implications and guidelines follow from the above discussion and the findings of our study. These should be taken in consideration by designers of electronic consulting system, paying particular attention to users untrained in the discipline of statistics. With the increase in statistically oriented studies in Information Systems, the empirical work of researchers and students, as well as commercial use of statistics in data mining, such guidelines are timely and relevant.

1. An electronic consulting system should aim at naive and intermediate users in understanding statistics. User profile shows that more than half (56%, Table 1) have only a basic course in statistics. Consulting should therefore be structured to support the researcher through the entire process. Aside from being friendly, which any system should be, a

statistical support system should be general in scope and able to help users at different levels of statistical knowledge.

2. Users generally don't know what they want and have a hard time describing the problem domain in which they work. This finding, evident from Table 1 showing that 68% use consultants to define and design the study, suggests that the electronic consulting system
   - Elicit the knowledge from users in a question/answering session coupled with explanations of statistical terms (e.g., interactive, with context sensitive HELP).
   - Pay attention to the initial research hypothesis; it is critical for subsequent stages and will narrow the problem space and range of relevant applicable methods.
3. The systems should provide a structured 'walk through" of the entire research process life cycle. In Table 4, 56% of the consultants indicate the purpose of consulting is describing the domain and research problem, up to interpretation of output results.
4. Most users want simple, standard data analysis methods and easy-to-interpret output (Table 3). Graphics is a must and should accompany all phases of the study.
5. To avoid information overload, minimize the output presented to users to show only relevant facts and filter out parts meaningless to the user, preferably by personalizing the tool to the user.
6. An electronic system should be able to propose additional avenues of analysis that the user may not have considered at the outset. Fifty-six percent of consultants see their role as pointing to new directions as part of the support they provide (Table 4). This aspect, supporting the discovery of new insight is a desirable feature of a system, in line with the Knowledge Discovery in Database (KDD) methods of recent years.

## VI. SUMMARY AND CONCLUSIONS

This paper presents an initial step in the direction of defining an electronic consulting system for the use of statistics in research studies. Profiles of users and consultants engaged in statistical studies were outlined by means of an empirical survey. The objective of the survey was to gain an understanding of the statistical consulting process, the different perceptions the two key players have, the tools and methods used, and the way computer results are interpreted. The profiles should aid designers and vendors of on-line surveys and computer advisory systems that can support users in the social sciences in performing their statistically oriented studies.

Applying statistics to empirical data is an unstructured process as is the interaction between users and statistical consultants. Yet, based on our study, basic guidelines for the design of a support system can be outlined. Targeting an electronic consulting system to a confined user profile and related implications of statistical use may better structure hypothesis formulation, selection of appropriate statistical models and tools, and interpretation of results.

Several findings came from the user and consultant profiles. We observed little interest in statistics on the part of users beyond the operational data analyses needed in their study. Such low interest in statistics, despite its wide use in the social sciences, was evident in both the user and consultant questionnaires. Another finding is the information overload in statistical output with facts and factors, many of which are of little use to users. Users find the output difficult to interpret without a consultant. Despite that difficulty, few users turn to the literature for help -- an indication of a gap between the presentation of statistical materials in books and the ability of the general user to understand them.

Further studies are suggested in the area of user profiling and personalization. One direction is to look more at the behavioral aspects and interaction, sometimes conflicting, between users, naive in statistical methods, and consultants who are experts in methods. Another direction, in line with personalization of many computer applications and Internet access, evident from "cookies" and related data collection, may be finding better ways to tailor-make on-line systems of supporting statistical use in social sciences by setting them to specific levels of users, problem domain, or nature of data available for the study.

Editor's Note: This article was received on March 6, 2000. It was with the authors approximately 1 month for one revision. It was published on May __, 2000

## REFERENCES

Adomavicius, G., and Tuzhilin, A., (1999) "User Profiling in Personalization Application Through Rule Discovery and Validation", Proc. Of ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 15-18, San Diego, CA.

Berg, A., and Spiegler, I., (1992) "Intelligent Knowledge Generation from Statistical Data Bases", Journal of Information System Technology (2)1, pp. 1-11.

Bogue, E.L., (1989) "The Difficult Client: Consulting Techniques from the Human Services", ACM SIGUCCS Conference on User Services, October 1-4, Bethesda, MD.

Butler, K.A. and Corter, J.E., (1986) "Use of Psychometric Tools for Knowledge Acquisition: A Case Study", in Gale, W.A., (Ed.) Artificial Intelligence and Statistics, Reading, MA:Addison Wesley.

Campbell, J.P., (1982) "Some Remarks from the Outgoing Editor", Journal of Applied Psychology", (67)6, pp. 691-700.

Charniak, E., and McDermott, D., (1985) Introduction to Artificial Intelligence, Reading, MA: Addison Wesley.

Chung, H.M,. and Gray, P., (1999) "Data Mining", Journal of Management Information Systems, (16)1, pp. 11-16.

Clayden, A.D. and Croft, M.R., (1989) "Statistical Consultation - Who's the Expert?", Proc. of 2nd International Workshop on Artificial Intelligence and Statistics, January.

Coppeto, T.J., Anderson, B.L., Geer, D.E., and Treese, G.W., “OLC: An Online Consulting System”, ACM SIGUCCS Conference on User Services, October 1-4, Bethesda, MD.

Gale, W.A., (1989) "Statistical Applications of Artificial Intelligence and Knowledge Engineering", Knowledge Engineering Review, (2).

Hand, D.J., (1984) "Statistical Expert Systems: Design", The Statistician, (33)10 pp. 351-369.

Hand, D.J., (1985) "Statistical Expert Systems: Necessary Attributes", Journal of Applied Statistics, (12)1.

Hand, D.J., (1986) "Patterns in Statistical Strategy", in Gale, W.A., (Ed.) Artificial Intelligence and Statistics, Reading, MA: Addison Wesley.

Johnson, E., Wehrs, W., Delfield, T., Imhoff, J., and Manter, V., (1989) “Consulting without Consultants: Expert Systems Applications in User Services”, ACM SIGUCCS Conference on User Services, October 1-4, Bethesda, MD.

Kay, B.K., and Johnson, T,J., (1999) “Research Methodology: Taming the Cyber Frontier – Techniques for Improving On Line Surveys”, Social Science Computer Review, (17)3, pp. 323-337.

O'Keefe, R., (1985) "Logic and Lattice for Statistical Advisor", Thesis, University of Edinburgh.

Quiroga, L.M., and Mostafa, J.,(1999) “Empirical Evaluation of Explicit versus Implicit Acquisition of User Profiles in Information Filtering System”, Proceedings of Fourth ACM Conference on Digital Libraries, August 11-14, Berkeley CA.

Ramakrishnan, N., and Grama, A.Y.,(1999) “Data Mining: From Serendipity to Science”, IEEE Computer, (32)8, pp. 34-37.

Sheddon, M., and Schroeder, W., (1991) “Electronic Consulting: Software Tools to Enhance Consulting at the San Diego Supercomputer Center”, Proc. Of

19th ACM SIGUCCS Conference on User Services, November 3-6, 1991, Seattle, WA.

Thisted, R.A., (1986) "Representing Statistical Knowledge for Expert Data Analysis System", in Gale, W.A., (Ed.) Artificial Intelligence and Statistics, Reding, MA: Addison Wesley.

## APPENDIX A

## USER QUESTIONNAIRE

Name:____________________ Date:___________________

1. What are the steps in conducting a study/test/survey?

   ________________________________________________

   ____________________________________________________________

2. During the study design, do you use a statistical consultant? (yes/no/why) ______________________________________________

   ____________________________________________________________

3. What are the objectives of your statistical processing?

   ____________________________________________________________

4. Data are put on magnetic media by:
   1. Yourself
   2. Yourself with statistics consultant or programmer
   3. Keyboard operator
   4. Keyboard operator and consultant

5. In storing data, do you design the file structure according to subsequent types of processing you intend to do? (yes/no)

6. How do you rank your knowledge in statistics? (1=novice, 7=expert) _______________________

7. How was that knowledge acquired?
   1. Advance course in statistics

2. Basic course in statistics

3. From working with consultants

4. Self learning from books

5. Other (state)____________________

8. Does your knowledge in statistics focus mainly on the tests you routinely perform? (yes/no)

9. In your study, do you

1. Perform statistical analysis yourself

2. Work with a statistical consultant

3. Let the consultant perform the analysis him/herself

4. Other (state) ____________________

10. Which option (of question 9) do you prefer? _______________

why? ________________________________________________

11. What are the four statistical tests/analyses you usually do?

(from more to less frequent)

1. _________________ 2.___________________

3. _________________ 4.___________________

12. For each test/analysis you stated in 11, which hypothesis do you test?

1. __________________ 2.___________________

3. __________________ 4.___________________

13. Are you interested in statistical assumptions/considerations when selecting the test or model? (yes/no)

14. State for each test in question 11, which parts of the output are of interest to you?

1. ____________________________________________

2. ____________________________________________

3. ____________________________________________

4. ____________________________________________

15. Do you

1. Understand the statistical meaning of the output

2. Understand the meaning only in your problem context

3. Understand vaguely only parts of output

4. Other (state) ________________________________

16. Do you turn to a consultant to interpret the results of computer output? (yes/no), why __________________________

17. Can you say that the consultant clarifies the ambiguities (yes/no/partly)

18. Communication problems between you and consultant result from:

1. Unfamiliar statistical or mathematical terms

2. Different points of view or ways of thinking

3. Both

4. Other (state) ________________________________

19. State up to <u>four</u> statistical packages, which you personally use

1. ____________________ 2. ______________________

3. ____________________ 4. ______________________

5. None used in person

20. Why are you using those packages (question 19)

_____________________________________________________

_____________________________________________________

## APPENDIX B

## STATISTICAL CONSULTANT QUESTIONNAIRE

Name:________________________ Date:________________

Areas of Consulting: ___________________________________________

1. What are the questions you ask the user at the outset in order to decide on the test or analysis?

1.__________________________ 2.__________________________

3.__________________________ 4.__________________________

2. State six typical answers to the above questions and the type of test you will recommend for each.

1.________________________ 2.________________________

3.________________________ 4.________________________

5.________________________ 6.________________________

3. In your opinion, the strategy for statistical method selection should be establish according to:

   1. Nature of data
   2. Research problem
   3. Combination of both
   4. Other (state)________________________

4. Do cases you meet in reality match your answer to question 3?

   1. Yes 2. No, why________________________________

5. Assuming the user will initially describe the data file, what is the mandatory information at this stage?

   ____________________________

   ___________________________

   ____________________________

   ___________________________

6. After the user described the data for analysis, would you suggest operations for data description (e.g. frequencies)

   Usually Yes No

   Reasons:____________________________________

7. Which part of the frequency output usually requires explanation?

1.__________________________ 2.________________________

3.__________________________ 4.________________________

8. What other data description tools you usually recommend?

1.__________________________for what________________________

2.__________________________for what________________________

3.__________________________for what________________________

9. State six types of the most common processes used (in order of frequency of use)

1.______________________________ 2.______________________________

3.______________________________ 4.______________________________

5.______________________________ 6.______________________________

10. In which conditions would you recommend the use of those processes?

1.______________________________ 2.______________________________

3.______________________________ 4.______________________________

5.______________________________ 6.______________________________

11. At times the selected analysis method is inappropriate. State for each of methods/process in question 9 when will you change to another and which other method would be selected?

1.______________________________other______________________________

2.______________________________other______________________________

3.______________________________other______________________________

4.______________________________other______________________________

5.______________________________other______________________________

6.______________________________other______________________________

12. Based on your experience, what is the number of statistical tests commonly performed on a data file?_________________________

13. What are the typical things you explain users in each of the processes mentioned in question 9?

1.______________________________ 2.______________________________

3.______________________________ 4.______________________________

5.______________________________ 6.______________________________

14. During the consulting process, are you?

1. Suggesting new directions based on obtained results
2. Match the analyses to the original problem only
3. Change the research problem based on obtained results

15. Are users

1. Interested in the statistical meaning of the analysis
2. Interested only in the operational issue
3. ________________________________________

16. Do users take part in your considerations when selecting the analysis method?

Yes/No, why___________________________________________

## ABOUT THE AUTHORS

**Ayelet Berg** holds undergraduate and graduate degrees in Economics and Statistics from the Hebrew University in Jerusalem, and a Ph.D. in Information Systems from Tel Aviv University. She is a veteran statistician, who worked for many years at the Central Bureau of Statistics of the Israeli government where she held positions as senior research statistician and analyst. Her work concentrated on the use of computers and information system in statistics. She taught at the Hebrew University and Tel Aviv University, and is involved in statistical databases.

**Israel Spiegler** is an associate professor and past chair of the Information Systems program at Tel Aviv University, Graduate School of Management. He holds an M.Sc and Ph.D degrees in Computers and Information Systems from UCLA. He was on faculty of Boston University, and is currently a visiting professor at the School of Information Science Claremont Graduate University. His areas of research are databases, artificial intelligence, and system analysis.