Data and Statistical Literacy for Librarians

by Ann S. Gray[1]

Introduction
Libraries are full of publications containing statistics and most offer databases with indexes to statistical publications. Statistics are important: they are essential for social and economic development; for understanding among peoples; and necessary for any society that seeks to understand itself and respect the rights of its citizens.

In order to provide a high level of support and service for these resources, librarians benefit significantly from having an interest in data and statistical literacy. The goal of the librarian is to direct users of electronic data and statistical resources toward useful information that reflects the nature of the real world and to help users avoid the possible misuse of data and statistics. Thus, as librarians, we need to:

- understand statistical publications and electronic resources of national and international statistical organizations which are the primary source of most statistics;
- deal with statistical publications and value-added commercial products which may actually hide statistical details from us;
- be able to judge the different uses the media make of statistics;
- and be able to make informed decisions regarding the use of mapping and other display tools such as charts, graphs, and other presentations of statistics.

This paper looks at why data and statistical literacy for librarians are relevant, how data and statistical resources are evaluated, the types of information about data and statistics that one needs to know to provide assistance in the use of statistical resources, and the possibility that training in statistics would be useful, if not necessary, to providing these services[2].

The Importance of Statistical Literacy
Data and statistical publications in electronic format were introduced in to U.S. academic libraries in the mid 1980s by U.S. federal statistical agencies. By the early 1990s data held on CD-ROM were fairly common in most libraries in the U.S. With this came the the growth of data services as a function of the university library. Data in its many forms had a foot in the door at many libraries and libraries were rapidly embracing computer technologies. In the U.S., 1996 saw the beginning of a loosely organized approach to centralized delivery of statistical information from the various federal statistical agencies. Lead by the Census Bureau, an Internet service known as FEDSTATS was set up to provide one-stop shopping for accessing federal statistics[3]. In Canada, progress was slowed by Statistics Canada's steep price increases in its data products during the 1980s. However, the Data Liberation Initiative (DLI) was launched formally in 1996, which contributed significantly to the growth of data support services within Canadian libraries.[4]

With the delivery of statistical tables using the Internet, there was a lot of concern over the potential misuse or misunderstanding of statistical data that would appear electronically without explanation. When summary tables and statistical reports were in print, it was expected that the report would include information about the tables that would allow for an evaluation of the information. But when tables became separated from the text and when the data became survey microdata, the question of utility and quality became more complex. Not only are data resources being delivered directly to the analytically 'unsophisticated' public, they are also being grabbed by the mass media and used in ways that may not always be honest or useful. Groups such as the Statistical Assessment Service (STATS) and the Center for Media and Public Affairs deal with the misuse of data by that sector, and many statistical organizations and government agencies are trying to solve the problems posed by direct access to statistical resources by establishing policies that reduce the chance of misuse.[5]

There are several reasons why statistical literacy has become important recently. We live in the Information Age with rapid distribution of news and content, where content is often overlooked in favor of images, and at a time when more and more statistics and data products are being made available to a larger and less data-literate audience.

Olenski (2003) suggests that we are also living in a time in which new democracies are emerging and recognizing the functions their national statistical agencies ought to play in the life of their countries and the global community[6]. In 1993 the United Nations' (UN) Economic Commission for Europe (ECE) adopted a list of fundamental principles of official statistics later endorsed by all as being of "universal significance."[7]

Noting the importance of official statistical information for development and the necessity that public trust should be based on scientific principles and professional ethics, the ECE adopted 10 principles. The fundamentals recognize the importance of the quality of the statistics but emphasize that correct interpretation depends on the presentation of information on the sources, methods and procedures of the statistics and that agencies are entitled to comment on erroneous interpretation and misuse of data. Agencies should protect the human rights of individuals, laws and regulations that govern the agencies should be public, and standards and concepts should be established to ensure comparability. The principles point to quality, timeliness, costs and respondent burden as key considerations in data collections.

Statistical literacy is also a concern of the International Statistical Institute (ISI), which has its own Literacy Project[8], and many national statistical organizations, such as Statistics Canada, the U.S. statistical agencies, and international agencies such as the International Monetary Fund (IMF), have policies and guidelines to support both quality and utility for statistics.

Librarians already act as intermediaries to statistical resources within their libraries and it seems logical that they could serve the same function for supporting a broader range of data and statistical products. Looking at how statistical publications are evaluated, there is a natural progression from this evaluative process to other quality measures that form part of the knowledge base of statistical literacy.

Evaluation of Statistical Publications
Print resources are evaluated based on their quality and objectivity, and on their utility and value. The publisher, author or peer reviewer is often used to determine quality and objectivity. The relevance, timeliness, scope, coverage, and presentation of the information are criteria that are typically used to judge utility and value of the piece within the context of the collection of the library and the purpose for holding the publication. This does not differ substantially from the guidelines set up by Statistics Canada in its *Elements of Quality* and its longer document *Quality Guidelines*.[9]

Statistics Canada Quality Guidelines
Statistics Canada stresses the relationship between quality and utility. In order to judge quality and determine utility one needs accuracy indicators. These are possible with a presentation that fosters understanding. The agency accepts responsibility to provide sufficient information about its statistics so that one can determine if the data or statistic can be used in a specific application or for a specific purpose. Documentation on methodology and on quality are vital to data analysis. Because neither the statistical agencies nor the data provider or data supporter can anticipate every future intended use of data, we must rely on the information already supplied about the data. Statistics Canada supplies information that covers the areas of relevance, accuracy, timeliness, accessibility, interpretability, and coherence. Still, statistical literacy is necessary to interprete the information that is provided in order to determine if the statistic is right for a specific purpose or application.

It is possible to rate quality in terms of accuracy using both expert opinion and measures of data accuracy. Accuracy measures typically describe error, such as bias or systematic error and variance, also known as random error. Documentation can also be judged based on its completeness and accuracy. "Statistics Canada's "Highlights, interpretations, statistical test results, and statements of trend, change or significance" are other ways to provide users with a notion of quality.[10] Interpretability, another cornerstone of statistical literacy, requires information necessary to interpret and utilize statistical information appropriately. This covers the concepts, variables, classifications that are used as well as the methodology of data collection and the accuracy indicators and statements relating to the strength and weaknesses of the data. Coherence refers to the use of standardized terms, classifications, and concepts among the data products, and can be described by information on how classification schemes are related to each other, e.g. SIC and NAICS, standard recodes and, if possible, aggregations.[11]

Quality Guidelines U.S. Census Bureau
The quality guidelines of the U.S. Census Bureau are very similar in nature[12]. They also include utility and objectivity as goals for the agency, state that they will provide indicators of quality in a timely fashion and indicate that the analytic results of their products should be reproducible following the prescribed methodology (i.e., exact same method of data collection and analysis). That said, in reality it would be impossible for users to actually reproduce the surveys since the underlying microdata would not be available due to confidentiality issues. Other U.S. Statistical agencies, following the direction of the Office of Budget and Management, have also tried to create policies and sites that promote good data practices[13].

Thus we can look for statements regarding accuracy and methods in order to interpret the utility of statistical resources, but is that enough? Some of the quality

measures that these agencies see as being important are given below. Because so much of our current data comes from sample surveys, the quality of the survey method plays an important role in the resulting quality of the data. The following measures are important:

- sample size is often regarded as the most likely indicator of quality reported in surveys. All things being equal a larger sample will produce more reliable results with less margin of error;
- the sample frame needs to be current and contain information necessary to draw the sample;
- response rate remains important even though there are studies that claim it may not. Many survey methodologists still regard a high response rate as necessary to reduce nonresponse error: in other words, would those in the sample who did not respond have responded with the same variation as those that did respond?;
- sampling error refers to deliberate undercounts or overcounts of groups of persons in the sample;
- coverage error looks at the omission of sets of the population from the sampling frame (such as takes place when phone surveys omit non-phone households);
- quality of inputs are important for derived or computed measures, such as estimates and indexes;
- other types of error, for example, measurement error and processing errors, are less easy to detect.

Regardless of the format of the publication, this information may or may not be available. In print sources, we often look for other types of quality indicators.

Two Commercial Print Publications

Two commercial publications were examined for quality indicators including: sources of data; sources of error; assumptions; adjustments; comparability; definitions; and methods. The first, *International Smoking Statistics,* is a collection of historical data from thirty economically developed countries[14]. It includes statements regarding coverage, sources of data, brief notices of possible sources of errors, time period to which the data refer (e.g. midpoint year), and a brief description of the survey scope and methodology, if available. It further covers adjustments made to render the data comparable among the constituent nations, and provides information on the length of cigarettes and thus the amount of tobacco within, and whether hand-rolled or manufactured. However, the publication is associated with the Wolfson Institute of Preventative Medicine , an organization that aims to reduce smoking. This could lead one to question its objectivity.

The second example, *Complete Economic and Demographic Data Source* is both a book and a set of tables available on CD-ROM[15]. The print version contains an entire chapter devoted to how estimates and forecasts are created, as well as definitions and sources. However, this chapter and the notes are not present in the CD-ROM version of the tables.

These two resources provide adequate information on sources, assumptions, comparability, definitions and methods provided one knows how and where to look for this type of information and how to evaluate it.

Data Resources on the Internet

Another frequent source of statistics are websites. An example is voter registration data assembled on a site at the University of California, Berkeley. If one begins at the webpage for Statewide Information by Assembly District[16], there is little indication of how the data came about. Data on registration and voting by race is not collected by the state of California nor the U.S. government, but the site has a report and data on the number of registered minorities in the 1996 California elections. The project homepage provides something of an introduction into this large project by a reliable source but there is no link from this page to its home. It appears that voter registration rolls furnished the name, location of residence, political party affiliation and vote history (if the person voted or not in various elections) The classification of ethnicity was based on an analysis of surname.

The validity of the latter type of classification should be further investigated by users, for example finding out more about the methodology of using residence and surname to classify African-Americans, Hispanics, and Jews . This particular Internet resource did not provide much in the way of supporting studies to justify these methods and thus may be of questionable use to the new user. However, an exhaustive search of the site found a citation to a publication about the use of surnames and contextual information to determine race (Lauderdale & Kestenbaum 2000)[17]. Clearly, a higher level of understanding of statistical methods would be necessary in order to interpret these types of resources effectively and to assist others in the selection of data or data resources for analysis.

Statistical Literacy

Iddo Gal (2002) made some observations on the components and attributes of statistical literacy[18]. Writing in the *International Statistical Review*, Gal looked at a person's ability to interpret and critically evaluate statistical information and his or her ability to discuss or communicate reactions to same. To that I add the ability to apply the method of analysis, that is to interpret the information in order to determine the best method of analysis. The continuum of skills runs from the ability to evaluate data and statistics to the ability to communicate

meaning and concerns.

Gal sees his definition of statistical literacy as requiring certain knowledge elements. These knowledge elements include literacy skills, statistical knowledge, mathematical knowledge, context knowledge, and critical questions. If we accept these elements as necessary components of statistical literacy, the question becomes one of degree. Questions include How much statistical knowledge must one have to be literate? How much mathematical knowledge is needed to be able to communicate about statistics? Should explanations of data quality and utility be framed in the scientific language of statistics?

Levels of Competence: The Standards for Success
In 1999 the Association of American Universities created a task force to "define the knowledge and skills students would need in their first year of college."[19] With support of the Pew Charitable Trusts, the Standards for Success project was launched in 2000 and published, and are available on the Internet (Conley 2003)[20]. For the social sciences, the standards include areas dealing with statistical literacy and state that incoming freshmen should know how to interpret data presented in tables and graphs, know the basics of probability theory and the concept of a sample, and know the difference between statistical and substantive significance. Successful students are expected to know how to find different sources of information and be able to analyze, evaluate the use them properly. Critical thinking is emphasized in evaluation based on quality of materials, credibility of information, presence of bias, and the ability to draw inferences. Students who plan on majoring in a social science are expected to have some basic knowledge of a statistical software package, and, depending on the field, understand and know how to use analytic tools, including basic statistics. Standards for statistical knowledge is included in the section on mathematics, although knowledge of statistics was not seen as a prerequisite for mathematical course work. Rather the knowledge of statistics was included for use in the social sciences, particularly economics, as important for entry level college work. These levels of competence could also work for librarians in the field of social science.

Understand and Use Summary Data
Summary data can be presented as counts, calculations, percentages, rates, or ratios in tables, graphs, histograms, polygons, pie charts, or maps. For data in print, statistical literacy would cover some types of evaluation and interpretation as well as the ability to formulate some opinion or concerns regarding the presentation. Much statistical information is presented as summary statistics, and exploratory data analysis – what one does prior to the actual tests, models, or other analytic tools – involves various summary statistics.

Interactive systems now allow us to create our own tables, graphs, charts, and begin the process of data analysis. Exploring the data might involve preparing summary statistics about individual variables.

The language of statistics refers to the examination of a single measure as univariate analysis. In this process, the distribution provides a sense of variance. The central tendency and dispersion also hint at the shape of the distribution. This might involve finding the maximum, minimum, mean, median, mode, and standard deviation. It should be possible to explain these measures using the language of statistics. For example, that the mode is useful where measures are expressed in categories, categorical variables such as 1=yes, 2=don't know, and 3=no. The category that occurs most frequently is the mode. For continuous variables, such as actual years of age, the mean or median is the statistic of choice. This is not calculus, but it is far from simple because the usefulness of each statistic depends upon the intended purpose as well as how the measure was constructed. But from this basic level, statistical processes can rapidly climb beyond simple mathematical computations. Today there are many online systems that allow users to manipulate data, create customized tables, and conduct higher level analysis. The popular online analysis tool, Survey Documenation and Analysis (SDA) allows users to determine the following measures:

- frequencies or cross tabulations
- comparison of means
- correlation matrix
- comparison or correlations
- multiple regression
- logit/probit
- statistics: Eta, R, Somer'd, Gamma, Tau-b, Tau-c, Chisq(P), Chisq (LR) dF

Here we get into advanced descriptive statistics and the statistics of relationships and a higher level of understanding is required. Even constructing and understanding a cross-tabulation can quickly become a source of confusion. Knowing which method can be used or should be used requires a greater practical knowledge of common use as well as some theoretical understanding. Although there are a number of online resources for students and teachers, including textbooks, glossaries and tutorials, the end-user probably wants some targeted, specific assistance when confronted with statistical information or data. It is my contention that this is best provided by a knowledgeable person who can engage the user in a structured dialog whereby he or she can determine the best statistic or measure based on "fitness for use." This human interaction is being abandoned in favor of

the development of online systems, probably with the belief that such systems lead to higher productivity and decreased cost. This is not to criticize those who promote improvements to online statistical resources, such as the UN and the ECE particularly when they emphasize the inclusion and organization of statistical metadata through publications and conferences, but rather to recognize that understanding of the need or use precedes the choice of statistic or method[21].

Online Help and the Digital Government Project

In an effort to help the general public understand the statistical information available from U.S. federal statistical agencies, the National Science Foundation and other federal statistical agencies have funded a number of projects that look at how statistical information is organized and presented. The GovStat project at the School of Library and Information Science of the University of North Carolina and the University of Maryland is multidimensional but part of the project involves creating some type of visual assistance for understanding statistical concepts.[22] This type of online help would be part of a larger architecture of information but it is uncertain as to how it translates into the correct use of statistics.

Conclusion

In addition to teaching students how to make help screens, library schools should teach statistical literacy. Librarians need to know to make assessments of quality and utility -- how to evaluate data and statistical publications. They should be able to furnish some guidance in interpretation of various types of statistical presentations and be able to point the way on how to interpret the results of analysis. As a group, they may never be statisticians, but they can perform a useful function in communicating meaning and concerns.

Statistical associations involved in statistical literacy have education programs aimed at high schools. Many online resources have been developed to teach statistics, but as Hans-Joachim Mittag noted, there is little "systematic cooperation." (Mittag 2000, p 6)[23]. There are many individual Internet sites devoted to teaching statistics to classroom students (Saporta 1999)[24]. There are very few aimed at data and statistical intermediaries, such as librarians. Sociometrics has a Data & Internet Literacy Series (Sociometrics 2005) that contains one section on understanding data in numbers, words, and pictures[25]. I have not seen it and cannot provide an evaluation. IASSIST would benefit from having some members qualified to teach and advise on the skill set we need to provide credible service in this area. The organization certainly has the ability to develop interactive, online, support to those who would be in the service of others.

Notes

[1] Contact: At the time this paper was presented, Ann S. Gray was the Data Reference Librarian at Princeton University Library.

[2] This paper is based on a talk given at the 2003 IASSIST Conference held in Ottawa on May 30, 2003.

[3] FEDSTATSs is an Internet Resource: http://www.fedstats.gov/

[4] DLI Update. Internet resource: http://www.statcan.ca.english/Dli/Document/update.htm

[5] STATS: Statistical Assessment Service, George Mason University, is an Internet Resource. http://www.stats.org/

[6] Olenski, Jozef (2004), "The Citizens' Right to Information and the Duties of a Democratic State in Modern IT Environment in the Light of the UN Fundamental Principles of Official Statistics and the ISI Declaration on Statistical Ethics," International Statistical Review, 71(1): 33-48.

[7] United Nations (1992), Economic Commission for Europe. "Fundamental Principles of Official Statistics in the Region of the Economic Commission for Europe", E1992/32 C(47) Available at http://www.unece.org/stats/archive/docs.fp.e.htm; United Nations (1993), Economic and Social Council. Statistical Commission, "Adoption of the Agenda and Other Organizational Matters: Fundamental Principles of Official Statistics." E/CN.3/1993/26; United Nations (1994), Statistics Division, "Fundamental Principles of Official Statistics" (1994), Washington, 1994.2.10. Available at: http://unstats.un.org/unsd/methods/statorg/FP-English.htm.

[8] ISI (2004), International Statistical Institute International Statistical Literacy Project (ISLP). Internet Resource: http://course1.winona.edu/cblumberg/islphome.htm.

[9] Statistics Canada (1998), "Statistics Canada Quality Guidelines", Third Edition, October 1998, Ottawa; Statisitics Canada (1998a), "Policy on Standards", Internet Resource: http://www.statcan.ca/english/concepts/policy-standards.htm.

[10] Statistics Canada (2000), "Policy on Informing Users of Data Quality and Methodology." Internet Resource: http://www.statcan.ca/english/about/policy-infousers.htm

[11] North American Industry Classification System (NAICS), was developed in cooperation with the U.S. Economic Classification Policy Committee, Statistics Canada, and Mexico's Instituto Nacional del Estadistica, Geografia e Informatica. See http://www.census.gov/epcd/www/naics.

html. In the U.S. it replaced the Standard Industrical Classification System (SIC). One version of the SIC codes and their meanings can be found in Standard Industrial Classification Manaual [rev. ed.], Washington, D.C. : Executive Office of the President, Office of Management and Budget; Springfield, Va. : National Technical Information Service, 1987 Standard Industrial Classification (SIC) System. See http://www.census.gov/epcd/www/sic.html.

[12] United States Census Bureau (2004), U.S. Census Bureau Section 515 Information Quality Guidelines. Internet Resource. http://www.census.gov/qdocs/www/quality_guidelines.htm

[13] Federal Register (2004), "Federal Statistical Organizations' Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Disseminated Information." 67(107): 38467. Available at: http://www.whitehouse.gov/omb/fedreg/reproducible.html/

[14] International Smoking Statisitics: a Collection of Historical Data from 30 Economically Developed Countries (2002), Forey, Barabara (ed.), London: Wolfson Institute of Preventive Medicine, Oxford: Oxford University Press.

[15] Complete Economic and Demographic Data Source (CEEDS) (2002), Woods & Poole Economics, Inc., Washington, D.C.

[16] Statewide Information by Assembly DistrictInternet Resource. http://swdb.berkeley.edu/info/statetext/staterpt.html#adstaterpt; Statewide Databases, Institute of Governmental Studies. University of California, Berkeley. Internet Resource. http://swdb.berkeley.edu/index.html

[17] Lauderdale, Diane S. and Bert Kestenbaum (2000), "Asian American ethnic identification by surname", Population Research and Policy Review (19(3) 283-300, June 2000.

[18] Gal, Iddo (2002), "Adults' Statistical Literacy: Meanings, Components, and Responsibilities", International Statistical Review 70(1):1-25.

[19] Standards for Success Organization (1999), Association of American Universities. Internet Resource: http://www.s4s,org/02_projectoverview/history.php

[20] Conley, David T. (2003) Understanding University Success.: a Report from Standards for Success: a Project of the Association of American Universities and the Pew Charitable Trusts, Eugene, Center for Educational Policy Research, 2003. Available at: http://www.s4s.org/03_viewproducts/ksus/ Mathematics: KSUS_math.pdf (page 10) ; Social Sciences: KSUS_social_sci.pdf.

[21] United Nations (2000), United Nations. Economic Commission for Europe. "Guidelines for Statistical Metadata on the Internet", UNSC/ECE-CES Statistical Standards and Studies, No. 52.

[22] GovStat (2005), University of North Carolina School of Library and Information Science, GovStat Program. Internet Resource: http://ils.unc.edu/govstat/

[23] Mittag, Hans-Hochim (2000), "Multimedia and Multimedia Databases for Teaching Statistics." Invited paper to be presented at the 9th International Conference on Mathematical Education, Makuhari/Tokyo, July 31 - August 6, 2000. Available at: http://www.stat.auckland.ac.nz/~iase/publications/10/ICME9_07.pdf

[24] Saporta, Gilbert (1999), "Teaching Statistics with Internet: a Survey of Available Resources and the ST@tNet Project". 52nd session of the International Statistical Institute, Helsinki, Finland. Available at: http://www.stat.auckland.ac/nx/~iase/publications/5/sapo0830.pdf

[25] Sociometrics (2005), "About Sociometrics Data & Internet Literacy Series" (DIL Series), Available at: http://www.socio.com/dil/